Linear Models with Response Functions Based on the Laplace Distribution: Statistical Formulae and Their Application to…

Article · November 2013

DOI: 10.1155/2013/496180

CITATIONS
0

READS
24

5 authors, including:

Catherine Z. W. Hassell Sweatman
Massey University
13 PUBLICATIONS **51** CITATIONS
SEE PROFILE

Graeme Wake
Massey University
222 PUBLICATIONS **2,458** CITATIONS
SEE PROFILE

A. B. Pleasants
Massey University
61 PUBLICATIONS **792** CITATIONS
SEE PROFILE

Cameron Angus Mclean
University of Auckland
18 PUBLICATIONS **941** CITATIONS
SEE PROFILE

Electronic Journal of Statistics
Vol. 0 (2006) 1–32
ISSN: 1935-7524

Linear models with perturbed and truncated Laplace response functions: the asymptotic theory of the MLE with application to epigenetics

C. Z. W. Hassell Sweatman* **and G. C. Wake*,†**

IIMS, Massey University, Albany Campus,
Private Bag 102-904, North Shore Mail Centre 0745
Auckland, New Zealand
e-mail: C.Z.W.Hassell-Sweatman@massey.ac.nz; g.c.wake@massey.ac.nz

A. B. Pleasants*,†

Ruakara Research Centre, Hamilton, New Zealand
e-mail: tony.pleasants@agresearch.co.nz

C. A. McLean and A. M. Sheppard†

Liggins Institute, The University of Auckland
Private Bag 92019, Victoria Street West
Auckland 1142, New Zealand
e-mail: ca.mclean@auckland.ac.nz; a.sheppard@auckland.ac.nz

Abstract: We extend the theory of maximum likelihood estimation for linear models to deal with response variables with distributions more general than the exponential family and the Laplace distribution. We consider perturbed and truncated versions of the Laplace distribution. These probability density functions have abrupt changes in gradient due to the presence of the modulus function. The link function is assumed to be the identity. This work arose in a biological context, the modelling of the distribution of errors in the proportions of chemical modification (methylation) on DNA, measured at specific genomic sites (CpG sites).

The perturbed Laplace probability density function has a sharp peak at its maximum. Maximun likelihood parameter estimation may be done by non-gradient methods. However, the usual classical expressions for the standard errors of the parameters, the information matrix and the log-likelihood ratio statistic do not apply due to lack of differentiability. We derive expressions for these quantities using generalized functions. The MLE is shown to be asymptotically normal. In the absence of truncation and perturbation of the Laplace probability density function, MLE corresponds to least absolute error regression. The theory is applied to find the standard errors for coefficients of a linear model, assuming the response function has a truncated Laplace distribution with added kurtosis.

AMS 2000 subject classifications: Primary 62E15, 62P10; secondary 92B15, 92D10.

*Liggins Institute, University of Auckland, Auckland, New Zealand
†National Research Centre for Growth and Development, New Zealand

Keywords and phrases: probability density functions with abrupt changes in gradient, Laplace distribution, maximum likelihood estimation, kurtosis, generalized calculus, epigenetics, Hermite polynomials, least absolute error regression.

Contents

1. Introduction and motivation

This work arose in a biological context, in epigenetics, namely the modelling of the distribution of errors in the proportions of chemical modification (methylation) on DNA, measured at specific genomic sites (CpG sites). It was observed that this error distribution may be suitably modelled by a truncated Laplace distribution with added kurtosis. Our focus became coefficient estimation for a linear model with a response variable distribution assumed to be a truncated and/or perturbed version of the Laplace distribution, estimating the standard errors of these coefficients and understanding the asymptotic theory.

The theory of generalized linear models as described in [1] covers the case of distributions from the exponential family. These distributions have probability density functions which are twice continuously differentiable ($\mathcal{C}^2$), everywhere on their support. The usual expressions for the standard errors of the model coefficients for the generalized linear models in [1] are derived using Taylor series and assume distributions with probability density functions which are $\mathcal{C}^2$, everywhere on their support. They cannot be applied to our model due to the presence of the modulus function.

We extend the theory of linear models as given in [1] to deal with response variables with distributions more general than the exponential family. We consider the Laplace distribution, and modifications thereof, with probability density functions which have abrupt changes in gradient due to the presence of the modulus function. We are concerned with response functions which are truncated and/or perturbed Laplace distributions. The theory in this paper corresponds to least absolute error (LAE) (or least absolute deviation (LAD)) regression [2, 3, 4], also called median regression [5], when the response function is the Laplace distribution without modification.

The modified Laplace probability density functions considered here have a sharp peak at the maximum. Maximum likelihood estimation (MLE) of coefficients may be done by non-gradient methods, such as the simplex method. However, the usual classical expressions for the standard errors of the coefficients, the information matrix and the log-likelihood ratio statistic do not apply due to lack of differentiability. We derive expressions for generalized versions of these quantities using generalized functions. Consequently, we show that the MLE is asymptotically normal.

The method we present to estimate these statistics could in principle be applied to other probability density functions exhibiting abrupt changes in gradi-

ent. Response function parameters are assumed known or previously estimated. The theory is applied to find the standard errors for coefficients of a linear model, assuming the response function has a truncated Laplace distribution with added kurtosis.

2. The model

2.1. The expectation is linear

Let

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T \in \mathbb{R}^n \tag{2.1}$$

be a vector of response variables,

$$\boldsymbol{X} = \boldsymbol{X}_{n,m} = \begin{pmatrix} 1 & x_{12} & \ldots & x_{1m} \\ 1 & x_{22} & \ldots & x_{2m} \\ & & \vdots & \\ 1 & x_{n2} & \ldots & x_{nm} \end{pmatrix} \tag{2.2}$$

be an $n \times m$ matrix of explanatory variables (real-valued). The subscripts denote the dimensions and will be omitted when these are assumed fixed (in §2 and §3). Let

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)^T \in \mathbb{R}^m \tag{2.3}$$

be a vector of coefficients for our linear model and assume that

$$E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}. \tag{2.4}$$

Then each component of the deviation (or error) vector

$$\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}. \tag{2.5}$$

has expectation zero. The explanatory variables may be continuous or discrete. We assume $n \geq m$ and that $\boldsymbol{X}$ has rank $r_{\boldsymbol{X}} \leq m$. In practice, we usually have n, the number of data points, much larger than m, the number of coefficients. Our goal is estimating the components of $\boldsymbol{\beta}$ by ML principles, and determining their standard errors, given a set of response variables $\boldsymbol{y}$, explanatory variables $\boldsymbol{X}$ and a response variable distribution based on the Laplace distribution as described below. In terms of generalized linear models, the link function is assumed to be the identity.

2.2. The distribution of the deviations - a modified Laplace distribution

Example 2.1 Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(z; p) = (p/Q(p))e^{(-p|z|)} \tag{2.6}$$

where $p > 0$ is a real-valued parameter and $Q(p)$ is a real-valued normalizing function defined so that

$$\int_{-\infty}^{\infty} f(z;p)dz = 1.$$

Then f is the probability density function for the Laplace distribution, with scale parameter p, centred at the origin and with unbounded support. It is not differentiable at the origin in the classical sense.

The method of MLE for the response function 2.6 corresponds to least absolute error (LAE) regression [2, 4, 3]. However, the theory of LAE regression is not sufficiently general for our epigenetic modelling problem. We next describe the more general response functions we require.

Example 2.2 Now consider the case of bounded support. For finite $B > 0$, define $f : [-B, B] \to \mathbb{R}$ by

$$f(z;p) = (p/Q(p;B))e^{(-p|z|)} \tag{2.7}$$

where $p > 0$ is a real-valued scale parameter and $Q(p;B)$ is a real-valued normalizing function defined so that

$$\int_{-B}^{B} f(z;p)dz = 1.$$

Then f is the probability density function for the truncated Laplace distribution with scale parameter p, centred at the origin and with bounded support $[-B, B]$.

Example 2.3 More generally, consider perturbations of the truncated Laplace probability density function of the following form. Let

$$f(z;p,\boldsymbol{q}) = (p/Q(p,g,\boldsymbol{q};B))e^{(-p|z|)}g(|z|;\boldsymbol{q}) \tag{2.8}$$

where real-valued $g(z;\boldsymbol{q})$ is equal to the constant map

$$g_1(z) = 1$$

plus a perturbation, $\boldsymbol{q}$ is a vector of parameters for g and parameter vector $\boldsymbol{p} = (p, \boldsymbol{q}) \in \mathbb{R}^r$, $r \geq 1$. (If $r = 1$, g has no parameters.) Here $Q(p, g, \boldsymbol{q}; B)$ is a real-valued normalizing function defined so that

$$\int_{-B}^{B} f(z;\boldsymbol{p})dz = 1.$$

We assume that there exists some $\epsilon > 0$ such that $g(z;\boldsymbol{q})$ is $\mathcal{C}^3$ in z on $(-\epsilon, B+\epsilon)$ and that $g(z;\boldsymbol{q}) > 0$ on $[0, B]$, for fixed parameter vector $\boldsymbol{q}$. Note that, as a consequence of using the modulus function, $g(|z|;\boldsymbol{q})$ will not be differentiable with respect to z, at $z = 0$, in general. As in Example 2.1, f is not differentiable at the origin due to the use of the modulus function.

Example 2.4 We could allow unbounded support if $\int_{-\infty}^{\infty} e^{(-p|z|)} g(|z|; \boldsymbol{q}) dz$ is finite.

Example 2.5 Now consider our motivating example, a truncated Laplace distribution with bounded support $[-1, 1]$, perturbed by adding kurtosis. Such a distribution is used to model the deviations in the proportions of methylation measured at gene promoter CpG sites. Specifically, to fit with observations, kurtosis is added to a Laplace probability density function with bounded support by adding a third order Hermite polynomial to give an amended version

$$f(z; \boldsymbol{p}) = (p/Q(p, g_2, q; B))e^{(-p|z|)} g_2(|z|; q). \tag{2.9}$$

Here $\boldsymbol{p} = (p, q)$, $q \geq 0$ is small, $B = 1$,

$$g_2(z; q) = 1 + qH_3(z) \tag{2.10}$$

and $H_3(z) = z^3 - 3z$ is the third order Hermite polynomial. Solving for Q yields

$$f(z; p, q) = \frac{p^4 e^{(-p|z|)}[1 + qH_3(|z|)]}{2[(p^3 - 3qp^2 + 6q) - e^{-p}(p^3(1 - 2q) + 6pq + 6q)]}. \tag{2.11}$$

Example 2.6 The functions $g_3(z; q) = 1 - qz$ and $g_4(z; q) = e^{-qz^2}$, for small positive q, with bounded support, could be used in equation (2.8) to model distributions similar to the Laplace but with thinner tails.

We restrict to symmetric distributions satisfying $f(z) = f(-z)$.

3. Maximum likelihood estimation

3.1. The log-likelihood function

Let

$$f(z_i; \boldsymbol{p})$$

be a probability density function for the deviations z_i, with parameter vector $\boldsymbol{p}$, as described in the previous section. The joint probability density function for a set of n such deviations is

$$f(z_1, \ldots, z_n; \boldsymbol{p}) = f(z_1(\boldsymbol{\beta}), \ldots, z_n(\boldsymbol{\beta}); \boldsymbol{p}) = \Pi_{i=1}^{n} f(z_i(\boldsymbol{\beta}); \boldsymbol{p}),$$

assuming independence. This is also the likelihood function

$$\mathrm{L}_{\boldsymbol{z}}(\boldsymbol{z}; \boldsymbol{p}) = \mathrm{L}_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \boldsymbol{p}; \boldsymbol{X}, \boldsymbol{y}) = \mathrm{L}(\boldsymbol{z}(\boldsymbol{\beta}); \boldsymbol{p}) = f(\boldsymbol{z}(\boldsymbol{\beta}); \boldsymbol{p}), \tag{3.1}$$

which may be regarded as a function of $\boldsymbol{z}$ or $\boldsymbol{\beta}$, here the subscript reflects our point of view. We use the log-likelihood function l in the estimation of $\boldsymbol{\beta}$ where, using various notation,

$$l_{\boldsymbol{z}}(\boldsymbol{z}; \boldsymbol{p}) = \log_e(\mathrm{L}_{\boldsymbol{z}}(\boldsymbol{z}; \boldsymbol{p})) = l_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \boldsymbol{p}; \boldsymbol{X}, \boldsymbol{y}) = l(\boldsymbol{z}(\boldsymbol{\beta}); \boldsymbol{p}). \tag{3.2}$$

Substituting measured values of y_i and known inputs x_{ij} into $l_{\boldsymbol{\beta}}$ we obtain a function of $\boldsymbol{\beta}$ and $\boldsymbol{p}$. The parameters $\boldsymbol{p}$ are assumed known, but if not, may be estimated separately. In our biological application, they are estimated by MLE and are assumed fixed for a particular measuring process. Hence we have a function of $\boldsymbol{\beta}$, the coefficients of our linear model.

Our aim is to find a maximum likelihood estimator (MLE) denoted $\hat{\boldsymbol{\beta}}_n \in \mathbb{R}^m$, that is, some point at which l attains its maximum value. The subscript n corresponds to the number of deviations. Now l is a continuous function. If B is finite, it has compact support in $\mathbb{R}^m$. Since a continuous function on a compact set attains its supremum, the existence of a MLE for $l_{\boldsymbol{\beta}}$ is guaranteed. Even if $B = \infty$, we may consider truncations with finite bounds $B_k = k$, $k = 1, 2, \ldots$, Since l is maximized when the z_i are small, truncating f to $[-k, k]$ for k large enough will not affect the set of points at which l attains its maximum. We show in §3.3 that a MLE $\hat{\boldsymbol{\beta}}_n$ is not necessarily unique.

Case 3.1 If $g(z; \boldsymbol{q}) = g_1(z) = 1$ and so f is the Laplace probability density function with parameter p as in (2.6) or (2.7), then, for $\boldsymbol{\beta} \in \mathbb{R}^m$ such that every $z_i(\boldsymbol{\beta})$ is in the support of f, then

$$\begin{aligned} l_{\boldsymbol{\beta}}(\boldsymbol{\beta}; p; \boldsymbol{X}, \boldsymbol{y}) &= l_{\boldsymbol{\beta},n,m}(\boldsymbol{\beta}; p; \boldsymbol{X}_{n,m}, \boldsymbol{y}) \\ &= -n\log(Q) + n\log(p) + \Sigma_{i=1}^{n}(-p|z_i(\boldsymbol{\beta})|) \\ &= -n\log(Q) + n\log(p) + \Sigma_{i=1}^{n}(-p|y_i - (\boldsymbol{X}\boldsymbol{\beta})_i)|). \end{aligned} \tag{3.3}$$

where $Q = 2$ if the support of f is $\mathbb{R}$ and if the support of f is $[-B, B]$

$$Q = Q(p, g_1; B) = 2(1 - e^{-pB}). \tag{3.4}$$

The theory of LAE regression (corresponding to MLE using Laplace distributions without modification as response functions) may be found in various texts eg [3]. Here it is proved that there exists a MLE $\hat{\boldsymbol{\beta}}_n$ corresponding to at least $r_{\boldsymbol{X}}$ zero errors. We are concerned with the extension of these ideas to the case of perturbed and truncated Laplace response functions. For the truncated Laplace distribution we prove that there exists a MLE $\hat{\boldsymbol{\beta}}_n$ corresponding to at least $r_{\boldsymbol{X}}$ zero errors. Consider the following more general case.

Case 3.2 If f is a perturbed Laplace probability density function with perturbing function $g(z; \boldsymbol{q})$ and bounded support as in (2.8), then for $\boldsymbol{\beta} \in \mathbb{R}^m$ such that $|z_i(\boldsymbol{\beta})| \leq B$, $i = 1, 2, \ldots, n$,

$$\begin{aligned} & l_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \boldsymbol{p}; \boldsymbol{X}, \boldsymbol{y}) \\ = \; & l_{\boldsymbol{\beta},n,m}(\boldsymbol{\beta}; p; \boldsymbol{X}_{n,m}, \boldsymbol{y}) \\ = \; & -n\log(Q(p, g, \boldsymbol{q}; B)) + n\log(p) + \Sigma_{i=1}^{n}(-p|z_i(\boldsymbol{\beta})|) + \Sigma_{i=1}^{n}\log(g(|z_i(\boldsymbol{\beta})|; \boldsymbol{q})) \\ = \; & -n\log(Q(p, g, \boldsymbol{q}; B)) + n\log(p) + \Sigma_{i=1}^{n}(-p|y_i - (\boldsymbol{X}\boldsymbol{\beta})_i)|) \\ + \; & \Sigma_{i=1}^{n}\log(g(|y_i - (\boldsymbol{X}\boldsymbol{\beta})_i)|; \boldsymbol{q})). \end{aligned} \tag{3.5}$$

Note since $g(z_i;\boldsymbol{q})$ is strictly positive on $[0,B]$, so is $f(z_i;\boldsymbol{p})$ and so $\log(f(z_i;\boldsymbol{p}))$ is well-defined on $[-B,B]$, $i\in 1,2,\ldots,n$.

