Supporting Theory and Data Analysis for "Likelihood Inference in Exponential Families and Directions of Recession"

By

Charles J. Geyer

Technical Report No. 672
School of Statistics
University of Minnesota
September 29, 2008

Abstract

When in a full exponential family the maximum likelihood estimate (MLE) does not exist, the MLE may exist in the Barndorff-Nielsen completion of the family (Barndorff-Nielsen, 1978; Brown, 1986; Geyer, 1990). A practical algorithm for finding the MLE in the completion using repeated linear programming was proposed in the author's unpublished thesis (Geyer, 1990) and used in Geyer and Thompson (1992). Now we propose a slightly different method, also using repeated linear programming with the R contributed package `rcdd` (Geyer and Meeden, 2008), which makes straightforward the calculation of the MLE in the Barndorff-Nielsen completion for any models satisfying a condition of Brown (1986) and for which some R function can calculate the MLE when it does exist, for example, generalized linear models (GLM) and aster models (Geyer et al., 2007; Geyer, 2008). In this technical report we give details of two GLM examples.

Likelihood ratio tests of model comparison are almost unchanged from the usual case. Only the degrees of freedom need be adjusted when the MLE for the null hypothesis lies in the completion rather than the original family. Confidence intervals are changed much more. When the MLE for the natural parameter does not exist, it can be thought of as having gone to infinity in a certain direction, which we call a generic direction of recession. Here we propose a new kind of one-sided confidence interval, not involving asymptotic approximation, for how close to infinity the true unknown natural parameter value may be. This maps to a one-sided confidence interval for the mean value parameter showing how close to the boundary of its support it may be.

1 R Package Rcdd

We use the R statistical computing environment (R Development Core Team, 2008) in our analysis. It is free software and can be obtained from `http://cran.r-project.org`. Precompiled binaries are available for Windows, Macintosh, and popular Linux distributions. We use the contributed package `rcdd`. If R has been installed, but this package has not yet been installed, do

```
install.packages("rcdd")
```

from the R command line (or do the equivalent using the GUI menus if on Macintosh or Windows). This may require root or administrator privileges.

If the `rcdd` package has been installed, we load it

```
> library(rcdd)
```

The version of the package used to make this document is 1.1-1. The version of R used to make this document is 2.7.2.

This entire document and all of the calculations shown were made using the R command `Sweave` and hence are exactly reproducible by anyone who has R and the R noweb (RNW) file from which it was created. Both the RNW file and and the PDF document produced from it are available at `http://www.stat.umn.edu/geyer/gdor`.

2 Introduction

When in a full exponential family the maximum likelihood estimate (MLE) does not exist, the MLE may exist in the Barndorff-Nielsen completion of the family (Barndorff-Nielsen, 1978; Brown, 1986; Geyer, 1990, and references cited at the beginning of Chapter 2 therein). In this case the MLE in the completion can be thought of as a pair $(\hat{\theta}, \delta)$ satisfying

$$\lim_{s \to \infty} l(\hat{\theta} + s\delta) = \sup_{\theta \in \Theta} l(\theta). \tag{1}$$

Although $\hat{\theta}$ and δ satisfying (1) are not necessarily unique, the distribution which is the limit as $s \to \infty$ of distributions having parameter values $\hat{\theta} + s\delta$ is unique.

A more complicated but also more interesting characterization of δ is that it is an element of the relative interior of the normal cone $N_C(y)$ of the convex support C at the point y, where C is the smallest closed convex set that contains the natural statistic Y almost surely for one of the distributions in the family and hence for all of them and y is the observed value of Y. This makes determination of δ an exercise in computational geometry. When C is a polyhedral convex set, δ can be determined using the R functions `linearity` and `lpcdd` in the `rcdd` contributed package (Geyer and Meeden, 2008) for the R statistical computing environment (R Development Core Team, 2008).

A more complicated but also more interesting characterization of $\hat{\theta}$ is that it is the MLE in the limiting conditional model, which is the family of distributions that are limits in distribution as $s \to \infty$ of distributions having parameters $\theta + s\delta$ for all $\theta \in \Theta$. This limiting conditional model, described in Section 3.4 below, is an exponential family, the given exponential family conditioned on the event $\langle Y - y, \delta \rangle = 0$, where Y is the natural statistic, y is its observed value, and $\langle \,\cdot\,, \,\cdot\, \rangle$ is the bilinear form given by (3) below. Thus computation of $\hat{\theta}$ is maximum likelihood estimation in an exponential family and can often be done using available software.

Some regularity conditions are necessary for the above theory to be correct. They are given in Section 3.7 below. These regularity conditions are not restrictive, being satisfied by all applications known to me involving full exponential families. Geyer (1990) has algorithms that work on non-full but convex families, meaning the set of natural parameter values is a convex subset of the natural parameter space that is closed relative to it, and Geyer and Thompson (1992) have an application of such families, but we do not discuss them here.

So what is new here? Geyer (1990) provided a method of finding the MLE in the Barndorff-Nielsen completion that was effective and, like the methods recommended here, was based on repeated linear programming. However, that work was never published, partly because it depended on high-quality linear programming software, which was not easy for statisticians to use and certainly not available in widely used statistical computing environments. With the advent of the `rcdd` package (Geyer and Meeden, 2008), high quality linear programming is now available in R, so it was time to revisit the issue. Some new ideas have been added, so we do not closely follow Geyer (1990). In particular, the algorithms presented in Sections 3.11 and 3.12 below, which are the heart of our methodology, differ from those of Geyer (1990). Finally, the hypothesis tests and confidence intervals proposed in Sections 3.15 and 3.16 below are new, although we cannot claim priority for our proposal for testing, which was suggested by S. Fienberg (personal communication).

Section 4 below contains explicit examples that show how all calculations related to our methodology are carried out in R.

3 Theory

3.1 Exponential Families

An exponential family of distributions (Barndorff-Nielsen, 1978; Brown, 1986; Geyer, 1990) is a statistical model having log likelihood

$$l(\theta) = \langle y, \theta \rangle - c(\theta), \tag{2}$$

where y is a vector statistic, θ is a vector parameter, and $(y, \theta) \mapsto \langle y, \theta \rangle$ is a bilinear form. In all computations, we will assume y and θ are elements of $\mathbb{R}^p$ and

$$\langle y, \theta \rangle = \sum_{i=1}^{p} y_i \theta_i, \tag{3}$$

but in theory any bilinear form works. A statistic y and parameter θ that give a log likelihood of this form are called *natural* or *canonical.* We will say *natural.* The function c is called the *cumulant function* of the family.

The distribution with parameter value θ has a density with respect to the distribution with parameter value ψ of the form

$$f_\theta(\omega) = e^{\langle Y(\omega), \theta - \psi \rangle - c(\theta) + c(\psi)}. \tag{4}$$

The requirement that this integrate to one determines the function c up to an additive constant

$$c(\theta) = c(\psi) + \log E_\psi\bigl(e^{\langle Y, \theta - \psi \rangle}\bigr). \tag{5}$$

We take (5) to be valid for all θ in $\mathbb{R}^p$, defining $c(\theta) = \infty$ for θ such that the expectation in (5) does not exist. Since the argument of the expectation operator in (5) is strictly positive, so is the expectation; hence the cumulant function takes values that are either real or $+\infty$ and the log likelihood function takes values that are either real or $-\infty$. Define

$$\Theta = \{\, \theta \in \mathbb{R}^p : c(\theta) < \infty \,\}. \tag{6}$$

The exponential family is *full* if its natural parameter space is (6). We shall be interested only in full exponential families.

3.2 Convex Support, Tangent Cone, and Normal Cone

The *convex support* of an exponential family is the smallest closed convex set that contains the natural statistic with probability one under some distribution in the family (Barndorff-Nielsen, 1978, p. 90), in which case this is true for all distributions in the family, because the distributions are mutually absolutely continuous because the densities (4) are everywhere nonzero.

The *tangent cone* of a convex set C at a point $y \in C$ is

$$T_C(y) = \operatorname{cl}\{\, s(w - y) : w \in C \text{ and } s \geq 0 \,\}, \tag{7}$$

where cl denotes the closure operation (Rockafellar and Wets, 2004, Theorem 6.9). The *normal cone* of a convex set C in $\mathbb{R}^p$ at a point $y \in C$ is

$$N_C(y) = \{\, \delta \in \mathbb{R}^p : \langle w - y, \delta \rangle \leq 0 \text{ for all } w \in C \,\}. \tag{8}$$

(Rockafellar and Wets, 2004, Theorem 6.9). Tangent and normal cones are polars of each other (Rockafellar and Wets, 2004, Theorem 6.9 and Corollary 6.30). Each determines the other.

3.3 Directions of Recession and Constancy

Directions of recession and constancy of convex and concave functions are defined by Rockafellar (1970, p. 69). We apply these notions to log likelihoods of full exponential families.

Proofs of all theorems are given in Section 6.

