THE UNIVERSITY OF ADELAIDE

School of Agriculture and Wine
BiometricsSA

Scale Parameter Modelling of the t-distribution

Doctor of Philosophy
2005

Julian Taylor
Supervisor: Arunas P. Verbyla

Contents

Acknowledgements

Firstly, I would like to acknowledge and thank my supervisor, Ari Verbyla, for the interesting topic and the generous help throughout my candidature.

Thanks to Adelchi Azzalini for supplying and assisting with an example used in this thesis and a subsequent paper.

I would also like to thank the BiometricsSA staff and postgraduates for their helpful suggestions and comments.

I am also grateful to my fellow Adelaide University postgraduate students, David McInerney and Stuart Johnson for their support in the first stages of this degree.

Many thanks to my ex-girlfriend Megan Haynes for her love and support throughout the first part of my PhD. Similarly, I'm very grateful to my ex-girlfriend Allison Flaherty for her love, support and friendship over the last two years.

Many thanks to all my family and friends who have kept me motivated through the extreme bad times and have helped propel this thesis to its conclusion.

Peace out!

Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made.

I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

SIGNED: ..... **DATE:** .....

Abstract

This thesis considers location and scale parameter modelling of the heteroscedastic t-distribution. This new distribution is an extension of the heteroscedastic Gaussian and provides robust analysis in the presence of outliers as well accommodates possible heteroscedasticity by flexibly modelling the scale parameter using covariates existing in the data.

To motivate components of work in this thesis the Gaussian linear mixed model is reviewed. The mixed model equations are derived for the location fixed and random effects and this model is then used to introduce Restricted Maximum Likelihood (REML). From this an algorithmic scheme to estimate the scale parameters is developed.

A review of location and scale parameter modelling of the heteroscedastic Gaussian distribution is presented. In this thesis, the scale parameters are a restricted to be a function of covariates existing in the data. Maximum Likelihood (ML) and REML estimation of the location and scale parameters is derived as well as an efficient computational algorithm and software are presented.

The Gaussian model is then extended by considering the heteroscedastic t distribution. Initially, the heteroscedastic t is restricted to known degrees of freedom. Scoring equations for the location and scale parameters are derived and their intimate connection to the prediction of the random scale effects is discussed. Tools for detecting and testing heteroscedasticity are also derived and a computational algorithm is presented. A mini software package "hett" using this algorithm is also discussed.

To derive a REML equivalent for the heteroscedastic t asymptotic likelihood theory is discussed. In this thesis an integral approximation, the Laplace approximation, is presented and two examples, with the inclusion of ML for the heteroscedastic t, are discussed. A new approximate integral technique called Partial Laplace is also discussed and is exemplified with linear mixed models. Approximate marginal likelihood techniques using Modified Profile Likelihood (MPL), Conditional Profile Likelihood (CPL) and Stably Adjusted Profile Likelihood (SAPL) are also presented and offer an alternative to the approximate integration techniques.

The asymptotic techniques are then applied to the heteroscedastic t when the degrees of freedom is known to form two distinct REMLs for the scale parameters. The first approximation uses the Partial Laplace approximation to form a REML for the scale parameters, whereas, the second uses the approximate marginal likelihood technique MPL.

For each, the estimation of the location and scale parameters is discussed and computational algorithms are presented. For comparison, the heteroscedastic t for known degrees of freedom using ML and the two new REML equivalents are illustrated with an example and a comparative simulation study.

The model is then extended to incorporate the estimation of the degrees of freedom parameter. The estimating equations for the location and scale parameters under ML are preserved and the estimation of the degrees of freedom parameter is integrated into the algorithm. The approximate REML techniques are also extended. For the Partial Laplace approximation the estimation of the degrees of freedom parameter is simultaneously estimated with the scale parameters and therefore the algorithm differs only slightly. The second approximation uses SAPL to estimate the parameters and produces approximate marginal likelihoods for the location, scale and degrees of freedom parameters. Computational algorithms for each of the techniques are also presented. Several extensive examples, as well as a comparative simulation study, are used to illustrate ML and the two REML equivalents for the heteroscedastic t with unknown degrees of freedom.

The thesis is concluded with a discussion of the new techniques derived for the heteroscedastic t distribution along with their advantages and disadvantages. Topics of further research are also discussed.

Chapter 1

Introduction

1.1 Background

Homoscedasticity of the scale parameter is a common assumption in linear and non-linear models. When the assumption of constant scale is violated the scale parameter is said to be heterogeneous. One approach to rectify this violation is to transform the response variable. For example, if the response is non-normally distributed a transformation can sometimes be chosen to ensure a Gaussian assumption of the residuals. Examples of this include the square root stabilizing transformation for a Poisson distributed response and an arcsine transformation of a response that is Binomially distributed. More general families of transformations that allow a degree of flexibility also exist (see Box & Meyer, 1986). A more comprehensive overview and discussion of methods to select the appropriate transformation can be found in Cook & Weisberg (1982). The applicability of these transformations, however, may be questionable if the location component of the model still depends on the scale. Furthermore, the normality and additivity of the model may be modified increasing the complexity of the analysis and interpretation.

Another type of heterogeneity, closely connected to this thesis, is when the variance is modelled as a function of the location component of the model. This assumption is a requirement for the class of Generalized Linear Models (see McCullagh & Nelder, 1989) used to parametrically model selected types of non-normal data. These models are not considered here.

In some cases modelling of the variance or scale parameter may be of some interest. In this thesis a flexible approach to account for the heterogeneity is considered where the scale parameter is modelled using covariates existing in the data. Previous research in this area has been confined to exponential families (see Smyth, 1989) and, in particular, much research has focussed on heteroscedastic regression using the Gaussian distribution (see

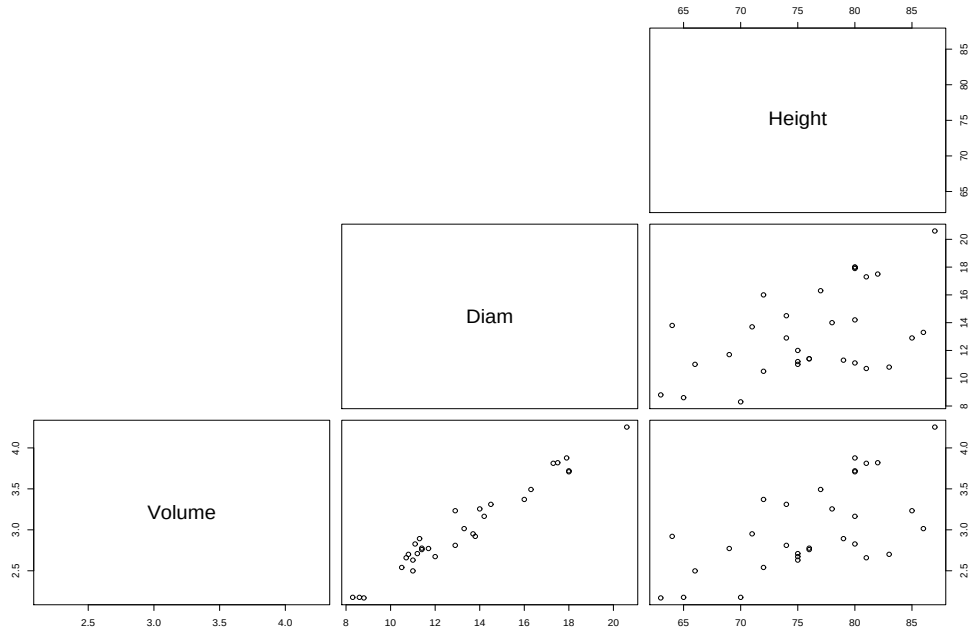

Figure 1.1: Pairs plot of the cube root of Volume with two explanatory variables, Diameter and Height

Aitkin, 1987; Verbyla, 1993; Smyth et al., 2001 and Smyth, 2002). This model, although flexible, may still be deficient in some cases. For example, outliers may be present in the data that cannot be accommodated by the heterogeneous scale parameter model. For the heteroscedastic Gaussian, Verbyla (1993) presents a comprehensive overview of diagnostics for determination of both location and scale parameter outliers. In some cases the outliers may be genuine and then their accommodation, rather than deletion, is important and a more robust modelling approach is required.

This thesis discusses a robust extension to the heteroscedastic Gaussian by considering scale parameter modelling of the t-distribution. Identical to the Gaussian equivalent, the location and scale parameters are modelled using covariates existing in the data providing a flexible and computationally efficient methodology for handling robust data as well as accommodating possible heteroscedasticity.

1.2 Motivation

1.2.1 Cherry Trees

The cherry tree data, Ryan et al. (1985), consists of volume measurements of harvestable timber from 31 cherry trees. The explanatory variables are the measured diameter and

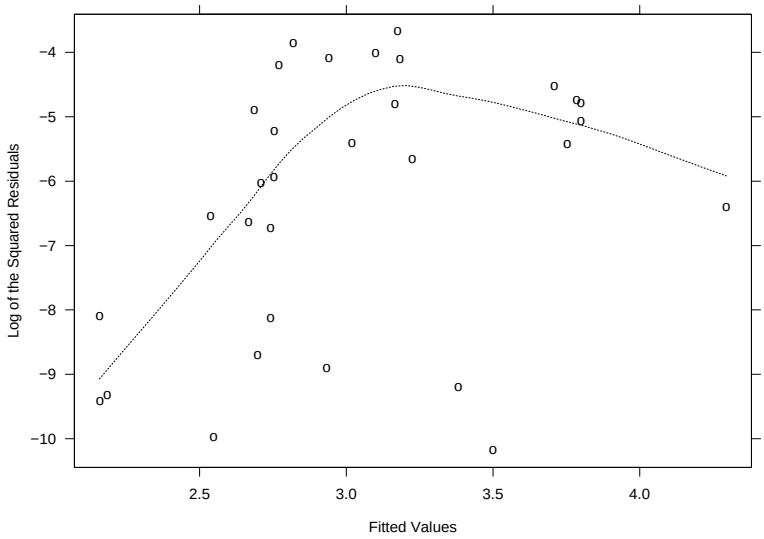

Figure 1.2: Scatter plot of the log of the squared residuals against the fitted values after fitting an additive location model. The dotted line is a local smoother to describe trend.

height of the trees. Early analysis (see Cook & Weisberg, 1982) suggests the cube root of volume is an appropriate transformation of the response. Figure 1.1 shows the transformed response with the two explanatory variables, Diameter and Height. As expected, the diameter and height are positively related to the cube root of the volume. Therefore an additive location model in Diameter and Height may be appropriate.

After fitting this simple location model the residuals can be explored. Figure 1.2 shows the log of the squared residuals against the fitted values from the additive location model. The dotted line is a local smoother to describe trend. The transformed residuals appear to be quadratically related to the location component of the model suggesting possible heterogeneity. The plot also displays potential outliers suggesting a robust approach to modelling of the location and scale parameters. These models are explored further in Chapter 7.

1.2.2 Rent for Land Planted With Alfalfa

In this example the relationship between the rent for agricultural land planted with alfalfa crops and density of cows from 67 counties of Minnesota in the year 1977 (see Weisberg, 1985) is investigated. The data set contains the average rent per acre for the land planted with alfalfa, average rent for all agricultural uses, density of cows per square mile (SC), proportion of pastoral land and whether liming was required on the land (LI). As alfalfa is a possible feed for dairy cows it is proposed that the rent for the land planted with alfalfa

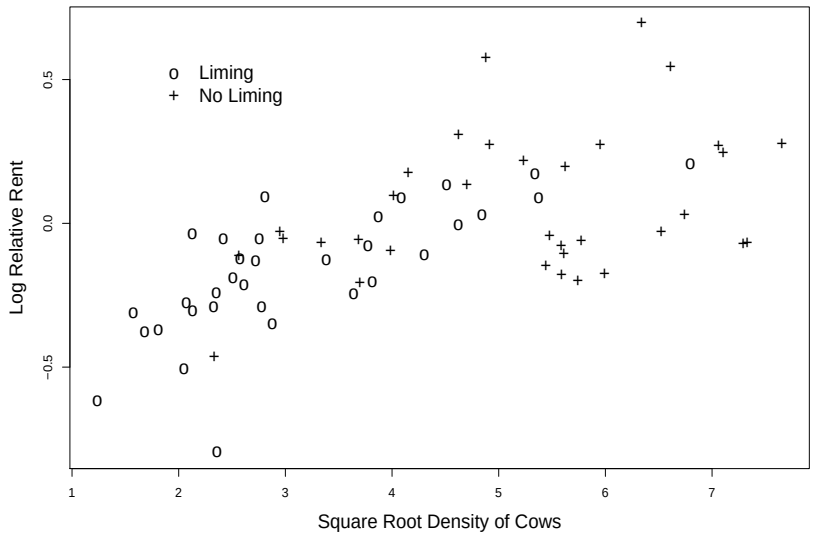

Figure 1.3: Scatter plot of the log relative rent of the land planted with alfalfa against the square root of the dairy cow density per square mile.

relative to the rent for the other agricultural uses would be higher in high density dairy cow areas. It is also proposed that this relative rent would also be lower in areas where liming for alfalfa has occurred due to the additional expenses incurred. The natural log of the relative rent against the square root of the density of cows is displayed in Figure 1.3.

A positive relationship is evident as well as heteroscedasticity that changes for each Liming type. One approach to model the location component would be to consider an interaction model for the square root density of the cows and each Liming type. The expected high correlation between the proportion of pastoral land and the density of the cow population allows the exclusion of pastoral land as a possible influencing covariate in the location parameter model. After fitting this model the residuals are presented in Figure 1.4. The skewness of the distributions suggests possible linear or non-linear heteroscedastic components in the scale component of the model. These models are explored further in Chapter 9.

1.2.3 Stack Loss data

The Brownlee (1965) stack loss data has been examined many times using various robust methods. The data consists of oxidation rates of ammonia from plants measured over a 21 day period. Possible influential explanatory variables, Water Temperature (WT), Air Flow (AF) and Acid Concentration (AC) of the ammonia were also measured to help

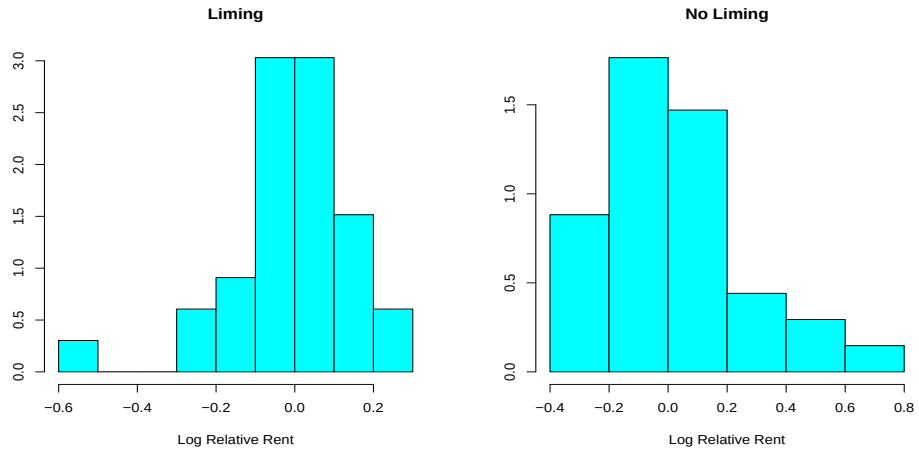

Figure 1.4: Histograms of the residuals for Liming and No Liming after fitting an interaction location model of log relative rent against square root of the dairy cow density per square mile for each Liming type.

explain the variation in the oxidation rate. A pairs plot of the stack loss against the three explanatory variables is given in Figure 1.5. The plot shows there is a positive relationship between the oxidation rate of the plant and the three measured explanatory variables. The variables also exhibit noticeable correlations between each other.

Previous methods of modelling the stack loss data include M-estimation by Andrews (1974) and Huber (1981) as well as trimmed least squares utilised by Ruppert & Carroll (1980). Nelder (2000) proposes a non-robust method by adopting a generalized linear model that indicates there are no outliers in the data. In an illuminating paper, Lange et al. (1989) suggests that the response may be t-distributed and analyse the data over a range of degrees of freedom. For this example, the t-regression model of Lange et al. (1989) is extended to show that the scale parameter is heteroscedastic in at least one of the explanatory variables.

1.2.4 Martin Marietta Data

The relationship of the excess rate of returns of the Marietta company and an index for the excess rate of return for the New York Exchange (CRSP) is the subject of this motivating example. Both the rate of returns for the company and the CRSP index were measured monthly over a period of five years and a scatterplot is provided in Figure 1.6. The plot suggests that there is a positive linear increase in the excess returns of

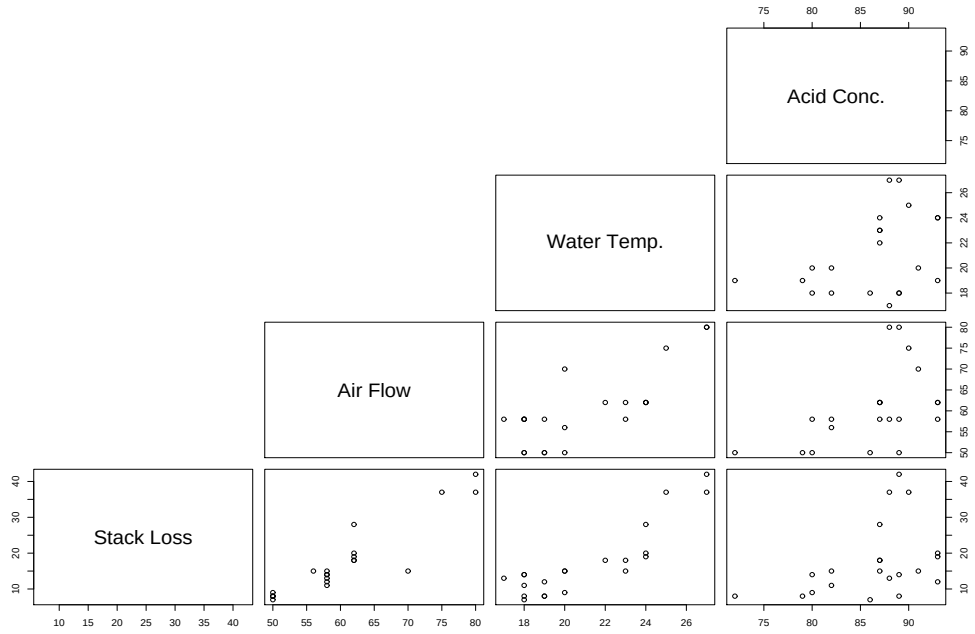

Figure 1.5: Pairs plot of the stack loss of ammonia against three explanatory variables; Air Flow, Water Temperature and Acid Concentration.

the company when the CRSP index is higher. At the highest CRSP index the company excess returns were more than twice the returns of any other period over the five years. This outlier cannot be rejected and therefore a robust approach to analysis is required.

After a least squares or homoscedastic Gaussian fit to the data the residuals are inspected and displayed in Figure 1.7. With the influence of an outlier the distribution of the residuals display heavy tails and skewness to the right. To allow for this several modelling approaches are possible. Jones (2001) and Azzalini & Capitanio (2003) have suggested a skew t-distribution, an extension of the skew normal distribution researched by Azzalini & Capitanio (1999). This approach is flexible when the residuals have a genuine pattern of skewness. In this thesis an alternative flexible approach is discussed where the skewness pattern of the residuals is modelled parametrically through the scale parameter of the t-distribution.

1.3 Literature Review

Modelling the scale parameter using covariates available in the data has been discussed in many areas of applied statistics. In the econometric literature, Park (1966) proposes a log-linear model for the scale parameter and describes the Gaussian model using a two-stage process to estimate the parameters. Harvey (1976) discusses Maximum Likelihood

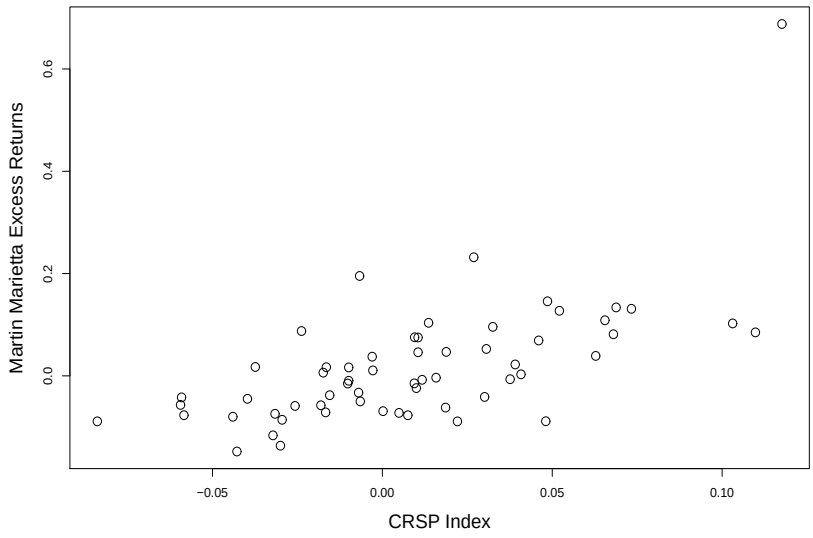

Figure 1.6: Scatter plot of the Martin Marietta company excess returns against the CRSP index for the whole market.

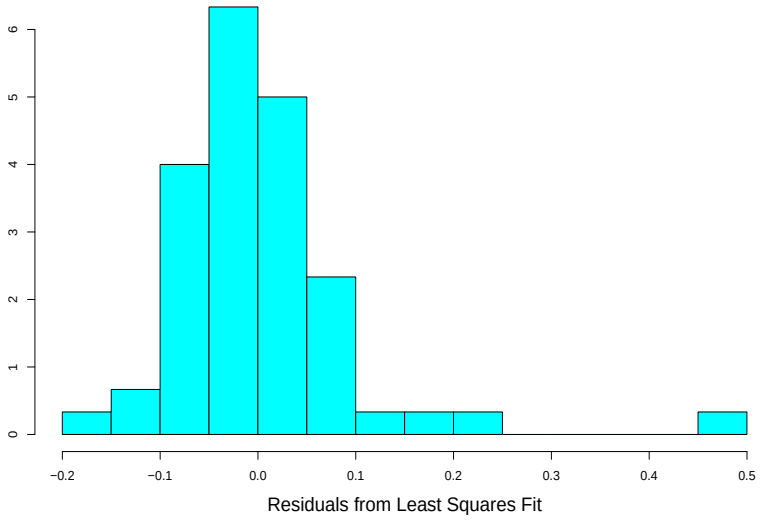

Figure 1.7: Histogram of the residuals from a Least Squares fit of Martin Marietta company excess returns against the CRSP index for the whole market.

(ML) estimation of the location and scale parameters and the subsequent likelihood ratio test under general conditions. Aitkin (1987) provides ML estimation for a joint location and scale model and applies it to the cherry tree data of Section 1.2.1. More recently scale parameter models have been utilised in industrial statistics for unreplicated experiments and process control (see Carroll & Ruppert, 1988; Nelder & Lee, 1991; Engel & Huele,

1996 and Lee & Nelder, 1998).

The Gaussian location and scale model can be extended in a number of ways. Verbyla (1993) estimates the parameters using Restricted Maximum Likelihood (REML) and provides leverage and influence diagnostics for ML and REML. This extension provides estimates of the scale parameters with reduced bias and allows the standard errors of both the location and scale effects to be determined more accurately. This is a current area of research and key references include Huele (1998), Huele & Engel (1998), Smyth et al. (2001) and Smyth (2002). In the latter two of these papers a more general form for the scale parameter is assumed. In particular, Smyth (2002) introduces a stepwise procedure for the efficient calculation of the components of the scoring algorithm to estimate the scale parameters. More general distributions from the family of generalized linear models are considered by Smyth (1989), Nelder & Lee (1991) and Smyth & Verbyla (1999). In these papers the location and scale parameters of the distribution are estimated using double generalized linear models. A more flexible approach is considered by Rigby & Stasinopoulos (1996a) and Rigby & Stasinopoulos (1996b) where the heterogeneity is modelled semi-parametrically using smoothing splines.

It is common for the observables to contain outliers. If present, the Normality assumption of the error distribution for the model is questionable and estimates of the parameters may be misleading. Robust approaches to regression and outlier detection have an extensive literature (see Huber, 1981; Cook & Weisberg, 1982; Atkinson, 1985 and Atkinson & Riani, 2000). If the outliers are considered to be genuine then their accommodation, rather than deletion, is important and can sometimes be achieved with a t-distribution. Using the t specification to model the observables has widely been considered a useful tool for robustifying an analysis. Fraser (1976), Fraser (1979) and West (1984) describe examples of its use with simple general univariate linear models. For the multivariate t, Rubin (1983) and Lange et al. (1989) discuss and exemplify the possibility of allowing for correlation between observations. These important contributions discuss the estimation process for a simple location and homogeneous scale parameter using the Expectation-Maximisation (EM) algorithm. Little (1988) extends this multivariate approach by allowing for missing values present in the response. These missing values can then be imputed by nesting them in a modified version of the EM algorithm. Liu & Rubin (1994), Liu & Rubin (1995) and Liu (1995) extend the EM algorithm further using the multivariate t-distribution as basis for their derivations. Liu (1997) and Meng & van Dyk (1997) also exemplify its use to coalesce the many modifications of the EM algorithm. Meng & van Dyk (1997) describe a technique of expanding the parameter space whilst using the EM algorithm (PX-EM) and apply this extension to the t-distribution. The PX-EM algorithm was then formalised by Liu et al. (1998) and its efficiency gains recognised for the multivariate t.

Extensions of the t-distribution are also available. James et al. (1993) discusses t-REML, a Restricted Maximum Likelihood approach to estimation of the parameters of the t-

distribution. In this paper the location parameters are approximately conditioned out of the marginal likelihood. It is found that this approach is equivalent to conditional profile likelihood (see Cox & Reid, 1987). Welsh & Richardson (1997) discusses various approaches of robust estimation using the t-distribution with the inclusion of random effects in the location component of the model. These models are also presented in Pinheiro et al. (2001), where the random effects are also distributed as t and therefore the potential presence of outliers in both the random effects and the errors is accommodated by a marginal t-distribution. The estimation of parameters is restricted to ML.

More recent advances include modelling using the multivariate skew t-distribution (see Jones, 2001). Key references in this area include Azzalini & Capitanio (1999), Jones & Faddy (2003) and Azzalini & Capitanio (2003).

In this thesis the t-distribution is extended by the inclusion of heteroscedasticity in the scale parameter. The ML estimation and inference of the heteroscedastic t-distribution is similar to Verbyla (1993). In particular, when the degrees of freedom is known, the heteroscedastic t models and the simpler heteroscedastic Gaussian models from Verbyla (1993) are derived from the *location-scale family*. The models diverge when the degrees of freedom is unknown. Similar models may also be derived using generalized additive models discussed in Rigby & Stasinopoulos (2005). A comprehensive overview of ML estimation of the parameters of the heteroscedastic t-distribution can be found in Taylor & Verbyla (2004).

To extend the t-distribution further two new approaches of obtaining an approximate t-REML when the scale parameter is heteroscedastic are presented. Firstly, the Laplace approximation is used to obtain an approximate marginal likelihood for the heteroscedastic t-distribution. This type of approximation was derived originally by Erdelyi (1956) and De Bruijn (1961) for numerical analysis. In the statistical literature, the Laplace approximation was found useful in many areas. In Bayesian statistics Lindley (1980), Tierney & Kadane (1986) and Tierney et al. (1989) discuss its use in evaluating posterior means and variances of parameters of interest. Solomon & Cox (1992), Wolfinger (1993), Breslow & Clayton (1993), McGilchrist (1994) and Engel & Keen (1994) use this method to approximate the marginal likelihood for non-linear and generalized linear mixed models. In these papers the location random effects are considered to be nuisance parameters that require integrating out of a pseudo joint likelihood. This area of research has flourished recently. Breslow & Lin (1995) and Lin & Breslow (1996) extend the work of Breslow & Clayton (1993) by refining the estimates obtained from the approximate marginal likelihood. Shun & McCallaugh (1995) and Shun (1997) obtain a more accurate representation of the Laplace approximation to the marginal likelihood for generalised linear mixed models by considering higher order terms. This has also been considered by Raudenbush et al. (2000) for a logistic mixed model but is easily generalised to other distributions of the exponential family. To obtain an approximate REML for the scale

parameters of the heteroscedastic t-distribution an extension of Laplace's method called Partial Laplace is used. This new technique exploits the component form of the integrand and allows the approximate marginal likelihood to be partitioned for separate estimation of the location and scale parameters.

Secondly, adjusted likelihood techniques are used to modify the marginal likelihood due to estimation of the nuisance parameters. An extensive overview of these techniques can be found in Barndorff-Nielsen & Cox (1989) and Barndorff-Nielsen & Cox (1994). The techniques considered here include Modified Profile Likelihood derived by Barndorff-Nielsen (1980) and Barndorff-Nielsen (1983). Connected to this is Conditional Profile Likelihood considered by Cox & Reid (1987). Both of these approaches require an ancillary statistic to be available to derive the modification terms. The latter of these two approaches assume an orthogonalization of the nuisance parameter and the parameter of interest to reduce the adjustment terms required (see the discussion of Cox & Reid, 1987). Comparisons of the two approximate marginal likelihood approaches has been discussed in Cox & Reid (1987), Cox & Reid (1992) and Barndorff-Nielsen & McCullagh (1993). Approximations to the Modified Profile Likelihood have also been derived by Severini (1998). An extension of Modified Profile Likelihood considered in this thesis is Stably Adjusted Profile Likelihood. This technique assumes no ancillary statistic is available and therefore allows more complex distributions to be modified in the presence of nuisance parameters. Important references in this area of research include Barndorff-Nielsen (1994), Barndorff-Nielsen & Chamberlain (1994) and Barndorff-Nielsen & Cox (1994). Stern (1997) discusses a more accurate extension of the Stable Adjusted Profile Likelihood by considering a higher order expansion of the modification term.

1.4 Outline

To motivate proceeding chapters of this thesis, Chapter two reviews simple Gaussian linear mixed models. The formulation of the marginal likelihood is discussed and the Mixed Model Equations (MME) are derived for the location fixed and random effects. To obtain an independent objective function for the variance (scale) parameters, the marginal likelihood is partitioned to form a Restricted Maximum Likelihood (REML). From this an iterative algorithm is derived to estimate the scale parameters.

Chapter three presents review theory on joint modelling of the location and scale parameters of the Gaussian distribution. Under a very general scale parameter model, ML and REML techniques are used to obtain an algorithm with which to jointly estimate the parameters. An asymptotic hypothesis test for heteroscedasticity is derived and a computational algorithm to efficiently estimate the parameters using ML and REML is supplied.

Joint modelling of the location and scale parameters of the heteroscedastic t-distribution when the degrees of freedom is known is the subject of the fourth chapter. This chapter is restricted ML estimation of the location and scale parameters. Firstly, an overview of the t-distribution and its connection to other distributions is provided. Scoring equations for the location and scale parameters are determined for the heterogeneous case when the degrees of freedom is known. An extension of the scoring algorithm is derived to incorporate a mechanism to estimate the degrees of freedom parameter and its convergence properties discussed. The prediction of the random scale effects is presented and parameter inference is discussed. Asymptotic properties and tests are derived for the location and scale parameters and a tool for determining heteroscdasticity is presented. The chapter is concluded with a computational algorithm to estimate the parameters and a description of the associated software.

Chapter five provides and overview of current asymptotic likelihood techniques required in proceeding chapters of this thesis. In particular, the Laplace approximation to an intractable integral is described for single and multidimensional nuisance parameters along with an example of its use for linear mixed models and the heteroscedastic t-distribution. An extension of the integral approximation that exploits the possible component form of the integrand, called the Partial Laplace approximation, is also derived. The likelihoods associated with the linear mixed model under REML are derived to exemplify its use. Modified profile likelihood (MPL) and its extensions are discussed. In particular, the Conditional Profile Likelihood (CPL) and Stably Adjusted Profile Likelihood (SAPL) are described along with their connection to MPL.

The sixth chapter uses the approximate likelihood techniques of the previous chapter to derive two alternate REML equivalents for the heteroscedastic t-distribution when the degrees of freedom is known. Firstly, the Partial Laplace approximation is used to obtain an approximate marginal likelihood which may be partitioned to estimate the location and scale parameters separately. Prediction of the random scale effects and an estimating equation for the location parameters is derived. A second approximate t-REML is derived using MPL. The connection of the t with known degrees of freedom to the location-scale family of distributions allows simplifications to occur and a simple computational algorithm is discussed.

The seventh chapter contains an example and simulations using the heteroscedastic t-distribution under ML and the approximate t-REML methods derived in chapter four and six respectively. The example compares the heteroscedastic Gaussian and t under ML for a variety of scale parameter models. Similarly, comparisons of heteroscedastic Gaussian and t under REML are performed and compared to their ML equivalents. Simulations are also performed to understand the properties of the location and scale estimators using ML and t-REML. Empirical means and standard errors for the parameters are provided and, for the latter case, compared to theoretical standard errors. Performance plots from

the simulations are also presented to display the bias adjustment to the scale parameter estimates obtained from the approximate t-REML methods.

The eighth chapter focuses on joint modelling of the location and scale parameters of the heteroscedastic t-distribution when the degrees of freedom is unknown. This extends the results of chapter four and six. Under ML, a parameter orthogonalization is derived to ensure the computational algorithm to estimate the location and scale parameters derived in chapter four can be maintained. Estimation of the degrees of freedom parameter is then appended to this algorithm. The Partial Laplace approximation to obtain a t-REML equivalent is discussed and requires only a small adjustment to the approximate t-REML derived by Partial Laplace in chapter six. A second t-REML is also derived using Stably Adjusted Profile Likelihood (SAPL). For each parameter of interest the marginal profile likelihood is adjusted for the nuisance parameters to form a set of Stably Adjusted Profile Likelihoods.

To illustrate the extensions derived in the previous chapter, chapter nine presents the analysis of several examples data sets. Comparisons between the ML and t-REML models are presented where possible and, for one example, profiling of the scale and degrees of freedom parameters is shown for the more complex t-REML methods derived in this thesis. A simulation study is also conducted to understand the properties of the parameters for the ML and approximate t-REML approaches of the previous chapter. Identical to chapter seven, empirical means and standard errors are presented for all methods and, for the latter, compared to the ML theoretical standard errors. Performance plots are presented for both approximate t-REML methods to determine the bias adjustment to the estimated scale parameters obtained from using a REML construct.

The thesis concludes with a summary and discussion of the advantages and disadvantages of the ML and approximate REML approaches derived for the heteroscedastic t-distribution as well as topics for further research.

Chapter 2

Linear Mixed Effects Models

This chapter describes a class of statistical models called linear mixed models. In ordinary linear models the explanatory variables are considered to be fixed, whereas, in mixed models some variables can be considered to have been sampled from a larger population with an assumed known distribution. Typically, these new variables are called *random effects*.

The mixed models considered in this chapter are used to motivate the forthcoming chapters of this thesis. In particular, Section 2.3 discusses Restricted Maximum Likelihood (REML). This methodology is used estimate the scale parameters associated with the random effects or error component of the linear mixed model whilst allowing for the loss of degrees of freedom when estimating the location parameters. In the forthcoming chapters this technique is applied to various models including the heteroscedastic Gaussian and the heteroscedastic t-distribution. This chapter also motivates two theoretical examples presented in Chapter 5 Section 5.1. These examples discuss the derivation of the marginal likelihood for the linear mixed models using an integration technique called the Laplace approximation. These derivations are shown to be equivalent to the derivations considered in this chapter.

2.1 Introduction and Notation

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a vector of responses and consider the general linear model defined by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \tag{2.1.1}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ to be $n \times p$ and $n \times r$ matrices of explanatory variables respectively, $\boldsymbol{\beta}$ is vector of length p of unknown parameters and the joint distribution of $(\boldsymbol{u}, \boldsymbol{e})$ is given by

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G}(\boldsymbol{\varphi}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}(\boldsymbol{\gamma}) \end{bmatrix} \right) \tag{2.1.2}$$

where $\boldsymbol{\varphi}$ and $\boldsymbol{\gamma}$ are scale parameters associated with $\boldsymbol{u}$ and $\boldsymbol{e}$ respectively. Generally, the random effects are of the form $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q)$ where $\boldsymbol{u}_i$ is a vector of length $r_i \times 1$ representing the ith random effect. The associated design matrices are then of the form $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_q)$. Following this, $\mathrm{Var}[\boldsymbol{u}_i] = \boldsymbol{G}_i$ for $i = 1, \ldots, q$ and, generally, $\mathrm{Cov}(\boldsymbol{u}_i, \boldsymbol{u}_j) = \boldsymbol{0}$ for all $i \neq j,\ i, j = 1, \ldots, q$ suggesting that $\boldsymbol{G}$ is block diagonal with ith matrix $\boldsymbol{G}_i$.

The form of the matrix $\boldsymbol{R}$ varies according to the dependence between the observations. For example, in a multi-site spatial analysis of a field trial, similar to the variance matrix for the random effects, $\mathrm{Var}[\boldsymbol{e}] = \mathrm{diag}(\boldsymbol{R}_j),\ j = 1, \ldots, s$ where s is the number of defined sites used in the experiment.

In this thesis a simplified model is used where $q = 1$ and $s = 1$.

2.1.1 Marginalising the Likelihood

Under the assumptions of the previous section the conditional distribution of $\boldsymbol{y}|\boldsymbol{u}$ and marginal distribution of $\boldsymbol{u}$ can be expressed as

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{u} &\sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \boldsymbol{R}) \\ \boldsymbol{u} &\sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G}) \end{aligned}$$

Let $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\varphi})$ be the vector of scale parameters from the matrices $\boldsymbol{R}$ and $\boldsymbol{G}$ respectively. The marginal likelihood can be expressed as the product of the conditional probability density function of $y_i|\boldsymbol{u},\ i = 1, \ldots, n$ and the probability distribution function of $\boldsymbol{u}$ integrated over the range of the unobserved random effects, namely

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}) = \int_{\mathcal{R}^r} p(\boldsymbol{y}|\boldsymbol{u}; \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{u}; \boldsymbol{\varphi}) d\boldsymbol{u}$$

where $\mathcal{R}^r$ is an r-dimensional subspace of $\mathcal{R}$ and

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{u}; \boldsymbol{\beta}, \boldsymbol{\theta}) &= (2\pi)^{-n/2} |\boldsymbol{R}|^{-1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^T \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})\} \\ p(\boldsymbol{u}; \boldsymbol{\varphi}) &= (2\pi)^{-r/2} |\boldsymbol{G}|^{-1/2} \exp\{-\tfrac{1}{2}\boldsymbol{u}^T \boldsymbol{G}^{-1} \boldsymbol{u}\} \end{aligned} \tag{2.1.3}$$

The marginal likelihood can then be expressed as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}) &= (2\pi)^{-(n+r)/2} |\boldsymbol{R}|^{-1/2} |\boldsymbol{G}|^{-1/2} \\ &\times \int_{\mathcal{R}^r} \exp\{ -\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^T \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) - \tfrac{1}{2}\boldsymbol{u}^T \boldsymbol{G}^{-1} \boldsymbol{u}\} d\boldsymbol{u} \end{aligned} \tag{2.1.4}$$

Let

$$\boldsymbol{y}^* = (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y}$$
$$\boldsymbol{X}^* = (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}$$

be transformations of the response vector and the explanatory design matrix for the location effects respectively. Then the marginal likelihood can be written as

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y}) &= (2\pi)^{-(n+r)/2}|\boldsymbol{R}|^{-1/2}|\boldsymbol{G}|^{-1/2} \\ &\times \exp\{-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\} \\ &\times \int_{\mathcal{R}^r}\exp\{-\tfrac{1}{2}(\boldsymbol{u}-(\boldsymbol{y}^*-\boldsymbol{X}^*\boldsymbol{\beta}))^T(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1})(\boldsymbol{u}-(\boldsymbol{y}^*-\boldsymbol{X}^*\boldsymbol{\beta}))\}d\boldsymbol{u} \end{aligned}$$

The term in the exponent of the integrand is the standard quadratic form for $\boldsymbol{u}$ with expectation $\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}$ and variance $(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1})$. Integrating over these effects and using Result A.3.1 allows the marginal likelihood given in (2.1.4) to be reduced to

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y}) &= (2\pi)^{-n/2}|\boldsymbol{G}|^{-1/2}|\boldsymbol{R}|^{-1/2}|\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1}|^{-1/2} \\ &\times \exp\{-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\} \end{aligned}$$

where $\boldsymbol{H} = (\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{R})$. This can be simplified again by amalgamating the determinants

$$\begin{aligned} & |\boldsymbol{G}|^{-1/2}|\boldsymbol{R}|^{-1/2}|\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1}|^{-1/2} \\ = \;& |\boldsymbol{R}|^{-1/2}|\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{G}+\boldsymbol{I}_r|^{-1/2} \\ = \;& |\boldsymbol{R}|^{-1/2}|\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T+\boldsymbol{I}_n|^{-1/2} \qquad \text{(using Result A.2.2)} \\ = \;& |\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T+\boldsymbol{R}|^{-1/2} \end{aligned}$$

The final form for the marginal log-likelihood can be expressed as

$$\ell(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{H}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \qquad (2.1.5)$$

so that $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{H})$. A comparative approach using the Laplace approximation is presented in Section 5.1.2.

2.2 Prediction and Estimation

Suppose for given $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ we wish to predict $\boldsymbol{a}_1^T\boldsymbol{\beta} + \boldsymbol{a}_2^T\boldsymbol{u}$ using a linear function of the data $\boldsymbol{y}$, that is by $\boldsymbol{\alpha}^T\boldsymbol{y}$ for some $\boldsymbol{\alpha}$. The $\boldsymbol{\alpha}$ is chosen such that the predictor has minimum mean square error (MMSE) among a class of unbiased linear predictors.

The unbiasedness constraint can be expressed as

$$\begin{aligned} & \mathrm{E}[\boldsymbol{\alpha}^T\boldsymbol{y}] = \mathrm{E}[\boldsymbol{a}_1^T\boldsymbol{\beta} + \boldsymbol{a}_2^T\boldsymbol{u}] \\ \Rightarrow \quad & \boldsymbol{\alpha}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{a}_1^T\boldsymbol{\beta} \\ \Rightarrow \quad & \boldsymbol{X}^T\boldsymbol{\alpha} = \boldsymbol{a}_1 \end{aligned} \tag{2.2.1}$$

The mean square error (MSE) is defined as

$$\begin{aligned} MSE &= \mathrm{E}[(\boldsymbol{\alpha}^T\boldsymbol{y} - \boldsymbol{a}_1^T\boldsymbol{\beta} - \boldsymbol{a}_2^T\boldsymbol{u})^2] \\ &= \mathrm{E}[(\boldsymbol{\alpha}^T\boldsymbol{y} - \boldsymbol{a}_2^T\boldsymbol{u})^2] + \mathrm{E}[(\boldsymbol{a}_1^T\boldsymbol{\beta})^2] - 2\mathrm{E}[\boldsymbol{\alpha}^T\boldsymbol{y} - \boldsymbol{a}_2^T\boldsymbol{u}]\mathrm{E}[\boldsymbol{a}_1^T\boldsymbol{\beta}] \\ &= \mathrm{Var}[\boldsymbol{\alpha}^T\boldsymbol{y} - \boldsymbol{a}_2^T\boldsymbol{u}] + (\mathrm{E}[\boldsymbol{\alpha}^T\boldsymbol{y} - \boldsymbol{a}_2^T\boldsymbol{u}])^2 + \mathrm{E}[(\boldsymbol{a}_1^T\boldsymbol{\beta})^2] - 2\mathrm{E}[\boldsymbol{\alpha}^T\boldsymbol{y} - \boldsymbol{a}_2^T\boldsymbol{u}]\mathrm{E}[\boldsymbol{a}_1^T\boldsymbol{\beta}] \\ &= \mathrm{Var}[\boldsymbol{\alpha}^T\boldsymbol{y} - \boldsymbol{a}_2^T\boldsymbol{u}] + (\boldsymbol{a}_1^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{a}_1^T\boldsymbol{\beta})^2 \\ &= \mathrm{Var}[\boldsymbol{\alpha}^T\boldsymbol{y} - \boldsymbol{a}_2^T\boldsymbol{u}] \qquad \text{(using (2.2.1))} \end{aligned}$$

Noting that $\mathrm{Cov}[\boldsymbol{y}, \boldsymbol{u}] = \boldsymbol{Z}\boldsymbol{G}$ the MSE can be expressed as

$$MSE = \boldsymbol{\alpha}^T\boldsymbol{H}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T\boldsymbol{Z}\boldsymbol{G}\boldsymbol{a}_2 + \boldsymbol{a}_2^T\boldsymbol{G}\boldsymbol{a}_2$$

Under the constraint of unbiasedness the function that requires minimisation uses Lagrange Multipliers and can be expressed as

$$CM = \boldsymbol{\alpha}^T\boldsymbol{H}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T\boldsymbol{Z}\boldsymbol{G}\boldsymbol{a}_2 + \boldsymbol{a}_2^T\boldsymbol{G}\boldsymbol{a}_2 + 2\boldsymbol{\lambda}^T(\boldsymbol{a}_1 - \boldsymbol{X}^T\boldsymbol{\alpha}),$$

where $\boldsymbol{\lambda}$ is a $n \times 1$ vector of Lagrange Multipliers. To minimise this Lagrange equation the derivatives with respect to $\boldsymbol{\alpha}$ and the Langragian multipliers, $\boldsymbol{\lambda}$ are required, namely

$$\begin{aligned} \frac{\partial CM}{\partial \boldsymbol{\alpha}} &= 2\big(\boldsymbol{H}\boldsymbol{\alpha} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{a}_2 - \boldsymbol{X}\boldsymbol{\lambda}\big) \\ \frac{\partial CM}{\partial \boldsymbol{\lambda}} &= 2\big(\boldsymbol{a}_1 - \boldsymbol{X}^T\boldsymbol{\alpha}\big) \end{aligned}$$

Equating these equations to zero and solving for $\boldsymbol{\alpha}$ and $\boldsymbol{a}$ gives

$$\boldsymbol{\alpha} = \boldsymbol{H}^{-1}(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{a}_2 + \boldsymbol{X}\boldsymbol{\lambda}) \tag{2.2.2}$$

$$\boldsymbol{a}_1 = \boldsymbol{X}\boldsymbol{\alpha} \tag{2.2.3}$$

Substituting (2.2.2) into the unbiasedness constraint, (2.2.3), the Lagrange Multiplier can be expressed as

$$\boldsymbol{\lambda} = (\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{a}_1 - \boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{a}_2)$$

Replacing $\boldsymbol{\lambda}$ in (2.2.2) gives

$$\boldsymbol{\alpha} = \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{a}_1 + (\boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1})\boldsymbol{Z}\boldsymbol{G}\boldsymbol{a}_2.$$

Therefore the linear estimator can be written in the form

$$\begin{aligned}\boldsymbol{\alpha}^T\boldsymbol{y} &= \boldsymbol{a}_1^T(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y} + \boldsymbol{a}_2^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{y} \\ &= \boldsymbol{a}_1^T\hat{\boldsymbol{\beta}} + \boldsymbol{a}_2^T\tilde{\boldsymbol{u}}\end{aligned}$$

where $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}$. Here, $\hat{\boldsymbol{\beta}}$ is known as the best linear unbiased estimator (**BLUE**) of the location effects parameter, $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{u}}$ is the best linear unbiased predictor (**BLUP**) for the random effect variable, $\boldsymbol{u}$. A comprehensive review of BLUP can be found in Robinson (1991).

2.2.1 Mixed Model Equations

Consider the BLUP for the random effects, $\tilde{\boldsymbol{u}}$, from the previous section. This may be expressed as

$$\begin{aligned}\tilde{\boldsymbol{u}} &= \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{y} && (2.2.4)\\ &= \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{H}^{-1}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1})\boldsymbol{y} \\ &= \boldsymbol{G}\boldsymbol{Z}^T(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{R})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \quad \text{(using Result A.3.2)} && (2.2.5)\end{aligned}$$

Similarly the location parameter estimator may be rearranged to give

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y} && (2.2.6)\\ &= (\boldsymbol{X}^T(\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1})\boldsymbol{X})^{-1} \\ &\times\ \boldsymbol{X}^T(\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1})\boldsymbol{y} \quad \text{(using Result A.3.1)} \\ &= (\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{u}}) \quad \text{(using (2.2.5))} && (2.2.7)\end{aligned}$$

Using (2.2.5) and (2.2.7) gives the mixed model equations (**MME**) as

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix} \tag{2.2.8}$$

These equations were originally derived by Henderson (1953) and have been widely used to understand mixed models. Henderson also derived a secondary justification for the estimates, $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}$, from the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{u}$. A maximising objective function for $(\boldsymbol{y}, \boldsymbol{u})$ can be expressed as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}) = L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{u})L(\boldsymbol{\varphi}; \boldsymbol{u})$$

where $L(\cdot; \boldsymbol{y}|\boldsymbol{u})$ and $L(\cdot; \boldsymbol{u})$ are defined by (2.1.3). Note, as $\boldsymbol{u}$ is unobserved, $L(\cdot; \boldsymbol{y}|\boldsymbol{u})$ is not a true likelihood for the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{u}$ and therefore neither

is $L(\cdot; \boldsymbol{y}, \boldsymbol{u})$. Omitting constants and using (2.1.3) this *pseudo* joint log-likelihood can be expressed as

$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}) &= -\frac{1}{2}\big\{\log|\boldsymbol{R}| + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})\big\} \\ &\quad -\frac{1}{2}\big\{\log|\boldsymbol{G}| + \boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}\big\}\end{aligned}$$

Taking derivatives with respect to $\boldsymbol{\beta}$ and $\boldsymbol{u}$ gives

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u})}{\partial\boldsymbol{\beta}} &= \boldsymbol{X}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) \\ \frac{\partial\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u})}{\partial\boldsymbol{u}} &= \boldsymbol{Z}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\boldsymbol{u}\end{aligned}$$

Equating the derivatives to zero provides a solution equivalent to the mixed model equations given by (2.2.8).

The residuals for the linear mixed effects model considered in (2.1.1) can be written as

$$\begin{aligned}\tilde{\boldsymbol{e}} &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\tilde{\boldsymbol{u}} \\ &= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{y} \\ &= \boldsymbol{H}(\boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1})\boldsymbol{y} - (\boldsymbol{H} - \boldsymbol{R})\boldsymbol{P}\boldsymbol{y} \\ &= \boldsymbol{R}\boldsymbol{P}\boldsymbol{y}\end{aligned}$$

2.3 Scale Parameter Estimation

Depending on the data the scale parameters may be estimated efficiently in several ways. If the data is balanced and the scale parameter structures are simplistic, ANOVA tables can be used to provide estimates for the scale parameters. When the data are unbalanced or the scale parameter structures are complex Maximum Likelihood (ML) can be used to estimate the parameters. It is well known that ML produces biased scale parameter estimates when the data are unbalanced. An alternative method of estimating the location and scale parameters is to use Residual/Restricted Maximum Likelihood (REML) (see Patterson & Thompson, 1971). REML allows for the loss of degrees of freedom when estimating the location parameters and therefore produces less biased scale parameter estimates than ML.

2.3.1 Restricted Maximum Likelihood

Several derivations of REML for mixed models are available. A convenient conditional derivation was presented by Verbyla (1990) and is presented in this section.

Verbyla (1990) considers a non-singular matrix $\boldsymbol{L} = \big[\boldsymbol{L}_1, \boldsymbol{L}_2\big]^T$ where $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ are $n \times p$ and $n \times (n-p)$ matrices respectively. These matrices are chosen to satisfy specific conditions, namely,

$$\boldsymbol{L}_1^T \boldsymbol{X} = \boldsymbol{I}_p \qquad \text{and} \qquad \boldsymbol{L}_2^T \boldsymbol{X} = \boldsymbol{0} \tag{2.3.1}$$

Transforming $\boldsymbol{y}$ the joint distribution of $(\boldsymbol{y}, \boldsymbol{u})$ can be expressed as

$$\begin{bmatrix} \boldsymbol{L}_1^T \boldsymbol{y} \\ \boldsymbol{L}_2^T \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \boldsymbol{u} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{L}_1^T \boldsymbol{R} \boldsymbol{L}_1 & \boldsymbol{L}_1^T \boldsymbol{R} \boldsymbol{L}_2 & \boldsymbol{L}_1^T \boldsymbol{Z} \boldsymbol{G} \\ \boldsymbol{L}_2^T \boldsymbol{R} \boldsymbol{L}_1 & \boldsymbol{L}_2^T \boldsymbol{R} \boldsymbol{L}_2 & \boldsymbol{L}_2^T \boldsymbol{Z} \boldsymbol{G} \\ \boldsymbol{G} \boldsymbol{Z}^T \boldsymbol{L}_1 & \boldsymbol{G} \boldsymbol{Z}^T \boldsymbol{L}_2 & \boldsymbol{Z} \boldsymbol{G} \boldsymbol{Z}^T \end{bmatrix} \right)$$

Integrating out the random effects the marginal distribution can be immediately written as

$$\begin{bmatrix} \boldsymbol{L}_1^T \boldsymbol{y} \\ \boldsymbol{L}_2^T \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{L}_1^T \boldsymbol{H} \boldsymbol{L}_1 & \boldsymbol{L}_1^T \boldsymbol{H} \boldsymbol{L}_2 \\ \boldsymbol{L}_2^T \boldsymbol{H} \boldsymbol{L}_1 & \boldsymbol{L}_2^T \boldsymbol{H} \boldsymbol{L}_2 \end{bmatrix} \right) \tag{2.3.2}$$

where $\boldsymbol{H}$ is defined in the previous section and is a function of the scale parameters, $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\varphi})$. This new partitioned response vector also has a Gaussian distribution and assumes a joint likelihood of the form,

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}_1, \boldsymbol{y}_2) = L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}_1 | \boldsymbol{y}_2) L(\boldsymbol{\theta}; \boldsymbol{y}_2) \tag{2.3.3}$$

where, using Result A.4.2, the conditional distribution of $\boldsymbol{y}_1$ given $\boldsymbol{y}_2$ has the form

$$\boldsymbol{y}_1 | \boldsymbol{y}_2 \sim \mathrm{N}\big(\boldsymbol{\beta} + \boldsymbol{L}_1^T \boldsymbol{H} \boldsymbol{L}_2 (\boldsymbol{L}_2^T \boldsymbol{H} \boldsymbol{L}_2)^{-1} \boldsymbol{y}_2, \boldsymbol{L}_1^T \boldsymbol{H} \boldsymbol{L}_1 - \boldsymbol{L}_1^T \boldsymbol{H} \boldsymbol{L}_2 (\boldsymbol{L}_2^T \boldsymbol{H} \boldsymbol{L}_2)^{-1} \boldsymbol{L}_2^T \boldsymbol{H} \boldsymbol{L}_1\big) \tag{2.3.4}$$

and the marginal distribution of $\boldsymbol{y}_2$ is

$$\boldsymbol{y}_2 \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{L}_2^T \boldsymbol{H} \boldsymbol{L}_2) \tag{2.3.5}$$

Noting (2.3.1) and Result A.5.1 the conditional distribution can be simplified to

$$\boldsymbol{y}_1 | \boldsymbol{y}_2 \sim \mathrm{N}\big(\boldsymbol{\beta} + \boldsymbol{y}_2^*, (\boldsymbol{X}^T \boldsymbol{H}^{-1} \boldsymbol{X})^{-1}\big).$$

where $\boldsymbol{y}_2^* = \boldsymbol{L}_1^T \boldsymbol{H} \boldsymbol{L}_2 (\boldsymbol{L}_2^T \boldsymbol{H} \boldsymbol{L}_2)^{-1} \boldsymbol{y}_2$. The log-likelihood, of this distribution omitting constant terms, is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}_1 | \boldsymbol{y}_2) = \tfrac{1}{2}\big\{\log |\boldsymbol{X}^T \boldsymbol{H}^{-1} \boldsymbol{X}| - (\boldsymbol{y}_1 - \boldsymbol{\beta} - \boldsymbol{y}_2^*)^T \boldsymbol{X}^T \boldsymbol{H}^{-1} \boldsymbol{X} (\boldsymbol{y}_1 - \boldsymbol{\beta} - \boldsymbol{y}_2^*)\big\} \tag{2.3.6}$$

To estimate the fixed location parameters, the derivative of this conditional likelihood is required,

$$\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}_1 | \boldsymbol{y}_2)}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{H}^{-1} \boldsymbol{X} (\boldsymbol{y}_1 - \boldsymbol{\beta} - \boldsymbol{y}_2^*).$$

Equating this equation to zero the location parameters can be expressed as

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \boldsymbol{L}_1^T(\boldsymbol{I} - \boldsymbol{H}\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T)\boldsymbol{y} \\ &= \boldsymbol{L}_1^T\{\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1})\}\boldsymbol{y} \qquad \text{(using Result A.5.1)} \\ &= (\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y} \qquad \text{(using (2.3.1))}\end{aligned}$$

This estimate for $\boldsymbol{\beta}$ is identical to the location parameter estimate derived in Section 2.2. This crucial step in the estimation of the location parameters describes Restricted Maximum Likelihood as more than a marginal likelihood approach to parameter maximisation.

In particular, the columns of the transformation matrix $\boldsymbol{L}_1$ define a set of location parameter contrast vectors, $\boldsymbol{l}_i, i = 1, \ldots, p$ to estimate the location fixed effects. Therefore p degrees of freedom have been utilised. Furthermore, under this transformation, the residuals from the conditional likelihood, $\boldsymbol{y}_1 - \hat{\boldsymbol{\beta}} - \boldsymbol{y}_2^* = \boldsymbol{0}_P$. Therefore the conditional likelihood cannot be used to estimate the remaining scale parameters $\boldsymbol{\theta}$ as the maximal set of contrasts have been applied.

Estimates of the scale parameters are found using the marginal likelihood, $\ell(\boldsymbol{\theta}; \boldsymbol{y}_2)$. Thus, the columns of $\boldsymbol{L}_2$ define a set of $n - p$ error or residual contrasts with expectation zero. Using (2.3.5) the marginal log-likelihood can be written as, omitting constant terms,

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}_2) = -\tfrac{1}{2}\big\{\log\big|\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2\big| + \boldsymbol{y}^T\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\boldsymbol{y}\big\}$$

Using the determinants from the conditional and the marginal likelihood it can be seen that

$$\begin{aligned}\log|\boldsymbol{L}^T\boldsymbol{H}\boldsymbol{L}| &= \log|\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}| - \log\big|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}\big| \\ \Rightarrow \qquad \log|\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2| &= \log|\boldsymbol{L}^T\boldsymbol{L}| + \log|\boldsymbol{H}| + \log|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}|\end{aligned}$$

If $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}$, then using Result A.5.1 the marginal log-likelihood, ignoring constants, can be rewritten as

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = -\tfrac{1}{2}\big\{\log|\boldsymbol{H}| + \log|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}| + \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}\big\} \qquad (2.3.7)$$

Noting that

$$\begin{aligned}\boldsymbol{H}\boldsymbol{P}\boldsymbol{y} &= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y} \\ &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\end{aligned}$$

the quadratic form of the Restricted Maximum Likelihood for $\boldsymbol{y}_2$ given in (2.3.7) can also be expressed as

$$\begin{aligned}\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} &= \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{H}\boldsymbol{P}\boldsymbol{y} \qquad \text{(using Result A.5.2)} \\ &= \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{H}\boldsymbol{H}^{-1}\boldsymbol{H}\boldsymbol{P}\boldsymbol{y} \\ &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\end{aligned}$$

This result will be used in the forthcoming chapters.

2.3.2 Estimation of the Scale Parameters

Estimates of the scale parameters under Restricted Maximum Likelihood are found by solving the REML score equations given by

$$\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \theta_i} = \boldsymbol{U}(\theta_i) = 0 \tag{2.3.8}$$

for $i = 1, \ldots, s$ where s is the number of scale parameters.

Let $\dot{\boldsymbol{H}}_i$ be the derivative of scale parameter matrix $\boldsymbol{H}$ with respect to θ_i. Then the partial derivatives of the determinant terms of (2.3.7) with respect to θ_i gives

$$\begin{aligned}
\frac{\partial}{\partial \theta_i}\big(\log|\boldsymbol{H}| + \log|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}|\big) &= \operatorname{tr}(\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_i) - \operatorname{tr}((\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_i\boldsymbol{H}^{-1}\boldsymbol{X}) \\
&= \operatorname{tr}(\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_i) - \operatorname{tr}(\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_i) \\
&= \operatorname{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_i)
\end{aligned}$$

The partial derivative of the quadratic form of (2.3.7) with respect to θ_i is

$$\frac{\partial}{\partial \theta_i}\big(\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}\big) = -\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y} \qquad \text{(using Result A.5.3)}$$

The REML score equations can then be expressed as

$$\boldsymbol{U}(\theta_i) = -\frac{1}{2}\operatorname{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_i) + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y} \tag{2.3.9}$$

for $i = 1, \ldots, n$. This system of equations is generally not explicitly solvable due to the non-linearity of the scale parameters in $\boldsymbol{U}(\cdot)$. Commonly, a linearisation of the score function is used to derive an iterative approach. The first term of a Taylor series expansion around the mth iterate, $\boldsymbol{\theta}^{(m)}$ becomes

$$\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{U}(\boldsymbol{\theta}^{(m)}) + \left.\frac{\partial \boldsymbol{U}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}} (\boldsymbol{\theta} - \boldsymbol{\theta}^{(m)}) \tag{2.3.10}$$

The score equation is then solved by equating the RHS to zero giving

$$\boldsymbol{\theta} = \boldsymbol{\theta}^{(m)} + \big(\mathcal{I}_o(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)})\big)^{-1}\boldsymbol{U}(\boldsymbol{\theta}^{(m)}) \tag{2.3.11}$$

where $\mathcal{I}_o(\boldsymbol{\theta}, \boldsymbol{\theta}) = -\partial \boldsymbol{U}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is the observed information for $\boldsymbol{\theta}$. This iterative process is known as the Newton-Raphson algorithm. Therefore given the mth iterate, $\boldsymbol{\theta}^{(m)}$, $\boldsymbol{\theta}$ is using the RHS of (2.3.11) to obtain $\boldsymbol{\theta}^{(m+1)}$. This process is repeated until $|\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}| < \epsilon$, where ϵ is some predefined tolerance.