In §3.3 we show that if the perturbing function g is such that $\log g(z;\boldsymbol{q})$ is convex and non-increasing on $[0,B]$, there exists a MLE corresponding to at least $r_{\boldsymbol{X}}$ data points. We give an upper bound on $|d\log g(z;\boldsymbol{q})/dz|$ on $[0,B]$ which, if not exceeded, ensures that there exists a MLE corresponding to at least one data point. We apply these results to our motivating example, the Laplace distribution with added kurtosis, described below.

Example 3.3 If $g(z;\boldsymbol{q})=g_2(z;q)=1+qH_3(z)$ and so f is a Laplace probability density function with parameter p with added kurtosis and bounded support as in (2.9), then for $\boldsymbol{\beta}\in\mathbb{R}^m$ such that $|z_i(\boldsymbol{\beta})|\leq 1$, $i=1,2,\ldots,n$,

$$\begin{aligned} & l_{\boldsymbol{\beta}}(\boldsymbol{\beta};p,q;\boldsymbol{X},\boldsymbol{y}) \\ = \;& l_{\boldsymbol{\beta},n,m}(\boldsymbol{\beta};p;\boldsymbol{X}_{n,m},\boldsymbol{y}) \\ = \;& -n\log(Q(p,g_2,q;1))+n\log(p)+\Sigma_{i=1}^{n}(-p|z_i(\boldsymbol{\beta})|) \\ + \;& \Sigma_{i=1}^{n}\log(1+qH_3(|z_i(\boldsymbol{\beta})|)) \\ = \;& -n\log(Q(p,g_2,q;1))+n\log(p)+\Sigma_{i=1}^{n}(-p|y_i-(\boldsymbol{X}\boldsymbol{\beta})_i)|) \\ + \;& \Sigma_{i=1}^{n}\log(1+qH_3(|y_i-(\boldsymbol{X}\boldsymbol{\beta})_i)|)). \end{aligned} \tag{3.6}$$

Now $l_{\boldsymbol{\beta}}$ is not differentiable in the classical sense with respect to the linear model coefficients β_j when any $z_i=0$. Hence we cannot assume that $l_{\boldsymbol{\beta}}$ is differentiable at a MLE. This paper addresses this issue firstly by proposing a non-gradient method of coefficient estimation and secondly (in §4) by using generalized functions to calculate statistical estimates including estimates of standard errors. In §5 we discuss the asymptotic theory of the MLE.

3.2. Coefficient estimation dealing with abrupt changes in gradient

Although $\mathrm{L}_{\boldsymbol{\beta}}$ and $l_{\boldsymbol{\beta}}$ are continuous functions of $\boldsymbol{\beta}$, their first derivatives are not. Consider the geometry of the coefficient space $\mathbb{R}^m$, where $\boldsymbol{\beta}\in\mathbb{R}^m$. For each index i, since the response function distribution is defined in terms of absolute values, we can find a hyperplane H_i^0 in $\mathbb{R}^m$ on which $\mathrm{L}=\mathrm{L}_{\boldsymbol{\beta}}$ and $l=l_{\boldsymbol{\beta}}$ are not differentiable, defined by setting $z_i=0$. Let $\boldsymbol{x}_i^T=(x_{i1},x_{i2},\ldots,x_{im})$, the i-th row of $\boldsymbol{X}$, it is never the zero vector since $x_{i1}=1$, $i=1,2,\ldots,n$. Choose $\boldsymbol{w}\in\mathbb{R}^m$ so that $\boldsymbol{x}_i^T\boldsymbol{w}=0$. Then

$$\boldsymbol{\beta}=\boldsymbol{w}+(y_i\boldsymbol{x}_i)/(\boldsymbol{x}_i^T\boldsymbol{x}_i)$$

yields $z_i=0$. Let H_i^0 be the set of all such $\boldsymbol{\beta}$. For example, for $m=2$, for each error term there is a line in $\mathbb{R}^2$ on which $\mathrm{L}_{\boldsymbol{\beta}}$ and $l_{\boldsymbol{\beta}}$ have a sharp ridge. By inspection of the geometry, we would expect the values of $\boldsymbol{\beta}$ which maximize $l_{\boldsymbol{\beta}}$ to be either on the union of the hyperplanes or very close to intersections of the hyperplanes H_i^0, $i=1,2,\ldots,n$. Imagine searching in $\boldsymbol{\beta}$-space near the

hyperplanes H_i^0. Even if $\mathrm{L}_{\boldsymbol{\beta}}$ has a local maximum near but not on the union of the hyperplanes, it would be difficult to use a method based on the gradient of either $\mathrm{L}_{\boldsymbol{\beta}}$ or $l_{\boldsymbol{\beta}}$ since the gradient changes sharply whenever we cross one of the H_i^0. The simplex method of coefficient estimation, which does not require any partial derivatives, suits this geometry.

3.3. The maximum likelihood estimator corresponds to a data point

3.3.1. Convex and non-increasing perturbations of the Laplace probability density function

Consider the probability density function (2.8)

$$f(z;p,\boldsymbol{q}) = (p/Q(p,g,\boldsymbol{q};B))e^{(-p|z|)}g(|z|;\boldsymbol{q}),$$

with support $[-B,B]$, for some finite $B>0$ as described in §2.2 (Example 2.3). Recall we assume that there exists some $\epsilon>0$ such that $g(z;\boldsymbol{q})$ is $\mathcal{C}^3$ in z on $(-\epsilon,B+\epsilon)$ and that $g(z;\boldsymbol{q})>0$ on $[0,B]$, for fixed parameter $\boldsymbol{q}$. Let

$$\Omega_B = \{\boldsymbol{z}\in\mathbb{R}^n : |z_i|\leq B, i=1,\ldots,n\} = [-B,B]^n.$$

We consider the log-likelihood function

$$l_{\boldsymbol{z}}(\boldsymbol{z};\boldsymbol{p}) : \Omega_B \to \mathbb{R}$$

(conditional on $\boldsymbol{p}$) with its domain restricted to

$$\begin{aligned}\mathcal{A} &= \{\boldsymbol{z}(\boldsymbol{\beta}) : \boldsymbol{\beta}\in\mathbb{R}^m, \}\cap\Omega_B,\\ &= \{\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta} : \boldsymbol{\beta}\in\mathbb{R}^m\}\cap\Omega_B,\end{aligned}$$

that is, constrained to $\mathcal{A}$, or equivalently, the log-likelihood function

$$l_{\boldsymbol{\beta}}(\boldsymbol{\beta};\boldsymbol{p};\boldsymbol{X},\boldsymbol{y}) : \mathcal{B}\to\mathbb{R},$$

constrained to $\mathcal{B}$, where $\mathcal{B}\subset\mathbb{R}^m$ is defined as $\{\boldsymbol{\beta}\in\mathbb{R}^m : z(\boldsymbol{\beta})\in\Omega_B\}$. Note that $\mathcal{A}=\mathcal{A}(\boldsymbol{X},\boldsymbol{y},B)$ and similarly $\mathcal{B}=\mathcal{B}(\boldsymbol{X},\boldsymbol{y},B)$. Also, if the function Z is defined by

$$\begin{aligned}Z : \mathbb{R}^m &\to \mathbb{R}^n\\ \boldsymbol{\beta} &\mapsto \boldsymbol{z}(\boldsymbol{\beta}) = \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\end{aligned}$$

then $Z(\mathcal{B})=\mathcal{A}$.

Lemma 3.1. *If* $\log(g) : (-\epsilon,B+\epsilon)\to\mathbb{R}$ *is convex and non-increasing (ie* $d\log g(z;\boldsymbol{q})/dz \leq 0$ *on* $[0,B]$*, then there exists a maximum of* $l_{\boldsymbol{\beta}} : \mathcal{B}\to\mathbb{R}$ *corresponding to at least* $r_{\boldsymbol{X}}$ *data points. That is, there exists* $\hat{\boldsymbol{\beta}}_n\in\mathcal{B}\subset\mathbb{R}^m$ *such that the constrained* $l_{\boldsymbol{\beta}}$ *attains its maximum at* $\hat{\boldsymbol{\beta}}_n$ *and there exists at least* $r_{\boldsymbol{X}}$ *indices* $i_j\in\{1,2,\ldots,n\}$ *such that* $z_{i_j}(\hat{\boldsymbol{\beta}}_n)=0$*,* $j\in\{1,2,\ldots n\}$*.*

Corollary 3.1. *Let $f(z;p)$ be the Laplace probability density function (2.7) with support $[-B,B]$ and parameter p. Then there exists a maximum of $l_{\boldsymbol{\beta}} : \mathcal{B} \to \mathbb{R}$ corresponding to at least $r_{\boldsymbol{X}}$ data points.*

Proof of Corollary 3.1 Let g be the constant map $g_1(z) = 1$, then $\log(g(z)) = 0$ and hence $\log(g(z))$ is convex and non-increasing on $[0, B]$. Corollary 3.1 follows from Lemma 3.1.

Note that Corollary 3.1 could be proved directly by linear programming theory. Linear programming has been applied to the problem of minimising the sum of absolute errors in various applications (see [6], a survey article, and also [7]). The convex analysis results we require are in Appendix A.

Proof of Lemma 3.1 To begin, assume that $\boldsymbol{X}$ has full rank m, recall $n \geq m$ and that f has bounded support. Then $\mathcal{A}$ is a compact convex subset of an m-dimensional affine subspace of $\mathbb{R}^n$. Since the mapping Z is linear and has full rank m, the inverse image $\mathcal{B} = Z^{-1}(\mathcal{A})$ is a compact convex subset of $\mathbb{R}^m$. We partition $\mathcal{B} \subset \mathbb{R}^m$, which is the support of $\mathrm{L}_{\boldsymbol{\beta}}$, into a finite collection of compact convex sets, so that, on each subset, $l_{\boldsymbol{\beta}}$ is convex.

Let H_i^{δ} be the hyperplane in $\mathbb{R}^m$ corresponding to the error term $z_i(\boldsymbol{\beta}) = \delta$. Then H_i^{-B} and H_i^{B} are the hyperplanes in $\mathbb{R}^m$ corresponding to errors $z_i = -B$ and $z_i = B$, respectively. It follows that the log-likelihood function

$$\begin{aligned} & l_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \boldsymbol{p}; \boldsymbol{X}, \boldsymbol{y}) \\ = \; & -n \log(Q(p, g, \boldsymbol{q}; B)) + n \log(p) + \Sigma_{i=1}^{n}(-p|z_i(\boldsymbol{\beta})|) \\ & + \Sigma_{i=1}^{n} \log(g(|z_i(\boldsymbol{\beta})|; \boldsymbol{q})) \\ = \; & -n \log(Q(p, g, \boldsymbol{q}; B)) + n \log(p) + \Sigma_{i=1}^{n}(-p|y_i - (\boldsymbol{X}\boldsymbol{\beta})_i|) \\ & + \Sigma_{i=1}^{n} \log(g(|y_i - (\boldsymbol{X}\boldsymbol{\beta})_i)|; \boldsymbol{q}) \end{aligned} \tag{3.7}$$

is a convex function in between the the hyperplanes H_i^{-B}, H_i^0 and H_i^B, $i = 1, \ldots, n$. These $3n$ hyperplanes divide the domain $\mathcal{B}$ in $\mathbb{R}^m$ into at most 2^n open sets bounded by (but not intersecting) the hyperplanes. Each such open set (and hence its closure) may be labelled by a set of n signs. For any $\boldsymbol{\beta} \in \mathbb{R}^m$ such that $0 < |z_i(\boldsymbol{\beta})| < B$, $i = 1, \ldots, n$; $(\mathrm{sgn}(z_1(\boldsymbol{\beta})), \mathrm{sgn}(z_2(\boldsymbol{\beta})), \ldots, \mathrm{sgn}(z_n(\boldsymbol{\beta})))$ labels the open set containing $\boldsymbol{\beta}$.

Next, consider $\mathbb{R}^n$ as the union of its orthants, which we denote $\mathcal{O}_k$, $k = 1, \ldots, 2^n$. We assume the orthants are closed sets. For example, the non-negative orthant is $\{z \in \mathbb{R}^n : z_i \geq 0,\ i = 1, \ldots, n\}$. In the interior of any $\mathcal{O}_k$, the sign of z_i does not change, $i = 1, \ldots, n$. Relabel the open subsets $\mathcal{B}_k = \mathcal{B}_k(\boldsymbol{X}, \boldsymbol{y}, B)$, where $k \in \{1, 2, \ldots, 2^n\}$, so that $Z(\mathcal{B}_k) \subset \mathrm{int}(\mathcal{O}_k)$. Let $\mathcal{A}_k = Z(\mathcal{B}_k) = \mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B)$.

Now, $\mathcal{B} = \cup_k \mathrm{cl}(\mathcal{B}_k)$ where $\mathrm{cl}(\mathcal{B}_k)$ denotes the closure of the set. Since $\mathrm{cl}(\mathcal{B}_k)$ is bounded by hyperplanes, it is convex. It is closed and bounded and hence compact. Choose $k \in \{1, \ldots, 2^n\}$, such that $\mathcal{B}_k$ is non-empty. Since continuous functions are bounded on compact sets, the supremum of $l_{\boldsymbol{\beta}}$, when restricted to $\mathrm{cl}(\mathcal{B}_k)$, must be attained at one or more points in $\mathrm{cl}(\mathcal{B}_k)$. By Corollary A.1, the supremum (in our case the maximum) of $l_{\boldsymbol{\beta}}$ on $\mathrm{cl}(\mathcal{B}_k)$ is attained on the whole set or on a union of faces of dimension less than m or at a vertex. Since

there are a finite number of sets to consider, the maximum of $l_{\boldsymbol{\beta}}$ must occur at a vertex but might occur, for example, on the whole of a set or on a union of faces. This is important to consider when using search algorithms such as the simplex method; as repeated application with different starting points may give a set of solutions which, for example, lie on a line segment. Note the following points.

- Assuming that $r_{\boldsymbol{X}} = m$, any vertex of the set in $\mathbb{R}^m$ at which $l_{\boldsymbol{\beta}}$ attains its maximum must correspond to at m data points, possibly more (degeneracy). This is due to the fact that in $\mathbb{R}^n$, the gradient $\nabla l_{\boldsymbol{z}}(\boldsymbol{z}; \boldsymbol{p})$ points in the direction of the boundary of the corresponding orthant and away from the boundary of Ω_B.
- A MLE is not necessarily unique.
- If $r_{\boldsymbol{X}} < m$, then we may apply the same reasonng to a subspace of $\mathbb{R}^m$ of dimension $r_{\boldsymbol{X}}$ on which the error mapping has full rank $r_{\boldsymbol{X}}$.
- Since at the MLE the absolute values of the deviations $|z_i(\hat{\boldsymbol{\beta}}_n)|$ will all be small, this proof for finite B may be extended to $B = \infty$.

Lemma 3.1 is useful but we need to know what happens for more general perturbing functions g. First we consider the Laplace distribution without perturbation.

3.3.2. The truncated Laplace probability density function

For $n = 1$, let

$$f(z; p) = (p/Q(p; 1))e^{(-p|z|)}$$

be the Laplace probability density function (2.7) with scale parameter p and with support $[-1, 1]$. Then

$$Q(p; 1) = 2(1 - e^{-p}), \tag{3.8}$$

$$f(z; p) = \frac{p}{2(1 - e^{-p})} e^{(-p|z|)}$$

and

$$\log(f(z; p)) = \log(\frac{p}{2(1 - e^{-p})}) - p|z|$$

which is a piecewise affine function in z. It has a maximum value of $\log(\frac{p}{2(1-e^{-p})})$ when $z = 0$. More generally, for $n > 1$ and independent deviations (error terms) $z_1, z_2, \ldots, z_n$,

$$l_{\boldsymbol{z}}(\boldsymbol{z}; p) = \Sigma_{i=1}^{n} \log(f(z_i; p)) = \Sigma_{i=1}^{n}(-p|z_i|) + n \log(\frac{p}{2(1 - e^{-p})}).$$

Hence, the log-likelihood function is, up to a constant term, a piecewise linear function in the error terms, with a maximum attained when all the errors are

zero. However, we must restrict our domain to $\mathcal{A} \subset \mathbb{R}^n$, or eqivalently to $\mathcal{B} \subset \mathbb{R}^m$. The log-likelihood function

$$l_{\boldsymbol{\beta}}(\boldsymbol{\beta}; p) = -p\Sigma_{i=1}^{n}|(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})_i| + n\log(\frac{p}{2(1-e^{-p})})$$

is a piecewise linear function, up to a constant term, in between the the hyperplanes H_i^{-1}, H_i^0 and H_i^{+1}, $i = 1, \ldots, n$. Let $\boldsymbol{X}_j$ denote the j-th column of $\boldsymbol{X}$, $j = 1, 2 \ldots, m$. Let

$$\mathrm{Sp}\{\boldsymbol{X}\} = \mathrm{Sp}\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m\} \subset \mathbb{R}^n$$

denote the span of the columns of $\boldsymbol{X}$.