Theorem 1. *For some vector δ and for a full exponential family with log likelihood* (2)*, natural parameter space Θ, convex support C, natural statistic Y, and observed value of the natural statistic y such that $y \in C$, the following are equivalent.*

(a) *There exists a $\theta \in \Theta$ such that $s \mapsto l(\theta + s\delta)$ is not a strictly concave function on the interval where it is finite.*

(b) *For all $\theta \in \Theta$ the function $s \mapsto l(\theta + s\delta)$ is constant on $\mathbb{R}$.*

(c) *The parameter values θ and $\theta + s\delta$ correspond to the same probability distribution for some $\theta \in \Theta$ and some $s \neq 0$.*

(d) *The parameter values θ and $\theta + s\delta$ correspond to the same probability distribution for all $\theta \in \Theta$ and all real s.*

(e) $\langle Y - y, \delta \rangle = 0$ *almost surely for some distribution in the family.*

(f) $\langle Y - y, \delta \rangle = 0$ *almost surely for all distributions in the family.*

(g) $\delta \in N_C(y)$ *and* $-\delta \in N_C(y)$.

(h) $\langle w, \delta \rangle = 0$*, for all $w \in T_C(y)$.*

Any vector δ that satisfies any one of the conditions of the theorem (and hence all of them) is called a *direction of constancy* of the log likelihood. The set of all directions of constancy is called the *constancy space* of the log likelihood. It is clear from (e) or (h) of the theorem that the constancy space is a vector subspace.

Corollary 2. *For a full exponential family, suppose $\hat{\theta}_1$ and $\hat{\theta}_2$ are maximum likelihood estimates. Then $\hat{\theta}_1 - \hat{\theta}_2$ is a direction of constancy.*

From the corollary and (d) of the theorem, we see that directions of constancy do not cause any problem for statistical inference, because all maximum likelihood estimates correspond to the same probability distribution. Thus we have uniqueness where it is important. Nonuniqueness of the MLE for the natural parameter is, at worst, merely a computational nuisance.

A family is said to be *minimal* if it has no directions of constancy. This can always be arranged by reparametrization (Barndorff-Nielsen, 1978, pp. 111–116; Brown, 1986, pp. 13–16; see also Geyer, 1990, Section 1.5). The R function `glm` always uses a minimal parametrization, dropping predictors to obtain a full rank model matrix. We take the view, however, that minimality is not necessary and that insisting on minimality can complicate other theoretical issues. Thus we never insist on minimality and do allow for directions of constancy.

Theorem 3. *For some vector δ and for a full exponential family with log likelihood* (2)*, natural parameter space Θ, convex support C, natural statistic Y, and observed value of the natural statistic y such that $y \in C$, the following are equivalent.*

(a) *There exists a $\theta \in \Theta$ such that the function $s \mapsto l(\theta + s\delta)$ is nondecreasing on $\mathbb{R}$.*

(b) *For all $\theta \in \Theta$ the function $s \mapsto l(\theta + s\delta)$ is nondecreasing on $\mathbb{R}$.*

(c) *$\langle Y - y, \delta \rangle \leq 0$ almost surely for some distribution in the family.*

(d) *$\langle Y - y, \delta \rangle \leq 0$ almost surely for all distributions in the family.*

(e) *$\delta \in N_C(y)$.*

(f) *$\langle w, \delta \rangle \leq 0$, for all $w \in T_C(y)$.*

Any vector δ that satisfies any one of the conditions of the theorem (and hence all of them) is called a *direction of recession* of the log likelihood. From now on we will simply say direction of recession or constancy to refer to directions of recession or constancy of the log likelihood (since we will not be interested in directions of recession or constancy of any other function). Note that every direction of constancy is a direction of recession.

In light of (e) of the theorem, we could also call directions of recession normal vectors (of the convex support at the observed value of the natural statistic). Since the word "normal" is already overused in statistics, we prefer the term that cannot be confused with other statistical notions.

Theorem 4. *For a full exponential family with convex support C and observed value of the natural statistic y such that $y \in C$, the following are equivalent.*

(a) *The MLE exists.*

(b) *Every direction of recession is a direction of constancy.*

(c) *$N_C(y)$ is a vector subspace.*

(d) *$T_C(y)$ is a vector subspace.*

This theorem provides a complete geometric solution to the problem of when the MLE exists in a full exponential family.

Corollary 5. *For a full exponential family with log likelihood* (2)*, natural parameter space Θ, convex support C, and observed value of the natural statistic y such that $y \in C$, if δ is a direction of recession that is not a direction of constancy, then for all $\theta \in \Theta$ the function $s \mapsto l(\theta + s\delta)$ is strictly increasing on the interval where it is finite.*

3.4 Limits in Directions of Recession

Theorem 6. *For a full exponential family having log likelihood* (2)*, densities* (4)*, natural statistic Y, observed value of the natural statistic y such that y is in the convex support, and natural parameter space Θ, if δ is a direction of recession that is not a direction of constancy,*

$$H = \{ w \in \mathbb{R}^p : \langle w - y, \delta \rangle = 0 \}, \tag{9}$$

and $\Pr(Y \in H) > 0$ *for some distribution in the family, and hence for all, then for all* $\theta \in \Theta$

$$\lim_{s \to \infty} f_{\theta + s\delta}(\omega) = \begin{cases} 0, & \langle Y(\omega) - y, \delta \rangle < 0 \\ f_\theta(\omega) / \Pr_\theta(Y \in H), & \langle Y(\omega) - y, \delta \rangle = 0 \\ +\infty, & \langle Y(\omega) - y, \delta \rangle > 0 \end{cases} \tag{10}$$

Moreover $s \mapsto \Pr_{\theta + s\delta}(Y \in H)$ *is continuous and strictly increasing, and* $\Pr_{\theta + s\delta}(Y \in H) \to 1$ *as* $s \to \infty$.

We note three things about the right-hand side of (10). First, it is a probability density with respect to the distribution having parameter value ψ. The set where it is $+\infty$ has probability zero by Theorem 3 (d), so this is not a problem. Second, it is the density of the conditional distribution given the event $Y \in H$ of the distribution having parameter value θ. Third, by Scheffé's lemma (Lehmann, 1959, p. 351) pointwise convergence of densities implies convergence in total variation, which implies convergence in distribution.

Denote the right-hand side of (10) by $f_\theta(\omega \mid Y \in H)$. It is clear that the family

$$\{ f_\theta(\cdot \mid Y \in H) : \theta \in \Theta \} \tag{11}$$

is an exponential family with the same natural statistic and natural parameter as the original family. Moreover, it is clear that the log likelihood for this conditional family

$$l_H(\theta) = \langle y, \theta \rangle - c(\theta) - \log \Pr_\theta(Y \in H)$$

satisfies

$$l(\theta) < l_H(\theta), \qquad \theta \in \Theta.$$

Thus, if an MLE exists for the conditional family (11), then it maximizes the likelihood in the family that is the union of (11) and the original family. When this happens, we say we have found an MLE in the Barndorff-Nielsen completion of the original exponential family.

If the conditional family (11) has a direction of recession that is not a direction of constancy, we can take limits again, and continue until we have a limiting family that has no directions of recession that are not directions of constancy, which must eventually occur since the dimension of the convex support of the limiting family must decrease in each limiting operation and can go no lower than zero, in which case the convex support contains a single point and the conditional family contains a single distribution which is trivially the MLE. This iterated limiting process is the one followed in Geyer (1990). It is actually much more general than the one we follow here.

Up to this point, we have been closely following Geyer (1990), but now we part company, imposing conditions that assure δ can be chosen so that one limit is enough. So far the only condition we have imposed is $\Pr(Y \in H) > 0$, which is required for the existence of the limiting conditional family (11). We will impose other conditions as we go along. All of the conditions we assume will be collected in one place in Section 3.7. Most of these conditions are not new, having been proposed by Brown (1986, pp. 193 and 197). Although these conditions limit the applicability of our procedure, they do apply in most applications, and they make construction of confidence intervals much easier.

The limiting conditional family (11) need not be full; the natural parameter space of the full family containing (11) is at least as large as

$$\Theta + \Gamma_{\text{lim}} = \{\, \theta + \gamma : \theta \in \Theta \text{ and } \gamma \in \Gamma_{\text{lim}} \,\}, \tag{12}$$

where Θ is the natural parameter space of the original family and Γ_{lim} is the constancy space of (11). We will assume that (12) is the natural parameter space of the full family containing (11). Although pathological examples can be constructed for which this assumption fails (Geyer, 1990, Example 2.1), we know of no realistic applications for which it fails.

3.5 Convex Polyhedra

One part of the condition of Brown (1986, p. 197) mentioned above is that the convex support is polyhedral (Rockafellar, 1970, Section 19). Since most applications satisfy this condition, it entails little loss of generality.

A set C is a *convex polyhedron* if it is the intersection of a finite collection of closed half-spaces (Rockafellar, 1970, Section 19), that is,

$$C = \{\, w \in \mathbb{R}^d : \langle w, \alpha_i \rangle = b_i,\ i \in E, \text{ and } \langle w, \alpha_i \rangle \le b_i,\ i \in I \,\}, \tag{13}$$

where the α_i are nonzero vectors, the b_i are scalars, and E and I are disjoint finite sets.

There is an alternative characterization of convex polyhedra; they are convex hulls of finite sets of points and directions, that is, sets of all linear combinations

$$\sum_{i \in E \cup I} b_i \alpha_i, \tag{14}$$

where the α_i are vectors, the b_i are scalars, E and I are disjoint finite sets, and the b_i satisfy

$$b_i \ge 0, \qquad i \in E \cup I \tag{15a}$$

and if I is nonempty

$$\sum_{i \in I} b_i = 1. \tag{15b}$$

The vectors α_i, $i \in E \cup I$ are also called the *generators* of C and C the set *generated* by the *points* α_i, $i \in I$ and the *directions* α_i, $i \in E$.