The elements of the observed information can be derived by taking the derivative of the score function derived in (2.3.9). The derivative of its trace term with respect to θ_j is

$$\frac{\partial}{\partial \theta_j}\big(\operatorname{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_i)\big) = \operatorname{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_{ij}) - \operatorname{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_i) \qquad \text{(using Result A.5.3)}$$

The derivative of the quadratic form remaining in (2.3.9) with respect to θ_j is

$$\frac{\partial}{\partial\theta_j}\big(\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y}\big) = \boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_{ij}\boldsymbol{P}\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y} \qquad \text{(using Result A.5.3)}$$

Therefore the ijth element of the observed information can be written as

$$\mathcal{I}_o(\theta_i,\theta_j) = \frac{1}{2}\mathrm{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_{ij}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_i) - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_{ij}\boldsymbol{P}\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y} \quad (2.3.12)$$

An alternative algorithmic process known as Fisher scoring is also often used to solve the REML score equations. This requires the elements of the observed information in (2.3.11) to be replaced with the elements of the expected information, where the expected information is defined by

$$\mathcal{I}_e(\theta_i,\theta_j) = \mathrm{E}\Big[-\frac{\partial^2\ell(\boldsymbol{\theta},\boldsymbol{y})}{\partial\theta_i\partial\theta_j}\Big] = \mathrm{E}\big[\mathcal{I}_o(\theta_i,\theta_j)\big]$$

for $i,j = 1,\ldots,n$. Taking expectations of the ijth element of the observed information the ijth element of the expected information becomes

$$\begin{aligned}
\mathcal{I}_e(\theta_i,\theta_j) &= \frac{1}{2}\mathrm{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_{ij}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_i) - \frac{1}{2}\mathrm{tr}(\boldsymbol{H}\boldsymbol{P}\dot{\boldsymbol{H}}_{ij}\boldsymbol{P}) - \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{P}\dot{\boldsymbol{H}}_{ij}\boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta} \\
&\quad + \mathrm{tr}(\boldsymbol{H}\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}) + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta} \qquad \text{(using Result A.4.1)} \\
&= \frac{1}{2}\mathrm{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_{ij}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_i) - \frac{1}{2}\mathrm{tr}(\boldsymbol{P}\boldsymbol{H}\boldsymbol{P}\dot{\boldsymbol{H}}_{ij}) + \mathrm{tr}(\boldsymbol{P}\boldsymbol{H}\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_i) \\
&= \frac{1}{2}\mathrm{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_i) \qquad \text{(using Result A.5.2)}
\end{aligned}$$

Chapter 3

Location and Scale Parameter Modelling of the heteroscedastic Gaussian Distribution

The general framework of the linear mixed model formulation of the previous chapter allows for variance or correlation structures to determine the dependence between the observations. In some instances the variance of the observations may change in magnitude according to covariates existing in the data. This type of heterogeneity can be flexibly modelled without any alteration to the distribution assumed for the response.

This chapter reviews location and scale parameter modelling of the Gaussian distribution using a very general scale parameter model. The models considered here are similar to the ones investigated in Smyth et al. (2001) and Smyth (2002). In particular, methodology for ML and REML estimation of the location and scale parameters is discussed and techniques for efficient calculation of their components is investigated.

As the heteroscedastic Gaussian and the heteroscedastic t with known degrees of freedom are members of the location-scale family this chapter motivates the research of the heteroscedastic t in the forthcoming chapters.

3.1 Introduction and Notation

Consider the linear model expressed by,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.1.1}$$

where y_i is the ith observed response, $i = 1, \ldots, n$, $\boldsymbol{X}$ is an $n \times p$ matrix of explanatory variables and $\epsilon_i \sim \mathrm{N}(0, \sigma_i^2)$.

Consider a new form for the scale parameter of the ith response.

$$\sigma_i^2 = \sigma^2(\boldsymbol{z}_i; \boldsymbol{\lambda}) \tag{3.1.2}$$

where $\boldsymbol{z}_i$ is a $1 \times q$ vector of explanatory variables that may have common components to $\boldsymbol{x}_i$ from the mean component of the model given in (3.1.1). This generality allows $\boldsymbol{\lambda}$ to be modelled non-linearly if required.

A common simplification for (3.1.2) is to use the natural log as a functional link to a linear predictor, namely

$$\log \sigma_i^2 = \boldsymbol{z}_i^T \boldsymbol{\lambda}, \qquad \text{or} \qquad \sigma_i^2 = \exp(\boldsymbol{z}_i^T \boldsymbol{\lambda}) \tag{3.1.3}$$

to provide positive definiteness to the diagonal scale matrix, $\boldsymbol{\Sigma}$. Common references for such models are Aitkin (1987), Verbyla (1993) and Smyth (2002).

3.2 Maximum Likelihood

The likelihood for (3.1.1) can be immediately written as

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\} \tag{3.2.1}$$

As the scale matrix is diagonal the log-likelihood, omitting constants, is

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y}) &= -\frac{1}{2} \log \prod_{i=1}^{n} \sigma^2(\boldsymbol{z}_i; \boldsymbol{\lambda}) - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2}{\sigma_i^2(\boldsymbol{z}_i; \boldsymbol{\lambda})} \\ &= -\frac{1}{2} \left\{ \sum_{i=1}^{n} \log \sigma_i^2 + \sum_{i=1}^{n} \frac{d_i}{\sigma_i^2} \right\} \end{aligned} \tag{3.2.2}$$

where $d_i = (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$ and $\sigma_i^2 = \sigma^2(\boldsymbol{z}_i; \boldsymbol{\lambda})$. This likelihood can be viewed in two distinct ways. Firstly, it represents a likelihood for a simple linear model where, for the ith response, y_i, it has location $\boldsymbol{x}_i^T \boldsymbol{\beta}$ and scale parameter σ_i^2. Secondly it can be viewed as a likelihood for a Gamma generalized linear model with ith response d_i, location σ_i^2 and a common scale parameter equal to 2.

3.2.1 Score Equations

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\lambda})$. The scale parameters appear non-linearly in the proposed likelihood given in (3.2.2). This suggests that an iterative approach to estimate the location and scale parameters is required. Using Section (2.3.2) estimates can be obtained by solving a system of score equations given as

$$\boldsymbol{U}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{U}(\boldsymbol{\beta}) \\ \boldsymbol{U}(\boldsymbol{\lambda}) \end{bmatrix} = \begin{bmatrix} \partial \ell(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y}) / \partial \boldsymbol{\beta} \\ \partial \ell(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y}) / \partial \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}$$

Partially differentiating (3.2.2) with respect to the lth location parameter β_l and the jth scale parameter λ_j gives

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y})}{\partial \beta_l} &= \sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})}{\sigma_i^2} x_{il} \qquad (3.2.3) \\
\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y})}{\partial \lambda_j} &= \frac{1}{2} \sum_{i=1}^{n} \left\{ \frac{d_i}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \lambda_j} - \frac{1}{\sigma_i^2} \frac{\partial \sigma_i^2}{\partial \lambda_j} \right\} \\
&= \frac{1}{2} \sum_{i=1}^{n} \left\{ \frac{\dot{s}_{ij}}{\sigma_i^2} \left(\frac{d_i}{\sigma_i^2} - 1 \right) \right\} \qquad (3.2.4)
\end{aligned}$$

where $\dot{s}_{ij} = \partial \sigma_i^2 / \partial \lambda_j$. Let $\dot{\boldsymbol{S}}$ be a $n \times q$ matrix of partial derivatives with ijth element $\dot{s}_{ij}$ and $\boldsymbol{d}$ be a vector of length n with ith element $d_i = (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$ then the score equations can be written as

$$\boldsymbol{U}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{U}(\boldsymbol{\beta}) \\ \boldsymbol{U}(\boldsymbol{\lambda}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ \frac{1}{2} \dot{\boldsymbol{S}}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\Sigma}^{-1} \boldsymbol{d} - \boldsymbol{1}_n) \end{bmatrix} \qquad (3.2.5)$$

As $\mathrm{E}[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta}$ and $\mathrm{E}[\boldsymbol{d}] = \boldsymbol{\Sigma}\boldsymbol{1}_n$ the score equations are unbiased.

3.2.2 Solving the Score Equations

Section 2.3.2 proposed a linearisation of the score functions to obtain an iterative solution known as the Newton-Raphson algorithm defined by where at the mth iterate, is defined by

$$\boldsymbol{\theta}_{(m+1)} = \boldsymbol{\theta}_m) + \left(\mathcal{I}_o(\boldsymbol{\theta}_m, \boldsymbol{\theta}_m) \right)^{-1} \boldsymbol{U}(\boldsymbol{\theta}_m) \qquad (3.2.6)$$

where $\mathcal{I}_o(\boldsymbol{\theta}, \boldsymbol{\theta})$ is the observed information,

$$\mathcal{I}_o(\boldsymbol{\theta}, \boldsymbol{\theta}) = \begin{bmatrix} \mathcal{I}_o(\boldsymbol{\beta}, \boldsymbol{\beta}) & \mathcal{I}_o(\boldsymbol{\beta}, \boldsymbol{\lambda}) \\ \mathcal{I}_o(\boldsymbol{\lambda}, \boldsymbol{\beta}) & \mathcal{I}_o(\boldsymbol{\lambda}, \boldsymbol{\lambda}) \end{bmatrix}$$

To obtain this the elements of the score functions (3.2.3) and (3.2.4) are partially differentiated with respect to β_m and λ_k. Let $\dot{s}_{i(jk)} = \partial s_{ij} / \partial \lambda_k$. The observed information elements can then be expressed as

$$\begin{aligned}
\mathcal{I}_o(\beta_l, \beta_m) &= -\partial \boldsymbol{U}(\beta_l) / \partial \beta_m = \sum_{i=1}^{n} \frac{x_{il} x_{im}}{\sigma_i^2} \\
\mathcal{I}_o(\lambda_j, \beta_m) &= -\partial \boldsymbol{U}(\lambda_l) / \partial \beta_m = \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})}{(\sigma_i^2)^2} x_{im} \dot{s}_{ij} \\
\mathcal{I}_o(\lambda_j, \lambda_k) &= -\partial \boldsymbol{U}(\lambda_j) / \partial \lambda_k = \frac{1}{2} \sum_{i=1}^{n} \left\{ \frac{2 d_i}{(\sigma_i^2)^3} \dot{s}_{ij} \dot{s}_{ik} - \frac{d_i}{(\sigma_i^2)^2} \dot{s}_{i(jk)} - \frac{1}{(\sigma_i^2)^2} \dot{s}_{ij} \dot{s}_{ik} + \frac{1}{\sigma_i^2} \dot{s}_{i(jk)} \right\} \\
&= \frac{1}{2} \sum_{i=1}^{n} \left\{ \frac{\dot{s}_{ij}}{\sigma_i^2} \left(\frac{2 d_i}{\sigma_i^2} - 1 \right) \frac{\dot{s}_{ik}}{\sigma_i^2} - \frac{\dot{s}_{i(jk)}}{\sigma_i^2} \left(\frac{d_i}{\sigma_i^2} - 1 \right) \right\}
\end{aligned}$$

In matrix notation the observed information can be written as

$$\mathcal{I}_o(\boldsymbol{\theta},\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{D}^{1/2}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}} \\ \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{D}^{1/2}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}} & (\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-1}(2\boldsymbol{\Sigma}^{-1}\boldsymbol{D}-I)\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}}+\boldsymbol{K})/2 \end{bmatrix}$$

where $\boldsymbol{D}$ is a diagonal matrix with ith diagonal element $(y-\boldsymbol{x}_i^T\boldsymbol{\beta})^2$ and

$$\boldsymbol{K} = [K_{jk}] = \sum_{i=1}^{n} \frac{\dot{s}_{i(jk)}}{\sigma_i^2}\left(\frac{d_i}{\sigma_i^2}-1\right)$$

As noted in Section 2.3.2 the observed information can be replaced with the expected information in (3.2.6) to obtain a Fisher scoring algorithm. Noting that $\mathrm{E}[\boldsymbol{D}] = \boldsymbol{\Sigma}$ and $\mathrm{E}[\boldsymbol{D}^{1/2}] = \boldsymbol{0}$ and taking expectations of the observed information the expected information for $(\boldsymbol{\beta},\boldsymbol{\lambda})$ can be immediately written as

$$\mathcal{I}_e(\boldsymbol{\theta},\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{0} & \dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-2}\dot{\boldsymbol{S}}/2 \end{bmatrix}.$$

Therefore the location and scale parameters are mutually orthogonal. The expected information, and this property of orthogonality, can also be derived by taking the variance of the score function defined in (3.2.5). The variance of the score function for $\boldsymbol{\beta}$ is then

$$\begin{aligned} \mathrm{Var}\big[\boldsymbol{U}(\boldsymbol{\beta})\big] &= \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\mathrm{Var}\left[\boldsymbol{y}\right]\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \\ &= \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} = \mathcal{I}_e(\boldsymbol{\beta},\boldsymbol{\beta}) \end{aligned}$$

Recognising higher moments of the multivariate normal distribution, the variance of the score function for $\boldsymbol{\lambda}$ is

$$\begin{aligned} \mathrm{Var}\left[\boldsymbol{U}(\boldsymbol{\lambda})\right] &= \frac{1}{4}\dot{\boldsymbol{S}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\mathrm{Var}\left[\boldsymbol{d}\right]\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}} \\ &= \frac{1}{4}\dot{\boldsymbol{S}}\boldsymbol{\Sigma}^{-2}\left\{\mathrm{E}\left[\boldsymbol{d}\boldsymbol{d}^T\right]-\mathrm{E}\left[\boldsymbol{d}\right]\mathrm{E}\left[\boldsymbol{d}^T\right]\right\}\boldsymbol{\Sigma}^{-2}\dot{\boldsymbol{S}} \\ &= \frac{1}{4}\dot{\boldsymbol{S}}\boldsymbol{\Sigma}^{-2}\left\{3\boldsymbol{\Sigma}^2-\boldsymbol{\Sigma}^2\right\}\boldsymbol{\Sigma}^{-2}\dot{\boldsymbol{S}} \\ &= \frac{1}{2}\dot{\boldsymbol{S}}\boldsymbol{\Sigma}^{-2}\dot{\boldsymbol{S}} \;=\; \mathcal{I}_e(\boldsymbol{\lambda},\boldsymbol{\lambda}) \end{aligned}$$

The orthogonality of the parameters $(\boldsymbol{\beta},\boldsymbol{\lambda})$ can be shown by taking the covariance of the score functions, namely

$$\begin{aligned} \mathrm{Cov}\left[\boldsymbol{U}(\boldsymbol{\beta}),\boldsymbol{U}(\boldsymbol{\lambda})\right] &= \mathrm{E}\left[\boldsymbol{U}(\boldsymbol{\beta})\boldsymbol{U}(\boldsymbol{\lambda})^T\right] \\ &= \frac{1}{2}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\mathrm{E}\left[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{1}_n-\boldsymbol{\Sigma}^{-1}\boldsymbol{d})^T\right]\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}} \\ &= \frac{1}{2}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\left\{\mathrm{E}\left[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{1}_n^T\right]-\mathrm{E}\left[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{d}^T\boldsymbol{\Sigma}^{-1}\right]\right\}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}} \\ &= \boldsymbol{0} = \mathcal{I}_e(\boldsymbol{\beta},\boldsymbol{\lambda}) \end{aligned}$$

Due to the orthogonality of the location and scale parameters the Fisher scoring algorithm can proceed independently for each parameter. Using (3.2.6), the scoring equations for the mth iteration can be written as

$$\begin{aligned}\boldsymbol{\beta}_{(m+1)} &= f_g(\boldsymbol{\beta}_m, \boldsymbol{\lambda}) = (\boldsymbol{X}^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{y} \qquad (3.2.7)\\ \boldsymbol{\lambda}_{(m+1)} &= \boldsymbol{\lambda}_m + (\dot{\boldsymbol{S}}_m^T\boldsymbol{\Sigma}_m^{-2}\dot{\boldsymbol{S}}_m)^{-1}\dot{\boldsymbol{S}}_m^T\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\Sigma}_m^{-1}\boldsymbol{d} - \boldsymbol{1}_n)\\ &= g_g(\boldsymbol{\beta}, \boldsymbol{\lambda}_m) = (\dot{\boldsymbol{S}}_m^T\boldsymbol{\Sigma}_m^{-2}\dot{\boldsymbol{S}})^{-1}\dot{\boldsymbol{S}}_m^T\boldsymbol{\Sigma}_m^{-2}\boldsymbol{d}_m \qquad (3.2.8)\end{aligned}$$

where

$$\boldsymbol{d}_m = \boldsymbol{d} - \boldsymbol{\Sigma}_m\boldsymbol{1}_n + \dot{\boldsymbol{S}}_m\boldsymbol{\lambda}_m$$

Thus, for given $\boldsymbol{\lambda}$, (3.2.7) is a weighted regression to estimate $\boldsymbol{\beta}$ with weights $(\sigma_i^2)^{-1}$. For given $\boldsymbol{\beta}$, (3.2.8) is an iteratively reweighted least squares procedure with weights $(\sigma_i^2)^{-2}$ and working variate $\boldsymbol{d}_m$ generated from a Gamma generalised linear model (see Verbyla, 1993; Smyth, 2002).

If $\boldsymbol{\lambda}$ is present in the scale parameter log-linearly then (3.1.2) can be expressed as (3.1.3). Under this simplified scale parameter model,

$$\frac{\partial\sigma_i^2}{\partial\lambda_j} = \sigma_i^2 z_{ij} = \dot{s}_{ij}$$

and therefore $\dot{\boldsymbol{S}} = \boldsymbol{\Sigma}\boldsymbol{Z}$ and the scoring equation for the $(m+1)$th iteration of the scale parameter model is reduced to

$$\boldsymbol{\lambda}_{(m+1)} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T(\boldsymbol{\Sigma}_m^{-1}\boldsymbol{d} - \boldsymbol{1}_n + \boldsymbol{Z}\boldsymbol{\lambda}_m) \qquad (3.2.9)$$

noted by Verbyla (1993).

3.3 Restricted Maximum Likelihood

The ML method described in the previous section requires iterative estimation of the scale parameters jointly with the location parameters. As discussed in Section 2.3, this method produces biased estimates of the scale parameters. This bias may be alleviated by estimation of the parameters using Restricted Maximum Likelihood (REML).

Consider the linear model, (3.1.1) where the scale parameter model is defined by (3.1.2). Using (2.3.7) from Section 2.3 the marginal log-likelihood for $\boldsymbol{\lambda}$, omitting constants, can be immediately written as

$$\begin{aligned}\ell(\boldsymbol{\lambda}; \boldsymbol{y}) &= -\frac{1}{2}\left\{\log|\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}| + \sum_{i=1}^n \log\sigma^2(\boldsymbol{z}_i;\boldsymbol{\lambda}) + \sum_{i=1}^n \frac{(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2}{\sigma_i^2(\boldsymbol{z}_i;\boldsymbol{\lambda})}\right\}\\ &= -\frac{1}{2}\left\{\log|\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}| + \sum_{i=1}^n \log\sigma_i^2 + \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}\right\}\end{aligned}$$

where $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}$ and $\hat{\boldsymbol{\beta}}$ is the REML estimate for $\boldsymbol{\beta}$ found by maximising the conditional log-likelihood, $\ell(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y}_1|\boldsymbol{y}_2)$ given by

$$\ell(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y}_1|\boldsymbol{y}_2) = \tfrac{1}{2}\big\{|\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}| - (\boldsymbol{y}_1 - \boldsymbol{\beta} - \boldsymbol{y}_2^*)^T\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{y}_1 - \boldsymbol{\beta} - \boldsymbol{y}_2^*)\big\}$$

Following from Section 2.3 the REML estimate for $\boldsymbol{\beta}$ is equivalent to the ML estimate given by (3.2.7).

3.3.1 REML Scoring Equation

To obtain an estimate for $\boldsymbol{\lambda}$ the REML scoring equations defined by (2.3.8) require solving. As the scale parameters for the location scale model in this chapter are functions of the scale matrix $\boldsymbol{\Sigma}$ the derivations of Section 2.3.2 may be used. Therefore, using (2.3.9), the score function for the jth scale parameter can be immediately written as

$$\boldsymbol{U}_r(\lambda_j) = -\frac{1}{2}\mathrm{tr}(\boldsymbol{P}\dot{\boldsymbol{\Sigma}}_j) + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{\Sigma}}_j\boldsymbol{P}\boldsymbol{y}$$

where $\dot{\boldsymbol{\Sigma}}_j$ is a diagonal matrix with ith diagonal element $\partial\sigma_i^2/\partial\lambda_j$. Expanding this the score function can be written as

$$\begin{aligned}\boldsymbol{U}_r(\lambda_j) &= -\frac{1}{2}\Big\{\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_j) - \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_j)\Big\} \\ &\quad +\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\end{aligned}$$

Let $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{\lambda}) = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1/2}$ be the "hat matrix" (see Cook & Weisberg, 1982 and Cook & Weisberg, 1983) and $h_{ii} = (\sigma_i^2)^{-1}\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i$ be its ith diagonal element. Then as the scale matrix is diagonal the score equation can be reduced to

$$\begin{aligned}\boldsymbol{U}_r(\lambda_j) &= -\frac{1}{2}\sum_{i=1}^n \frac{1}{\sigma_i^2}\frac{\partial\sigma_i^2}{\partial\lambda_j} + \frac{1}{2}\sum_{i=1}^n \frac{h_{ii}}{\sigma_i^2}\frac{\partial\sigma_i^2}{\partial\lambda_j} + \frac{1}{2}\sum_{i=1}^n \frac{\hat{d}_i}{(\sigma_i^2)^2}\frac{\partial\sigma_i^2}{\partial\lambda_j} \qquad (3.3.1) \\ &= -\frac{1}{2}\sum_{i=1}^n \frac{1}{\sigma_i^2}\dot{s}_{ij} + \frac{1}{2}\sum_{i=1}^n \frac{h_{ii}}{\sigma_i^2}\dot{s}_{ij} + \frac{1}{2}\sum_{i=1}^n \frac{\hat{d}_i}{(\sigma_i^2)^2}\dot{s}_{ij}\end{aligned}$$

where $\dot{s}_{ij} = \partial\sigma_i^2/\partial\lambda_j$, $\hat{d}_i = (y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2$, and h_{ii} is the ith diagonal element of $\boldsymbol{H}$. Let $\dot{\boldsymbol{s}}_i$ be the ith row of the matrix $\dot{\boldsymbol{S}}$ defined in the previous section. The score for the vector of scale parameters can then be expressed as

$$\begin{aligned}\boldsymbol{U}_r(\boldsymbol{\lambda}) &= \frac{1}{2}\sum_{i=1}^n \frac{\dot{\boldsymbol{s}}_i}{\sigma_i^2}\Big(\frac{\hat{d}_i}{\sigma_i^2} + h_{ii} - 1\Big) \\ &= \frac{1}{2}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-1}\Big(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{d}} - \boldsymbol{1}_n + \boldsymbol{h}\Big) \qquad (3.3.2)\end{aligned}$$

where $\hat{\boldsymbol{d}}$ has ith element $\hat{d}_i$ and $\boldsymbol{h} = (h_{11}, \ldots, h_{nn})$.

3.3.2 Adjusted Profile Score

McCullagh & Tibshirani (1990) show that the REML score equation (3.3.2) can be derived by adjusting the profile score equation for $\boldsymbol{\lambda}$. Replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$ in the lower partition of (3.2.5) gives the profile score equation as

$$\boldsymbol{U}_p(\boldsymbol{\lambda}) = \frac{1}{2}\sum_{i=1}^{n} \frac{\dot{\boldsymbol{s}}_i}{\sigma_i^2}\Big\{\frac{(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2}{\sigma_i^2} - 1\Big\} \tag{3.3.3}$$

This equation lacks the fundamental property of the Maximum Likelihood score, i.e $\mathrm{E}[\boldsymbol{U}(\boldsymbol{\lambda})] = \boldsymbol{0}_q$. To correct this the expectation of (3.3.3) is required. The expectation of $(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2,\ i = 1, \ldots, n$ requires the diagonal elements of

$$\mathrm{E}\left[(\boldsymbol{y} - \boldsymbol{X}^T\hat{\boldsymbol{\beta}})(\boldsymbol{y} - \boldsymbol{X}^T\hat{\boldsymbol{\beta}})^T\right] \;=\; \mathrm{Var}[\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}]$$

and

$$\begin{aligned}
\mathrm{Var}[\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}] &= (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1})\mathrm{Var}[\boldsymbol{y}](\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1})^T \\
&= \boldsymbol{\Sigma} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T \\
&= \boldsymbol{\Sigma}^{1/2}(\boldsymbol{I} - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1/2})\boldsymbol{\Sigma}^{1/2} \\
&= \boldsymbol{\Sigma}^{1/2}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\Sigma}^{1/2}
\end{aligned} \tag{3.3.4}$$

The ith diagonal element can be expressed as

$$\mathrm{E}\left[(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2\right] = \sigma_i^2(1 - h_{ii})$$

and the *adjusted profile score* can be written as

$$\boldsymbol{U}_a(\boldsymbol{\lambda}) = \frac{1}{2}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{d}} - \boldsymbol{1}_n + \boldsymbol{h}\right)$$

This score, now adjusted for the estimation of the location parameters, is identical to the REML score given in (3.3.2).

3.3.3 Solving the REML Equations

Applying the techniques of Section 2.3 the scoring equations for $\boldsymbol{\lambda}$ can be iteratively solved. At the mth iteration a Newton-Raphson algorithm can be written

$$\boldsymbol{\lambda}_{(m+1)} = \boldsymbol{\lambda}_m + (\mathcal{I}_o(\boldsymbol{\lambda}_m, \boldsymbol{\lambda}_m))^{-1}\boldsymbol{U}(\boldsymbol{\lambda}_m)$$

where $\mathcal{I}_o(\cdot, \cdot)$ is the observed information for $\boldsymbol{\lambda}$. Again, to obtain a Fisher scoring algorithm the observed information may be replaced with the expected information.

To obtain the observed information components the the derivative of the REML score equation, (3.3.1), is required. The components of the score, h_{ii} and $\hat{d}_i$ are functions of the scale parameters, $\boldsymbol{\lambda}$, and therefore careful differentiation is required. Differentiating the first component of this score with respect to λ_k gives

$$\begin{aligned}\frac{\partial}{\partial\lambda_k}\left\{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\frac{\partial\sigma_i^2}{\partial\lambda_j}\right\} &= -\sum_{i=1}^{n}\left\{\frac{1}{(\sigma_i^2)^2}\frac{\partial\sigma_i^2}{\partial\lambda_j}\frac{\partial\sigma_i^2}{\partial\lambda_k}-\frac{1}{\sigma_i^2}\frac{\partial^2\sigma_i^2}{\partial\lambda_j\partial\lambda_k}\right\}\\ &= -\sum_{i=1}^{n}\left\{\frac{\dot{s}_{ij}\dot{s}_{ik}}{(\sigma_i^2)^2}-\frac{\dot{s}_{i(jk)}}{\sigma_i^2}\right\} \qquad (3.3.5)\end{aligned}$$

The differentiation of the second component of the score with respect to λ_k becomes

$$\frac{\partial}{\partial\lambda_k}\left\{\sum_{i=1}^{n}\frac{h_{ii}}{\sigma_i^2}\frac{\partial\sigma_i^2}{\partial\lambda_j}\right\} = \sum_{i=1}^{n}\left\{\frac{\partial h_{ii}}{\partial\lambda_k}\frac{1}{\sigma_i^2}\frac{\partial\sigma_i^2}{\partial\lambda_j}-\frac{h_{ii}}{(\sigma_i^2)^2}\frac{\partial\sigma_i^2}{\partial\lambda_j}\frac{\partial\sigma_i^2}{\partial\lambda_k}+\frac{h_{ii}}{\sigma_i^2}\frac{\partial^2\sigma_i^2}{\partial\lambda_j\partial\lambda_k}\right\}$$

where

$$\begin{aligned}\frac{\partial h_{ii}}{\partial\lambda_k} &= \frac{\partial}{\partial\lambda_k}\left\{\frac{\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i}{\sigma_i^2}\right\}\\ &= -\frac{h_{ii}}{\sigma_i^2}\frac{\partial\sigma_i^2}{\partial\lambda_k}+(\sigma_i^2)^{-1}\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_k\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i\\ &= \sum_{m=1}^{n}\frac{h_{im}\dot{s}_{mk}h_{mi}}{\sigma_m^2}-\frac{\dot{s}_{ik}h_{ii}}{\sigma_i^2}\end{aligned}$$

Recombining this the derivative of the second component is

$$\frac{\partial}{\partial\lambda_k}\left\{\sum_{i=1}^{n}\frac{h_{ii}}{\sigma_i^2}\frac{\partial\sigma_i^2}{\partial\lambda_j}\right\}=\sum_{i=1}^{n}\left\{\sum_{m=1}^{n}\left(\frac{h_{im}\dot{s}_{mk}h_{mi}}{\sigma_m^2}\right)\frac{\dot{s}_{ij}}{\sigma_i^2}-2\frac{\dot{s}_{ij}\dot{s}_{ik}}{(\sigma_i^2)^2}h_{ii}+\frac{\dot{s}_{i(jk)}}{\sigma_i^2}h_{ii}\right\} \qquad (3.3.6)$$

The derivative of the third component of the score with respect to λ_k is

$$\frac{\partial}{\partial\lambda_k}\left\{\sum_{i=1}^{n}\frac{\hat{d}_i}{(\sigma_i^2)^2}\frac{\partial\sigma_i^2}{\partial\lambda_j}\right\} = \sum_{i=1}^{n}\left\{\frac{\partial\hat{d}_i}{\partial\lambda_k}\frac{1}{(\sigma_i^2)^2}\frac{\partial\sigma_i^2}{\partial\lambda_j}-2\frac{\hat{d}_i}{(\sigma_i^2)^3}\frac{\partial\sigma_i^2}{\partial\lambda_j}\frac{\partial\sigma_i^2}{\partial\lambda_k}+\frac{\hat{d}_i}{(\sigma_i^2)^2}\frac{\partial^2\sigma_i^2}{\partial\lambda_j\partial\lambda_k}\right\}$$

where

$$\frac{\partial\hat{d}_i}{\partial\lambda_k} = -2(y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})\boldsymbol{x}_i^T\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\lambda_k}$$

and

$$\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\lambda_k} = -(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_k\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$$

so that

$$\frac{\partial\hat{d}_i}{\partial\lambda_k} = 2\sigma_i r_i\sum_{m=1}^{n}\dot{s}_{mk}\frac{h_{im}}{\sigma_m^3}r_m$$

where $r_i = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ is the ith residual. Recombining this the derivative of the third component becomes

$$\frac{\partial}{\partial\lambda_k}\left\{\sum_{i=1}^{n}\frac{\hat{d}_i}{(\sigma_i^2)^2}\frac{\partial\sigma_i^2}{\partial\lambda_j}\right\} = \sum_{i=1}^{n}\left\{\sum_{m=1}^{n}2\left(\frac{\dot{s}_{mk}}{\sigma_m^3}r_m h_{im}\right)\frac{\dot{s}_{ij}}{\sigma_i^3}r_i - 2\frac{\dot{s}_{ij}\dot{s}_{ik}}{(\sigma_i^2)^3}d_i + \frac{\dot{s}_{i(jk)}}{(\sigma_i^2)^2}\hat{d}_i\right\} \quad (3.3.7)$$

Combining (3.3.5), (3.3.6) and (3.3.7) the jkth element of the observed information can be expressed as

$$\begin{aligned}\mathcal{I}_o(\lambda_j,\lambda_k) &= \frac{1}{2}\sum_{i=1}^{n}\frac{\dot{s}_{ij}\dot{s}_{ik}}{(\sigma_i^2)^2}\left\{\frac{2d_i}{\sigma_i^2}+2h_{ii}-1\right\} - \sum_{i=1}^{n}\sum_{m=1}^{n}\frac{r_i}{\sigma_i^3}\dot{s}_{ij}h_{im}\dot{s}_{mk}\frac{r_m}{\sigma_m^3}\\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{m=1}^{n}\frac{\dot{s}_{ij}}{\sigma_i^2}\frac{\dot{s}_{mk}}{\sigma_m^2}h_{im}^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{\dot{s}_{i(jk)}}{\sigma_i^2}\left\{\frac{d_i}{\sigma_i^2}+h_{ii}-1\right\}\end{aligned}$$

Taking expectations ensures the final term is zero and using (3.3.4) the jkth element of the expected information can be written as

$$\mathcal{I}_e(\lambda_j,\lambda_k) = \begin{cases}\dfrac{1}{2}\displaystyle\sum_{i=1}^{n}\frac{\dot{s}_{ij}\dot{s}_{ik}}{(\sigma_i^2)^2}(1-2h_{ii}+h_{ii}^2) & (i=m)\\ \dfrac{1}{2}\displaystyle\sum_{i=1}^{n}\sum_{m=1}^{n}\frac{\dot{s}_{ij}\dot{s}_{mk}}{\sigma_i^2\sigma_m^2}h_{im}^2 & (i\neq m)\end{cases}$$

Combining these two terms the expected information for $\boldsymbol{\lambda}$ becomes

$$\mathcal{I}_e(\boldsymbol{\lambda},\boldsymbol{\lambda}) = \frac{1}{2}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}} \quad (3.3.8)$$

where $\boldsymbol{V}$ has diagonal elements $(1-h_{ii})^2$ and off-diagonal elements h_{ij}^2.

Identical to the previous section this result may also be derived by considering the variance of the score.

$$\begin{aligned}\mathrm{Var}\left[U_r(\boldsymbol{\lambda})\right] &= \frac{1}{4}\dot{\boldsymbol{S}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\mathrm{Var}\big[\hat{\boldsymbol{d}}\big]\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}}\\ &= \frac{1}{4}\dot{\boldsymbol{S}}\boldsymbol{\Sigma}^{-2}\left\{\mathrm{E}\big[\hat{\boldsymbol{d}}\hat{\boldsymbol{d}}^T\big] - \mathrm{E}\big[\hat{\boldsymbol{d}}\big]\mathrm{E}\big[\hat{\boldsymbol{d}}^T\big]\right\}\boldsymbol{\Sigma}^{-2}\dot{\boldsymbol{S}}\\ &= \frac{1}{4}\dot{\boldsymbol{S}}\boldsymbol{\Sigma}^{-2}\left\{3\boldsymbol{\Sigma}(\boldsymbol{I}-\boldsymbol{H})^2\boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\boldsymbol{I}-\boldsymbol{H})^2\boldsymbol{\Sigma}\right\}\boldsymbol{\Sigma}^{-2}\dot{\boldsymbol{S}} \qquad \text{(using 3.3.4)}\\ &= \frac{1}{2}\dot{\boldsymbol{S}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{I}-\boldsymbol{H})^2\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}} \;=\; \mathcal{I}_e(\boldsymbol{\lambda},\boldsymbol{\lambda})\end{aligned}$$

The scoring algorithm to estimate $\boldsymbol{\lambda}$ can be expressed as

$$\begin{aligned}\boldsymbol{\lambda}_{(m+1)} &= \boldsymbol{\lambda}_m + (\dot{\boldsymbol{S}}^T\boldsymbol{V}_m^*\dot{\boldsymbol{S}})^{-1}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\Sigma}_m^{-1}\hat{\boldsymbol{d}}_m - \boldsymbol{1}_n + \boldsymbol{h}_m)\\ &= g_g^*(\boldsymbol{\beta},\boldsymbol{\lambda}_m) = (\dot{\boldsymbol{S}}^T\boldsymbol{V}_m^*\dot{\boldsymbol{S}})^{-1}\dot{\boldsymbol{S}}^T\boldsymbol{V}_m^*\boldsymbol{d}_m^* \quad (3.3.9)\end{aligned}$$

where $\boldsymbol{V}^*$ is a $q \times q$ matrix with diagonal entries $\left((1-h_{ii})/\sigma_i^2\right)^2$ and off-diagonal entries $(h_{ij}/\sigma_i\sigma_j)^2$ and

$$\boldsymbol{d}_m^* = \boldsymbol{V}_m^{*-1}\boldsymbol{\Sigma}_m^{-1}\left(\boldsymbol{\Sigma}_m^{-1}\hat{\boldsymbol{d}}_m - \mathbf{1}_n + \boldsymbol{h}_m\right) + \dot{\boldsymbol{S}}\boldsymbol{\lambda}_m$$

The $\boldsymbol{V}^*$ is a dense $n \times n$ matrix suggesting that (3.3.9) cannot be viewed as a simple least squares but a general non-linear iteration.

Under a simplified log-linear scale parameter model defined by (3.1.3) the scoring algorithm for $\boldsymbol{\lambda}$ can be reduced to

$$\boldsymbol{\lambda}_{(m+1)} = (\boldsymbol{Z}^T\boldsymbol{V}_m\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{V}_m\left\{\boldsymbol{V}_m^{-1}\left(\boldsymbol{\Sigma}_m^{-1}\hat{\boldsymbol{d}}_m - \mathbf{1}_n + \boldsymbol{h}_m\right) + \boldsymbol{Z}\boldsymbol{\lambda}_m\right\}$$

This result was also noted in Verbyla (1993).

3.3.4 Efficient Calculation of the REML Information

The scoring algorithm defined by (3.3.9) requires the calculation of the dense $n \times n$ matrix, $\boldsymbol{V}$. For large data sets this computation is restrictive. In the past authors have avoided its computation by approximations such as $\boldsymbol{V}_d = \text{diag}\{\boldsymbol{V}\}$ (see Verbyla, 1993 and Huele, 1998). Smyth (2002) shows that, although the elements of the off diagonals of $\boldsymbol{V}$ will be of smaller order, $\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{V}_d\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}}$ does not converge to $\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}}$ as $n \to \infty$. This difference reduces the efficiency of the scoring algorithm, (3.3.9), and introduces relative errors to the computed standard errors of $\boldsymbol{\lambda}$.

Smyth (2002) also shows that the information matrix for $\boldsymbol{\lambda}$ can be calculated efficiently using the properties of the hat matrix, $\boldsymbol{H}$. The matrix $\boldsymbol{V}$ inside the information can also be written as

$$\boldsymbol{V} = \boldsymbol{I} - 2\boldsymbol{H}^* + \boldsymbol{H}^2$$

where $\boldsymbol{H}^*$ is a diagonal matrix with ith element h_{ii} and $\boldsymbol{H}^2$ has ijth element h_{ij}^2. As the first term is diagonal the computational burden is with the calculation of $\boldsymbol{H}^2$. A decomposition of this matrix is available and is derived below (for more details see Smyth, 2002).

Let $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ where $\boldsymbol{Q}$ is $n \times p$ matrix such that $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$ and $\boldsymbol{R}$ is an $p \times p$ is a non-singular upper triangular matrix. This partitioning is known as the QR decomposition.

Then

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{Q}\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{R})^{-1}\boldsymbol{R}^T\boldsymbol{Q} \\ &= \boldsymbol{Q}\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T\boldsymbol{Q} \\ &= \boldsymbol{Q}\boldsymbol{R}\boldsymbol{R}^{-1}(\boldsymbol{R}^T)^{-1}\boldsymbol{R}^T\boldsymbol{Q}^T \\ &= \boldsymbol{Q}\boldsymbol{Q}^T \\ &= \sum_{k=1}^{p}\boldsymbol{q}_k\boldsymbol{q}_k^T \end{aligned} \tag{3.3.10}$$

and therefore the ijth element of $\boldsymbol{H}$ may be written as

$$h_{ij} = \sum_{k=1}^{p} q_{ki}q_{kj} \tag{3.3.11}$$

Moreover, the columns $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_p$ are an independent set of eigenvectors with eigenvalues equal to one forming an orthonormal basis over the range space of $\boldsymbol{H}$. Therefore $\boldsymbol{H}$ has rank p equal to the column rank of the matrix $\boldsymbol{X}$. Using (3.3.11) the ijth element of $\boldsymbol{H}^2$ can be expressed as

$$\begin{aligned} h_{ij}^2 &= (q_{1i}q_{1j} + \cdots + q_{pi}q_{pj})^2 \\ &= \sum_{k=1}^{p} q_{ki}^2 q_{kj}^2 + 2\sum_{1\le k\le l\le p} q_{ki}q_{kj}q_{li}q_{lj} \end{aligned}$$

Therefore the matrix $\boldsymbol{H}^2$ may be written as

$$\boldsymbol{H}^2 = \sum_{m=1}^{p}\boldsymbol{s}_m\boldsymbol{s}_m^T + 2\sum_{m=1}^{p(p-1)/2}\boldsymbol{t}_m\boldsymbol{t}_m^T$$

where $\boldsymbol{s}_m$ has ith element q_{mi}^2 and $\boldsymbol{t}_m$ has ith element $q_{ki}q_{li}$. This suggests that $\boldsymbol{H}^2$ can be constructed using a set of $p + p(p-1)/2$ rank one matrices and therefore is at most rank $p(p+1)/2$.

Let $\boldsymbol{S}$ be a $n \times p$ matrix with mth row $\boldsymbol{s}_m$ and $\boldsymbol{T}$ a $n \times p(p-1)/2$ matrix with mth row $\boldsymbol{t}_m$ then the REML information matrix for $\boldsymbol{\lambda}$ can then be written as

$$\begin{aligned} \mathcal{I}_e(\boldsymbol{\lambda},\boldsymbol{\lambda}) &= \frac{1}{2}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-1}\left\{\boldsymbol{I} - 2\boldsymbol{H}^* + \boldsymbol{S}^T\boldsymbol{S} + 4\boldsymbol{T}^T\boldsymbol{T}\right\}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}} \\ &= \frac{1}{2}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-1}\left\{\boldsymbol{I} - 2\boldsymbol{H}^*\right\}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}} + \frac{1}{2}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}} \end{aligned}$$

where $\boldsymbol{W} = [\boldsymbol{S}\ 2\boldsymbol{T}]^T$. Smyth (2002) recognised that the extra burden to compute the last term of this information will only grow linearly as the data size increases.

3.4 Inference on Parameters

It is of interest to obtain measures of precision and test hypotheses for the location and scale parameter models.

An obvious choice for the precision of the chosen estimates is to use the prediction errors defined as

$$\mathrm{Var}\begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda} \end{bmatrix} = \mathrm{Var}\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\lambda}} \end{bmatrix}$$

For given $\boldsymbol{\lambda}$ the estimate of $\boldsymbol{\beta}$ is identical to (2.2.6) and therefore

$$\mathrm{Var}\big[\hat{\boldsymbol{\beta}}\big] = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$$

This result is identical for ML and REML estimation of $\boldsymbol{\beta}$ discussed in this chapter.

The scale parameters, $\boldsymbol{\lambda}$, appear non-linearly in the ML and REML scoring equations given by (3.2.8) and (3.3.9) respectively. The distribution of these estimators is unknown. Commonly, for non-linear models such as GLMs, large sample inference is used to ascertain approximate properties of the estimators (see Cox & Hinkley, 1974). Under large sample theory the estimated scale parameters are said to converge to the distribution

$$\hat{\boldsymbol{\lambda}} \sim N\big(\boldsymbol{\lambda}, (\mathcal{I}_e(\boldsymbol{\lambda}, \boldsymbol{\lambda})^{-1}\big)$$

For the models considered in this chapter the variance for the ML estimate of $\boldsymbol{\lambda}$ can be immediately written as

$$\mathrm{Var}\big[\hat{\boldsymbol{\lambda}}\big] = 2(\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-2}\dot{\boldsymbol{S}})^{-1}$$

and for REML

$$\mathrm{Var}\big[\hat{\boldsymbol{\lambda}}\big] = 2(\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{S}})^{-1}$$

3.4.1 Tests of Hypothesis

For given $\boldsymbol{\beta}$, the non-linearity of the ML and REML scoring equations for the scale parameters suggests the use of an asymptotic hypothesis test. A common asymptotic test used for scale parameter models considered in this chapter is the score test (see Cox & Hinkley, 1974).

The test is constructed generally by considering a null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ where $\{\boldsymbol{\theta}, \boldsymbol{\theta}_0\} \in \boldsymbol{\Theta}$. Let $\boldsymbol{U}(\boldsymbol{\theta})$ and $\mathcal{I}_o(\boldsymbol{\theta}, \boldsymbol{\theta})$ be the score and observed information for $\boldsymbol{\theta}$. Then

for a response $\boldsymbol{y} = (y_1, \ldots, y_n)$, the log-likelihood ratio statistic can be expanded in a Taylor series around $\boldsymbol{\theta}_0$

$$\mathcal{S} = 2\Big(\ell(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \boldsymbol{U}(\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathcal{I}_o(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\Big) \quad (3.4.1)$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate achieved by allowing $\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})/\partial \boldsymbol{\theta} = \boldsymbol{U}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}$ and solving for $\boldsymbol{\theta}$. Using (2.3.11) and replacing the observed information with the expected information, a first order linearisation of the score equation around its true value $\boldsymbol{\theta}_0$ gives

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = (\mathcal{I}_e(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0))^{-1} \boldsymbol{U}(\boldsymbol{\theta}_0)$$

Substituting this into (3.4.1), and replacing the observed information with the expected, the final asymptotic score statistic can be written as

$$\mathcal{S} = \boldsymbol{U}(\boldsymbol{\theta}_0)^T \mathcal{I}_e(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)^{-1} \boldsymbol{U}(\boldsymbol{\theta}_0) \quad (3.4.2)$$

From Cox & Hinkley (1974) this statistic has a limiting chi-square distribution with degrees of freedom equal to $dim(\hat{\boldsymbol{\theta}}) - dim(\boldsymbol{\theta}_0)$, the difference between the number of estimated parameters and the number of parameters contained in the null hypothesis. Further details are available in Cox & Hinkley (1974).

Following Breusch & Pagan (1979) and Cook & Weisberg (1983) a score test for homogeneity of the scale parameter in the Gaussian case is constructed as follows. Let the null hypothesis be $H_0;\ \lambda_j = 0,\ j = 1, \ldots, q-1$. Thus, under H_0, $\sigma^2(\boldsymbol{z}_i; \boldsymbol{\lambda}) = \sigma_0^2,\ i = 1, \ldots, n$. Let $d_i = (y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})^2$ where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{y}$ and $\boldsymbol{\Sigma}_0^{-1}$ has ith diagonal element $1/\hat{\sigma}_0^2$. Here, $\hat{\sigma}_0^2 = n^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$. The score test statistic for testing homogeneity of the scale model under ML is

$$\mathcal{S} = \frac{1}{2\hat{\sigma}_0^2} \boldsymbol{a}^T \dot{\boldsymbol{S}} (\dot{\boldsymbol{S}}^T \dot{\boldsymbol{S}})^{-1} \dot{\boldsymbol{S}}^T \boldsymbol{a} \quad (3.4.3)$$

where $\boldsymbol{a}$ is a vector of length n with ith component $d_i/\hat{\sigma}_0^2 - 1$. Following Cook & Weisberg (1983), $\mathcal{S}$ can be viewed as one-half of the regression sum of squares of $\bar{\boldsymbol{d}}/\hat{\sigma}_0^2 - \boldsymbol{1}_n$ on $\dot{\boldsymbol{S}}$ and has an asymptotic χ^2 distribution with $q-1$ degrees of freedom.

Similarly, the score test statistic for testing homogeneity under REML is

$$\mathcal{S}_r = \frac{1}{2\hat{\sigma}_0^2} \boldsymbol{a}^{*T} \dot{\boldsymbol{S}} (\dot{\boldsymbol{S}}^T \boldsymbol{V}_0 \dot{\boldsymbol{S}})^{-1} \dot{\boldsymbol{S}}^T \boldsymbol{a}^* \quad (3.4.4)$$

where $\boldsymbol{a}^*$ has ith component $d_i/\hat{\sigma}_0^2 - 1 + h_{0ii}$. Here, $\hat{\sigma}_0^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/(n-p)$, the unbiased estimator of σ^2 and $\boldsymbol{V}_0$ has diagonals $(1 - h_{0ii})^2$ and off diagonals h_{0ij} where $h_{0ij} = \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_j$. The score statistics (3.4.3) and (3.4.4) generalize the statistics derived by Verbyla (1993).

3.5 Computation and Software

The process for computing the estimates of the parameters for ML and REML is iterative. At the mth iteration

- For given $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(m)}$ update $\boldsymbol{\beta}$ using $\boldsymbol{\beta}^{(m+1)} = f_g(\boldsymbol{\beta}^{(m)}, \boldsymbol{\lambda}^{(m)})$ where $f_g(\cdot)$ is given by (3.2.7).
- **ML** : For given $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m+1)}$ update $\boldsymbol{\lambda}$ using $\boldsymbol{\lambda}^{(m+1)} = g_g(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\lambda}^{(m)})$ where $g_g(\cdot)$ is given by (3.2.8).

  OR
- **REML** : For given $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m+1)}$ update $\boldsymbol{\lambda}$ using $\boldsymbol{\lambda}^{(m+1)} = g_g^*(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\lambda}^{(m)})$ where $g_g^*(\cdot)$ is given by (3.3.9).

Software implementing this particular algorithm for ML and REML has been developed by Gordon Smyth and can be found at

`http://www.stasci.org/s/dglm.html`

The software contains functions to estimate the location and scale parameters of the heteroscedastic Gaussian using ML or REML. A variety of other location and scale parameter models can also be fitted (see documentation for `dglm()`). Summary functions provide appropriate standard errors for both location and scale parameters and testing between nested models is also available. Documentation for the major function `dglm()` can be found online at the same location as above. This software is only available in S-Plus but, at the time of writing this thesis, is being ported to R (see Ihaka & Gentleman, 1996).

Chapter 4

Heteroscedastic t-ML with known degrees of freedom

If outliers are present in the data then the parameter estimates obtained from the heteroscedastic Gaussian model presented in the previous chapter may be misleading. To provide a more robust approach to estimation of the location and scale parameters in the presence of heteroscedasticity a new formulation of the t-distribution is considered where, in a similar manner to the previous chapter, the scale parameter is modelled using covariates in the data.

Scale parameter modelling using the heteroscedastic t-distribution is a new area of research and in this chapter methods for ML estimation of the location and scale parameters will be derived for known degrees of freedom. Methodology to estimate the location, scale and degrees of freedom parameters using ML in the presence of heteroscdasticity are explored in Section 8.1 of this thesis. The chapter also presents an extension of the asymptotic tests considered in Section 3.4.1 to include tests of hypotheses of the location and scale parameters for the heteroscedastic t. A further extension to allow the detection of heteroscedasticity when the response is t distributed is also developed.

Fixing the degrees of freedom of the t-distribution has been a common approach in robust regression. Lange et al. (1989) exploit this technique to profile the log-likelihood for a given ν. The authors also suggest that for small data sets the degrees of freedom should be fixed to avoid estimation problems (see Section 8.1.1). This tactic is utilised by James et al. (1993) by setting $\nu = 3$, which thereby provides a degree of robustness. Fixing the degrees of freedom also ensures that the t-distribution is a member of the location-scale family and therefore has attractive properties (see Barndorff-Nielsen, 1994). These properties are maintained under the heteroscedastic model examined in this thesis.

4.1 Properties of the t distribution

The univariate t-distribution was first discussed by "Student" (1908). Fisher (1925) considers the random variable defined by the form

$$t_\nu = U\left(\chi^2_\nu/\nu\right)^{-1/2}$$

where U is a unit standard normal and independent of χ^2_ν. The distribution of this random variable is called the t-distribution. This t-distribution which is denoted by $t(0,1,\nu)$ can be extended as follows. If $y|\omega \sim \mathrm{N}(\mu, \sigma^2/\omega)$ and $\omega \sim \chi^2_\nu/v$ then $y \sim t(\mu, \sigma^2, \nu)$. This t-distribution includes a location parameter, μ and a scale parameter, σ^2. The following properties of $t(\mu, \sigma^2, \nu)$ are used in the development of this and future chapters.

Property 1 *The probability density function for y is*

$$\frac{\Gamma((\nu+1)/2)}{(\Gamma(1/2))\Gamma(\nu/2)(\sigma^2\nu)^{1/2}}\left(1+d/\sigma^2\nu\right)^{-(\frac{\nu+1}{2})}$$

where $d = (y-\mu)^2$.

Property 2

$$\frac{d}{\sigma^2} \sim F_{1,\nu}.$$

where d is defined in Property 1.

Property 3

$$\frac{d}{\sigma^2\nu} \sim B_2\Big(\frac{1}{2}, \frac{\nu}{2}\Big)$$

where $B_2(\cdot,\cdot)$ is a Beta distribution of the second kind.

Property 4 *Using Property 2*

$$\frac{d/\sigma^2\nu}{1+d/\sigma^2\nu} \sim B_1\Big(\frac{1}{2}, \frac{\nu}{2}\Big) \tag{4.1.1}$$

where $B_1(\cdot,\cdot)$ is a Beta distribution of the first kind.

Property 5 $\omega|y \sim \chi^2_{\nu+1}/(\nu + d/\sigma^2\nu)$

Property 6 $E(y) = \mu\ (\nu > 1), \quad Var(y) = \nu\sigma^2/(\nu-2)\ (\nu > 2)$

Property 7 *If $\nu \to \infty$ then*

$$t(\mu, \sigma^2, \nu) \to N(\mu, \sigma^2) \tag{4.1.2}$$

For a more comprehensive overview see Johnson et al. (1995). These properties will be used in the proceeding sections to estimate the location and scale parameters as well as to predict the scale random effects, ω associated with the t-distribution.

4.2 Notation

Consider an extension relevant to a sample from independent t-distributions

$$y_i|\omega_i \sim \mathrm{N}\left(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma_i^2/\omega_i\right), \qquad i = 1, \ldots, n \tag{4.2.1}$$

where y_i is the ith observed response, $\boldsymbol{x}_i$ is a $p \times 1$ vector of explanatory variables, $\boldsymbol{\beta}$ is a set of unknown parameters and $\omega_i, i = 1, \ldots, n$ represents a set of random variables with distributional form

$$\omega_i \sim \mathrm{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \qquad i = 1, \ldots, n \tag{4.2.2}$$

where ν is an unknown parameter, alternatively

$$\omega_i \sim \chi^2_\nu/\nu, \qquad i = 1, \ldots, n$$

As in the Gaussian case of Chapter 3 the scale parameters are assumed to be of the form

$$\varphi_i = \sigma_i^2/\omega_i = \sigma^2(\boldsymbol{z}_i; \boldsymbol{\lambda})/\omega_i, \qquad 1 = 1, \ldots, n \tag{4.2.3}$$

where $\boldsymbol{z}_i$ is a $q \times 1$ vector of explanatory variables (with possibly some components in common with $\boldsymbol{x}_i$) and $\boldsymbol{\lambda}$ is a set of unknown parameters.

Under such a hierarchy the marginal likelihood can be expressed as

$$\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y}) = \int_{\mathcal{R}^n} \prod_{i=1}^n p(y_i|\omega_i; \boldsymbol{\beta}, \boldsymbol{\lambda}) p(\omega_i; \nu) d\boldsymbol{\omega} \tag{4.2.4}$$

where

$$\prod_{i=1}^n p(y_i|\omega_i; \boldsymbol{\beta}, \boldsymbol{\lambda}) = (2\pi)^{-n/2}|\boldsymbol{\Psi}|^{-1/2} \exp\Big\{-\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\Big\} \tag{4.2.5}$$

$$\prod_{i=1}^n p(\omega_i; \nu) = \frac{(\nu/2)^{n\nu/2}}{(\Gamma(\nu/2))^n} \exp\Big\{-\tfrac{\nu}{2}\sum_{i=1}^n \omega_i\Big\} \prod_{i=1}^n \omega_i^{\nu/2-1} \tag{4.2.6}$$

and $\boldsymbol{\Psi}$ is a diagonal matrix with ith diagonal element σ_i^2/ω_i.

In this special case the multi-dimensional integral is tractable and clearly, marginally,

$$y_i \sim t(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma_i^2, \nu) \qquad i = 1, \ldots, n \tag{4.2.7}$$

the t-distribution with ν degrees of freedom, mean $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, $(\nu > 1)$, variance $\nu\sigma_i^2/(\nu - 2)$, $(\nu > 2)$. Let the ith diagonal element of $\boldsymbol{\Sigma}$ be σ_i^2 then (4.2.4) becomes

$$\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y}) = |\boldsymbol{\Sigma}|^{-1/2} \left\{\frac{\Gamma((\nu+1)/2)}{(\Gamma(1/2))\Gamma(\nu/2)\nu^{1/2}}\right\}^n \prod_{i=n}^n \left\{1 + \frac{d_i}{\sigma_i^2\nu}\right\}^{-(\frac{\nu+1}{2})} \tag{4.2.8}$$

where

$$d_i = (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$$

Note that as $\nu \to \infty$ the random variable, y_i tends to a Gaussian distribution with location $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ and scale parameters $\sigma_i^2,\ i = 1, \ldots, n$. The log-likelihood of (4.2.8) is

$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y}) &= n\log\big(\Gamma((\nu+1)/2)\big) - n\log\big(\Gamma(1/2)\big) - n\log\big(\Gamma(\nu/2)\big) \\ &- \frac{n}{2}\log\nu - \frac{1}{2}\sum_{i=1}^{n}\log\sigma_i^2 - \frac{\nu+1}{2}\sum_{i=1}^{n}\log\left\{1 + \frac{d_i}{\sigma_i^2\nu}\right\} \qquad (4.2.9)\end{aligned}$$

4.3 Estimation of Parameters

4.3.1 Score Equations

If the degrees of freedom, ν, is known only the last two terms of (4.2.9) are of interest and these terms form the kernel of the log-likelihood used in estimating the location and scale parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\lambda})$. Thus consider

$$\ell(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y}) = -\frac{1}{2}\sum_{i=1}^{n}\log\sigma^2(\boldsymbol{z}_i; \boldsymbol{\lambda}) - \frac{\nu+1}{2}\sum_{i=1}^{n}\log\left\{1 + d_i/\sigma^2(\boldsymbol{z}_i; \boldsymbol{\lambda})\nu\right\} \qquad (4.3.1)$$

Differentiating (4.3.1) with respect to the lth location parameter β_l gives

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y})}{\partial\beta_l} &= \sum_{i=1}^{n}\left(\frac{\nu+1}{1 + d_i/\sigma^2\nu}\right)\frac{(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})}{\sigma_i^2\nu}x_{il} \\ &= \sum_{i=1}^{n}\frac{\bar{\omega}_i}{\sigma_i^2}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})x_{il}\end{aligned}$$

where

$$\bar{\omega}_i = \frac{\nu+1}{\nu + d_i/\sigma_i^2} \qquad (4.3.2)$$

Differentiating (4.3.1) with respect to the jth scale parameter λ_j gives

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y})}{\partial\lambda_j} &= \frac{1}{2}\sum_{i=1}^{n}\left\{\left(\frac{\nu+1}{1 + d_i/\sigma^2\nu}\right)\frac{d_i}{(\sigma_i^2)^2\nu}\frac{\partial\sigma_i^2}{\partial\lambda_j} - \frac{1}{\sigma_i^2}\frac{\partial\sigma_i^2}{\partial\lambda_j}\right\} \\ &= \frac{1}{2}\sum_{i=1}^{n}\left\{\frac{\dot{s}_{ij}}{\sigma_i^2}\left(\frac{\bar{\omega}_i d_i}{\sigma_i^2} - 1\right)\right\}\end{aligned}$$

where $\dot{s}_{ij} = \partial\sigma_i^2/\partial\lambda_j$.