Now $l_{\boldsymbol{z}} : \mathrm{cl}(\mathcal{A}_k) \mapsto \mathbb{R}$ has a critical point at $\boldsymbol{z}(\boldsymbol{\beta}) \in \mathcal{A}_k$ if and only if the gradient $\nabla l_{\boldsymbol{z}} = (\partial l/\partial z_1, \ldots, \partial l/\partial z_n)^T$ (evaluated at $\boldsymbol{z}(\boldsymbol{\beta})$) is orthogonal to $\mathrm{Sp}\{\boldsymbol{X}\}$. These column vectors are linearly independent tangent vectors to $\mathcal{A}$ at this point. This gradient is constant in the interior of any orthant. If we travel along a straight line path in any orthant, $l_{\boldsymbol{z}}$ either always increases, always decreases or remains constant. Hence we will not find an isolated local maximum or minimum in $\mathcal{A}_k$, an open set, for the constrained $l_{\boldsymbol{z}}$.

We need to be aware of the case where $\mathcal{A}_k$ lies in or very close to a level set of $l_{\boldsymbol{z}}$. We might need to test for this. This happens when the sign vector $(\mathrm{sgn}(z_1), \ldots, \mathrm{sgn}(z_n))^T$ is orthogonal to $\mathrm{Sp}\{\boldsymbol{X}\}$, or nearly so. In the former case, the constrained l is constant on $\mathrm{cl}(\mathcal{A}_k)$. In the latter case, the constrained l will differ very little around the maximum on $\mathrm{cl}(\mathcal{A}_k)$ If this is the case for all the $\mathcal{A}_k$, then the ML values for the coefficients β_j will not be sharply defined (will have large variance).

3.3.3. More general perturbations of the Laplace probability density function

The question is, given a non-trivial perturbing function $g(z; \boldsymbol{q})$, does the log-likelihood function attain its maximum at a data point? We have given conditions on g in Lemma 3.1 which are sufficient to ensure the maximum is attained at a data point. We give a more general criterion in Lemma 3.2.

Assume that $\log(g)$ is non-linear in any orthant. Otherwise we can write g in the form of a scaled Laplace distribution and apply Lemma 3.1. Then l is the sum of an affine function and a non-linear function in any orthant. This affine function is

$$l_{p,B} = -n\log(Q(p, g_1; B)) + n\log(p) + \Sigma_{i=1}^{n}(-p|z_i|), \tag{3.9}$$

the log-likelihood function corresponding to the Laplace distribution, The non-linear function is

$$l_{\mathrm{nlin}} = -n\log(Q(p, g, \boldsymbol{q}; B)) + n\log(Q(p, g_1; B)) + \Sigma_{i=1}^{n}\log g(|z_i|; \boldsymbol{q}). \tag{3.10}$$

Then $l = l_{p,B} + l_{\mathrm{nlin}}$, and so $\nabla l = \nabla l_{p,B} + \nabla l_{\mathrm{nlin}}$, where

$$\nabla l_{p,B} = -p(\mathrm{sgn}(z_1), \mathrm{sgn}(z_2), \ldots, \mathrm{sgn}(z_n))^T \tag{3.11}$$

and

$$\nabla l_{\text{nlin}} = (\frac{d\log(g(z_1;\boldsymbol{q}))}{dz_1}, \frac{d\log(g(z_2;\boldsymbol{q}))}{dz_2}, \ldots, \frac{d\log(g(z_n;\boldsymbol{q}))}{dz_n})^T. \tag{3.12}$$

It is possible that there exist orthants $\mathcal{O}_k$ in which the set $\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B)$ is orthogonal to the gradient $\nabla l_{p,B}$. It is possible that $l_{p,B}$ attains its global maximum (with respect to $\mathcal{A}(\boldsymbol{X}, \boldsymbol{y}, B)$) on the whole of $\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B)$. Hence it is important to consider the behaviour of l_{nlin} in such orthants.

Lemma 3.2. *Assume bounded support and let*

$$\gamma = \sup_{0\le z\le B}\{|(dg(z;\boldsymbol{q})/dz)/g(z;\boldsymbol{q})|\} = \sup_{0\le z\le B}\{|d\log(g(z;\boldsymbol{q}))/dz|\},$$

where sup *denotes supremum. Then if $\gamma < p$, the supremum or maximum of l is attained at a data point. In the special case that $Sp\{\boldsymbol{X}\}$ is orthogonal to $(-\text{sgn}(z_1), -\text{sgn}(z_2), \ldots, -\text{sgn}(z_n))^T$ in any orthant, it may be that the supremum is also attained elsewhere.*

Proof of Lemma 3.2 Choose $\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B) = \mathcal{A}(\boldsymbol{X}, \boldsymbol{y}, B) \cap \mathcal{O}_k$, where $k \in \{1, 2 \ldots, 2^n\}$, such that the set $\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B)$ is non-empty, choose $\boldsymbol{z} \in \mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B)$ (an open set relative to $\mathcal{A}$), and let $\boldsymbol{w}(k, \boldsymbol{X})$ be the projection of

$$\frac{\nabla l_{p,B}}{p} = (-\text{sgn}(z_1), -\text{sgn}(z_2), \ldots, -\text{sgn}(z_n))^T$$

onto $Sp\{\boldsymbol{X}\}$. Consider the case $\boldsymbol{w}(k, \boldsymbol{X})$ is non-zero. Then the affine function $l_{p,B}$ is not constant on $\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B)$. Given any $\boldsymbol{v}$ in $\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B)$, there exists some $\epsilon > 0$ such that the map

$$\begin{aligned} h_{p,B} : (-\epsilon, \epsilon) &\to \mathbb{R} \\ t &\mapsto l_{p,B}(\boldsymbol{v} + t\boldsymbol{w}(k, \boldsymbol{X})) \end{aligned}$$

is increasing. Now consider

$$\begin{aligned} h : (-\epsilon, \epsilon) &\to \mathbb{R} \\ t &\mapsto l(\boldsymbol{v} + t\boldsymbol{w}(k, \boldsymbol{X})). \end{aligned}$$

Since $g(z;\boldsymbol{q})$ is continuous and strictly positive on $[0, B]$, γ is finite. We show that if $\gamma < p$, $dh/dt > 0$ at $t = 0$, that is at $\boldsymbol{v}$. Specifically,

$$\begin{aligned} dh/dt &= -p\Sigma_{i=1}^n(\text{sgn}(v_i + tw_i(k, \boldsymbol{X}))w_i(k, \boldsymbol{X})) \\ &+ \Sigma_{i=1}^n \frac{(dg(|z_i|;\boldsymbol{q})/d|z_i|)_{h(t)}}{g(|v_i + tw_i(k, \boldsymbol{X})|;\boldsymbol{q})}(\text{sgn}(v_i + tw_i(k, \boldsymbol{X})))w_i(k, \boldsymbol{X}) \\ &= p(w_i(k, \boldsymbol{X}))^T w_i(k, \boldsymbol{X}) \\ &+ \Sigma_{i=1}^n \frac{(dg(|z_i|;\boldsymbol{q})/d|z_i|)_{h(t)}}{g(|v_i + tw_i(k, \boldsymbol{X})|;\boldsymbol{q})}(\text{sgn}(v_i + tw_i(k, \boldsymbol{X})))w_i(k, \boldsymbol{X}) \\ &\ge p(w_i(k, \boldsymbol{X}))^T w_i(k, \boldsymbol{X}) - \gamma(w_i(k, \boldsymbol{X}))^T w_i(k, \boldsymbol{X}) \\ &= (p - \gamma)(w_i(k, \boldsymbol{X}))^T w_i(k, \boldsymbol{X}) \\ &> 0 \end{aligned}$$

if $\gamma < p$. Then the supremum of l, constrained to $\mathrm{cl}(\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B))$, must be attained on the boundary of $\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B)$, relative to $\mathcal{A}$, and this must be at a data point due to the direction of ∇l.

In the limiting case, where $\boldsymbol{w}(k, \boldsymbol{X}) = \boldsymbol{0}$, but we still have $\gamma < p$, we find that the supremum of l, constrained to $\mathrm{cl}(\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B))$ (relative to $\mathcal{A}$), must be attained at a data point but may be attained at other points in $\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B)$ as well. Assume this is not the case, that is, there exists no data point at which l attains its supremum when constrained to $\mathrm{cl}(\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B))$. We find a contradiction as follows. Assume there exists some $\boldsymbol{v} \in \mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B)$, an open set (relative to $\mathcal{A}$), such that

$$l(\boldsymbol{v}) - \sup\{l(\boldsymbol{z}) : \boldsymbol{z} \in \mathrm{bd}(\mathcal{O}_k) \cap \mathrm{cl}(\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B))\} = \tau > 0,$$

where bd denotes boundary. Imagine perturbing the columns of $\boldsymbol{X}$ slightly (and continuously) to obtain $\boldsymbol{X}'$ so that $l_{\boldsymbol{\beta}}$, a continuous function, conditional on $\boldsymbol{X}$, changes by at most $\tau/3$, that is,

$$|l(\boldsymbol{\beta}; \boldsymbol{p}; \boldsymbol{X}, \boldsymbol{y}) - l(\boldsymbol{\beta}; \boldsymbol{p}; \boldsymbol{X}', \boldsymbol{y})| < \tau/3$$

for all $\boldsymbol{\beta} \in \mathcal{B}_k(\boldsymbol{X}, \boldsymbol{y}, B) \cap \mathcal{B}_k(\boldsymbol{X}', \boldsymbol{y}, B)$; and so that now $\boldsymbol{w}(k, \boldsymbol{X}') \neq \boldsymbol{0}$. We also require the perturbation small enough that

$$\begin{aligned} & |\sup\{l_{\boldsymbol{z}}(\boldsymbol{z}; \boldsymbol{p}) : \boldsymbol{z} \in \mathrm{bd}(\mathcal{O}_k) \cap \mathrm{cl}(\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B))\} \\ - \quad & \sup\{l_{\boldsymbol{z}}(\boldsymbol{z}; \boldsymbol{p}) : \boldsymbol{z} \in \mathrm{bd}(\mathcal{O}_k) \cap \mathrm{cl}(\mathcal{A}_k(\boldsymbol{X}', \boldsymbol{y}, B))\}| < \tau/3. \end{aligned}$$

Now $\boldsymbol{v} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{v}}$ for some unique $\boldsymbol{\beta}_{\boldsymbol{v}}$. Let $\boldsymbol{v}' = \boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\beta}_{\boldsymbol{v}}$. If the perturbation of $\boldsymbol{X}$ is small enough, then $\boldsymbol{\beta}_{\boldsymbol{v}} \in \mathcal{B}_k(\boldsymbol{X}', \boldsymbol{y}, B)$ and so $\boldsymbol{v}' \in \mathcal{A}_k(\boldsymbol{X}', \boldsymbol{y}, B)$.

Now, $|l(\boldsymbol{v}') - l(\boldsymbol{v})| < \tau/3$. Hence,

$$l(\boldsymbol{v}') - \sup\{l(\boldsymbol{z}) : \boldsymbol{z} \in \mathrm{bd}(\mathcal{O}_k) \cap \mathrm{cl}(\mathcal{A}_k(\boldsymbol{X}', \boldsymbol{y}, B))\} > \tau/3 > 0$$

which is not possible since $\boldsymbol{w}(k, \boldsymbol{X}') \neq \boldsymbol{0}$. Hence we have a contradiction. Now, for completeness, assume there exists some $\boldsymbol{r} \in \mathrm{bd}(\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B))$, $\boldsymbol{r}$ not a data point (that is some $r_i = \pm B$) such that

$$l(\boldsymbol{r}) - \sup\{l(\boldsymbol{z}) : \boldsymbol{z} \in \mathrm{bd}(\mathcal{O}_k) \cap \mathrm{cl}(\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B))\} = \tau > 0.$$

Then there exists $\boldsymbol{v} \in \mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B)$ such that

$$l(\boldsymbol{r}) - l(\boldsymbol{v}) < \tau/6.$$

Hence,

$$l(\boldsymbol{v}) - \sup\{l(\boldsymbol{z}) : \boldsymbol{z} \in \mathrm{bd}(\mathcal{O}_k) \cap \mathrm{cl}(\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, B))\} > 5\tau/6 > 0.$$

The contradiction follows as above. Hence we have proved the lemma.

3.3.4. An amended Laplace probability density function with added kurtosis

We may apply Lemma 3.2 to show that, for the amended Laplace probability density function with added kurtosis (Example 2.5), for realistic values of p and q, the maximum of the log-likelihood function must be attained at a data point. Here $g(z;q) = g_2(z;q) = 1 + q(z^3 - 3z) = 1 + qH_3(z)$ on $[0,1]$, that is, $g_2(|z|;q) = 1 + q(H_3(|z|))$ on $[-1,1]$. In n dimensions,

$$\begin{aligned} & l(\boldsymbol{\beta};\boldsymbol{p};\boldsymbol{X},\boldsymbol{y}) \\ = \; & -n\log(Q(p,g_2,q;1)) + n\log(p) + \Sigma_{i=1}^{n}(-p|y_i - (\boldsymbol{X}\boldsymbol{\beta})_i|) \\ + \; & \Sigma_{i=1}^{n}\log(1 + qH_3(|y_i - (\boldsymbol{X}\boldsymbol{\beta})_i|)) \\ = \; & -n\log(Q(p,g_2,q;1)) + n\log(p) + \Sigma_{i=1}^{n}(-p|y_i - (\boldsymbol{X}\boldsymbol{\beta})_i|) \\ + \; & \Sigma_{i=1}^{n}\log(1 + q(|y_i - (\boldsymbol{X}\boldsymbol{\beta})_i|^3 - 3|y_i - (\boldsymbol{X}\boldsymbol{\beta})_i|)). \end{aligned} \tag{3.13}$$

Consider, for $z \in [0,1]$,

$$\frac{d\log(g_2(z;q))}{dz} = \frac{d(\log(1 + q(z^3 - 3z)))}{dz} = \frac{3q(z^2-1)}{1 + q(z^3 - 3z)}$$

and

$$\frac{d^2\log(g_2(z;q))}{dz^2} = \frac{d^2(\log(1 + q(z^3 - 3z)))}{dz^2} = \frac{(3q)(2z - q(3 + z^4))}{(1 + q(z^3 - 3z))^2}.$$

When $0 < q < 0.5$, $d^2\log(g_2(z;q))/dz^2$ is negative when $z = 0$ and positive when $z = 1$ and so the continuous monotonic function $\log(g_2)$ has a point of inflection where $d^2\log(g_2(z;q))/dz^2 = 0$ in $(0,1)$, at say $\omega(q)$. Hence, $\log(g_2)$ and l are both concave and convex on $[0,1]$. When q is small, the concavity occurs close to the origin and the functions are convex on most of $(0,1)$. Moreover, $d\log(g_2(z;q))/dz$ is always negative on $(0,1)$ and its limit as $z \to 0$ is also negative so we will not get isolated local maxima in one dimension. The function $\log(g_2(z;q))$ and its first and second derivatives are plotted in Figures 1, 2 and 3, respectively; setting $q = 0.025$, a typical value. Reading from Figure 2, the upper bound γ is approximately 0.075 when $q = 0.025$. Since $p \in [3,40]$ typically, the criterion $(\gamma < p)$ for Lemma 3.2 is satisfied.

3.3.5. Non-increasing perturbations both concave and convex

In certain situations we might find the criteria for both Lemma 3.1 and Lemma 3.2 are not satisfied but that $\log g(z;\boldsymbol{q})$ is convex on most of $[0,B]$ and non-increasing. Since a sum of convex functions is convex ([8]), l will be convex wherever $\log g(|z_i|;\boldsymbol{q})$ is convex for $i = 1,2,\ldots,n$. For example, for $g = g_2$ and $B = 1$, we can prove a partial result as follows. By restricting to the domain

$$\mathcal{B}^q(\boldsymbol{X},\boldsymbol{y},1) = \cap_{i=1}^{n}\{\boldsymbol{\beta} \in \mathbb{R}^m : \omega(q) \le |z_i(\boldsymbol{\beta})| \le 1\} \subset \mathcal{B}(\boldsymbol{X},\boldsymbol{y},1)$$

or equivalently, by substituting for each set $\mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, 1)$ the subset

$$\{\boldsymbol{z} \in \cap \mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, 1) : \omega(q) < |z_i| < 1, i = 1, \ldots, n\} \subset \mathcal{A}_k(\boldsymbol{X}, \boldsymbol{y}, 1),$$

and applying convex function theory as for Lemma 3.1 we can prove that there exists $\boldsymbol{\beta}_1 \in \mathcal{B}^q(\boldsymbol{X}, \boldsymbol{y}, 1)$ at which l attains its maximum and there exists at least one index i such that $z_i(\boldsymbol{\beta}_1) = \omega(q)$. Hence there exists some $\boldsymbol{\beta}_2 \in \mathcal{B}(\boldsymbol{X}, \boldsymbol{y}, 1)$ at which l attains its maximum and there exists at least one index i such that $z_i(\boldsymbol{\beta}_2) \leq \omega(q)$. In other words, there exists some point at which l has a maximum at which at least one of the errors is very close to zero. It makes sense to search for the maxima of l near or at vertices.

For $g = g_2$ we may apply Lemma 3.2 and so do not need this partial result, but for a perturbation similar in shape to g_2, non-increasing everywhere on $[0, 1]$, with $\log g(z; \boldsymbol{q})$ convex everywhere on $(\omega(g; \boldsymbol{q}), 1]$, concave everywhere on $[0, \omega(g; \boldsymbol{q}))$, for some small positive $\omega(g; \boldsymbol{q})$, and with steep slope at the point of inflection ($z = \omega(g; \boldsymbol{q})$), such analysis would be useful.