The equivalence of these two characterizations is called the Minkowski-Weyl theorem (Rockafellar, 1970, Theorem 19.1). The R function `scdd` in the contributed package `rcdd` (Geyer and Meeden, 2008) converts between these two representations of a convex polyhedron, which it calls the H-representation and V-representation, respectively.

Let P denote the set $\{ \alpha_i : i \in I \}$ of points and D denote the set $\{ \alpha_i : i \in E \}$ of directions. When $P \neq \varnothing$, we write

$$C = \operatorname{con}(P) + \operatorname{con}(\operatorname{pos} D)$$

to denote the relationship between the convex polyhedron C and its sets of generators P and D. When $P = \varnothing$, we write

$$C = \operatorname{con}(\operatorname{pos} D)$$

to denote the relationship between C and D. This notation follows Rockafellar and Wets (2004, Sections 2E and 3G).

When C is a convex polyhedron, $N_C(y)$ and $T_C(y)$ are also convex polyhedrons for each $y \in C$ and are given in terms of the H-representation of C by simple formulas (Rockafellar and Wets, 2004, Theorem 6.46). Moreover, the closure operation in (7) is unnecessary when C is a convex polyhedron. When $T_C(y)$ or any convex cone is polyhedral, its V-representation can consist of directions only, so can be of the form $\operatorname{con}(\operatorname{pos} V)$ for some finite set V.

3.6 Generic Directions of Recession

The *relative interior* of a convex set C, denoted $\operatorname{rint} C$, is its interior relative to its affine hull (Rockafellar, 1970, Chapter 6). Every nonempty convex set has a nonempty relative interior (Rockafellar, 1970, Theorem 6.2).

We say a vector δ is a *generic direction of recession* (GDOR) if $\delta \in \operatorname{rint} N_C(y)$ and $N_C(y)$ is not a vector subspace, where C is the convex support and y an observed value of the natural statistic such that $y \in C$. Since the relative interior is always nonempty, a GDOR exists if and only if none of the conditions of Theorem 4 hold.

Theorem 7. *For a full exponential family having polyhedral convex support C and observed value of the natural statistic y such that $y \in C$, let $T_C(y) = \operatorname{con}(\operatorname{pos} V)$, and define*

$$L = \{ v \in V : -v \in T_C(y) \}.$$

Then a generic direction of recession exists if and only if $L \neq V$, in which case a vector δ is a generic direction of recession if and only if

$$\langle w, \delta \rangle = 0, \qquad w \in L \tag{16a}$$

$$\langle w, \delta \rangle < 0, \qquad w \in V \setminus L \tag{16b}$$

Corollary 8. *Under the assumptions of the theorem, a generic direction of recession is not a direction of constancy.*

If B is a set of vectors, let $\operatorname{span} B$ denote the smallest vector subspace containing B. Also for any vector x, let $x + \operatorname{span} B = \{ x + v : v \in \operatorname{span} B \}$.

Corollary 9. *Under the assumptions of the theorem, suppose δ is a generic direction of recession, and H is defined by (9). Then $T_{C \cap H}(y) = \operatorname{span} L$, and $C \cap H = C \cap (y + \operatorname{span} L)$.*

The theorem and corollaries explain the purpose of generic directions of recession. By Corollary 8, a GDOR implies the MLE does not exist in the conventional sense, so we seek it in the limiting family described in (3.4). Suppose that $C \cap H$ is the convex support of the limiting family. This is another part of the conditions of Brown (1986, pp. 193) referred to above. Then $T_{C \cap H}(y)$ being a vector subspace implies that the MLE in the limiting family exists by Theorem 4 (c). Thus finding one GDOR allows us to find the MLE in the Barndorff-Nielsen completion.

At this point our theory is essentially complete; only computational issues remain. So we take some time to explain the connection between this theory and the pre-existing theory of Barndorff-Nielsen completion (Barndorff-Nielsen, 1978; Brown, 1986; Geyer, 1990). The pre-existing theory says the MLE lies in the limiting conditional family whose convex support, what we are calling $C \cap H$, is the unique face of C containing y in its relative interior (Geyer, 1990, Chapter 4 generalizes this). Thus the pre-existing approach makes it clear that the limiting conditional family containing the MLE is unique and does not depend on the GDOR, which is in general not unique. In our approach, uniqueness comes from the assertion $C \cap H = C \cap (y + \operatorname{span} L)$ in Corollary 9. This makes it clear that, although the hyperplane H does depend on the GDOR δ used to define it, the convex support $C \cap H$ of the limiting distribution does not depend on H, hence does not depend on δ. Since we do not need the pre-existing theory, we shall not bother with a proof that $C \cap H$ is the face of C containing y in its relative interior (the proof is almost immediate from Lemma A.1 in Geyer, 1990), and merely assure the reader that this is indeed the case, and our new theory describes the same mathematical structure as the pre-existing theory. We have not rewritten the theory of Barndorff-Nielsen completion merely for amusement. As we shall see, the new theory based on the GDOR concept is much better suited for computation and statistical inference than the pre-existing theory based on faces of a convex set.

Before moving to computational issues, we would like to write down somewhere one interesting issue. Corollary 9 would be false without the assumption that C is polyhedral. Consider the following example. C is the set in $\mathbb{R}^2$ consisting of the closed unit disk and all points with one coordinate negative and the other less than or equal to one. This is a closed convex set, but not polyhedral. The points $(0, 1)$ and $(1, 0)$ are peculiar in that they are faces of C that are not exposed in the terminology of Rockafellar (1970, pp. 162–163). Suppose $y = (1, 0)$. Then $N_C(y) = \{\, (s, 0) : s \geq 0 \,\}$. Since this normal cone is not a vector subspace, there is a generic direction of recession; one is $\delta = (1, 0)$. Then $H = \{\, (1, s) : s \in \mathbb{R} \,\}$ and $C \cap H = \{\, (1, s) : s \leq 0 \,\}$. Then $T_{C \cap H}(y) = \{\, (0, s) : s \leq 0 \,\}$ is not a vector subspace. In this sort of situation the more general theory of Geyer (1990) may apply, but our theory based on generic directions of recession cannot. The condition that C be polyhedral could clearly be weakened to C being locally polyhedral (every point has a convex polyhedral neighborhood B such that $B \cap C$ is polyhedral), but we know of no application of such a condition, hence do not use it.

3.7 Assumptions

We summarize the assumptions we have made above. We deal with a full exponential family. If every direction of recession is a direction of constancy, then we need no further assumptions. Otherwise, let δ be a generic direction of recession, let C be the convex support, let Y be the natural statistic, let y be an observed value of the natural statistic satisfying $y \in C$, and let H be defined by (9). We assume the event $Y \in H$ has positive probability so the limiting conditional family defined in Section 3.4 exists. We further assume that $C \cap H$ is the convex support of this limiting conditional family, so that by Theorem 9 the MLE in this limiting conditional family exists. We further assume that the natural parameter space of the full family containing the limiting conditional family is given by (12), so that the confidence interval construction in Section 3.16 below is valid. Finally, we assume that C is a convex polyhedron. We need no other assumptions.

3.8 Natural Affine Submodels

In most applications of exponential family theory, we start with a very large exponential family, which we call *saturated* and which has too many parameters to estimate well. Then we consider *natural affine submodels*, parametrized by

$$\theta = a + M\beta,$$

where θ is the natural parameter of the saturated model, β is the natural parameter of the natural affine submodel, a is a known vector, and M is a known matrix. In the terminology of the R function `glm`, a is called the *offset vector* and M is called the *model matrix*.

Observe that

$$\langle y, a + M\beta \rangle = \langle y, a \rangle + \langle M^T y, \beta \rangle,$$

where the two bilinear forms on the right-hand side have different dimensions. Since the first term on the right-hand side does not contain the parameter and can be dropped from the log likelihood, the submodel is itself an exponential family with natural statistic $M^T y$ and natural parameter β. Thus everything said above applies to natural affine submodels, we just work with the convex support of $M^T Y$ rather than of Y.

3.9 Relating Tangent Cones of Models and Affine Submodels

Let C_{sat} denote the convex support of the saturated model and C_{sub} that of the natural affine submodel. By Theorems 6.43 and 6.46 in Rockafellar and Wets (2004),

$$T_{C_{\text{sub}}}(M^T y) = \operatorname{cl}\{\, M^T w : w \in T_{C_{\text{sat}}}(y) \,\} \tag{17}$$

and the closure operation is not necessary if C_{sat} is polyhedral. Moreover, it is clear that if $T_{C_{\text{sat}}}(y) = \operatorname{con}(\operatorname{pos} V_{\text{sat}})$, then $T_{C_{\text{sub}}}(y) = \operatorname{con}(\operatorname{pos} V_{\text{sub}})$, where

$$V_{\text{sub}} = \{\, M^T w : w \in V_{\text{sat}} \,\}. \tag{18}$$

The induced mapping for normal cones is not so simple, requiring linear programming. This explains our starting with tangent vectors and V-representations rather than normal vectors and H-representations.