As in Section 3.2.1, let $\dot{\boldsymbol{S}}$ be a matrix of partial derivatives with ijth element $\dot{s}_{ij}$ and $\bar{\boldsymbol{\Omega}}$ be an $n \times n$ diagonal matrix with ith diagonal element $\bar{\omega}_i$. The score equations are then

$$\boldsymbol{U}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{U}(\boldsymbol{\beta}) \\ \boldsymbol{U}(\boldsymbol{\lambda}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{\Omega}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ \frac{1}{2}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{\Omega}}\boldsymbol{d} - \boldsymbol{1}_n) \end{bmatrix} \tag{4.3.3}$$

Note that the score equations for $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ defined by (4.3.3) may also be written as

$$\boldsymbol{U}(\boldsymbol{\beta}) = (\nu+1)\sum_{i=1}^{n} \frac{\boldsymbol{x}_i}{\sigma_i^2\nu}(1 - B_i)(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) \tag{4.3.4}$$

$$\boldsymbol{U}(\boldsymbol{\lambda}) = \frac{1}{2}\sum_{i=1}^{n} \frac{\dot{\boldsymbol{s}}_i}{\sigma_i^2}\{(\nu+1)B_i - 1\} \tag{4.3.5}$$

where

$$B_i = \frac{d_i/\sigma_i^2\nu}{1 + d_i/\sigma_i^2\nu} \tag{4.3.6}$$

Using Property (4) the distribution of this ith term is then

$$B_i \sim B_1\Big(\frac{1}{2}, \frac{\nu}{2}\Big)$$

where $B_1(\cdot,\cdot)$ describes a Beta distribution of the first kind. Its first two moments can be immediately expressed as

$$\mathrm{E}\big[B_i\big] = \frac{1}{\nu+1} \tag{4.3.7}$$

$$\mathrm{E}\left[B_i^2\right] = \frac{3}{(\nu+1)(\nu+3)} \tag{4.3.8}$$

The expectation of the score for $\boldsymbol{\beta}$ can then be written as

$$\begin{aligned} \mathrm{E}[\boldsymbol{U}(\boldsymbol{\beta})] &= (\nu+1)\sum_{i=1}^{n} \frac{\boldsymbol{x}_i}{\sigma_i^2\nu}\mathrm{E}\big[(1 - B_i)(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\big] \\ &= (\nu+1)\sum_{i=1}^{n} \frac{\boldsymbol{x}_i}{\sigma_i^2\nu}\left\{\mathrm{E}\big[(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\big] + \mathrm{E}\big[B_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\big]\right\} \\ &= 0 \end{aligned}$$

The function in the second expectation term is an odd function of y_i and therefore, around a symmetric interval, such as the t-distribution, will have expectation zero. Similarly, the expectation of the score for $\boldsymbol{\lambda}$ is

$$\begin{aligned} \mathrm{E}[\boldsymbol{U}(\boldsymbol{\lambda})] &= \frac{1}{2}\sum_{i=1}^{n} \frac{\dot{\boldsymbol{s}}_i}{\sigma_i^2}\big\{(\nu+1)\mathrm{E}\big[B_i\big] - 1\big\} \\ &= 0 \end{aligned}$$

4.3.2 Solving the Score Equations

As in Section 3.2.2 the score equations given by (4.3.3) must be solved by an iterative scheme (3.2.6). The observed information,

$$\mathcal{I}_o(\boldsymbol{\theta},\boldsymbol{\theta}) = \begin{bmatrix} \mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\beta}) & \mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\lambda}) \\ \mathcal{I}_o(\boldsymbol{\lambda},\boldsymbol{\beta}) & \mathcal{I}_o(\boldsymbol{\lambda},\boldsymbol{\lambda}) \end{bmatrix}$$

is required. To obtain this (4.3.4) and (4.3.5) are partially differentiated with respect to β_m and λ_k. The lmth observed information element for $\boldsymbol{\beta}$ can be then expressed as

$$\begin{aligned} \mathcal{I}_o(\beta_l,\beta_m) &= (\nu+1)\sum_{i=1}^{n}\frac{x_{il}x_{im}}{\sigma_i^2\nu}\left\{\left(\frac{1}{1+d_i/\sigma_i^2\nu}\right)-\left(\frac{2}{1+d_i/\sigma_i^2\nu}\right)\left(\frac{d_i/\sigma_i^2\nu}{1+d_i/\sigma_i^2\nu}\right)\right\} \\ &= (\nu+1)\sum_{i=1}^{n}\frac{x_{il}x_{im}}{\sigma_i^2\nu}\left\{(1-B_i)-2(1-B_i)B_i\right\} \\ &= (\nu+1)\sum_{i=1}^{n}\frac{x_{il}x_{im}}{\sigma_i^2\nu}\left\{(1-B_i)(1-2B_i)\right\} \end{aligned} \tag{4.3.9}$$

Let $\dot{s}_{i(jk)} = \partial\dot{s}_{ij}/\partial\lambda_k$. The jkth observed information component for $\boldsymbol{\lambda}$ can be expressed as

$$\mathcal{I}_o(\lambda_j,\lambda_k) = \frac{1}{2}\sum_{i=1}^{n}\left\{\left(\frac{\dot{s}_{ij}\dot{s}_{ik}}{(\sigma_i^2)^2}-\frac{\dot{s}_{i(jk)}}{\sigma_i^2}\right)((\nu+1)B_i-1)-\frac{\dot{s}_{ij}}{\sigma_i^2}(\nu+1)\frac{\partial B_i}{\partial\lambda_k}\right\} \tag{4.3.10}$$

where

$$\begin{aligned} \frac{\partial B_i}{\partial\lambda_k} &= \left(\frac{d_i/\sigma_i^2\nu}{(1+d_i/\sigma_i^2\nu)^2}\right)\frac{d_i}{(\sigma_i^2)^2\nu}\dot{s}_{ik}-\left(\frac{d_i/(\sigma_i^2)^2\nu}{1+d_i/\sigma_i^2\nu}\right)\dot{s}_{ik} \\ &= \left(\frac{d_i/\sigma_i^2\nu}{1+d_i/\sigma_i^2\nu}\right)^2\frac{\dot{s}_{ik}}{\sigma_i^2}-\left(\frac{d_i/\sigma_i^2\nu}{1+d_i/\sigma_i^2\nu}\right)\frac{\dot{s}_{ik}}{\sigma_i^2} \\ &= B_i(B_i-1)\frac{\dot{s}_{ik}}{\sigma_i^2} \end{aligned}$$

Substituting this into (4.3.10) the jkth observed information component for $\boldsymbol{\lambda}$ is

$$\mathcal{I}_o(\lambda_j,\lambda_k) = \frac{1}{2}\sum_{i=1}^{n}\left\{\frac{\dot{s}_{i(jk)}}{\sigma_i^2}((\nu+1)B_i-1)+\frac{\dot{s}_{ij}\dot{s}_{ik}}{(\sigma_i^2)^2}\left((\nu+1)(2B_i-B_i^2)-1\right)\right\} \tag{4.3.11}$$

The lkth co-information is obtained by considering the partial derivative of (4.3.4) with respect to λ_k. This gives

$$\begin{aligned}
\mathcal{I}_o(\beta_l, \lambda_k) &= (\nu+1)\sum_{i=1}^{n}\left\{\left(\frac{1}{1+d_i/\sigma_i^2\nu}\right)\frac{x_{il}\dot{s}_{ik}}{(\sigma_i^2)^2\nu}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\right. \\
&- \left.\left(\frac{1}{1+d_i/\sigma_i^2\nu}\right)\left(\frac{d_i/\sigma_i^2\nu}{1+d_i/\sigma_i^2\nu}\right)\frac{x_{il}\dot{s}_{ik}}{(\sigma_i^2)^2\nu}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\right\} \\
&= (\nu+1)\sum_{i=1}^{n}\frac{x_{il}\dot{s}_{ik}}{(\sigma_i^2)^2\nu}\left\{1 - B_i - (1-B_i)B_i\right\}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) \\
&= (\nu+1)\sum_{i=1}^{n}\frac{x_{il}\dot{s}_{ik}}{(\sigma_i^2)^2\nu}\left\{(1-B_i)(1-B_i)\right\}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) \qquad (4.3.12)
\end{aligned}$$

To obtain a Fisher scoring algorithm the expected information must be found. Noting (4.3.7) and (4.3.8) and taking the expectation of (4.3.9) the lmth element of the expected information for $\boldsymbol{\beta}$ can be expressed as

$$\begin{aligned}
\mathcal{I}_e(\beta_l, \beta_m) &= (\nu+1)\sum_{i=1}^{n}\frac{x_{il}x_{im}}{\sigma_i^2\nu}\left\{(1 - \frac{3}{\nu+1} + \frac{6}{(\nu+1)(\nu+3)}\right\} \\
&= \sum_{i=1}^{n}\frac{x_{il}x_{im}}{\sigma_i^2\nu}\left\{\nu - 2 + \frac{6}{\nu+3}\right\} \\
&= \sum_{i=1}^{n}\frac{x_{il}x_{im}}{\sigma_i^2}\left\{\frac{\nu+1}{\nu+3}\right\} \qquad (4.3.13)
\end{aligned}$$

Similarly, taking the expectation of (4.3.11) the jkth element of the expected information for $\boldsymbol{\lambda}$ can be expressed as

$$\begin{aligned}
\mathcal{I}_e(\lambda_j, \lambda_k) &= \frac{1}{2}\sum_{i=1}^{n}\frac{\dot{s}_{ij}\dot{s}_{ik}}{(\sigma_i^2)^2}\left\{(\nu+1)\left(\frac{2}{\nu+1} - \frac{3}{(\nu+1)(\nu+3)}\right) - 1\right\} \\
&= \frac{1}{2}\sum_{i=1}^{n}\frac{\dot{s}_{ij}\dot{s}_{ik}}{(\sigma_i^2)^2}\left\{1 - \frac{3}{\nu+3}\right\} \\
&= \frac{1}{2}\sum_{i=1}^{n}\frac{\dot{s}_{ij}\dot{s}_{ik}}{(\sigma_i^2)^2}\left\{\frac{\nu}{\nu+3}\right\}
\end{aligned}$$

Lastly the lkth element of the expected co-information for $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ can be expressed as

$$\mathcal{I}_e(\beta_l, \lambda_k) = (\nu+1)\sum_{i=1}^{n}\frac{x_{il}\dot{s}_{ik}}{(\sigma_i^2)^2\nu}\mathrm{E}\left[\left\{B_i^2 - 2B_i\right\}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\right]$$

The terms remaining in the expectation are odd functions of y_i and therefore, around a symmetrical interval, will have zero expectation. In matrix notation the expected

information for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\lambda})$ can be written as

$$\begin{aligned} \mathcal{I}_e(\boldsymbol{\theta}, \boldsymbol{\theta}) &= \begin{bmatrix} \mathcal{I}_e(\boldsymbol{\beta}, \boldsymbol{\beta}) & \mathcal{I}_e(\boldsymbol{\beta}, \boldsymbol{\lambda}) \\ \mathcal{I}_e(\boldsymbol{\lambda}, \boldsymbol{\beta}) & \mathcal{I}_e(\boldsymbol{\lambda}, \boldsymbol{\lambda}) \end{bmatrix} && (4.3.14) \\ &= \begin{bmatrix} \frac{\nu+1}{\nu+3} \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X} & \mathbf{0} \\ \mathbf{0} & \frac{\nu}{2(\nu+3)} \dot{\boldsymbol{S}}^T \boldsymbol{\Sigma}^{-2} \dot{\boldsymbol{S}} \end{bmatrix} && (4.3.15) \end{aligned}$$

The expected information can also be derived from taking the variance of the score function, $\boldsymbol{U}(\boldsymbol{\theta})$. Noting that

$$\begin{aligned} \mathrm{Var}\big[B_i\big] &= \mathrm{E}\big[B_i^2\big] - \mathrm{E}\big[B_i\big]\mathrm{E}\big[B_i\big] \\ &= \frac{3}{(\nu+1)(\nu+3)} - \left(\frac{1}{\nu+1}\right)^2 \\ &= \frac{2\nu}{(\nu+1)^2(\nu+3)} \end{aligned}$$

the variance of the score for $\boldsymbol{\beta}$ can be expressed as

$$\begin{aligned} \mathrm{Var}\big[\boldsymbol{U}(\boldsymbol{\beta})\big] &= (\nu+1)^2 \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_i^2 \nu} \mathrm{E}\left\{ \frac{d_i/\sigma_i^2 \nu}{(1 + d_i/\sigma^2 \nu)^2} \right\} \\ &= (\nu+1)^2 \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_i^2 \nu} \mathrm{E}\left\{ B_i - B_i^2 \right\} \\ &= (\nu+1)^2 \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_i^2 \nu} \left\{ \frac{1}{\nu+1} - \frac{3}{(\nu+1)(\nu+3)} \right\} \\ &= (\nu+1) \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_i^2 \nu} \left\{ 1 - \frac{3}{\nu+3} \right\} \\ &= \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_i^2} \left\{ \frac{\nu+1}{\nu+3} \right\} \;=\; \mathcal{I}_e(\boldsymbol{\beta}, \boldsymbol{\beta}) \end{aligned}$$

Similarly taking the variance of the score for $\boldsymbol{\lambda}$ gives

$$\begin{aligned} \mathrm{Var}\big[\boldsymbol{U}(\boldsymbol{\lambda})\big] &= \frac{(\nu+1)^2}{4} \sum_{i=1}^{n} \frac{\dot{\boldsymbol{s}}_i \dot{\boldsymbol{s}}_i^T}{(\sigma_i^2)^2} \mathrm{Var}\left\{ B_i \right\} \\ &= \frac{(\nu+1)^2}{4} \sum_{i=1}^{n} \frac{\dot{\boldsymbol{s}}_i \dot{\boldsymbol{s}}_i^T}{(\sigma_i^2)^2} \left\{ \frac{2\nu}{(\nu+1)^2(\nu+3)} \right\} \\ &= \frac{1}{2} \sum_{i=1}^{n} \frac{\dot{\boldsymbol{s}}_i \dot{\boldsymbol{s}}_i^T}{(\sigma_i^2)^2} \left\{ \frac{\nu}{\nu+3} \right\} \;=\; \mathcal{I}_e(\boldsymbol{\lambda}, \boldsymbol{\lambda}) \end{aligned}$$

The orthogonality of the parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ can be shown by taking the covariance of the

score functions, namely

$$\begin{aligned}\mathrm{Cov}\left[\boldsymbol{U}(\boldsymbol{\beta}),\boldsymbol{U}(\boldsymbol{\lambda})\right] &= \mathrm{E}\left[\boldsymbol{U}(\boldsymbol{\beta})\boldsymbol{U}(\boldsymbol{\lambda})^T\right] \\ &= \frac{\nu+1}{2}\sum_{i=1}^{n}\frac{\boldsymbol{x}_i\boldsymbol{s}_i^T}{(\sigma_i^2)^2\nu}\left\{\mathrm{E}\left[((\nu+1)B_i-1)(1-B_i)(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})\right]\right\} \\ &= 0\end{aligned}$$

The orthogonality of the location and scale parameters suggests that each parameter may be scored independently. Using (3.2.6) the $(m+1)$th iterate of the Fisher scoring algorithm for the location parameter can be written as

$$\boldsymbol{\beta}_{(m+1)} = \boldsymbol{\beta}_m + \tfrac{\nu+3}{\nu+1}(\boldsymbol{X}^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_m^{-1}\bar{\boldsymbol{\Omega}}_m(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_m)$$

Subsuming $\boldsymbol{\beta}_m$, the implicit equation for the location parameters can be reduced to

$$\boldsymbol{\beta}_{(m+1)} = f_t(\boldsymbol{\beta}_m,\boldsymbol{\lambda},\nu) = (\boldsymbol{X}^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{y}_m^* \tag{4.3.16}$$

where

$$\boldsymbol{y}_m^* = \tfrac{\nu+3}{\nu+1}\bar{\boldsymbol{\Omega}}_m(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_m)+\boldsymbol{X}\boldsymbol{\beta}_m$$

Similar to Section 3.2.2, for given $(\boldsymbol{\lambda},\nu)$, (4.3.16) can be viewed as an iterative reweighted least squares (IRLS) algorithm with $1/\sigma_i^2$ as its ith diagonal weight and working vector $\boldsymbol{y}^*$.

Solving the score equation for $\boldsymbol{\beta}$ given in (4.3.3) directly allows a second estimating equation to be immediately written as

$$\boldsymbol{\beta}_{(m+1)} = (\boldsymbol{X}^T\bar{\boldsymbol{\Psi}}_m^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\bar{\boldsymbol{\Psi}}_m^{-1}\boldsymbol{y} \tag{4.3.17}$$

where $\bar{\boldsymbol{\Psi}}$ is a diagonal matrix with ith diagonal element $\bar{\varphi}_i = \bar{\omega}_i/\sigma_i^2$. This is also an iteratively reweighted least squares algorithm with ith weight $\bar{\varphi}_i$. Notice as (4.3.16) and (4.3.17) are derived from the same score equation they are theoretically equivalent.

The $(m+1)$th iterate of the Fisher scoring algorithm for the fixed scale parameters can be written as

$$\boldsymbol{\lambda}_{(m+1)} = \boldsymbol{\lambda}_m + (\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}_m^{-2}\dot{\boldsymbol{S}})^{-1}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\Sigma}_m^{-1}\bar{\boldsymbol{\Omega}}_m\boldsymbol{d}-\boldsymbol{1}_n)$$

Subsuming $\boldsymbol{\lambda}_m$ the implicit scoring equation for the scale parameters can be expressed as

$$\boldsymbol{\lambda}_{(m+1)} = g_t(\boldsymbol{\beta},\boldsymbol{\lambda}_m,\nu) = (\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}_m^{-2}\dot{\boldsymbol{S}})^{-1}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}_m^{-2}\boldsymbol{d}_m^*, \tag{4.3.18}$$

where

$$\boldsymbol{d}_m^* = \frac{\nu+3}{\nu}(\bar{\boldsymbol{\Omega}}_m\boldsymbol{d}-\boldsymbol{\Sigma}_m\boldsymbol{1}_n+\dot{\boldsymbol{S}}\boldsymbol{\lambda}_m) \tag{4.3.19}$$

Therefore, for given $(\boldsymbol{\beta}, \nu)$, (4.3.18) can be viewed as an iterative reweighted least squares procedure with equal weights $1/(\sigma_i^2)^2$ and working vector $\boldsymbol{d}^*$.

In particular if a simplified log-linear scale parameter model defined by (3.1.3) is assumed then the scoring algorithm for $\boldsymbol{\lambda}$ reduces to

$$\boldsymbol{\lambda}_{(m+1)} = \frac{\nu+3}{\nu}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T(\bar{\boldsymbol{\Omega}}_m^{-1}\boldsymbol{d} - \mathbf{1}_n + \boldsymbol{Z}\boldsymbol{\lambda}_m)$$

4.4 Prediction

The random effects $\omega_i,\ i = 1, \ldots, n$ are not present in the marginal distribution for $\boldsymbol{y}$ and therefore must be predicted using an alternate distribution. Standard prediction theory, Searle et al. (1992), suggests choosing the mean of the conditional distribution, $w_i|y_i$ to obtain the best unbiased predictor (BUP) of ω_i. For the ith component of the conditional distribution the density can expressed as

$$p(\omega_i|y_i; \boldsymbol{\beta}, \sigma_i^2, \nu) = \frac{p(y_i|\omega_i; \boldsymbol{\beta}, \sigma_i^2)p(\omega_i; \nu)}{\int p(y_i|\omega_i; \boldsymbol{\beta}, \sigma_i^2)p(\omega_i; \nu)d\omega_i} \tag{4.4.1}$$

where

$$\begin{aligned} p(y_i|\omega_i; \boldsymbol{\beta}, \sigma_i^2) &= (2\pi\sigma_i^2)^{-1/2}\omega_i^{1/2}\exp(-\omega_i d_i/2\sigma_i^2) \\ p(\omega_i; \nu) &= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\omega_i^{\nu/2-1}\exp(-\nu\omega_i/2) \end{aligned}$$

The denominator of the RHS is is the marginal density for the ith response of the heteroscedastic t-distribution. Using Property (1) of Section 4.1 the denominator can be expressed as

$$p(y_i; \boldsymbol{\beta}, \sigma_i^2, \nu) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\Gamma(1/2)}(\sigma_i^2\nu)^{-1/2}(\nu + d_i/\sigma_i^2)^{-(\nu+1)/2}$$

Combining these (4.4.1) can be written as

$$p(\omega_i|y_i; \boldsymbol{\beta}, \sigma_i^2, \nu) = \frac{(2\pi)^{-1/2}\Gamma(1/2)\omega^{(\nu-1)/2}\exp\left(-\omega_i\nu/2 - \omega_i d_i/\sigma_i^2/2\right)}{\Gamma((\nu+1)/2)\big(\nu + d_i/\sigma_i^2\big)^{-(\nu+1)/2}}$$

and the density can be expressed

$$p(\omega_i|y_i; \boldsymbol{\beta}, \sigma_i^2, \nu) = \frac{\big(\nu + d_i/\sigma_i^2\big)^{(\nu+1)/2}}{\Gamma((\nu+1)/2)}\omega_i^{(\nu+1)/2-1}\exp\left(-\omega_i(\nu + \omega_i d_i/\sigma_i^2)/2\right)$$

Therefore the conditional distribution of $\omega_i|y_i$ can be expressed as

$$\omega_i|y_i \sim \text{Gamma}\left(\frac{\nu+1}{2}, \frac{\nu + d_i/\sigma_i^2}{2}\right) \tag{4.4.2}$$

This derivation is equivalent to Property (5) from Section 4.1. The mean of this distribution is equivalent to its first moment and therefore the BUP for the random scale effects is

$$\mathrm{E}(\omega_i|y_i) = \tilde{\omega}_i = \frac{\nu+1}{\nu + d_i/\sigma_i^2} \tag{4.4.3}$$

Using (4.3.7), the unbiasedness of the prediction can be shown by taking expectations of both sides,

$$\mathrm{E}(\tilde{\omega}_i) = \frac{\nu+1}{\nu}\mathrm{E}(1-B_i) = 1$$

and therefore $\mathrm{E}[\tilde{\omega}_i] = \mathrm{E}_{y_i}\left(\mathrm{E}[\omega_i|y_i]\right) = 1 = \mathrm{E}[\omega_i]$. As $\tilde{\omega}_i = \bar{\omega}_i$, it is clear these predicted random effects are intimately connected to the scoring equations for the location and scale parameters given in (4.3.16) and (4.3.18) and hence prediction is a byproduct of the scoring algorithm.

4.5 Parameter Inference

It is of interest to obtain measures of precision for the estimates of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ for the heteroscedastic t-distribution.

The scoring equations (4.3.16) and (4.3.18) to estimate the parameters are iteratively reweighted least squares. Under such an estimation process the distribution of the estimators are unknown and therefore large sample inference is required. The asymptotic distribution of the estimators can be immediately written as

$$\begin{aligned} \hat{\boldsymbol{\beta}} &\sim N(\boldsymbol{\beta}, \mathcal{I}_e(\boldsymbol{\beta},\boldsymbol{\beta})^{-1}) \\ \hat{\boldsymbol{\lambda}} &\sim N(\boldsymbol{\lambda}, \mathcal{I}_e(\boldsymbol{\lambda},\boldsymbol{\lambda})^{-1}) \end{aligned}$$

where

$$\begin{aligned} \mathcal{I}_e(\boldsymbol{\beta},\boldsymbol{\beta}) &= \tfrac{\nu+1}{\nu+3}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \\ \mathcal{I}_e(\boldsymbol{\lambda},\boldsymbol{\lambda}) &= \tfrac{\nu}{2(\nu+3)}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-2}\dot{\boldsymbol{S}} \end{aligned}$$

4.5.1 Asymptotic Properties of the Estimators

Property 6 from section 4.1 shows that as the degrees of freedom increase for the t-distribution the distribution tends to a Gaussian distribution. Rewriting $\bar{\omega}_i$ as

$$\bar{\omega}_i = \frac{1}{1 - 1/(\nu+1) + d_i/\sigma_i^2(\nu+1)} \tag{4.5.1}$$

and allowing $\nu \to \infty$ then $\bar{\omega_i} \to 1$. Using this property it can be seen that the score and expected information for $\boldsymbol{\beta}$ have the asymptotic property

$$\lim_{\nu\to\infty} \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{\Omega}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$$
$$\lim_{\nu\to\infty} \left(1-\frac{2}{\nu+3}\right)\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$$

Therefore (4.3.16) and (4.3.17) reduce to the ordinary Gaussian ML estimator for the location parameter $\boldsymbol{\beta}$ given by (3.2.7).

Similarly the score and expected information for the scale parameter have the property

$$\begin{aligned}\lim_{\nu\to\infty} \frac{1}{2}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{\Omega}}\boldsymbol{d}-\mathbf{1}_n) &= \frac{1}{2}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{d}-\mathbf{1}_n)\\ \lim_{\nu\to\infty}\left(1-\frac{3}{\nu+3}\right)\frac{1}{2}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-2}\dot{\boldsymbol{S}} &= \frac{1}{2}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-2}\dot{\boldsymbol{S}}\end{aligned}$$

Therefore (4.3.18) reduces to the ordinary Gaussian ML estimator for the scale parameter $\boldsymbol{\lambda}$ given by (3.2.8).

4.5.2 Tests of Hypotheses

The non-normality of the t-distribution suggests that asymptotic tests are required for both location and scale parameters. Following from Section 3.4.1 the asymptotic test discussed here is the score test. For the location parameter, given the maximum likelihood estimates for the variance parameters, $\hat{\sigma}_i^2 = \sigma^2(\boldsymbol{z}_i;\hat{\boldsymbol{\lambda}}),\ i=1,\ldots,n$ obtained from the scoring equation given in (4.3.18), consider the conformal partitions $(\boldsymbol{X}_1,\boldsymbol{X}_2)$ and $(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)$ of $\boldsymbol{X}$ and $\boldsymbol{\beta}$ respectively. Under the null hypothesis, $H_0:\boldsymbol{\beta}_2=\mathbf{0}$, the location model is fitted using the scoring algorithm (4.3.16) with $\boldsymbol{\beta}_1$ only. To form the score test the asymptotic variance of $\hat{\boldsymbol{\beta}}_2|\hat{\boldsymbol{\beta}}_1$ is required. The asymptotic variance matrix for the partitioned estimates can be expressed as

$$\text{Var}\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{bmatrix} = \frac{\nu+3}{\nu+1}\begin{bmatrix}\boldsymbol{X}_1^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}_1 & \boldsymbol{X}_1^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}_2\\ \boldsymbol{X}_2^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}_1 & \boldsymbol{X}_2^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}_2\end{bmatrix}^{-1}$$

Using Result A.3.3 the asymptotic variance matrix for $\hat{\boldsymbol{\beta}}_2|\hat{\boldsymbol{\beta}}_1$ is then

$$\text{Var}[\hat{\boldsymbol{\beta}}_2|\hat{\boldsymbol{\beta}}_1] = \boldsymbol{X}_2^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}_2 - \boldsymbol{X}_2^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}_1(\boldsymbol{X}_1^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}_2$$

Then if $\boldsymbol{\Omega}_0^{-1} = \text{diag}\{\omega_{i,0}/\hat{\sigma}_i^2\}$ where $\omega_{i,0} = (\nu+1)/(\nu+d_{i1}/\hat{\sigma}_i^2)$ and $d_{i1} = (y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_1)$, the score statistic for the null hypothesis is given as

$$\frac{\nu+3}{\nu+1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_1)\boldsymbol{\Omega}_0^{-1}\boldsymbol{X}_2\boldsymbol{C}^{-1}\boldsymbol{X}_2^T\boldsymbol{\Omega}_0^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_1) \sim \chi^2(p_1)$$

where

$$\boldsymbol{C} = \boldsymbol{X}_2^T \boldsymbol{P}_1 \boldsymbol{X}_2 \qquad \text{and} \qquad \boldsymbol{P}_1 = \hat{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}_1(\boldsymbol{X}_1^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\hat{\boldsymbol{\Sigma}}^{-1}$$

Following Breusch & Pagan (1979) and Cook & Weisberg (1983) this statistic can be viewed as the half the sum of squares regression of $\boldsymbol{\Omega}_0^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1)$ on $\boldsymbol{X}_2$.

Similarly the scale parameter sub-model can be tested using the score and information components of the scoring algorithm given in (4.3.18). Let $d_i = (y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2$ where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{\Omega}_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Omega}_0^{-1}\boldsymbol{y}$ and $\boldsymbol{\Omega}_0^{-1}$ has ith diagonal element $w_{i,0} = (\nu+1)/(\nu + d_i/\hat{\sigma}_0^2)$. Here, $\hat{\sigma}_0^2 = n^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Omega}_0^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ is obtained iteratively. The score test statistic for testing homogeneity of the scale model for the t-distribution is then

$$\frac{\nu+3}{2\nu\hat{\sigma}_0^2}\boldsymbol{a}^{*T}\dot{\boldsymbol{S}}(\dot{\boldsymbol{S}}^T\dot{\boldsymbol{S}})^{-1}\dot{\boldsymbol{S}}^T\boldsymbol{a}^* \sim \chi^2(q-1)$$

where $\boldsymbol{a}^*$ is a vector of length n with ith component $d_i\omega_{i,0}/\hat{\sigma}_0^2 - 1$. This extends the work of Section 3.4.1 to an asymptotic test for the location and scale parameters for the heteroscedastic t-distribution.

4.6 Detecting Heteroscedasticity

It is useful to informally investigate dependence of covariates on the scale parameter for the t-distribution. Following Cook & Weisberg (1983) and Verbyla (1993), let $\bar{d}_i = (y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2$ where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ is the ordinary least squares estimator for $\boldsymbol{\beta}$. For the t-distribution, the estimation of $\boldsymbol{\beta}$ requires iterative reweighting. The distribution of the latter estimator is, exactly and asymptotically, unknown, whereas conditionally, the least squares estimator has a known distribution and facilitates the exact theory presented below.

For the t-distribution, using Property 2 and Property 6 of Section 4.1 $\bar{d}_i \sim \phi_i^2 F_{i\,,\nu}$, where

$$\phi_i^2 = \frac{\nu}{\nu-2}\left\{(1-h_{ii})^2\sigma_i^2 + \sum_{j\neq i} h_{ij}^2\sigma_j^2\right\} \tag{4.6.1}$$

and h_{ij} are elements of the hat matrix $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. If the scale parameter model is log-linear and assuming h_{ij} is small for all $i, j,\ i \neq j$, then

$$\log\big(d_i/(1-h_{ii})^2\big) - \log\nu + \log(\nu-2) = \log\sigma_i^2$$

Following Harvey (1976), Smyth (1989) and Verbyla (1993) this requires an approximate stabilizing transformation of the squared residuals. For $r_i = \log(d_i/\phi_i^2)$ the moment generating function is

$$\begin{aligned} M_{r_i}(t) &= \mathrm{E}\left[\exp\{\log(d_i/\phi_i^2)t\}\right] \\ &= \mathrm{E}\left[(d_i/\phi_i^2)^t\right] \end{aligned}$$

Therefore the transformation requires the first two uncentred moments of the $F_{1,\,\nu}$ distribution. The moment generating function is

$$M_{r_i}(t) = \exp(t \log \nu)\frac{\Gamma(1/2+t)\Gamma(\nu_2/2-t)}{\Gamma(1/2)\Gamma(\nu_2/2)}$$

Taking derivatives with respect to t and evaluating them at $t = 0$ provides the first and second moments of the distribution, namely

$$\begin{aligned} \mathrm{E}(r_i) &= \log\tfrac{\nu}{2} - \log\tfrac{1}{2} + \psi(\tfrac{1}{2}) - \psi(\tfrac{\nu}{2}) \\ \mathrm{E}(r_i^2) &= \big(\log\tfrac{\nu}{2} + \psi(\tfrac{1}{2}) - \psi(\tfrac{\nu}{2})\big)^2 + \dot{\psi}(\tfrac{1}{2}) + \dot{\psi}(\tfrac{\nu}{2}) \end{aligned}$$

where $\psi(\cdot)$ is the digamma function and $\dot{\psi}(\cdot)$ is the trigamma function. This allows the approximate stabilizing transformation to be expressed as

$$\begin{aligned} \mathrm{E}\big\{\log\big(d_i/(1-h_{ii})^2\big) + \log\tfrac{1}{2} - \psi(\tfrac{1}{2}) + s(\nu)\big\} &= \boldsymbol{z}_i^T\boldsymbol{\lambda} \\ \mathrm{var}\big\{\log\big(d_i/(1-h_{ii})^2\big) + \log\tfrac{1}{2} - \psi(\tfrac{1}{2}) - \log\tfrac{\nu}{2} + \psi(\tfrac{\nu}{2})\big\} &= \dot{\psi}(\tfrac{1}{2}) + \dot{\psi}(\tfrac{\nu}{2}) \end{aligned} \tag{4.6.2}$$

where $s(\nu) = \psi(\frac{\nu}{2}) - \log\frac{\nu}{2} - \log\,\nu + \log(\nu-2)$. Therefore the component inside the expectation has constant scale and a mean that is linear in the covariates of the scale parameter model. The terms $s(\nu)$ and $\dot{\psi}(\frac{\nu}{2})$ therefore provide an adjustment to the equations (5) in Verbyla (1993). Asymptotic theory also suggests that as $\nu \to \infty$, $\psi(\frac{v}{2}) \to \log\frac{\nu}{2}$ and $\dot{\psi}(\frac{v}{2}) \to 0$, ensuring (4.6.2) will be identical to the approximate transformation presented in Verbyla (1993).

As in Verbyla (1993), the adjusted residuals given in (4.6.2) may be used as an exploratory tool for graphical investigation of an initial scale model. Examples of its use in this context can be found in Chapters 7 and 9. In addition, similar to Smyth (1989), (4.6.2) can be used to obtain starting values for the computational algorithm given in Section 4.7. Furthermore, for known degrees of freedom, the adjustments given in (4.6.2) are constant suggesting that the simpler Gaussian equations from Verbyla (1993) can also be utilised when the response is t-distributed. This useful when $\nu \leq 2$, as in this case the variance for the t-distribution and the the adjustment derived here are not defined.

4.7 Computation and Software

The process for computing the random effects, $\boldsymbol{\omega}$ and estimating the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\lambda})$ can be simplified by using a modified scoring algorithm. For the $(m+1)$th iterate of the algorithm the BUP for the random effects can be expressed as

$$\omega_i^{(m+1)} = \mathrm{E}(w_i|y_i, \boldsymbol{\theta}^{(m)}) = \frac{\nu^{(m)}+1}{\nu^{(m)} + d_i^{(m)}/\sigma_i^{2(m)}} \tag{4.7.1}$$

The location and scale parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, are updated using the scoring equations (4.3.16) and (4.3.18) respectively. With a two parameter location and scale model, Smyth (1989) discusses the computational requirements for different iterative processes for non-normal models, concluding that only *alternating* iterations between the location and scale estimating equations are required for maximum efficiency. Thus, $\boldsymbol{\theta}$ can be updated using

- **Score Step 1:** For given $\omega_i^{(m+1)}$, $1 = 1, \ldots, n$, $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(m)}$ and ν update $\boldsymbol{\beta}$ using $\boldsymbol{\beta}^{(m+1)} = f_t(\boldsymbol{\beta}^{(m)}, \boldsymbol{\lambda}^{(m)}, \nu)$, where $f_t(\cdot)$ is given in (4.3.16).
- **Score Step 2:** For given $\omega_i^{(m+1)}$, $1 = 1, \ldots, n$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m+1)}$ and ν update $\boldsymbol{\lambda}$ using $\boldsymbol{\lambda}^{(m+1)} = g_t(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\lambda}^{(m)}, \nu)$, where $g_t(\cdot)$ is given in (4.3.18).

Further enhancements to the efficiency of the algorithm can be achieved by using the information from score steps 1 and 2 as soon as it is available. For example, the random effects are then updated in an intermediate step using the prediction (4.7.1) with $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ fixed at $(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\lambda}^{(m)})$.

The prediction of the random effects given in (4.7.1) is identical to evaluating the E-step of the Expectation-Maximisation (EM) algorithm for the t-distribution with known or unknown degrees of freedom (see Rubin, 1983; Lange et al., 1989; Meng & van Dyk, 1997; Liu et al., 1998 and Pinheiro et al., 2001). However the non-linearity of the location and scale parameters in the estimating equations suggests that Fisher scoring steps are computationally preferable for each iteration.

A software library "hett" implementing this particular scoring algorithm has been developed by the author for general use in the R package (see Ihaka & Gentleman, 1996). It contains functions for the estimation and summary of the parameters as well as simple score tests for hypothesis testing. Documentation is also available for the important functions in this package (see Appendix B). The library is available online at

`http://www.biometricssa.adelaide.edu.au/staff/staff_jtaylor.shtml/hett`

or it can be downloaded directly from CRAN (Comprehensive R Archive Network) and installed using `install.packages("hett")`.

Chapter 5

Approximate Likelihood Techniques

The derivation of REML for the Gaussian linear mixed model presented in Section 2.3 requires the conditional distribution of $\boldsymbol{y}_1|\boldsymbol{y}_2$ to be available to provide neat factorisation of the marginal likelihood. The REML component of the factorisation, $L(\cdot;\boldsymbol{y}_2)$, can then be used for inference about the remaining scale parameters in the model. For the heteroscedastic t-distribution this conditional distribution cannot be derived explicitly. Therefore obtaining a REML equivalent is non-trivial and approximate techniques are required. In this thesis, two approximate approaches have been researched.

The hierarchical nature of the heteroscedastic t allows the marginal likelihood to be expressed as an integral given by (4.2.4) where, over n dimensional space, the random scale effects require integrating out. This suggests that approximate integration techniques may be used to obtain an approximate marginal likelihood. The first approximate technique discussed in Section 5.1 derives an approximation to an integral regardless of its intractability. This is called the Laplace approximation. To illustrate its use two theoretical examples are presented. For the linear mixed model presented in Chapter 2, the marginal likelihood is derived by integrating the location random effects out using the Laplace approximation. This is found to be analogous to the methodology presented in Section 2.1.1. A second extensive example using the hierarchical structure of the heteroscedastic t-distribution is also discussed. Here, the random scale effects are considered to be nuisance parameters and integrated out to obtain an approximate marginal likelihood. This marginal likelihood can then be used for inference on the location and scale parameters. Notice that this constitutes an approximate likelihood of the maximum likelihood form. It is not REML.

To obtain an approximate REML for the heteroscedastic t an extension of the Laplace approximation is required. This extension is called the Partial Laplace approximation and is presented in Section 5.1.3. This approximation requires that the integrand consist of at least two disjoint functions. The word "partial" is then used to explain the requirement to

maximise only partially over the integrand to obtain the appropriate approximation. Its use is illustrated with the linear mixed model and is shown to lead to REML as presented in Section 2.3. More formally, in Chapter 6, the new integral approximation is used to obtain an approximate REML for the heteroscedastic t.

The second comparative approximate approach presented in 5.2 uses adjustments to the exact marginal likelihood in the presence of nuisance parameters. In particular, the approximations are based on modifications to the profile likelihood and provide a flexible framework for approximating the marginal likelihood for the parameter of interest. The adjusted likelihood techniques derived in this chapter are motivated by the forms of the heteroscedastic t. When the degrees of freedom is known, the heteroscedastic t is a member of the location-scale family and ancillary and sufficient statistics are available. This motivates the theoretical derivation of Modified Profile Likelihood (MPL) for a parameter of interest in Section 5.2.1. If the degrees of freedom is not known the attractive properties of sufficiency and ancillarity are lost. Stably adjusted Profile Likelihood (SAPL) circumvents this as it does not require these statistics to be available. Its derivation is outside the scope of this thesis but an overview is provided in Section 5.2.4. In following chapters both of these techniques are applied to the heteroscedastic t to obtain different approximate restricted likelihoods for the scale parameters.

5.1 The Laplace Approximation

Let $(\theta_1, \ldots, \theta_p, \phi_1, \ldots, \phi_r) = (\boldsymbol{\theta}, \boldsymbol{\phi})$ be a vector of variables of length p and r respectively contained in the space $\boldsymbol{\Theta}$. If the variable of interest is $\boldsymbol{\theta}$, then $\boldsymbol{\phi}$ are nuisance variables that may be integrated out. The integral for consideration is of the form

$$M(\boldsymbol{\theta}) = \int_{\mathcal{R}^r} \exp\big(h(\boldsymbol{\theta}, \boldsymbol{\phi})\big) d\boldsymbol{\phi}, \tag{5.1.1}$$

where $h(\cdot)$ is an arbitrary function and $\mathcal{R}^p$ describes a multidimensional subspace in $\boldsymbol{\Theta}$ over which the function must be integrated. This is a similar integral considered by Erdelyi (1956) and De Bruijn (1961) who propose the Laplace approximation.

5.1.1 Uniparameter

Let ϕ and θ be scalar variables. Let $h^k(\theta, \phi)$ be the kth partial derivative of $h(\theta, \phi)$ with respect to ϕ. Expanding the functional component of the exponent inside the integrand

of (5.1.1) in a Taylor series around some maximum $\hat{\phi}$ gives

$$\begin{aligned} M(\theta) &= \int \exp(h(\theta,\phi))d\phi \approx \int \exp\Big(h(\theta,\hat{\phi}) + (\phi-\hat{\phi})h^1(\theta,\phi)\Big|_{\phi=\hat{\phi}} \\ &+ \frac{1}{2}(\phi-\hat{\phi})^2 h^2(\theta,\phi)\Big|_{\phi=\hat{\phi}} + \frac{1}{6}(\phi-\hat{\phi})^3 h^3(\theta,\phi)\Big|_{\phi=\hat{\phi}} + \cdots\Big) \end{aligned}$$

Let $\hat{\phi}$ be the maximum of the function $h(\cdot)$ so that $h^1(\theta,\hat{\phi}) = 0$. This reduces to

$$M(\theta) \approx \exp(h(\hat{\phi},\theta)) \int \exp\Big(-\frac{1}{2}(\phi-\hat{\phi})^2(-h^2(\theta,\hat{\phi}))\Big)\exp(G(\theta,\phi))d\phi$$

where

$$G(\theta,\phi) = \sum_{k=3}^{\infty} \frac{1}{k!}(\phi-\hat{\phi})^k h^k(\theta,\hat{\phi})$$

The first exponential term of the integrand mimics the quadratic form for a Gaussian distribution with expected value 0 and variance $(-h^2(\theta,\hat{\phi}))^{-1}$. Replacing the missing components of this distribution in the integrand gives

$$\begin{aligned} M(\theta) &\approx (2\pi)^{1/2}(-h^2(\theta,\hat{\phi}))^{-1/2}\exp(h(\theta,\hat{\phi})) \\ &\times \int (2\pi)^{-1/2}(-h^2(\hat{\phi},\theta))^{1/2}\exp\Big(-\frac{1}{2}(\phi-\hat{\phi})^2(-h^2(\theta,\hat{\phi}))\Big)\exp(G(\phi,\theta))d\phi \quad (5.1.2) \end{aligned}$$

Disregarding the higher order terms in the last exponent term of the integrand the first order Laplace approximation is

$$M(\theta) \approx (2\pi)^{1/2}(-h^2(\hat{\phi},\theta))^{-1/2}\exp(h(\theta,\hat{\phi})) \quad (5.1.3)$$

5.1.2 Multiparameter

Let the integral of interest be described by (5.1.1) where $\boldsymbol{\phi}$ are considered to be nuisance variables and $\boldsymbol{\theta}$ the variable of interest. Let $\boldsymbol{h}(\boldsymbol{\theta},\boldsymbol{\phi})$ be the vector of partial derivatives of $h(\cdot)$ with ith element $\partial h(\boldsymbol{\theta},\boldsymbol{\phi})/\partial\phi_i$ and $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta},\boldsymbol{\phi})$ be the matrix of partial second derivatives with ijth element $\partial^2 h(\boldsymbol{\theta},\boldsymbol{\phi})/\partial\phi_i\partial\phi_j$ then expanding the functional component of the exponent in the integrand using a Taylor series to the second order around some maximum $\hat{\boldsymbol{\phi}}$ gives

$$\begin{aligned} M(\boldsymbol{\theta}) &\approx \exp(h(\boldsymbol{\theta},\hat{\boldsymbol{\phi}})) \int_{\mathcal{R}^r} \exp\Big((\boldsymbol{\phi}-\hat{\boldsymbol{\phi}})^T\boldsymbol{h}(\boldsymbol{\theta},\hat{\boldsymbol{\phi}}) \\ &\qquad -\frac{1}{2}(\boldsymbol{\phi}-\hat{\boldsymbol{\phi}})^T(-\boldsymbol{\mathcal{H}}(\boldsymbol{\theta},\hat{\boldsymbol{\phi}}))(\boldsymbol{\phi}-\hat{\boldsymbol{\phi}})\Big)d\boldsymbol{\phi} \end{aligned}$$

If $\hat{\boldsymbol{\phi}}$ is the maximiser of $h(\boldsymbol{\phi},\theta)$ then, equivalent to the single parameter case, $\boldsymbol{h}(\hat{\boldsymbol{\phi}},\boldsymbol{\theta}) = \mathbf{0}$ and the integrand is reduced to

$$M(\boldsymbol{\theta}) \approx \exp(h(\boldsymbol{\theta},\hat{\boldsymbol{\phi}})) \int_{\mathcal{R}^p} \exp\left(-\frac{1}{2}(\boldsymbol{\phi}-\hat{\boldsymbol{\phi}})^T(-\boldsymbol{\mathcal{H}}(\boldsymbol{\theta},\hat{\boldsymbol{\phi}}))(\boldsymbol{\phi}-\hat{\boldsymbol{\phi}})\right)d\boldsymbol{\phi}$$

The remaining exponent in the integrand is the multivariate Gaussian distribution with with expected value $\mathbf{0}$ and variance $(-\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}))^{-1}$. Following from the previous section this can be immediately reduced to

$$M(\boldsymbol{\theta}) \approx (2\pi)^{p/2} \exp(h(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}))| - \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}})|^{-1/2}, \tag{5.1.4}$$

where $|-\boldsymbol{\mathcal{H}}(\cdot)|$ is the determinant of the negative partial second derivatives of the original function, $h(\boldsymbol{\theta}, \boldsymbol{\phi})$ with respect to $\boldsymbol{\phi}$ evaluated at $\hat{\boldsymbol{\phi}}$. An identical result is derived by Leonard (1982), Tierney & Kadane (1986) and Tierney et al. (1989).

Identical to the previous section a more accurate approximation can be found by considering higher terms of the Taylor series expansion. For brevity this has been omitted from this thesis. Details of this approximation can be found in Barndorff-Nielsen & Cox (1989) and for a more detailed derivation see Raudenbush et al. (2000).

Example: Laplace and Linear Mixed Models

The marginal likelihood derived in Section 2.1.1 can also be derived using the Laplace approximation from Section 5.1.2. To perform this approximation the random effects, $\boldsymbol{u}$, are treated as nuisance variables that require integrating out.

Let

$$h(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{u}) = \log p(\boldsymbol{y}|\boldsymbol{u}; \boldsymbol{\beta}, \boldsymbol{\theta}) + \log p(\boldsymbol{u}; \boldsymbol{\varphi}) \tag{5.1.5}$$

where

$$\begin{aligned} \log p(\boldsymbol{y}|\boldsymbol{u}; \boldsymbol{\beta}, \boldsymbol{\theta}) &= -\tfrac{1}{2}\big\{n\log(2\pi) + \log|\boldsymbol{R}| + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})\big\} \\ \log p(\boldsymbol{u}; \boldsymbol{\varphi}) &= -\tfrac{1}{2}\big\{r\log(2\pi) + \log|\boldsymbol{G}| + \boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}\big\} \end{aligned}$$

Then the marginal likelihood for $\boldsymbol{y}$ can be expressed as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}) = \int_{\mathcal{R}^r} \exp\big(h(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{u})\big) d\boldsymbol{u}$$

Let $\boldsymbol{h}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{u})$ be a r length vector of partial derivatives with ith element $\partial h(\cdot)/\partial u_i$ and $\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{u})$ be a $r \times r$ matrix of partial second derivatives with ijth element $\partial^2 h(\cdot)/\partial u_i \partial u_j$. Let $\tilde{\boldsymbol{u}}$ be a maximum of $\boldsymbol{u}$ obtained by solving $\boldsymbol{h}(\cdot) = \mathbf{0}$. Using (5.1.4) the Laplace approximation to the marginal likelihood can be immediately written as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}) \approx (2\pi)^{r/2}| - \boldsymbol{\mathcal{H}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \tilde{\boldsymbol{u}})| \exp(h(\boldsymbol{\beta}, \boldsymbol{\theta}, \tilde{\boldsymbol{u}})) \tag{5.1.6}$$

To obtain the terms of the approximate marginal likelihood the first two derivatives of (5.1.5) are required, namely

$$\begin{aligned} \boldsymbol{h}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{u}) = \frac{\partial h(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{u})}{\partial \boldsymbol{u}} &= \boldsymbol{Z}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) - \boldsymbol{G}^{-1}\boldsymbol{u} \\ \boldsymbol{\mathcal{H}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{u}) = \frac{\partial^2 h(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{u})}{\partial \boldsymbol{u} \partial \boldsymbol{u}^T} &= -\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} - \boldsymbol{G}^{-1} \end{aligned}$$

Setting the first derivative to zero and solving gives the maximised random effects as $\tilde{\boldsymbol{u}}$ as

$$\begin{aligned}\tilde{\boldsymbol{u}} &= (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{R}+\boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \qquad \text{(using Result (A.3.2))}\end{aligned}$$

Using (5.1.4), the Laplace approximation to the marginal likelihood can be immediately expressed as

$$L(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y}) = (2\pi)^{-n/2}|\boldsymbol{G}|^{-1/2}|\boldsymbol{R}|^{-1/2}|\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1}|^{-1/2}\exp\left(h^*(\boldsymbol{\beta},\boldsymbol{\theta},\tilde{\boldsymbol{u}})\right)$$

where, upon substitution of $\tilde{\boldsymbol{u}}$

$$\begin{aligned}h^*(\boldsymbol{\beta},\boldsymbol{\theta},\tilde{\boldsymbol{u}}) &= -\tfrac{1}{2}\big[(\boldsymbol{I}-\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{H}^{-1})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\big]^T\boldsymbol{R}^{-1}\big[(\boldsymbol{I}-\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{H}^{-1})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\big] \\ &- \tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{G}^{-1}\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\ &= -\tfrac{1}{2}\big[\boldsymbol{R}\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\big]^T\boldsymbol{R}^{-1}\big[\boldsymbol{R}\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\big] \\ &- \tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{H}^{-1}(\boldsymbol{H}-\boldsymbol{R})\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\ &= -\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\end{aligned}$$

Following Section (2.1.1) the determinants can be amalgamated by using Result A.2.2. The approximate marginal likelihood can then be expressed as

$$L(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y}) = (2\pi)^{-n/2}|\boldsymbol{H}|^{-1/2}\exp\big(-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\big)$$

This is identical to the marginal likelihood (2.1.5) derived in Section 2.1.1. This equivalence is due to the quadratic nature of the exponential component of the pseudo-joint likelihood required for expansion.

Heteroscedastic t-ML

Consider the model defined by (3.1.1) where conditional on the random scale effects, $\omega_i,\ i=1,\ldots,n$ the response is distributed by (4.2.1) and the random scale effects have the distribution defined by (4.2.2).

For this section a conditional scale parameter model defined by (4.2.3) is assumed. The heteroscedastic t-distribution has a marginal likelihood that is defined by (4.2.4). Section 5.1.2 suggests that this marginal likelihood can be approximated using the Laplace approximation. In this particular case the random scale effects, $\omega_i,\ 1,\ldots,n$ are considered to be nuisance variables that require integrating out of the integrand of (4.2.4). Here the unknown parameters, $(\boldsymbol{\beta},\boldsymbol{\lambda},\nu)$, are the parameters of interest.

Let

$$h(\boldsymbol{\beta},\boldsymbol{\lambda},\nu,\boldsymbol{\omega}) = \log p(\boldsymbol{y}|\boldsymbol{\omega};\boldsymbol{\beta},\boldsymbol{\lambda},\nu) + \log p(\boldsymbol{\omega};\nu) = \log\left\{\prod_{i=1}^{n} p(y_i|\omega_i;\boldsymbol{\beta},\boldsymbol{\lambda})p(\omega_i;\nu)\right\} \quad (5.1.7)$$

where

$$\log p(\boldsymbol{y}|\boldsymbol{\omega};\boldsymbol{\beta},\boldsymbol{\lambda},\nu) = -\tfrac{1}{2}\left\{n\log(2\pi) + \log|\boldsymbol{\Psi}| + (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\} \tag{5.1.8}$$

$$\log p(\boldsymbol{\omega};\nu) = \frac{n\nu}{2}\log(\nu/2) - n\log(\Gamma(\nu/2)) + \sum_{i=1}^{n}\left\{\left(\frac{\nu}{2}-1\right)\log\omega_i - \frac{\nu}{2}\omega_i\right\} \tag{5.1.9}$$

The marginal likelihood for the heteroscedastic t is

$$L(\boldsymbol{\beta},\boldsymbol{\lambda},\nu;\boldsymbol{y}) \;=\; \int_{\mathcal{R}^n}\exp(h(\boldsymbol{\beta},\boldsymbol{\lambda},\nu,\boldsymbol{\omega}))d\boldsymbol{\omega} \tag{5.1.10}$$

Let $\boldsymbol{h}(\boldsymbol{\beta},\boldsymbol{\lambda},\nu,\boldsymbol{\omega})$ be a n length vector of partial derivatives with ith element $\partial h(\cdot)/\partial\omega_i$ and let $\boldsymbol{\mathcal{H}}(\boldsymbol{\beta},\boldsymbol{\lambda},\nu,\boldsymbol{\omega})$ be an $n\times n$ matrix of partial second derivatives with ijth element $\partial^2 h(\cdot)/\partial\omega_i\partial\omega_j$. To integrate out the random scale effects, ω_i , $1,\ldots,n$ the function in the exponent of the integrand is expanded in a Taylor series about $\tilde{\boldsymbol{\omega}}$, a maximum value of $\boldsymbol{\omega}$ found by solving $\boldsymbol{h}(\cdot)=\boldsymbol{0}$, and the Laplace approximation is applied. Using (5.1.4) the Laplace approximation to the marginal likelihood is

$$L(\boldsymbol{\beta},\boldsymbol{\lambda},\nu;\boldsymbol{y}) \approx (2\pi)^{n/2}|-\boldsymbol{\mathcal{H}}(\boldsymbol{\beta},\boldsymbol{\lambda},\nu,\tilde{\boldsymbol{\omega}})|\exp(h(\boldsymbol{\beta},\boldsymbol{\lambda},\nu,\tilde{\boldsymbol{\omega}})) \tag{5.1.11}$$

The derivatives required for the approximation can be derived as follows. Taking the first derivative of $h(\cdot)$ with respect to ω_i gives

$$\frac{\partial h(\cdot)}{\partial\omega_i} = -\frac{1}{2}\left\{\mathrm{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i) - (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\} + \left(\frac{\nu}{2}-1\right)\frac{1}{\omega_i} - \frac{\nu}{2} \tag{5.1.12}$$

where $\dot{\boldsymbol{\Psi}}_i = \partial\boldsymbol{\Psi}/\partial\omega_i$ Taking the derivative of (5.1.12) with respect to ω_i gives

$$\begin{aligned}\frac{\partial^2 h(\cdot)}{(\partial\omega_i)^2} &= -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ii}) - (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\\ &+ \frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ii}\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) - \left(\frac{\nu}{2}-1\right)\frac{1}{\omega_i^2}\end{aligned} \tag{5.1.13}$$

where $\dot{\boldsymbol{\Psi}}_{ij} = \partial^2\boldsymbol{\Psi}/(\partial\omega_i)^2$. Taking the derivative of (5.1.12) with respect to ω_j gives

$$\begin{aligned}\frac{\partial^2 h(\cdot)}{\partial\omega_i\partial\omega_j} &= -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ij}) - (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_j\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\\ &+ \frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_j) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ij}\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\end{aligned}$$

where $\dot{\boldsymbol{\Psi}}_{ij} = \partial^2\boldsymbol{\Psi}/\partial\omega_i\partial\omega_j$. As $\boldsymbol{\Psi}$ is a diagonal matrix the first and second derivatives of its ith element with respect to ω_i can be expressed as

$$\frac{\partial\varphi_i}{\partial\omega_i} = -\frac{\varphi_i}{\omega_i} = -\frac{\sigma_i^2}{\omega_i^2}, \qquad \frac{\partial^2\varphi_i}{(\partial\omega_i)^2} = \frac{2\varphi_i}{\omega_i^2} = \frac{2\sigma_i^2}{\omega_i^3} \tag{5.1.14}$$

The diagonality of the scale matrix, $\boldsymbol{\Psi}$, also ensures that

$$\frac{\partial^2 \varphi_i}{\partial \omega_i \partial \omega_j} = 0, \qquad i \neq j, i = 1, \ldots, n;\ j = 1, \ldots, n \tag{5.1.15}$$

and therefore $\dot{\boldsymbol{\Psi}}_{ij} = \mathbf{0}$. This ensures, for this particular case, that the second derivative $\partial^2 h(\cdot)/\partial\omega_i\partial\omega_j = 0$ and $\boldsymbol{\mathcal{H}}(\cdot)$ is a $n \times n$ diagonal matrix with ith diagonal element $\partial h(\cdot)/(\partial\omega_i)^2$ and off diagonals equal to zero. Considering the trace terms of (5.1.13) and using (5.1.14)

$$\begin{aligned} \operatorname{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ii}) &= \frac{2}{\omega_i^2} \\ \operatorname{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{i}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{i}) &= \frac{1}{\omega_i^2} \end{aligned}$$

The quadratic terms of (5.1.13) can also be reduced to

$$\begin{aligned} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ii}\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &= \frac{2d_i}{\sigma_i^2\omega_i} \\ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{i}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{i}\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &= \frac{d_i}{\sigma_i^2\omega_i} \end{aligned} \tag{5.1.16}$$

Combining these with the last term of (5.1.13) the second derivative can be expressed as

$$\begin{aligned} \frac{\partial^2 h(\cdot)}{(\partial\omega_i)^2} &= -\frac{1}{2\omega_i^2} - \Big(\frac{\nu}{2} - 1\Big)\frac{1}{\omega_i^2} \\ &= -\frac{\nu - 1}{2\omega_i^2} \end{aligned} \tag{5.1.17}$$

Therefore the Laplace approximation given by (5.1.11) is

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y}) &= |\tilde{\boldsymbol{\Omega}}||\tilde{\boldsymbol{\Psi}}|^{-1/2} \exp\big\{-\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \tilde{\boldsymbol{\Psi}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big\} \\ &\times \ ((\nu-1)/2)^{-n/2} \frac{(\nu/2)^{n\nu/2}}{(\Gamma(\nu/2))^n} \exp\Big\{-\tfrac{\nu}{2}\sum_{i=1}^{n}\tilde{\omega}_i\Big\} \prod_{i=1}^{n} \tilde{\omega}_i^{\nu/2-1} \end{aligned} \tag{5.1.18}$$

where $\boldsymbol{\Omega}$ is a diagonal matrix with ith diagonal element ω_i. Terms that have a $\tilde{}$ represent the respective term evaluated at $\tilde{\boldsymbol{\omega}}$.

The approximate marginal likelihood given by (5.1.18) can be reduced further by considering $\tilde{\omega}_i,\ i = 1, \ldots, n$, the maximum of the random scale effects required for the approximation. The ith random scale effect, $\tilde{\omega}_i$, is evaluated by considering the first derivative of $h(\cdot)$ with respect to ω_i given by (5.1.12). Using (5.1.14) this can be reduced to

$$\frac{\partial h(\cdot)}{\partial \omega_i} = \frac{1}{2\omega_i} - \frac{1}{2}\frac{d_i}{\sigma_i^2} + \Big(\frac{\nu}{2} - 1\Big)\frac{1}{\omega_i} - \frac{\nu}{2} \tag{5.1.19}$$

Setting this to zero and solving for ω_i gives

$$\tilde{\omega}_i = \frac{\nu - 1}{\nu + d_i/\sigma_i^2} \tag{5.1.20}$$

Note this is not equivalent to the predicted random scale effects (4.4.3) derived from the conditional distribution of $\omega_i|y_i$.

Taking natural logs of (5.1.18) the approximate marginal log-likelihood is

$$\begin{aligned}
\ell(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y}) &= \frac{n\nu}{2}\log(\nu/2) - n\log\Gamma(\nu/2) - \frac{n}{2}\log((\nu-1)/2) \\
&- \frac{1}{2}\sum_{i=1}^{n}\Big\{\log\sigma_i^2 + \frac{d_i\tilde{\omega}_i}{\sigma_i^2} - (\nu+1)\log\tilde{\omega}_i + \nu\tilde{\omega}_i\Big\}
\end{aligned}$$

Substituting (5.1.20) and using the relation, $\nu - 1 = \nu\tilde{\omega}_i + d_i\tilde{\omega}_i/\sigma_i^2$, from the same equation the approximate marginal log-likelihood becomes

$$\begin{aligned}
\ell(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y}) &= \frac{n\nu}{2}\log(\nu/2) - n\log\Gamma(\nu/2) - \frac{n}{2}\log((\nu-1)/2) - \frac{n}{2}(\nu-1) \\
&- \frac{1}{2}\sum_{i=1}^{n}\Big\{\log\sigma_i^2 + (\nu+1)\log\Big(\frac{\nu + d_i/\sigma_i^2}{\nu-1}\Big)\Big\} \\
&= \frac{n\nu}{2}\log((\nu-1)/2) - n\log\Gamma(\nu/2) - \frac{n}{2}\log(\nu/2) - \frac{n}{2}(\nu-1) \\
&- \frac{1}{2}\sum_{i=1}^{n}\Big\{\log\sigma_i^2 + (\nu+1)\log\Big(1 + \frac{d_i}{\sigma_i^2\nu}\Big)\Big\}
\end{aligned}$$

Therefore when the degrees of freedom is known the Laplace approximation reproduces the kernel of the heteroscedastic t-distribution for the location and scale parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda})$.