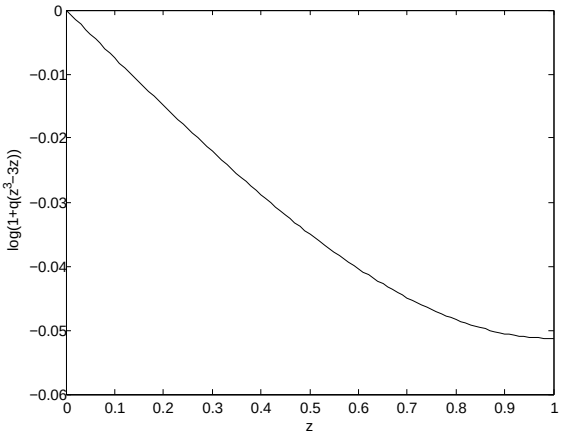

FIG 1. *The non-linear part of the log-likelihood function,* $n = 1$*;* $\log(g(z; q))$*,* $q = 0.025$

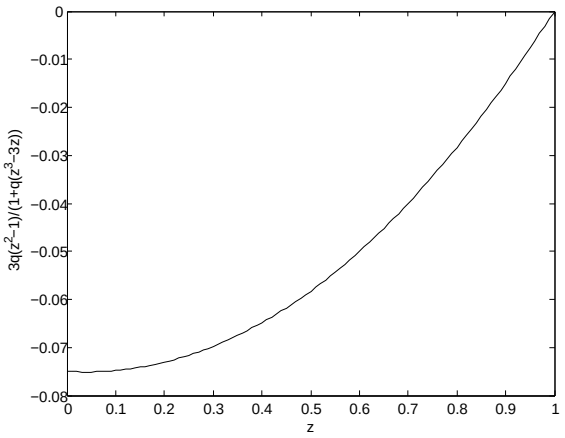

FIG 2. *The first derivative of* $\log(g(z;q))$, $q = 0.025$

4. Statistics for linear model coefficients assuming perturbed and truncated Laplace response functions

4.1. Dealing wth abrupt changes in gradient

The inclusion of the modulus (absolute value) function in the Laplace probability density function (2.6) (and variations thereof) is the cause of abrupt changes in the gradient of the log-likelihood function $l_{\boldsymbol{\beta}}$ (see §4.2). This section is devoted to dealing with the problems which are associated with these abrupt changes, encountered when deriving statistical formulae, for example, for standard errors.

The fact that $l_{\boldsymbol{\beta}}$ is not differentiable in the classical sense at a local maximum means that the assumptions made in the derivation of the usual classical formulae for the information matrix, the expected value of the Hessian of the log-likelihood function and the variance-covariance matrix for the model coefficients β_j, $j = 1, 2, \ldots, m$, are not met. For $\mathcal{C}^2$ probability density functions (and $\mathcal{C}^2$ log-likelihood functions), these formulae are derived using Taylor series. We

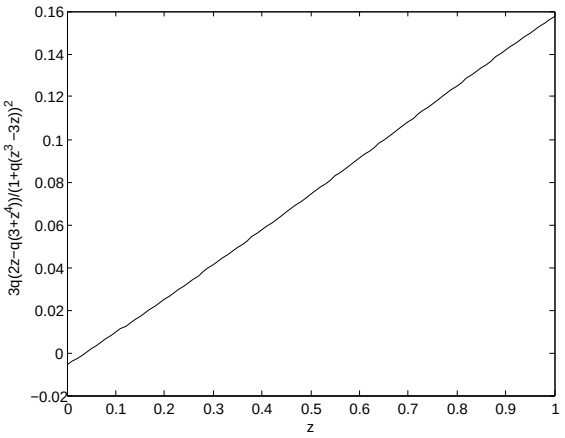

FIG 3. *The second derivative of* $\log(g(z;q))$, $q = 0.025$

find alternative expressions for these quantities assuming the truncated and/or perturbed Laplace response functions (as defined in §2) which are $\mathcal{C}^3$ where the modulus function is non-zero. In §5 these expressions will be used to prove the asymptotic convergence of our MLE to a random variable with a normal distribution.

4.2. Differentiation in a generalized sense

The following generalized functions are required to determine the first and second partial derivatives of the log-likelihood function $l_{\boldsymbol{\beta}}$, with respect to the coefficients β_j. These derivatives are needed for the calculation of the standard errors. We require

$$\operatorname{sgn}(z) = \begin{cases} 1 & z > 0 \\ 0 & z = 0 \\ -1 & z < 0 \end{cases} \tag{4.1}$$

and $\delta(z)$ which is the delta function, that is, $\delta(z) = 0$ except at $z = 0$ and $\int_{-\infty}^{\infty} \delta(z)dz = 1$. These expressions and the modulus function are connected by

$$\frac{d|z|}{dz} = \text{sgn}(z) \tag{4.2}$$

and

$$\frac{d\text{sgn}(z)}{dz} = 2\delta(z) \tag{4.3}$$

where the differentiation is taken in the generalized sense ([9, 10]). Hence, for $z_i \in [-B, B]$, the generalized derivative

$$\frac{dg(|z_i|;\boldsymbol{q})}{dz_i} = \text{sgn}(z_i)\frac{dg(|z_i|;\boldsymbol{q})}{d|z_i|}.$$

Also, the derivative of the delta function may be defined via integration by parts, assuming $h : \mathbb{R} \to \mathbb{R}$ is $\mathcal{C}^1$, we have

$$\int_{-\infty}^{\infty} \delta'(t)h(t)dt = -\int_{-\infty}^{\infty} \delta(t)h'(t)dt = -h'(0). \tag{4.4}$$

In §5 we investigate the behaviour of our model as $n \to \infty$ and so use subscripts to clarify the variables under consideration ($\boldsymbol{z}$ or $\boldsymbol{\beta}$) and/or the dimension of the space(s) under consideration. Using (2.8) and (4.2) , it follows that

$$\begin{aligned} \frac{\partial l}{\partial z_i} = \frac{\partial l_{\boldsymbol{z},n}}{\partial z_i} &= -p\text{sgn}(z_i) + \frac{d}{dz_i}\log(g(|z_i|;\boldsymbol{q})) \\ &= -p\text{sgn}(z_i) + \frac{\text{sgn}(z_i)}{g(|z_i|;\boldsymbol{q})}\left(\frac{dg(|z_i|;\boldsymbol{q})}{d|z_i|}\right). \end{aligned} \tag{4.5}$$

Since $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{X}_{n,m}\boldsymbol{\beta}$ (see (2.5)),

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l_{\boldsymbol{\beta},n,m}}{\partial \beta_j} = p\Sigma_{i=1}^{n} x_{ij}\text{sgn}(z_i) - \Sigma_{i=1}^{n} x_{ij}\frac{\text{sgn}(z_i)}{g(|z_i|;\boldsymbol{q})}\frac{dg(|z_i|;\boldsymbol{q})}{d|z_i|}. \tag{4.6}$$

Letting $\nabla l_{\boldsymbol{\beta},n,m}$ denote the gradient of $l_{\boldsymbol{\beta},n,m}$ and letting $\nabla l_{\boldsymbol{z},n}$ denote the gradient of $l_{\boldsymbol{z},n}$, we have

$$\nabla l_{\boldsymbol{\beta},n,m} = -\boldsymbol{X}_{n,m}^T \nabla l_{\boldsymbol{z},n}. \tag{4.7}$$

In addition, using (4.3) and omitting the dependence of g upon its parameters

for brevity,

$$\begin{aligned}
\frac{\partial^2 l_{\boldsymbol{z},n}}{\partial {z_i}^2} &= -2p\delta(z_i) + 2\delta(z_i)\frac{1}{g(|z_i|)}(\frac{dg(|z_i|)}{d|z_i|}) + \mathrm{sgn}(z_i)\frac{d}{dz_i}(\frac{1}{g(|z_i|)}\frac{dg(|z_i|)}{d|z_i|}) \\
&= -2p\delta(z_i) + 2\delta(z_i)\frac{1}{g(|z_i|)}(\frac{dg(|z_i|)}{d|z_i|}) + (\mathrm{sgn}(z_i))^2\frac{d}{d|z_i|}(\frac{1}{g(|z_i|)}\frac{dg(|z_i|)}{d|z_i|}) \\
&= -2p\delta(z_i) + 2\delta(z_i)\frac{1}{g(|z_i|)}(\frac{dg(|z_i|)}{d|z_i|}) \\
&\quad +(\mathrm{sgn}(z_i))^2[\frac{1}{g(|z_i|)}\frac{d^2g(|z_i|)}{d|z_i|^2} - (\frac{1}{g(|z_i|)^2})(\frac{dg(|z_i|)}{d|z_i|})^2] \\
&= -2p\delta(z_i) + 2\delta(z_i)\frac{1}{g(|z_i|)}(\frac{dg(|z_i|)}{d|z_i|}) \\
&\quad +\frac{(\mathrm{sgn}(z_i))^2}{(g(|z_i|))^2}[g(|z_i|)\frac{d^2g(|z_i|)}{d|z_i|^2} - (\frac{dg(|z_i|)}{d|z_i|})^2]
\end{aligned} \tag{4.8}$$

and, if $i \neq j$,

$$\frac{\partial^2 l_{\boldsymbol{z},n}}{\partial z_i \partial z_j} = 0. \tag{4.9}$$

Let $\mathcal{H}_{\boldsymbol{z},n}$ denote the generalized Hessian of $l_{\boldsymbol{z},n}$, where

$$\mathcal{H}_{\boldsymbol{z},n} = \begin{pmatrix} \frac{\partial^2 l}{\partial {z_1}^2} & \frac{\partial^2 l}{\partial z_1 \partial z_2} & \cdots & \frac{\partial^2 l}{\partial z_1 \partial z_n} \\ \frac{\partial^2 l}{\partial z_2 \partial z_1} & \frac{\partial^2 l}{\partial z_2^2} & \cdots & \frac{\partial^2 l}{\partial z_2 \partial z_n} \\ & & \vdots & \\ \frac{\partial^2 l}{\partial z_n \partial z_1} & \frac{\partial^2 l}{\partial z_n \partial z_2} & \cdots & \frac{\partial^2 l}{\partial z_n^2} \end{pmatrix} \tag{4.10}$$

and let $E(\mathcal{H}_{\boldsymbol{z},\boldsymbol{n}})$ denote its expected value. Then $\mathcal{H}_{\boldsymbol{z},n}$ and $E(\mathcal{H}_{\boldsymbol{z},n})$ are diagonal matrices. Since the diagonal elements of $E(\mathcal{H}_{\boldsymbol{z},n})$ are all equal (see §4.6), this matrix is a multiple of the identity. We have

$$E(\mathcal{H}_{\boldsymbol{z},n}) = E(\frac{\partial^2 l_{\boldsymbol{z},n}}{\partial {z_1}^2})I_n$$

where I_n denotes the $n \times n$ identity matrix. Let $\mathcal{H}_{\boldsymbol{\beta},n,m}$ denote the generalized Hessian of $l_{\boldsymbol{\beta},n,m}$, where

$$\mathcal{H}_{\boldsymbol{\beta},n,m} = \begin{pmatrix} \frac{\partial^2 l}{\partial {\beta_1}^2} & \frac{\partial^2 l}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 l}{\partial \beta_1 \partial \beta_m} \\ \frac{\partial^2 l}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 l}{\partial \beta_2^2} & \cdots & \frac{\partial^2 l}{\partial \beta_2 \partial \beta_m} \\ & & \vdots & \\ \frac{\partial^2 l}{\partial \beta_m \partial \beta_1} & \frac{\partial^2 l}{\partial \beta_m \partial \beta_2} & \cdots & \frac{\partial^2 l}{\partial \beta_m^2} \end{pmatrix} \tag{4.11}$$

and let $E(\mathcal{H}_{\boldsymbol{\beta},n,m})$ denote its expected value. Then

$$\mathcal{H}_{\boldsymbol{\beta},n,m} = \boldsymbol{X}_{n,m}^T \mathcal{H}_{\boldsymbol{z},n} \boldsymbol{X}_{n,m} \tag{4.12}$$

and

$$E(\mathcal{H}_{\boldsymbol{\beta},n,m}) = \boldsymbol{X}_{n,m}^T E(\mathcal{H}_{\boldsymbol{z},n})\boldsymbol{X}_{n,m} = E(\frac{\partial^2 l_{\boldsymbol{z},n}}{\partial {z_1}^2})\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m}. \tag{4.13}$$

If any $z_i = 0$, then the i-th diagonal element of $\mathcal{H}_{\boldsymbol{z},n}$ is infinite. In this case $\mathcal{H}_{\boldsymbol{\beta},n,m}$ has infinite components. We prove in §4.6 that $E(\partial^2 l_{\boldsymbol{z},n}/\partial {z_1}^2)$ is finite.

4.3. A classical relation to be generalized

Let $\mathcal{J}_{\boldsymbol{\beta},n,m}$ denote the $m \times m$ Fisher information matrix where

$$(\mathcal{J}_{\boldsymbol{\beta},n,m})_{jk} = E[(\frac{\partial l_{\boldsymbol{\beta},n,m}}{\partial \beta_j} - E(\frac{\partial l_{\boldsymbol{\beta},n,m}}{\partial \beta_j}))(\frac{\partial l_{\boldsymbol{\beta},n,m}}{\partial \beta_k} - E(\frac{\partial l_{\boldsymbol{\beta},n,m}}{\partial \beta_k}))]. \tag{4.14}$$

The components of our MLE $\hat{\boldsymbol{\beta}}_n$ depend on the errors z_i and have a distribution whose variance-covariance matrix is denoted $\mathcal{V}_{\boldsymbol{\beta},n,m}$ where

$$(\mathcal{V}_{\boldsymbol{\beta},n,m})_{jk} = E[(\beta_j - E(\beta_j))(\beta_k - E(\beta_k))], \tag{4.15}$$

$j = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, m$. If the log-likelihood function l was sufficiently smooth around the region of interest (that is, around its maximum value), then Taylor series expansions could be used to derive a relationship between $E(\mathcal{H}_{\boldsymbol{\beta},n,m})$, $\mathcal{J}_{\boldsymbol{\beta},n,m}$ and $\mathcal{V}_{\boldsymbol{\beta},n,m}$, namely

$$\mathcal{V}_{\boldsymbol{\beta},n,m} = {\mathcal{J}_{\boldsymbol{\beta},n,m}}^{-1} = (-E(\mathcal{H}_{\boldsymbol{\beta},n,m}))^{-1}, \tag{4.16}$$

(see [1]). However, our l is not sufficiently smooth and so we cannot make use of this relationship without further justification. In general, equation (4.16) above does not hold, assuming a truncated (and possibly perturbed) Laplace distribution.

In §4.5 and §4.6 we calculate the expected values of the first and second partial derivatives of the log-likelihood function $l_{\boldsymbol{\beta},n,m}$ using generalized functions, this enables us to derive a generalized Taylor series expansion for the log-likelihood function about a maximum even when the maximum is, for example, on a ridge or at a vertex. Also, this enables us to derive an expression for the generalized variance-covariance matrix for the MLEs of the model coefficients and an expression for the generalized log-likelihood ratio statistic. These formulae differ from the standard formulae for the case of smooth log-likelihood functions, although their form is similar. Specifically, in our case, we prove that $\mathcal{J}_{\boldsymbol{\beta},n,m}$ is a multiple of $E(\mathcal{H}_{\boldsymbol{\beta},n,m})$, but that the multiple is not -1, rather, it is a negative real number that depends on p, the perturbation g, its parameters $\boldsymbol{q}$ and the bound B. We prove that our generalized $\mathcal{V}_{\boldsymbol{\beta},n,m}$ is a multiple of ${\mathcal{J}_{\boldsymbol{\beta},n,m}}^{-1}$, where the multiple is a positive real number depending on p, g, $\boldsymbol{q}$ and B. We assume independent error distributions.