3.10 Tangent Cones of Saturated Models

In saturated families the convex support is often easy to calculate. In logistic regression, each component of the response vector is Bernoulli and $C_{\text{sat}} = [0, 1]^p$. In Poisson regression, each component of the response vector is Poisson and $C_{\text{sat}} = [0, \infty)^p$. In categorical data analysis with Poisson response, $C_{\text{sat}} = [0, \infty)^p$, just as in Poisson regression. In categorical data analysis with multinomial or product-multinomial response, the model can be derived from the Poisson model by conditioning on an affine subspace, hence the convex support is the intersection of $[0, \infty)^p$ with this affine subspace. As we shall see (Section 4.2.2 below), this allows us to use the solution in the Poisson response problem to calculate the solutions in the other problems.

When C_{sat} is a Cartesian product, as in the examples mentioned, $T_{C_{\text{sat}}}(y)$ can be calculated coordinatewise (Rockafellar and Wets, 2004, Proposition 6.41). This is the only situation we will use for our examples. Let e_i denote the unit vector in the i-th coordinate direction (every coordinate is zero except for the i-th, which is one). Then e_i is a tangent vector at y if y_i is not at the upper bound of its range, and $-e_i$ is a tangent vector at y if y_i is not at the lower bound. In this case, the set V_{sat} of the preceding section contains e_i or $-e_i$ or both for each i.

We will not do any examples where C_{sat} is not a Cartesian product, except for categorical data analysis with multinomial or product multinomial response, where we will derive the solution from the solution for the Poisson response case where C_{sat} is a Cartesian product. An application where the convex support is not a Cartesian product is provided by unconditional aster models (Geyer et al., 2007). We merely note that all of the theory developed in this technical report applies to case where C_{sat} is not a Cartesian product. Moreover, most of the computational procedures described below also apply to this case. Only at the beginning and the end of this process does the non-Cartesian-product case present additional issues. At the beginning we need to determine a set V_{sat} that generates $T_{C_{\text{sat}}}(y)$, and this may be more difficult than in the Cartesian product case. At the end (Section 3.13.2 below), we need to determine the convex support $C_{\text{sat}} \cap H_{\text{sat}}$ of the limiting conditional model and compute the MLE in this model, and this may also be more difficult than in the Cartesian product case.

3.11 Calculating the Linearity

Next we determine the *linearity* of V_{sub}

$$L_{\text{sub}} = \{\, w \in V_{\text{sub}} : -w \in \operatorname{con}(\operatorname{pos} V_{\text{sub}}) \,\}. \tag{19}$$

This sounds like a complicated operation, and it is, but the `rcdd` package has a function `linearity` that does it by repeated linear programming.

Having found the linearity, we have solved the problem of when the MLE exists in the original family. It exists if and only if $L_{\text{sub}} = V_{\text{sub}}$.

All functions in the `rcdd` package use two forms of arithmetic. One is the default computer arithmetic used by all other R functions. Answers produced using that arithmetic are inexact, so one is uncertain whether the L_{sub} produced is actually correct. The other form of arithmetic is exact, infinite-precision, rational arithmetic. Answers produced using that arithmetic are exact, so one is certain that the L_{sub} produced is actually correct, but only if the vectors in V_{sub} are also produced exactly using either integer arithmetic or rational arithmetic.

For readers curious about how the `linearity` function works we give the following description. Others should skip the rest of this section. According to comments in the source code (starting at line 3062 of the file `cddlp.c` of the source code for the `rcdd` package, version 1.1-1, which comes from the `cddlib` library, version 0.94f, written by K. Fukuda) for each $w \in V_{\text{sub}}$ it solves the linear programming problem

$$\begin{aligned} &\text{maximize} \\ &\qquad \langle w, \delta \rangle \\ &\text{subject to} \\ &\qquad \langle v, \delta \rangle \geq 0, \qquad v \in V_{\text{sub}} \setminus \{w\} \end{aligned} \tag{20}$$

where δ is the state vector of the linear programming problem.

Theorem 10. *A vector w is in the linearity* (19) *if and only if the optimal value of the linear program* (20) *is nonpositive.*

It may be the case that some $w \in V_{\text{sub}}$ are known a priori to be in L_{sub}. This can be specified in the input to the `linearity` function, so the work verifying this need not be done.

3.12 Calculating Generic Directions of Recession

If $L_{\text{sub}} \neq V_{\text{sub}}$, then $T_{C_{\text{sub}}}(M^T y)$ is not a vector subspace, hence there exists a generic direction of recession. By Theorem 7, δ is a GDOR if and only if

$$\langle w, \delta \rangle = 0, \qquad w \in L_{\text{sub}} \tag{21a}$$

$$\langle w, \delta \rangle < 0, \qquad w \in V_{\text{sub}} \setminus L_{\text{sub}} \tag{21b}$$

Hence we can find one such δ by solving the following linear programming problem

$$\begin{aligned} &\text{maximize} \\ &\qquad \epsilon \\ &\text{subject to} \\ &\qquad \qquad \epsilon \leq 1 \\ &\qquad \langle v, \delta \rangle = 0, \qquad v \in L_{\text{sub}} \\ &\qquad \langle v, \delta \rangle \leq -\epsilon, \qquad v \in V_{\text{sub}} \setminus L_{\text{sub}} \end{aligned}$$

where δ is a p-vector, ϵ is a scalar, and (δ, ϵ) is the state vector of the linear programming problem (so the dimension is $p + 1$). The δ part of the solution is a generic direction of recession. The ϵ part does not matter.

The idea for using this particular linear program came from the documentation for the `dd_ExistsRestrictedFace2` function in the `cddlib` library, which is the computational geometry library (written by K. Fukuda) to which `rcdd` provides an incomplete interface. The `rcdd` package does not provide an interface to this `cddlib` function, but it does provide a function `lpcdd` that does linear programming and can be used to solve this linear program.

3.13 Calculating Maximum Likelihood Estimates

3.13.1 In the Original Family

There is little to be said about calculating the MLE in the original family. When we have found that $L_{\text{sub}} = V_{\text{sub}}$, then we know the MLE exists and can use available software to find it. We will use the R function `glm` for our examples.

There is one issue worth mentioning. If the model is non-identifiable, so the MLE is non-unique, the R function `glm` is smart enough to drop enough predictors to produce an identifiable model. However, its method of doing so is not guaranteed because of inexactness of the default computer arithmetic.

We can use the `redundant` function in the `rcdd` package applied to the columns of M to reduce to a linearly independent set. If M was calculated using integer or rational arithmetic, and `redundant` uses rational arithmetic, then this operation will be exact.

3.13.2 In the Completion

When we have found that $L_{\text{sub}} \neq V_{\text{sub}}$ and have found a generic direction of recession δ, we still need to characterize the support of the limiting conditional family. We can characterize this two ways using either of

$$
\begin{aligned}
H_{\text{sub}} &= \{\, w \in \mathbb{R}^q : \langle w - M^T y, \delta \rangle = 0 \,\} \\
H_{\text{sat}} &= \{\, w \in \mathbb{R}^p : \langle w - y, M\delta \rangle = 0 \,\}
\end{aligned}
$$

where p and q are the dimensions of the saturated model and affine submodel, respectively. Then the limiting conditional model conditions on the event $M^T Y \in H_{\text{sub}}$ or $Y \in H_{\text{sat}}$, which is the same event characterized two different ways, the latter usually simpler.

Theorem 11. *In the setup of Sections 3.11 through 3.13, define*

$$L_{\text{sat}} = \{\, v \in V_{\text{sat}} : M^T v \in L_{\text{sub}} \,\}.$$

Then the support of the limiting conditional family is

$$C_{\text{sat}} \cap H_{\text{sat}} = C_{\text{sat}} \cap (y + \operatorname{span} L_{\text{sat}}) \tag{22}$$

when referred to the saturated model.

In the case where the convex support of the saturated model is a Cartesian product, the support of the limiting model simply constrains $Y_i = y_i$ for i such that $e_i \notin L_{\text{sat}}$, that is, the i-th component of the response is unconstrained if $i \in L_{\text{sat}}$ and is constrained to be equal to its observed value if $i \notin L_{\text{sat}}$.

This finishes our analysis of maximum likelihood estimation in the Barndorff-Nielsen completion. At least in the Cartesian product case, the maximum likelihood problem in the completion is of the same form as the original problem. The only difference is that we constrain certain components of the response vector to their observed values. This can be achieved by removing those components from the response vector and proceeding as if the resulting subvector were the entire response vector. If, for example, we are using the R function `glm` to fit models, we merely delete certain elements of the response vector and the corresponding rows of the model matrix (or the data frame containing the data if we are using a formula to specify the model) and proceed normally. This conditional model (with some components deleted) will always be non-identifiable, because δ will always be a direction of constancy and there may be other directions of constancy. The `glm` function, however, can deal with this issue. Furthermore, even in the rare case when the `glm` function may be confused, we can find a full rank model matrix having the same column space as the original model matrix using the function `redundant` in the `rcdd` package, as described in the preceding section.

3.14 Phase I and Phase II

Geyer (1990) coined the term "phase I maximum likelihood problem" to refer to the process of determining whether the likelihood function has any local or global maxima and if not what to do about it. This term was also used by Geyer and Thompson (1992). It was coined by analogy with the phase I linear programming problem, which is to find a feasible point, if any exists, or determine that none exist. If one exists, then this is used to start the phase II problem, finding optimal values.