The estimated random scale effects, $\tilde{\omega}_i,\ i = 1, \ldots, n$, required to form the Laplace approximation are obtained by maximising the integrand of (5.1.20). This integrand represents a pseudo joint likelihood for $(\boldsymbol{y}, \boldsymbol{\omega})$ and therefore this maximisation is equivalent to a Hierarchical Generalised Linear Modelling (HGLM) (see Lee & Nelder, 1996 and Lee & Nelder, 2001) approach to random effect estimation. A potential problem with this method is the non-invariance of the estimated random effects from HGLM when the scale of the random effects distribution is transformed. Although the heteroscedastic t-distribution is not a member of the exponential family it is used to highlight this problem.

Under a heteroscedastic Gaussian distribution the kernel for the scale parameters is represented by a Gamma generalized linear model (see Section 3.2 or Smyth, 1989 and Verbyla, 1993) with log link for the location parameters, σ_i^2. Similarly, the heteroscedastic t-distribution contains the conditional Gaussian (5.1.8) as a component of the integrand. Under the log link the random scale effects contained in this component are naturally logged and occur linearly in the predictor. Lee & Nelder (1996) and Lee & Nelder (2001)

allude that the use of the correct scale, the scale on which the random effects occur linearly, ensures the estimated random effects obtained from HGLM will be equivalent to the predicted random effects obtained from the conditional distribution of $\omega_i|y_i$. For this reason a transformation of the form $\omega_i^* = \log \omega_i$ is used.

The new well known log Gamma distribution has a log-likelihood of the form

$$\ell(\boldsymbol{\omega}^*; \nu) = \frac{n\nu}{2}\log(\nu/2) - n\log\Gamma(\nu/2) + \sum_{i=1}^{n}\Big\{\frac{\nu}{2}\omega_i^* - \frac{\nu}{2}\exp\omega_i^*\Big\} \tag{5.1.21}$$

Substituting this into the integrand of (5.1.10) the integral is approximated again by the Laplace approximation given in (5.1.11). Following identically to the previous section the approximation requires the derivatives of $q(\cdot)$. Noting that

$$\frac{\partial\varphi_i}{\partial\omega_i^*} = -\frac{\sigma_i^2}{\exp\omega_i^*}, \qquad \frac{\partial^2\varphi_i}{(\partial\omega_i^*)^2} = \frac{\sigma_i^2}{\exp\omega_i^*} \tag{5.1.22}$$

The first and second derivative with respect to ω_i^* can be immediately written as

$$\begin{aligned}\frac{\partial h(\cdot)}{\partial\omega_i^*} &= \frac{1}{2} - \frac{1}{2}\frac{d_i\exp\omega_i^*}{\sigma_i^2} + \frac{\nu}{2} - \frac{\nu}{2}\exp\omega_i^* \\ \frac{\partial^2 h(\cdot)}{(\partial\omega_i^*)^2} &= -\frac{1}{2}\frac{d_i\exp\omega_i^*}{\sigma_i^2} - \frac{\nu}{2}\exp\omega_i^*\end{aligned} \tag{5.1.23}$$

To obtain the maximised random scale effects the first derivative is set to zero and solved for $\omega_i = \exp\omega_i^*$ giving

$$\tilde{\omega}_i = \frac{\nu+1}{\nu + d_i/\sigma_i^2}$$

Note this maximisation is equivalent to the predicted random scale effects (4.4.3) obtained from the conditional distribution $\omega_i|y_i$ and differs from the maximised random scale effects (5.1.20) obtained from using the ordinary Gamma kernel (5.1.9) in the integrand of the Laplace approximation. This suggests that changing the scale of the random effects distribution alters the invariance of the random effects under HGLM.

Substitution of these new maximised effects allows the second derivative to be reduced to

$$\frac{\partial^2 h(\cdot)}{(\partial\omega_i^*)^2} = -\frac{\nu+1}{2}$$

The final approximate marginal log-likelihood is

$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y}) &= \frac{n\nu}{2}\log(\nu/2) - n\log\Gamma(\nu/2) - \frac{n}{2}\log((\nu+1)/2) - \frac{n}{2}(\nu+1) \\ &- \frac{1}{2}\sum_{i=1}^{n}\Big\{\log\sigma_i^2 + (\nu+1)\log\Big(\frac{\nu + d_i/\sigma_i^2}{\nu+1}\Big)\Big\} \\ &= \frac{n\nu}{2}\log((\nu+1)/2) - n\log\Gamma(\nu/2) - \frac{n}{2}\log(\nu/2) - \frac{n}{2}(\nu+1) \\ &- \frac{1}{2}\sum_{i=1}^{n}\Big\{\log\sigma_i^2 + (\nu+1)\log\Big(1 + \frac{d_i}{\sigma_i^2\nu}\Big)\Big\}\end{aligned}$$

Again, this is equivalent to the kernel of the heteroscedastic t-distribution for the location and scale parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda})$. In this particular case, the approximate marginal likelihood obtained by the Laplace approximation is parameter invariant under transformation of the random scale effects distribution in the integrand.

5.1.3 Partial Laplace Approximation

Consider $(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi})$ from Section 5.1.2 to be a p, q and r length vectors of variables respectively where, again, $\boldsymbol{\phi}$ is the nuisance variable and $(\boldsymbol{\theta}, \boldsymbol{\psi})$ are the variables of interest. Let $h_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi})$ and $h_2(\boldsymbol{\psi}, \boldsymbol{\phi})$ be a arbitrary functions of the three variables and two variables respectively. The proposed integral is then

$$P(\boldsymbol{\theta}, \boldsymbol{\psi}) = \int_{\mathcal{R}^r} \exp\big(h_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi}) + h_2(\boldsymbol{\psi}, \boldsymbol{\phi})\big) d\boldsymbol{\phi} \tag{5.1.24}$$

Let $\boldsymbol{h}_2(\boldsymbol{\psi}, \boldsymbol{\phi})$ be an r length vector of partial derivatives with ith element $\partial h_2(\cdot)/\partial \phi_i$ and let $\boldsymbol{\mathcal{H}}_2(\boldsymbol{\psi}, \boldsymbol{\phi})$ be an $r \times r$ matrix of partial second derivatives with ijth element $\partial^2 h_2(\cdot)/\partial \phi_i \partial \phi_j$. Furthermore, let $\boldsymbol{h}_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi})$ be an r length vector of partial derivatives with ith element $\partial h_1(\cdot)/\partial \phi_i$ and let $\boldsymbol{\mathcal{H}}_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi})$ be an $r \times r$ matrix of partial second derivatives with ijth element $\partial^2 h_1(\cdot)/\partial \phi_i \partial \phi_j$. If $\hat{\boldsymbol{\phi}}$ is a maximum of $\boldsymbol{\phi}$ such that $\boldsymbol{h}_2(\boldsymbol{\psi}, \hat{\boldsymbol{\phi}}) = \mathbf{0}$ then expanding the integrand of (5.1.24) to the second order using an ordinary Taylor series around the maximum $\hat{\boldsymbol{\phi}}$ gives

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{\psi}) \;\approx\; & \exp\big(h_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}}) + h_2(\boldsymbol{\psi}, \hat{\boldsymbol{\phi}})\big) \int_{\mathcal{R}^r} \exp\Big\{(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})^T (\boldsymbol{h}_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}})) \\ - \;& \frac{1}{2}(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})^T (-\boldsymbol{\mathcal{H}}_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}}))(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}) - \frac{1}{2}(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})^T (-\boldsymbol{\mathcal{H}}_2(\boldsymbol{\psi}, \hat{\boldsymbol{\phi}}))(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})\Big\} d\boldsymbol{\phi} \end{aligned}$$

The integrand can then be reduced by completing the square of the components

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{\psi}) \;\approx\; & \exp\big(h_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}}) + h_2(\boldsymbol{\psi}, \hat{\boldsymbol{\phi}})\big) \\ \times \;& \exp\Big\{\frac{1}{2}\boldsymbol{h}_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}})^T \big(\boldsymbol{\mathcal{M}}(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}})\big)^{-1} \boldsymbol{h}_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}})\Big\} \\ \times \;& \int_{\mathcal{R}^r} \exp\Big\{-\frac{1}{2}\boldsymbol{a}(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi})^T \boldsymbol{\mathcal{M}}(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}}) \boldsymbol{a}(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi})\Big\} d\boldsymbol{\phi} \end{aligned} \tag{5.1.25}$$

where

$$\begin{aligned} \boldsymbol{a}(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi}) &= (\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}) - \big(\boldsymbol{\mathcal{M}}(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}})\big)^{-1} \boldsymbol{h}_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}}) \\ \boldsymbol{\mathcal{M}}(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}}) &= -\boldsymbol{\mathcal{H}}_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}}) - \boldsymbol{\mathcal{H}}_2(\boldsymbol{\psi}, \hat{\boldsymbol{\phi}}) \end{aligned}$$

The integral given by (5.1.25) is in standard form and therefore the Partial Laplacian approximation to (5.1.24) is

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{\psi}) \;\approx\; & (2\pi)^{r/2} |\boldsymbol{\mathcal{M}}(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}})|^{-1/2} \exp\big(h_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}}) + h_2(\boldsymbol{\psi}, \hat{\boldsymbol{\phi}})\big) \\ \times \;& \exp\Big\{\frac{1}{2}\boldsymbol{h}_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}})^T \big(\boldsymbol{\mathcal{M}}(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}})\big)^{-1} \boldsymbol{h}_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}})\Big\} \end{aligned} \tag{5.1.26}$$

If $\hat{\boldsymbol{\phi}}$ is in the neighbourhood of the maximiser for $h_1(\cdot) + h_2(\cdot)$ then the second exponent term approaches zero and the Laplace approximation reverts to the ordinary first order approximation (see Section 5.1.2). This approximation can also be partitioned into two approximately disjoint functions. The Partial Laplace approximation can then be expressed as

$$P(\boldsymbol{\theta}, \boldsymbol{\psi}) = (2\pi)^{r/2} h_1^*(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}}) h_2^*(\boldsymbol{\psi}, \hat{\boldsymbol{\phi}}) \tag{5.1.27}$$

where

$$\begin{aligned} h_1^*(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}}) &= |\boldsymbol{I} + (-\boldsymbol{\mathcal{H}}_2(\boldsymbol{\psi}, \hat{\boldsymbol{\phi}}))^{-1}(-\boldsymbol{\mathcal{H}}_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}}))|^{-1/2} \exp\big(h_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}})\big) \\ &\times \exp\Big\{\frac{1}{2}\boldsymbol{h}_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi})^T \big(\boldsymbol{\mathcal{M}}(\boldsymbol{\theta}, \boldsymbol{\psi}, \hat{\boldsymbol{\phi}})\big)^{-1} \boldsymbol{h}_1(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi})\Big\} \\ h_2^*(\boldsymbol{\psi}, \hat{\boldsymbol{\phi}}) &= |-\boldsymbol{\mathcal{H}}_2(\boldsymbol{\psi}, \hat{\boldsymbol{\phi}})|^{-1/2} \exp\big(h_2(\boldsymbol{\psi}, \hat{\boldsymbol{\phi}})\big) \end{aligned}$$

An example of this is given in the next section.

Example: REML for Linear Mixed Models

The REML for linear mixed models can also be derived using the Partial Laplace approximation. For this example the notation of Section (2.3) is used. Consider the conditional transformed response vector given the location random effects, namely

$$\begin{bmatrix} \boldsymbol{L}_1^T \boldsymbol{y} \\ \boldsymbol{L}_2^T \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} \Big| \boldsymbol{u} \sim N\left(\begin{bmatrix} \boldsymbol{\beta} + \boldsymbol{L}_1^T \boldsymbol{Z}\boldsymbol{u} \\ \boldsymbol{L}_2^T \boldsymbol{Z}\boldsymbol{u} \end{bmatrix}, \begin{bmatrix} \boldsymbol{L}_1^T \boldsymbol{R} \boldsymbol{L}_1 & \boldsymbol{L}_1^T \boldsymbol{R} \boldsymbol{L}_2 \\ \boldsymbol{L}_2^T \boldsymbol{R} \boldsymbol{L}_1 & \boldsymbol{L}_2^T \boldsymbol{R} \boldsymbol{L}_2 \end{bmatrix} \right)$$

Using Result A.4.2 the conditional distributions can be expressed as

$$\begin{aligned} \boldsymbol{y}_1 | (\boldsymbol{y}_2, \boldsymbol{u}) \sim N \big(& \boldsymbol{\beta} + \boldsymbol{L}_1^T \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{L}_1^T \boldsymbol{R} \boldsymbol{L}_2 (\boldsymbol{L}_2^T \boldsymbol{R} \boldsymbol{L}_2)^{-1} (\boldsymbol{y}_2 - \boldsymbol{L}_2^T \boldsymbol{Z}\boldsymbol{u}) \,, \\ & \boldsymbol{L}_1^T \boldsymbol{R} \boldsymbol{L}_1 - \boldsymbol{L}_1^T \boldsymbol{R} \boldsymbol{L}_2 (\boldsymbol{L}_2^T \boldsymbol{R} \boldsymbol{L}_2)^{-1} \boldsymbol{L}_2^T \boldsymbol{R} \boldsymbol{L}_1 \big) \end{aligned} \tag{5.1.28}$$

$$\boldsymbol{y}_2 | \boldsymbol{u} \sim N(\boldsymbol{0} + \boldsymbol{L}_2^T \boldsymbol{Z}\boldsymbol{u}, \boldsymbol{L}_2^T \boldsymbol{R} \boldsymbol{L}_2) \tag{5.1.29}$$

$$\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G}) \tag{5.1.30}$$

The marginal likelihood has the form

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}) = \int_{\mathcal{R}^r} p(\boldsymbol{y}_1 | \boldsymbol{y}_2, \boldsymbol{u}; \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{y}_2 | \boldsymbol{u}; \boldsymbol{\theta}) p(\boldsymbol{u}; \boldsymbol{\varphi}) d\boldsymbol{u} \tag{5.1.31}$$

The integrand in (5.1.31) is analogous to the separation of components proposed in Section 2.3. Therefore, given an observed vector, $\boldsymbol{y}$ and the random effects $p(\boldsymbol{y}_1 | \boldsymbol{y}_2 | \boldsymbol{u}; \cdot)$ represents a conditional probability density function for $\boldsymbol{\beta}$ that contains no information about the scale parameters. Similarly, $p(\boldsymbol{y}_2 | \boldsymbol{u}; \cdot)$ and $p(\boldsymbol{u}; \cdot)$ are probability density functions for $\boldsymbol{\theta}$ and contain no information about $\boldsymbol{\beta}$.

Using Result A.5.1,

$$\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2^T)^{-1}\boldsymbol{L}_2^T = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}^{-1} = \boldsymbol{S}$$

and this result with (2.3.6) allows the probability density function for (5.1.28) to be expressed as

$$p(\boldsymbol{y}_1|\boldsymbol{y}_2,\boldsymbol{u};\boldsymbol{\beta},\boldsymbol{\theta}) = (2\pi)^{-p/2}|\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}|^{1/2} \times \exp\left\{-\tfrac{1}{2}(\boldsymbol{y}_1-\boldsymbol{\beta}-\boldsymbol{y}_2^*)^T\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{y}_1-\boldsymbol{\beta}-\boldsymbol{y}_2^*)\right\}$$

where $y_2^* = \boldsymbol{L}_1^T(\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{R}\boldsymbol{S}(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{u}))$. Using (2.3.7) the probability density function for (5.1.29) can be expressed as

$$p(\boldsymbol{y}_2|\boldsymbol{u};\boldsymbol{\theta}) = (2\pi)^{-(n-p)/2}|\boldsymbol{R}|^{-1/2}|\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}|^{-1/2}\exp\left\{-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{u})^T\boldsymbol{S}(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{u})\right\}$$

and the probability density function for (5.1.30) can be expressed as

$$p(\boldsymbol{u};\boldsymbol{\varphi}) = (2\pi)^{-r/2}|\boldsymbol{G}|^{-1/2}\exp\left\{-\tfrac{1}{2}\boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}\right\}$$

Let

$$\begin{aligned} h_1(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\theta}) &= \log p(\boldsymbol{y}_1|\boldsymbol{y}_2,\boldsymbol{u};\boldsymbol{\beta},\boldsymbol{\theta}) \\ h_2(\boldsymbol{u},\boldsymbol{\theta}) &= \log p(\boldsymbol{y}_2|\boldsymbol{u};\boldsymbol{\theta}) + \log p(\boldsymbol{u};\boldsymbol{\varphi}) \end{aligned}$$

then the marginal likelihood, (5.1.31) can be expressed as

$$L(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y}) = \int_{\mathcal{R}^r}\exp\left(h_1(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\theta}) + h_2(\boldsymbol{u},\boldsymbol{\theta})\right)d\boldsymbol{u} \tag{5.1.32}$$

As $h_1(\cdot)$ contains no information about the scale parameters, the integration of the random effects requires that the integrand be expanded in a Taylor series around some maximum value of $\boldsymbol{u}$, say $\tilde{\boldsymbol{u}}$ of $h_2(\cdot)$ only. Following this, the Partial Laplace approximation from section 5.1.3 can be used. Let $\boldsymbol{h}_1(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\theta})$ be an r length vector of partial derivatives with ith element $\partial h_1(\cdot)/\partial u_i$ and let $\boldsymbol{\mathcal{H}}_1(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\theta})$ be a $r\times r$ matrix of partial second derivatives with ijth element $\partial^2 h_1(\cdot)/\partial u_i\partial u_j$. Furthermore, let $\boldsymbol{h}_2(\boldsymbol{u},\boldsymbol{\theta})$ be a r length vector of partial derivatives with ith element $\partial h_2(\cdot)/\partial u_i$ and let $\boldsymbol{\mathcal{H}}_2(\boldsymbol{u},\boldsymbol{\theta})$ be a $r\times r$ matrix of partial second derivatives with ijth element $\partial^2 h_2(\cdot)/\partial u_i\partial u_j$. Noting (5.1.27), the marginal likelihood, (5.1.32), can be immediately approximated by the Partial Laplace approximation

$$L(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y}) = (2\pi)^{r/2}L_1(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y})L_2(\boldsymbol{\theta};\boldsymbol{y}) \tag{5.1.33}$$

where

$$\begin{aligned} L_1(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y}) &= |\boldsymbol{I} + (-\boldsymbol{\mathcal{H}}_2(\tilde{\boldsymbol{u}},\boldsymbol{\theta}))^{-1}(-\boldsymbol{\mathcal{H}}_1(\boldsymbol{\beta},\tilde{\boldsymbol{u}},\boldsymbol{\theta}))|^{-1/2}\exp\big(h_1(\boldsymbol{\beta},\tilde{\boldsymbol{u}},\boldsymbol{\theta})\big) \\ &\times \exp\Big\{\tfrac{1}{2}\boldsymbol{h}_1(\boldsymbol{\beta},\tilde{\boldsymbol{u}},\boldsymbol{\theta})^T\big(-\boldsymbol{\mathcal{H}}_1(\boldsymbol{\beta},\tilde{\boldsymbol{u}},\boldsymbol{\theta}) - \boldsymbol{\mathcal{H}}_2(\tilde{\boldsymbol{u}},\boldsymbol{\theta})\big)^{-1}\boldsymbol{h}_1(\boldsymbol{\beta},\tilde{\boldsymbol{u}},\boldsymbol{\theta})\Big\} \\ L_2(\boldsymbol{\theta};\boldsymbol{y}) &= \exp\big(h_2(\tilde{\boldsymbol{u}},\boldsymbol{\theta})\big)|-\boldsymbol{\mathcal{H}}_2(\tilde{\boldsymbol{u}},\boldsymbol{\theta})|^{-1/2} \end{aligned}$$

where, noting that $\boldsymbol{L}_1^T\boldsymbol{X} = \boldsymbol{I}$ and $\boldsymbol{S}\boldsymbol{X} = \boldsymbol{0}$,

$$\begin{aligned}\boldsymbol{h}_1(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\theta}) &= \frac{\partial h_1(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\theta})}{\partial \boldsymbol{u}} = \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{y}_1 - \boldsymbol{\beta} - \boldsymbol{y}_2^*)\\ \boldsymbol{\mathcal{H}}_1(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\theta}) &= \frac{\partial^2 h_1(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\theta})}{\partial \boldsymbol{u}\partial \boldsymbol{u}^T} = -\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\end{aligned} \tag{5.1.34}$$

and

$$\begin{aligned}\boldsymbol{h}_2(\boldsymbol{u},\boldsymbol{\theta}) &= \frac{\partial h_2(\boldsymbol{u},\boldsymbol{\theta})}{\partial \boldsymbol{u}} = \boldsymbol{Z}^T\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{u}) - \boldsymbol{G}^{-1}\boldsymbol{u}\\ \boldsymbol{\mathcal{H}}_2(\boldsymbol{u},\boldsymbol{\theta}) &= \frac{\partial^2 h_2(\boldsymbol{u},\boldsymbol{\theta})}{\partial \boldsymbol{u}\partial \boldsymbol{u}^T} = -\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z} - \boldsymbol{G}^{-1}\end{aligned} \tag{5.1.35}$$

For completeness, combining (5.1.34) and (5.1.35) gives

$$\boldsymbol{\mathcal{H}}_1(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\theta}) + \boldsymbol{\mathcal{H}}_2(\boldsymbol{u},\boldsymbol{\theta}) = -\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} - \boldsymbol{G}^{-1}$$

The terms of the approximate marginal likelihood, (5.1.33) can be written as

$$\begin{aligned}L_1(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y}) = (2\pi)^{-p/2}&|\boldsymbol{I} + \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}|^{-1/2}\\ &\times\ |\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}|^{1/2}\exp\Big\{-\tfrac{1}{2}(\boldsymbol{y}_1 - \boldsymbol{\beta} - \tilde{\boldsymbol{y}}_2^*)^T\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{y}_1 - \boldsymbol{\beta} - \tilde{\boldsymbol{y}}_2^*)\Big\}\\ \times\ \exp\Big\{&\tfrac{1}{2}(\boldsymbol{y}_1 - \boldsymbol{\beta} - \tilde{\boldsymbol{y}}_2^*)^T\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{y}_1 - \boldsymbol{\beta} - \tilde{\boldsymbol{y}}_2^*)\Big\}\end{aligned} \tag{5.1.36}$$

$$\begin{aligned}L_2(\boldsymbol{\theta};\boldsymbol{y}) = (2\pi)^{-(n-p)/2}&|\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z} + \boldsymbol{G}^{-1}|^{-1/2}|\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}|^{-1/2}|\boldsymbol{G}|^{-1/2}|\boldsymbol{R}|^{-1/2}\\ &\times\ \exp\left\{-\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{u}})^T\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{u}}) - \tfrac{1}{2}\tilde{\boldsymbol{u}}^T\boldsymbol{G}^{-1}\tilde{\boldsymbol{u}}\right\}\end{aligned} \tag{5.1.37}$$

where $\tilde{\boldsymbol{y}}_2^*$ is $\boldsymbol{y}_2^*$ evaluated at $\tilde{\boldsymbol{u}}$. $L_1(\cdot)$ is viewed as an approximate objective function analogous to a conditional likelihood which is free of the random effects and is used to estimate the location parameters $\boldsymbol{\beta}$. As $L_2(\cdot)$ does not contain $\boldsymbol{\beta}$ it is viewed as an approximate Restricted Likelihood used to estimate the remaining scale parameters, $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}$.

The estimated random effects required for the approximation, $\tilde{\boldsymbol{u}}$, are found by maximising $h_2(\boldsymbol{u},\boldsymbol{\theta})$ with respect to $\boldsymbol{u}$. Setting $\boldsymbol{h}_2(\boldsymbol{u},\boldsymbol{\theta}) = \boldsymbol{0}$ and solving for $\boldsymbol{u}$ gives

$$\begin{aligned}\boldsymbol{0} &= \boldsymbol{Z}^T\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{u}}) - \boldsymbol{G}^{-1}\tilde{\boldsymbol{u}}\\ \tilde{\boldsymbol{u}} &= (\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{y}\end{aligned} \tag{5.1.38}$$

These estimated random effects can also be written as

$$\begin{aligned}\tilde{\boldsymbol{u}} &= \boldsymbol{G}(\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z})(\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{y}\\ &= \boldsymbol{G}(\boldsymbol{I} - \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1})\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{y}\\ &= \boldsymbol{G}\boldsymbol{Z}^T(\boldsymbol{S} - \boldsymbol{S}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{S})\boldsymbol{y}\\ &= \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{y} \qquad \text{(using Result A.5.4)}\end{aligned}$$

where $\boldsymbol{P}$ is defined in section 2.3. Notice these random effects are identical to the predicted random effects (2.2.4) derived in Section 2.2. Using (5.1.38) and (5.1.39), $\tilde{\boldsymbol{y}}_2^*$ can be simplified

$$\begin{aligned}
\tilde{\boldsymbol{y}}_2^* &= \boldsymbol{L}_1^T(\boldsymbol{Z}\tilde{\boldsymbol{u}} + \boldsymbol{RS}(\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{u}})) \\
&= \boldsymbol{L}_1^T\left\{\boldsymbol{ZGZ}^T\boldsymbol{Py} + \boldsymbol{RS}(\boldsymbol{y} - \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{SZ} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{Sy})\right\} \\
&= \boldsymbol{L}_1^T\left\{(\boldsymbol{H} - \boldsymbol{R})\boldsymbol{Py} + \boldsymbol{R}(\boldsymbol{S} - \boldsymbol{SZ}(\boldsymbol{Z}^T\boldsymbol{SZ} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{S})\boldsymbol{y}\right\} \\
&= \boldsymbol{L}_1^T\left\{(\boldsymbol{H} - \boldsymbol{R})\boldsymbol{Py} + \boldsymbol{RPy}\right\} \qquad \text{(using Result A.5.4)} \\
&= \boldsymbol{L}_1^T\boldsymbol{HPy}
\end{aligned}$$

The exponent term of (5.1.36) can be expressed as

$$\begin{aligned}
& -\frac{1}{2}(\boldsymbol{y}_1 - \boldsymbol{\beta} - \boldsymbol{y}_2^*)^T\boldsymbol{X}^T(\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1})\boldsymbol{X}(\boldsymbol{y}_1 - \boldsymbol{\beta} - \boldsymbol{y}_2^*) \\
= \quad & -\frac{1}{2}(\boldsymbol{y}_1 - \boldsymbol{\beta} - \boldsymbol{y}_2^*)^T\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{y}_1 - \boldsymbol{\beta} - \boldsymbol{y}_2^*) \qquad \text{(using Result A.3.1)}
\end{aligned}$$

The determinant terms of (5.1.36) can be written as

$$\begin{aligned}
& |\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}|^{1/2}|\boldsymbol{I} + \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{SZ} + \boldsymbol{G}^{-1})^{-1}|^{-1/2} \\
= \quad & |\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}|^{1/2}|\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}|^{-1/2}|\boldsymbol{Z}^T\boldsymbol{SZ} + \boldsymbol{G}^{-1}|^{1/2} \\
= \quad & |\boldsymbol{C}|^{1/2}|\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}|^{-1/2} \qquad \text{(using Result A.5.5)} \\
= \quad & |\boldsymbol{C}|^{1/2}|\boldsymbol{R}|^{1/2}|\boldsymbol{G}|^{1/2}|\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}|^{-1/2}|\boldsymbol{R}|^{-1/2}|\boldsymbol{G}|^{-1/2} \\
= \quad & |\boldsymbol{H}|^{1/2}|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}|^{1/2}|\boldsymbol{H}|^{-1/2} \qquad \text{(using Result A.5.5 and A.2.2)} \\
= \quad & |\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}|^{1/2}
\end{aligned}$$

Therefore (5.1.36) is equivalent to the conditional likelihood (2.3.6) derived in Section 2.3.

Substituting the random effects into the components of the exponent term of (5.1.37) allows it to be expressed as

$$\begin{aligned}
& -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{SZ} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{Sy})^T\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{SZ} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{Sy}) \\
& -\frac{1}{2}\boldsymbol{y}^T\boldsymbol{PZGG}^{-1}\boldsymbol{GZ}^T\boldsymbol{Py} \\
= \quad & -\frac{1}{2}\boldsymbol{yPS}^{-1}\boldsymbol{Py} - \frac{1}{2}\boldsymbol{yP}(\boldsymbol{H} - \boldsymbol{R})\boldsymbol{Py} \qquad \text{(using Result A.5.4)} \\
= \quad & -\frac{1}{2}\boldsymbol{y}^T\boldsymbol{PRPy} - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{P}(\boldsymbol{H} - \boldsymbol{R})\boldsymbol{Py} \qquad (\boldsymbol{PX} = \boldsymbol{0}) \\
= \quad & -\frac{1}{2}\boldsymbol{y}^T\boldsymbol{Py} \qquad \text{(using Result A.5.2)}
\end{aligned}$$

The determinants of (5.1.37) can be expressed as

$$\begin{aligned} & |\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z}+\boldsymbol{G}^{-1}|^{-1/2}|\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}|^{-1/2}|\boldsymbol{G}|^{-1/2}|\boldsymbol{R}|^{-1/2} \\ = \quad & |\boldsymbol{C}|^{-1/2}|\boldsymbol{G}|^{-1/2}|\boldsymbol{R}|^{-1/2} \qquad \text{(using Result A.5.5)} \\ = \quad & |\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}|^{-1/2}|\boldsymbol{H}|^{-1/2} \qquad \text{(using Result A.2.2)} \end{aligned}$$

Therefore (5.1.37) is equivalent to the REML, (2.3.7), derived in Section 2.3.

5.2 Adjusted Likelihood Techniques

Profile Likelihoods have been widely used in the statistical literature for inference on parameters of interest. Let $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ be p and q length parameter vectors. For given $\boldsymbol{y}=(y_1,\ldots,y_n)$ the profile likelihood for $\boldsymbol{\theta}$ is defined as

$$L_P(\boldsymbol{\theta})=\exp(\ell(\boldsymbol{\theta},\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}};\boldsymbol{y})) \tag{5.2.1}$$

where $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$ is the ML estimate of $\boldsymbol{\phi}$ for a given $\boldsymbol{\theta}$. A common use of this profile likelihood is to calculate confidence limits for the parameter of interest (see Venzon & Mollgavkar, 1988). However, the profile likelihood is not a true likelihood. In particular, the moment properties of the profile score statistic $\partial L_P(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ do not share the same moment properties as the usual score statistic. An approach that accounts for the presence of nuisance parameters more appropriately is to consider adjustments to the profile likelihood function.

Inferential adjustments to a profile likelihood in the presence of nuisance parameters has been a long standing area of research. An overview of many techniques and their applications are described by Barndorff-Nielsen & Cox (1994). In particular, these authors build on previous work by Barndorff-Nielsen (1980) and Barndorff-Nielsen (1983) which extends the seminal work of Fisher (1934). Approximate techniques that are discussed in this section include *Modified Profile Likelihood* (MPL), *Conditional Profile Likelihood* (CPL) and *Stably Adjusted Profile Likelihood* (SAPL). An essential reference for all these approaches are found in Barndorff-Nielsen & Cox (1994).

5.2.1 Modified Profile Likelihood

The underlying principles for the discussion of MPL can be found in Fisher (1934) and is revised in Barndorff-Nielsen (1980). The latter of these authors considers the joint distribution of the parameter estimates $\hat{\boldsymbol{\psi}}=(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\phi}})$, given an ancillary statistic $\boldsymbol{a}$, factorised in the following form

$$p(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\phi}};\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{a})=p(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}|\boldsymbol{a})p(\hat{\boldsymbol{\phi}};\boldsymbol{\theta},\boldsymbol{\phi}|\hat{\boldsymbol{\theta}},\boldsymbol{a}) \tag{5.2.2}$$

For the parameter of interest, $\boldsymbol{\theta}$, the MPL is determined from the marginal distribution, $p(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}|\boldsymbol{a})$, as it is the most likely candidate for inference about $\boldsymbol{\theta}$. Consider the transformation $\hat{\boldsymbol{\phi}} \rightarrow \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$. Applying this to the second term of the RHS of (5.2.2) gives

$$p(\hat{\boldsymbol{\phi}}; \boldsymbol{\theta}, \boldsymbol{\phi}|\hat{\boldsymbol{\theta}}, \boldsymbol{a}) = p(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{\theta}, \boldsymbol{\phi}|\hat{\boldsymbol{\theta}}, \boldsymbol{a}) \left| \frac{\partial \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}}{\partial \hat{\boldsymbol{\phi}}^T} \right| \tag{5.2.3}$$

where the determinant term is the Jacobian required from the transformation. Applying the $p^* - formula$ (see Barndorff-Nielsen, 1980; Barndorff-Nielsen, 1983 and Barndorff-Nielsen & Cox, 1994, Section 6.2) to the first term of the RHS of (5.2.3) and the LHS of (5.2.2) gives

$$\begin{aligned} p^*(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}; \boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{a}) &\approx c(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{a}) |\mathcal{I}_o(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\psi}})|^{1/2} \exp\big(\ell(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}; \boldsymbol{y})\big) \\ p^*(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{\theta}, \boldsymbol{\phi}|\hat{\boldsymbol{\theta}}, \boldsymbol{a}) &\approx c(\boldsymbol{\phi}|\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}, \boldsymbol{a}) |\mathcal{I}_o(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\phi}})|^{1/2} \exp\big(\ell(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{y}) - \ell(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y})\big) \end{aligned}$$

Barndorff-Nielsen & Cox (1994) note the validity of the second equation is contingent on the assumption that the marginal distribution of $\hat{\boldsymbol{\theta}}$ does not depend on $\boldsymbol{\phi}$. The implication is that for fixed $\boldsymbol{\theta}$, $(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}, \boldsymbol{a})$ remains sufficient and $(\hat{\boldsymbol{\theta}}, \boldsymbol{a})$ can be considered to be ancillary, thus allowing the $p^* - formula$ to be applied appropriately (see Barndorff-Nielsen, 1983 and Barndorff-Nielsen & Cox, 1994, Section 8.2, pg. 266 for further details). Discarding norming constants, the conditional distribution of $\hat{\boldsymbol{\theta}}$ given $\boldsymbol{a}$ can then be written as

$$p(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}|\boldsymbol{a}) \approx \frac{|\mathcal{I}_o(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\psi}})|^{1/2}}{|\mathcal{I}_o(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})|^{1/2}} \left| \frac{\partial \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}}{\partial \hat{\boldsymbol{\phi}}^T} \right|^{-1} \exp\big(\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}; \boldsymbol{y})\big) \tag{5.2.4}$$

By discarding terms that are functions of the data only this may be simplified to provide an MPL for $\boldsymbol{\theta}$, namely

$$L_{MP}(\boldsymbol{\theta}) = |\mathcal{I}_o(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})|^{-1/2} \left| \frac{\partial \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}}{\partial \hat{\boldsymbol{\phi}}^T} \right|^{-1} \exp\big(\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y})\big) \tag{5.2.5}$$

An alternative expression for MPL can be derived. Let $\ell_{\boldsymbol{\phi}}(\cdot)$ be the derivative of the full likelihood with respect to ϕ. Evaluating this score equation at $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$ and setting it to zero gives

$$\ell_{\boldsymbol{\phi}}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y}) = 0$$

Taking the second derivative of this identity with respect to $\hat{\phi}$ gives

$$\ell_{\boldsymbol{\phi};\boldsymbol{\phi}}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y}) \frac{\partial \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}}{\partial \hat{\boldsymbol{\phi}}^T} + \ell_{\boldsymbol{\phi};\hat{\boldsymbol{\phi}}}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y}) = 0$$

where $\ell_{\boldsymbol{\phi};\hat{\boldsymbol{\phi}}}(\cdot)$ represents the matrix of second derivatives of the log-likelihood function with respect to the parameters $\boldsymbol{\phi}$ and $\hat{\boldsymbol{\phi}}$ respectively. Rearranging

$$\frac{\partial \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}}{\partial \hat{\boldsymbol{\phi}}^T} = \frac{\ell_{\boldsymbol{\phi};\hat{\boldsymbol{\phi}}}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y})}{\mathcal{I}_o(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})}$$

Substituting this into (5.2.5) the MPL is found to be

$$L_{MP}(\boldsymbol{\theta}) = \frac{|\mathcal{I}_o(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})|^{1/2}}{|\ell_{\boldsymbol{\phi};\hat{\boldsymbol{\phi}}}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y})|} \exp\big(\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y})\big) \tag{5.2.6}$$

A consequence of the term, $|\partial\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}/\partial\hat{\boldsymbol{\phi}}^T|$ on the RHS of (5.2.3), generated by the Jacobian from the transformation is that if $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$ is independent of $\boldsymbol{\theta}$ then it becomes unity and MPL can be reduced to

$$L^o_{MP}(\boldsymbol{\theta}) = |\mathcal{I}_o(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})|^{-1/2} \exp\big(\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y})\big) \tag{5.2.7}$$

Following Barndorff-Nielsen & Cox (1994) the MPLs (5.2.6) and (5.2.7) have the advantageous property of invariance under interest respecting parameter transformations. Furthermore, the accuracy of each approximation is $\mathcal{O}(n^{-3/2})$. Barndorff-Nielsen (1983) also states that the parameter invariance is maintained if the observed information for $\boldsymbol{\phi}$ is replaced by its expected information producing the MPL

$$L^e_{MP}(\boldsymbol{\theta}) = |\mathcal{I}_e(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})|^{-1/2} \exp\big(\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y})\big) \tag{5.2.8}$$

5.2.2 Parameter Orthogonality and Conditional Profile Likelihood

To ensure simplicity of inference and estimation techniques it is sometimes useful to orthogonalise parameters. Following Cox & Reid (1987) let $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)$ be a set of nuisance parameters orthogonal to θ, a scalar parameter of interest. Thus the original p parameters are defined $(\phi_1(\theta, \boldsymbol{\psi}), \ldots, \phi_p(\theta, \boldsymbol{\psi}))$. For given data $\boldsymbol{y}$, the likelihood for the parameterization can be defined by

$$\ell(\theta, \boldsymbol{\psi}; \boldsymbol{y}) = \ell_o(\theta, \phi_1(\theta, \boldsymbol{\psi}), \ldots, \phi_p(\theta, \boldsymbol{\psi}); \boldsymbol{y})$$

Taking derivatives of both sides of this equation with respect to θ gives

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial \ell_o}{\partial \theta} + \sum_{i=1}^{p} \frac{\partial \ell_o}{\partial \phi_i} \frac{\partial \phi_i}{\partial \theta}$$

To obtain the co-information between the parameter of interest and the new orthogonal parameters the next derivative is taken with respect to the kth orthogonal parameter ψ_k, namely

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \theta \partial \psi_k} &= \sum_{j=1}^{p} \frac{\partial^2 \ell_o}{\partial \theta \partial \phi_j} \frac{\partial \phi_j}{\partial \psi_k} + \sum_{j=1}^{p} \sum_{i=1}^{p} \frac{\partial^2 \ell_o}{\partial \phi_i \partial \phi_j} \frac{\partial \phi_i}{\partial \theta} \frac{\partial \phi_j}{\partial \psi_k} + \sum_{i=1}^{p} \frac{\partial \ell_o}{\partial \phi_i} \frac{\partial^2 \phi_i}{\partial \theta \partial \psi_k} \\
&= \sum_{j=1}^{p} \left\{ \frac{\partial^2 \ell_o}{\partial \theta \partial \phi_j} + \sum_{i=1}^{p} \frac{\partial^2 \ell_o}{\partial \phi_i \partial \phi_j} \frac{\partial \phi_i}{\partial \theta} \right\} \frac{\partial \phi_j}{\partial \psi_k} + \sum_{i=1}^{p} \frac{\partial \ell_o}{\partial \phi_i} \frac{\partial^2 \phi_i}{\partial \theta \partial \psi_k}
\end{aligned}$$

$\mathrm{E}(\partial \ell_o / \partial \phi_i) = \mathbf{0},\ i = 1, \ldots, n$ due to the unbiasedness of the likelihood score function and therefore taking expectations of both sides gives

$$\mathcal{I}_e(\theta, \psi_k) = \sum_{i=j}^{p} \left\{ \mathcal{I}_e(\theta, \phi_j) + \sum_{i=1}^{p} \mathcal{I}_e(\phi_i, \phi_j) \frac{\partial \phi_i}{\partial \theta} \right\} \frac{\partial \phi_j}{\partial \psi_k}, \qquad k = 1, \ldots, p$$

Assuming orthogonality between θ and $\psi_k,\ k = 1, \ldots, p$ the LHS becomes zero. The Jacobian of the transformation $\partial \phi_i / \partial \psi_k,\ i, j = 1, \ldots, p$ is non-zero and therefore the orthogonality equations can be reduced to

$$\sum_{j=1}^{p} \sum_{i=1}^{p} \mathcal{I}_e(\phi_i, \phi_j) \frac{\partial \phi_i}{\partial \theta} = -\sum_{j=1}^{p} \mathcal{I}_e(\theta, \phi_j)$$

and in matrix notation

$$\mathcal{I}_e(\boldsymbol{\phi}, \boldsymbol{\phi}) \frac{\partial \boldsymbol{\phi}}{\partial \theta} = -\mathcal{I}_e(\theta, \boldsymbol{\phi}) \tag{5.2.9}$$

It is clear that this equation does not depend on the new orthogonal parameters but provides a mechanism for evaluation of the dependence between θ and $\boldsymbol{\phi}$. Therefore the choice of the functional relationship between $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ is somewhat arbitrary (see Cox & Reid, 1987 for details). It also must be noted that the orthogonality equations (5.2.9) only hold when the parameter of interest, θ, is scalar (see Cox & Reid, 1987 or Barndorff-Nielsen & Cox, 1994, Section 2.7, pg. 50).

If $\boldsymbol{\psi}$ and θ are orthogonal parameters then from (iv) in Section 2.2 of Cox & Reid (1987) the maximum likelihood estimate of $\boldsymbol{\psi}$ when θ is given, $\hat{\boldsymbol{\psi}}_\theta$, varies slowly with θ (see Cox & Reid, 1987 or Barndorff-Nielsen & Cox, 1994, Section 2.7 for details). Therefore, using (5.2.5) and forming the modified profile likelihood for θ, the determinant term containing the Jacobian of the transformation can be ignored and the CPL of θ can be expressed as

$$L_{CP}(\boldsymbol{\theta}) = |\mathcal{I}_o(\hat{\boldsymbol{\psi}}_\theta, \hat{\boldsymbol{\psi}}_\theta)|^{-1/2} \exp\big(\ell(\theta, \hat{\boldsymbol{\psi}}_\theta; \boldsymbol{y})\big) \tag{5.2.10}$$

In the discussion of Cox & Reid (1987) and Barndorff-Nielsen & Cox (1994) and in more detail by Cox & Reid (1992) and Barndorff-Nielsen & McCullagh (1993) it is found that this approximate likelihood approach unfortunately lacks the parameter invariance of MPL, (5.2.7), derived in the previous section. Furthermore, due to the omission of the Jacobian, its accuracy is $\mathcal{O}(n^{-1})$ in comparison to (5.2.7) which is of $\mathcal{O}(n^{-3/2})$.

5.2.3 Laplace's method, MPL and CPL

The adjustments to the profile likelihood in Section 5.2.1 and 5.2.2 were derived under specialised conditional distribution theory. In particular, the ancillary statistic, $\boldsymbol{a}$, has to be available to provide neat factorisation of the joint distribution of $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ in (5.2.2).

The Laplace approximation techniques of Section 5.1.2 do not require an available ancillary statistic for inference to proceed. This section illustrates the connection of Laplace's method and the derived modified and conditional profile likelihoods of the previous sections.

Let $\boldsymbol{\theta}$ be the parameter of interest and $\boldsymbol{\phi}$ the nuisance parameter then if the log-likelihood has the form $\ell(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{y})$ then using section 5.1.2 the marginal likelihood can be expressed as

$$L_{LA}(\boldsymbol{\theta}) = \int_{\mathcal{R}^r} \exp(\ell(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{y}))\pi(\boldsymbol{\phi})d\boldsymbol{\phi}$$

where $\pi(\boldsymbol{\phi})$ is the prior density of the nuisance parameter. Following Section 5.1.2 by expanding the integrand in a Taylor series at some maximum of $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$ and using (5.1.4), the Laplace approximation to the marginal likelihood can be expressed as

$$L_{LA}(\boldsymbol{\theta}) = (2\pi)^{q/2}|\mathcal{I}_o(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})|^{-1/2}\pi(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}) \exp(\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y}))$$

If the range of interest for $\boldsymbol{\theta}$ varies within $\mathcal{O}(n^{-1/2})$ then $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}} - \boldsymbol{\phi}$ varies within $\mathcal{O}(n^{-1})$. From this the prior density for $\boldsymbol{\phi}$ evaluated at $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$, $\pi(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})$, is also $\mathcal{O}(n^{-1})$ and can be ignored. Omitting constants, the Laplace approximation can then be expressed as

$$L_{LA}(\boldsymbol{\theta}) = |\mathcal{I}_o(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})|^{-1/2} \exp(\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y})) \tag{5.2.11}$$

with accuracy $\mathcal{O}(n^{-1})$. From (5.2.10) it can be seen that the $L_{CP}(\cdot) = L_{LA}(\cdot)$ to $\mathcal{O}(n^{-1})$ and identical to CPL the Laplacian approximation also suffers from a lack of parameter invariance when $\boldsymbol{\theta}$ is not independent of $\boldsymbol{\phi}$.

For clarity, the equivalence relationship between the asymptotic approximations derived here can be presented as follows. When $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are independent the following relationship holds to the accuracy $\mathcal{O}(n^{-1})$,

$$\boldsymbol{L}^o_{MP}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}; \boldsymbol{y}) \doteq \boldsymbol{L}_{CP}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}; \boldsymbol{y}) \doteq \boldsymbol{L}_{LA}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}; \boldsymbol{y})$$

Furthermore, $\boldsymbol{L}^o_{MP}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}; \boldsymbol{y})$ and $\boldsymbol{L}_{CP}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}; \boldsymbol{y})$ are equivalent to $\mathcal{O}(n^{-3/2})$. The independence between the parameters, in this particular case, also ensures that exact parameter invariance is maintained in all three approximations.

5.2.4 Extending the Modified Profile likelihood

This section discusses an alternative adjusted profile likelihood that retains the nice properties of MPL without the requirement of an explicit ancillary statistic. In particular, the *Stably Adjusted Profile Likelihood* (SAPL) is invariant under transformation of the parameter of interest and is accurate to the order $\mathcal{O}(n^{-1})$. The technical derivation of this

approximate likelihood is given in Barndorff-Nielsen (1994), Barndorff-Nielsen & Chamberlain (1994) and Barndorff-Nielsen & Cox (1994), Section 8.3 and the theory presented below follows this very closely.

The MPL given by (5.2.6) can be also be expressed as

$$L_{MP}(\boldsymbol{\theta}) = K(\boldsymbol{\theta})\big\{|\mathcal{I}_e(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})|/|\mathcal{I}_o(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})|\big\}^{1/2} \exp(\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y})) \tag{5.2.12}$$

where

$$K(\boldsymbol{\theta}) = |\mathcal{I}_o(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})|/\big\{|\ell_{\boldsymbol{\phi};\hat{\boldsymbol{\phi}}}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y})||\mathcal{I}_e(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})|\big\}^{1/2}$$

If an ancillary statistic is not available then, in general, it is very difficult to determine the appropriate differentiation for the component $\ell_{\boldsymbol{\phi};\hat{\boldsymbol{\phi}}}(\cdot)$. To circumvent this a linearisation of the parameter of interest is performed on $\log K(\boldsymbol{\theta}) = k(\boldsymbol{\theta})$ around the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$. Let $\mathcal{I}_e^{(\theta_j)}(\cdot,\cdot)$ be the derivative of a given expected information with respect to θ_j. The linearisation can then be expressed as

$$k^*(\boldsymbol{\theta}) = k(\boldsymbol{\theta}) - k(\hat{\boldsymbol{\theta}}) = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{k}'(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}) \tag{5.2.13}$$

where

$$\boldsymbol{k}'(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}) = (\mathrm{tr}(\boldsymbol{A}_1), \ldots, \mathrm{tr}(\boldsymbol{A}_p))^T$$

and

$$\boldsymbol{A}_j = (\mathcal{I}_e(\boldsymbol{\phi}, \boldsymbol{\phi}))^{-1}\Big\{\mathcal{I}_e^{(\boldsymbol{\phi})}(\boldsymbol{\phi}, \theta_j) - \mathcal{I}_e^{(\boldsymbol{\phi})}(\boldsymbol{\phi}, \boldsymbol{\phi})(\mathcal{I}_e(\boldsymbol{\phi}, \boldsymbol{\phi}))^{-1}\mathcal{I}_e(\boldsymbol{\phi}, \theta_j) - \frac{1}{2}\boldsymbol{B}_j\Big\}\Big|_{\theta_j=\hat{\theta}_j;\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}$$

with

$$\boldsymbol{B}_j = \mathcal{I}_e^{(\theta_j)}(\boldsymbol{\phi}, \boldsymbol{\phi}) - \mathcal{I}_e^{(\boldsymbol{\phi})}(\boldsymbol{\phi}, \boldsymbol{\phi})\mathcal{I}_e(\boldsymbol{\phi}, \boldsymbol{\phi})\mathcal{I}_e(\boldsymbol{\phi}, \theta_j)$$

The technical derivation of this result is outside the scope of this thesis and therefore has been omitted (see Barndorff-Nielsen, 1994 or Barndorff-Nielsen & Cox, 1994, Section 8.3 for the derivation and details). Following Barndorff-Nielsen & Cox (1994), Section 8.3, pg. 270-271, the first position where $\mathcal{I}_e^{(\boldsymbol{\phi})}(\boldsymbol{\phi}, \boldsymbol{\phi})$ occurs the matrix multiplication is with respect to the second $\boldsymbol{\phi}$ and the second position it occurs the matrix multiplication is with respect to the third $\boldsymbol{\phi}$.

Under parameter orthogonality, $\mathcal{I}_e(\boldsymbol{\phi}, \boldsymbol{\theta}) = \boldsymbol{0}$, and therefore

$$\boldsymbol{A}_j = -\frac{1}{2}(\mathcal{I}_e(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}))^{-1}\mathcal{I}_e^{(\theta_j)}(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})\Big|_{\theta_j=\hat{\theta}_j}$$

allowing for considerable simplification. It is then argued in Barndorff-Nielsen (1994) and Barndorff-Nielsen & Cox (1994) that under transformations of the parameter of interest, $\boldsymbol{\theta}$, $k^*(\boldsymbol{\theta})$ is invariant. Therefore the SAPL proposed can be expressed as

$$\boldsymbol{L}_S(\boldsymbol{\theta}) = \exp(k^*(\boldsymbol{\theta}))\big\{|\mathcal{I}_e(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})|/|\mathcal{I}_o(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})|\big\}^{1/2} \exp(\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}; \boldsymbol{y})) \tag{5.2.14}$$

and is parameter invariant under transformation of $\boldsymbol{\theta}$.

Chapter 6

Heteroscedastic t-REML with known degrees of freedom

Chapter 4 focussed on ML estimation of the heteroscedastic t-distribution when the degrees of freedom is known. This chapter extends these results by deriving an approximate REML for the heteroscedastic t-distribution using approximate likelihood techniques of the previous chapter.

The first of these approximations requires the use of the Partial Laplace approximation (see Section 5.1.3) to the heteroscedastic t-distribution when the degrees of freedom is known. This technique exploits the component form of the integrand given by (4.2.4). The methodology of REML derived in Section 2.3 is applied to the conditional Gaussian component of the integrand before the random scale effects are integrated out. The approximate marginal likelihood derived can then be partitioned to obtain approximate REML equivalents for the location and scale parameters separately.

As the heteroscedastic t with known degrees of freedom is from the location-scale family the second approximation uses the adjusted likelihood technique MPL (see Section 5.2.1) to obtain an approximate REML equivalent. The derivation supplied shows that this is equivalent to the approximate likelihood derived in James et al. (1993) and similar to the Gaussian REML derived in Chapter 2.

6.1 Heteroscedastic t-REML using the Partial Laplace approximation

Consider the model defined by (3.1.1) where conditional on the random scale effects, $\omega_i,\ i = 1, \ldots, n$ the response is distributed as given by (4.2.1) and the random scale

effects have distribution defined by (4.2.2). In this section the scale parameter model is defined by (3.1.2).

6.1.1 Notation

Again, the marginal likelihood for the heteroscedastic t-distribution is defined by (4.2.4) where the components of the integrand are the likelihood for the conditional response given by (4.2.5) and the likelihood for the random scale effects given by (4.2.6). Given matrices $\boldsymbol{L} = [\boldsymbol{L}_1, \boldsymbol{L}_2]^T$ that satisfy the conditions (2.3.1) allows the conditional response to be transformed to $\boldsymbol{y} = (\boldsymbol{L}_1^T\boldsymbol{y}, \boldsymbol{L}_2^T\boldsymbol{y})^T = (\boldsymbol{y}_1, \boldsymbol{y}_2)^T$. Using (2.3.2) the conditional distribution of this transformed response becomes

$$\begin{bmatrix} \boldsymbol{L}_1^T\boldsymbol{y} \\ \boldsymbol{L}_2^T\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} \Big| \boldsymbol{\omega} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{L}_1^T\boldsymbol{\Psi}\boldsymbol{L}_1 & \boldsymbol{L}_1^T\boldsymbol{\Psi}\boldsymbol{L}_2 \\ \boldsymbol{L}_2^T\boldsymbol{\Psi}\boldsymbol{L}_1 & \boldsymbol{L}_2^T\boldsymbol{\Psi}\boldsymbol{L}_2 \end{bmatrix} \right) \tag{6.1.1}$$

From this the appropriate conditional REML likelihood can be expressed as

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y}|\boldsymbol{\omega}) = L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y}_1|\boldsymbol{y}_2, \boldsymbol{\omega}) L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y}_2|\boldsymbol{\omega})$$

where from (2.3.6) and (2.3.7),

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y}_1|\boldsymbol{y}_2, \boldsymbol{\omega}) &= (2\pi)^{-p/2} |\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X}|^{1/2} \\ &\times \exp\left\{-\tfrac{1}{2}(\boldsymbol{y}_1 - \boldsymbol{\beta} - \boldsymbol{y}_2^*)^T \boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{y}_1 - \boldsymbol{\beta} - \boldsymbol{y}_2^*)\right\} \quad (6.1.2) \\ L(\boldsymbol{\lambda}; \boldsymbol{y}_2|\boldsymbol{\omega}) &= (2\pi)^{-(n-p)/2} |\boldsymbol{\Psi}|^{-1/2} |\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X}|^{-1/2} \exp\left\{-\tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{S}\boldsymbol{y}\right\} \quad (6.1.3) \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{S} = \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{\Psi}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T &= \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Psi}^{-1} \quad (6.1.4) \\ \boldsymbol{y}_2^* &= \boldsymbol{L}_1^T\boldsymbol{\Psi}\boldsymbol{S}\boldsymbol{y} \end{aligned}$$

Identical to Section 2.3, the columns of the transformation matrix $\boldsymbol{L}_1$ define a p set of location contrasts associated with the location parameters. Therefore, conditionally on the random scale effects, $\omega_i,\ i = 1, \ldots, n$, (6.1.2) can only be used to estimate the location parameters, $\boldsymbol{\beta}$. Thus the columns of $\boldsymbol{L}_2$ define an $n - p$ set of contrasts and therefore (6.1.3) is used to conditionally estimate the scale parameters, $\boldsymbol{\lambda}$. Based on results from Section 2.3 the conditional likelihood for the location fixed effects, (6.1.2) can also be expressed as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y}_1|\boldsymbol{y}_2, \boldsymbol{\omega}) =& (2\pi)^{-p/2} |\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X}|^{1/2} \\ &\times \exp\left\{-\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{S}\boldsymbol{y}\right\} \end{aligned}$$

It is clear that multiplying with (6.1.3) the likelihood reverts to the standard conditional likelihood given in (4.2.5).