4.4. The mean and variance of the partial derivatives of the log-likelihood function

The mean and the variance of $\partial l/\partial z_i$ are required in the calculation of $\mathcal{J}$. Recall $f(z_1,\dots,z_n;\boldsymbol{p}) = \Pi_{i=1}^n f(z_i;\boldsymbol{p})$ is the joint probability density function for the independent deviations (errors), and that

$$l = l_{\boldsymbol{z},n} = \Sigma_{i=1}^n \log(f(z_i;\boldsymbol{p}))$$

and so

$$\frac{\partial l}{\partial z_i} = \frac{\partial l_{\boldsymbol{z},n}}{\partial z_i} = \frac{\partial \log(f(z_i;\boldsymbol{p}))}{\partial z_i}.$$

Let

$$\begin{aligned}\mu_i(p,g,\boldsymbol{q};B) &= E(\frac{\partial l_{\boldsymbol{z},n}}{\partial z_i})\\ &= \int_{\Omega_B}(\frac{\partial \log(f(z_i;\boldsymbol{p}))}{\partial z_i})(f(\boldsymbol{z};\boldsymbol{p}))dz_1\dots dz_n\\ &= \int_{-B}^{B}(\frac{\partial \log(f(z_i;\boldsymbol{p}))}{\partial z_i})(f(z_i;\boldsymbol{p}))dz_i,\end{aligned}$$

$i = 1,\dots,n$, then $\mu(p,g,\boldsymbol{q};B) = \mu_i(p,g,\boldsymbol{q};B)$ is independent of index i. Since $f(z_i;\boldsymbol{p}) = f(-z_i;\boldsymbol{p})$, $i = 1,\dots,n$, L and l are symmetric about the origin and so

$$\mu = \mu(p,g,\boldsymbol{q};B) = 0$$

for any choice of p, g, $\boldsymbol{q}$ and B. Let

$$\begin{aligned}\nu_i(p,g,\boldsymbol{q};B) &= \text{var}(\frac{\partial l_{\boldsymbol{z},n}}{\partial z_i})\\ &= \int_{\Omega_B}(\frac{\partial \log(f(z_i;\boldsymbol{p}))}{\partial z_i})^2(f(\boldsymbol{z};\boldsymbol{p}))dz_1\dots dz_n\\ &= \int_{-B}^{B}(\frac{\partial \log(f(z_i;\boldsymbol{p}))}{\partial z_i})^2(f(z_i;\boldsymbol{p}))dz_i,\end{aligned}$$

where var denotes variance. Using expression (4.5) for $\partial l_{\boldsymbol{z},n}/\partial z_i$,

$$\nu_i(p,g,\boldsymbol{q};B) = 2\int_0^B (F(z_i;\boldsymbol{p}))^2 f(z_i;\boldsymbol{p})dz_i \tag{4.17}$$

where, for $z \in (0,B]$,

$$F(z;\boldsymbol{p}) = F(z;p,g,\boldsymbol{q}) = -p + \frac{1}{g(z)}\frac{dg(z;\boldsymbol{q})}{dz} \tag{4.18}$$

and $F(0;\boldsymbol{p}) = 0$. Since $\nu_i(p,g,\boldsymbol{q};B)$ is independent of index i, we omit the subscript.Hence

$$\mathcal{J}_{\boldsymbol{z},n} = \nu I_n. \tag{4.19}$$

If $g(z;\boldsymbol{q}) = g_1(z) = 1$, then $F(z;p) = F(z;p,g_1) = -p$, and so

$$\nu(p, g_1; B) = p^2. \tag{4.20}$$

For the non-trivial perturbing function g_2, for fixed p and q, one can show that ν depends on B by direct calculation.

4.5. The information matrix

We calculate the information matrix $\mathcal{J}_{\boldsymbol{\beta},n,m}$ (conditional on p, g, $\boldsymbol{q}$ and B). We are trying to quantify the steepness of the slope of $l_{\boldsymbol{\beta},n,m}$ around a maximum, in the directions represented by the coefficients β_j. If $l_{\boldsymbol{\beta},n,m}$ is very flat in one direction, then the model coefficient representing that direction is not well-defined (will have large variance). When calculating $\mathcal{J}_{\boldsymbol{\beta},n,m}$, we are taking into account the behaviour of the gradient of $l_{\boldsymbol{\beta},n,m}$ on a whole neighbourhood of the MLE (how it differs from the expected value) and discontinuities on sets of measure zero can be accommodated. Recall equation(4.6), for $j = 1, 2, \ldots, m$,

$$\frac{\partial l_{\boldsymbol{\beta},n,m}}{\partial \beta_j} = p\Sigma_{i=1}^n x_{ij}\mathrm{sgn}(z_i) - \Sigma_{i=1}^n x_{ij}\frac{\mathrm{sgn}(z_i)}{g(|z_i|;\boldsymbol{q})}(\frac{dg(|z_i|;\boldsymbol{q})}{d|z_i|}.$$

Hence, omitting some subscripts on l for brevity,

$$\begin{aligned}
E(\frac{\partial l_{\boldsymbol{\beta},n,m}}{\partial \beta_j}) &= \int_{\Omega_B} \frac{\partial l}{\partial \beta_j}(z_1,\ldots,z_n) f(z_1,\ldots,z_n;\boldsymbol{p}) dz_1 \ldots dz_n \\
&= \int_{\Omega_B} \Big(\Sigma_{i=1}^n (\frac{\partial l}{\partial z_i}\frac{\partial z_i}{\partial \beta_j})\Big) f(z_1;\boldsymbol{p})\ldots f(z_n;\boldsymbol{p}) dz_1 \ldots dz_n \\
&= \Sigma_{i=1}^n (-x_{ij}) \int_{-B}^{B} (\frac{\partial l}{\partial z_i}) f(z_i;\boldsymbol{p}) dz_i \\
&= \Sigma_{i=1}^n (-x_{ij}) E(\frac{\partial l}{\partial z_i}) \\
&= \Sigma_{i=1}^n (-x_{ij})\mu = 0.
\end{aligned}$$

Since the expectations of the partial derivatives of $l_{\boldsymbol{\beta},n,m}$ are zero, the diagonal elements of $\mathcal{J}_{\boldsymbol{\beta},n,m}$ are an indication of the steepness of the gradient around the

maximum likelihood estimate. Now,

$$\begin{aligned}
(\mathcal{J}_{\boldsymbol{\beta},n,m})_{jk} &= E[(\frac{\partial l}{\partial \beta_j} - E(\frac{\partial l}{\partial \beta_j}))(\frac{\partial l}{\partial \beta_k} - E(\frac{\partial l}{\partial \beta_k}))] \\
&= E[(\frac{\partial l}{\partial \beta_j})(\frac{\partial l}{\partial \beta_k})] \\
&= \int_{\Omega_B} \Big(\frac{\partial l}{\partial \beta_j}\Big)\Big(\frac{\partial l}{\partial \beta_k}\Big) f(\boldsymbol{z};\boldsymbol{p}) dz_1 \dots dz_n \\
&= \int_{\Omega_B} \Big(\Sigma_{i=1}^n (\frac{\partial l}{\partial z_i}\frac{\partial z_i}{\partial \beta_j})\Big)\Big(\Sigma_{t=1}^n (\frac{\partial l}{\partial z_t}\frac{\partial z_t}{\partial \beta_k})\Big) f(\boldsymbol{z};\boldsymbol{p}) dz_1 \dots dz_n \\
&= \int_{\Omega_B} \Big(\Sigma_{i=1}^n (-x_{ij}\frac{\partial l}{\partial z_i})\Big)\Big(\Sigma_{t=1}^n (-x_{tk}\frac{\partial l}{\partial z_t})\Big) f(z_1;\boldsymbol{p}) \dots f(z_n;\boldsymbol{p}) dz_1 \dots dz_n \\
&= \int_{\Omega_B} \Big(\Sigma_{i=1}^n (x_{ij}\frac{\partial l}{\partial z_i})(x_{ik}\frac{\partial l}{\partial z_i})\Big) f(z_1;\boldsymbol{p}) \dots f(z_n;\boldsymbol{p}) dz_1 \dots dz_n
\end{aligned}$$

since the cross terms indexed by $i \neq t$ equal zero by symmetry, that is,

$$\int_{z_i=-B,z_t=-B}^{z_i=B,z_t=B} (x_{ij}\frac{\partial l}{\partial z_i})(x_{tk}\frac{\partial l}{\partial z_t}) f(z_i;\boldsymbol{p}) f(z_t;\boldsymbol{p}) dz_i\, dz_t = x_{ij}x_{tk}\mu^2 = 0.$$

So,

$$\begin{aligned}
(\mathcal{J}_{\boldsymbol{\beta},n,m})_{jk} &= \int_{\Omega_B} \Big(\Sigma_{i=1}^n (x_{ij}x_{ik}(\frac{\partial l}{\partial z_i})^2\Big) f(z_1;\boldsymbol{p}) \dots f(z_n;\boldsymbol{p}) dz_1 \dots dz_n \\
&= \Sigma_{i=1}^n (x_{ij}x_{ik}) \int_{-B}^{B} (\frac{\partial l}{\partial z_i})^2 f(z_i;\boldsymbol{p}) dz_i \\
&= \nu(p,g,\boldsymbol{q};B)\Sigma_{i=1}^n (x_{ij}x_{ik}).
\end{aligned}$$

Hence,

$$\mathcal{J}_{\boldsymbol{\beta},n,m} = \boldsymbol{X}_{n,m}^T \mathcal{J}_{\boldsymbol{z},n} \boldsymbol{X}_{n,m} = \nu(p,g,\boldsymbol{q};B)\boldsymbol{X}_{n,m}^T \boldsymbol{X}_{n,m}. \tag{4.21}$$

If $g(z;\boldsymbol{q}) = g_1(z) = 1$ then $\mathcal{J}_{\boldsymbol{\beta},n,m} = p^2 \boldsymbol{X}_{n,m}^T \boldsymbol{X}_{n,m}$.

4.6. The expected value of the generalized Hessian

In order to calculate the expected value of the generalized Hessian we require equation (4.13)

$$E(\mathcal{H}_{\boldsymbol{\beta},n,m}) = \boldsymbol{X}_{n,m}^T E(\mathcal{H}_{\boldsymbol{z},n}) \boldsymbol{X}_{n,m} = E(\frac{\partial^2 l_{\boldsymbol{z},n}}{\partial {z_1}^2}) \boldsymbol{X}_{n,m}^T \boldsymbol{X}_{n,m}$$

and equation (4.8)

$$\begin{aligned}
\frac{\partial^2 l_{\boldsymbol{z},n}}{\partial {z_i}^2} &= -2p\delta(z_i) + 2\delta(z_i)\frac{1}{g(|z_i|;\boldsymbol{q})}(\frac{dg(|z_i|;\boldsymbol{q})}{d|z_i|}) \\
&\quad + \frac{(\mathrm{sgn}(z_i))^2}{(g(|z_i|;\boldsymbol{q}))^2}[g(|z_i|;\boldsymbol{q})\frac{d^2 g(|z_i|;\boldsymbol{q})}{d|z_i|^2} - (\frac{dg(|z_i|;\boldsymbol{q})}{d|z_i|})^2].
\end{aligned}$$

Let

$$\begin{aligned}\zeta_i(p,g,\boldsymbol{q};B) &= E(\frac{\partial^2 l_{\boldsymbol{z},n}}{\partial {z_i}^2}) = \int_{\Omega_B} \frac{\partial^2 l}{\partial {z_i}^2} f(\boldsymbol{z};\boldsymbol{p}) dz_1 \dots z_n \\ &= \int_{-B}^{B} \frac{\partial^2 l}{\partial {z_i}^2} f(z_i;\boldsymbol{p}) dz_i \\ &= -2pf(0;\boldsymbol{p}) + 2f(0;\boldsymbol{p}) \frac{1}{g(0;\boldsymbol{q})} (\frac{dg(|z_i|;\boldsymbol{q})}{d|z_i|})_{z_i=0} \\ &\quad + \int_{-B}^{B} \frac{1}{g(|z_i|;\boldsymbol{q})} \frac{d^2 g(|z_i|;\boldsymbol{q})}{d|z_i|^2} f(z_i;\boldsymbol{p}) dz_i \\ &\quad - \int_{-B}^{B} \frac{1}{(g(|z_i|;\boldsymbol{q}))^2} (\frac{dg(|z_i|;\boldsymbol{q})}{d|z_i|})^2 f(z_i;\boldsymbol{p}) dz_i \end{aligned} \tag{4.22}$$

since $(dg(|z_i|;\boldsymbol{q})/d|z_i|)_{z_i=0} = (dg(u;\boldsymbol{q})/du)_{u=0}$ where $u = |z_i|$. The quantity $\zeta_i(p,g,\boldsymbol{q};B)$ does not depend on the index i and so we omit it. Hence,

$$E(\mathcal{H}_{\boldsymbol{\beta},n,m}) = \zeta(p,g,\boldsymbol{q};B) \boldsymbol{X}_{n,m}^T \boldsymbol{X}_{n,m}. \tag{4.23}$$

If $g(z;\boldsymbol{q}) = g_1(z) = 1$, then

$$\zeta(p,g_1;B) = -2p^2/Q(p,g_1;B) = -p^2/(1-e^{-pB}) \tag{4.24}$$

and so although the generalised Hessian $\mathcal{H}_{\boldsymbol{\beta},n,m}$ has infinite elements, its expected value is negative definite. By continuity, if $g(z;\boldsymbol{q})$ is close enough to the constant map g_1, the expected value $E(\mathcal{H}_{\boldsymbol{\beta},n,m})$ will still be negative definite. Note that if g has negative slope at the origin, the peak of $l_{\boldsymbol{z},n}$ at the origin becomes sharper, compared to that for the case $g = g_1$. If g has positive slope at the origin, we see the opposite effect.

4.7. The generalized variance-covariance matrix for the model coefficients

We use a generalized Taylor series expansion (in the coefficients β_j) to approximate $l_{\boldsymbol{\beta},n,m}$ by a negative definite quadratic function about a local maximum. Although we know that, for finite n, this approximation is not exact, we show in §5 that we would expect it to become more accurate as $n \to \infty$. Assuming $\boldsymbol{X}_{n,m}$ and $\boldsymbol{y} \in \mathbb{R}^n$ are given, conditional on p and $\boldsymbol{q}$, we could write a Taylor series approximation for $l_{\boldsymbol{\beta},n,m}$ about a ML estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_n \in \mathcal{B}$ as follows

$$\begin{aligned} l_{\boldsymbol{\beta},n,m}(\boldsymbol{\beta}) &\approx l_{\boldsymbol{\beta},n,m}(\hat{\boldsymbol{\beta}}_n) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)^T (\nabla l_{\boldsymbol{\beta}})_{\hat{\boldsymbol{\beta}}_n} \\ &+ (1/2)(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)^T \mathcal{H}_{\hat{\boldsymbol{\beta}}_n} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n) + \dots \end{aligned}$$

if f and hence $\mathrm{L}_{\boldsymbol{\beta},n,m}$ and $l_{\boldsymbol{\beta},n,m}$ were sufficiently smooth. Now our probability density function f and hence $l_{\boldsymbol{\beta},n,m}$ are not sufficiently smooth but we can

replace the second derivative of $l_{\boldsymbol{\beta},n,m}$ by its expected value using generalized functions. Let $\Delta\boldsymbol{\beta} = \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n$. This yields the approximation

$$l_{\boldsymbol{\beta},n,m}(\boldsymbol{\beta}) \approx l_{\boldsymbol{\beta},n,m}(\hat{\boldsymbol{\beta}}_n) + (1/2)(\Delta\boldsymbol{\beta})^T E(\mathcal{H}_{\boldsymbol{\beta},n,m})\Delta\boldsymbol{\beta}. \tag{4.25}$$

Assuming $E(\mathcal{H}_{\boldsymbol{\beta},n,m})$ is non-singular, which is true when $g = g_1$, we ignore higher order terms. Equation (4.25) provides an indication of the behaviour of $l_{\boldsymbol{\beta},n,m}$ about a maximum since, for example, if $g = g_1$ then $E(\mathcal{H}_{\boldsymbol{\beta},n,m})$ is negative definite.

Next we consider the score function $\nabla l_{\boldsymbol{\beta},n,m}$ and use a Taylor series approximation incorporating generalized functions (about a local maximum $\hat{\boldsymbol{\beta}}_n$) to derive a relationship between the expected value of the generalized Hessian $E(\mathcal{H}_{\boldsymbol{\beta},n,m})$, the information matrix $\mathcal{J}_{\boldsymbol{\beta},n,m}$ and the generalized variance-covariance matrix $\mathcal{V}_{\boldsymbol{\beta},n,m}$. This approximation is

$$\begin{aligned}(\nabla l_{\boldsymbol{\beta},n,m})_{\boldsymbol{\beta}} &\approx (\nabla l_{\boldsymbol{\beta},n,,m})_{\hat{\boldsymbol{\beta}}_n} + E(\mathcal{H}_{\boldsymbol{\beta},n,m})\Delta\boldsymbol{\beta} \\ &= E(\mathcal{H}_{\boldsymbol{\beta},n,m})\Delta\boldsymbol{\beta}.\end{aligned}$$

Since $E(\nabla l_{\boldsymbol{\beta},n,m}) = \mathbf{0}$ and $E(\mathcal{H}_{\boldsymbol{\beta},n,m})$ has full rank, $E(\hat{\boldsymbol{\beta}}_n) = \boldsymbol{\beta}$. We multiply each side by its own transpose and take expected values to obtain

$$E[(\nabla l_{\boldsymbol{\beta},n,m})_{\boldsymbol{\beta}}(\nabla l_{\boldsymbol{\beta},n,m})_{\boldsymbol{\beta}}^T] = (E(\mathcal{H}_{\boldsymbol{\beta},n,m}))E[\Delta\boldsymbol{\beta}(\Delta\boldsymbol{\beta})^T](E(\mathcal{H}_{\boldsymbol{\beta},n,m}))^T.$$

Hence

$$\mathcal{J}_{\boldsymbol{\beta},n,m} = (E(\mathcal{H}_{\boldsymbol{\beta},n,m}))\mathcal{V}_{\boldsymbol{\beta},n,m}(E(\mathcal{H}_{\boldsymbol{\beta},n,m}))^T = (E(\mathcal{H}_{\boldsymbol{\beta},n,m}))\mathcal{V}_{\boldsymbol{\beta},n,m}(E(\mathcal{H}_{\boldsymbol{\beta},n,m})) \tag{4.26}$$

and so

$$\begin{aligned}\mathcal{V}_{\boldsymbol{\beta},n,m} = \mathcal{V}_{\boldsymbol{\beta},n,m}(p,g,\boldsymbol{q},\boldsymbol{X}_{n,m};B) &= (E(\mathcal{H}_{\boldsymbol{\beta},n,m}))^{-1}\mathcal{J}_{\boldsymbol{\beta},n,m}(E(\mathcal{H}_{\boldsymbol{\beta},n,m}))^{-1} \\ &= \frac{\nu(p,g,\boldsymbol{q};B)}{(\zeta(p,g,\boldsymbol{q};B))^2}(\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m})^{-1}. \end{aligned} \tag{4.27}$$

Here we require $\boldsymbol{X}_{n,m}$ to have full rank and that $\zeta(p,g,\boldsymbol{q};B) \neq 0$ which is certainly true when $g = g_1$.