Little, if anything is known about the phase I maximum likelihood problem except in one special case: exponential families, where it is completely understood theoretically. As we have seen, the phase I problem for full exponential families satisfying the condition of Brown (1986) consists entirely in determining the set L_{sat} and a GDOR δ, and this is done by repeated linear programming using one call to the function `linearity` in the `rcdd` package and one call to the function `lpcdd` in the same package. The phase I algorithm can be done using exact infinite-precision rational arithmetic, in which case the result is exact.

The rest of the maximum likelihood problem, can be called the phase II problem: finding the MLE. In Section 3.13.1 we saw that, when the phase I problem has established $L_{\text{sat}} = V_{\text{sat}}$, the MLE exists in the usual sense and is found using the usual algorithm. In Section 3.13.2 we saw that, when the phase I problem has established $L_{\text{sat}} \neq V_{\text{sat}}$, the MLE exists in the Barndorff-Nielsen completion and is found using the usual algorithm applied to modified data, the modification being determined by L_{sat}.

3.15 Likelihood Ratio Tests

Given two nested natural affine submodels, the maximum value of the log likelihood can be calculated for each submodel even without solving the phase I algorithm. Available software, such as the R function `glm`, will go uphill on the log likelihood until reaching a point where the log likelihood is nearly flat, in which case the value of the log likelihood is nearly the maximum. If the MLE does not exist in the conventional sense, then the natural parameter estimates will be large but not infinite, and the `glm` function may or may not give a warning about lack of convergence. If the MLE does not exist in the conventional sense, then the natural parameter estimates are infinitely wrong, but the value of the maximized log likelihood is nearly correct. Thus we can correctly calculate the likelihood ratio test statistic without solving the phase I problem.

This does no good, however, because the usual asymptotics of the likelihood ratio test (Wilks' theorem) do not hold in the case where the MLE for the null model does not exist in the conventional sense. In this case, the following simple correction, suggested by S.

Fienberg (personal communication) seems reasonable. Solve the phase I problem for the null model, determining $L_{\text{sat}} \neq V_{\text{sat}}$. Let M_0 and M_1 be the model matrices for the null and alternative natural affine submodels (M_0 was used in the phase I calculation).

If we apply Wilks' theorem to the limiting conditional model for the null hypothesis, we obtain the result that the deviance (twice the log likelihood ratio) is approximately chi-squared with degrees of freedom which is the difference in dimension of $M_1^T(\operatorname{span} L_{\text{sat}})$ and of $M_0^T(\operatorname{span} L_{\text{sat}})$. Assuming that M_0, M_1, and L_{sat} were determined exactly using rational arithmetic, the degrees of freedom can be determined exactly by applying the `redundant` function in the `rcdd` package to the sets $\{ M_i^T w : w \in L_{\text{sat}} \}$, $i = 0, 1$.

This asymptotic approximation may or may not hold depending on the sample size and on how close the observed value of the natural statistic is to the boundary of the convex support of the limiting conditional model. By construction, it cannot be on the boundary, but if it is close the asymptotic approximation can be bad. If one is worried about the validity of the asymptotic approximation, one can always do a parametric bootstrap calculation based on the limiting conditional model for the null hypothesis.

3.16 Confidence Intervals

Confidence intervals are more complicated than hypothesis tests. Confidence intervals for both natural and mean value parameters are of interest. The R function `predict.glm` provides either, depending on the value of its `type` argument. We will also provide either.

Before getting into details, we should first note that confidence intervals are often inappropriate from a theoretical point of view. When there is not a single scalar parameter of interest, a confidence region for the vector parameter of interest should, theoretically, be provided. However, high-dimensional confidence regions are unvisualizable, hence uninterpretable and of no interest to users. Thus statisticians usually provide something users think they can interpret, which is multiple confidence intervals, often not adjusted for simultaneous coverage. That is what, for example, `predict.glm` provides. We will follow the usual practice, providing multiple confidence intervals in our examples and restricting our treatment of confidence regions to a few comments.

In the case where the MLE does not exist in the conventional sense but is found in the Barndorff-Nielsen completion, we have two natural parameter spaces in play, the natural parameter space Θ of the original natural affine submodel and the natural parameter space Θ_{lim} of the full family containing the limiting conditional model, which by assumption is given by (12). The MLE is in the latter, but confidence intervals or confidence regions must be in the former. Thus we need to fully understand the relationship between the two. Both are subsets of $\mathbb{R}^p$ and considered as such $\Theta \subset \Theta_{\text{lim}}$, but we should not consider them as such because a point θ in both sets corresponds to different distributions in the two models. Let δ be a GDOR, and let Γ_{lim} denote the constancy space of the limiting conditional model. Then we know $\delta \in \Gamma_{\text{lim}} \subset \Theta_{\text{lim}}$. We also know that in the limiting conditional model the parameter is not identifiable: for every $\gamma \in \Gamma_{\text{lim}}$ the parameter values θ and $\theta + \gamma$ correspond to the same distribution. Thus distributions do not correspond to parameter points θ but to equivalence classes of points

$$\theta + \Gamma_{\text{lim}} = \{ \theta + \gamma : \gamma \in \Gamma_{\text{lim}} \}.$$

Of course, the same issue applies to the original model. If its constancy space is Γ, then equivalence classes $\theta + \Gamma$ correspond to distributions in the original model. Since Γ_{lim} contains δ, it is a nontrivial subspace. We cannot reparametrize to make the limiting conditional model identifiable without losing the connection between the two models.

Points θ and $\theta + \gamma$ with $\gamma \in \Gamma_{\text{lim}}$ correspond to the same distribution in the limiting conditional model but may correspond to different distributions in the original model (do correspond to different distributions unless $\gamma \in \Gamma$). The relationship between the two models is that the distributions in the original model corresponding to $\theta + s\delta$ and $\theta + \gamma + s\delta$ converge as $s \to \infty$ to the (single) distribution in the limiting conditional model corresponding to both parameter values θ and $\theta + \gamma$.

3.16.1 In the Limiting Conditional Model

In the limiting conditional model we have the "usual asymptotics" of maximum likelihood. The MLE $\hat{\theta}$ is asymptotically normal with variance inverse Fisher information, or would be except for two issues: the Fisher information matrix is singular with null space Γ_{lim} and we may not believe any distribution in the limiting conditional model is correct, so a confidence region in the limiting conditional model may be nonsense. Nevertheless, such confidence regions may be useful as a tool for constructing confidence regions in the original model that do make sense. Moreover, the non-identifiability issue in the limiting conditional model can be dealt with either by using a pseudo-inverse for Fisher information or by constraining $\hat{\theta}$ to lie in a subspace of $\mathbb{R}^p$ such that the limiting conditional model is identifiable when this constraint is imposed. We usually take the latter approach, since the R function `glm` does this automatically, dropping parameters to obtain identifiability.

Suppose we have such a confidence region R_{lim} for θ in the limiting conditional model. Because we used constraints, we do not automatically get $R_{\text{lim}} \supset \Gamma_{\text{lim}}$. Hence, if we wish to relate this confidence region to the original model, which we do, then we need to consider $R_{\text{lim}} + \Gamma_{\text{lim}} = \{\, \theta + \gamma : \theta \in R_{\text{lim}},\ \gamma \in \Gamma_{\text{lim}} \,\}$ the "actual" confidence region.

3.16.2 In the Original Model

Now for each $\theta \in R_{\text{lim}}$ and each $\gamma \in \Gamma_{\text{lim}}$, the distribution in the limiting conditional model corresponding to $\theta + \gamma$ is the limit as $s \to \infty$ of the distributions in the original model corresponding to $\theta + \gamma + s\delta$. Thus it remains to be decided how large s may be, that is, we need a confidence interval for s, which will necessarily be one-sided, of the form $(\hat{s}_{\theta+\gamma}, \infty)$. As our notation suggests, we make one such confidence interval for each point $\theta + \gamma$ we need to consider.

We base our interval on the statistic $\langle Y, \delta \rangle$, the observed value of which $\langle y, \delta \rangle$ is its maximum possible value. Since $\langle Y, \delta \rangle$ is discrete where it counts — the value $\langle y, \delta \rangle$ is assumed to have positive probability so that the limiting conditional model described in Section 3.4 exists — it is not possible to get an exact confidence interval unless one uses randomized or fuzzy intervals, described by Geyer and Meeden (2005). It follows immediately from the theory of uniformly most powerful tests (Lehmann, 1959) and the definitions in Geyer and Meeden (2005), that the exact $1 - \alpha$ fuzzy confidence interval in this case has membership function

$$I(s) = \max[0, 1 - \alpha / \Pr_{\theta+\gamma+s\delta}(Y \in H)], \tag{23}$$

which has a direct interpretation as a confidence statement: $I(s)$ gives the degree to which s should be considered to be in the confidence interval (Geyer and Meeden, 2005). Also (23) can be used to construct randomized confidence intervals; if U is a Uniform$(0, 1)$ random variate independent of the data, then

$$^{U}I(s) = \{\, s : I(s) \geq U \,\}$$

is an exact $1-\alpha$ randomized confidence interval Geyer and Meeden (2005, Section 2.1). The main point of this technical report is to advocate the use of generic directions of recession to calculate maximum likelihood estimates in the Barndorff-Nielsen completion and related hypothesis tests and confidence intervals. Advocacy of fuzzy confidence intervals is not our purpose (see Geyer and Meeden, 2005, for that). However, if one wants an exact procedure, one must use the fuzzy confidence interval, so we have mentioned it.