6.1.2 Laplace Approximation

The marginal likelihood can be written as the multidimensional integral

$$\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y}) = \int_{\mathcal{R}^n} p(\boldsymbol{y}_1|\boldsymbol{y}_2, \boldsymbol{\omega}; \boldsymbol{\beta}, \boldsymbol{\lambda}) p(\boldsymbol{y}_2|\boldsymbol{\omega}; \boldsymbol{\lambda}) p(\boldsymbol{\omega}; \nu) d\boldsymbol{\omega} \tag{6.1.5}$$

Identical to Section 5.1.2 the random scale effects are treated as nuisance parameters and therefore must be integrated out. Let

$$\begin{aligned} h_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\omega}) &= \log p(\boldsymbol{y}_1|\boldsymbol{y}_2, \boldsymbol{\omega}; \boldsymbol{\beta}, \boldsymbol{\lambda}) \\ h_2(\boldsymbol{\lambda}, \boldsymbol{\omega}, \nu) &= \log p(\boldsymbol{y}_2|\boldsymbol{\omega}; \boldsymbol{\lambda}) + \log p(\boldsymbol{\omega}; \nu) \end{aligned}$$

where

$$\begin{aligned} \log p(\boldsymbol{y}_1|\boldsymbol{y}_2, \boldsymbol{\omega}; \boldsymbol{\beta}, \boldsymbol{\lambda}) &= -\tfrac{1}{2}\big\{(n-p)\log(2\pi) - \log|\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X}| \\ &+ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{y}^T\boldsymbol{S}\boldsymbol{y}\big\} \\ \log p(\boldsymbol{y}_2|\boldsymbol{\omega}; \boldsymbol{\lambda}) &= -\tfrac{1}{2}\big\{p\log(2\pi) + \log|\boldsymbol{\Psi}| + \log|\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X}| + \boldsymbol{y}^T\boldsymbol{S}\boldsymbol{y}\big\} \end{aligned}$$

and $p(\boldsymbol{\omega}; \nu)$ is defined by (4.2.6). The marginal likelihood given in (6.1.5) can now be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y}) = \int_{\mathcal{R}^n} \exp\left\{h_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\omega}) + h_2(\boldsymbol{\lambda}, \boldsymbol{\omega}, \nu)\right\} d\boldsymbol{\omega} \tag{6.1.6}$$

As $h_1(\cdot)$ contains no information about the scale parameters it seems reasonable that to integrate out the random scale effects the integrand is expanded in a Taylor series around some maximum value of $\boldsymbol{\omega}$, say $\tilde{\boldsymbol{\omega}}$ of $h_2(\cdot)$ only. This allows the Partial Laplace approximation of Section 5.1.3 to be used. Let $\boldsymbol{h}_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\omega})$ and $\boldsymbol{\mathcal{H}}_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\omega})$ be the first and second derivative of $h_1(\cdot)$ with respect to $\boldsymbol{\omega}$ and let $\boldsymbol{h}_2(\boldsymbol{\lambda}, \boldsymbol{\omega}, \nu)$ and $\boldsymbol{\mathcal{H}}_2(\boldsymbol{\lambda}, \boldsymbol{\omega}, \nu)$ be the first and second derivative of $h_2(\cdot)$ with respect to $\boldsymbol{\omega}$. Using (5.1.27) the Partial Laplace approximation to (6.1.6) can be expressed as

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y}) = (2\pi)^{n/2} L_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y}) L_2(\boldsymbol{\lambda}, \nu; \boldsymbol{y})$$

where

$$\begin{aligned} L_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y}) &= \big|\boldsymbol{I} + (-\boldsymbol{\mathcal{H}}_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}}))(-\boldsymbol{\mathcal{H}}_2(\boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}}, \nu)^{-1})\big|^{-1/2} \exp(h_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}})) \\ &\times \exp\big\{\frac{1}{2}\boldsymbol{h}_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}})^T\big(-\boldsymbol{\mathcal{H}}_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}}) - \boldsymbol{\mathcal{H}}_2(\boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}}, \nu)\big)^{-1}\boldsymbol{h}_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}})\big\} \end{aligned} \tag{6.1.7}$$

$$L_2(\boldsymbol{\lambda}, \nu; \boldsymbol{y}) = \exp\big(h_2(\boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}}, \nu)\big)\big|-\boldsymbol{\mathcal{H}}_2(\boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}}, \nu)\big|^{-1/2} \tag{6.1.8}$$

The derivatives required for the approximation can be found using Section 2.3. Taking the first derivative of $h_2(\cdot)$ with respect to ω_i gives

$$\frac{\partial h_2(\cdot)}{\partial \omega_i} = -\frac{1}{2}\mathrm{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i) + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\boldsymbol{y} + \left(\frac{\nu}{2} - 1\right)\frac{1}{\omega_i} - \frac{\nu}{2} \tag{6.1.9}$$

where $\dot{\boldsymbol{\Psi}}_i = \partial\boldsymbol{\Psi}/\partial\omega_i$. Taking the derivative of (6.1.9) with respect to ω_i gives

$$\begin{aligned}\frac{\partial^2 h_2(\cdot)}{\partial\omega_i\partial\omega_i} &= -\frac{1}{2}\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_{ii}) + \frac{1}{2}\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i) + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_{ii}\boldsymbol{S}\boldsymbol{y} \\ &\quad -\boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\boldsymbol{y} - \left(\frac{\nu}{2}-1\right)\frac{1}{\omega_i^2} \qquad (6.1.10)\end{aligned}$$

where $\dot{\boldsymbol{\Psi}}_{ii} = \partial^2\boldsymbol{\Psi}/(\partial\omega_i)^2$. Taking the derivative of (6.1.9) with respect to ω_j gives

$$\frac{\partial^2 h_2(\cdot)}{\partial\omega_i\partial\omega_j} = -\frac{1}{2}\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_{ij}) + \frac{1}{2}\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_j) + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_{ij}\boldsymbol{S}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_j\boldsymbol{S}\boldsymbol{y} \qquad (6.1.11)$$

where $\dot{\boldsymbol{\Psi}}_{ij} = \partial^2\boldsymbol{\Psi}/\partial\omega_i\partial\omega_j$. Using (6.1.9) and (6.1.11) the first and second derivatives of $h_1(\cdot)$ can be immediately written as

$$\begin{aligned}\frac{\partial h_1(\cdot)}{\partial\boldsymbol{\omega}_i} &= \frac{1}{2}\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i) - \frac{1}{2}\text{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i) - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\boldsymbol{y} \\ &+ \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \qquad (6.1.12) \\ \frac{\partial^2 h_1(\cdot)}{\partial\omega_i\partial\omega_j} &= \frac{1}{2}\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_{ij}) - \frac{1}{2}\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_j) - \frac{1}{2}\text{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ij}) + \frac{1}{2}\text{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_j) \\ &- \frac{1}{2}\boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_{ij}\boldsymbol{S}\boldsymbol{y} + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ij}\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\ &+ \boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_j\boldsymbol{S}\boldsymbol{y} - (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_j\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \qquad (6.1.13)\end{aligned}$$

Consider the iith diagonal element of the second derivative of $h_2(\cdot)$. Let h_{ii} be the ith diagonal element of the hat matrix $\boldsymbol{H} = \boldsymbol{\Psi}^{-1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Psi}^{-1/2}$ Using (5.1.14) and (5.1.15) and considering the trace terms of (6.1.10) separately

$$\begin{aligned}\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_{ii}) &= \text{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ii}) - \text{tr}(\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ii}) \\ &= \text{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ii}) - \text{tr}(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ii}) \\ &= \frac{2}{\omega_i^2} - \frac{2h_{ii}}{\omega_i^2}\end{aligned}$$

and

$$\begin{aligned}\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i) &= \text{tr}(\boldsymbol{\Psi}^{-1/2}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\Psi}^{-1/2}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1/2}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\Psi}^{-1/2}) \\ &= \text{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i) - 2\text{tr}(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i) + \text{tr}(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i) \\ &= \frac{1}{\omega_i^2} - \frac{2h_{ii}}{\omega_i^2} + \frac{h_{ii}^2}{\omega_i^2}\end{aligned}$$

Let $\hat{\boldsymbol{\beta}}_c = (\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{y}$ be a conditional estimator for the location parameters and $d_{i,c} = (y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_c)^2$. The non-trace terms of (6.1.10) can be expressed as

$$\begin{aligned}\boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_{ii}\boldsymbol{S}\boldsymbol{y} &= (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_c)^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ii}\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_c) \\ &= \frac{2(y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_c)^2}{\sigma_i^2\omega_i} \\ &= \frac{d_{i,c}}{\sigma_i^2\omega_i}\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\boldsymbol{y} &= (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_c)^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_c) \\
&= (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_c)^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1/2}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\Psi}^{-1/2}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_c) \\
&= (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&- (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1/2}\boldsymbol{H}\boldsymbol{\Psi}^{-1/2}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&= \frac{(y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_c)^2}{\sigma_i^2\omega_i} - \frac{(y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_c)^2h_{ii}}{\sigma_i^2\omega_i} \\
&= \frac{d_{i,c}}{\sigma_i^2\omega_i} - \frac{d_{i,c}h_{ii}}{\sigma_i^2\omega_i}
\end{aligned}$$

The *off diagonal* ijth element of the second derivative of $h_2(\cdot)$ contains two terms. From the trace terms

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_j) &= \mathrm{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_j) - 2\mathrm{tr}(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_j) + \mathrm{tr}(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_j) \\
&= \mathrm{tr}(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_j) \\
&= \frac{h_{ij}^2}{\omega_i\omega_j} \qquad (6.1.14)
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{y}^T\boldsymbol{S}\boldsymbol{\Psi}_i\boldsymbol{S}\boldsymbol{\Psi}_j\boldsymbol{S}\boldsymbol{y} &= (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&- (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_c)^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1/2}\boldsymbol{H}\boldsymbol{\Psi}^{-1/2}\dot{\boldsymbol{\Psi}}_j\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_c) \\
&= -\frac{(y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_c)h_{ij}(y_j-\boldsymbol{x}_j^T\hat{\boldsymbol{\beta}}_c)}{\sigma_i\sigma_j(\omega_i\omega_j)^{1/2}} \\
&= -\frac{r_{i,c}}{\sigma_i(\omega_i)^{1/2}}h_{ij}\frac{r_{j,c}}{\sigma_j(\omega_j)^{1/2}} \qquad (6.1.15)
\end{aligned}$$

where $r_{i,c} = (d_{i,c})^{1/2}$ is the ith residual. Combining all the terms the iith diagonal element of the second derivative of $h_2(\cdot)$ evaluated at $\tilde{\boldsymbol{\omega}}$ becomes,

$$\left.\frac{\partial^2 h_2(\cdot)}{(\partial\omega_i)^2}\right|_{\boldsymbol{\omega}=\tilde{\boldsymbol{\omega}}} = \frac{1}{2\tilde{\omega}_i^2}\left(1-\nu+\tilde{h}_{ii}^2+2\frac{\tilde{d}_{i,c}\,\tilde{h}_{ii}\,\tilde{\omega}_i}{\sigma_i^2}\right)$$

where $\tilde{h}_{ii}$ is the iith diagonal element of the hat matrix evaluated at $\tilde{\boldsymbol{\omega}}$. Similarly the ijth element of the the second derivative of $h_2(\cdot)$ evaluated at $\tilde{\boldsymbol{\omega}}$ is

$$\left.\frac{\partial^2 h_2(\cdot)}{\partial\omega_i\partial\omega_j}\right|_{\boldsymbol{\omega}=\tilde{\boldsymbol{\omega}}} = \frac{1}{2\tilde{\omega}_i}\left(\tilde{h}_{ij}+2\frac{\tilde{r}_{i,c}\,\tilde{h}_{ij}\,\tilde{r}_{j,c}(\tilde{\omega}_i\tilde{\omega}_j)^{1/2}}{\sigma_i\sigma_j}\right)\frac{1}{\tilde{\omega}_j} \qquad (6.1.16)$$

where $\tilde{r}_{i,c}$ is the ith conditional residual evaluated at $\tilde{\boldsymbol{\omega}}$. In matrix notation

$$\boldsymbol{\mathcal{H}}_2(\boldsymbol{\lambda},\tilde{\boldsymbol{\omega}},\nu) = -\frac{1}{2}\tilde{\boldsymbol{\Omega}}^{-1}(\boldsymbol{V}-\tilde{\boldsymbol{H}}^2-2\tilde{\boldsymbol{D}}_c^{1/2}\tilde{\boldsymbol{H}}\tilde{\boldsymbol{D}}_c^{1/2})\tilde{\boldsymbol{\Omega}}^{-1}$$

where

$$\tilde{\boldsymbol{D}}_c = \text{diag}\Big\{\frac{\tilde{\omega}_i \tilde{d}_{i,c}}{\sigma_i^2}\Big\}, \qquad \boldsymbol{V} = (\nu - 1)\boldsymbol{I}$$

$\tilde{\boldsymbol{H}}^2$ contains the squared elements of the hat matrix, $\tilde{\boldsymbol{H}}$ and $\tilde{\boldsymbol{\Omega}}$ is a diagonal matrix with ith element $\tilde{\omega}_i$.

Noting (5.1.16) suggests that the two terms containing the location parameters, $\boldsymbol{\beta}$, in (6.1.13) are equal and the second derivative of $h_1(\cdot)$ can be immediately written as

$$\boldsymbol{\mathcal{H}}_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}}) = -\frac{1}{2}\tilde{\boldsymbol{\Omega}}^{-1}(\tilde{\boldsymbol{H}}^2 + 2\tilde{\boldsymbol{D}}_c^{1/2}\tilde{\boldsymbol{H}}\tilde{\boldsymbol{D}}_c^{1/2})\tilde{\boldsymbol{\Omega}}^{-1}$$

Therefore

$$\boldsymbol{\mathcal{H}}_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}}) + \boldsymbol{\mathcal{H}}_2(\boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}}, \nu) = -\frac{\nu - 1}{2}\tilde{\boldsymbol{\Omega}}^{-2}$$

The first derivative of $h_1(\cdot)$ can be written as

$$\boldsymbol{h}_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}}) = \frac{1}{2}\tilde{\boldsymbol{\Omega}}^{-1}(\tilde{\boldsymbol{H}}^* - \tilde{\boldsymbol{D}} + \tilde{\boldsymbol{D}}_c)\mathbf{1}_n$$

where

$$\tilde{\boldsymbol{D}} = \text{diag}\left\{\frac{d_i\tilde{\omega}_i}{\sigma_i^2}\right\}$$

and $\tilde{\boldsymbol{H}}^*$ is a diagonal matrix with ith element, $\tilde{h}_{ii}$. Using (6.1.7) the approximate conditional likelihood is

$$\begin{aligned} L_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y}) = {} & |\boldsymbol{I} + (\boldsymbol{V} - \tilde{\boldsymbol{H}}^2 - 2\tilde{\boldsymbol{D}}_c^{1/2}\tilde{\boldsymbol{H}}\tilde{\boldsymbol{D}}_c^{1/2})^{-1}(\tilde{\boldsymbol{H}}^2 + 2\tilde{\boldsymbol{D}}_c^{1/2}\tilde{\boldsymbol{H}}\tilde{\boldsymbol{D}}_c^{1/2})|^{-1/2} \\ & \times (2\pi)^{-p/2}|\boldsymbol{X}^T\tilde{\boldsymbol{\Psi}}^{-1}\boldsymbol{X}|^{1/2} \times \exp\Big\{-\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\tilde{\boldsymbol{\Psi}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \tfrac{1}{2}\boldsymbol{y}^T\tilde{\boldsymbol{S}}\boldsymbol{y}\Big\} \\ & \times \exp\Big\{-\tfrac{1}{4(\nu-1)}\Big((\tilde{\boldsymbol{H}}^* - \tilde{\boldsymbol{D}} + \tilde{\boldsymbol{D}}_c)\mathbf{1}_n\Big)^T\Big((\tilde{\boldsymbol{H}}^* - \tilde{\boldsymbol{D}} + \tilde{\boldsymbol{D}}_c)\mathbf{1}_n\Big)\Big\} \end{aligned} \tag{6.1.17}$$

where $\tilde{\boldsymbol{\Psi}}$ is the diagonal scale matrix with ith diagonal element $\sigma_i^2/\tilde{\omega}_i$ and $\tilde{\boldsymbol{S}}$ is the projection matrix defined by (6.1.4) evaluated at $\tilde{\boldsymbol{\omega}}$. Using (6.1.8) the approximate Restricted Maximum Likelihood or t-REML becomes

$$\begin{aligned} L_2(\boldsymbol{\lambda}, \nu; \boldsymbol{y}) = {} & |\tilde{\boldsymbol{\Psi}}|^{-1/2}|\boldsymbol{X}^T\tilde{\boldsymbol{\Psi}}^{-1}\boldsymbol{X}|^{-1/2}|\tilde{\boldsymbol{\Omega}}||\tfrac{1}{2}(\boldsymbol{V} - \tilde{\boldsymbol{H}}^2 - 2\tilde{\boldsymbol{D}}_c^{1/2}\tilde{\boldsymbol{H}}\tilde{\boldsymbol{D}}_c^{1/2})|^{-1/2} \\ & \times (2\pi)^{-(n-p)/2}\exp\Big\{-\tfrac{1}{2}\boldsymbol{y}^T\tilde{\boldsymbol{S}}\boldsymbol{y}\Big\}\frac{(\nu/2)^{n\nu/2}}{(\Gamma(\nu/2))^n}\exp\Big\{-\tfrac{\nu}{2}\sum_{i=1}^n\tilde{\omega}_i\Big\}\prod_{i=1}^n\tilde{\omega}_i^{\nu/2-1} \end{aligned} \tag{6.1.18}$$

6.1.3 Random scale effects

For the approximation to proceed the random scale effects, $\boldsymbol{\omega}$, are maximised by considering the first derivative of $h_2(\cdot)$ given by, (6.1.9),

$$\begin{aligned}\frac{\partial h_2(\boldsymbol{\lambda},\boldsymbol{\omega},\nu)}{\partial \omega_i} &= -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}_i)+\frac{1}{2}\mathrm{tr}(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}_i)+\Big(\frac{\nu}{2}-1\Big)\frac{1}{\omega_i}-\frac{\nu}{2}\\ &\qquad +\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_c)^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}_i\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_c)\\ &= \frac{1}{2\omega_i}-\frac{h_{ii}}{2\omega_i}-\frac{d_{i,c}}{2\sigma_i^2}+\frac{\nu-2}{2\omega_i}-\frac{\nu}{2}.\end{aligned}$$

Setting this to zero and evaluating at $\boldsymbol{\omega}=\tilde{\boldsymbol{\omega}}$ the ith maximised random effect can be written as an implicit equation given by

$$\tilde{\omega}_i=\frac{\nu-1-\tilde{h}_{ii}}{\nu+\tilde{d}_{i,c}/\sigma_i^2}. \tag{6.1.19}$$

Note that on the RHS of this equation $\tilde{h}_{ii}$ and $\tilde{d}_{i,c}$ are a function of the random effects $\tilde{\omega}_i, i=1,\ldots,n$.

This equation has several disadvantages. Firstly each of the terms given in the approximate REML, (6.1.18), are evaluated at $\omega_i=\tilde{\omega}_i,\ i=1,\ldots,n$ and therefore increases the complexity of estimating the remaining scale parameters, $\boldsymbol{\lambda}$, from the approximate REML given by (6.1.18). Secondly, for low degrees of freedom, $\nu<2$ it is possible that (6.1.19) may produce negative values. The approximate REML requires positivity of the estimated random scale effects due to the presence of the Gamma kernel. The next section discusses a change of scale for the random effects which ensures the maximised random scale effects will be less likely to have these problems.

For given $(\boldsymbol{\lambda},\nu)$ the mth iterate for the random scale effects can be written as

$$\tilde{\omega}_i^{(m+1)}=k(\boldsymbol{\omega}^{(m)},\boldsymbol{\lambda},\nu)=\frac{\nu-1-\tilde{h}_{ii}^{(m)}}{\nu+\tilde{d}_{i,c}^{(m)}/\sigma_i^2}. \tag{6.1.20}$$

6.1.4 Changing the scale of the random effects

In Section 5.1.2 it was shown that changing the scale of the random scale effects distribution in the integrand of (4.2.4) produced different estimates for the maximised random scale effects. In particular, under ML, transforming the random effects using the natural log, $\omega_i^*=\log\omega_i,\ i=1,\ldots,n$ produced estimates equivalent to the predicted values obtained from the conditional distribution of $\omega_i|y_i,\ i=1,\ldots,n$

Applying this same methodology the log Gamma likelihood (5.1.21) is substituted into (6.1.5) and the Partial Laplace approximation is then reapplied. Following identically to

Section 6.1 the derivatives of $h_1(\cdot)$ and $h_2(\cdot)$ are required. The derivatives of $h_2(\cdot)$ with respect to ω_i^* are given by

$$\begin{aligned}
\frac{\partial h_2(\cdot)}{\partial \omega_i^*} &= -\frac{1}{2}\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i) + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\boldsymbol{y} + \frac{\nu}{2} - \frac{\nu}{2}\exp\,\omega_i^* \qquad (6.1.21)\\
\frac{\partial^2 h_2(\cdot)}{(\partial \omega_i^*)^2} &= -\frac{1}{2}\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_{ii}) + \frac{1}{2}\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i) + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_{ii}\boldsymbol{S}\boldsymbol{y}\\
&\quad -\boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\boldsymbol{y} - \frac{\nu}{2}\exp\,\omega_i^* \qquad (6.1.22)
\end{aligned}$$

where, here, $\dot{\boldsymbol{\Psi}}_i = \partial\boldsymbol{\Psi}/\partial\omega_i^*$. The derivative of (6.1.21) with respect to ω_j^* is given by (6.1.11). The first and second derivative of $h_1(\cdot)$ are given by (6.1.12) and (6.1.13) respectively. Using (5.1.22) the terms of (6.1.22) can be reduced to

$$\begin{aligned}
\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_{ii}) &= \text{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ii}) - \text{tr}(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ii})\\
&= 1 - h_{ii}\\
\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i) &= \text{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i) - 2\text{tr}(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i) + \text{tr}(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i)\\
&= 1 - 2h_{ii} + h_{ii}^2\\
\boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_{ii}\boldsymbol{S}\boldsymbol{y} &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_c)^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_{ii}\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_c)\\
&= \frac{d_{i,c}\exp\,\omega_i^*}{\sigma_i^2}\\
\boldsymbol{y}^T\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\boldsymbol{y} &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\\
&- (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1/2}\boldsymbol{H}\boldsymbol{\Psi}^{-1/2}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\\
&= \frac{d_{i,c}\exp\,\omega_i^*}{\sigma_i^2} - \frac{d_{i,c}\exp\,\omega_i^* h_{ii}}{\sigma_i^2}
\end{aligned}$$

The terms of (6.1.11) can also be reduced

$$\begin{aligned}
\text{tr}(\boldsymbol{S}\dot{\boldsymbol{\Psi}}_i\boldsymbol{S}\dot{\boldsymbol{\Psi}}_j) &= \text{tr}(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{H}\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_j)\\
&= h_{ij}^2\\
\boldsymbol{y}^T\boldsymbol{S}\boldsymbol{\Psi}_i\boldsymbol{S}\boldsymbol{\Psi}_j\boldsymbol{S}\boldsymbol{y} &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\\
&- (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_c)^T\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1/2}\boldsymbol{H}\boldsymbol{\Psi}^{-1/2}\dot{\boldsymbol{\Psi}}_j\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_c)\\
&= -\frac{r_{i,c}(\exp\,\omega_i^*)^{1/2}}{\sigma_i}h_{ij}\frac{r_{j,c}(\exp\,\omega_j^*)^{1/2}}{\sigma_i}
\end{aligned}$$

Combining all terms and after some algebra the second derivative of $h_2(\cdot)$ evaluated at $\tilde{\omega}_i = \exp\,\tilde{\omega}_i^*$ is given by

$$\boldsymbol{\mathcal{H}}_2(\boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}}, \nu) = -\frac{1}{2}(\boldsymbol{V}^* - \tilde{\boldsymbol{H}}^2 - 2\tilde{\boldsymbol{D}}_c^{1/2}\tilde{\boldsymbol{H}}\tilde{\boldsymbol{D}}_c^{1/2}) \qquad (6.1.23)$$

where $\boldsymbol{V}^* = (\nu + 1)\boldsymbol{I}$. Using (5.1.23) the second derivative of $h_1(\cdot)$ evaluated at $\tilde{\omega}_i = \exp\,\tilde{\omega}_i^*$ can be immediately written as

$$\boldsymbol{\mathcal{H}}_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}}) = -\frac{1}{2}(\tilde{\boldsymbol{H}}^2 + \tilde{\boldsymbol{D}} + 2\tilde{\boldsymbol{D}}_c^{1/2}\tilde{\boldsymbol{H}}\tilde{\boldsymbol{D}}_c^{1/2})$$

and therefore

$$\boldsymbol{\mathcal{H}}_1(\boldsymbol{\beta},\boldsymbol{\lambda},\tilde{\boldsymbol{\omega}})+\boldsymbol{\mathcal{H}}_2(\boldsymbol{\lambda},\tilde{\boldsymbol{\omega}},\nu)=-\frac{\nu}{2}\tilde{\boldsymbol{\Omega}}-\frac{1}{2}\tilde{\boldsymbol{D}}$$

The first derivative of $h_1(\cdot)$ can also be immediately written as

$$\boldsymbol{h}_1(\boldsymbol{\beta},\boldsymbol{\lambda},\tilde{\boldsymbol{\omega}})=\frac{1}{2}(\tilde{\boldsymbol{H}}^*-\tilde{\boldsymbol{D}}+\tilde{\boldsymbol{D}}_c)\mathbf{1}_n$$

Using (6.1.17) the approximate conditional likelihood for the location parameters is

$$\begin{aligned}L_1(\boldsymbol{\beta},\boldsymbol{\lambda},\nu;\boldsymbol{y})&=|\boldsymbol{I}+(\boldsymbol{V}^*-\tilde{\boldsymbol{H}}^2-2\tilde{\boldsymbol{D}}_c^{1/2}\tilde{\boldsymbol{H}}\tilde{\boldsymbol{D}}_c^{1/2})^{-1}(\tilde{\boldsymbol{H}}^2+2\tilde{\boldsymbol{D}}_c^{1/2}\tilde{\boldsymbol{H}}\tilde{\boldsymbol{D}}_c^{1/2})|^{-1/2}\\&\times(2\pi)^{-p/2}|\boldsymbol{X}^T\tilde{\boldsymbol{\Psi}}^{-1}\boldsymbol{X}|^{1/2}\exp\Big\{-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\tilde{\boldsymbol{\Psi}}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})+\tfrac{1}{2}\boldsymbol{y}^T\tilde{\boldsymbol{S}}\boldsymbol{y}\Big\}\\&\times\exp\Big\{-\tfrac{1}{4}\Big((\tilde{\boldsymbol{H}}^*-\tilde{\boldsymbol{D}}+\tilde{\boldsymbol{D}}_c)\mathbf{1}_n\Big)^T(\boldsymbol{V}^*-\tilde{\boldsymbol{H}}^*)\Big((\tilde{\boldsymbol{H}}^*-\tilde{\boldsymbol{D}}+\tilde{\boldsymbol{D}}_c)\mathbf{1}_n\Big)\Big\}\end{aligned}\tag{6.1.24}$$

Using (6.1.18) the approximate REML for the remaining scale parameters is

$$\begin{aligned}L_2(\boldsymbol{\lambda},\nu;\boldsymbol{y})&=|\tilde{\boldsymbol{\Psi}}|^{-1/2}|\boldsymbol{X}^T\tilde{\boldsymbol{\Psi}}^{-1}\boldsymbol{X}|^{-1/2}|\tfrac{1}{2}(\boldsymbol{V}^*-\tilde{\boldsymbol{H}}^2-2\tilde{\boldsymbol{D}}_c^{1/2}\tilde{\boldsymbol{H}}\tilde{\boldsymbol{D}}_c^{1/2})|^{-1/2}\\&\times(2\pi)^{-(n-p)/2}\exp\Big\{-\tfrac{1}{2}\boldsymbol{y}^T\tilde{\boldsymbol{S}}\boldsymbol{y}\Big\}\frac{(\nu/2)^{n\nu/2}}{(\Gamma(\nu/2))^n}\exp\Big\{-\tfrac{\nu}{2}\sum_{i=1}^n\tilde{\omega}_i\Big\}\prod_{i=1}^n\tilde{\omega}_i^{\nu/2}\end{aligned}\tag{6.1.25}$$

The random scale effects are maximised by setting this first derivative of $h_2(\cdot)$ to zero and solving for $\omega_i=\exp\omega_i^*$, namely

$$\tilde{\omega}_i=\frac{\nu+1-\tilde{h}_{ii}}{\nu+\tilde{d}_{i,c}/\sigma_i^2}$$

Note this estimate is not equivalent to the estimates given by (6.1.19). The location parameters, $\boldsymbol{\beta}$ are not present in the maximised random scale effects but as the last exponent term of (6.1.24) is inherently different than the last exponent term of (6.1.17) the approximate conditional likelihood is not invariant for the location parameters if the scale of the random effects is altered. Similarly, as the scale parameters are present in the estimated random scale effects and the extra determinant term in (6.1.25) is inherently different from the extra determinant term in (6.1.18) the approximate REML is not invariant for the scale parameters when the scale of the random effects is changed.

For examples and simulations in this thesis, (6.1.17) and (6.1.18) are used as objective functions to estimate the location and scale parameters respectively.

6.1.5 Estimating the Location Parameter

The approximate t-REML given by (6.1.18) is free of the location parameter $\boldsymbol{\beta}$. Therefore an objective function to estimate $\boldsymbol{\beta}$ is given by the approximate conditional likelihood in

(6.1.17). As the estimated random effects, (6.1.19) are free of $\boldsymbol{\beta}$ the initial determinant terms of (6.1.17) containing the second derivatives $\boldsymbol{\mathcal{H}}_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \tilde{\omega})$ and $\boldsymbol{\mathcal{H}}_2(\boldsymbol{\lambda}, \tilde{\boldsymbol{\omega}}, \nu)$ are also free of the location parameters. Therefore only the multiplicative exponential terms of (6.1.17) contain $\boldsymbol{\beta}$. Omitting constants this approximate conditional log-likelihood used to estimate the location parameters can be reduced to

$$\begin{aligned} \ell_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y}) &= -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \tilde{\boldsymbol{\Psi}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &- \frac{1}{4(\nu-1)}\{(\tilde{\boldsymbol{H}}^* - \tilde{\boldsymbol{D}} + \tilde{\boldsymbol{D}}_c)\mathbf{1}_n\}^T\{(\tilde{\boldsymbol{H}}^* - \tilde{\boldsymbol{D}} + \tilde{\boldsymbol{D}}_c)\mathbf{1}_n\} \end{aligned} \quad (6.1.26)$$

The second quadratic form in the exponential can be written as

$$\sum_{i=1}^{n}\left\{h_{ii} - \frac{d_i\tilde{\omega}_i}{\sigma_i^2} + \frac{\tilde{d}_{i,c}\tilde{\omega}_i}{\sigma_i^2}\right\}^2 \quad (6.1.27)$$

Noting that (6.1.19) can be rearranged to be

$$\frac{\tilde{d}_{i,c}\tilde{\omega}_i}{\sigma_i^2} + h_{ii} = \nu - 1 - \nu\tilde{\omega}_i$$

Subtituting this into (6.1.27) gives

$$\begin{aligned} &\sum_{i=1}^{n}\left\{(\nu-1) - \nu\tilde{\omega}_i - \frac{d_i\tilde{\omega}_i}{\sigma_i^2}\right\}^2 \\ = &\sum_{i=1}^{n}\left\{\big((\nu-1) - \nu\tilde{\omega}_i\big)^2 - 2\big((\nu-1) - \nu\tilde{\omega}_i\big)\frac{d_i\tilde{\omega}_i}{\sigma_i^2} + \left(\frac{d_i\tilde{\omega}_i}{\sigma_i^2}\right)^2\right\} \end{aligned}$$

To obtain the score equation for the location parameters the approximate conditional log-likelihood (6.1.26) is differentiated with respect to $\boldsymbol{\beta}$ giving

$$\begin{aligned} \frac{\partial \ell_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu; \boldsymbol{y})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}\frac{\tilde{\omega}_i}{\sigma_i^2}\left\{1 - \frac{1}{\nu-1}\left((\nu-1) - \nu\tilde{\omega}_i - \frac{d_i\tilde{\omega}_i}{\sigma_i^2}\right)\right\}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i \\ &= \frac{1}{\nu-1}\sum_{i=1}^{n}\frac{\tilde{\omega}_i}{\sigma_i^2}\left\{\tilde{\omega}_i\left(\nu + \frac{d_i}{\sigma_i^2}\right)\right\}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i \\ &= \frac{\nu+1}{\nu-1}\sum_{i=1}^{n}\frac{\tilde{\omega}_i^2}{\sigma_i^2\bar{\omega}_i}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i \end{aligned}$$

where $\bar{\omega}_i$ is the ith estimated random scale effect (4.4.3) obtained from ML. Setting this score to zero and solving implicitly for $\boldsymbol{\beta}$ an algorithm for the REML estimation of the location parameters can be derived. The $(m+1)$th iterate is

$$\hat{\boldsymbol{\beta}}_{m+1} = f_t^*(\boldsymbol{\beta}_m, \boldsymbol{\lambda}, \nu) = (\boldsymbol{X}^T\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{W}}_m\boldsymbol{X})^{-1}\boldsymbol{X}^T\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{W}}_m\boldsymbol{y} \quad (6.1.28)$$

where $\tilde{\boldsymbol{W}}$ is a diagonal matrix with ith element $(\nu+1)\tilde{\omega}_i/(\nu-1)\bar{\omega}_i$. Given the scale parameters, $(\boldsymbol{\lambda}, \nu)$, (6.1.28) is an iteratively reweighted least squares procedure with which to obtain an approximate REML estimate for the location parameters, $\boldsymbol{\beta}$.

6.1.6 Estimating the Scale Parameters

The approximate conditional likelihood (6.1.17) is used as an objective function to estimate the location parameters and therefore approximately contains no information about the scale parameters $\boldsymbol{\lambda}$. Therefore the second component of the approximate marginal likelihood given by (6.1.18) is used as an objective function to estimate the scale parameters. This approximate Restricted Maximum log-likelihood can be simplified further. Under a simplified scale parameter model (3.1.3) and omitting constants the log-likelihood can be reduced to

$$\begin{aligned}\ell_2(\boldsymbol{\lambda},\nu;\boldsymbol{y}) &= -\frac{1}{2}\log|\tilde{\boldsymbol{Q}}| - \frac{1}{2}\log|\boldsymbol{X}^T\tilde{\boldsymbol{\Psi}}^{-1}\boldsymbol{X}| \\ &- \frac{1}{2}\sum_{i=1}^{n}\left\{\log\sigma_i^2 - \frac{\tilde{d}_{i,c}\tilde{\omega}_i}{\sigma_i^2}\right\} + \sum_{i=1}^{n}\left\{\left(\frac{\nu+1}{2}\right)\log\tilde{\omega}_i - \frac{\nu}{2}\tilde{\omega}_i\right\} \quad (6.1.29)\end{aligned}$$

where

$$\tilde{\boldsymbol{Q}} = \tfrac{1}{2}(\boldsymbol{V} - \tilde{\boldsymbol{H}}^2 - 2\tilde{\boldsymbol{D}}_c^{1/2}\tilde{\boldsymbol{H}}\tilde{\boldsymbol{D}}_c^{1/2}) \quad (6.1.30)$$

The estimated random effects, $\tilde{\omega}_i,\ 1,\ldots,n$ are a complex function of the remaining scale parameters. Furthermore, $\tilde{\boldsymbol{Q}}$ is also complex function of the remaining parameters and the predicted random effects. For this reason the maximisation of (6.1.29) is handled numerically.

The numerical maximisation requires the calculation of the determinant of $\tilde{\boldsymbol{Q}}$. This matrix is a function of the observed responses and therefore may be semi-negative definite. To circumvent this problem and ensure positive definiteness, the observed components contained in $\tilde{\boldsymbol{Q}}$ are eliminated by considering its expected value. Let $\bar{\omega} = \tilde{\omega}$ be fixed, then the conditional expectation

$$\mathrm{E}\left[\tfrac{1}{2}(\boldsymbol{V} - \boldsymbol{H}^2 - 2\boldsymbol{D}_c^{1/2}\boldsymbol{H}\boldsymbol{D}_c^{1/2})|\bar{\omega}\right]$$

The expected value of this term is found by taking the expectations of the last term only and for its ijth element becomes

$$\mathrm{E}\left[r_{i,c}r_{j,c}\right]\frac{\bar{\omega}_i^{1/2}h_{ij}\bar{\omega}_j^{1/2}}{\sigma_i\sigma_j} = \begin{cases}(1-h_{ii})h_{ii} & i=j\\ -h_{ij}^2 & i\neq j\end{cases}$$

Substituting this back into $\boldsymbol{Q}$ the approximation becomes

$$\tilde{\boldsymbol{Q}}^* = \tfrac{1}{2}(\boldsymbol{V} + \tilde{\boldsymbol{H}}^2 - 2\tilde{\boldsymbol{H}}^*) \quad (6.1.31)$$

This is then substituted into (6.1.29) as the replacement for $\tilde{\boldsymbol{Q}}$.

6.1.7 Asymptotics

The asymptotic properties of the approximate Restricted Maximum likelihood (6.1.29) can be checked. Of particular interest is the form of the marginal likelihood for the scale parameters, $\boldsymbol{\lambda}$, as $\nu \to \infty$. Noting that $\tilde{\boldsymbol{Q}}$ can be expressed as

$$\tilde{\boldsymbol{Q}} = \frac{\nu}{2}\big\{(1-1/\nu)\boldsymbol{I} - \tilde{\boldsymbol{H}}^2/\nu - 2\tilde{\boldsymbol{D}}_r\tilde{\boldsymbol{H}}\tilde{\boldsymbol{D}}_r/\nu\big\}$$

and, similar to the ML case, the estimated random effects can be written as

$$\tilde{\omega}_i^* = \frac{1-(1-\tilde{h}_{ii})/\nu}{1+\tilde{d}_{i,c}/\sigma_i^2\nu}$$

Therefore as $\nu \to \infty$, $\tilde{\omega}_i \to 1$, $\tilde{\boldsymbol{Q}} \to \boldsymbol{I}$ and, similarly, $\tilde{\boldsymbol{Q}}^* \to \boldsymbol{I}$. Omitting constants, the marginal likelihood can be expressed as

$$\ell_2(\boldsymbol{\lambda}, \nu; \boldsymbol{y}) \;\; = \;\; -\frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}| - \frac{1}{2}\sum_{i=1}^{n}\Big\{\log\sigma_i^2 + \frac{d_{i,c}}{\sigma_i^2}\Big\}$$

This is equivalent to the Gaussian REML under a simplified scale parameter model defined by (3.1.3). Therefore in the limit ($\nu \to \infty$) the heteroscedastic Gaussian REML is nested in the approximate heteroscedastic t-REML approach defined here.

Similarly, as $\nu \to \infty$ then $\tilde{\boldsymbol{W}} \to \boldsymbol{I}$ and therefore the approximate REML estimator given by (6.1.19) for the location fixed effects also tends to the ordinary REML or ML estimator for the Gaussian case.

6.1.8 Computations

The Partial Laplace approximation to the marginal likelihood of the heteroscedastic t-distribution gives two approximately disjoint likelihoods given by (6.1.17) and (6.1.18). The subsequent estimation of the parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ can therefore be viewed as a sequential computational algorithm.

- **Estimation of $\boldsymbol{\lambda}$:** The approximate t-REML estimates for $\boldsymbol{\lambda}$ are obtained iteratively. At the mth iteration
  - For given $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(m)}$ and ν update $\omega_i,\ 1, \ldots, n$ using $\omega_i^{(m+1)} = k(\boldsymbol{\omega}^{(m)}, \boldsymbol{\lambda}^{(m)}, \nu)$, where $k(\cdot)$ is given in (6.1.20).
  - For given $\omega_i^{(m+1)},\ i = 1, \ldots, n$ update $\boldsymbol{\lambda}$ using approximate t-REML defined by (6.1.18), namely

$$\boldsymbol{\lambda}^{(m+1)} = \max\big\{\boldsymbol{\lambda}; L_2(\boldsymbol{\lambda}^{(m)}, \nu; \boldsymbol{y})\big\}$$

- **Estimation of $\boldsymbol{\beta}$:** The approximate t-REML estimate for $\boldsymbol{\beta}$ can be obtained iteratively. At the mth iteration

  - For given $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$ and $\omega_i = \omega_i(\hat{\boldsymbol{\lambda}}, \nu),\ 1, \ldots, n$, the approximate t-REML estimates for the scale parameters, $\boldsymbol{\lambda}$ obtained from Step 1 of the algorithm, update $\boldsymbol{\beta}$ using $\boldsymbol{\beta} = f_t^*(\boldsymbol{\beta}^{(m)}, \hat{\boldsymbol{\lambda}}, \nu)$, where $f_t^*(\cdot)$ is given in (6.1.28).

As the approximate REML for the scale parameters is free of the location parameters there is no requirement for the recalculation of $\boldsymbol{\beta}$ in Step 1 of the algorithm. This ensures that Step 1 may be reiterated until the approximate REML estimates, $\hat{\boldsymbol{\lambda}}$ are obtained. These estimates are then the logical choice to be substituted into the iterative procedure to obtain an approximate REML estimate for the location parameters.

For the purpose of brevity for proceeding chapters of this thesis, this algorithm and the estimators obtained from it will be defined as t-REML I with known degrees of freedom.

6.2 Heteroscedastic t-REML using Modified Profile Likelihood

In this section the MPL techniques from Section 5.2.1 are used as a basis for deriving an approximate REML for the heteroscedastic t-distribution with known degrees of freedom.

Consider the model defined by (3.1.1) where the distribution of the response is defined by (4.2.7). For this particular section let the scale parameter model be defined by (3.1.2).

For known degrees of freedom the heteroscedastic t-distribution only requires the estimation of the location and scale parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda})$. Therefore a profile likelihood for the scale parameters, $\sigma^2(\boldsymbol{z}_i; \boldsymbol{\lambda}),\ i = 1, \ldots, n$, given some maximal estimate of $\boldsymbol{\beta}$,

$$L_p(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}; \boldsymbol{y}) = \left\{ \frac{\Gamma((\nu+1)/2)}{(\Gamma(1/2))\Gamma(\nu/2)\nu^{1/2}} \right\}^n |\boldsymbol{\Sigma}|^{-1/2} \prod_{i=n}^{n} \left\{ 1 + \frac{r_i^2}{\sigma_i^2 \nu} \right\}^{-(\frac{\nu+1}{2})} \tag{6.2.1}$$

where $r_i = (y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$ obtained by maximising (4.2.9) for given $\boldsymbol{\lambda}$ using equation (4.3.16) from Section 4.3.2. This profile likelihood is not adjusted for the estimation of the location parameters.

6.2.1 Modifying the Profile Likelihood

Section 5.2.1 suggests that to adjust (6.2.1) a modified profile likelihood approach is required. Using (5.2.5) an MPL function can be immediately written as

$$L_{MP}(\boldsymbol{\lambda};\boldsymbol{y}) = \Big|\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}}\Big||\mathcal{I}_o(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}},\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})|^{-1/2}\exp(\ell(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}},\boldsymbol{\lambda};\boldsymbol{y})) \tag{6.2.2}$$

where $\ell(\cdot)$ is the the natural logarithm of (6.2.1) and $\mathcal{I}_o(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}},\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})$ is the observed information for $\boldsymbol{\beta}$ defined by (4.3.9) evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$. The first determinant term on the RHS is the Jacobian due to the transformation of $\hat{\boldsymbol{\beta}} \rightarrow \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$. Using (5.2.6) from the same section this can also be expressed as

$$L_{MP}(\boldsymbol{\lambda};\boldsymbol{y}) = \frac{|\mathcal{I}_o(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}},\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})|^{1/2}}{|\ell_{\boldsymbol{\beta};\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}},\boldsymbol{\lambda};\boldsymbol{y})|}\exp(\ell(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}},\boldsymbol{\lambda};\boldsymbol{y})) \tag{6.2.3}$$

where $\ell_{\boldsymbol{\beta};\hat{\boldsymbol{\beta}}}(\cdot)$ is the first derivative of the log-likelihood function (4.2.9) with respect to $\boldsymbol{\beta}$ and its second derivative with respect to the maximum likelihood estimator of the location parameters, $\hat{\boldsymbol{\beta}}$.

As the heteroscedastic t is a member of the location-scale family when the degrees of freedom is known the specification of an ancillary statistic is possible. Consider the denominator of the first term of (6.2.3) and let

$$d_i = r_i^2 + 2r_i\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) + (\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))^2$$

where $r_i = (y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$. Then for known degrees of freedom consider kernel of the log-likelihood for $\boldsymbol{\beta}$ derived from (4.2.9) is

$$-\Big(\frac{\nu+1}{2}\Big)\sum_{i=1}^{n}\log\Big\{1+\frac{(\hat{\sigma}_i^2)^2}{\sigma_i^2\nu}\Big(a_i^2+\frac{2a_i\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\hat{\sigma}_i^2}+\frac{(\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))^2}{(\hat{\sigma}_i^2)^2}\Big)\Big\}$$

where $\boldsymbol{a}$ is a vector of ancillary statistics with ith element $a_i = r_i/\hat{\sigma}_i^2$ and $\hat{\sigma}_i^2 = \sigma^2(\boldsymbol{z}_i;\hat{\boldsymbol{\lambda}})$. Taking the first derivative of this function with respect to $\boldsymbol{\beta}$ gives

$$\frac{\partial\ell(\boldsymbol{\beta},\boldsymbol{\lambda};\boldsymbol{y})}{\partial\boldsymbol{\beta}} = \Big(\frac{\nu+1}{2}\Big)\sum_{i=1}^{n}\frac{\boldsymbol{x}_i(\hat{\sigma}_i^2)^2}{\sigma_i^2\nu}\Big(\frac{1}{1+d_i/\sigma_i^2\nu}\Big)\Big\{\frac{2a_i}{\hat{\sigma}_i^2}+\frac{2\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(\hat{\sigma}_i^2)^2}\Big\}$$

Taking the second derivative of this function with respect to $\hat{\boldsymbol{\beta}}$ gives

$$\begin{aligned}\ell_{\boldsymbol{\beta};\hat{\boldsymbol{\beta}}}(\boldsymbol{\beta},\boldsymbol{\lambda};\boldsymbol{y}) &= \frac{\partial^2\ell(\boldsymbol{\beta},\boldsymbol{\lambda};\boldsymbol{y})}{\partial\boldsymbol{\beta}\partial\hat{\boldsymbol{\beta}}} = \Big(\frac{\nu+1}{2}\Big)\sum_{i=1}^{n}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T(\hat{\sigma}_i^2)^2}{\sigma_i^2\nu}\Big(\frac{1}{1+d_i/\sigma_i^2\nu}\Big)\Big\{\frac{2}{(\hat{\sigma}_i^2)^2}\Big\}\\ &- \Big(\frac{\nu+1}{2}\Big)\sum_{i=1}^{n}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T(\hat{\sigma}_i^2)^4}{(\sigma_i^2\nu)^2}\Big(\frac{1}{1+d_i/\sigma_i^2\nu}\Big)^2\Big\{\frac{2a_i}{\hat{\sigma}_i^2}+\frac{2\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(\hat{\sigma}_i^2)^2}\Big\}^2\end{aligned}$$

Noting that

$$\begin{aligned}\left\{\frac{a_i}{\hat{\sigma}_i^2}+\frac{\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(\hat{\sigma}_i^2)^2}\right\}^2 &= \frac{a_i^2}{(\hat{\sigma}_i^2)^2}+\frac{2a_i\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(\hat{\sigma}_i^2)^3}+\frac{(\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))^2}{(\hat{\sigma}_i^2)^4}\\ &= \frac{r_i^2}{(\hat{\sigma}_i^2)^4}+\frac{2r_i\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(\hat{\sigma}_i^2)^4}+\frac{(\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))^2}{(\hat{\sigma}_i^2)^4}\\ &= \frac{d_i}{(\hat{\sigma}_i^2)^4}\end{aligned}$$

and using (4.3.6) this may be reduced to

$$\begin{aligned}\ell_{\boldsymbol{\beta};\hat{\boldsymbol{\beta}}}(\boldsymbol{\beta},\boldsymbol{\lambda};\boldsymbol{y}) &= (\nu+1)\sum_{i=1}^n\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\sigma_i^2\nu}\left\{\left(\frac{1}{1+d_i/\sigma_i^2\nu}\right)-\left(\frac{d_i/\sigma_i^2\nu}{1+d_i/\sigma_i^2\nu}\right)\left(\frac{2}{1+d_i/\sigma_i^2\nu}\right)\right\}\\ &= (\nu+1)\sum_{i=1}^n\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\sigma_i^2\nu}\left\{(1-B_i)-2(1-B_i)B_i\right\}\end{aligned}$$

Using (4.3.9) the adjustment terms of (6.2.3) are related by $\ell_{\boldsymbol{\beta};\hat{\boldsymbol{\beta}}}(\boldsymbol{\beta},\boldsymbol{\lambda};\boldsymbol{y}) = \mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\beta})$. Then evaluating at the maximum likelihood estimate for the location parameters, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ the t-REML MPL for $\sigma^2(\boldsymbol{z}_i;\boldsymbol{\lambda}),\ i=1,\ldots,n$ can be reduced to

$$L_{MP}^o(\boldsymbol{\lambda};\boldsymbol{y}) = |\mathcal{I}_o(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}},\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})|^{-1/2}\exp(\ell(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}},\boldsymbol{\lambda};\boldsymbol{y})) \tag{6.2.4}$$

Using the results from section 5.2.3, and noting that $(\boldsymbol{\beta},\boldsymbol{\lambda})$ are orthogonal, the leading term of (6.2.2) is unity and an equivalent result is obtained. Thus, in this particular case, $L_{MP}(\boldsymbol{\lambda};\boldsymbol{y}) = L_{CP}(\boldsymbol{\lambda};\boldsymbol{y})$ to the accuracy $\mathcal{O}(n^{-3/2})$. Barndorff-Nielsen (1983) and Barndorff-Nielsen & Cox (1994) recognise these results as a special case due to its connection to the location-scale family.

The observed information $\mathcal{I}_o(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}},\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})$ in the determinant term of (6.2.4) contains the observation vector and therefore may be semi negative-definite. This will produce negative eigenvalues and the determinant will not be calculable. To ensure positive definiteness an approximation to the observed information is used. The observed information for $\boldsymbol{\beta}$ can also be expressed as

$$\mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\beta}) = \sum_{i=1}^n\frac{\bar{\omega}_i}{\sigma_i^2}\boldsymbol{x}_i\boldsymbol{x}_i^T\left\{1-\left(\frac{1}{\nu+1}\right)\frac{2d_i\bar{\omega}_i}{\sigma_i^2}\right\}$$

If $\bar{\omega}_i$ is assumed to be fixed then noting, $\mathrm{E}[d_i|\bar{\omega}_i] = \sigma_i^2/\bar{\omega}_i$ and taking expectations of this matrix given $\bar{\omega}_i$ is

$$\mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\beta}) = \sum_{i=1}\frac{\bar{\omega}_i}{\sigma_i^2}\boldsymbol{x}_i\boldsymbol{x}_i^T\left\{\frac{\nu-1}{\nu+1}\right\}$$

This allows the modified profile likelihood for the scale parameters of the heteroscedastic t-distribution for known degrees of freedom to be written as

$$L^{o}_{MP}(\boldsymbol{\lambda};\boldsymbol{y}) = \Big(\frac{\nu-1}{\nu+1}\Big)^{-p/2}|\boldsymbol{X}^T\bar{\boldsymbol{\Psi}}^{-1}\boldsymbol{X}|^{-1/2}|\boldsymbol{\Sigma}|^{-1/2}\prod_{i=n}^{n}\left\{1+\frac{r_i^2}{\sigma_i^2\nu}\right\}^{-(\frac{\nu+1}{2})} \tag{6.2.5}$$

where $\bar{\boldsymbol{\Psi}}$ is a diagonal matrix with ith diagonal element $\sigma_i^2/\bar{\omega}_i$. As the degrees of freedom is known apriori the first term of the RHS is also a constant.

Barndorff-Nielsen (1983) also suggests that the expected information can be used a replacement for the observed information. Using (4.3.13) and omitting constants the MPL can be immediately written as

$$L^{e}_{MP}(\boldsymbol{\lambda};\boldsymbol{y}) = \Big(\frac{\nu+1}{\nu+3}\Big)^{-p/2}|\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}|^{-1/2}|\boldsymbol{\Sigma}|^{-1/2}\prod_{i=n}^{n}\left\{1+\frac{r_i^2}{\sigma_i^2\nu}\right\}^{-(\frac{\nu+1}{2})} \tag{6.2.6}$$

Again the first term of RHS is a constant and therefore this version of the modified profile likelihood only contains an adjustment term identical to the term required for Gaussian REML.

The asymptotic nature of (6.2.5) and (6.2.6) can be checked. As $\nu \to \infty$, $\bar{\omega}_i \to 1$, $\bar{\Psi} \to \boldsymbol{\Sigma}$ and the leading constant terms become unity. Therefore due to Property (7), (6.2.5) and (6.2.6) approach the heteroscedastic Gaussian REML. Thus, in the limit ($\nu \to \infty$), the simpler heteroscedastic Gaussian model is nested in the approximate heteroscedastic t-REML approaches defined here. For the examples and simulations in this thesis (6.2.5) is used.

6.2.2 Computations

When the degrees of freedom is fixed the estimation of the location and scale parameters can be achieved iteratively using the following algorithm.

- **Estimation of $(\boldsymbol{\beta},\boldsymbol{\lambda})$:** For the mth iteration the parameters are updated using
  - For given $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(m)}$ and ν update $\boldsymbol{\beta}$ using $\boldsymbol{\beta}^{(m+1)} = f_t(\boldsymbol{\beta}^{(m)},\boldsymbol{\lambda}^{(m)},\nu)$, where $f_t(\cdot)$ is given in (4.3.16).
  - For given $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m+1)}$ and ν update $\boldsymbol{\lambda}$ using t-REML defined by (6.2.5), namely,

$$\boldsymbol{\lambda}^{(m+1)} = \max\{\boldsymbol{\lambda}; L(\boldsymbol{\beta}^{(m+1)}(\boldsymbol{\lambda}^{(m)},\nu),\boldsymbol{\lambda}^{(m)};\boldsymbol{y})\}$$

For the purpose of brevity for proceeding chapters of this thesis this algorithm and the estimators obtained from it will be defined as t-REML II with known degrees of freedom.

Chapter 7

Examples and Simulations

To illustrate the ML techniques from Chapter 4 and the approximate REML techniques from Chapter 6 the Cherry Tree data set is considered in the next section. Complex scale parameter models similar to models considered in Aitkin (1987) and Verbyla (1993) are investigated under ML, t-REML I and t-REML II with homogeneity tests where possible.

To understand the properties of the estimators for the heteroscedastic t for known degrees of freedom under ML, and the two approximate t-REML techniques, a comparative simulation study is conducted in Section 7.2.

7.1 Examples

7.1.1 Cherry Trees

The cherry tree data was introduced in Section 1.2.1 as a potential candidate for robust modelling of the scale parameter. After fitting an additive model in the two explanatory variables, Diameter and Height, for the location, Cook & Weisberg (1983) proposed a simple scale parameter model. Aitkin (1987) and Verbyla (1993) model the data more extensively proposing a quadratic model in diameter for the Gaussian scale parameter. The latter author investigates various scale parameter models, estimating the location and scale parameters using Gaussian ML and REML.

The outliers in Figure 1.2 suggest the possible use of the heteroscedastic t to model the location and scale parameters. For all models proposed here the scale parameter is defined by (3.1.3) and, for this particular example, the degrees of freedom parameter is fixed at $\nu = 3$. This ensures that the variance for the t is defined and a highly robust fit is achieved. To explore the heteroscedasticity graphically, an initial homoscedastic t regression is fitted and (4.6.2) from Section 4.6 is used. Figure 7.1 shows the added variable plots (see

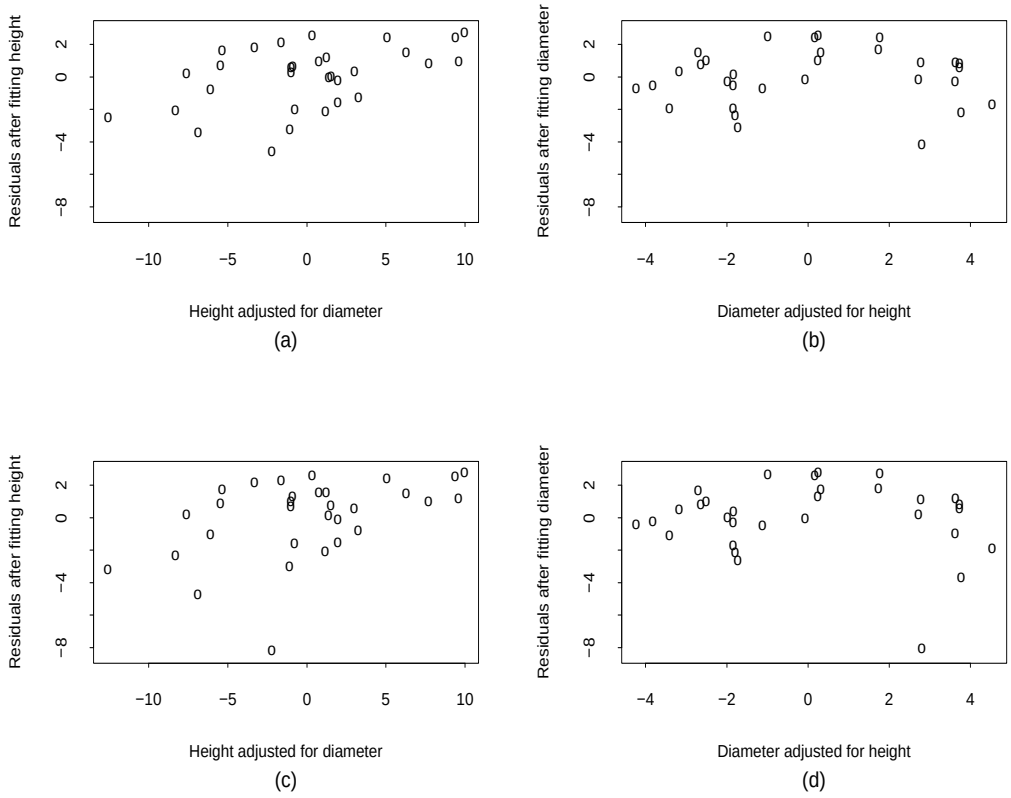

Figure 7.1: Added variable plots for the Cherry Tree data under the Gaussian and t-distribution (a) added variable plot for diameter under Gaussian; (b) added variable plot for height under Gaussian; (c) added variable plot for diameter under t; (d) added variable plot for height under t

Cook & Weisberg, 1982) for the adjusted residuals from the left hand side (4.6.2), that is $\log\{\bar{d}_i/(1-h_{ii})^2\} - \log\frac{1}{2} + \psi(\frac{1}{2}) - 1.4675$ against Diameter and Height. The added variable plots for the Gaussian model, given in Verbyla (1993), are also displayed for comparison. The Gaussian and t added variable plots show similar patterns for each of the explanatory variables suggesting that either may be used as a tool to determine possible heteroscedasticity. In comparison to the adjusted residuals under the Gaussian, the adjusted residuals for the added variable plots of the t are larger due to the longer tails of the t_3 distribution. For both models, a positive linear component for the Height is evident as well as a possible quadratic trend in Diameter. An initial additive scale parameter model with these components is adopted here.

Table 7.1 presents the various scale parameter models considered for analyses along with their associated score statistics and log-likelihood values under ML. The Gaussian and t models with fixed degrees of freedom have the same number of parameters and therefore precludes hypothesis testing between alternate models. However, in comparison to the log-likelihood values for scale parameter models under the t specification, the log-

	Homogeneity Test			
	Score		*2LogL*	
Model	Gaussian	t-ML	Gaussian	t-ML
1	-	-	69.62	65.48
H	3.24	5.09	74.73	72.24
D	0.47	1.43	70.52	67.26
H, D	3.32	5.09	74.80	72.31
D, D^2	3.70	5.60	85.49	82.05
H, D, D^2	6.14	8.73	87.63	82.79
H, D, D^2, HD	8.26	10.71	88.00	84.61

Table 7.1: Homogeneity Test for the Cherry Tree data under ML

likelihood values are larger for all Gaussian scale parameter models. This suggests that the simpler Gaussian model is preferred over the t specification for all heteroscedastic models presented here. Verbyla (1993) and 7.1 suggests that the Gaussian model with the squared diameter is sufficient for the scale parameter. The heteroscedastic t model proposed here is identical with slight changes in the numerical estimates. For this final model, in comparison to the scale parameter model under the t-specification, the larger log-likelihood value of the ML heteroscedastic Gaussian suggests that the Gaussian is preferred after accounting for the scale parameter heterogeneity.

Table 7.2 presents log-likelihood values for the identical models considered in Table 7.1 using Gaussian REML, t-REML I and t-REML II fits to the data. Section 6.1.7 and 6.2.1 show that as $\nu \to \infty$ the approximate REML obtained under t-REML I and t-REML II, asymptotically approaches the Gaussian REML. However, as the degrees of freedom is fixed hypothesis testing is not available between alternate models. Table 7.2 confirms the squared diameter as the scale parameter model required for all REML methods used here. Furthermore, similar to the final model under ML, the simpler Gaussian REML is preferred over the two approximate REML approaches under the heteroscedastic t-specification.

For comparison, the heteroscedastic Gaussian ML, REML, heteroscedastic t-ML, t-REML I and t-REML II estimates for the location and scale parameters are given, with standard errors, in Table 7.3. The estimates of the location parameters for all the methods presented here are very similar suggesting excellent stability in the location model. The ML estimates of the scale parameters for the Gaussian and t also show similarity. For the REML methods presented here the scale parameter estimates differ slightly. As $\nu = 3$

	Homogeneity Test: 2LogL		
Model	Gaussian	t-REML I	t-REML II
1	38.608	35.805	35.265
H	42.678	40.537	40.087
D	39.399	37.210	36.731
H, D	42.735	40.626	40.173
D, D^2	47.751	44.262	43.976
H, D, D^2	50.350	46.242	45.869
H, D, D^2, HD	50.861	47.198	46.557

Table 7.2: Homogeneity Test for the Cherry Tree data under REML

the standard errors for the ML scale parameter estimates for the t, in comparison to ML under Gaussian, are larger. However, the standard errors for the ML location parameter estimates are smaller for the t. This is most likely due to the outliers in the data being accommodated more efficiently by the scale parameter model under the t-specification. The t-REML I method produced slightly larger standard errors for the scale parameters than its Gaussian equivalent, whereas, t-REML II were slightly less.

	Scale Parameter Estimates			*Location Parameter Estimates*		
Method	1	D	D^2	1	D	H
Gaussian ML	-41.386	5.1705	-0.1768	0.0954	0.1527	0.0117
	(4.7590)	(0.6986)	(0.0247)	(0.0531)	(0.0017)	(0.0010)
Gaussian REML	-29.427	3.4196	-0.1151	0.0302	0.1513	0.0128
	(7.9516)	(1.1653)	(0.0414)	(0.0889)	(0.0030)	(0.0016)
t-ML	-44.771	5.5821	-0.1903	0.0962	0.1529	0.0116
	(6.7308)	(0.9880)	(0.0349)	(0.0439)	(0.0014)	(0.0008)
t-REML I	-30.179	3.4411	-0.1143	0.0543	0.1522	0.0123
	(8.3883)	(1.2202)	(0.0431)	(0.0783)	(0.0030)	(0.0014)
t-REML II	-30.890	3.5487	-0.1182	0.0780	0.1525	0.0119
	(7.6961)	(1.1245)	(0.0398)	(0.0824)	(0.0030)	(0.0015)

Table 7.3: Parameter estimates of Cherry Tree data for the heteroscedastic Gaussian and heteroscedastic t models. Standard errors are in parentheses.

Figure 7.2 shows the log of the squared residuals from the homogeneous scale parameter

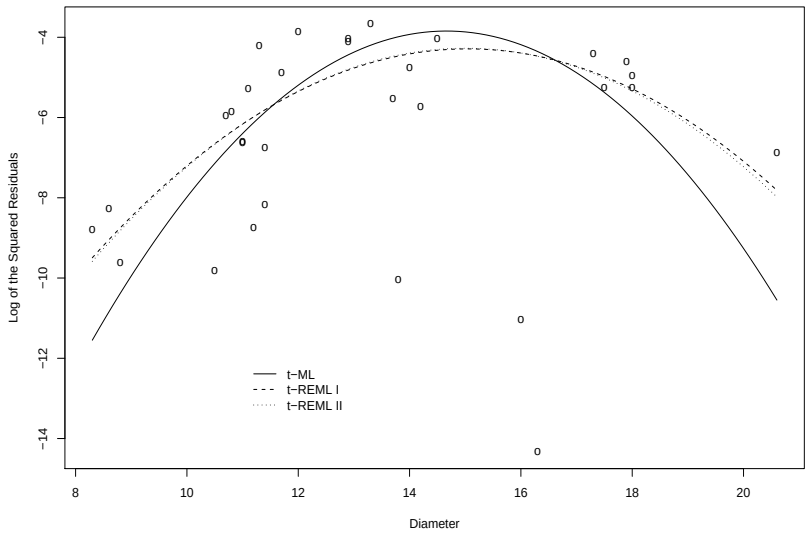

Figure 7.2: Log of the squared residuals from the additive location model with constant scale and $\nu = 3$ against Diameter; t-ML, t-REML I and t-REML II fitted lines for the quadratic scale parameter model are also displayed.

model. The fitted quadratic lines for the final scale parameter models under t-ML, t-REML I and t-REML II are overlayed. The two scale parameter models from approximate REML techniques are less convex than the fitted ML scale parameter model and provide almost coincidental lines.