4.8. Generalized statistical expressions and relations

We have shown that the expected value of the generalized Hessian $E(\mathcal{H}_{\boldsymbol{\beta},n,m})$ and the information matrix $\mathcal{J}_{\boldsymbol{\beta},n,m}$ are multiples of $\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m}$. Specifically,

$$\begin{aligned}E(\mathcal{H}_{\boldsymbol{\beta},n,m}) &= \boldsymbol{X}_{n,m}^T E(\mathcal{H}_{\boldsymbol{z},n})\boldsymbol{X}_{n,m} = E(\frac{\partial^2 l(z_1;p,g,\boldsymbol{q})}{\partial {z_1}^2})\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m} \\ &= \zeta(p,g,\boldsymbol{q};B)\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m}, \\ \mathcal{J}_{\boldsymbol{\beta},n,m} &= E\Big((\frac{\partial l(z_1;p,g,\boldsymbol{q})}{\partial z_1})^2\Big)\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m} = \nu(p,g,\boldsymbol{q};B)\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m}\end{aligned}$$

(see equations (4.23) and (4.21)) and so if $\nu(p,g,\boldsymbol{q};B)\neq 0$ (true if $g=g_1$),

$$E(\mathcal{H}_{\boldsymbol{\beta},n,m})=\frac{E(\frac{\partial^2 l(z_1;p,g,\boldsymbol{q})}{\partial {z_1}^2})}{E\Big((\frac{\partial l(z_1;p,g,\boldsymbol{q})}{\partial z_1})^2\Big)}\mathcal{J}_{\boldsymbol{\beta},n,m}=\frac{\zeta(p,g,\boldsymbol{q};B)}{\nu(p,g,\boldsymbol{q};B)}\mathcal{J}_{\boldsymbol{\beta},n,m}. \tag{4.28}$$

In addition, assuming that the scalar $\zeta(p,g,\boldsymbol{q};B)$ is also non-zero (true if $g=g_1$) and that $\boldsymbol{X}_{n,m}$ has full rank m, we have proved the following relations

$$\begin{aligned}(\mathcal{V}_{\boldsymbol{\beta},n,m}(p,g,\boldsymbol{q},\boldsymbol{X}_{n,m};B))^{-1} &= \frac{(\zeta(p,g,\boldsymbol{q};B))^2}{\nu(p,g,\boldsymbol{q};B)}\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m} &(4.29)\\ &= \frac{\zeta(p,g,\boldsymbol{q};B)}{\nu(p,g,\boldsymbol{q};B)}E(\mathcal{H}_{\boldsymbol{\beta},n,m}) &(4.30)\\ &= \frac{(\zeta(p,g,\boldsymbol{q};B))^2}{(\nu(p,g,\boldsymbol{q};B))^2}\mathcal{J}_{\boldsymbol{\beta},n,m}, &(4.31)\end{aligned}$$

used in the derivation of the generalized log-likelihood ratio statistic (§4.11).

4.9. Statistical relations for the Laplace distribution

If $g(z,\boldsymbol{q})=g_1(z)=1$, then using equations (3.4), (4.20) and (4.24),

$$Q(p,g_1;B)=2(1-e^{-pB}),$$

$$\nu(p,g_1;B)=p^2$$

and

$$\zeta(p,g_1;B)=-2p^2/Q(p,g_1;B)=-p^2/(1-e^{-pB}).$$

Hence,

$$\mathcal{J}_{\boldsymbol{\beta},n,m}=p^2\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m}$$

and

$$E(\mathcal{H}_{\boldsymbol{\beta},n,m})=\frac{(-p^2)}{(1-e^{-pB})}\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m}$$

and so

$$\mathcal{V}_{\boldsymbol{\beta},n,m}=\frac{(1-e^{-pB})^2}{p^2}(\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m})^{-1}.$$

Here

$$\mathcal{J}_{\boldsymbol{\beta},n,m}=-(1-e^{-pB})E(\mathcal{H}_{\boldsymbol{\beta},n,m})$$

and

$$\mathcal{V}_{\boldsymbol{\beta},n,m}=(1-e^{-pB})^2{\mathcal{J}_{\boldsymbol{\beta},n,m}}^{-1}, \tag{4.32}$$

where $p>0$, $B>0$ and $\boldsymbol{X}_{n,m}$ has full rank m. Hence equation (4.16), derived for $\mathcal{C}^2$ distributions, does not hold for the Laplace distribution with bounded support. However, equation (4.16) describes the limiting behaviour, as $B\to\infty$.

4.10. Statistical relations for a Laplace distribution with added kurtosis

Now consider our motivating example, a Laplace distribution with bounded support $[-1,1]$, amended by adding kurtosis. In this case $g(z,\boldsymbol{q}) = g_2(z;q) = 1+qH_3(z)$ (see equation(2.10)), $B=1$,

$$Q(p,q) = \frac{2[(p^3-3qp^2+6q) - e^{-p}(p^3(1-2q)+6pq+6q)]}{p^3}, \tag{4.33}$$

and

$$f(z;p,q) = \frac{p^4 e^{(-p|z|)}[1+qH_3(|z|)]}{2[(p^3-3qp^2+6q) - e^{-p}(p^3(1-2q)+6pq+6q)]}.$$

Also, by equations (4.18) and (4.17),

$$F(z;p,q) = F(z;p,g_2,q) = -p + \frac{3q(z^2-1)}{1+q(z^3-3z)}, \tag{4.34}$$

$$\nu(p,g_2,q;1) = 2\int_0^1 (F(z;p,q))^2 f(z;p,q)dz$$

and, using (4.22),

$$\begin{aligned}
\zeta(p,g_2,q;1) &= -2pf(0;p,q) + 2f(0;p,q)\frac{1}{g(0;q)}\left(\frac{dg(|z|;q)}{d|z|}\right)_{z=0} \\
&\quad + \int_{-1}^1 \frac{1}{g(|z|;q)}\frac{d^2g(|z|;q)}{d|z|^2}f(z;p,q)dz \\
&\quad - \int_{-1}^1 \frac{1}{(g(|z|;q))^2}\left(\frac{dg(|z|;q)}{d|z|}\right)^2 f(z;p,q)dz \\
&= \frac{-2p^2}{Q(p,q)} + \frac{-6pq}{Q(p,q)} \\
&+ 2\int_0^1 \frac{1}{g(z;q)}\frac{d^2g(z;q)}{dz^2}f(z;p,q)dz \\
&\quad -2\int_0^1 \frac{1}{(g(z;q))^2}\left(\frac{dg(z;q)}{dz}\right)^2 f(z;p,q)dz \\
&= \frac{-2p^2-6pq}{Q(p,q)} \\
&+ 2\int_0^1 \frac{6qz}{(1+q(z^3-3z)}f(z;p,q)dz \\
&\quad -2\int_0^1 \frac{1}{(1+q(z^3-3z)^2}(3q(z^2-1))^2 f(z;p,q)dz.
\end{aligned}$$

For typical values $p = 5.254$ and $q = 0.025$, numerical integration gives $\nu = 28.3561$ and $\zeta = -28.4957$, to four decimal places. In this case,

$$\mathcal{J}_{\boldsymbol{\beta},n,m} = 28.3561\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m}$$

and

$$E(\mathcal{H}_{\boldsymbol{\beta},n,m}) = -28.4957\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m}$$

and so

$$\mathcal{V}_{\boldsymbol{\beta},n,m} = \frac{28.3561}{(-28.4957)^2}(\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m})^{-1}.$$

Here

$$\mathcal{J}_{\boldsymbol{\beta},n,m} = \frac{28.3561}{-28.4957}E(\mathcal{H}_{\boldsymbol{\beta},n,m}) \tag{4.35}$$

and

$$\mathcal{V}_{\boldsymbol{\beta},n,m} = \frac{(28.3561)^2}{(-28.4957)^2}{\mathcal{J}_{\boldsymbol{\beta},n,m}}^{-1}. \tag{4.36}$$

In this example, equation (4.16) could be used as an approximation, but does not exactly describe the relationship between $E(\mathcal{H}_{\boldsymbol{\beta},n,m})$, $\mathcal{J}_{\boldsymbol{\beta},n,m}$ and $\mathcal{V}_{\boldsymbol{\beta},n,m}$.

4.11. The generalized log-likelihood ratio statistic

The log-likelihood ratio statistic enables us to assess the adequacy of a model. It enables us to compare a model with M coefficients (parameters) with a model of interest which differs only in that it has fewer coefficients, say P, with $M > P \geq 1$, see [1] for example. We wish to compare a linear model with m coefficients, the β_j, for $j = 1, 2, \ldots, m$ with a lesser linear model with fewer coefficients. The aim is to decide whether or not the excluded coefficients are useful. Our comparison is conditional on the parameters p and $\boldsymbol{q}$.

Let $\mathrm{L}_{\boldsymbol{\rho},n,M}(\rho_1, \ldots, \rho_M; p, \boldsymbol{q}; \boldsymbol{X}_{n,M}, \boldsymbol{y})$ denote the likelihood function for the model with M parameters and let $\mathrm{L}_{\boldsymbol{\psi},n,P}(\psi_1, \ldots, \psi_P; p, \boldsymbol{q}; \boldsymbol{X}_{n,P}, \boldsymbol{y})$ denote the likelihood function for the lesser model. Let $\hat{\boldsymbol{\rho}}$ (or $\hat{\boldsymbol{\rho}_n}$) be the maximum likelihood estimator of $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_M)^T$ and let $\hat{\boldsymbol{\psi}}$ (or $\hat{\boldsymbol{\psi}_n}$) be the maximum likelihood estimator of $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_P)^T$. Then the likelihood ratio

$$\lambda = \frac{\mathrm{L}_{\boldsymbol{\rho},n,M}(\hat{\boldsymbol{\rho}}; p, \boldsymbol{q}; \boldsymbol{X}_{n,M}, \boldsymbol{y})}{\mathrm{L}_{\boldsymbol{\psi},n,P}(\hat{\boldsymbol{\psi}}; p, \boldsymbol{q}; \boldsymbol{X}_{n,P}, \boldsymbol{y})},$$

is a ratio of two probabilities and will be greater than one since the model with M parameters provides the more complete description of the model. In our application, $M = m$, $\boldsymbol{\rho} = \boldsymbol{\beta}$ and usually (but not necessarily) $P = m - 1$. We show that a multiple (conditional on p and $\boldsymbol{q}$) of

$$\log(\lambda) = \log(\mathrm{L}_{\boldsymbol{\rho},n,M}(\hat{\boldsymbol{\rho}}; p, \boldsymbol{q}; \boldsymbol{X}_{n,M}, \boldsymbol{y})) - \log(\mathrm{L}_{\boldsymbol{\psi},n,P}(\hat{\boldsymbol{\psi}}; p, \boldsymbol{q}; \boldsymbol{X}_{n,P}, \boldsymbol{y}))$$

has a chi-squared distribution as follows.

The derivation of the log-likelihood ratio statistic (for smooth functions) may be found in the textbooks. For example, for generalized linear models ([1]) where the log-likelihood function is smooth in a neighbourhood of its maximum, $D = 2\log(\lambda)$ is distributed approximately as $\chi^2(M - P, \delta_D)$. Here δ_D is a non-centrality parameter, a positive constant which will be near zero if the lesser

model fits the data almost as well as the model with more coefficients. Consider our situation in which the log-likelihood function is continuous at a maximum but where this maximum occurs at a vertex or possibly on a ridge in $(\boldsymbol{\beta}, l)$-space and generalized calculus is required to consider Taylor series expansions. Then, as in equation (4.25), around $\hat{\boldsymbol{\rho}}$, replacing the Hessian of $\log(\mathrm{L}_M) = l_M$ by its expected value, we obtain the following approximation

$$\begin{aligned} & \log(\mathrm{L}_{\boldsymbol{\rho},n,M}(E(\hat{\boldsymbol{\rho}}); p, \boldsymbol{q}; \boldsymbol{X}_{n,M}, \boldsymbol{y})) - \log(\mathrm{L}_{\boldsymbol{\rho},n,M}(\hat{\boldsymbol{\rho}}; p, \boldsymbol{q}; \boldsymbol{X}_{n,M}, \boldsymbol{y})) \\ = \;& \log(\mathrm{L}_{\boldsymbol{\rho},n,M}(E(\hat{\boldsymbol{\rho}}))) - \log(\mathrm{L}_{\boldsymbol{\rho},n,M}(\hat{\boldsymbol{\rho}})) \\ \approx \;& \frac{1}{2}(E(\hat{\boldsymbol{\rho}}) - \hat{\boldsymbol{\rho}})^T E(\mathcal{H}_{\boldsymbol{\rho},n,M})(E(\hat{\boldsymbol{\rho}}) - \hat{\boldsymbol{\rho}}). \end{aligned} \tag{4.37}$$

So, by equation (4.30) assuming that the scalar factors $\nu(p, g, \boldsymbol{q}; B)$ and $\zeta(p, g, \boldsymbol{q}; B)$ are both non-zero, and that $\boldsymbol{X}_{n,M}$ has full rank,

$$\begin{aligned} & 2\frac{E(\frac{\partial^2 l(z_1;p,g,\boldsymbol{q})}{\partial {z_1}^2})}{E\left((\frac{\partial l(z_1;p,g,\boldsymbol{q})}{\partial z_1})^2\right)}(l_{\boldsymbol{\rho},n,M}(E(\hat{\boldsymbol{\rho}})) - l_{\boldsymbol{\rho},n,M}(\hat{\boldsymbol{\rho}})) \\ \approx \;& (E(\hat{\boldsymbol{\rho}}) - \hat{\boldsymbol{\rho}})^T \frac{\zeta(p, g, \boldsymbol{q}; B)}{\nu(p, g, \boldsymbol{q}; B)} E(\mathcal{H}_{\boldsymbol{\rho},n,M})(E(\hat{\boldsymbol{\rho}}) - \hat{\boldsymbol{\rho}}) \\ = \;& (E(\hat{\boldsymbol{\rho}}) - \hat{\boldsymbol{\rho}})^T \mathcal{V}_{\boldsymbol{\rho},n,M}^{-1}(E(\hat{\boldsymbol{\rho}}) - \hat{\boldsymbol{\rho}}) \end{aligned}$$

which has the distribution $\chi^2(M)$ if the MLE has a normal distribution. We show in §5 that the distribution of the MLE is asymptotically normal. Hence if n is large enough, this will be a good approximation. Similarly,

$$\begin{aligned} & 2\frac{E(\frac{\partial^2 l(z_1;p,g,\boldsymbol{q})}{\partial {z_1}^2})}{E\left((\frac{\partial l(z_1;p,g,\boldsymbol{q})}{\partial z_1})^2\right)}(l_{\boldsymbol{\psi},n,P}(E(\hat{\boldsymbol{\psi}})) - l_{\boldsymbol{\psi},n,P}(\hat{\boldsymbol{\psi}})) \\ \approx \;& (E(\hat{\boldsymbol{\psi}}) - \hat{\boldsymbol{\psi}})^T \frac{\zeta(p, g, \boldsymbol{q}; B)}{\nu(p, g, \boldsymbol{q}; B)} E(\mathcal{H}_{\boldsymbol{\psi},n,P})(E(\hat{\boldsymbol{\psi}}) - \hat{\boldsymbol{\psi}}) \\ = \;& (E(\hat{\boldsymbol{\psi}}) - \hat{\boldsymbol{\psi}})^T \mathcal{V}_{\boldsymbol{\psi},n,P}^{-1}(E(\hat{\boldsymbol{\psi}}) - \hat{\boldsymbol{\psi}}) \end{aligned}$$

which has the distribution $\chi^2(P)$, approximately. Noting that

$$\frac{\zeta(p, g, \boldsymbol{q}; B)}{\nu(p, g, \boldsymbol{q}; B)} = \frac{-1}{(1 - e^{-pB})} < 0 \tag{4.38}$$

when $g = g_1$, let

$$D_{\mathrm{gen}} = -2\frac{\zeta(p, g, \boldsymbol{q}; B)}{\nu(p, g, \boldsymbol{q}; B)} \log(\lambda). \tag{4.39}$$