For those that want a conventional confidence interval, the *support* of the fuzzy interval

$$(\hat{s}_{\theta+\gamma}, \infty) = \operatorname{supp} I = \{\, s : I(s) > 0 \,\} \tag{24}$$

Geyer and Meeden (2005, Section 1.3) is a conservative $1 - \alpha$ confidence interval for s. Clearly,

$$\hat{s}_{\theta+\gamma} = \inf\{\, s \in \mathbb{R} : \theta + \gamma + s\delta \in \Theta \text{ and } \Pr_{\theta+\gamma+s\delta}(Y \in H) > \alpha \,\}.$$

Since $s \mapsto \Pr_{\theta+\gamma+s\delta}(Y \in H)$ is continuous and strictly increasing by Theorem 6, usually $\hat{s}_{\theta+\gamma}$ is the unique s such that $\Pr_{\theta+\gamma+s\delta}(Y \in H) = \alpha$. Only when the data have very little information about the parameter would it be the case that $\Pr_{\theta+\gamma+s\delta}(Y \in H) > \alpha$ for all s in the allowed range, in which case $\hat{s}_{\theta+\gamma}$ would be the lower endpoint of this range.

We should remark that there is a simple argument leading directly to these conventional confidence intervals (24) without going through fuzzy confidence intervals. The conventional conservative P-value for the upper-tailed test having test statistic $\langle Y, \delta \rangle$, null hypothesis $\theta + \gamma + s\delta$, and observed data y is

$$\Pr_{\theta+\gamma+s\delta}(\langle Y - y, \delta \rangle \geq 0). \tag{25}$$

A conventional level α test rejects when (25) is less than or equal to α. A conservative $1-\alpha$ confidence interval consists of the set of s values that are not rejected at level α. In the only case of interest to us, where $\langle y, \delta \rangle$ is the largest possible value of $\langle Y, \delta \rangle$ so the event $\langle Y - y, \delta \rangle \geq 0$ is the same as $Y \in H$ up to a set of measure zero, this comes to the same interval as (24).

3.16.3 A Combination of the Two

Because all of our one-sided confidence intervals for s, one for each parameter point $\theta+\gamma$, $\theta \in R_{\text{lim}}$, $\gamma \in \Gamma_{\text{lim}}$, use the same test statistic $\langle Y, \delta \rangle$, it follows that they have simultaneous coverage probability at least $1 - \alpha$. Thus under the assumption that R_{lim} was a $1 - \alpha^*$ confidence region in the conditional limiting model, we see that

$$\{\, \theta + \gamma + s\delta : \theta \in R_{\text{lim}},\ \gamma \in \Gamma_{\text{lim}},\ s \geq \hat{s}_{\theta+\gamma} \,\}$$

is a $1 - \alpha - \alpha^*$ confidence region for the natural parameter of the original model.

However, we are rarely interested in providing a confidence region. Thus we will instead consider

$$\{\, \hat{\theta} + \gamma + s\delta : \ \gamma \in \Gamma_{\text{lim}}, \ s \geq \hat{s}_{\hat{\theta}+\gamma} \,\}$$

to be a $1-\alpha$ confidence region that captures the part of the uncertainty relating to "how close to infinity" the natural parameter may be. This must be combined with R_{lim} or with non-simultaneous confidence intervals based on the limiting conditional model. Simultaneous coverage will not be achieved unless Bonferroni or other correction is applied.

3.16.4 The Cartesian Product Case

In the case where the convex support of the saturated model is a Cartesian product, all of this simplifies somewhat. We know that a confidence region for the mean value parameters of the limiting conditional model only involves the response variables that are not constrained to be at their observed values in the limiting conditional model. We take such a confidence region or separate confidence intervals, such as those provided by the R function `predict.glm` applied to the limiting conditional model, to be adequate for describing those components of the response.

Our one-sided intervals come into play in computing one-sided confidence intervals for the mean value parameters of the other components of the response. Since the MLE of their mean value parameters are on the boundary, one-sided intervals are the only kind that make sense. We distinguish two cases.

The first case, is where $\Gamma_{\text{lim}} = \{\, s\delta : s \in \mathbb{R} \,\}$. In this case, all intervals $(\hat{s}_{\theta+\gamma}, \infty)$ for the same θ but different $\gamma \in \Gamma_{\text{lim}}$ correspond to the same natural parameter values. Hence we can ignore γ. We can hope that one interval $(\hat{s}_{\hat{\theta}}, \infty)$ adequately describes the variability in mean value parameters for components of the response having MLE at the boundary.

The second case, is where Γ_{lim} contains vectors not proportional to δ. Then we use our general formula, but can still hope that the intervals $(\hat{s}_{\hat{\theta}+\gamma}, \infty)$, $\gamma \in \Gamma_{\text{lim}}$ adequately describe the variability in mean value parameters for components of the response having MLE at the boundary.

In both cases we only use the intervals corresponding to $\hat{\theta}$ rather than all $\theta \in R_{\text{lim}}$. This is clearly not exactly correct, however, we do have the following argument. The idea is that we are separating variation into two components, one along the direction of recession and the other across the direction of recession. To a first approximation, our one-sided intervals are about variation along the direction of recession and conventional intervals for the limiting conditional model are about variation across the direction of recession. In the spirit of sloppiness that allows us to provide separate confidence intervals rather than confidence regions, we consider this division not too bad. From the discussion about the distributions for $\theta + s\delta$ and $\theta + \gamma + s\delta$ converging to the same limiting conditional distribution, we see that this division cannot be exactly correct, but it may do for practical purposes.

In any event, statisticians have made do up to now with no tools whatsoever for handling this issue. Approximately correct tools will be a great improvement.

3.16.5 Calculating the Constancy Space

By Theorems 1 and 4 the constancy space is $N_C(y)$ in the case where every direction of recession is a direction of constancy. By Corollary 9 the tangent space $T_{C \cap H}(y)$ in the limiting conditional model is span L. Hence the constancy space is

$$N_{C \cap H}(y) = \{\, \delta \in \mathbb{R}^p : \langle v, \delta \rangle = 0, \ v \in L \,\}.$$

In the case of a natural affine submodel this becomes

$$\Gamma_{\text{lim}} = N_{C_{\text{sub}} \cap H_{\text{sub}}}(M^T y) = \{\, \delta \in \mathbb{R}^q : \langle v, \delta \rangle = 0, \ v \in L_{\text{sub}} \,\}. \tag{26}$$

where q is the dimension of the submodel.

We have already seen in Section 3.11 how to calculate L_{sub}. Now we merely note that L_{sub} is a V-representation of its span, and a call to the function `scdd` in the `rcdd` package will compute an H-representation of its span which is also a V-representation for (26), that is, a basis for the constancy space of the limiting conditional model. The examples that follow do not exhibit a need for this operation, but the need would arise in other examples.

When calculating (26) for the purpose of generating one-sided confidence intervals it is clear that we do not need to let γ range over the whole constancy space (26) because points γ and $\gamma + s\delta$ lead to the same one-sided confidence intervals. Hence it is enough to use the subspace of Γ_{lim} orthogonal to δ, which is calculated by feeding $L_{\text{sub}} \cup \{\delta\}$ to the R function `scdd` for conversion to an H-representation, which will also be a basis of the desired subspace.

4 Examples

4.1 A Logistic Regression Example

We start with a logistic regression. Suppose we observe a vector y whose components are Bernoulli with means forming a vector p. The natural parameter is $\theta = \text{logit}(p)$, where logit operates componentwise $\theta_i = \text{logit}(p_i)$. Suppose we also have one covariate vector x and we want to fit a quadratic model

$$\theta_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2.$$

Finally, suppose x_i takes the values 1, ..., 30 and $y_i = 0$ for $x_i \leq 12$ or $x_i \geq 24$ and $y_i = 1$ otherwise.

The following R statements create the data and attempt to fit the model.

```
> x <- 1:30
> y <- c(rep(0, 12), rep(1, 11), rep(0, 7))
> out1 <- suppressWarnings(glm(y ~ x + I(x^2), family = binomial,
+     x = TRUE))
```

It seems difficult if not impossible to capture warning messages to an Sweave file, so we have suppressed them. The warnings given by R version 2.7.0 are

```
algorithm did not converge
fitted probabilities numerically 0 or 1 occurred
```

(two separate warnings). This is the `glm` function's way of indicating that the MLE may not exist. However, because we gave the argument `x = TRUE` we have obtained the model matrix

```
> M <- out1$x
```

Now we are ready to carry out the computation of Section 3.11, determining L_{sub}.

```
> tanv <- M
> tanv[y == 1, ] <- (-tanv[y == 1, ])
> vrep <- cbind(0, 0, tanv)
> lout <- linearity(vrep, rep = "V")
> lout

integer(0)
```

The rows of the matrix `tanv` are the elements of V_{sub}. The matrix `vrep` is the form in which the `rcdd` package encodes the V-representation of $\text{con}(\text{pos}\, V_{\text{sub}}) = T_{C_{\text{sub}}}(M^T y)$. The vector `lout` gives the indices of the elements of V_{sub} that are in L_{sub}. It having length zero indicates that $L_{\text{sub}} = \varnothing$, and this in turn indicates by Corollary 9 that the support of the limiting conditional model is $C_{\text{sat}} \cap H_{\text{sat}} = \{y\}$. Hence there is only one distribution in the limiting conditional model, which is the distribution concentrated at y, and this trivially must be the MLE in the limiting conditional model. This also implies that every direction is a direction of constancy in the limiting conditional model, that is $\Gamma_{\text{lim}} = \mathbb{R}^3$.