7.2 Simulations

The examples of the previous section show marked differences in the estimates of the parameters of the heteroscedastic t using ML and the two approximate REML methods. In this section a simulation study is conducted to understand the properties of the estimators of the heteroscedastic t-distribution known degrees of freedom obtained from ML), REML using the Partial Laplace approximation (t-REML I) and Restricted Maximum Likelihood using Modified Profile Likelihood (t-REML II).

For all simulations in this chapter the covariates used to describe the location and scale components are defined by

$$\begin{aligned} \boldsymbol{x}_{i1} &= \boldsymbol{z}_{i1} = 1, \qquad i = 1, \ldots, n \\ \boldsymbol{x}_{i2} &= \boldsymbol{z}_{i2} = 0.1 + 9.9(i-1)/(n-1), \qquad i = 1, \ldots, n \end{aligned} \tag{7.2.1}$$

The scale covariate, $\boldsymbol{z}_2$, was centered around its mean. The model (4.2.7) was used where the location parameters follow a linear form and the scale parameters follow a log-linear

		t-model ($\nu = 3$)			*t-model* ($\nu = 8$)			*t-model* ($\nu = 15$)		
		t-ML	t-RL I	t-RL II	t-ML	t-RL I	t-RL II	t-ML	t-RL I	t-RL II
20	$\hat{\beta}_0$	-0.509	-0.510	-0.509	-0.506	-0.505	-0.505	-0.515	-0.514	-0.515
	$\hat{\beta}_1$	2.012	2.011	2.011	1.996	1.996	1.995	2.009	2.009	2.009
	$\hat{\lambda}_0$	0.278	0.465	0.464	0.258	0.392	0.391	0.295	0.416	0.416
	$\hat{\lambda}_1$	0.525	0.479	0.481	0.546	0.511	0.511	0.532	0.501	0.501
50	$\hat{\beta}_0$	-0.492	-0.492	-0.492	-0.506	-0.506	-0.506	-0.501	-0.500	-0.500
	$\hat{\beta}_0$	2.001	2.001	2.001	2.002	2.002	2.002	2.004	2.003	2.003
	$\hat{\lambda}_0$	0.449	0.519	0.519	0.430	0.481	0.481	0.407	0.453	0.453
	$\hat{\lambda}_1$	0.503	0.488	0.488	0.510	0.498	0.498	0.500	0.490	0.490
100	$\hat{\beta}_0$	-0.495	-0.495	-0.495	-0.502	-0.502	-0.502	-0.507	-0.507	-0.507
	$\hat{\beta}_1$	1.997	1.997	1.997	2.000	2.000	2.000	2.000	2.000	2.000
	$\hat{\lambda}_0$	0.460	0.494	0.494	0.482	0.508	0.508	0.466	0.489	0.489
	$\hat{\lambda}_1$	0.507	0.500	0.500	0.502	0.497	0.497	0.506	0.501	0.501
200	$\hat{\beta}_0$	-0.506	-0.506	-0.506	-0.497	-0.497	-0.497	-0.498	-0.498	-0.498
	$\hat{\beta}_1$	2.001	2.001	2.001	2.001	2.001	2.001	1.999	1.999	1.999
	$\hat{\lambda}_0$	0.481	0.498	0.498	0.488	0.500	0.500	0.486	0.498	0.498
	$\hat{\lambda}_1$	0.503	0.499	0.499	0.502	0.500	0.500	0.501	0.499	0.499

Table 7.4: Mean estimates for $\boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)$ under simulated distribution $y_i \sim t(\boldsymbol{x}_i^T\boldsymbol{\beta}, \exp(\boldsymbol{z}_i^T\boldsymbol{\lambda}), \nu),\ i = 1, \ldots, n$ using ML, t-REML I and t-REML II and fixed degrees of freedom $\nu = (3, 8, 15)$ and $n = (20, 50, 100, 200)$.

form, namely

$$\begin{aligned} \mu_i &= \boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1\boldsymbol{x}_{i2}, \qquad i = 1, \ldots, n \\ \log\ \sigma_i^2 &= \boldsymbol{z}_i^T\boldsymbol{\lambda} = \lambda_0 + \lambda_1\boldsymbol{z}_{i2} \qquad i = 1, \ldots, n \end{aligned} \tag{7.2.2}$$

and

$$\begin{aligned} \boldsymbol{\beta}^T &= (\beta_0, \beta_1) = (-0.5, 2.0) \\ \boldsymbol{\lambda}^T &= (\lambda_0, \lambda_1) = (0.5, 0.5) \end{aligned} \tag{7.2.3}$$

are the target values for the fixed effects of the location and scale parameter models.

The estimates of the location and scale parameters, $(\boldsymbol{\beta}, \boldsymbol{\lambda})$, are obtained from ML, t-REML I and t-REML II using the computational algorithms from Sections 4.7, 6.1.8 and 6.2.2 respectively.

To ensure defined first and second moments ($\nu > 2$) of the true t-distribution under heteroscedasticity, the known values for the degrees of freedom used in this particular

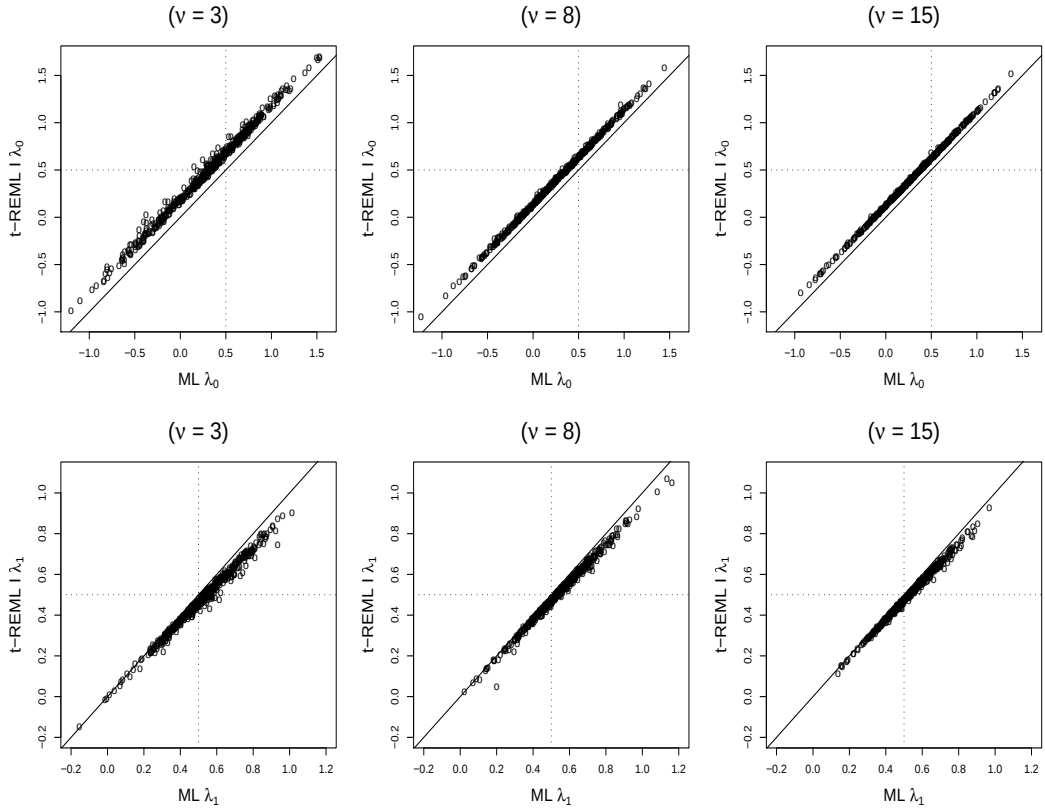

Figure 7.3: The t-REML I estimates of the scale parameters $(\hat{\lambda}_0, \hat{\lambda}_1)$ vs the ML equivalents for 500 simulations under the distribution $y_i \sim t(\boldsymbol{x}_i^T\boldsymbol{\beta}, \exp(\boldsymbol{z}_i^T\boldsymbol{\lambda}), \nu),\ i = 1, \ldots, n$ for $\nu = (3, 8, 15)$ and $n = 20$.

study were $\nu = (3, 8, 15)$. The simulation was run with sample sizes $n = (20, 50, 100, 200)$ to gauge the effect on the properties of the parameter estimates for an increasing number of observations. A total of 500 replications of each combination of (ν, n) was simulated for the three approaches. The convergence criterion in each case was $|\ell(\boldsymbol{\theta}^{(m+1)}) - \ell(\boldsymbol{\theta}^m)| < \epsilon$ where $\boldsymbol{\theta}$ is the parameter of interest and $\epsilon = 10^{-8}$.

Table 7.4 presents the means of the estimates for the fixed location and scale parameters over the 500 simulations for all sample sizes and degrees of freedom combinations. For each of the three methods used in this simulation study, the intercept parameter for the location model is consistently estimated for all sample sizes and degrees of freedom values. The location slope parameter estimates have also been efficiently estimated for all sample size, degrees of freedom and method of estimation combinations suggesting excellent stability in the location component of the model.

It is well known that the variance is biased under ML for standard Gaussian models. Similarly, using ML under the t specification produces biased estimates for the intercept parameters of the scale model. Table 7.4 suggests that this bias increases as the sample

ν	SS	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\lambda}_0$	$\hat{\lambda}_1$	SS	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\lambda}_0$	$\hat{\lambda}_1$
3	20	0.295	0.118	0.447	0.147	50	0.199	0.077	0.283	0.096
8		0.266	0.106	0.371	0.122		0.180	0.069	0.235	0.080
15		0.256	0.102	0.346	0.114		0.173	0.067	0.219	0.074
3	100	0.144	0.055	0.200	0.069	200	0.103	0.039	0.141	0.049
8		0.130	0.050	0.166	0.057		0.093	0.035	0.117	0.040
15		0.125	0.048	0.155	0.053		0.089	0.034	0.110	0.038

Table 7.5: Table of theoretical standard errors for $\boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)$ for the true simulated distribution $y_i \sim t(\boldsymbol{x}_i^T\boldsymbol{\beta}, \exp(\boldsymbol{z}_i^T\boldsymbol{\lambda}), \nu),\ i = 1, \ldots, n$, under ML.

size decreases. The approximate t-REML I and t-REML II likelihoods derived in this thesis suggest that this bias may be reduced. This reduction can be seen in Table 7.4 for the lowest sample size and degrees of freedom combination ($n = 20, \nu = 3$). The bias reduction of the intercept scale parameter decreases for higher degrees of freedom values and low sample sizes but this diminishes as the sample size is increased. For this particular simulation study, the t-REML I and t-REML II methods produce almost identical scale parameter estimates for all degrees of freedom values and sample sizes.

Under ML, the slope parameter for the scale model is also biased for low sample sizes and all degrees of freedom values. This bias decreases as the sample size is increased. The estimates of the slope for the scale parameter model under the two t-REML constructs reduce this bias. This reduction becomes negligible as the sample size increases.

It is of interest to understand the bias reduction of the estimated scale parameters obtained from the approximate REML constructs derived in this thesis in comparison to ML when the degrees of freedom is known. Figure 7.3 shows the 500 simulated empirical t-REML I estimates for the scale parameters against the ML equivalents for $\nu = (3, 8, 15)$ and $n = 20$. It can be seen for all fixed degrees of freedom t-REML I estimates of the scale parameter exhibit less bias than ML below the target value $\lambda_0 = 0.5$. Conversely, for ML estimates above the required value t-REML I increases the bias further. For estimated values of the scale slope parameter above the target value, t-REML I reduces the bias in comparison to ML. For estimated values below $\lambda_1 = 0.5$, the bias is not reduced and ML performs slightly better than the t-REML I equivalents. As the graphical results of the scale parameter estimates for t-REML II in comparison to ML are similar to those presented in Figure 7.3 they have been omitted for brevity.

The asymptotic score hypothesis tests derived in Section 4.5.2 require the use of the variance matrix for the location and scale parameter estimates. The approximate theoretical asymptotic variability of the location and scale parameter under Maximum Likelihood

		t-model ($\nu = 3$)			*t-model* ($\nu = 8$)			*t-model* ($\nu = 15$)		
		t-ML	t-RL I	t-RL II	t-ML	t-RL I	t-RL II	t-ML	t-RL I	t-RL II
20	$\hat{\beta}_0$	0.344	0.346	0.343	0.304	0.305	0.304	0.273	0.273	0.273
	$\hat{\beta}_1$	0.132	0.130	0.131	0.115	0.114	0.114	0.103	0.103	0.103
	$\hat{\lambda}_0$	0.479	0.475	0.476	0.429	0.428	0.428	0.384	0.383	0.383
	$\hat{\lambda}_1$	0.174	0.158	0.159	0.161	0.150	0.151	0.135	0.126	0.126
50	$\hat{\beta}_0$	0.225	0.225	0.225	0.191	0.191	0.191	0.176	0.176	0.176
	$\hat{\beta}_1$	0.078	0.078	0.078	0.071	0.071	0.071	0.068	0.068	0.068
	$\hat{\lambda}_0$	0.282	0.281	0.281	0.232	0.233	0.233	0.235	0.235	0.235
	$\hat{\lambda}_1$	0.100	0.097	0.097	0.089	0.087	0.087	0.082	0.080	0.080
100	$\hat{\beta}_0$	0.146	0.146	0.146	0.141	0.141	0.141	0.132	0.132	0.132
	$\hat{\beta}_1$	0.056	0.056	0.056	0.047	0.047	0.047	0.049	0.049	0.049
	$\hat{\lambda}_0$	0.202	0.202	0.202	0.169	0.169	0.169	0.158	0.158	0.158
	$\hat{\lambda}_1$	0.072	0.071	0.071	0.057	0.057	0.057	0.054	0.054	0.054
200	$\hat{\beta}_0$	0.112	0.112	0.112	0.097	0.097	0.097	0.088	0.088	0.088
	$\hat{\beta}_1$	0.041	0.041	0.041	0.036	0.036	0.036	0.031	0.031	0.031
	$\hat{\lambda}_0$	0.134	0.134	0.134	0.117	0.117	0.117	0.110	0.110	0.110
	$\hat{\lambda}_1$	0.047	0.047	0.047	0.043	0.042	0.042	0.038	0.038	0.038

Table 7.6: Empirical standard errors of the estimates $\boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)$ under simulated distribution $y_i \sim t(\boldsymbol{x}_i^T\boldsymbol{\beta}, \exp(\boldsymbol{z}_i^T\boldsymbol{\lambda}), \nu),\ i = 1, \ldots, n$ using ML, t-REML I and t-REML II and fixed degrees of freedom $\nu = (3, 8, 15)$ and $n = (20, 50, 100, 200)$.

estimation was derived in Section 4.5.1, namely

$$\text{Var}\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \frac{\nu+3}{\nu+1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} & \mathbf{0} \\ \mathbf{0} & \frac{2(\nu+3)}{\nu}\boldsymbol{Z}^T\boldsymbol{Z} \end{bmatrix}$$

These variances suggest that as the degrees of freedom parameter value increases the variability in the estimates should decrease. Moreover, if $\nu = 3$, then $\text{Var}(\hat{\boldsymbol{\lambda}}) = 4\boldsymbol{Z}^T\boldsymbol{Z}$, which is twice the normal variability when $\hat{\boldsymbol{\lambda}}$ is Gaussian. For all sample sizes, this decrease in the theoretical standard errors, can be seen in Table 7.5. For all fixed degrees of freedom values used in this simulation study the theoretical standard errors decreased as the sample size increased.

For comparison, the empirical standard errors for the simulated distribution of the estimates, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}})$, for all known degrees of freedom values and sample sizes is presented in Table 7.6. Similar to the theoretical standard errors derived under ML, the empirical standard errors under ML, t-REML I, and t-REML II decrease as the fixed degrees of freedom and the sample size increases. For all parameters the theoretical standard errors

are marginally less than their empirical equivalents and this difference increases as the sample size decreases suggesting that the empirical standard errors should be used with caution. Furthermore, in comparison to ML, t-REML I and t-REML II empirical standard errors for the slope of the scale parameter model are slightly less for small degrees of freedom and small sample sizes. This difference diminishes as the sample size and degrees of freedom increase. This suggests that t-REML I and t-REML II are more efficient than ML in obtaining the target values for the slope parameter of the scale model when the sample size is low and the the degrees of freedom is fixed at a small value in advance.

Chapter 8

Heteroscedastic t-distribution with unknown degrees of freedom

In previous chapters the degrees of freedom for the heteroscedastic t-distribution was fixed or known in advance. In this chapter ν is allowed to vary according to the data and techniques for its estimation are discussed.

The ML techniques of Chapter 4 are extended by incorporating a mechanism for the estimation of the degrees of freedom parameter. To ensure the estimating equations for the location and scale parameters remain identical an orthogonal transformation of the scale parameter is performed. As the new orthogonal parameter is a function of the degrees of freedom appropriate adjustments are derived. To increase the flexibility of the model further in this chapter the approximate t-REML I technique discussed in Chapter 6 Section 6.1 is extended to simultaneously estimate the parameters $(\boldsymbol{\lambda}, \nu)$. Furthermore, similar to the t-REML II technique discussed in Chapter 6 Section 6.2, a flexible extension of Modified Profile Likelihood called Stably Adjusted Profile Likelihood is considered that allows the separate estimation of the parameters of the heteroscedastic t in the presence of multiple nuisance parameters.

8.1 Heteroscedastic t-ML

Consider the linear model (3.1.1) where the distribution of the response is defined by (4.2.7). For this particular section let the scale parameter model be defined by (3.1.2).

8.1.1 Estimating the Degrees of Freedom

Estimating the degrees of freedom of the univariate and multivariate t-distribution under maximum likelihood has been attempted by many authors. Lange et al. (1989), Liu & Rubin (1994), Liu & Rubin (1995) and Pinheiro et al. (2001) include the estimation of ν in an Expectation-Maximisation framework. In particular, Liu & Rubin (1994), Liu & Rubin (1995) and Pinheiro et al. (2001) discuss an extension of the EM which allows the degrees of freedom parameter to be estimated from the marginal likelihood derived from the t-distribution constrained at estimates $(\boldsymbol{\beta}, \sigma^2(\boldsymbol{z}_i; \boldsymbol{\lambda})) = (\hat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{z}_i; \hat{\boldsymbol{\lambda}}))$. Method of moments estimators have also been utilised by Sutradhar & Ali (1986) and Singh (1988) to overcome potential problems with maximum likelihood estimation of ν for the multivariate t-distribution (see Breusch et al., 1997 for more details).

A simple score function for ν can be derived by taking the first derivative of (4.2.9) with respect to ν, namely

$$\boldsymbol{U}(\nu) \;=\; \frac{n}{2}\psi\Big(\frac{\nu+1}{2}\Big) \;-\; \frac{n}{2}\psi\Big(\frac{\nu}{2}\Big) \;-\; \frac{n}{2\nu} \;-\; \frac{1}{2}\sum_{i=1}^{n}\log\big(1 \;+\; d_i/\sigma_i^2\nu\big) \;+\; \frac{\nu+1}{2\nu}\sum_{i=1}^{n}B_i$$

where $\psi(\theta) = \partial \log\Gamma(\theta)/\partial\theta$ is the digamma function and B_i is defined by (4.3.6).

To estimate ν from this score equation and maintain the reweighted least squares approach to estimation of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ requires independence of the parameters. This dependence can be checked by obtaining the co-information components from the full information matrix. Let $\boldsymbol{\theta}^* = (\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu)$ then the full observed information can be expressed as

$$\mathcal{I}_o(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = \begin{bmatrix} \mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\beta}) & \mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\lambda}) & \mathcal{I}_o(\boldsymbol{\beta},\nu) \\ \mathcal{I}_o(\boldsymbol{\lambda},\boldsymbol{\beta}) & \mathcal{I}_o(\boldsymbol{\lambda},\boldsymbol{\lambda}) & \mathcal{I}_o(\boldsymbol{\lambda},\nu) \\ \mathcal{I}_o(\nu,\boldsymbol{\beta}) & \mathcal{I}_o(\nu,\boldsymbol{\lambda}) & \mathcal{I}_o(\nu,\nu) \end{bmatrix}$$

Differentiating (4.3.4) with respect to ν allows the lth co-information element for $(\boldsymbol{\beta}, \nu)$ to be expressed as

$$\mathcal{I}_o(\beta_l, \nu) \;=\; \sum_{i=1}^{n}\frac{x_{il}}{\sigma_i^2}\left\{\left(\frac{\nu+1}{\nu^2} - \frac{1}{\nu}\right)(1 - B_i) + \frac{\nu+1}{\nu}\frac{\partial B_i}{\partial \nu}\right\}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) \quad (8.1.1)$$

(8.1.2)

where

$$\begin{aligned} \frac{\partial B_i}{\partial \nu} &= \left(\frac{d_i/\sigma_i^2\nu}{(1 + d_i/\sigma_i^2\nu)^2}\right)\frac{d_i}{\sigma_i^2(\nu)^2} - \left(\frac{d_i/\sigma_i^2(\nu)^2}{1 + d_i/\sigma_i^2\nu}\right) \\ &= \left(\frac{d_i/\sigma_i^2\nu}{1 + d_i/\sigma_i^2\nu}\right)^2\frac{1}{\nu} - \left(\frac{d_i/\sigma_i^2\nu}{1 + d_i/\sigma_i^2\nu}\right)\frac{1}{\nu} \\ &= (B_i^2 - B_i)\frac{1}{\nu} \end{aligned} \quad (8.1.3)$$

Substitution of this into (8.1.1) gives

$$\begin{aligned}\mathcal{I}_o(\beta_l,\nu) &= \sum_{i=1}^{n}\frac{x_{il}}{\sigma_i^2}\left\{\frac{\nu+1}{\nu^2}(B_i^2-B_i)-\frac{1}{\nu^2}(1-B_i)\right\}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})\\ &= \sum_{i=1}^{n}\frac{x_{il}}{\sigma_i^2\nu^2}\left\{(\nu+1)B_i^2-\nu B_i-1\right\}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}) \qquad (8.1.4)\end{aligned}$$

Differentiating (4.3.5) with respect to ν allows the jth co-information element $(\boldsymbol{\lambda},\nu)$ to be expressed as

$$\begin{aligned}\mathcal{I}_o(\lambda_j,\nu) &= \frac{1}{2}\sum_{i=1}^{n}\frac{s_{ij}}{\sigma_i^2}\left\{-B_i-(\nu+1)\frac{\partial B_i}{\partial\nu}\right\}\\ &= \frac{1}{2}\sum_{i=1}^{n}\frac{s_{ij}}{\sigma_i^2}\left\{\left(\frac{\nu+1}{\nu}-1\right)B_i-\frac{\nu+1}{\nu}B_i^2\right\}\\ &= \frac{1}{2}\sum_{i=1}^{n}\frac{s_{ij}}{\sigma_i^2\nu}\left\{B_i-(\nu+1)B_i^2\right\} \qquad (8.1.5)\end{aligned}$$

The observed information element for ν is

$$\begin{aligned}\mathcal{I}_o(\nu,\nu) &= \frac{n}{4}\dot{\psi}\Big(\frac{\nu}{2}\Big)-\frac{n}{4}\dot{\psi}\Big(\frac{\nu+1}{2}\Big)-\frac{n}{2\nu^2}-\frac{1}{2\nu}\sum_{i=1}^{n}B_i\\ &- \left\{\frac{1}{2\nu}-\frac{\nu+1}{2\nu^2}\right\}\sum_{i=1}^{n}B_i-\frac{\nu+1}{2\nu}\sum_{i=1}^{n}\frac{\partial B_i}{\partial\nu} \qquad (8.1.6)\end{aligned}$$

where $\dot{\psi}(\theta)=\partial^2\log\Gamma(\theta)/(\partial\theta)^2$ is the trigamma function. Substituting (8.1.3) the observed information element for ν becomes

$$\mathcal{I}_o(\nu,\nu)=\frac{n}{4}\dot{\psi}\Big(\frac{\nu}{2}\Big)-\frac{n}{4}\dot{\psi}\Big(\frac{\nu+1}{2}\Big)-\frac{n}{2\nu^2}-\frac{1}{2\nu^2}\sum_{i=1}^{n}\left\{(\nu+1)B_i^2-2B_i\right\} \qquad (8.1.7)$$

The full expected information matrix can be expressed as

$$\mathcal{I}_e(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)=\begin{bmatrix}\mathcal{I}_e(\boldsymbol{\beta},\boldsymbol{\beta}) & \mathcal{I}_e(\boldsymbol{\beta},\boldsymbol{\lambda}) & \mathcal{I}_e(\boldsymbol{\beta},\nu)\\ \mathcal{I}_e(\boldsymbol{\lambda},\boldsymbol{\beta}) & \mathcal{I}_e(\boldsymbol{\lambda},\boldsymbol{\lambda}) & \mathcal{I}_e(\boldsymbol{\lambda},\nu)\\ \mathcal{I}_e(\nu,\boldsymbol{\beta}) & \mathcal{I}_e(\nu,\boldsymbol{\lambda}) & \mathcal{I}_e(\nu,\nu)\end{bmatrix}$$

Taking expectations of (8.1.4), the lth element of the expected co-information for $(\boldsymbol{\beta},\nu)$ becomes

$$\mathcal{I}_e(\beta_l,\nu)=\sum_{i=1}^{n}\frac{x_{il}}{\sigma_i^2\nu^2}\mathrm{E}\left\{(\nu+1)B_i^2(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})-\nu B_i(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})\right\}$$

Again, the two terms in parentheses are odd functions of y_i and therefore, under expectation, are zero. Noting (4.3.7) and (4.3.8) and taking expectations of (8.1.5) the jth

element of the expected information for $(\boldsymbol{\lambda}, \nu)$ becomes

$$\begin{aligned}\mathcal{I}_e(\lambda_j, \nu) &= \frac{1}{2}\sum_{i=1}^{n}\frac{s_{ij}}{\sigma_i^2\nu}\left\{\frac{1}{\nu+1}-\frac{3}{\nu+3}\right\} \\ &= -\sum_{i=1}^{n}\frac{s_{ij}}{\sigma_i^2}\left\{\frac{1}{(\nu+1)(\nu+3)}\right\}\end{aligned}$$

Taking expectations of (8.1.7) the expected information element for ν becomes

$$\begin{aligned}\mathcal{I}_e(\nu,\nu) &= \frac{n}{4}\dot{\psi}\Big(\frac{\nu}{2}\Big)-\frac{n}{4}\dot{\psi}\Big(\frac{\nu+1}{2}\Big)-\frac{n}{2\nu^2}-\frac{1}{2\nu^2}\sum_{i=1}^{n}\left\{\frac{3}{\nu+3}-\frac{2}{\nu+1}\right\} \\ &= \frac{n}{4}\dot{\psi}\Big(\frac{\nu}{2}\Big)-\frac{n}{4}\dot{\psi}\Big(\frac{\nu+1}{2}\Big)-\frac{n}{2\nu^2}\left\{\frac{3}{\nu+3}-\frac{2}{\nu+1}+1\right\} \\ &= \frac{n}{4}\dot{\psi}\Big(\frac{\nu}{2}\Big)-\frac{n}{4}\dot{\psi}\Big(\frac{\nu+1}{2}\Big)-\frac{n}{2\nu}\frac{(\nu+5)}{(\nu+1)(\nu+3)}\end{aligned}$$

In matrix notation the 2×2 sub block of the expected information gives

$$\begin{bmatrix}\mathcal{I}_e(\boldsymbol{\lambda},\boldsymbol{\lambda}) & \mathcal{I}_e(\boldsymbol{\lambda},\nu)\\ \mathcal{I}_e(\nu,\boldsymbol{\lambda}) & \mathcal{I}_e(\nu,\nu)\end{bmatrix} = \begin{bmatrix}\frac{\nu}{2(\nu+3)}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-2}\dot{\boldsymbol{S}} & -\frac{1}{(\nu+1)(\nu+3)}\dot{\boldsymbol{S}}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_n \\ -\frac{1}{(\nu+1)(\nu+3)}\mathbf{1}_n^T\boldsymbol{\Sigma}^{-1}\boldsymbol{S} & \frac{n}{4}\dot{\psi}(\frac{\nu}{2})-\frac{n}{4}\dot{\psi}(\frac{\nu+1}{2})-\frac{n}{2\nu}\frac{(\nu+5)}{(\nu+1)(\nu+3)}\end{bmatrix}$$

The non-zero components of the co-information reveal the dependency of ν and $\boldsymbol{\lambda}$. Furthermore it is clear that the information for ν accumulates without bound as $n\to\infty$ (see Breusch et al., 1997 for more details). Therefore the univariate maximum likelihood estimation procedure provides a framework for consistent estimation of the degrees of freedom parameter.

8.1.2 Orthogonal Transformation

To obtain independence between the co-dependent parameters an orthogonal transformation is required. Let $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ be an unknown vector of parameters of length $p+q$ orthogonal to ν. For simplification let the scale parameter sub-model be defined by (3.1.3). From Cox & Reid (1987) and Section 5.2.2 the new orthogonal parameters are derived using the partial differential equations

$$\mathcal{I}(\boldsymbol{\theta},\boldsymbol{\theta})\frac{\partial\boldsymbol{\theta}}{\partial\nu} = -\mathcal{I}(\boldsymbol{\theta},\nu) \tag{8.1.8}$$

As $\boldsymbol{\beta}$ is orthogonal to ν, $\boldsymbol{\delta}_1 = \boldsymbol{\beta}$, whereas, to orthogonalize the scale parameters, the following partial differential equation must be solved by integration

$$\frac{\nu}{2(\nu+3)}\boldsymbol{Z}^T\boldsymbol{Z}\frac{\partial\boldsymbol{\lambda}}{\partial\nu} = \frac{1}{(\nu+1)(\nu+3)}\boldsymbol{Z}^T\mathbf{1}_n$$

Choosing the constant of integration $a(\boldsymbol{\delta}_2) = \boldsymbol{\delta}_2$, gives

$$\boldsymbol{\delta}_2 = \boldsymbol{\lambda} - (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\mathbf{1}_n\big\{2(\log\nu - \log(\nu+1))\big\}$$

as the orthogonal set of parameters. Note that in Taylor & Verbyla (2004) the orthogonal transformation is incorrectly reported. It is clear that the second term of this new transformation does not contain $\sigma_i^2,\ i = 1,\ldots,n$ and therefore estimation of the scale parameters remains identical to the previous section. For a homogeneous model, $\sigma^2 = \exp\lambda_0$, and the orthogonal parameter reduces to $(\sigma(\nu+1)/\nu)^2$. This result is similar to the reparameterization discussed in Jones & Faddy (2003). For the heteroscedastic case and functions of the scale parameters, $\sigma_i^2,\ i = 1,\ldots,n$, other than the reciprocal or natural logarithm the partial differential equations given by (8.1.8) are not easily solvable.

Using (4.2.9) and for known $(\boldsymbol{\beta}, \boldsymbol{\delta}_2)$, a maximising objective function for ν would be

$$\begin{aligned}\ell(\nu;\boldsymbol{y}) &= n\log\big(\Gamma((\nu+1)/2)\big) - n\log\big(\Gamma(1/2)\big) - n\log\big(\Gamma(\nu/2)\big) - \frac{n}{2}\log\nu \\ &- \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{z}_i^T(\boldsymbol{\delta}_2 + (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\mathbf{1}_n\big\{2(\log\nu - \log(\nu+1))\big\} - \frac{\nu+1}{2}\sum_{i=1}^{n}\log\left\{1 + \frac{d_i}{\sigma_i^2\nu}\right\}\end{aligned} \tag{8.1.9}$$

As the parameters are now orthogonal a maximum likelihood estimate for ν can then be found using a simple one dimensional search. For given $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ the $(m+1)$th iterate becomes

$$\nu_{m+1} = h_t(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu_m) = \nu_m + \boldsymbol{U}^*(\nu_m)/\mathcal{I}_e^*(\nu_m, \nu_m) \tag{8.1.10}$$

where $\boldsymbol{U}^*(\cdot)$ and $\mathcal{I}_e^*(\cdot,\cdot)$ are an adjusted score and expected information element respectively derived from the likelihood (8.1.9) containing the orthogonally transformed parameters. Let $z_i^* = \boldsymbol{z}_i^T(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\mathbf{1}_n$. The the adjusted score, $\boldsymbol{U}^*$, is found by differentiating (8.1.9) with respect to ν giving

$$\begin{aligned}\boldsymbol{U}^*(\nu) &= \boldsymbol{U}(\nu) - \frac{1}{\nu(\nu+1)}\sum_{i=1}^{n} z_i^* + \frac{1}{\nu}\sum_{i=1}^{n} B_i z_i^* \\ &= \boldsymbol{U}(\nu) - \frac{1}{\nu(\nu+1)}\sum_{i=1}^{n}\left\{1 - \frac{d_i\bar{\omega}_i}{\sigma_i^2}\right\} z_i^* \\ &= \boldsymbol{U}(\nu) - \frac{1}{\nu(\nu+1)}(\mathbf{1}_n^T - \boldsymbol{d}^T\bar{\boldsymbol{\Omega}}\boldsymbol{\Sigma}^{-1})\boldsymbol{P}_z\mathbf{1}_n\end{aligned} \tag{8.1.11}$$

where $\boldsymbol{P}_z = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$. Noting (8.1.3), the adjusted observed information can be found by differentiating (8.1.11) with respect to ν giving

$$\mathcal{I}_o^*(\nu,\nu) = \mathcal{I}_o(\nu,\nu) + \boldsymbol{A}(\nu)$$

where

$$\boldsymbol{A}(\nu) = -\frac{2\nu+1}{\nu^2(\nu+1)^2}\sum_{i=1}^{n} z_i^* - \frac{1}{\nu^2}\sum_{i=1}^{n}\Big\{2B_i^2 - \frac{2\nu+3}{\nu+1}B_i\Big\}z_i^* - \frac{2}{\nu^2(\nu+1)}\sum_{i=1}^{n}(B_i^2 - B_i)z_i^{*2}$$

Taking expectations of this gives

$$\begin{aligned}
\mathrm{E}\big[\boldsymbol{A}(\nu)\big] &= -\sum_{i=1}^{n}\left\{\frac{2\nu+1}{\nu^2(\nu+1)^2}z_i^* + \frac{1}{\nu^2(\nu+1)}\left(\frac{6(\nu+1)}{\nu+3} - (2\nu+3)\right)z_i^* \right. \\
&\qquad \left. + \frac{2}{\nu^2(\nu+1)^2}\left(\frac{3}{\nu+3} - 1\right)z_i^{*2}\right\} \\
&= -\sum_{i=1}^{n}\frac{1}{\nu^2(\nu+1)^2(\nu+3)}\left\{(-2(\nu+3) + 6(\nu+1))\,z_i^* - 2\nu z_i^{*2}\right\}
\end{aligned}$$

As

$$\begin{aligned}
\sum_{i=1}^{n} z_i^{*2} &= \sum_{i=1}^{n}\mathbf{1}_n^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{z}_i\boldsymbol{z}_i^T(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\mathbf{1}_n \\
&= \mathbf{1}_n^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\Big\{\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{z}_i^T\Big\}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\mathbf{1}_n \\
&= \mathbf{1}_n^T\boldsymbol{P}_z\boldsymbol{P}_z\mathbf{1}_n \\
&= \mathbf{1}_n^T\boldsymbol{P}_z\mathbf{1}_n = \sum_{i=1}^{n} z_i^*
\end{aligned}$$

the adjusted expected information for ν becomes

$$\mathcal{I}_e^*(\nu,\nu) = \mathcal{I}_e(\nu,\nu) - \frac{2}{\nu(\nu+1)^2(\nu+3)}\mathbf{1}^T\boldsymbol{P}_z\mathbf{1}_n$$

Notice for large degrees of freedom the contribution from this component will be negligible.

As discussed in Fernandez & Steel (1999), the estimate for ν obtained by iterating (8.1.10) will only locally maximise (8.1.9). This is due to the likelihood becoming unbounded if ν is allowed to vary over the whole parameter space. For the univariate case proposed here, this occurs when $\nu = \nu_0 < s(\boldsymbol{\beta}_0)/(n - s(\boldsymbol{\beta}_0))$ where $s(\boldsymbol{\beta}_0)$ is the number of observations, given $\boldsymbol{\beta}_0$ such that $y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_0 = 0,\ i = 1,\ldots,n$. In general this threshold, ν_0, is small (generally less than one) and infers that areas of likelihood unboundedness are most likely to occur as $\nu \to 0$ (see Fernandez & Steel, 1999; Azzalini & Capitanio, 2003 and Jones & Faddy, 2003 for further details). For the simulations and examples in this thesis these regions of the likelihood were avoided.

8.1.3 Computation and Software

The inclusion of the degrees of freedom parameter in the estimation process requires a simple modification of the scoring algorithm given in Section 4.7. For the $(m+1)$th iterate the scoring algorithm becomes

- **Score Step 1:** For given $\omega_i^{(m+1)}$, $1 = 1, \ldots, n$, $\boldsymbol{\delta} = \boldsymbol{\delta}^{(m)}$ and $\nu = \nu^{(m)}$ update $\boldsymbol{\beta}$ using $\boldsymbol{\beta}^{(m+1)} = f_t(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}^{(m)}, \nu^{(m)})$, where $f_t(\cdot)$ is given in (4.3.16).
- **Score Step 2:** For given $\omega_i^{(m+1)}$, $1 = 1, \ldots, n$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m+1)}$ and $\nu = \nu^{(m)}$ update $\boldsymbol{\delta}$ using $\boldsymbol{\delta}^{(m+1)} = g_t(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\delta}^{(m)}, \nu^{(m)})$, where $g_t(\cdot)$ is given in (4.3.18).
- **Score Step 3:** For given $\omega_i^{(m+1)}$, $1 = 1, \ldots, n$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m+1)}$ and $\boldsymbol{\delta} = \boldsymbol{\delta}^{(m+1)}$ update ν using $\nu^{(m+1)} = h_t(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\delta}^{(m+1)}, \nu^{(m)})$, where $h_t(\cdot)$ is given in (8.1.10).

Note that the estimation of $\boldsymbol{\delta}$ at score step 2 only requires the original estimation equation used to update $\boldsymbol{\lambda}$.

An implementation of this algorithm is available in the mini software library "hett". In particular, initial estimates for the degrees of freedom parameter can be given and an estimate as well as its standard error are returned. See the documentation at

`http://www.biometricssa.adelaide.edu.au/staff/hett/index.html`

or Appendix B for more details.

8.2 Heteroscedastic t-REML using the Partial Laplace approximation

In Section 6.1 an approximate marginal likelihood was derived that allowed a separate approximate conditional likelihood for the location parameters to be derived as well as an approximate heteroscedastic t-REML to estimate the scale parameters when the degrees of freedom is known in advance.

When the degrees of freedom parameter is unknown the approximate conditional likelihood and t-REML likelihood remains identical to (6.1.17) and (6.1.18). Therefore estimation of the location parameters is achieved by the least squares process defined by (6.1.28). The estimation of the scale parameters, $(\boldsymbol{\lambda}, \nu)$, is achieved by maximising the natural log of the integrand of (6.1.18), namely

$$\begin{aligned} \ell_2(\boldsymbol{\lambda}, \nu; \boldsymbol{y}) &= \frac{n\nu}{2}\log(\nu/2) - n\Gamma(\nu/2) - \frac{1}{2}\log|\tilde{\boldsymbol{Q}}| - \frac{1}{2}\log|\boldsymbol{X}^T\tilde{\boldsymbol{\Psi}}^{-1}\boldsymbol{X}| \\ &- \frac{1}{2}\sum_{i=1}^{n}\left\{\log \boldsymbol{z}_i^T\boldsymbol{\lambda} - \frac{d_{i,c}\tilde{\omega}_i}{\sigma_i^2}\right\} - \sum_{i=1}^{n}\left\{\left(\frac{\nu+1}{2}\right)\log\omega_i - \frac{\nu}{2}\omega_i\right\} \quad (8.2.1) \end{aligned}$$

where $\tilde{\boldsymbol{Q}}$ is defined by (6.1.30). Identical to Section 6.1.6 $\tilde{\boldsymbol{Q}}$ is a complex function of the remaining parameters, $(\boldsymbol{\lambda}, \nu)$ and therefore the estimation is handled numerically. To ensure positive definiteness $\tilde{\boldsymbol{Q}}$ is replaced with the approximation $\tilde{\boldsymbol{Q}}^*$ defined by (6.1.31).

8.2.1 Computations

For unknown degrees of freedom the estimation of the parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu)$ requires only a modification of the computational algorithm provided in Section 6.1.8.

- **Estimation of $(\boldsymbol{\lambda}, \nu)$:** The approximate REML estimates for $(\boldsymbol{\lambda}, \nu)$ are simultaneously obtained using an iterative process. At the mth iteration
  - For given $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(m)}$ and $\nu = \nu^{(m)}$ update $\omega_i,\ 1, \ldots, n$ using the prediction $\omega_i^{(m+1)} = k(\boldsymbol{\omega}^{(m)}, \boldsymbol{\lambda}^{(m)}, \nu^{(m)})$, where $k(\cdot)$ is given in (6.1.20).
  - For given $\omega_i^{(m+1)},\ i = 1, \ldots, n$ update $(\boldsymbol{\lambda}, \nu)$ using approximate t-REML defined by (8.2.1), namely

$$(\boldsymbol{\lambda}^{(m+1)}, \nu^{(m+1)}) = \max\big\{(\boldsymbol{\lambda}, \nu); L_2(\boldsymbol{\lambda}^{(m)}, \nu^{(m)}; \boldsymbol{y})\big\}$$

- **Estimation of $\boldsymbol{\beta}$:** The approximate REML estimate for $\boldsymbol{\beta}$ can be obtained using
  - For given $(\boldsymbol{\lambda}, \nu) = (\hat{\boldsymbol{\lambda}}, \hat{\nu})$, the approximate REML estimates for the scale parameters, $(\boldsymbol{\lambda}, \nu)$ obtained from Step 1 of the algorithm, and $\omega_i = \omega_i(\hat{\boldsymbol{\lambda}}, \hat{\nu})$, $1, \ldots, n$, update $\boldsymbol{\beta}$ using $\boldsymbol{\beta} = f_t^*(\boldsymbol{\beta}^{(m)}, \hat{\boldsymbol{\lambda}}, \hat{\nu})$, where $f_t^*(\cdot)$ is given in (6.1.28).

For the purpose of brevity the computational algorithm and the estimators obtained from it will be known as t-REML I with unknown degrees of freedom.

8.3 Heteroscedastic t-REML using Stably Adjusted Profile Likelihood

It seems appropriate that to find an MPL for any one of the parameters of the heteroscedastic t-distribution when the degrees of freedom is unknown, the original marginal likelihood needs adjusting for the estimation of the two other nuisance parameters. This would suggest that three modified profile likelihoods are possible for each of the three parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu)$ respectively.

However, the derivation of MPL given in Section 5.2.1 suggests that to form the adjustment requires the availability of ancillary statistics. When the degrees of freedom parameter requires estimation the heteroscedastic t-distribution is not a member of the simple location-scale family and therefore there is no ancillary statistic directly available. Furthermore, the implicit and non-linear nature of the maximum likelihood estimation procedure for $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu)$ given by (4.3.16), (4.3.18) and (8.1.10) respectively suggests difficulty in obtaining the determinant term of (5.2.5) containing the Jacobian of the

transformation to form the complete MPL. From the discussion by Barndorff-Nielsen in Cox & Reid (1987) and Chapter 5 Section 5.2.2, excluding this term would imply that MPL is not invariant under interest respecting parameter transformations. In this case the heteroscedastic t-distribution contains a transformation of the scale parameters $\sigma_i^2 = \sigma^2(\boldsymbol{z}_i; \boldsymbol{\lambda}),\ i = 1, \ldots, n$ and therefore maintaining invariance is preferable.

In Chapter 5 Section 5.2.4 methodology is discussed to obtain a parameter invariant modified or Stably Adjusted Profile Likelihood (SAPL) without the explicit use of ancillary statistics. This is used in the following sections to obtain an alternative form for the adjustment terms obtained from MPL and produce a set of SAPLs, one for each parameter.

Consider the model defined by (3.1.1) where the distribution of the response is defined by (4.2.7). Following from Section 5.2.4 the form of the stably adjusted profile likelihood can be simplified if the parameters are orthogonal. If the scale parameter model follows the simplified log-linear form given by (3.1.3) then following from Section 8.1.1 the scale parameters, $\boldsymbol{\lambda}$, can be orthogonalized by the parameter transformation $\boldsymbol{\delta} = \boldsymbol{\lambda} - (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\mathbf{1}_n\big\{2(\log\nu - \log(\nu+1))\big\}$. This ensures that the expected information for $(\boldsymbol{\beta}, \boldsymbol{\delta}, \nu)$ can be expressed as a block diagonal matrix of the form

$$\begin{bmatrix} \mathcal{I}_e(\boldsymbol{\beta},\boldsymbol{\beta}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_e(\boldsymbol{\delta},\boldsymbol{\delta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathcal{I}_e(\nu,\nu) \end{bmatrix} \tag{8.3.1}$$

where

$$\mathcal{I}_e(\boldsymbol{\beta},\boldsymbol{\beta}) = \frac{\nu+1}{\nu+3}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \tag{8.3.2}$$

$$\mathcal{I}_e(\boldsymbol{\delta},\boldsymbol{\delta}) = \frac{\nu}{2(\nu+3)}\boldsymbol{Z}^T\boldsymbol{Z} \tag{8.3.3}$$

$$\mathcal{I}_e^*(\nu,\nu) = -\frac{n}{4}\dot{\psi}(\tfrac{\nu}{2}) + \frac{n}{4}\dot{\psi}(\tfrac{\nu+1}{2}) - \frac{n}{2\nu}\frac{(\nu+5)}{(\nu+1)(\nu+3)} - \frac{2}{\nu(\nu+1)^2(\nu+3)}\mathbf{1}^T\boldsymbol{P}_z\mathbf{1}_n \tag{8.3.4}$$

8.3.1 Adjusting for $\boldsymbol{\beta}$ and ν

Let $\boldsymbol{\theta}^* = (\boldsymbol{\beta}, \nu)$. Following Chapter 5 Section 5.2.4, and noting the parameter orthogonalization, the SAPL for $\boldsymbol{\delta}$ can be expressed as

$$L_S(\boldsymbol{\delta}; \boldsymbol{y}) = \exp(k^*(\boldsymbol{\delta}))\big\{|\mathcal{I}_e(\hat{\boldsymbol{\theta}}^*_{\boldsymbol{\delta}}, \hat{\boldsymbol{\theta}}^*_{\boldsymbol{\delta}})|/|\mathcal{I}_o(\hat{\boldsymbol{\theta}}^*_{\boldsymbol{\delta}}, \hat{\boldsymbol{\theta}}^*_{\boldsymbol{\delta}})|\big\}^{1/2}\exp(\ell(\boldsymbol{\delta}, \hat{\boldsymbol{\theta}}^*_{\boldsymbol{\delta}}; \boldsymbol{y})) \tag{8.3.5}$$

where $\hat{\boldsymbol{\theta}}^*_{\boldsymbol{\delta}} = (\hat{\boldsymbol{\beta}}(\boldsymbol{\delta}, \hat{\nu}), \hat{\nu}(\hat{\boldsymbol{\beta}}, \boldsymbol{\delta}))$ are the maximum likelihood estimates for $\boldsymbol{\beta}$ and ν derived by iteratively solving (4.3.16) and (8.1.10) for given $\boldsymbol{\delta}$. The profile likelihood has the form

$$\exp(\ell(\boldsymbol{\delta}, \hat{\boldsymbol{\theta}}^*_{\boldsymbol{\delta}}; \boldsymbol{y})) = \left\{\frac{\Gamma((\hat{\nu}+1)/2)}{(\Gamma(1/2))\Gamma(\hat{\nu}/2)\hat{\nu}^{1/2}}\right\}^n |\boldsymbol{\Sigma}|^{-1/2}\prod_{i=n}^{n}\left\{1 + \frac{r_i^2}{\sigma_i^2\hat{\nu}}\right\}^{-(\hat{\nu}+1)/2}$$

where $r_i = (y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{\boldsymbol{\delta}})$. The two determinant terms are defined by

$$\mathcal{I}_e(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = \begin{bmatrix} \mathcal{I}_e(\boldsymbol{\beta}, \boldsymbol{\beta}) & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_e^*(\nu, \nu) \end{bmatrix} \tag{8.3.6}$$

$$\mathcal{I}_o(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = \begin{bmatrix} \mathcal{I}_o(\boldsymbol{\beta}, \boldsymbol{\beta}) & \mathcal{I}_o^*(\boldsymbol{\beta}, \nu) \\ \mathcal{I}_o^*(\nu, \boldsymbol{\beta}) & \mathcal{I}_o^*(\nu, \nu) \end{bmatrix} \tag{8.3.7}$$

Here, $\mathcal{I}_e(\boldsymbol{\beta}, \boldsymbol{\beta})$ and $\mathcal{I}_e^*(\nu, \nu)$ are defined by (8.3.2) and (8.3.4) respectively. As $\mathcal{I}_e^*(\nu, \nu)$ does not depend on $\boldsymbol{\delta}$, $\mathcal{I}_e(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = \mathcal{I}_e(\boldsymbol{\beta}, \boldsymbol{\beta})$ providing a moderate simplification to the SAPL (8.3.5).

Under the simplified scale parameter model the observed information for $\boldsymbol{\beta}$, $\mathcal{I}_o(\boldsymbol{\beta}, \boldsymbol{\beta})$, and adjusted observed information for ν, $\mathcal{I}_o^*(\nu, \nu)$, are given by (4.3.9) and (8.1.6) respectively. Differentiating (4.3.4) with respect to ν and recognising the parameter orthogonality the lth adjusted observed co-information term for $(\boldsymbol{\beta}, \nu)$ can be written as

$$\begin{aligned}
\mathcal{I}_o^*(\beta_l, \nu) =& \sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) x_{il}}{\sigma_i^2 \nu} \left\{ \left(\frac{\nu+1}{\nu} + \frac{2z_i^*}{\nu} - 1 \right) \left(\frac{1}{1 + d_i/\sigma_i^2 \nu} \right) \right. \\
& \qquad \left. - \left(\frac{\nu+1}{\nu} + \frac{2z_i^*}{\nu} \right) \frac{d_i/\sigma_i^2 \nu}{(1 + d_i/\sigma_i^2 \nu)^2} \right\} \\
=& \sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) x_{il}}{\sigma_i^2 \nu} \left\{ \left(\frac{\nu+1}{\nu} + \frac{2z_i^*}{\nu} - 1 \right) (1 - B_i) \right. \\
& \qquad \left. - \left(\frac{\nu+1}{\nu} + \frac{2z_i^*}{\nu} \right) B_i (1 - B_i) \right\} \\
=& \sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) x_{il}}{\sigma_i^2 \nu} \left\{ \left(\frac{\nu+1}{\nu} + \frac{2z_i^*}{\nu} \right) (B_i^2 - 2B_i + 1) + B_i - 1 \right\} \\
=& \sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) x_{il}}{\sigma_i^2 \nu} \left\{ \left(\frac{\nu+1}{\nu} + \frac{2z_i^*}{\nu} \right) (B_i - 1)(B_i - 1) + B_i - 1 \right\}
\end{aligned}$$

where $z_i^* = \boldsymbol{z}_i^T (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \mathbf{1}_n$ and B_i is defined by (4.3.6). Using Chapter 5 Section 5.2.4, $k^*(\boldsymbol{\delta})$ in the extra exponent term of (8.3.5) can be expressed as

$$k^*(\boldsymbol{\delta}) = (\boldsymbol{\delta} - \hat{\boldsymbol{\delta}})^T \boldsymbol{k}'(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\theta}}^*)$$

where

$$\boldsymbol{k}'(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\theta}}^*) = (\mathrm{tr}(\boldsymbol{A}_1), \ldots, \mathrm{tr}(\boldsymbol{A}_q))^T$$

and

$$\boldsymbol{A}_j = -\frac{1}{2} (\mathcal{I}_e(\hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\theta}}^*))^{-1} \mathcal{I}_e^{(\delta_j)}(\hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\theta}}^*) \Big|_{\boldsymbol{\delta} = \hat{\boldsymbol{\delta}}}$$

Here, $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\theta}}^*$ are the ML estimates derived by iteratively solving (4.3.18), (4.3.16) and (8.1.10) using the computational algorithm defined in Section 8.1.3. The first term of this expression is defined by (8.3.6) evaluated at the maximum likelihood estimates for $\boldsymbol{\delta}$ and $\boldsymbol{\theta}^*$. The second term can be derived by taking the derivative of (8.3.6) with respect to δ_j. In this particular case the expected information component for the degrees of freedom parameter does not contain $\boldsymbol{\delta}$ and therefore only the derivative of the information for the location parameters is required, namely

$$\mathcal{I}_e^{(\delta_j)}(\boldsymbol{\beta},\boldsymbol{\beta}) = \frac{\partial \mathcal{I}_e(\boldsymbol{\beta},\boldsymbol{\beta})}{\partial \delta_j} = -\frac{\nu+1}{\nu+3}\sum_{i=1}^{n}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\sigma_i^2}z_{ij}$$

and therefore

$$\mathrm{tr}(\boldsymbol{A}_j) = \frac{1}{2}\sum_{i=1}^{n} h_{ii}z_{ij}$$

where h_{ii} is the iith diagonal of $\boldsymbol{H} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1/2}$, the hat matrix. The final adjustment term can then be expressed as the exponent of

$$k^*(\boldsymbol{\delta}) = \frac{1}{2}(\boldsymbol{\delta}-\hat{\boldsymbol{\delta}})^T\boldsymbol{Z}^T\hat{\boldsymbol{h}}$$

where $\hat{\boldsymbol{h}} = (h_{11}(\hat{\boldsymbol{\delta}}),\ldots,\hat{h}_{nn}(\hat{\boldsymbol{\delta}})))$.

Typically, (8.3.5) defines a stably adjusted profile likelihood for $\boldsymbol{\delta}$, adjusted for the location parameters, $\boldsymbol{\beta}$, and the scale parameter, ν. This SAPL can be considered to be equivalent to an *extended* t-REML likelihood for $\boldsymbol{\delta}$.

8.3.2 Adjusting for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$

Let $\boldsymbol{\theta}^+ = (\boldsymbol{\beta},\boldsymbol{\delta})$. Proceeding identically to the previous section the SAPL for ν can be expressed as

$$L_S(\nu;\boldsymbol{y}) = \exp(k^*(\nu))\big\{|\mathcal{I}_e(\hat{\boldsymbol{\theta}}_\nu^+,\hat{\boldsymbol{\theta}}_\nu^+)|/\mathcal{I}_o(\hat{\boldsymbol{\theta}}_\nu^+,\hat{\boldsymbol{\theta}}_\nu^+)|\big\}^{1/2}\exp(\ell(\nu,\hat{\boldsymbol{\theta}}_\nu^+;\boldsymbol{y})) \tag{8.3.8}$$

where $\hat{\boldsymbol{\theta}}_\nu^+ = (\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}},\nu),\hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}},\nu))$ are the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ obtained by iteratively solving (4.3.16) and (4.3.18) respectively, for given ν. Here, the profile likelihood can be expressed as

$$\exp(\ell(\nu,\hat{\boldsymbol{\theta}}_\nu^+;\boldsymbol{y})) = \left\{\frac{\Gamma(\nu+1)/2)}{(\Gamma(1/2))\Gamma(\nu/2)\nu^{1/2}}\right\}^n|\hat{\boldsymbol{\Sigma}}|^{-1/2}\prod_{i=n}^{n}\left\{1+\frac{r_i^2}{\hat{\sigma}_i^2\nu}\right\}^{-(\nu+1)/2}$$

and the determinant terms can be immediately written as

$$\mathcal{I}_e(\boldsymbol{\theta}^+, \boldsymbol{\theta}^+) = \begin{bmatrix} \mathcal{I}_e(\boldsymbol{\beta}, \boldsymbol{\beta}) & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_e(\boldsymbol{\delta}, \boldsymbol{\delta}) \end{bmatrix} \tag{8.3.9}$$

$$\mathcal{I}_o(\boldsymbol{\theta}^+, \boldsymbol{\theta}^+) = \begin{bmatrix} \mathcal{I}_o(\boldsymbol{\beta}, \boldsymbol{\beta}) & \mathcal{I}_o(\boldsymbol{\beta}, \boldsymbol{\delta}) \\ \mathcal{I}_o(\boldsymbol{\delta}, \boldsymbol{\beta}) & \mathcal{I}_o(\boldsymbol{\delta}, \boldsymbol{\delta}) \end{bmatrix} \tag{8.3.10}$$

The expected information components $\mathcal{I}_e(\boldsymbol{\beta}, \boldsymbol{\beta})$ and $\mathcal{I}_e(\boldsymbol{\delta}, \boldsymbol{\delta})$ are defined by (8.3.2) and (8.3.3) respectively. Under the simplified scale model the observed information for $\boldsymbol{\beta}$ is given by (4.3.9). Using (4.3.11) and (4.3.12) the remaining observed information components can be expressed as

$$\mathcal{I}_o(\boldsymbol{\delta}, \boldsymbol{\delta}) = \frac{\nu + 1}{2} \sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^T \left\{ B_i - B_i^2 \right\}$$

$$\mathcal{I}_o(\boldsymbol{\beta}, \boldsymbol{\delta}) = (\nu + 1) \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{z}_i^T}{\sigma_i^2 \nu} \left\{ (1 - B_i)(1 - B_i) \right\} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})$$

Again, proceeding identically to the previous section, $k^*(\nu)$ in the extra exponent term can be expressed as

$$k^*(\nu) = (\nu - \hat{\nu}) k'(\hat{\nu}, \hat{\boldsymbol{\theta}}^+)$$

where

$$k'(\hat{\nu}, \hat{\boldsymbol{\theta}}^+) = -\frac{1}{2} \text{tr} \left((\mathcal{I}_e(\hat{\boldsymbol{\theta}}^+, \hat{\boldsymbol{\theta}}^+))^{-1} \mathcal{I}_e^{(\nu)}(\hat{\boldsymbol{\theta}}^+, \hat{\boldsymbol{\theta}}^+) \right) \Big|_{\nu = \hat{\nu}} \tag{8.3.11}$$

where $\hat{\nu}$ and $\hat{\boldsymbol{\theta}}^+$ are the ML estimates found by iteratively solving (8.1.10), (4.3.16) and (4.3.18) using the computational algorithm defined in Section 8.1.3. The second component of the trace term requires the derivative of the expected information of $\boldsymbol{\theta}^+$ with respect to ν, namely

$$\mathcal{I}_e^{(\nu)}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^+) = \begin{bmatrix} \mathcal{I}_e^{(\nu)}(\boldsymbol{\beta}, \boldsymbol{\beta}) & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_e^{(\nu)}(\boldsymbol{\delta}, \boldsymbol{\delta}) \end{bmatrix}$$

where

$$\mathcal{I}_e^{(\nu)}(\boldsymbol{\beta}, \boldsymbol{\beta}) = \frac{\partial \mathcal{I}_e(\boldsymbol{\beta}, \boldsymbol{\beta})}{\partial \nu} = \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_i^2} \left\{ \frac{2}{(\nu + 3)^2} - \frac{2 z_i^*}{\nu(\nu + 3)} \right\}$$

$$\mathcal{I}_e^{(\nu)}(\boldsymbol{\delta}, \boldsymbol{\delta}) = \frac{\partial \mathcal{I}_e(\boldsymbol{\delta}, \boldsymbol{\delta})}{\partial \nu} = \frac{3}{2(\nu + 3)^2} \boldsymbol{Z}^T \boldsymbol{Z}$$

and $z_i^* = \boldsymbol{z}_i^T (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \mathbf{1}_n$. Combining this with the first trace term in (8.3.11) and evaluating at the maximum likelihood estimate $(\hat{\nu}, \hat{\boldsymbol{\theta}})$ the adjustment can be expressed as

$$k^*(\nu) = \frac{1}{2}(\nu - \hat{\nu}) \left\{ \frac{2}{\hat{\nu}(\hat{\nu} + 1)} \hat{\boldsymbol{h}}^T \boldsymbol{P}_{\boldsymbol{z}} \mathbf{1}_n - \frac{2}{(\hat{\nu} + 1)(\hat{\nu} + 3)} \text{tr}(\hat{\boldsymbol{H}}) - \frac{3}{\hat{\nu}(\hat{\nu} + 3)} \text{tr}(\boldsymbol{P}_{\boldsymbol{z}}) \right\}$$

where $\hat{\boldsymbol{H}}$ is the hat matrix defined in the previous section, $\hat{\boldsymbol{h}}$ is a $n \times 1$ vector with ith element $h_{ii}(\hat{\boldsymbol{\delta}})$, and $\boldsymbol{P_z} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$.

Typically, (8.3.8) defines a stably adjusted profile likelihood for ν adjusted for the location parameters, $\boldsymbol{\beta}$ and the orthogonalized scale parameters, $\boldsymbol{\delta}$. This SAPL can be considered to be equivalent to an *extended* t-REML likelihood for ν.