Then D_{gen}, the log-likelihood ratio statistic calculated with generalized func-

tions, a positive number, may be expressed as

$$\begin{aligned} & D_{\text{gen}} \qquad (4.40)\\ = \; & -2\frac{\zeta(p,g,\boldsymbol{q};B)}{\nu(p,g,\boldsymbol{q};B)}(l_{\boldsymbol{\rho},n,M}(\hat{\boldsymbol{\rho}};p,\boldsymbol{q};\boldsymbol{X}_{n,M},\boldsymbol{y})-l_{\boldsymbol{\psi},n,P}(\hat{\boldsymbol{\psi}};p,\boldsymbol{q};\boldsymbol{X}_{n,P},\boldsymbol{y}))\\ = \; & -2\frac{\zeta(p,g,\boldsymbol{q};B)}{\nu(p,g,\boldsymbol{q};B)}(l_{\boldsymbol{\rho},n,M}(\hat{\boldsymbol{\rho}};p,\boldsymbol{q};\boldsymbol{X}_{n,M},\boldsymbol{y})-l_{\boldsymbol{\rho},n,M}(E(\hat{\boldsymbol{\rho}});p,\boldsymbol{q};\boldsymbol{X}_{n,M},\boldsymbol{y}))\\ + \; & 2\frac{\zeta(p,g,\boldsymbol{q};B)}{\nu(p,g,\boldsymbol{q};B)}(l_{\boldsymbol{\psi},n,P}(\hat{\boldsymbol{\psi}};p,\boldsymbol{q};\boldsymbol{X}_{n,P},\boldsymbol{y})-l_{\boldsymbol{\psi},n,P}(E(\hat{\boldsymbol{\psi}});p,\boldsymbol{q};\boldsymbol{X}_{n,P},\boldsymbol{y}))\\ - \; & 2\frac{\zeta(p,g,\boldsymbol{q};B)}{\nu(p,g,\boldsymbol{q};B)}(l_{\boldsymbol{\rho},n,M}(E(\hat{\boldsymbol{\rho}});p,\boldsymbol{q};\boldsymbol{X}_{n,M},\boldsymbol{y})-l_{\boldsymbol{\psi},n,P}(E(\hat{\boldsymbol{\psi}});p,\boldsymbol{q};\boldsymbol{X}_{n,P},\boldsymbol{y})). \end{aligned}$$

Hence D_{gen} is the sum of three terms, the first (positive) has the distribution $\chi^2(M)$ (approximately). The second (negative) has the distribution $\chi^2(P)$ (approximately). The third is a positive constant (say δ_{gen}) that depends on p and $\boldsymbol{q}$. Hence, D_{gen} is distributed approximately as $\chi^2(M-P,\delta_{\text{gen}})$. If the lesser model gives a good description of the data then δ_{gen} will be small. The generalized log-likelihood ratio statistic D_{gen} is easily calculated and is hence a potentially useful statistic for assessing our linear model for which the log-likelihood function is not $\mathcal{C}^2$. Note if $g=g_1$, $D_{\text{gen}}=(2\log(\lambda))/(1-e^{-pB})$ and so $D_{\text{gen}}\to 2\log(\lambda)$, as $B\to\infty$.

5. The maximum likelihood estimator is consistent and asymptotically normal

The generalized expressions derived in §4 will be used to prove the asymptotic convergence of our MLE to a random variable with a normal distribution. Recall the following assumptions.

- Our model is linear (equation (2.4)).
- The response function f is a Laplace probability density function, generally perturbed and/or truncated, as given in equation (2.8).

We make the following further assumptions.

- There exists a unique true vector of coefficients $\boldsymbol{\beta}_0\in\mathbb{R}^m$ whose value we are trying to estimate.
- The matrix $\boldsymbol{X}_{n,m}$ has full rank m, so that $\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m}$ has full rank m.
- The $\lim_{n\to\infty}(1/n)\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m}=\boldsymbol{W}_m$, a positive definite matrix.
- Assuming fixed m, for $n\geq m$, denote by $\hat{\boldsymbol{\beta}}_n$ a (not necessarily unique) MLE of the true value $\boldsymbol{\beta}_0$, corresponding to the explanatory variables in $\boldsymbol{X}_{n,m}$.

Lemma 5.1. *The ML estimates $\hat{\boldsymbol{\beta}}_n$ exist, for $n\geq m$.*

Proof of Lemma 5.1 Since a continuous function on a compact set attains it maximum, the existence of a maximum of the log-likelihood function $l_{\boldsymbol{\beta},n,m}$ is

guaranteed for finite bound B. Even if the domain is not bounded, we can work with finite bound B and then let $B \to \infty$.

Theorem 5.1. *The random variable $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ converges in distribution to an m-dimensional normally distributed random vector with mean $\mathbf{0}$ and covariance matrix $(\nu/\zeta^2)\boldsymbol{W}_m^{-1}$, that is, as $n \to \infty$,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, (\nu/\zeta^2)\boldsymbol{W}_m^{-1}). \tag{5.1}$$

Lemma 5.2. *The random variable $(1/\sqrt{n})\nabla l_{\boldsymbol{\beta},n,m}$ converges in distribution to an m-dimensional normally distributed random vector with mean $\mathbf{0}$ and covariance matrix $\nu\boldsymbol{W}_m$, that is, as $n \to \infty$,*

$$(1/\sqrt{n})\nabla l_{\boldsymbol{\beta},n,m} \xrightarrow{D} N(\mathbf{0}, \nu\boldsymbol{W}_m). \tag{5.2}$$

Hence, $(1/n)\nabla l_{\boldsymbol{\beta},n,m}$ converges in probability to $\mathbf{0}$.

Proof of Lemma 5.2 Consider the random variable $\nabla l_{\boldsymbol{\beta},n,m} = -\boldsymbol{X}_{n,m}^T \nabla l_{z,n}$. If we repeat the sampling of $n \geq m$ data points, using $\boldsymbol{X}_{n,m}$ a total of t times, we may write

$$\begin{aligned} \nabla l_{\boldsymbol{\beta},nt,m} &= -\boldsymbol{X}_{nt,m}^T \nabla l_{z,nt} && (5.3)\\ &= -t\boldsymbol{X}_{n,m}^T \overline{\nabla l_{z,n}} && (5.4) \end{aligned}$$

where$\overline{\nabla l_{z,n}}$ denotes the average of t samples of the random vector $\nabla l_{z,n}$. Now $\nabla l_{z,n}$ has mean $\mathbf{0}$ and covariance matrix νI_n and so by the multivariate Central Limit Theorem, as $t \to \infty$,

$$\sqrt{t}\,\overline{\nabla l_{z,n}} \xrightarrow{D} N(\mathbf{0}, \nu I_n), \tag{5.5}$$

$$\boldsymbol{X}_{n,m}^T \sqrt{t}\,\overline{\nabla l_{z,n}} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{X}_{n,m}^T \nu \boldsymbol{X}_{n,m}), \tag{5.6}$$

$$(1/\sqrt{t})\,\nabla l_{\boldsymbol{\beta},nt,m} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{X}_{n,m}^T \nu \boldsymbol{X}_{n,m}) \tag{5.7}$$

and

$$(1/\sqrt{nt})\,\nabla l_{\boldsymbol{\beta},nt,m} \xrightarrow{D} N(\mathbf{0}, \frac{\nu}{n}\boldsymbol{X}_{n,m}^T \boldsymbol{X}_{n,m}). \tag{5.8}$$

Hence as $n \to \infty$ and $t \to \infty$

$$(1/\sqrt{nt})\,\nabla l_{\boldsymbol{\beta},nt,m} \xrightarrow{D} N(\mathbf{0}, \nu\boldsymbol{W}_m). \tag{5.9}$$

Hence as $n \to \infty$

$$(1/\sqrt{n})\,\nabla l_{\boldsymbol{\beta},n,m} \xrightarrow{D} N(\mathbf{0}, \nu\boldsymbol{W}_m). \tag{5.10}$$

Lemma 5.3. *Although for finite n, the log-likelihood function $l_{\boldsymbol{\beta},n,m}$ is not differentiable at a maximum, $(1/n)l_{\boldsymbol{\beta},n,m}$ converges in distribution to a negative definite quadratic function centred at $\boldsymbol{\beta}_0$. Hence the MLEs $\hat{\boldsymbol{\beta}}_n$ are consistent, that is, they converge in probability to the true value $\boldsymbol{\beta}_0$.*

Proof of Lemma 5.3 Using a generalized Taylor series expansion about $\boldsymbol{\beta}_0$ and noting that by Lemma 5.2, $(1/n)\nabla l_{\boldsymbol{\beta},n,m}$ converges in probability to $\mathbf{0}$, we can write

$$\begin{aligned}\lim_{n\to\infty}\frac{l_{\boldsymbol{\beta},n,m}(\boldsymbol{\beta})-l_{\boldsymbol{\beta},n,m}(\boldsymbol{\beta}_0)}{n} &= \lim_{n\to\infty}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\frac{E(\mathcal{H}_{\boldsymbol{\beta},n,m})}{n}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\\ &= (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T(\lim_{n\to\infty}\frac{\zeta\boldsymbol{X}_{n,m}^T\boldsymbol{X}_{n,m}}{n})(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\\ &= \zeta(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\boldsymbol{W}_m(\boldsymbol{\beta}-\boldsymbol{\beta}_0) \qquad (5.11)\end{aligned}$$

This shows that, in the limit as $n\to\infty$, the log-likelihood function $l_{\boldsymbol{\beta},n,m}$ has an isolated maximum at $\boldsymbol{\beta}_0$ and so a sequence of MLEs $\hat{\boldsymbol{\beta}}_n$ must converge to $\boldsymbol{\beta}_0$ in probability (consistency).

In the proof of Lemma 5.3 we ignored the third partial derivatives of $l_{\boldsymbol{\beta},n,m}$ in the generalized Taylor series expansion. The justification is as follows. Since g is assumed to be $\mathcal{C}^3$ on an open interval which contains $[0,B]$, $|d^3g/dz^3|$ must be bounded above by some positive real number G_3 on $[0,B]$. Hence, the absolute values of the third partial derivatives of l must be bounded above by some positive real number except at points $\boldsymbol{z}$ where some $z_i=0$ (data points). Note that $E(\partial^3 l/\partial z_i^3)=0$, $i=1,2,\ldots,n$. This follows from equation (4.8), using the symmetry introduced by the modulus function and equation (4.4). Hence the third partial derivative terms will be small compared to the second partial derivative terms near a critical point and so we may ignore them in our generalized Taylor series expansion for l, when the expected value of the Hessian has full rank.

Proof of Theorem 5.1 Consider the first degree approximation

$$\nabla l_{\boldsymbol{\beta},n,m}(\boldsymbol{\beta})\sim E(\mathcal{H}_{\boldsymbol{\beta},n,m})(\boldsymbol{\beta}-\boldsymbol{\beta}_0). \qquad (5.12)$$

Since $E(\mathcal{H}_{\boldsymbol{\beta},n,m})$ has full rank

$$(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\sim(E(\mathcal{H}_{\boldsymbol{\beta},n,m}))^{-1}\nabla l_{\boldsymbol{\beta},n,m}(\boldsymbol{\beta}) \qquad (5.13)$$

and so

$$\frac{n}{\sqrt{n}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\sim(\frac{E(\mathcal{H}_{\boldsymbol{\beta},n,m})}{n})^{-1}\frac{\nabla l_{\boldsymbol{\beta},n,m}(\boldsymbol{\beta})}{\sqrt{n}} \qquad (5.14)$$

By Lemma 4.1, as $n\to\infty$

$$(1/\sqrt{n})\,\nabla l_{\boldsymbol{\beta},n,m}\xrightarrow{D}N(\mathbf{0},\nu\boldsymbol{W}_m). \qquad (5.15)$$

Hence, as $n\to\infty$

$$\sqrt{n}\,(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\xrightarrow{D}N(\mathbf{0},(\zeta\boldsymbol{W}_m)^{-1}\nu\boldsymbol{W}_m(\zeta\boldsymbol{W}_m)^{-T}). \qquad (5.16)$$

Hence, as $n\to\infty$

$$\sqrt{n}\,(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\xrightarrow{D}N(\mathbf{0},(\nu/\zeta^2)\boldsymbol{W_m}^{-1}). \qquad (5.17)$$

6. Real and simulated data illustrations

6.1. Empirical distribution of methylation proportion deviations

Quantitative analysis of DNA methylation at specific genomic sites (known as CpG sites) was carried out with the Sequenom MassARRAY Compact System (www.sequenom.com/). Briefly, this involves accurate determination and comparison of the mass of transcription cleavage products following chemical modification of the DNA which is dependent upon the *a priori* methylation status, using MALDI-TOF mass spectrometry (Bruker-Sequenom) ([11]). Quantitative CpG methylation was assessed using proprietary EpiTyper software v1.0.5 (www.sequenom.com/). Sequenom measurements of 1440 CpG sites in each of 41 human umbilical cord tissue samples were performed in duplicate and the difference between the measurements recorded. This difference represents the deviation in the measurement of CpG methylation due to sample nano-dispensing and MALDI-TOF mass spectrometry detection. Figure 4 is a histogram of the deviations of these 1440 repeated measurements. Although not obvious from the histogram, about 1.04% of values are greater than 0.2 in absolute value. The data was shown not to conform to a normal distribution using the test proposed by [12] (p-value < 0.00001).

6.2. Simulated data example using methylation proportion deviations

In order to illustrate the application of the theory developed, a sample of 40 deviations was chosen at random from the total pool of 1440 available CpG methylation proportion deviations. A constant value of 0.48 was added to 20 of these samples and designated treatment H, while a constant value of 0.45 was added to the other 20 samples and designated treatment L. A unifom random variable sampled between -0.01 and 0.01 was added to each value to simulate the additional differences expected to occur between individuals.

We analysed the data using our amended Laplace distribution (2.9), with $g = g_2$, $p = 37.2129$, $q = 0.0437$ (machine characteristics) and $B = 1$. We coded a low value treatment (L) by setting $x_{i2} = -1$, $i = 1, 2, \ldots, 20$ and a high value treatment (H) by setting $x_{i2} = 1$, $i = 21, 22, \ldots, 40$. Estimates for β_1 and β_2 were calculated by maximum likelihood estimation, using the simplex method. The standard errors of the coefficient estimates were calculated using our generalized $\mathcal{V}$. Hence the estimated treatment means $(\beta_1 \pm \beta_2)$ and their standard errors were calculated. A p-value was obtained using D_{gen}, which is assumed distributed $\chi^2(1)$. This simulation was performed twice and the results are given in Table 1. Note for each simulation, the p-value is less than 0.01, indicating a significant difference between the means.

For both simulations $\nu(p, g_2, q; 1) = 1394.59$ and $\zeta(p, g_2, q; 1) = -1394.59$ (see §4.10). Also,

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} 40 & 0 \\ 0 & 40 \end{pmatrix},$$

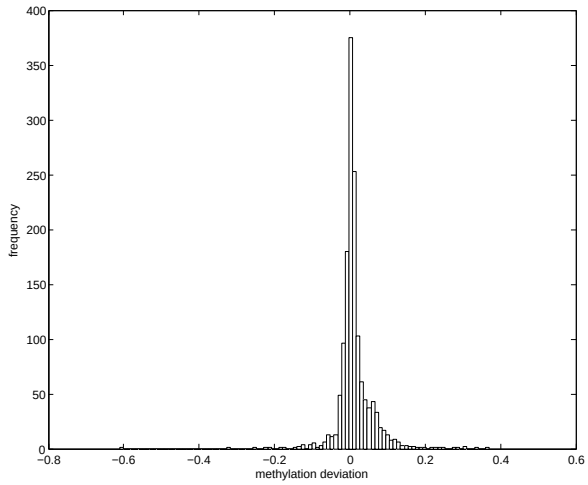

FIG 4. *Deviations in methylation proportion*

and rounding to five significant figures,

$$\mathrm{E}(\mathcal{H}) = \begin{pmatrix} -55784 & 0 \\ 0 & -55784 \end{pmatrix},$$

$$\mathcal{J} = \begin{pmatrix} 55784 & 0 \\ 0 & 55784 \end{pmatrix},$$

and

$$\mathcal{V} = \begin{pmatrix} 1.7926e-05 & 0 \\ 0 & 0.1.7926e-05 \end{pmatrix}.$$

Here

$$\mathrm{E}(\mathcal{H}) = -1.0000\mathcal{J},$$
$$\mathcal{V}^{-1} = 1.0000\mathcal{J}$$

and

$$D_{\mathrm{gen}} = 2.0000 \log(\lambda).$$

TABLE 1
Two simulation results, adding high (H) or low (L) treatments (T) to DNA methylation proportions

	amended Laplace distribution			LAE regression			normal distribution		
T	mean	std err	p-value	mean	std err	p-value	mean	std err	p-value
H	0.4705	0 0060		0.4708	0.0060		0.4806	0.0148	
L	0.4538	0.0060		0.4536	0.0060		0.4545	0.0148	
			0.005587			0.00845			0.0875
H	0.4829	0.0060		0.4831	0.0060		0.4873	0.0197	
L	0.4554	0.0060		0.4562	0.0060		0.4648	0.0197	
			0.003400			0.003466			0.3475

These expressions correspond to the classical expressions for smooth functions, up to our numerical tolerance. For the first simulation, the maximum likelihood estimator for $\boldsymbol{\beta}$ was $(0.462189, 0.00834866)^T$ and the generalized log-likelihood ratio statistic D_{gen} was 7.67895. For the second simulation, the maximum likelihood estimator for $\boldsymbol{\beta}$ was $(0.469161, 0.013757)^T$ and the generalized log-likelihood ratio statistic D_{gen} was 8.57957. For comparison, the results of a standard analysis of variance, assuming the deviations have a normal distribution, are also given in Table 1, for each of the two simulations. Using the p-values obtained, 0.0875 and 0.3475, we would not reject the null hypothesis (that the means are equal) at any usual level of significance, using a least squares approach. The two data sets analysed are given in Tables 3 and 4 in Appendix B. However, our algorithm based on the amended Laplace distribution correctly identified the structure of the simulated data set, separating two means which were fairly close. The standard least squares algorithm failed to do this.