Now we are ready to carry out the computation of Section 3.12, calculating δ.

```
> hrep <- cbind(-vrep, -1)
> hrep <- rbind(hrep, c(0, 1, rep(0, 3), -1))
> objv <- c(rep(0, 3), 1)
> pout <- lpcdd(hrep, objv, minimize = FALSE)
> names(pout)

[1] "solution.type"   "primal.solution" "dual.solution"
[4] "optimal.value"

> pout$solution.type

[1] "Optimal"
```

The list `pout` is the solution of the linear programming problem described in Section 3.12. We obtain the GDOR and check its validity

```
> gdor <- pout$primal.solution[-length(pout$primal.solution)]
> all(tanv %*% gdor < 0)

[1] TRUE

> gdor

[1] -53.3636364   6.5454545  -0.1818182
```

All of the computation to this point has used ordinary inexact computer arithmetic. We also illustrate exact infinite-precision rational arithmetic computations

```
> pout.exact <- lpcdd(d2q(hrep), d2q(objv), minimize = FALSE)
> gdor.exact <- pout.exact$primal.solution[-length(pout$primal.solution)]
> gdor.exact

[1] "-587/11" "72/11"   "-2/11"

> q2d(gdor.exact)

[1] -53.3636364   6.5454545  -0.1818182
```

We are now ready to find one-sided confidence intervals. We use the R function `uniroot` to solve equations involved in this process. We start by defining a function that evaluates $\Pr_{\beta+s\delta}(Y \in H)$ for any β and δ.

```
> invlogit <- function(x) 1/(1 + exp(-x))
> prob.face <- function(beta, s) {
+     moo <- M %*% (beta + s * gdor)
+     moo <- as.vector(moo)
+     prod(ifelse(y == 1, invlogit(moo), invlogit(-moo)))
+ }
```

Now we try it out on the point $\beta = 0$.

```
> alpha <- 0.05
> beta.hat <- rep(0, length(gdor))
> foo <- function(s) prob.face(beta.hat, s) - alpha
> fred <- uniroot(foo, lower = -10, upper = 10)
> lowbnd <- fred$root
> c(lowbnd, Inf)

[1] 0.5715858       Inf

> fred$estim.prec

[1] 6.396655e-05
```

The meaning of this confidence interval is not obvious just from looking at the numbers. Hence we make a plot showing the mean values that correspond to $s\delta$ for s in the interval.

```
> eta.gdor <- as.vector(M %*% gdor)
> par(mar = c(5, 4, 1, 1) + 0.1)
> plot(x, invlogit(lowbnd * eta.gdor), ylim = c(0,
+     1), ylab = "mean of y")
> points(x, y)
> segments(x, y, x, invlogit(lowbnd * eta.gdor))
```

The confidence intervals shown in Figure 1 are certainly better than what was previously available, which was nothing. Figure 1 clearly shows that some mean values are much better estimated than others, and gives a rough idea of the variability. However, Figure 1 is not

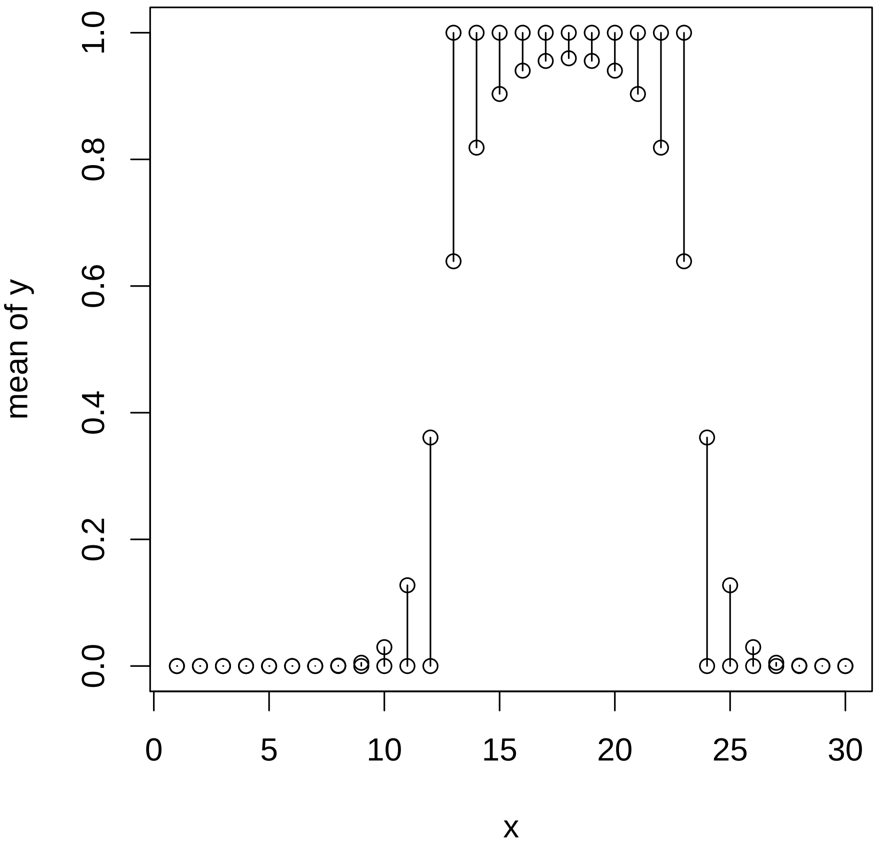

Figure 1: Quick and Dirty 95% Confidence Intervals for Regression Function. Compare with Figure 2.

the Right Thing, which was explained in Section 3.16. We now proceed to an approximation of that.

We should for each $\gamma \in \mathbb{R}^3$ find the $\hat{s}_\gamma$ such that $\Pr_{\gamma + \hat{s}_\gamma \delta}(Y \in H) = \alpha$. Then we should form the confidence region $\{\, \gamma + s\delta : \gamma \in \mathbb{R}^3,\ s \geq \hat{s}_\gamma \,\}$. Finally, we should find the set of mean value parameters corresponding to this confidence region. Clearly, we cannot do that, as it would entail an infinite amount of work. We can hope that we can make do with just a few γ values. Let us try that.

```
> set.seed(42)
> nboot <- 100
> scale <- 0.25
> eta.up <- rep(-Inf, length(y))
> eta.dn <- rep(Inf, length(y))
> for (iboot in 1:nboot) {
+     beta.hat <- rnorm(length(gdor)) * scale
+     foo <- function(s) prob.face(beta.hat, s) - alpha
+     fred <- uniroot(foo, lower = -20, upper = 500)
+     lowbnd <- fred$root
+     eta.hat <- as.vector(M %*% (beta.hat + lowbnd *
+         gdor))
+     eta.up <- pmax(eta.up, eta.hat)
+     eta.dn <- pmin(eta.dn, eta.hat)
+ }
> mu.up <- ifelse(y == 1, 1, invlogit(eta.up))
> mu.dn <- ifelse(y == 0, 0, invlogit(eta.dn))
```

Then we make a plot showing these confidence intervals.

```
> plot(x, mu.up, ylim = c(0, 1), ylab = "mean of y")
> points(x, mu.dn)
> segments(x, mu.up, x, mu.dn)
```

We claim (hope) that these confidence intervals do give good simultaneous coverage. Our methods are not ideal in that we probably should not use the `uniroot` function, which requires as input an interval bounding the solution and does not use the fact that the function is strictly increasing. Some method designed explicitly for strictly increasing functions would be better. If such a method needs first or second derivatives, these can be calculated explicitly. We do not give the formulas because we do not propose a particular method.

Our choice of random starting points is also perhaps not ideal for two reasons. The constant 0.25 in the algorithm is obviously arbitrary. Increasing it requires increasing the width of the interval given to the `uniroot` function. A better function for solving equations would give more flexibility in choosing this constant. The other reason our choice is less than ideal is that we could use the computation discussed in Section 3.16.5 to let our points `beta.hat` vary over the two-dimensional subspace perpendicular to δ, but we have not bothered to do this, since it is a mere optimization that may be more efficient but does not change the result calculated.