8.3.3 Adjusting for $\boldsymbol{\delta}$ and ν

Let $\boldsymbol{\delta}^* = (\boldsymbol{\delta}, \nu)$. The SAPL for $\boldsymbol{\beta}$ adjusted for $\boldsymbol{\delta}^*$ can be expressed as

$$L_S(\boldsymbol{\beta}; \boldsymbol{y}) = \exp(h^*(\boldsymbol{\beta}))\big\{|\mathcal{I}_e(\hat{\boldsymbol{\delta}}^*_{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}^*_{\boldsymbol{\beta}})|/|\mathcal{I}_o(\hat{\boldsymbol{\delta}}^*_{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}^*_{\boldsymbol{\beta}})|\big\}^{1/2} \exp(\ell(\boldsymbol{\beta}, \hat{\boldsymbol{\delta}}^*_{\boldsymbol{\beta}}; \boldsymbol{y})) \tag{8.3.12}$$

where $\hat{\boldsymbol{\delta}}^*_{\boldsymbol{\beta}} = (\hat{\boldsymbol{\delta}}(\boldsymbol{\beta}, \hat{\nu}), \hat{\nu}(\boldsymbol{\beta}, \hat{\boldsymbol{\delta}}))$ are defined by iteratively solving the scoring equations (4.3.18) and (8.1.10) for a given $\boldsymbol{\beta}$. The expected information components for $\boldsymbol{\delta}$ and ν given by (8.3.3) and (8.3.4) respectively do not contain $\boldsymbol{\beta}$ and therefore their derivative with respect to $\boldsymbol{\beta}$, $\mathcal{I}_e^{(\boldsymbol{\beta})}(\boldsymbol{\delta}^*, \boldsymbol{\delta}^*) = \boldsymbol{0}$. Then $k^*(\boldsymbol{\beta}) = 0$ and the first exponent term is unity. Considering only the terms that contain $\boldsymbol{\beta}$ the SAPL for $\boldsymbol{\beta}$ can be expressed as

$$L(\boldsymbol{\beta}; \boldsymbol{y}) = |\mathcal{I}_o(\hat{\boldsymbol{\delta}}^*_{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}^*_{\boldsymbol{\beta}})|^{-1/2} \exp(\ell(\boldsymbol{\beta}, \hat{\boldsymbol{\delta}}^*_{\boldsymbol{\beta}}; \boldsymbol{y})$$

Noting (5.2.8), the observed information can be replaced by the expected information, which does not depend on $\boldsymbol{\beta}$, and therefore (8.3.12) reduces to

$$L_S(\boldsymbol{\beta}; \boldsymbol{y}) = \exp(\ell(\boldsymbol{\beta}, \hat{\boldsymbol{\delta}}^*_{\boldsymbol{\beta}}; \boldsymbol{y})) \tag{8.3.13}$$

the ordinary profile likelihood for $\boldsymbol{\beta}$ given the maximum likelihood estimates $\hat{\boldsymbol{\delta}}^*_{\boldsymbol{\beta}} = (\hat{\boldsymbol{\delta}}, \hat{\nu})$. The estimate for $\boldsymbol{\beta}$ from this is clearly just the maximum likelihood estimate obtained from the ML computation (8.1.3).

8.3.4 Computations

If the degrees of freedom is unknown the SAPLs defined by (8.3.5), (8.3.8) and (8.3.13) can be used as a sequential computational algorithm to determine an adjusted estimate for each of the parameters, $(\boldsymbol{\beta}, \boldsymbol{\delta}, \nu)$.

- **Estimation of $\boldsymbol{\beta}$:** The estimate for $\boldsymbol{\beta}$ can be obtained iteratively using the ML algorithm from Section 8.1.3. This algorithm also supplies the ML estimators for $\hat{\boldsymbol{\delta}}$ and $\hat{\nu}$ to be used in the subsequent steps.

- **Estimation of $\boldsymbol{\delta}$:** Let the two adjustment parameters be denoted by $\boldsymbol{\beta}_{\boldsymbol{\delta}}$ and $\nu_{\boldsymbol{\delta}}$. The adjusted estimate for $\boldsymbol{\delta}$ can be obtained iteratively where at the mth iteration the parameters are updated using

– For given $\boldsymbol{\delta} = \boldsymbol{\delta}^{(m)}$ and $\nu_{\boldsymbol{\delta}} = \nu^{(m)}$ update $\boldsymbol{\beta}_{\boldsymbol{\delta}}$ using $\boldsymbol{\beta}^{(m+1)} = f_t(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}^{(m)}, \nu^{(m)})$, where $f_t(\cdot)$ is given in (4.3.16).

– For given $\boldsymbol{\beta}_{\boldsymbol{\delta}} = \boldsymbol{\beta}^{(m+1)}$ and $\boldsymbol{\delta} = \boldsymbol{\delta}^{(m)}$ update $\nu_{\boldsymbol{\delta}}$ using $\nu^{(m+1)} = h_t(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\delta}^{(m)}, \nu^{(m)})$, where $h_t(\cdot)$ is given in (8.1.10).

– For given $\boldsymbol{\beta}_{\boldsymbol{\delta}} = \boldsymbol{\beta}^{(m+1)}$ and $\nu_{\boldsymbol{\delta}} = \nu^{(m+1)}$ update $\boldsymbol{\delta}$ using extended t-REML defined by (8.3.5), namely,

$$\boldsymbol{\delta}^{(m+1)} = \max\big\{\boldsymbol{\delta}; L(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\delta}^{(m)}, \nu^{(m+1)}; \boldsymbol{y})\big\}$$

- **Estimation of ν:** Let the two adjustment parameters be denoted by $\boldsymbol{\beta}_\nu$ and $\boldsymbol{\delta}_\nu$. The adjusted estimate for ν can be obtained iteratively where at the mth iteration the parameters are updated using

– For given $\boldsymbol{\delta}_\nu = \boldsymbol{\delta}^{(m)}$ and $\nu = \nu^{(m)}$ update $\boldsymbol{\beta}_\nu$ using $\boldsymbol{\beta}^{(m+1)} = f_t(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}^{(m)}, \nu^{(m)})$, where $f_t(\cdot)$ is given in (4.3.16).

– For given $\boldsymbol{\beta}_\nu = \boldsymbol{\beta}^{(m+1)}$ and $\nu = \nu^{(m)}$ update $\boldsymbol{\delta}_\nu$ using $\boldsymbol{\delta}^{(m+1)} = g_t(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\delta}^{(m)}, \nu^{(m)})$, where $g_t(\cdot)$ is given in (4.3.18).

– For given $\boldsymbol{\beta}_\nu = \boldsymbol{\beta}^{(m+1)}$ and $\boldsymbol{\delta}_\nu = \boldsymbol{\delta}^{(m+1)}$ update ν using extended t-REML defined by (8.3.8), namely,

$$\nu^{(m+1)} = \max\big\{\nu; L(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\delta}^{(m+1)}, \nu^{(m)}; \boldsymbol{y})\big\}$$

The sequential nature of the algorithm suggests that some or all of the steps may be interchanged. The algorithm requires the unconstrained ML estimates of the parameters and therefore the first step remains fixed, whereas, the estimation of $\boldsymbol{\delta}$ and ν in the last two steps may be interchanged.

It can be seen that the estimation algorithms for $\boldsymbol{\delta}$ and ν require an iterative ML estimation for the adjustment parameters. As (8.3.5) and (8.3.8) are numerically maximised, this requires that the ML procedures are nested inside the numerical routine. The ML procedure for the estimation of ν is also a numerical algorithm and therefore the estimation of $\boldsymbol{\delta}$ using (8.3.5) contains a nested numerical procedure. This complex non-linearity is likely to increase the instability of the algorithm. Similarly, the estimation of ν using (8.3.8) contains the iterative reweighted least squares procedures to obtain the ML estimates for $(\boldsymbol{\beta}, \boldsymbol{\delta})$. This is also likely to increase the instability of the numerical algorithm.

For the purpose of brevity the computational algorithm and the estimators obtained from it will be known as t-REML II with unknown degrees of freedom.

Chapter 9

Examples and Simulations

To illustrate the ML and approximate REML techniques derived in the previous chapter several examples are considered. The first example considers complex scale parameter models for the Alfalfa data. For this particular data, due to the inherent instability of the t-REML I and t-REML II methods, the analysis was restricted to ML. The second example presents the Stack Loss data and illustrates the difficulty with simple scale parameter modelling under t-REML I and t-REML II for small data sets. For some scale parameter models considered in this example t-REML II did not converge and, in all cases, t-REML I did not converge. Under ML, the final model is shown to be heteroscedastic in one of the covariates used in the location model. The final example, the Martin Marietta Data, presents an extensive comparison of the the estimation techniques derived for the heteroscedastic t with unknown degrees of freedom in this thesis. In particular, ML, t-REML I and t-REML II estimators for the location, scale and degrees of freedom parameters are compared, where possible, to the Gaussian equivalents. Due to the instability of the t-REML I and t-REML II algorithms profile surfaces are constructed for the scale parameters to ensure global parameter estimates were obtained. Furthermore, for comparison, the estimators and model fit from the new skew t distribution derived by Azzalini & Capitanio (2003) are compared to the ML estimators and model fit derived from the heteroscedastic t.

To understand the properties of the estimators for the heteroscedastic t for unknown degrees of freedom under ML, t-REML I and t-REML II a comparative simulation study is conducted in Section 9.2.

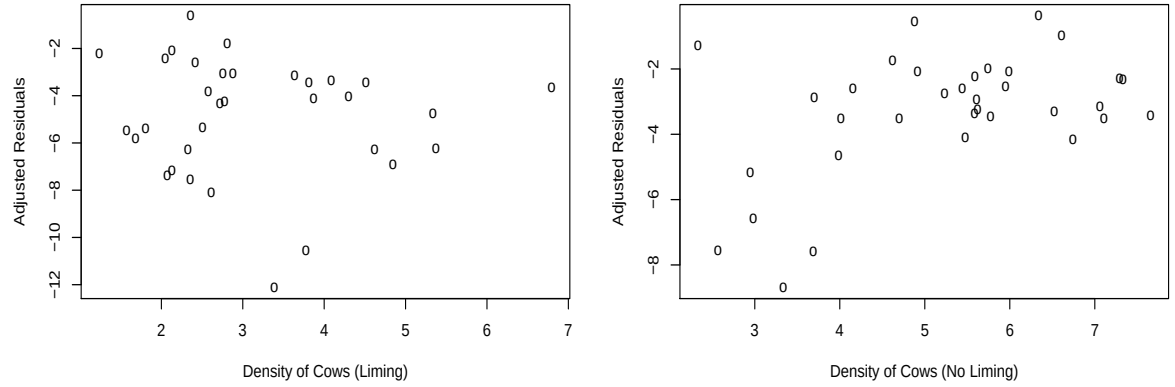

Figure 9.1: Scatter plots of the adjusted residuals against the square root of the density of dairy cows for the two liming types.

9.1 Examples

9.1.1 Rent for land Planted to Alfalfa: An ML example

The alfalfa data was introduced in Section 1.2.2 as a possible candidate data set for robust scale parameter modelling. To explore the scale parameter model the location parameter model is constructed first while allowing the degrees of freedom to vary according to the data. Section 1.2.2 and Figure 1.3 suggests that the increase in square root density of cows is vastly different for the two liming types. In addition, the square root density of the cows levels out for the land that is not limed. An appropriate location model would require an interaction between the square root density of cows and the liming type. This model is adopted here.

Assume a scale parameter model of the form (3.1.3). To understand whether the scale parameter is heteroscedastic (4.6.2) is used from Section 4.6. Under a homogeneous model, $\hat{\nu} = 5.37$ and Figure 9.1 shows the adjusted residuals, $\bar{d}_i/(1-h_{ii})^2 + 0.607$, from the fit of the saturated location model against the square root of the density of dairy cows for the two Liming types. The plots suggest that the non-homogeneity is different across the transformed cow densities for the two liming types. The plots also indicate that the heterogeneity for each liming type is possibly non-linear.

Table 9.1 presents the various scale parameter models considered for analyses along with their associated score test statistics and log-likelihood values. Hypothesis testing of t-specified models against the simpler Gaussian equivalent is difficult as the exact null distribution is a complex mixture of two chi-squared distributions. For the rest of this thesis a conservative approach is used where the null distribution is considered to be χ^2_1.

	Homogeneity Test			
	Score		*2LogL*	
Scale Parameter Model	Gaussian	t	Gaussian	t
1	-	-	29.94	32.68 (5.38)
SC	2.46	2.61	32.42	35.32 (4.92)
LI	5.10	5.97	35.52	38.96 (5.49)
$SC + LI$	5.15	5.99	35.93	39.11 (5.80)
$SC * LI$	7.68	8.76	43.51	44.17 (10.75)
$SC * LI + SC^2$	8.18	9.22	44.45	45.18 (9.11)
$(SC + SC^2) * LI$	8.55	9.63	44.87	45.65 (8.02)
$(SC + SC^2) * LI + SC^3$	9.38	10.45	52.69	56.98 (2.86)

Table 9.1: Homogeneity Test for the Alfalfa data. Estimated degrees of freedom for each t model are given in parentheses.

Thus assuming a constant scale parameter the model based on the t-distribution is not significantly different from the Gaussian equivalent. The addition of Liming type in the heteroscedastic t produces a log-likelihood ratio statistic of 6.28 on χ^2_1 and is therefore significant. A test under the Gaussian equivalent produces similar results. The score test concurs for both models. With a log-likelihood ratio statistic of 5.06 on χ^2_1 the table confirms that the interaction of square root density of cows with Liming type is significant for the heteroscedastic t and preferred over the simpler additive model. The score test produces a statistic of 8.76 on χ^2_3 and therefore also suggesting significance at the 5% level. Contrastly, for the Gaussian model, the score test indicates only marginal significance, whereas, the difference in log-likelihoods implies high significance. For this particular model the heteroscedastic t is not significantly different from the simpler Gaussian equivalent. The table also shows that the quadratic nature of the square root density of cows does not change for different liming types. However, for both the heteroscedastic Gaussian and t, the difference in log-likelihoods suggests retaining the quadratic form for the square root density of the cows for the two liming types and adding an overall cubic term to the scale model is highly significant. For this final model the heteroscedastic t is significantly different to the simpler Gaussian equivalent.

9.1.2 Stack-Loss Data

The stack loss data was introduced in Section 1.2.3 as a data set that requires robust analysis. Lange et al. (1989) shows, that under a t-distributed response, the distribution

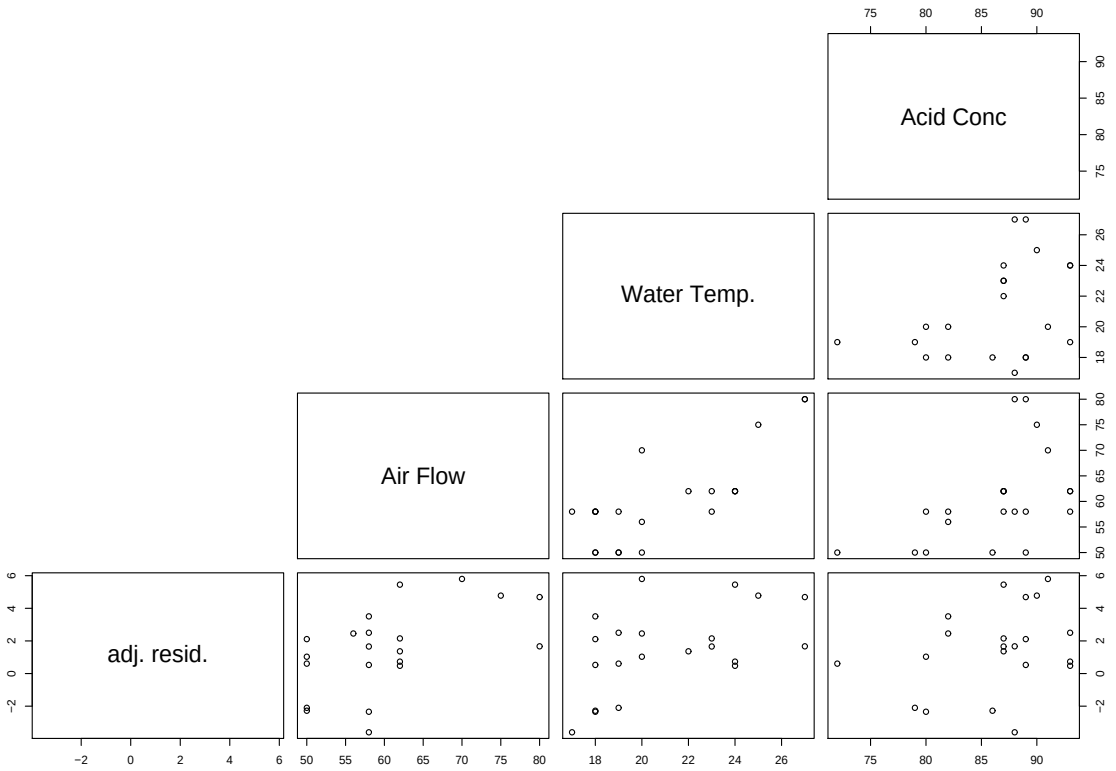

Figure 9.2: Pairs plot of the adjusted residuals from the homogeneous scale parameter model with the three explanatory variables.

of the residuals from an additive location model is heavy-tailed suggesting a small estimate for the degrees of freedom, $\hat{\nu} = 1.08$. Under ML, the likelihood associated with this particular homogeneous model demonstrated difficulty in obtaining convergence. One possible reason for this difficulty is the low degrees of freedom estimate may be close to a pole as discussed in Section 8.1.2 (also see Jones & Faddy, 2003 and Azzalini & Capitanio, 2003 for more details). For the homogeneous model, Fernandez & Steel (1999) show that this pole occurs at $\nu_0 = 0.615$, which is in close proximity to the estimated value of $\hat{\nu} = 1.08$.

To help identify possible heteroscedasticity in the scale parameter of the model the adjusted residuals defined by (4.6.2) are used from Section 4.6. As the estimate for the degrees of freedom is $\hat{\nu} < 2$, the estimated variance and the adjusted residuals are not defined for the t. It was suggested in Section 4.6 that (4.6.2) be replaced with the Gaussian derivation in Verbyla (1993). Fig. 9.2 presents a pairs plot of the Gaussian adjusted residuals, that is $\bar{d}_i/(1-h_{ii})^2 + 1.27406$, alongside the three regressors used additively in the location model. All three show a positive linear or non-linear trend and therefore a model that assumes heteroscedasticity would be preferable.

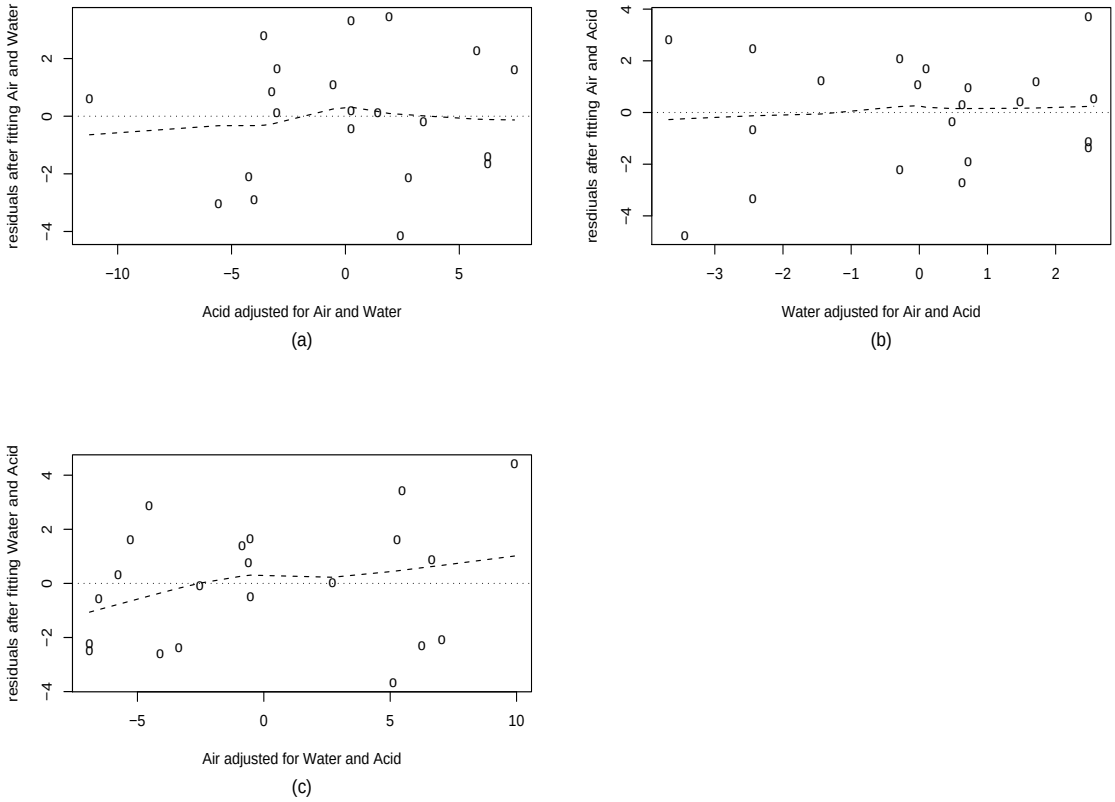

Figure 9.3: Detection of heteroscedasticity in the stack loss data(a) added variable plot for Acid; (b) added variable plot for Water; (c) added variable plot for Air.

Figure 9.3 shows the added variable plots for the three regressors. For Acid Concentration and Water Temperature the adjusted squared residuals appear to be constantly dispersed around zero, whereas Air Flow shows a positive trend in the adjusted squared residuals. This positive linearity suggests a simple Air Flow component as an initial covariate in the scale parameter model.

As the stack loss data contains only 21 observations the score test derived in Section 4.5.2 may be unreliable and is not used here. Table 9.2 presents the numerical likelihood for various ML heteroscedastic t models given a saturated additive model. The estimates for the degrees of freedom are given in parentheses. Under ML, the likelihood associated with the additive scale parameter model containing Water Temperature and Acid Concentration did not converge and therefore has been omitted from the table. Furthermore, for this particular model, the algorithm exhibited cyclicity as the estimated degrees of freedom approached low values ($\nu < 2$).

For comparison t-REML II log-likelihood values for the orthogonalized scale parameters are also given. The degrees of freedom estimate for each of these models is presented in parentheses. The congruence of the missing values for the additive Water Temperature

Homogeneity Test			
Variance Model	*-2LogL*		
	Gaussian	t-ML	t-REML II
1	104.6	99.14 (1.08)	102.2 (2.07)
AF	90.59	87.88 (2.90)	88.43 (2.96)
WT	101.4	94.25 (1.19)	89.64 (1.51)
AC	94.42	94.42 (∞)	94.42 (∞)
AF, WT	89.72	86.96 (2.33)	87.38 (+)
AF, AC	89.86	87.59 (2.70)	86.17 (2.45)
WT, AC	93.96	*	*
AF, WT, AC	89.26	86.75 (2.27)	86.77 (+)
AF, AF^2	80.69	80.69 (∞)	80.69 (∞)

Table 9.2: Homogeneity Test for the Stack Loss Data. Estimates for the degrees of freedom are given in parentheses. *'s represent a likelihood that would not converge; +'s represent a SAPL for ν that would not converge

and Acid Concentration scale parameter model is due to the inability to obtain ML estimates to substitute into the t-REML II algorithm. The SAPL associated with the degrees of freedom parameter from the t-REML II models that contained Air Flow and Water Temperature additively in the scale parameter model did not converge. Figure 9.4 shows the SAPL or profile likelihoods for the two models. A distinct jump discontinuity is present in the log-likelihood for low degrees of freedom values providing the cause for the convergence problems. As the SAPL for each model is an adjustment to the marginal profile likelihood in the presence of nuisance parameters associated with the location and scale parameter components of the model, it is highly likely that these discontinuities are related to the poles discussed in Section 8.1.2. t-REML I was also applied to the models in Table 9.2 and failed to converge in all cases.

For the constant scale parameter model the increase in the likelihood by estimating ν under ML is significant and therefore preferred over the simpler Gaussian location-scale model. Under ML, the addition of Air Flow to the scale parameter as a log-linear covariate produces a likelihood ratio statistic of 11.26 and therefore is very significant. Furthermore, the likelihood is also increased by the inclusion of a quadratic Air Flow component to the scale parameter model. The t-REML II model also concurs with this result. Table 9.2 also reveals that there is a possible linear Water Temperature component that may be added to the scale parameter model. However, Fig. 9.2 shows that this trend is correlated with Air Flow and this is verified by the added variable plot of the adjusted

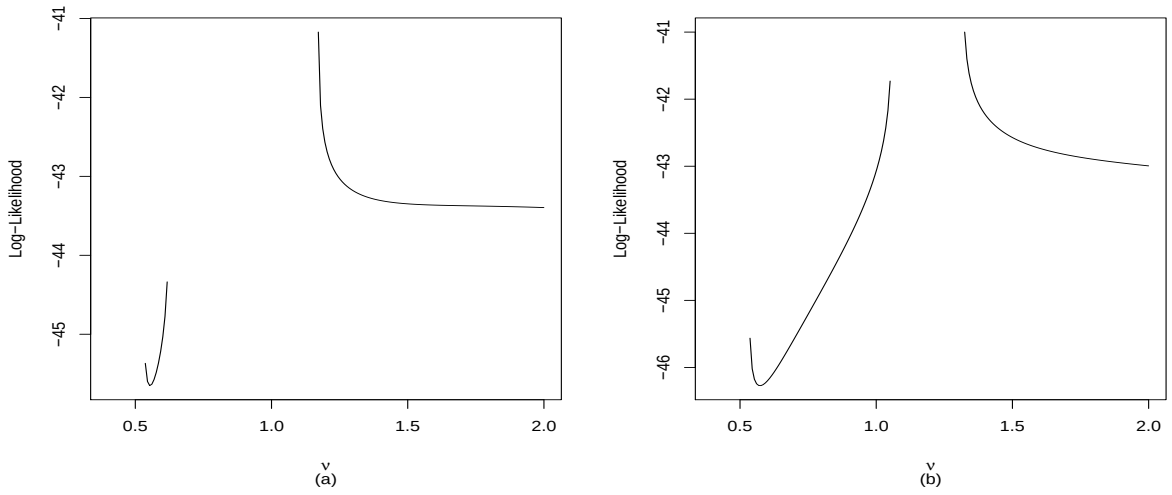

Figure 9.4: Stably Adjusted Profile Likelihoods of the degrees of freedom parameter for the scale parameter models (a) $AF + WT$; (b) $AF + WT + AC$

Water Temperature given in Fig. 9.3.

From Table 9.2 the addition of Air Flow at the linear and quadratic level is found to be not significantly different to the Gaussian location-scale model. Therefore for this particular data the Gaussian model is preferred over the t-ML and t-REML II models after accounting for the heteroscedasticity. The final fitted model is given in Figure 9.5.

9.1.3 Martin Marietta Data

The Martin Marietta data was introduced in Section 1.2.4. It was shown that after fitting a simple least squares process to the location component of the model the residuals are skewed to the right due to an influential outlier (see Figure 1.7).

Maximum Likelihood

Table 9.3 shows the ML estimates from fitting least squares (homoscedastic Gaussian) along with estimates from the heteroscedastic Gaussian, homoscedastic t and heteroscedastic t fits to the data. The location intercept parameter for all models is comparable. Conversely, the location slope parameter from least squares is greater than the compet-

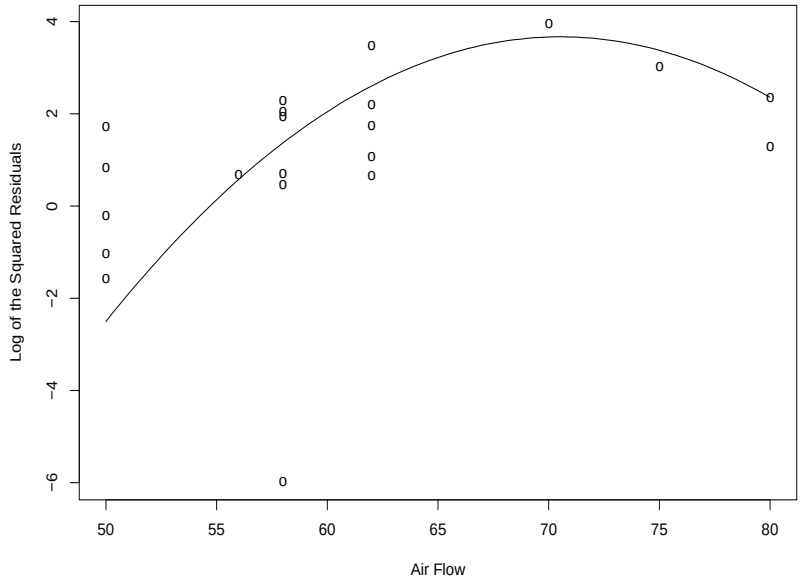

Figure 9.5: Log of the squared residuals from the Gaussian ML additive location model with constant scale against Air Flow; the fitted line for the quadratic Gaussian ML model is also displayed

ing models due to the influence of the outlier discussed earlier.

Assuming a constant scale parameter, the applicability of the homoscedastic t model under ML is tested. The null hypothesis is the homoscedastic Gaussian with a likelihood ratio statistic of 28.96 on χ^2_1, suggesting that the addition of the degrees of freedom parameter is warranted. It seems plausible that as the rate of the returns of the market increase the excess rate of returns of the company may become more variable. If the scale parameters are now defined by (3.1.3), then using (4.6.2) and the estimates from the homoscedastic t model, the data can be explored for heteroscedasticity. Figure 9.6 is a plot of the adjusted residuals from the left hand side of (4.6.2), that is $\log\{\bar{d}_i/(1-h_{ii})^2\} - \log\frac{1}{2} - \psi(\frac{1}{2}) + 1.241,\ i = 1, \ldots, n$, against the corresponding CRSP indexes. A local smoother was used as an exploratory tool to investigate the presence of a trend. A weak positive trend is evident in Figure 9.6. The addition of the slope parameter in the scale model yields a likelihood ratio test statistic of 3.11 on χ^2 with one degree of freedom (p-value = 0.0778) and is therefore marginally significant. To highlight the requirement of the heteroscedastic t model the inclusion of the degrees of freedom parameter is tested. In this case the heteroscedastic Gaussian ($\nu \to \infty$) model is the null hypothesis and the likelihood ratio statistic is 4.69. Again, this statistic has an asymptotic null distribution of χ^2_1, so the use of the heteroscedastic t-distribution appears justified.

Method	*2logL*	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\lambda}_0$	$\hat{\lambda}_1$	$\hat{\nu}$	$\hat{\alpha}$
Homo G	114.990	0.001	1.803	-4.754	-	-	-
		(0.012)	(0.282)	(0.183)	-	-	-
Hetero G	142.049	-0.009	1.319	-5.367	18.46	-	-
		(0.009)	(0.232)	(0.187)	(4.342)	-	-
Homo t	143.638	-0.007	1.264	-5.983	-	2.837	-
		(0.008)	(0.190)	(0.262)	-	(0.907)	-
Hetero t	146.740	-0.007	1.207	-5.960	13.12	3.759	-
		(0.008)	(0.206)	(0.250)	(5.823)	(1.442)	-
Skew t	146.166	-0.051	1.248	-5.450	-	3.320	1.246
		(0.023)	(0.190)	(0.289)	-	(1.430)	(0.653)

Table 9.3: ML estimates and their associated standard errors for the Martin Marietta data using various models. α refers to the skew parameter defined in Azzalini and Capitanio (2003).

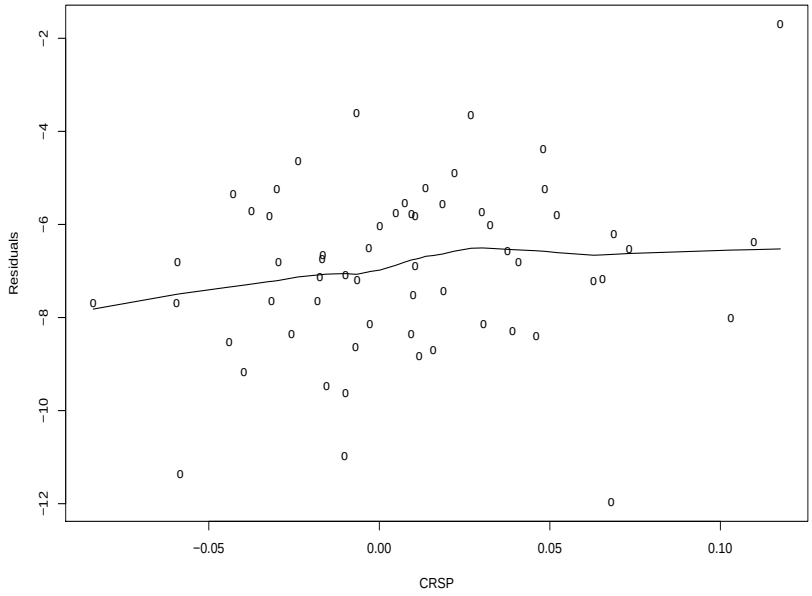

Figure 9.6: Heteroscedasticity plot for the adjusted residuals against the CRSP indexes. Fitted line is a local smoother to describe trend.

For comparison with the ML heteroscedastic t model, Table 9.3 also contains the skew t estimates from Azzalini & Capitanio (2003). Following Azzalini & Capitanio (2003), the estimated adjusted intercept can be expressed as

$$\hat{\beta}_0^* = \hat{\beta}_0 + \hat{\alpha}\left\{\frac{\hat{\nu}\hat{\sigma}^2}{\pi(1+\hat{\alpha}^2)}\right\}^{1/2}\frac{\Gamma\left(\frac{1}{2}(\hat{\nu}-1)\right)}{\Gamma\left(\frac{1}{2}\hat{\nu}\right)} = 0.0029$$

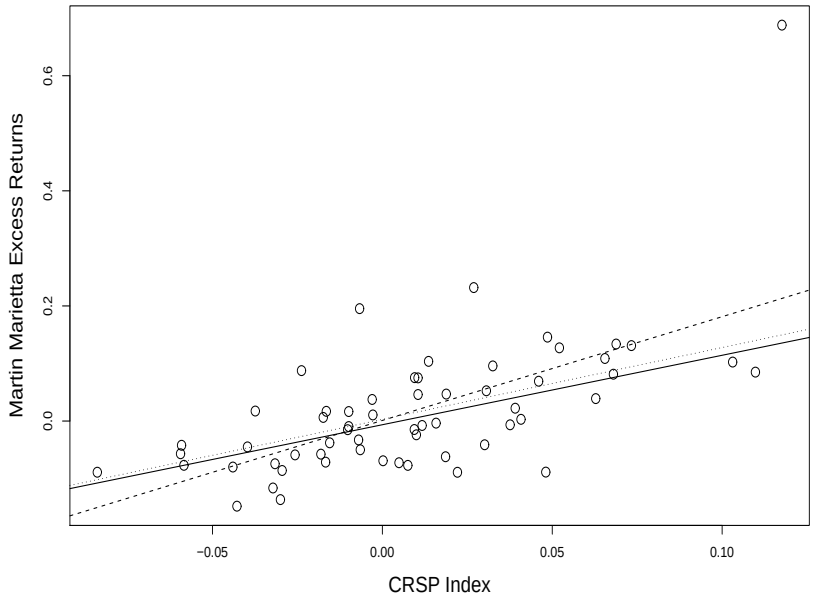

Figure 9.7: Scatter plot of the Martin Marietta company excess returns against the CRSP index for the whole market. Least squares line (dashed); skew t model fit (dotted); heteroscedastic t model fit (solid)

The fitted line for the skew t, along with the least squares line and the ML heteroscedastic t fit, are given in Figure 9.7. The fitted line for the heteroscedastic Gaussian is omitted due to its similarity to the fitted lines presented. Although the models and their estimates are not directly comparable, the concurrence of the skew t and heteroscedastic t fitted lines suggests these models provide similar location components given alternate specifications for the data.

Restricted Maximum Likelihood

Table 9.4 shows the REML estimates for the homoscedastic Gaussian, heteroscedastic Gaussian, homoscedastic t-REML I, heteroscedastic t-REML II, homoscedastic t-REML II and the heteroscedastic t-REML II fits to the data. Again, the location parameter intercepts are comparable. The increased value for the location slope parameter in the homoscedastic Gaussian model suggests that the outlier is still influential under REML. The estimated degrees of freedom values for t-REML II are comparable to the equivalent ML estimate, whereas, the t-REML I degrees of freedom estimate is significantly larger.

The numerical maximisation for heteroscedastic t-REML I and t-REML II showed instability while attempting to obtain estimates for $(\boldsymbol{\lambda}, \nu)$ or $(\boldsymbol{\delta}, \nu)$. For this reason, the log-likelihoods are investigated for possible anomalies. As the t-REML I likelihood, (8.2.1),

Method	*2LogL*	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\lambda}_0$	$\hat{\lambda}_1$	$\hat{\nu}$
Homo G	103.664	0.001	1.803	-4.721	-	-
		(0.012)	(0.289)	(0.189)	-	-
Hetero G	129.598	-0.001	1.310	-5.331	18.14	-
		(0.009)	(0.237)	(0.194)	(4.608)	-
Homo t-REML I	127.791	-0.006	1.301	-5.716	-	4.794
		(0.007)	(0.182)	(0.271)	-	(1.589)
Hetero t-REML I	133.941	-0.004	1.259	-5.561	16.73	12.01
		(0.008)	(0.211)	(0.311)	(5.362)	(13.50)
Homo t-REML II	$141.546_{\boldsymbol{\delta}}$	-0.007	1.264	-5.964	-	2.836
	143.579_{ν}	(0.008)	(0.190)	(0.231)	-	(1.037)
Hetero t-REML II	$145.402_{\boldsymbol{\delta}}$	-0.006	1.207	-5.905	14.64	3.582
	147.021_{ν}	(0.008)	(0.206)	(0.223*)	(NA)	(1.634)

Table 9.4: REML Estimates and their associated standard errors for the Martin Marietta data using various models. Standard errors are in parentheses. *'s represent standard errors of the associated orthogonal parameters, (δ_1, δ_2).

simultaneously maximises $(\boldsymbol{\lambda}, \nu)$, to investigate the log-likelihood surface for the scale parameters only the degrees of freedom is fixed at its maximum, $\hat{\nu} = 12.01$. The scale parameters are then profiled near the suggested maximum, $(\hat{\lambda}_0, \hat{\lambda}_1) = (-5.561, 16.73)$. Figure 9.8 shows a 3D perspective plot and contour plot of the surface of the log-likelihood for (λ_0, λ_1). Both plots suggest that the log-likelihood is convex around the maximum $\hat{\boldsymbol{\lambda}}$. For the slope scale parameter, λ_1 the surface exhibits flatness relative to the intercept parameter, λ_0. As the numerical maximisation uses a gradient method to estimate the parameters an incorrect step length across the flat surface may halt the algorithm prematurely. Therefore, adjustments to the arguments of the algorithm were required to ensure that a global maximum for $\boldsymbol{\lambda}$ had been obtained.

The t-REML II log-likelihood surface for the orthogonalized scale parameters is obtainable by profiling (8.3.5) near the suggested maximum $(\delta_1, \delta_2) = (-5.905, 14.64)$ and is given in Figure 9.9. This surface also shows convexity around $\hat{\boldsymbol{\delta}}$. Again, relative to the intercept scale parameter, the profile of the log-likelihood surface for the slope scale parameter is flat. Identical to t-REML I, adjustments to the arguments of the numerical maximisation procedure were required to ensure a global maximum had be obtained.

The unexpected bias in the estimates of the degrees of freedom under heteroscedastic t-REML I in comparison to the estimates obtained under ML and t-REML II, suggests a requirement to understand the log-likelihood surface near the applicable maximum for ν. For t-REML I the scale parameters are held at their maximum, $(\hat{\lambda}_0, \hat{\lambda}_1) = (-5.561, 16.73)$

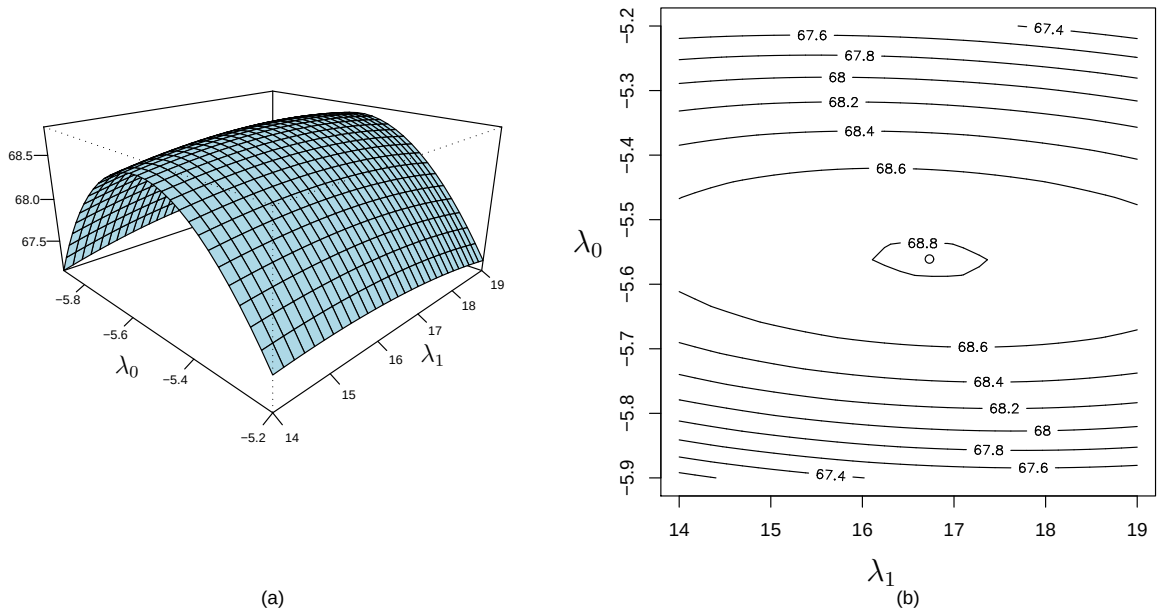

Figure 9.8: t-REML I log-Likelihood profile surface for the scale parameters (λ_0, λ_1) (a) 3D plot of the log-likelihood surface; (b) contour plot of the log-likelihood surface.

and the log-likelihood, (8.2.1) is profiled over a range of values for ν. For t-REML II, (8.3.8) is used to obtain a profile log-likelihood near the suggested maximum $\hat{\nu} = 3.582$. Figure 9.10 (a) and (b) shows the profile log-likelihood for heteroscedastic t-ML, t-REML I and t-REML II. The profile likelihood for t-REML I is not comparable to the two other profile likelihoods and therefore is displayed separately. The t-ML and t-REML II profile log-likelihoods for ν are very similar and show convexity near their associated maximum. The convexity of the profile log-likelihood for ν under t-REML I is less pronounced. Therefore the profile log-likelihoods for ν and the intercept scale parameter, λ_0, both exhibit flatness relative to the remaining parameter. Under t-REML I the approximate REML given by (8.2.1) simultaneously estimates the degrees of freedom and scale parameters and therefore this flatness may be a contributing factor to the instability of the non-linear algorithm for the heteroscedastic model considered here.

The standard errors of location parameter estimates for each of the approximate t-REML methods are comparable. The process to obtain scale parameter and degrees of freedom estimates using t-REML I and t-REML II is non-linear and therefore the standard errors for the scale and degrees of freedom parameter estimates obtained from these algorithms are obtained by inverting a numerical second derivative or hessian at the point where the associated likelihood was maximised. As expected, the standard errors for the scale parameter estimates for t-REML I and t-REML II, in comparison to Gaussian REML, show inflated values due to the inclusion of the degrees of freedom. At the maximum of the

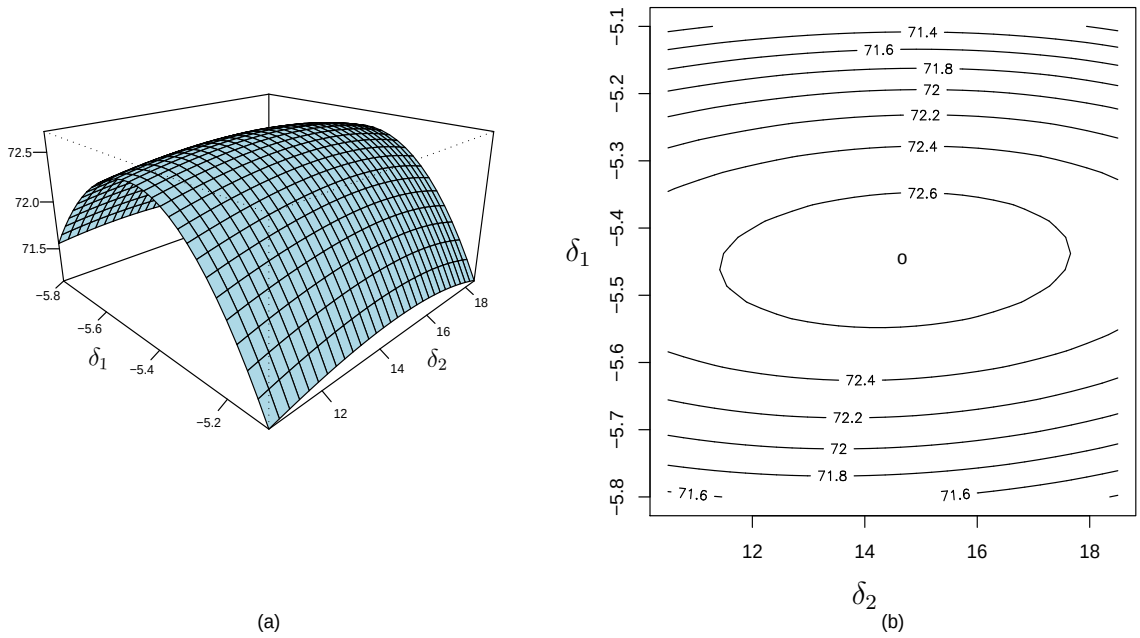

Figure 9.9: t-REML II log-Likelihood profile surface for the orthogonalized scale parameters (δ_1, δ_2) (a) 3D plot of the log-likelihood surface; (b) contour plot of the log-likelihood surface.

objective function, (8.3.5), for the heteroscedastic t-REML II model the numerical second derivative produced a negative value for δ_1. Therefore, upon inversion, the standard error for the slope scale parameter was not available. It is highly likely this is caused by the flatness of the surface for the slope scale parameter discussed previously.

As $\nu \to \infty$, t-REML I and t-REML II with unknown degrees of freedom do not approach the simpler heteroscedastic Gaussian REML Likelihood. Therefore testing of the inclusion of the degrees of freedom parameter under REML is a non-nested hypothesis. Although, the large log-likelihood ratio statistic in the ML case suggests that these tests are not necessary. The heteroscedasticity of the scale parameter can be tested for t-REML I and t-REML II. Under t-REML I the log-likelihood ratio statistic for the null hypothesis of $\lambda_1 = 0$ is 6.029 and therefore significant. Under t-REML II only the stably adjusted likelihood for $\boldsymbol{\delta}$ may be used. The log-likelihood ratio statistic for the null hypothesis of $\delta_1 = 0$ is 3.856 and is therefore, similar to the ML case, marginally significant at the 5% level.

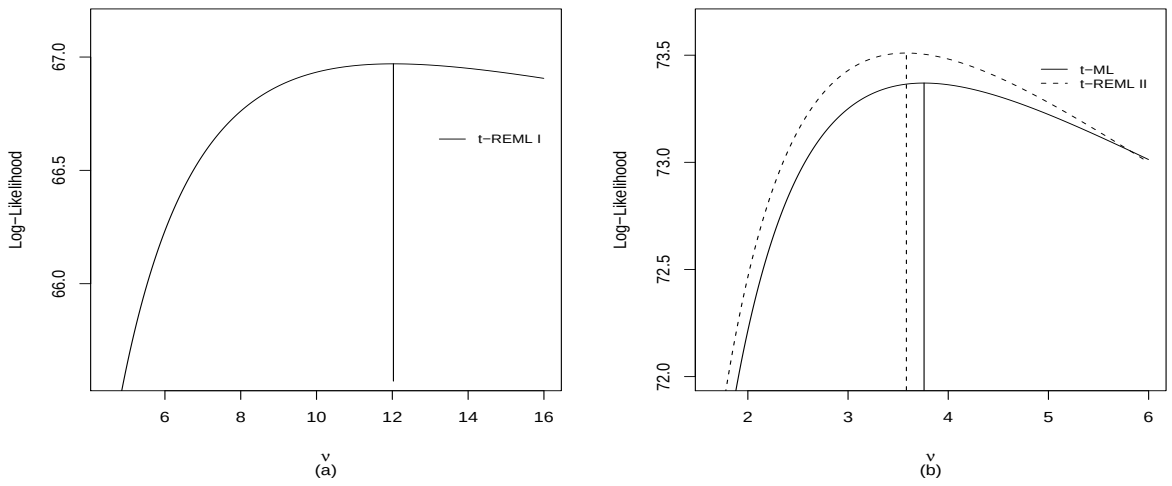

Figure 9.10: Profile log-likelihoods for ν (a) profile log-likelihood of ν for t-REML I; (b) profile log-likelihood of ν for t-ML and t-REML II

9.2 Simulation Study

The simulation study in this chapter mimics the simulation study in Chapter 7 with the exception that the degrees of freedom parameter of the heteroscedastic t-distribution is required to be estimated. The covariates for the location and scale components are defined by (7.2.1). The model (4.2.7) was used and the location and scale parameters follow the form defined by (7.2.2). The target values for the parameters are given by (7.2.3). In this particular set of simulations, for ML and t-REML II, the parameters have been orthogonalized using the derivations of Section 8.1.1. Thus, the estimates for the location, scale and degrees of freedom parameters, $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu)$ are obtained from ML, t-REML I and t-REML II using the computational algorithms from Sections 8.1.3, 8.2.1 and 8.3.4 respectively.

For this particular simulation study the target degrees freedom used were $\nu = (3, 5, 8)$. The simulation was run with sample sizes $n = (50, 100, 200)$ to gauge the effect on the properties of the parameter estimates for an increasing number of observations. A total of 500 replications of each combination of (ν, n) was obtained for the three approaches. In this particular case, for ML, the iterative scoring method defined in Section 8.1.3 was allowed an arbitrary maximum of 300 iterations to obtain convergence of the estimates. The numerical maximisation process used the parameters from t-REML I and t-REML

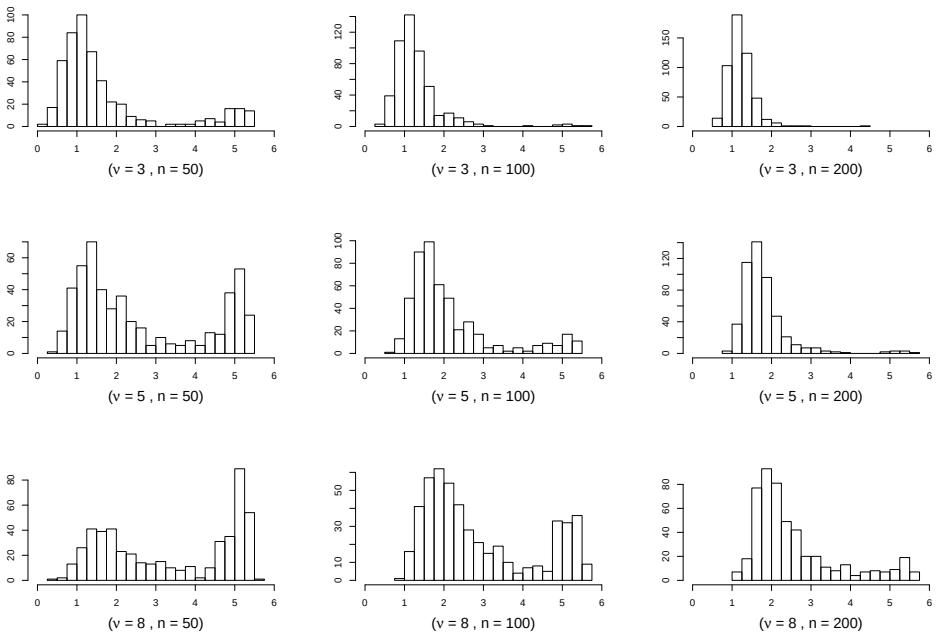

Figure 9.11: Histograms of the *log estimated degrees of freedom* parameter, ν, for 500 simulations under the distribution $y_i \sim t(\boldsymbol{x}_i^T\boldsymbol{\beta}, \exp(\boldsymbol{z}_i^T\boldsymbol{\lambda}), \nu),\ i = 1, \ldots, n$ where the target degrees of freedom is $\nu = (3, 5, 8)$, estimated using ML, and $n = (50, 100, 200)$.

II did not require a comparable mechanism. The convergence criterion in each case was $|\ell(\boldsymbol{\theta}^{(m+1)}) - \ell(\boldsymbol{\theta}^m)| < \epsilon$ where $\boldsymbol{\theta}$ is the parameter of interest and $\epsilon = 10^{-8}$.

For the ML case it was found that as the sample size is decreased and the target degrees of freedom increased the algorithm approaches maximum iterations more frequently. In addition, although the change in likelihood was negligible, in most of these cases the estimate for the degrees of freedom was tending to infinity. As an example of this bias, Figure 9.11 shows the histograms of the *log estimated degrees of freedom* obtained from all simulated combinations using ML. For a low target degrees of freedom and high sample size the distribution of the log estimated degrees of freedom is close to Gaussian. When the required degrees of freedom value is increased and the sample size is lowered the distribution shows bimodality indicating that a substantial proportion of the estimates are large values. Although t-REML I and t-REML II use numerical maximisation to obtain its estimates the extreme bias of the degrees of freedom parameter estimates was also visible. Figure 9.12 and Figure 9.13 show the histograms of the log estimated degrees of freedom obtained from all simulation combinations using t-REML I and t-REML II respectively. Under t-REML I, for the lowest target degrees of freedom and small sample sizes, the distribution of the log estimated degrees of freedom shows bimodality. This bimodality dissipates as the sample size is increased. For the higher degrees of freedom and small sample sizes this extreme skewness of the distributions of the log estimated

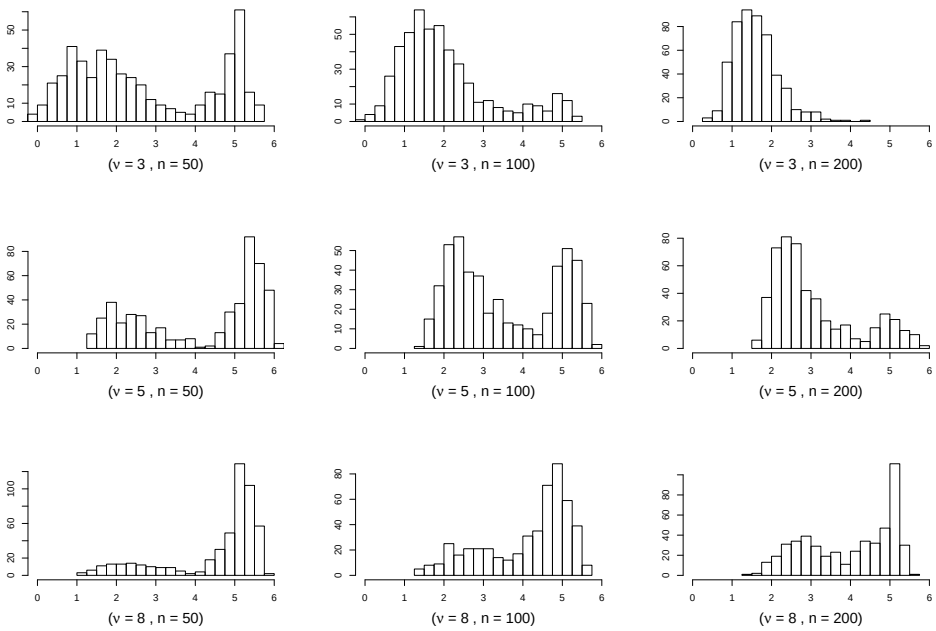

Figure 9.12: Histograms of the *log estimated degrees of freedom* parameter, ν, for 500 simulations under the distribution $y_i \sim t(\boldsymbol{x}_i^T\boldsymbol{\beta}, \exp(\boldsymbol{z}_i^T\boldsymbol{\lambda}), \nu),\ i = 1, \ldots, n$ where the target degrees of freedom is $\nu = (3, 5, 8)$, estimated using t-REML I, and $n = (50, 100, 200)$.

degrees suggests a large proportion of the simulations returned Gaussian estimates for the t-REML I parameters. In fact, for the extreme case $(\nu, n) = (8, 50)$, more than 85% of the simulations returned estimated degrees of freedom values > 16. This percentage is reduced to 75% when the sample size is increased to 200. Under t-REML II the bias of the estimated degrees of freedom parameter is also visible but less pronounced than bias obtained under ML or t-REML I. In particular, the skewness of the distributions is decreased for all sample sizes and degrees of freedom combinations suggesting that t-REML II estimated much lower values for the degrees of freedom than its ML equivalent. Due to the bias of the estimated degrees of freedom across all estimation methods the median value was chosen as an appropriate estimator for ν for each set of simulations.

Table 9.5 presents the means of the estimates for the fixed location and scale parameters and median estimates for degrees of freedom parameter over the 500 simulations for all sample sizes and degrees of freedom combinations. The table confirms the extreme bias of the degrees of freedom parameter exhibited in Figure 9.11, 9.12 and 9.13. Furthermore, the median estimates for the degrees of freedom under t-REML II are much less biased than the ML equivalents. Table 9.5 indicates that the location slope parameter for the has been efficiently estimated for all heteroscedastic t models suggesting excellent stability even for a small number of responses. For simulated models containing the lowest target degrees of freedom the intercept parameter for the location model is inconsistently

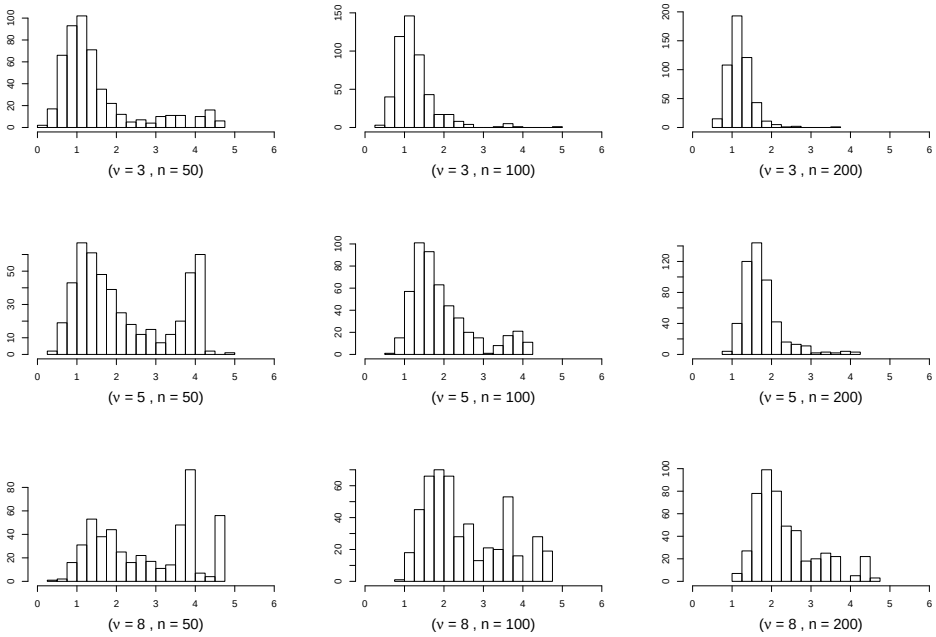

Figure 9.13: Histograms of the *log estimated degrees of freedom* parameter, ν, for 500 simulations under the distribution $y_i \sim t(\boldsymbol{x}_i^T\boldsymbol{\beta}, \exp(\boldsymbol{z}_i^T\boldsymbol{\lambda}), \nu),\ i = 1, \ldots, n$ where the target degrees of freedom is $\nu = (3, 5, 8)$, estimated using t-REML II, and $n = (50, 100, 200)$.

estimated for all sample sizes.

The simulation study conducted in Section 7.2 produced downwards biased estimates for the intercept parameter of the scale model under ML (see Table 7.4). Notably, the bias for the intercept scale parameter encountered in Table 9.5 is not as extreme as the bias encountered for the intercept scale parameter in Table 7.4. Also, in contrast to the empirical values for the intercept scale parameter in Table 7.4, this bias also increases as the degrees of freedom increases. This bias is alleviated as the sample size increases. The t-REML I estimates for the intercept scale parameter display extreme upwards bias. This bias decreases as the target degrees of freedom and the sample size is increased. Similar to ML, the t-REML II estimates for the intercept scale parameter display upwards bias for the smallest degrees of freedom and all sample sizes. This upwards bias dissipates as the degrees of freedom increases but is sporadic in comparison to ML estimates. The t-REML II estimates for the intercept scale parameter show upwards bias for all sample sizes and degrees of freedom combinations. In contrast to ML, this bias decreases as the degrees of freedom increases, but similar to ML, also decreases as the sample size increases.