For comparison we included LAE regression, estimating the model coefficients assuming the response function is the Laplace probability density function without modification or truncation. Since $(1 - e^{-pB})$ is so close to 1 in this example, the effect of truncation on the standard errors and D_{gen} (for p-values) is negligible. (If the machine characteristic p was smaller, say $1 \leq p \leq 3$, assuming $B = 1$, this effect would become more significant.) We see that including the perturbing Hermite polynomial improves (decreases) the p-values, meaning we can be even more confident, than when using LAE regression, that the means are different.

6.3. Primiparous versus multiparous effects on DNA methylation proportion at the promoter of the H19 gene

The CpG methylation at two CpG sites in the promoter of the H19 gene was measured in umbilical cord tissues collected as part of an ongoing prospective birth cohort study. Phenotype variables in this population include birth order or parity, defined as first born child (primiparous) or later born (multiparous). We have analysed the relationship between H19 gene methylation status and birth order in this study, using our amended Laplace distribution (2.9), with $g = g_2$, $p = 37.2129$, $q = 0.0437$ and $B = 1$, and for comparison, the usual least squares algorithm.

TABLE 2
Primiparous (p) versus multiparous (m) effects on DNA methylation proportion at the promoter of the H19 gene

		amended Laplace distribution			normal distribution		
site	parity	mean	st'd error	p-value	mean	st'd error	p-value
CpG9	p	0.180	0.006		0.300	0.059	
CpG9	m	0.480	0.006		0.441	0.059	
CpG9				$< 1.0e-09$			0.117857
CpG13	p	0.230	0.006		0.326	0.061	
CpG13	m	0.570	0.006		0.523	0.061	
CpG13				$< 1.0e-09$			0.0377835

The problem was coded by substituting $x_{i2} = 1$ for primiparous and $x_{i2} = -1$ for multiparous. Estimates for β_1 and β_2 and their standard errors were calculated. The data used is given in Table 5 in Appendix B. The estimated means $(\beta_1 \pm \beta_2)$ and their standard errors are given in Table 2. The small p-values associated with the amended Laplace distribution (2.9), calculated using D_{gen}, identify a difference between the mean methylation proportions (primiparous versus multiparous) at a given CpG site. This demonstrates the power of accounting properly for the distributional properties of the methylation errors and hence enables clearer inference of the epigenetic mechanisms underlying these biological phenomena.

In this example, LAE regression yields similar means and variances to our MLE (the same to two decimal places) and also gives small p-values, although not as small as for our MLE. For example, for LAE regression, $D_{\text{gen}} = 97.4978$ and for our MLE $D_{\text{gen}} = 97.8609$. This is because q is small and pB is large and so e^{-pB} is small. For smaller values of pB, and larger values of q, the value of taking into account the truncation and perturbation increases.

7. Discussion

7.1. A comparison with LAE regression

The original MLE theory and methods in this paper were developed assuming the response function is a modified version of the Laplace probability density function, that is, assuming non-trivial perturbation and/or truncation to compact support $[-B, B]$. Such response functions have been observed in measurement data generated by nearly all the analytical platforms currently used to assess DNA methylation, including the Sequenom EpiTyper, Infinium Mass Array and Restricted Representation Bisulphite Sequencing platforms [17].

In the absence of perturbation or truncation of the response function, the results in this paper correspond to the theory of LAE (or median) regression as found in [2, 3, 4]. That is, if $g(z) = 1$ and $B = \infty$, then our MLE method corresponds with LAE regression. In this case, $Q = 2$, $\nu = p^2$, $\zeta = -p^2$, $f(z) = (p/2)e^{-p|z|}$ (in one dimension) and so

$$\frac{\nu}{\zeta^2} = \frac{p^2}{p^4} = \frac{1}{p^2} = \frac{1}{(2f(0))^2}, \tag{7.1}$$

which is the asymptotic variance of the ordinary sample median for f [2].

We present an original and practical method of obtaining the covariance matrix for the model coefficients. This involves evaluating $(\boldsymbol{X}_{n,m}^T \boldsymbol{X}_{n,m})^{-1}$, where although n might be large, m generally is not, and using generalized functions to numerically evaluate two one-dimensional integrals (to find ν and ζ). The calculation of ν and ζ takes into account the characteristics of the response function (truncation and perturbation) which would be ignored if we used median regression. This is possible when we know the response function parameters, or have fairly accurate estimates, as in our epigenetic application, modelling DNA methylation proportion deviations.

For LAE regression, other methods of determining approximations to this covariance matrix may be found in the literature. In particular, in the method of quantile regression [5] implemented in the statistical package R, the covariance matrix for the MLE is calculated by resampling techniques, by bootstrapping or by using hierarchical spline models [13].

We prove that, even for truncated and perturbed Laplace response functions, subject to certain restrictions, the maximum of the log-likelihood function occurs at a data point. This result is well-known in the case of LAE regression. A proof that the LAE estimator passes through at least $r_{\boldsymbol{X}}$ data points may be found in references [3, 4].

Three asymptotically equivalent test statistics for LAE regression may be found in [14], namely a likelihood ratio test statistic, a Wald test statistic, and a Lagrange multiplier test statistic. Our likelihood ratio test statistic is an original modification of the former, applicable to our general case (not restricted to LAE regression), calculated using generalized functions. An F test statistic for LAE regression is found in [4].

When working with a model for which the response function is assumed to be a truncated Laplace probability density function, we could ignore the truncation to $[-B, B]$. However, taking the truncation into account reduces the variance in the model coefficient estimates by a factor $(1 - e^{-pB})^2$ and increases the log-likelihood ratio statistic by a factor $(1 - e^{-pB})^{-1}$. This effect is small if e^{-pB} is small but becomes more significant as pB decreases, that is, as more of the density function is truncated. Refer to §4.9, equation (4.32) and §4.11 equations (4.38) and (4.39) which show that, for example, when $g = g_1$,

$$D_{\text{gen}} = \frac{2}{(1 - e^{-pB})} \log(\lambda).$$

Hence, by taking into account the truncation, we can be more confident of our coefficient estimates and the value of appropriate beta coefficients than the standard formulae for LAE regression indicate. This effect will also be seen for small perturbations of the density function.

7.2. Summary

The Laplace distribution is the basis of many mathematical models (see [15]). Our focus has been modelling the distributions of errors in the proportions of

DNA methylation measured at genomic CpG sites.

Molecular biology deals with complex interactions both in terms of the physiology of the processes of interest and in the instrumentation required to measure these effects. The nonlinearity of these processes can result in frequency distributions that are far from normal, so that application of 'standard' methods of statistical inference based on least squares may be inadequate. Methods which deal with the form of the frequency distribution directly such as maximum likelihood are necessary for adequate inference to be made.

The Laplace or double exponential distribution considered here has been observed in molecular biology studies where a significant proportion of high deviations appear to occur regularly [16, 17]. The extension by Hermite polynomials considered here provides flexibility for describing the tails of the distribution. However, as noted, the use of the Laplace distribution as the 'key' function introduces problems in finding maximum likelihood estimators, and particularly their standard errors. This paper presents both a practical method for dealing with these problems, and the underlying asymptotic theory.

Appendix A: Useful convex analysis results

We prove Lemma 3.1 by applying results from convex analysis ([8]). The following definitions are taken from [8]. A face of a convex set $C \subset \mathbb{R}^n$ is a convex subset C' of C such that every (closed) line segment in C with a relative interior point in C' has both endpoints in C'. The empty set and C itself are faces of C. The zero-dimensional faces of C are called the extreme points of C. The relative interior of a convex set $C \subset \mathbb{R}^n$ is defined as the interior which results when C is regarded as a subset of its affine hull. The affine hull of a set $S \subset \mathbb{R}^n$ is the unique smallest affine set containing S. An alternative definition of an extreme point of a convex set C is a point $z \in C$ that cannot be written as $z = \theta u + (1-\theta)v$ with $0 < \theta < 1$, $u \in C$, $v \in C$, and $u \neq v$ [18, p686].

Theorem A.1. *Let C be a compact convex set in $\mathbb{R}^n$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a linear function. The maximum and minimum of f are attained at extreme points of C.*

Theorem A.1 (on the Maximum/Minimum Property ([18])) is useful but does not give a complete characterisation of the set of points in C at which f has a maximum.

Rockafellar [8] gives a more general definition of a convex function than we require. It is enough for our purposes to say that if the domain of real-valued function f is a convex set in $\mathbb{R}^n$ and if for any u and v in this domain, $f(\lambda u + (1-\lambda)v) \leq \lambda f(u) + (1-\lambda)f(v)$ for all $\lambda \in [0, 1]$, then f is convex.

Theorem A.2. (Theorem 32.1 [8]) *Let f be a convex function, and let C be a convex set contained in the effective domain of f. If f attains its supremum relative to C at some point of the relative interior of C, then f is actually constant throughout C.*

TABLE 3
First simulation data, treatments either H ($x_i = 1$) or L ($x_i = -1$) plus randomly sampled methylation deviances and randomly sampled uniformly distributed individual variation

T	y_i	T	y_i	T	y_i	T	y_i
L	0.4511	L	0.4433	H	0.4866	H	0.4503
L	0.4548	L	0.4608	H	0.7061	H	0.3597
L	0.4054	L	0.4540	H	0.4712	H	0.4933
L	0.4485	L	0.4533	H	0.4980	H	0.4621
L	0.4610	L	0.4736	H	0.4705	H	0.5074
L	0.4589	L	0.4426	H	0.4638	H	0.4340
L	0.5184	L	0.4700	H	0.4831	H	0.4698
L	0.4597	L	0.4434	H	0.4326	H	0.4968
L	0.4360	L	0.4507	H	0.4758	H	0.5190
L	0.4340	L	0.4712	H	0.4686	H	0.4584

For our purposes, the effective domain of f is the domain of f since the functions we consider are finite-valued. (See [8] for definitions.)

Corollary A.1. (Corollary 32.1.1 [8]) *Let f be a convex function, and let C be a convex set contained in the effective domain of f. Let W be the set of points (if any) where the supremum of f relative to C is attained. Then W is a union of faces of C.*

Corollary A.2. (Corollary 32.3.2 [8]) *Let f be a convex function, and let C be a non-empty closed bounded convex set contained in the relative interior of the effective domain of f. Then the supremum of f relative to C is finite, and it is attained at some extreme point of C.*

Theorem A.3. (Theorem 5.4 [8]) *Let f be a twice continuously differentiable real-valued function on a open convex set C in $\mathbb{R}^n$. Then f is convex on C if and only if its Hessian matrix is positive semidefinite for every $z \in C$.*

Appendix B: Data sets for §5

The simulated high (H) and low (L) treatment data analysed in §5.2 are given in Table 3. The CpG methylation measurements analysed in §5.3 are given in Table 5.

Appendix C: Chebyshev's Theorem

Theorem C.1. *Let ζ_1, ζ_2, be random variables, and let m_n and σ_n denote the mean and standard deviation of ζ_n. If $\sigma_n \to 0$ as $n \to \infty$, then $\zeta_n - m_n$ converges in probability to zero [19].*

Acknowledgements

The authors wish to acknowledege funding support provided by the National Research Centre for Growth and Development, New Zealand (GW, AP, AS), and

TABLE 4
Second simulation data, treatments either H ($x_i = 1$) or L ($x_i = -1$) plus randomly sampled methylation deviances and randomly sampled uniformly distributed individual variation

T	y_i	T	y_i	T	y_i	T	y_i
L	0.7204	L	0.4592	H	0.4837	H	0.5066
L	0.4784	L	0.4215	H	0.4829	H	0.3554
L	0.4524	L	0.5057	H	0.5371	H	0.4643
L	0.4490	L	0.4698	H	0.4771	H	0.6453
L	0.4554	L	0.4694	H	0.4799	H	0.4773
L	0.5179	L	0.4614	H	0.4856	H	0.4731
L	0.4379	L	0.4455	H	0.4255	H	0.5133
L	0.4955	L	0.4326	H	0.4993	H	0.4979
L	0.4782	L	0.4431	H	0.5298	H	0.4889
L	0.3303	L	0.4454	H	0.4617	H	0.4603

TABLE 5
CpG methylation measurements at sites 9 and 13 on the promoter of the H19 gene versus primiparous (p) or multiparous (m)

CpG9	p/m	CpG9	p/m	CpG13	p/m	CpG13	p/m
1.00	p	0.16	p	0.30	p	0.16	p
0.08	p	0.19	p	0.00	p	0.36	p
0.04	p	0.15	p	0.03	p	0.02	p
0.17	p	0.35	p	0.25	p	0.60	p
0.46	p	0.04	p	0.80	p	0.01	p
1.00	p	0.27	p	0.71	p	0.70	p
0.18	p	0.32	p	0.17	p	0.56	p
0.33	m	0.37	p	0.56	m	0.70	p
0.28	m	0.05	p	0.40	m	0.00	p
0.82	m	0.39	p	0.57	m	0.61	p
0.20	p	0.07	p	0.18	p	0.02	p
0.08	p	0.17	p	0.03	p	0.23	p
0.15	p	0.14	m	0.09	p	0.99	m
1.00	p	0.61	m	0.96	p	0.60	m
0.10	m	0.53	m	0.79	m	0.35	m
0.89	m	0.45	m	0.83	m	0.63	m
0.07	m	0.09	m	0.02	m	0.07	m
0.62	m	0.57	m	0.38	m	0.53	m
0.48	m	0.73	m	0.68	m	0.72	m
0.31	m	0.30	m	0.27	m	0.22	m
0.62	m			0.80	m		

the Foundation of Research Science and Technology, New Zealand (UOAX0808, AS). Further, we acknowledge our collaborative link with the GUSTO birth cohort, led by Professors P. D. Gluckman, University of Auckland, and Yap-Seng Chong, National University of Singapore.

References

[1] Dobson, A. J. & Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. 3rd edn. Chapman and Hall/CRC Press.

[2] Bassett, G. & Koenker, R.(1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association* **73**, Number 363, Theory and Methods Section, 618-662.

[3] Birkes, D. & Dodge Y. (1993). *Alternative Methods of Regression*. John Wiley and Sons, Inc.

[4] Bloomfield, P. & Steiger W. (1983). *Least Absolute Deviations, Theory Applications and Algorithms*. Birkhauser, Boston, Inc.

[5] Koenker, R. & Bassett, G. (1978). Regression quantiles. *Econometrica* **46**, No. 1, 33-50.

[6] Narula S. C. & Wellington J. F. (1982). The minimum sum of absolute errors regression : A state of the art survey. *International Statistical Review* **50** (3), 317-326.

[7] Norton, R. M. (1984). The double exponential distribution: using calculus to find a maximum likelihood estimator. *The American Statistician* **38**:(2), 135-136.

[8] Rockafellar, R. T. (1970). *Convex Analysis*. 10th edn. Princeton: Princeton University Press.

[9] Lighthill, M. J. (1958). *Introduction to Fourier Analysis and Generalized Functions*. Cambridge: Cambrige University Press.

[10] Stakgold, I. (1967). *Boundary value problems of mathematical physics, Vol. 1*. New York: Macmillan.

[11] Ehrich, M., Nelson, M. R., Stanssens, P., Zabeau, M., Liloglou, T., Xinarianos, G., Cantor, C. R., Field, J. K. & van den Boom, D. (2005). Quantitative high-temperature analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proc Natl Acad Sci USA* **102**(44), 15785-15790.

[12] Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52** (3-4): 591-611.

[13] Hendricks, W. & Koenker, R.(1978). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association* **87**, Number 417, 58-68.

[14] Koenker, R. & Bassett, G. (1982). Tests of linear hypotheses ond l_1 estimation. *Econometrica* **50**, 1157-1583.

[15] Kotz, S., Kozubowski, T. J. & Podgorski, K. (2001). *The Laplace Distribution and Generalizations*. Birkhauser, Boston.

[16] Purdom, E. & Holmes, S. P.(2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology* **4**(1), Article 16.
[17] Pleasants, et al., unpublished observations.
[18] Kincaid, D. & Cheney W. (2002). *Mathematics of Scientific Computing.* 3rd edn. Brooks/Cole.
[19] Cramer, H. (1958). *Mathematical methods of statistics.* Princeton University Press.

Linear models with perturbed and truncated Laplace response functions: The asymptotic theory of MLE with application to epigenetics

Hassell-Sweatman CZW

2012-11-20