Under ML, the slope scale parameter estimates exhibit slight upwards bias for all sample size and degrees of freedom combinations. In comparison, the t-REML I estimates for

		t-model ($\nu = 3$)			*t-model* ($\nu = 5$)			*t-model* ($\nu = 8$)		
	SS	50	100	200	50	100	200	50	100	200
t-ML	$\hat{\beta}_0$	-0.489	-0.510	-0.499	-0.509	-0.506	-0.507	-0.484	-0.498	-0.499
	$\hat{\beta}_1$	1.992	2.003	2.00	2.002	2.001	2.002	1.996	2.000	2.000
	$\hat{\lambda}_0$	0.495	0.506	0.515	0.464	0.502	0.498	0.436	0.491	0.496
	$\hat{\lambda}_1$	0.512	0.507	0.503	0.513	0.500	0.504	0.505	0.504	0.500
	$\hat{\nu}$	3.356	3.217	3.216	7.511	5.692	5.209	27.94	10.29	8.796
t-REML I	$\hat{\beta}_0$	-0.485	-0.513	-0.498	-0.509	-0.505	-0.507	-0.486	-0.498	-0.499
	$\hat{\beta}_1$	1.991	2.004	1.999	2.002	2.001	2.001	1.997	2.001	2.000
	$\hat{\lambda}_0$	0.903	0.869	0.839	0.755	0.757	0.748	0.638	0.673	0.68
	$\hat{\lambda}_0$	0.495	0.500	0.500	0.501	0.494	0.502	0.493	0.499	0.498
	$\hat{\nu}$	9.579	7.138	6.830	149.9	24.55	14.31	207.7	118.4	97.33
t-REML II	$\hat{\beta}_0$	-0.489	-0.510	-0.499	-0.509	-0.506	-0.507	-0.484	-0.498	-0.499
	$\hat{\beta}_1$	1.992	2.003	2.000	2.002	2.001	2.002	1.996	2.000	2.000
	$\hat{\lambda}_0$	0.591	0.545	0.531	0.545	0.537	0.513	0.502	0.521	0.512
	$\hat{\lambda}_1$	0.498	0.501	0.500	0.503	0.495	0.502	0.494	0.501	0.498
	$\hat{\nu}$	3.187	3.135	3.176	6.273	5.400	5.078	16.12	9.050	8.313

Table 9.5: Mean estimates for $\boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)$ and median estimates for ν under simulated distribution $y_i \sim t(\boldsymbol{x}_i^T\boldsymbol{\beta}, \exp(\boldsymbol{z}_i^T\boldsymbol{\lambda}), \nu),\ i = 1, \ldots, n$ using ML, t-REML I and t-REML II with unknown degrees of freedom with target values, $\nu = (3, 5, 8)$ and $n = (50, 100, 200)$

the slope scale parameter display negligible bias for all degrees of freedom and sample size combinations. For this particular simulation study, t-REML II is the most efficient at estimating the slope for the scale parameter model.

Identical to the simulation study conducted in Section 7.2 it is of interest to understand the bias reduction or increase of the estimated scale parameters obtained from the approximate REML techniques derived in this thesis in comparison to ML when the degrees is unknown. Figure 9.14 shows the 500 simulated empirical t-REML I estimates for the scale parameters against the ML equivalents for target degrees of freedom $\nu = (3, 5, 8)$ and $n = 50$. The extreme bias of the intercept scale parameter is prevalent for all target degrees freedom values. For the lowest target degrees of freedom, $\nu = 3$, nearly all the estimated values under t-REML I are above the required intercept value, $\lambda_0 = 0.5$. The simulations for which the estimated values for the intercept scale parameter are almost coincidental for the two methods ML and t-REML I, occurred when the degrees of freedom parameter was extremely upwards biased. Consequently, the number of coincidental points increased as the target degrees of freedom increased. As expected the slope for

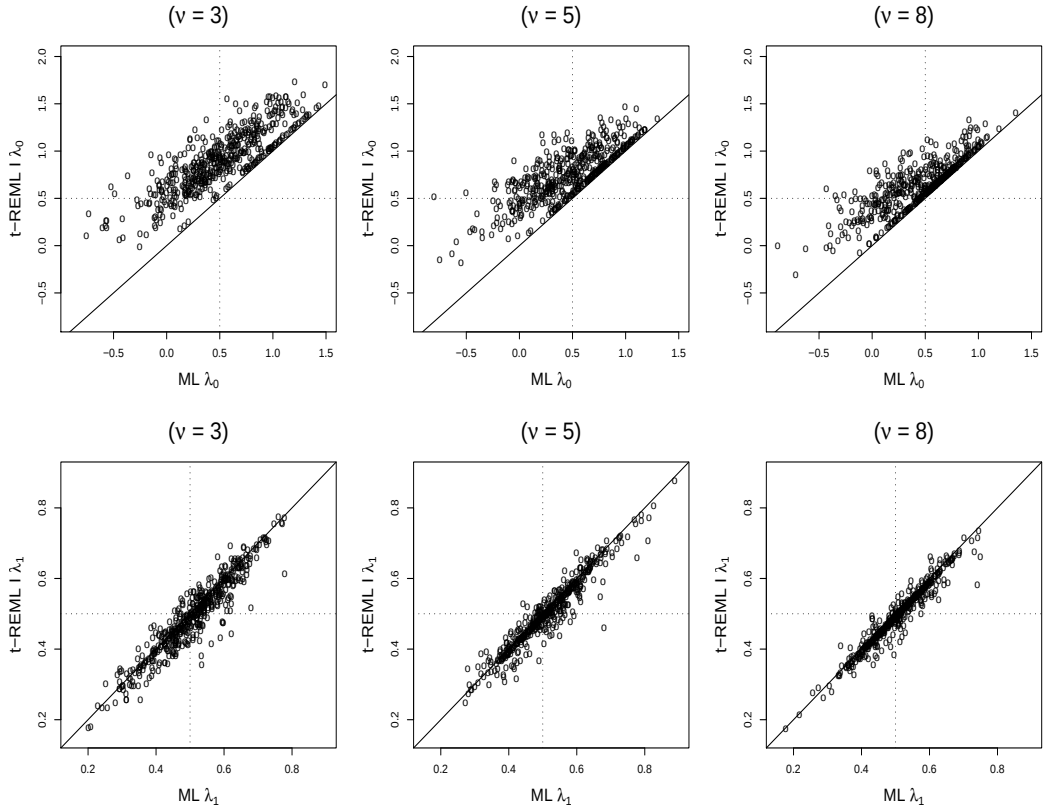

Figure 9.14: The t-REML I estimates of the scale parameters $(\hat{\lambda}_0, \hat{\lambda}_1)$ against the ML equivalents for 500 simulations under the distribution $y_i \sim t(\boldsymbol{x}_i^T\boldsymbol{\beta}, \exp(\boldsymbol{z}_i^T\boldsymbol{\lambda}), \nu),\ i = 1, \ldots, n$ where the target degrees of freedom is $\nu = (3, 5, 8)$ and $n = 50$.

the scale parameter model under t-REML I, in comparison to ML, is more consistent. Although, in a minor percentage of the simulations for each degrees of freedom combinations the t-REML I estimate for the slope parameter is downwards biased in comparison to ML.

Figure 9.15 displays the 500 simulated empirical t-REML II estimates against the ML equivalents for target degrees of freedom $\nu = (3, 5, 8)$ and $n = 50$. For estimates of the intercept scale parameter below the target value $\lambda_0 = 0.5$, t-REML II produces less biased estimates in comparison to ML. Similar to Figure 7.3, for ML estimates of the intercept scale parameter above the target value t-REML II increases this bias further. This upwards bias also notably decreases as the degrees of freedom increases. For empirical ML estimates of the slope scale parameter under the required value, $\lambda_1 = 0.5$, t-REML II slightly increases the downwards bias. For estimates over the target value, the majority of the estimates lay on the right hand side of the target line suggesting t-REML II provides slightly less biased results than ML for all degrees of freedom combinations used here.

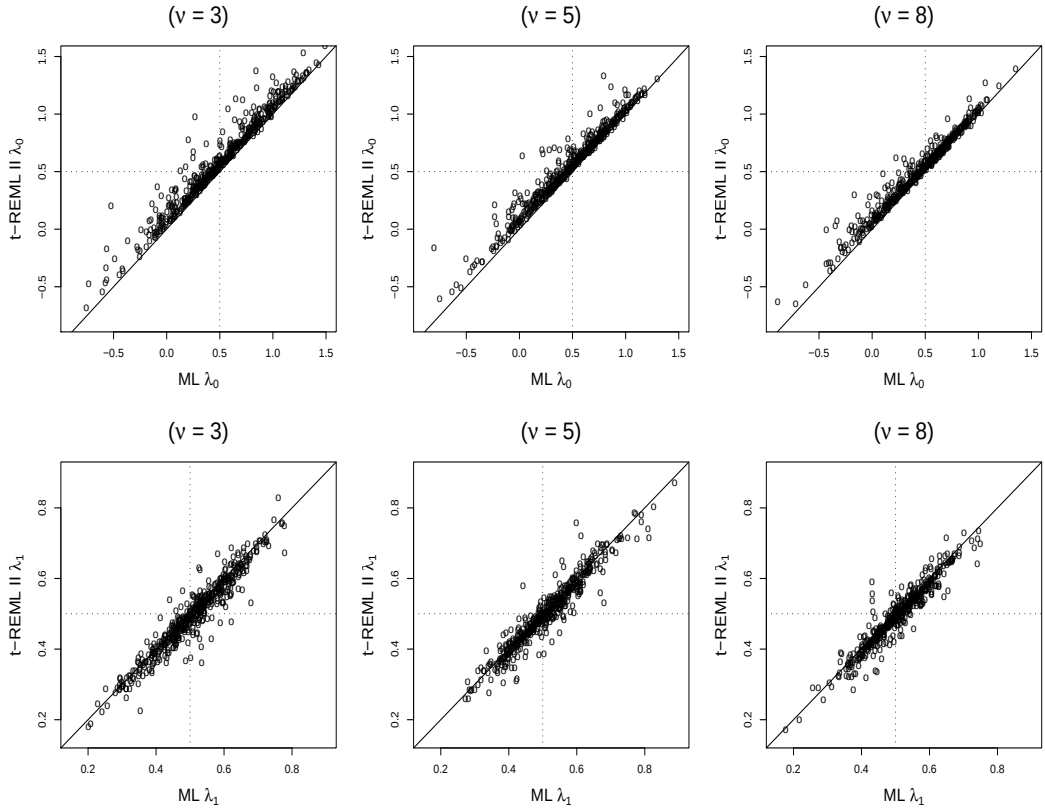

Figure 9.15: The t-REML II estimates of the scale parameters $(\hat{\lambda}_0, \hat{\lambda}_1)$ against the ML equivalents for 500 simulations under the distribution $y_i \sim t(\boldsymbol{x}_i^T\boldsymbol{\beta}, \exp(\boldsymbol{z}_i^T\boldsymbol{\lambda}), \nu)$, $i = 1, \ldots, n$ where the target degrees of freedom is $\nu = (3, 5, 8)$ and $n = 50$.

As the target values for the fixed scale parameter and degrees of freedom parameter are known, the theoretical standard errors for the location and scale parameters for ML were calculated and presented in Table 9.7. For comparison, the empirical standard errors of the location and scale parameter estimates using ML, t-REML I and t-REML II are obtained from the simulations and presented in Table 9.6. The standard errors of the estimated degrees of freedom parameter were omitted from both tables due to its extreme bias. As expected, both tables exhibit a trend of smaller standard errors for increasing degrees of freedom and sample sizes. In comparison to the empirical standard errors for all three methods, the asymptotic theoretical standard errors for the location parameters are slightly smaller. This difference is more prominent for low degrees of freedom and small sample sizes but diminishes as each of these are increased. In comparison to the empirical standard errors of the locations parameters from ML and t-REML II, for low degrees of freedom and small sample sizes, the the empirical standard errors for the location parameters under t-REML I are higher. As the sample size and degrees of freedom are increased this is reduced considerably.

		t-model($\nu = 3$)			*t-model* ($\nu = 5$)			*t-model*($\nu = 8$)		
	SS	50	100	200	50	100	200	50	100	200
t-ML	$\hat{\beta}_1$	0.219	0.146	0.111	0.205	0.146	0.107	0.190	0.145	0.099
	$\hat{\beta}_1$	0.082	0.057	0.039	0.077	0.054	0.038	0.071	0.054	0.036
	$\hat{\lambda}_0$	0.405	0.284	0.185	0.363	0.246	0.177	0.339	0.232	0.172
	$\hat{\lambda}_1$	0.104	0.075	0.052	0.098	0.064	0.045	0.084	0.06	0.041
t-REML I	$\hat{\beta}_0$	0.230	0.154	0.115	0.209	0.149	0.111	0.190	0.145	0.101
	$\hat{\beta}_1$	0.085	0.060	0.040	0.079	0.055	0.039	0.071	0.054	0.037
	$\hat{\lambda}_0$	0.325	0.247	0.164	0.272	0.197	0.156	0.258	0.185	0.135
	$\hat{\lambda}_1$	0.105	0.077	0.053	0.096	0.065	0.047	0.083	0.060	0.041
t-REML II	$\hat{\beta}_0$	0.219	0.146	0.111	0.205	0.146	0.107	0.190	0.145	0.099
	$\hat{\beta}_1$	0.082	0.057	0.039	0.077	0.054	0.038	0.071	0.054	0.036
	$\hat{\lambda}_0$	0.396	0.284	0.185	0.345	0.243	0.178	0.316	0.228	0.171
	$\hat{\lambda}_1$	0.105	0.075	0.052	0.099	0.065	0.046	0.085	0.061	0.041

Table 9.6: Empirical standard errors for $\hat{\boldsymbol{\theta}}^T = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\lambda}}^T)$ under simulated distribution $y_i \sim t(\boldsymbol{x}_i^T\boldsymbol{\beta}, \exp(\boldsymbol{z}_i^T\boldsymbol{\lambda}), \nu)$ using ML, t-REML I and t-REML II with unknown degrees of freedom and target values $\nu = (3, 5, 8)$ and $n = (50, 100, 200)$.

	t-model ($\nu = 3$)			*t-model* ($\nu = 5$)			*t-model* ($\nu = 8$)		
SS	50	100	200	50	100	200	50	100	200
β_0	0.199	0.144	0.103	0.188	0.136	0.097	0.180	0.130	0.093
β_1	0.077	0.055	0.039	0.072	0.052	0.037	0.069	0.050	0.035
λ_0	0.283	0.200	0.141	0.253	0.179	0.126	0.235	0.166	0.117
λ_1	0.096	0.069	0.049	0.086	0.061	0.044	0.080	0.057	0.040

Table 9.7: Theoretical asymptotic standard errors for $\boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)$ for the true simulated distribution $y_i \sim t(\boldsymbol{x}_i^T\boldsymbol{\beta}, \exp(\boldsymbol{z}_i^T\boldsymbol{\lambda}), \nu),\ i = 1, \ldots, n$, under ML.

The theoretical standard errors for the intercept scale parameter are much less than its empirical counterpart for all degrees of freedom and sample sizes under ML and t-REML II. This difference is reduced if the degrees of freedom or the sample size are increased. However, for the largest degrees of freedom used in this simulation study the difference is appreciable. Although the estimates for the intercept scale parameter using t-REML I are extremely biased, its empirical standard errors are comparative to its theoretical ML equivalent. This suggests that the estimation of the intercept scale parameter obtained

from t-REML I is problematic. In particular, the upwards bias may be caused by the extreme bias of the estimated degrees of freedom parameter as well as the simultaneous estimation of the degrees of freedom and scale parameters. The theoretical standard errors for the scale slope parameter under ML are slightly less than the corresponding empirical values for all three methods used here. For the simulations presented here the empirical standard errors for the scale slope parameter using t-REML II are marginally less than ML or t-REML I. The difference decreases as the sample size and the target degrees of freedom increase. In addition, it is interesting to note that the empirical standard errors for the location and scale parameters would increase further if the degrees of freedom parameter was less biased. This suggests that the empirical standard errors for the parameters of the heteroscedastic t-distribution with unknown degrees of freedom for all estimation methods considered here should be used with caution.

Chapter 10

Discussion and Conclusions

10.1 Discussion and Summary

10.1.1 Known degrees of freedom

When the degrees of freedom is known, the ML estimation method derived in Chapter 4 produces simplified scoring algorithms to estimate the location and scale parameters of the heteroscedastic t-distribution. These algorithms are slightly more computationally intensive than the Gaussian model and only require iterative weighted least squares procedures that is available in most commercial statistical software. The theory to detect heteroscedasticity and form asymptotic tests for the heteroscedastic t has also be generalised from the heteroscedastic Gaussian model presented in Verbyla (1993).

The t-REML I methodology derived in Section 6.1 of exploits the hierarchy of the heteroscedastic t-distribution. In particular, it exploits the form of the conditional Gaussian distribution required as a component of the integrand before the random scale effects have been integrated out. The approximate conditional likelihood returned from this method allows an implicit form for the location parameters, $\boldsymbol{\beta}$, to be derived which only requires an iterative reweighted least squares process. However, the remaining approximate REML used to estimate the scale parameters requires numerical maximisation due to the complexity of the extra determinant term from the Laplace approximation. This term also requires the calculation of a determinant of a $n \times n$ matrix and is therefore computationally cumbersome for large sample sizes.

As the degrees of freedom is known, the MPL used to derive t-REML II in Section 6.2 is simplified due to the existence of sufficient and ancillary statistics of the heteroscedastic t-distribution. The adjustment to the marginal profile likelihood is then similar to the adjustment required for the REML formulation under the heteroscedastic Gaussian

distribution.

Using a simplified log-linear scale parameter model the simulation study conducted in Section 7.2 reveals that, under ML, the intercept scale parameters are downwards biased. On average, both t-REML I and t-REML II estimators for the intercept scale parameters alleviated most of this bias. Graphical results of the simulations in Figure 7.3 revealed that when ML upwards biased an estimate for the intercept scale parameter t-REML I and t-REML II increased the bias further. Conversely, in comparison to the intercept scale parameter, the empirical estimates for the slope scale parameter are upwards biased under ML. Again, for all sample sizes and degrees of freedom used in this thesis t-REML I and t-REML II alleviated this bias. This alleviation can also be seen graphically in Figure 7.3. All three methods produce almost identical results for the location parameter estimates suggesting excellent stability for the location component of the model. The empirical standard errors for each method produced similar results for all parameters but larger than the ML theoretical counterpart.

10.1.2 Unknown degrees of freedom

When the degrees of freedom is unknown the ML procedure discussed in Section 8.1 requires the non-linear estimation of ν to be incorporated into the algorithmic process to estimate the parameters. To maintain the least squares algorithm for the parameters independence between the scale parameters and ν is required. This is achieved by considering an orthogonal transformation. Under a simplified log-linear scale parameter model, this orthogonal representation allows the estimation of the scale parameters to remain identical to ML estimation when the degrees of freedom is known. Although the new orthogonal parameters are a function of the degrees of freedom parameter, the score and Fisher information components for ν can be explicitly derived. This allows a simple scoring algorithm for ν to be achieved and independently incorporated into the estimation process with the location and scale parameters.

The Partial Laplace approximation (t-REML I) to the marginal likelihood when the degrees of freedom is unknown derived in Section 8.2 is similar to the approximation derived in Section 6.1. To complete the approximation the missing components of the Gamma kernel that contain ν are replaced. This does not affect the estimating equation for the location parameters. However, the approximate REML used to jointly estimate the parameters $(\boldsymbol{\lambda}, \nu)$ increases in complexity. The simulations show that, this minor increase in complexity has a substantial effect on its ability to efficiently estimate the scale and degrees of freedom parameters.

As the heteroscedastic t does not come from the location-scale family when the degrees of freedom is unknown it does not have explicit ancillary or sufficient statistics available.

The Stably Adjusted Profile Likelihood uses a modification of MPL that does not require ancillary statistics. This t-REML II method discussed in Section 8.3 requires three separate SAPLs to be maximised, one for each parameter adjusted for each set of nuisance parameters, but with the simplification that the location parameters only require estimation using ML. A further simplification occurs when the parameters are orthogonalized. The complexity of the estimation of $\boldsymbol{\delta}$ and ν from their respective SAPL is increased by the nuisance parameters containing the parameter of interest. Each of these requires non-linear maximisation that is readily available in statistical software.

The simulation study in Section 9.2 reveals that ML, t-REML I and t-REML II with unknown degrees of freedom are unstable for low sample sizes and moderate degrees of freedom. ML and t-REML II displayed similar but extreme bias in the estimates for ν. When the target degrees of freedom was high, ($\nu = 8$), many simulations tended to large values suggesting that the model was a heteroscedastic Gaussian. This suggests that the ML, t-REML I and t-REML II methods derived in this thesis could not numerically distinguish between a heteroscedastic Gaussian or t distribution for the response when the target degrees of freedom was high. This seems reasonable due to the negligible difference between the distributions when the degrees of freedom is increased. t-REML I also displayed bias in the degrees of freedom larger than ML or t-REML II. Under ML, the downwards bias of the estimated intercept scale parameter is not as extreme as the simulation study using the heteroscedastic t with known degrees of freedom and, contrastly, increases as the degrees freedom increases. t-REML II also displays similar values to ML but slightly upwards biased. t-REML I shows extreme upwards bias of the intercept scale parameter which is also revealed graphically in Figure 9.14 suggesting that this particular REML approximation to the heteroscedastic t-distribution should be used with caution. The slope for the scale parameter model has been efficiently estimated for all methods and combination of target degrees of freedom and sample sizes. As expected the location parameters have also been efficiently estimated for all methods. Although t-REML I produced the most bias in the estimated intercept scale parameters it produced the smallest empirical standard errors, in comparison to the remaining two methods, and the closest to the theoretical standard errors. This suggests that estimation of the intercept scale parameter associated with the t-REML I is problematic. In particular, the upwards bias asscoiated with thyis parameter may be caused by the simultaneous estimation of the degrees of freedom and the scale parameters and an extremely upwards biased estimate for ν. The empirical standard errors for the other parameters were similar for the three methods discussed in this simulation study.

10.2 Further Research

10.2.1 Link functions

The process to estimate the parameters of the heteroscedastic t-distribution when the degrees of freedom is unknown in Section 8.1.3 assumes that the parameters are mutually orthogonal. From Section 8.1.2 this orthogonal parameterization is only possible when the link function for the scale parameters is the natural log or the reciprocal. For link functions other than these (8.1.8) is not easily solvable. This may be overcome by jointly estimating $(\boldsymbol{\lambda}, \nu)$ using a non-linear maximiser. When the degrees of freedom is known $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ are mutually orthogonal and do not require reparameterizing. This ensures that the estimating equations (4.3.16) and (4.3.18) hold regardless of the link function and therefore allow for a very general scale parameter model. Furthermore, t-REML I and t-REML II with known degrees of freedom researched in this thesis also hold for general link functions of the scale parameter model.

10.2.2 Random scale effects

Assume a log-linear scale parameter model of the form (3.1.3). The random scale effects are required to have a χ^2_ν/ν distribution to ensure that the response is marginally distributed as a heteroscedastic t. As a component of this hierarchy the conditional Gaussian distribution has a scale parameter model that assumes the form $\log\varphi_i = \boldsymbol{z}_i^T\boldsymbol{\lambda} - \log\omega_i,\ i = 1, \ldots, n$. Therefore the random scale effects are naturally logged and are restrictive in some sense. A much broader approach would be to assume a more general class of random effects models for the scale parameter. For example, assume the random scale effects have a Gaussian distribution $\omega_i \sim N(0, \phi),\ i = 1, \ldots, n$ and are on the scale of the linear predictor for the scale parameter model. The conditional scale parameter model is

$$\log\varphi_i = \boldsymbol{z}_i^T\boldsymbol{\lambda} - \omega_i$$

and the marginal likelihood can be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}, \phi; \boldsymbol{y}) = \int_{\mathcal{R}^n} \prod_{i=1}^{n} p(y_i|\omega_i; \boldsymbol{\beta}, \boldsymbol{\lambda}) p(\omega_i; \phi) d\boldsymbol{\omega} \tag{10.2.1}$$

where

$$\prod_{i=1}^{n} p(\omega_i; \phi) = (2\pi\phi)^{-n/2} \exp\Big\{-\tfrac{1}{2}\sum_{i=1}^{n} \omega_i^2/\phi\Big\}$$

and $\prod_{i=1}^{n} p(y_i|\omega_i; \cdot)$ is defined by (4.2.5). From Section 3.2 the conditional Gaussian component of the integrand (10.2.1) can also be viewed as a likelihood for a Gamma

generalized linear model with response $d_i = (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$, location σ_i^2 and fixed scale parameter equal to 2. Therefore, although the Gamma component is restrictive, this hierarchy is defined as a Generalized Linear Mixed Model (GLMM). It is well known that the integral to obtain the marginal likelihood for these particular models is, in general, intractable and therefore requires approximation. One approximate approach is the Laplace approximation discussed in this thesis (see Wolfinger, 1993, Breslow & Clayton, 1993 and Breslow & Lin, 1995 or Chapter 5 Section 5.1). This connection between the heteroscedastic t and GLMMs suggest that the approximate REML approaches derived in this thesis may be applied to a class of Generalized Linear Mixed Models. In particular, if the conditional component of the hierarchy given in (10.2.1) is distributed as a Gaussian then partitioning of this component, identical to (6.1.5), is available regardless of the form of the marginal distribution chosen for the random effects. The Partial Laplace approximation can then be applied and an approximate conditional likelihood and an approximate REML with which to estimate location and scale parameters respectively can be derived. This is a subject for further research.

The random scale effects are predicted from the mean of the conditional distribution $\omega_i|y_i,\ i = 1, \ldots, n$. This distribution has a natural form given by (4.4.1) and is discussed in Section 4.4. The application of the Laplace approximation to the integral (4.2.4) requires estimates of the random effects obtained by maximising the pseudo joint likelihood or the components of the integrand. It was shown that choosing a different function of the random scale effects changes this maximisation. For the heteroscedastic t, the natural log scale, the scale on which the random effects appear linearly in the conditional component, produces estimates equivalent to the ones obtained from the conditional distribution $\boldsymbol{\omega}|\boldsymbol{y}$. For known degrees of freedom, the Laplace approximation to the marginal likelihood also reproduces the kernel for the location and scale parameters of the heteroscedastic t distribution. The Partial Laplace methodology derived in Section 6.1 used to obtain an approximate t-REML for the heteroscedastic t requires that only the component of the integrand (6.1.5) that is free of the location parameters be maximised to obtain the estimated random scale effects. It was shown in Section 6.1.4 that if the natural log scale for the random scale effects is chosen the approximate conditional likelihood to estimate the location parameters and the approximate t-REML to estimate the scale parameters differs from the separable likelihoods that are obtained if the scale of the random effects remains constant. This non-invariance will alter the estimates of the location and scale parameters. Further research is required to determine the differences and appropriateness of the approximate t-REMLs obtained in each case.

10.2.3 Other Miscellaneous Extensions

The Laplace and Partial Laplace approximation to the heteroscedastic t considered in this thesis is only first order. The Laplace approximation to the heteroscedastic t considered in Section 5.1.2 reproduces the kernel of the t when the degrees of freedom is known and therefore does not require higher order terms to be present. For the Partial Laplace approximation of the heteroscedastic t considered in Section 6.1, the integrand may be expanded using higher order terms and the accuracy of the approximation sharpened. Consequently, the approximate REML obtained from this approximation will be more accurate. This is a subject for further research.

When the degrees of freedom is unknown, the stably adjusted profile likelihoods derived in Section 8.3 rely on the parameter orthogonalization of Section 8.1.2. Section 5.2.4 and Barndorff-Nielsen & Cox (1994) discuss a first order linearisation to obtain the correction component, $k^*(\boldsymbol{\theta})$, that does not require orthogonal parameters to be available. Extensions of this are also available. Stern (1997) discusses a second order linearisation of an identical term that does not require parameter orthogonalization. For the SAPLs derived for the parameters $\boldsymbol{\delta}$ and ν of the heteroscedastic t, these two issues are a subject for further research.

Appendix A
Matrix Results

A.1 Introduction

This appendix gives definitions and derivations of matrix results that are connected directly to the content of this thesis. Popular references for such results can be found in Magnus & Neudecker (1988) and Lütkepohl (1996).

A.2 Determinant Results

Result A.2.1 *If $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ and $\boldsymbol{D}$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are non-singular,*

$$\begin{vmatrix} \boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{D} & \boldsymbol{B} \end{vmatrix} = |\boldsymbol{A}||\boldsymbol{B} - \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{C}| = |\boldsymbol{B}||\boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{D}|$$

Result A.2.2 *If $\boldsymbol{A}$ is $p \times p$ and $\boldsymbol{B}, \boldsymbol{C}$ are $n \times p$ and $p \times n$ respectively then using*

$$|\boldsymbol{A} + \boldsymbol{B}\boldsymbol{C}| = |\boldsymbol{A}||\boldsymbol{I}_p + \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{C}| = |\boldsymbol{A}||\boldsymbol{I}_n + \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}|$$

A.3 Inverse Results

Result A.3.1 *If $\boldsymbol{A}$ is $n \times n$, $\boldsymbol{B}$ is $p \times p$ and $\boldsymbol{C}$, $\boldsymbol{D}$ are $n \times p$ and $p \times n$ matrices respectively then*

$$(\boldsymbol{A} + \boldsymbol{C}\boldsymbol{B}\boldsymbol{D})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{C}(\boldsymbol{B}^{-1} + \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{C})^{-1}\boldsymbol{D}\boldsymbol{A}^{-1} \tag{A.3.1}$$

Proof: This can be verified by multiplying the LHS of (A.3.1) with the inverse of the RHS and checking for the identity,

$$\begin{aligned} \boldsymbol{I} &= \big(\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{C}(\boldsymbol{B}^{-1} + \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{C})^{-1}\boldsymbol{D}\boldsymbol{A}^{-1}\big)(\boldsymbol{A} + \boldsymbol{C}\boldsymbol{B}\boldsymbol{D}) \\ &= \boldsymbol{I} + \boldsymbol{A}^{-1}\boldsymbol{C}\big(\boldsymbol{B}^{-1} + \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{C}\big)^{-1}\big((\boldsymbol{B}^{-1} + \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{C})\boldsymbol{B}\boldsymbol{D} - \boldsymbol{D} - \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{B}\boldsymbol{D}\big). \end{aligned}$$

Expanding the last term in parentheses ensures that the RHS is the identity. Other useful identities can be formed from this result.

Result A.3.2 *Using the same matrices as result A.3.1, then*

$$\boldsymbol{B}\boldsymbol{D}(\boldsymbol{A} + \boldsymbol{C}\boldsymbol{B}\boldsymbol{D})^{-1} = (\boldsymbol{B}^{-1} + \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{C})^{-1}\boldsymbol{D}\boldsymbol{A}^{-1}$$

Proof: Using result A.3.1 and pre-multiplying the LHS and the RHS by $\boldsymbol{B}\boldsymbol{D}$ proves the result.

Result A.3.3 *Using the same matrices as A.3.1,*

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{D} & \boldsymbol{B} \end{bmatrix}^{-1} = \begin{bmatrix} (\boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{D})^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{C}(\boldsymbol{B} - \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{C})^{-1} \\ -\boldsymbol{B}^{-1}\boldsymbol{D}(\boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{D})^{-1} & (\boldsymbol{B} - \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{C})^{-1} \end{bmatrix} \quad \text{(A.3.2)}$$

Proof: Post-multiplying the LHS of (A.3.2) by $\boldsymbol{I}_{n+p}$ and simultaneously solving produces the desired result.

A.4 Distributional Matrix Results

Result A.4.1 *If $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and for any non-singular matrix $\boldsymbol{D}$,*

$$E(\boldsymbol{y}^T\boldsymbol{D}\boldsymbol{y}) = tr(\boldsymbol{\Sigma}\boldsymbol{D}) + \boldsymbol{\mu}^T\boldsymbol{D}\boldsymbol{\mu}$$

and

$$Var(\boldsymbol{y}^T\boldsymbol{D}\boldsymbol{y}) = tr\big((\boldsymbol{\Sigma}\boldsymbol{D})^2\big) + 4\boldsymbol{\mu}^T\boldsymbol{D}\boldsymbol{\mu}$$

Result A.4.2 *If $\boldsymbol{y}_1 \sim N(\boldsymbol{\mu}_1, \Sigma_{11})$ and $\boldsymbol{y}_2 \sim N(\boldsymbol{\mu}_2, \Sigma_{22})$ and $Cov(\boldsymbol{y}_1, \boldsymbol{y}_2) = \Sigma_{12}$ then*

$$\boldsymbol{y}_1|\boldsymbol{y}_2 \sim N\big(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}\big)$$

A.5 Miscellaneous Matrix Results

Result A.5.1 *If $\boldsymbol{H}$ is non-singular and $\boldsymbol{L}_2^T\boldsymbol{X} = \boldsymbol{0}$ then*

$$\boldsymbol{H} - \boldsymbol{H}\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

Proof: Using the inverse symmetric square root of $\boldsymbol{H}$, $\boldsymbol{X}$ can be transformed to $\boldsymbol{X}^* = \boldsymbol{H}^{-1/2}\boldsymbol{X}$. As $\boldsymbol{L}_2^T\boldsymbol{X} = \boldsymbol{0}$ the orthogonal complement to the column space of $\boldsymbol{X}^*$, $R(\boldsymbol{X}^*)$ is the column space of $\boldsymbol{L}^* = \boldsymbol{H}^{1/2}\boldsymbol{L}_2$. Simple orthogonal projections provide

$$\begin{aligned}\boldsymbol{L}_2^*(\boldsymbol{L}_2^{*T}\boldsymbol{L}_2^*)^{-1}\boldsymbol{L}_2^{*T} &= \boldsymbol{I} - \boldsymbol{X}^*(\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T} \\ \boldsymbol{H}^{1/2}\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\boldsymbol{H}^{1/2} &= \boldsymbol{I} - \boldsymbol{H}^{-1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1/2}\end{aligned}$$

Post and pre-multiplication of both sides by $\boldsymbol{H}^{1/2}$ completes the proof.

Result A.5.2 *If $\boldsymbol{H}$ is non-singular and $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}$ then $\boldsymbol{P}\boldsymbol{H}\boldsymbol{P} = \boldsymbol{P}$.*

Proof:

$$\begin{aligned}\boldsymbol{P}\boldsymbol{H}\boldsymbol{P} &= \boldsymbol{P}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}) \\ &= \boldsymbol{P} - \boldsymbol{P}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1} \\ &= \boldsymbol{P} \qquad (\text{as } \boldsymbol{P}\boldsymbol{X} = \boldsymbol{0})\end{aligned}$$

Result A.5.3 *If $\boldsymbol{H}$ is non-singular and $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}$ then*

$$\frac{\partial \boldsymbol{P}}{\partial \theta_i} = -\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}$$

Proof:

$$\begin{aligned}\frac{\partial \boldsymbol{P}}{\partial \theta_i} &= -\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_i\boldsymbol{H}^{-1} + \boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_i\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1} \\ &- \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_i\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1} \\ &+ \boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_i\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1} \\ &= -(\boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}) \\ &\times \dot{\boldsymbol{H}}_i(\boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}) \\ &= -\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\end{aligned}$$

Result A.5.4 *If $\boldsymbol{H}$ is non-singular and $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}$ then*

$$\boldsymbol{P} = \boldsymbol{S} - \boldsymbol{S}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z} + \boldsymbol{G})^{-1}\boldsymbol{Z}^T\boldsymbol{S}$$

where $\boldsymbol{S} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}^{-1}$

Proof: If $\boldsymbol{L}^T\boldsymbol{X} = \boldsymbol{0}$ then by Result A.5.1

$$\begin{aligned}\boldsymbol{P} &= \boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{H}\boldsymbol{L})^{-1}\boldsymbol{L}^T \\ &= \boldsymbol{L}(\boldsymbol{L}^T(\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)\boldsymbol{L})^{-1}\boldsymbol{L}^T \\ &= \boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{R}\boldsymbol{L} + \boldsymbol{L}^T\boldsymbol{Z}\boldsymbol{G}^{-1}\boldsymbol{Z}^T\boldsymbol{L})^{-1}\boldsymbol{L}^T \\ &= \boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{R}\boldsymbol{L})^{-1}\boldsymbol{L}^T - \boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{R}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{R}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1} \\ &\qquad \times \boldsymbol{Z}^T\boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{R}\boldsymbol{L})^{-1}\boldsymbol{L}^T \\ &= \boldsymbol{S} - \boldsymbol{S}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z} + \boldsymbol{G})^{-1}\boldsymbol{Z}^T\boldsymbol{S}\end{aligned}$$

as $\boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{R}\boldsymbol{L})^{-1}\boldsymbol{L}^T = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}^{-1} = \boldsymbol{S}$.

Result A.5.5 *If $\boldsymbol{C}$ is the coefficient matrix defined in (2.2.8) then its determinant can be written as*

$$|\boldsymbol{C}| = |\boldsymbol{R}|^{-1}|\boldsymbol{G}|^{-1}|\boldsymbol{H}||\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}|$$

Proof: Using Result A.2.1

$$\begin{aligned}
|\boldsymbol{C}| &= |\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}||\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} - \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}| \\
&= |\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}||\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}| \qquad \text{(using Result A.3.1)} \\
&= |\boldsymbol{G}|^{-1}|\boldsymbol{R}|^{-1}|\boldsymbol{H}||\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}| \qquad \text{(using Result A.2.2)}
\end{aligned}$$

Note using Result A.2.1 the determinant can also be written as

$$|\boldsymbol{C}| = |\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}||\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z} + \boldsymbol{G}^{-1}|$$

where $\boldsymbol{S} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}^{-1}$

Appendix B
Hett Documentation

`dof.profile` *Internal profile likelihood function*

Description

Internal profile log-likelihood function for `tlm` function

Usage

```
dof.profile(dof, n, sqResid, orthoI, X, Z)
```

Arguments

`dof`	degrees freedom value
`n`	number of values in the response vector
`sqResid`	squared residuals
`orthoI`	orthogonalized scale parameters
`X`	design matrix of explanatory variables for the location model
`Z`	design matrix of explanatory variables for the scale model

Details

This function is *not* intended to be directly called by users.

Value

a profile log-likelihood value

`mm` *Excess returns for Martin Marietta company*

Description

Data from the Martin Marietta company collected over a period of 5 years on a monthly basis

Usage

```
data(mm)
```

Format

A data frame with 60 observations on the following 4 variables.

date the month the data was collected

am.can a numeric vector

m.marietta excess returns from the Martin Marietta company

CRSP an index for the excess rate returns for the New York stock exchange

Source

Bulter et al (1990). Robust and partly adaptive estimation of regression models. *Review of Economic Statistics*, **72**, 321-327.

Examples

```
data(mm, package = "hett")
attach(mm)
plot(CRSP, m.marietta)
lines(CRSP, fitted(lm(m.marietta ~ CRSP)), lty = 2)
```

`rent` *Rent for Land PLanted to Alfalfa*

Description

Dataset collected in 1977 from Minnesota to study the variation in land rented for growing alfalfa

Usage

```
data(rent)
```

Format

A data frame with 67 observations on the following 5 variables.

Rent a numeric vector average rent per acre.

AllRent a numeric vector describing average rent paid for all tillable land.

Cows a numeric vector describing the density of dairy cows (number per square mile).

Pasture a numeric vector describing the proportion of farmland used as pasture.

Liming a factor with levels `No` if no liming is required to grow alfalfa and `Yes` if it does.

Source

Weisberg, S (1985). *Applied Linear Regression* Wiley: New York

Examples

```
library(lattice)
data(rent, package = "hett")
attach(rent)
xyplot(log(Rent/AllRent) ~ sqrt(Cows), groups = Liming,
panel = panel.superpose)
```

`summary.tlm` *summary method for class "tlm"*

Description

Summarizes the heteroscedastic t regression object

Usage

```
## S3 method for class 'tlm':
summary(object, correlation = FALSE, ...)
## S3 method for class 'summary.tlm':
print(x, ...)
```

Arguments

`object`	heteroscedastic t regression object called from `tlm()`
`x`	an object of class "`summary.tlm`" containing the values below
`correlation`	should the calculation of the parameter correlation matrix be suppressed. If the fit includes a location and a scale formula then both correlation matrices are printed. The default is `FALSE`.
`...`	arguments passed to or from other methods

Details

The table summary produced by this function should be used with caution. A more appropriate test between nested models is to use the score statistic function `tscore`.

Value

a list containing the following components:

`loc.summary`	an object containing a list of objects that summarize the location model
`scale.summary`	an object containing a list of objects that summarize the scale model
`iter`	the number of iterations of the algorithm
`dof`	value of the fixed or estimated degrees of freedom
`dofse`	the standard error associated with the degrees of freedom if estimated
`logLik`	the maximised log-likelihood
`method`	the method used to maximize the likelihood
`endTime`	the time taken for the algorithm to converge

See Also

`tsum`, `tlm`

Examples

```
data(mm, package = "hett")
attach(mm)

## fit a model with heteroscedasticity and estimating
## the degrees of freedom

tfit2 <- tlm(m.marietta ~ CRSP, ~ CRSP, data = mm, start = list(dof =
3), estDof = TRUE)
summary(tfit2)
```

`tlm` *Maximum likelihood estimation for heteroscedastic t regression*

Description

Fits a heteroscedastic t regression to given data for known and unknown degrees of freedom.

Usage

```
tlm(lform, sform =  ~ 1, data = sys.parent(), subset = NULL,
    contrasts = NULL, na.action = na.fail, start = NULL,
    control = tlm.control(...), obs = FALSE, estDof = FALSE, ... )

## S3 method for class 'tlm':
print(x, ...)
```

Arguments

`x`	an object of class "`tlm`"
`lform`	a formula of the type `response ~ terms`, where `terms` can be of the form, for example, `first + second` or `first*second`(see `lm` for details)
`sform`	a formula of the type `~ terms`, where `terms` can be of the form, for example, `first + second` or `first*second`(see `lm` for details).
`data`	the data in the form of a `data.frame` where the column names can be matched to the variable names supplied in `lform` and `sform`
`subset`	numerical vector to subset the `data` argument
`contrasts`	set of contrasts for the location model (see `contrasts.arg` for details)
`na.action`	the action to proceed with in the event of `NA`'s in the response. Currently `NA`'s are not allowed and therefore `na.fail` is the sole argument.
`start`	is a list of possibly four named components, ("`beta`", "`lambda`", "`dof`", "`omega`"), for the location, scale, degrees of freedom parameters and random scale effects respectively. Each component must be of the appropriate length.
`control`	is an argument to a function that maintains the control of the algorithm. The `tlm.control()`function contains the arguments, `epsilon` to determine how small the relative difference of likelihoods should be for convergence (default is 1e-06), `maxit` to determine the maximum iterations required (default = 50), `trace` if the user requires printing of estimates etc. as algorithm runs (default = `FALSE`), `verboseLev` to determine the amount of verbose printing to the screen as the algorithm runs (verboseLev = 1 displays location scale and dof estimates and the likelihood, verboseLev = 2 displays all of 1 plus the random scale effects)

`obs`	should the location parameters be calculated using the observed or expected information(default = `FALSE`). (Note: using the observed information does not calculate the appropriate standard errors, see DETAILS)
`estDof`	should the degrees of freedom parameter be estimated or not. If `FALSE` then the value given for `dof` in the `start` argument will be the fixed value used for the algorithm. If `TRUE` then the value given for `dof` in the `start` argument supplies an initial value only.
`...`	arguments passed to `tlm.control()` or to the `print` method

Details

When the degrees of freedom is unknown the code uses the non-linear optimiser `nlm`. If the data is tending toward the Gaussian this optimisation will still converge but with very high degrees of freedom.

To obtain the appropriate standard errors from `summary` the user must specify the argument `obs = F` to ensure that the location parameter is calculated using the expected information component.

Value

a list containing the following components:

`loc.fit`	an object containing the estimated location parameters and other elements associated with the location parameter model
`scale.fit`	an object containing the estimated scale parameters and other elements associated with the scale parameter model
`random`	the random scale effects
`dof`	fixed or estimated degrees of freedom
`dofse`	the standard error associated with the degrees of freedom
`iter`	the number of iterations of the algorithm
`logLik`	the maximised log-likelihood
`endTime`	the time taken for the algorithm to converge

Background

The theoretical background for this function can be found in Taylor and Verbyla (2004)

References

Taylor, J. D. & Verbyla, A. P (2004). Joint modelling of the location and scale parameters of the t-distribution. *Statistical Modelling* **4**, 91-112.

See Also

`summary.tlm`

Examples

```
data(mm, package = "hett")
attach(mm)

## fit a model with no heteroscedasticity and fixed degrees of freedom

tfit <- tlm(m.marietta ~ CRSP, data = mm, start = list(dof = 3))

## fit a model with heteroscedasticity and fixed degrees of freedom

tfit1 <- tlm(m.marietta ~ CRSP, ~ CRSP, data = mm, start = list(dof = 3))

## fit a model with heteroscedasticity and estimating
## the degrees of freedom

tfit2 <- tlm(m.marietta ~ CRSP, ~ CRSP, data = mm,
start = list(dof = 3), estDof = TRUE)
```

`tlm.control` *Auxiliary for Controlling tlm Fitting*

Description

Auxiliary function for fitting `tlm` model. Generally only used when calling `tlm`

Usage

```
tlm.control(epsilon = 1e-07, maxit = 50, trace = FALSE,
            verboseLev = 1)
```

Arguments

`epsilon`	positive convergence tolerance value. The iterations converge when [newlik - oldlik] < epsilon/2
`maxit`	integer giving the maximum iterations allowable for the routine
`trace`	logical. If `TRUE` output is printed to the screen during each iteration
`verboseLev`	integer. If 1 then print according to `trace`. If 2 then print random scale effects also.

Details

Value

A list with the argument as values

See Also

`tlm`

Examples

```
data(mm, package = "hett")
attach(mm)

## change the maximum amount of iterations for the algorithm

fit1 <- tlm(m.marietta ~ CRSP, ~ 1, data = mm, start = list(dof = 3),
estDof = TRUE, control = tlm.control(maxit = 100))
```

`tscore` *Score test for heteroscedastic t models*

Description

Provides a score test for the location and scale parameters of the heteroscedastic t regression model.

Usage

```
tscore(..., data = NULL, scale = FALSE)
```

Arguments

`...`	Any number of arguments containing nested model fits from `tlm()` (see Details)
`data`	the data used to fit the models involved
`scale`	logical. If `TRUE` the scale model is tested

Details

The user must supply nested models that test, *either*, the scale or the location component of the model. The model objects *must* be nested from left to right. Currently there are no traps if the arguments are not given in this order.

The models must also have either, all fixed degrees of freedom or estimated degrees of freedom.

Value

Output containing the hypothesis, the score statistic, degrees of freedom for the test and the p-value are printed to the screen.

....

See Also

`tlm`

Examples

```
data(mm, package = "hett")
attach(mm)
tfit1 <- tlm(m.marietta ~ CRSP, ~ 1, data = mm, start = list(dof = 3),
estDof = TRUE)

tfit2 <- tlm(m.marietta ~ CRSP, ~ CRSP, data = mm, start = list(dof =
3), estDof = TRUE)

tscore(tfit1, tfit2, data = mm, scale = TRUE)
```

`tsum` *Summary function for the scale or location component of a heteroscedastic t model*

Description

Summarizes the location *or* scale components of a heteroscedastic t model

Usage

```
tsum(object, dispersion = NULL, correlation = FALSE,
             symbolic.cor = FALSE, ...)

## S3 method for class 'tsum':
print(x, digits = max(3, getOption("digits") - 3), symbolic.cor =
      x$symbolic.cor, signif.stars = getOption("show.signif.stars"),
      scale = TRUE, ...)
```

Arguments

`object`	either the location *or* scale object created by fitting a heteroscedastic t object with `tlm`
`x`	an object of class "`tsum`"
`dispersion`	1 if summarizing the location model; 2 if summarizing the scale model (see Details)

`correlation` logical; if `TRUE`, the correlation matrix of the estimated parameters is returned and printed.

`digits` the number of significant digits to be printed.

`symbolic.cor`
logical. If `TRUE`, print the correlations in a symbolic form (see 'symnum') rather than as numbers.

`signif.stars`
logical. if `TRUE`, "significance stars" are printed for each coefficient.

`scale` logical. If `TRUE` then the dispersion is known in advance (2), and is printed accordingly.

`...` further arguments passed to or from other methods.

Details

The argument supplied to `dispersion` must be either 1 (location model) or 2 (scale model). The reason for this is because the fitting of the model has already scaled the covariance matrix for the location coefficients. Hence the scaled and unscaled versions of covariance matrix for the location model are identical.

This function will not be generally called by the user as it will only summarize the location or scale model but not both. Instead the user should refer to `summary.tlm` to print a summary of both models.

Value

`tsum` returns an object of class "`tsum`", a list with components

`call` the component from `object`

`df.residual` the component from `object`

`coefficients`
the matrix of coefficients, standard errors, z-values and p-values

`dispersion` the supplied dispersion argument

`df` a 2-vector of the rank of the model and the number of residual degrees of freedom

`cov.unscaled`
the unscaled (`dispersion = 1`) estimated covariance matrix of the estimated coefficients

`cov.scaled` ditto, scaled by `dispersion`

`correlation` (only if `correlation` is true.) The estimated correlations of the estimated coefficients

`symbolic.cor` (only if `correlation` is true.) The value of the argument `symbolic.cor`

See Also

`summary.tlm`, `tlm`

Examples

```
data(mm, package = "hett")
attach(mm)
tfit <- tlm(m.marietta ~ CRSP, ~ CRSP, data = mm, start = list(dof = 3),
estDof = TRUE)
tsum(tfit$loc.fit, dispersion = 1)
```

Bibliography

Aitkin, M. A. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Appl. Statist.* **36**, 332–339.

Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics* **16**, 523–531.

Atkinson, A. C. (1985). *Plots, Transformations and Regressions*. Oxford: Clarendon.

Atkinson, A. C. & Riani, M. (2000). *Robust Diagnostic Regression Analysis*. New York: Springer.

Azzalini, a. & Capitanio, A. (1999). Multivariate skew normal distribution. *Journal of the Royal Statistical Society, Series B* **61**, 579–602.

Azzalini, A. & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society, Series B* **65**, 367–389.

Barndorff-Nielsen, O. (1980). Conditionality resolutions. *Biometrika* **67**, 293–310.

Barndorff-Nielsen, O. (1983). Distribution of the maximum likelihood estimator. *Biometrika* **70**, 343–65.

Barndorff-Nielsen, O. E. (1994). Adjusted versions of profile likelihood and directed likelihood, and extended likelihood. *Journal of the Royal Statistical Society, Series B* **56**, 125–140.

Barndorff-Nielsen, O. E. & Chamberlain, S. R. (1994). Stable and invariant adjusted directed likelihoods. *Biometrika* **81**, 485–499.

Barndorff-Nielsen, O. E. & Cox, D. R. (1989). *Asymptotic Techniques for use in Statistics*. Chapman and Hall: London.

Barndorff-Nielsen, O. E. & Cox, D. R. (1994). *Inference and Asymptotics*. Chapman and Hall: London.

Barndorff-Nielsen, O. E. & McCullagh, P. (1993). A note on the relation between modified profile likelihood and the Cox-Reid adjusted profile likelihood. *Biometrika* **80**, 321–328.

Box, G. E. P. & Meyer, R. D. (1986). Dispersion effects from factorial designs. *Technometrics* **28**, 19–27.

Breslow, N. & Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* **82**, 81–91.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Breusch, T. S. & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica* **47**, 1287–1294.

Breusch, T. S., Robertson, J. C., & Welsh, A. H. (1997). The emperor's new clothes: a critique of the new multivariate t regression model. *Satistica Neerlandica* **51**, 269–286.

Brownlee, K. A. (1965). *Statistical theory and methodology in science and engineering*. New York: John Wiley.

Carroll, R. J. & Ruppert, D. (1988). *Transformations and Weighting in Regression*. New York: Chapman and Hall.

Cook, R. D. & Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.

Cook, R. D. & Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika* **70**, 1–10.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society, Series B* **49**, 1–39.

Cox, D. R. & Reid, N. (1992). A note on the difference between profile and modified profile likelihood. *Biometrika* **79**, 408–11.

De Bruijn, N. G. (1961). *Asymptotic Methods in Analysis*. Amsterdam: North-Holland.

Engel, B. & Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Staistica Neerlandica* **48**, 1–22.

Engel, J. & Huele, A. F. (1996). A generalized modelling approach to robust design. *Technometrics* **38**, 365–373.

Erdelyi, A. (1956). *Asymptotic Expansions*. New York: Dover Publications.

Fernandez, C. & Steel, M. F. J. (1999). Multivariate student-t regression models: pitfalls and inference. *Biometrika* **86**, 153–167.

Fisher, R. A. (1925). Applications of "student's" distribution. *Metron* **5**, 90–104.

Fisher, R. A. (1934). Two new properties of the mathematical likelihood. *Proceedings of the Royal Statistical Society A* **144**, 285–307.

Fraser, D. A. S. (1976). Necessary analysis and adaptive regression (with discussion). *Journal of the American Statistical Association* **71**, 99–113.

Fraser, D. A. S. (1979). *Inference and Linear Models*. New York: McGraw-Hill.

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica* **44**, 460–465.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics* **9**, 226–252.

Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley.

Huele, A. F. (1998). *Statistical Robust Design*. PhD thesis, Faculteit der Wiskunde, Informatica, Natuurkande en Sterrenkunde, Kortweg-de Vries Instituut voor Wiskunde, Amsterdam.

Huele, A. F. & Engel, J. (1998). Response to: joint modelling of mean and dispersion by Nelder and Lee. *Canadian Journal of Statistics* **36**, 95–105.

Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**.

James, A. T., Wiskich, J. T., & Conyers, R. A. (1993). t-REML for robust heteroscedastic regression analysis of mitochondrial power. *Biometrics* **49**, 339–356.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions*, volume 2. Wiley-Interscience, 2nd edition.

Jones, M. C. (2001). A skew t-distribution. In *Probability and Statistical Models with Applications: a Volume in Honor of Theophilus Cacoullos.*pp. 269-278. London: Chapman and Hall.

Jones, M. C. & Faddy, M. J. (2003). A skew extension of the t-distrbution with applications. *Journal of the Royal Statistical Society, Series B* **65**, 159–174.

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modelling using the t-distribution. *Journal of the American Statistical Association* **84**, 881–896.

Lee, Y. & Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Association, Series B* **58**, 619–678.

Lee, Y. & Nelder, J. A. (1998). Generalized linear models for the analysis of quality improvement experiments. *Canadian Journal of Statistics* **26**, 95–105.

Lee, Y. & Nelder, J. A. (2001). Hierarchical generalized linear models: a synthesis of generalized linar models, random-effect models and structured dispersions. *Biometrika* **88**, 987–1006.

Leonard, T. (1982). Comment on A simple predictive density function. *Journal of the American Staistical Association* **77**, 657–658.

Lin, X. & Breslow, N. (1996). Bias correction in generalized linear models with multiple components of dispersion. *Journal of the American Statistical Association* **91**, 1007–1016.

Lindley, D. V. (1980). Approximate Bayesian Methods. In*Bayesian Statistics*. Valencia, Spain: University Press.

Little, R. J. A. (1988). Robust estimation of mean and covariance matrix from data with missing values. *Applied Statistics* **37**, 23–39.

Liu, C., Rubin, D. B., & Wu, Y. W. (1998). Parameter expansion to accelarate the EM: The PX-EM algorithm. *Biometrika* **85**, 755–770.

Liu, C. H. (1995). Missing data imputation using the the multivariate t-distribution. *Journal of Multivariate Analysis* **53**, 139–158.

Liu, C. H. (1997). ML estimation of the multivariate t-distribution and the EM algo-

rithms. *Journal of Multivariate Analysis* **63**, 296–312.

Liu, C. H. & Rubin, D. R. (1994). The ECME algorithm: An extension of the EM and ECM with faster monotone convergence. *Biometrika* **81**, 633–648.

Liu, C. H. & Rubin, D. R. (1995). ML estimation of the multivariate t-distribution using EM and its extensions, ECM and ECME. *Statistica Sinica* **5**, 19–39.

Lütkepohl, H. (1996). *Handbook of Matrices*. New York: Wiley.

Magnus, J. R. & Neudecker, H. (1988). *Matrix Differential Calculus with Application in Statistics and Econometrics*. Chichester: John Wiley.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall, second edition.

McCullagh, P. & Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society, Series B* **52**, 325–344.

McGilchrist, C. A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society, Series B* **56**, 61–69.

Meng, X. & van Dyk, D. (1997). The EM algorithm - an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B* **59**, 511–567.

Nelder, J. A. (2000). There are not outliers in the stack-loss data. *Student* **3**, 211–216.

Nelder, J. A. & Lee, Y. (1991). Generalized linear models for the analysis of Taguchi-type experiments. *Applied Stochastic Modelling and Data Analysis* **7**, 107–120.

Park, R. E. (1966). Estimation with heteroscedastic terms. *Econometrica* **34**, 888.

Patterson, H. D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58**, 545–554.

Pinheiro, J. C., Liu, C., & Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effecs models using the multivariate t-distribution. *Journal of Computational and Graphical Statistics* **10**, 249–276.

Raudenbush, S. W., Yang, M., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of Computational and Graphical Statistics* **9**, 141–157.

Rigby, R. A. & Stasinopoulos, M. D. (1996a). Mean and dispersion additive models. In Hardle, W. & Schimek, M. G., editors, *Statistical Theory and Computational Aspects of Smoothing*, pages 215–230. Physica-Verlag.

Rigby, R. A. & Stasinopoulos, M. D. (1996b). A semi-parametric additive model for variance heterogeneity. *Statistics and Computing* **6**, 57–65.

Rigby, R. A. & Stasinopoulos, M. K. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics* **54**, 507–554.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science* **6**, 15–51.

Rubin, D. B. (1983). Iteratively reweighted least squares. In *Encyclopedia of Statistical Sciences*, volume 4. New York: John Wiley.

Ruppert, D. & Carroll, R. J. (1980). Trimmed least squares estimation in the linear

model. *Journal of the American Statistical Association* **75**, 828–838.

Ryan, T. A. J., Joiner, B. L., & Ryan, B. F. (1985). *Minitab Student Handbook*. Boston: Duxbury.

Searle, S. R., Casella, G., & McCullogh, C. E. (1992). *Variance components*. New York: Wiley.

Severini, T. A. (1998). An approximation to the modified profile likelihood function. *Biometrika* **85**, 403–411.

Shun, Z. (1997). Another loook at the salamander data: a modified laplace approximation approach. *Journal of the American Statistical Association* **92**, 341–349.

Shun, Z. & McCallaugh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B* **57**, 749–760.

Singh, R. S. (1988). Estimation of error variance in linear regression models with errors having multivariate student t-distribution with unknown degrees of freedom. *Economics Letters* **27**, 47–53.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society, Series B* **51**, 47–60.

Smyth, G. K. (2002). An efficient algorithm for REML in heteroscedastic regression. *Journal of Computational and Graphical Statistics* **11**, 836–847.

Smyth, G. K., Huele, A. F., & Verbyla, A. P. (2001). Exact and approximate REML for heteroscedastic regression. *Statistical Modelling* **1**, 161–175.

Smyth, G. K. & Verbyla, A. P. (1999). Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics* **10**, 696–709.

Solomon, P. J. & Cox, D. R. (1992). Non-linear component of variance models. *Biometrika* **79**, 1–11.

Stern, S. E. (1997). A second-order adjustemnt to the profile likelihood in the case of a multidimensional nuisance paremeter of interest. *Journal of the Royal Statistical Society, Series B* **59**, 653–665.

"Student" (1908). On the probable error of the mean. *Biometrika* **6**, 1–25.

Sutradhar, B. C. & Ali, M. M. (1986). Estimation of the parameters of a regression model with a multivariate t error variable. *Communications in Statistics: Theory and Methods* **15**, 429–450.

Taylor, J. D. & Verbyla, A. P. (2004). Joint modelling of the location and scale parameters of the t-distribution. *Statistical Modelling* **4**, 91–112.

Tierney, L. & Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86.

Tierney, L., Kass, R., & Kadane, J. (1989). Fully exponential laplace approximations to expectations and variances of non-positive functions. *Journal of the American Statistical Association* **84**, 710–716.

Venzon, D. J. & Mollgavkar, S. H. (1988). A method for computing profile-likelihood based confidence intervals. *Applied Statistics* **37**, 87–94.

Verbyla, A. P. (1990). A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics* **32**. 227-230.

Verbyla, A. P. (1993). Modelling variance heterogeneity: residual maximum likelihood and diagnostics. *Journal of the Royal Statistics Society, Series B* **55**, 493–508.

Weisberg, S. (1985). *Applied Linear Regression*. New York: John Wiley.

Welsh, A. T. & Richardson, A. M. (1997). Approaches to the robust estimation of mixed models. *Handbook of Statistics* **15**, 343–384.

West, M. (1984). Outlier models and prior distributions in bayesian linear regression. *Journal of the Royal Statistical Society, Series B* **46**, 431–439.

Wolfinger, R. (1993). Laplace's approximation for non-linear mixed models. *Biometrika* **80**, 791–795.