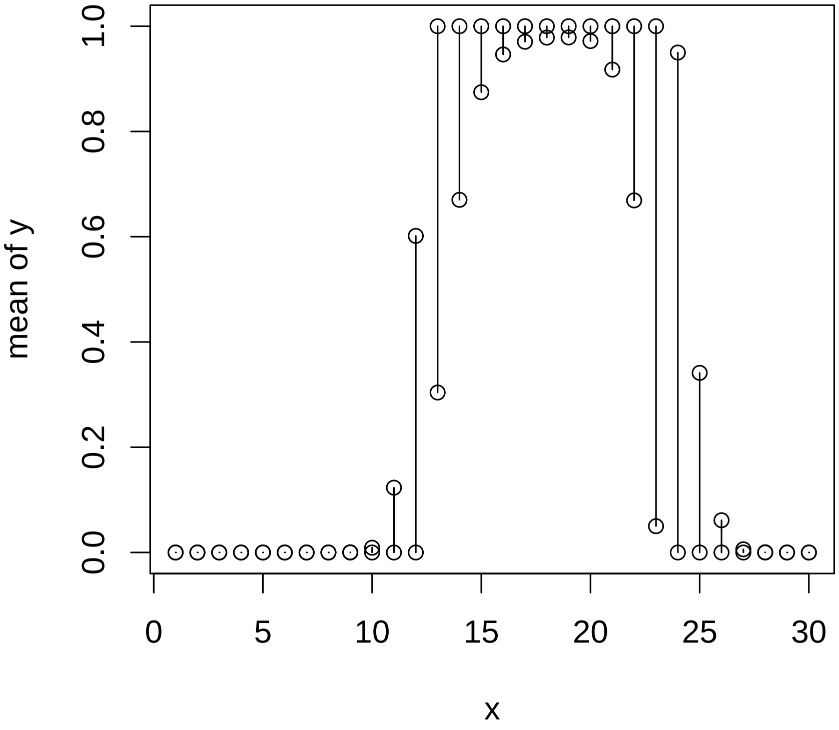

Figure 2: Simultaneous 95% Confidence Intervals for Regression Function.

4.2 A Contingency Table Example

This example is a $2 \times 2 \times \cdots \times 2$ contingency table with seven dimensions hence $2^7 = 128$ cells. The data are

```
> dat <- read.table(url("http://www.stat.umn.edu/geyer/gdor/catrec.txt"),
+     header = TRUE)
> dim(dat)

[1] 128   8

> names(dat)

[1] "v1" "v2" "v3" "v4" "v5" "v6" "v7" "y"
```

which presents the data as eight vectors, seven categorical predictors v_1, ..., v_7 that specify the cells of the contingency table and one response y that gives the cell counts.

4.2.1 Poisson Sampling

We start by fitting two models assuming Poisson sampling, one with all two-way interactions and no higher interactions and one with all three-way interactions and no higher interactions, and attempt to compare them using a test of model comparison.

```
> out2 <- glm(y ~ (v1 + v2 + v3 + v4 + v5 + v6 + v7)^2,
+     family = poisson, data = dat, x = TRUE)
> out3 <- glm(y ~ (v1 + v2 + v3 + v4 + v5 + v6 + v7)^3,
+     family = poisson, data = dat, x = TRUE)
> anova(out2, out3, test = "Chisq")

Analysis of Deviance Table

Model 1: y ~ (v1 + v2 + v3 + v4 + v5 + v6 + v7)^2
Model 2: y ~ (v1 + v2 + v3 + v4 + v5 + v6 + v7)^3
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        99    191.629
2        64     31.291 35  160.338 5.819e-18
```

Unlike the logistic regression example, the `glm` function gives no warning here about lack of convergence or nonexistence of the MLE. However, as we shall see, the MLE does not exist in the conventional sense in the larger model `out3`.

First we determine the linearity for model `out2`.

```
> tanv <- out2$x
> vrep <- cbind(0, 0, tanv)
> vrep[dat$y > 0, 1] <- 1
> lout <- linearity(d2q(vrep), rep = "V")
> linear <- dat$y > 0
> linear[lout] <- TRUE
> all(linear)
```

```
[1] TRUE
```

Thus the MLE does exist in the conventional sense for this model, and the `glm` function presumably finds it.

Second we determine the linearity for model `out3`.

```
> tanv <- out3$x
> vrep <- cbind(0, 0, tanv)
> vrep[dat$y > 0, 1] <- 1
> lout <- linearity(d2q(vrep), rep = "V")
> linear <- dat$y > 0
> linear[lout] <- TRUE
> all(linear)
```

```
[1] FALSE
```

Thus the MLE does not exist in the conventional sense for this model, and the `glm` function has produced nonsense with no error or warning.

Since the MLE does exist for the null hypothesis, the test done by the `anova` function above is correct (Section 3.15). Thus we have a problem for which available software provides no solution. The model `out2` that we can fit with available software clearly does not fit the data, but the model `out3` that appears to fit the data we cannot fit with available software. Hence the proposals of this technical report!

Next we determine a GDOR.

```
> hrep <- cbind(0, 0, -tanv, 0)
> hrep[!linear, ncol(hrep)] <- (-1)
> hrep[linear, 1] <- 1
> hrep <- rbind(hrep, c(0, 1, rep(0, ncol(out3$x)),
+     -1))
> objv <- c(rep(0, ncol(out3$x)), 1)
> pout <- lpcdd(d2q(hrep), d2q(objv), minimize = FALSE)
> gdor <- pout$primal.solution[-length(pout$primal.solution)]
```

and do some checks to try to understand what we have found

```
> foo <- gdor
> names(foo) <- names(out3$coef)
> print(cbind(foo[foo != "0"]))
```

```
            [,1]
(Intercept) "-1"
v1          "1"
v2          "1"
v3          "1"
v5          "1"
v1:v2       "-1"
v1:v3       "-1"
```

```
v1:v5       "-1"
v2:v3       "-1"
v2:v5       "-1"
v3:v5       "-1"
v1:v2:v3    "1"
v1:v3:v5    "1"
v2:v3:v5    "1"

> eta.gdor <- as.vector(qmatmult(tanv, cbind(gdor)))
> all(qsign(eta.gdor) <= 0)

[1] TRUE

> all(qsign(eta.gdor[linear]) == 0)

[1] TRUE

> all(qsign(eta.gdor[!linear]) < 0)

[1] TRUE
```

First we print out the nonzero components of the GDOR, not that this tells us much. Then we check that the GDOR actually satisfies the conditions (21a) and (21b).

Then we figure out the convex support of the limiting conditional family. Which of the cells that have observed count zero are fixed at zero in the limiting conditional family?

```
> sum(!linear)

[1] 16

> sum(dat$y == 0)

[1] 17

> all(dat$y == 0 | linear)

[1] TRUE
```

we see that of the 17 cells that have zero count in the observed data 16 are conditioned to be zero in the limiting conditional model.

Our next task is to fit the limiting conditional model.

```
> dat.cond <- dat[linear, ]
> out3.cond <- glm(y ~ (v1 + v2 + v3 + v4 + v5 + v6 +
+     v7)^3, family = poisson, data = dat.cond)
> summary(out3.cond)
```

```
Call:
glm(formula = y ~ (v1 + v2 + v3 + v4 + v5 + v6 + v7)^3, family = poisson,
    data = dat.cond)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.63571  -0.30009  -0.02353   0.27258   1.42540

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.150481   0.585423   3.673 0.000239 ***
v1           0.069795   0.587067   0.119 0.905364
v2          -0.524215   0.513583  -1.021 0.307396
v3           0.052966   0.551965   0.096 0.923552
v4          -0.709525   0.580147  -1.223 0.221326
v5           0.243002   0.548686   0.443 0.657853
v6          -1.163256   0.563668  -2.064 0.039044 *
v7          -0.990704   0.597335  -1.659 0.097208 .
v1:v2        0.384345   0.543024   0.708 0.479079
v1:v3       -0.630375   0.570151  -1.106 0.268888
v1:v4        0.008801   0.511458   0.017 0.986271
v1:v5       -1.022805   0.570440  -1.793 0.072971 .
v1:v6        0.540164   0.493879   1.094 0.274079
v1:v7        0.097178   0.536628   0.181 0.856297
v2:v3        0.602411   0.437371   1.377 0.168405
v2:v4        0.748226   0.486811   1.537 0.124295
v2:v5       -0.068926   0.428100  -0.161 0.872090
v2:v6        0.297165   0.487409   0.610 0.542071
v2:v7        0.274198   0.508369   0.539 0.589634
v3:v4       -0.124465   0.541056  -0.230 0.818060
v3:v5       -0.439354   0.468418  -0.938 0.348268
v3:v6        0.024399   0.530220   0.046 0.963296
v3:v7       -0.104400   0.556960  -0.187 0.851310
v4:v5       -0.169421   0.521323  -0.325 0.745194
v4:v6        0.756513   0.474213   1.595 0.110644
v4:v7        0.780671   0.500911   1.559 0.119114
v5:v6        1.245629   0.510770   2.439 0.014739 *
v5:v7       -0.262620   0.523125  -0.502 0.615652
v6:v7        0.697014   0.489957   1.423 0.154852
v1:v2:v3    -0.349902   0.483330  -0.724 0.469102
v1:v2:v4     0.101569   0.389778   0.261 0.794416
v1:v2:v5     0.655208   0.493737   1.327 0.184496
v1:v2:v6    -0.329286   0.390979  -0.842 0.399670
v1:v2:v7    -0.520368   0.393042  -1.324 0.185520
v1:v3:v4     0.353292   0.406623   0.869 0.384932
```

```
v1:v3:v5     0.638711   0.484979   1.317 0.187843
v1:v3:v6     0.352694   0.402715   0.876 0.381143
v1:v3:v7    -0.001586   0.413554  -0.004 0.996941
v1:v4:v5     0.664745   0.400212   1.661 0.096717 .
v1:v4:v6    -0.463885   0.368214  -1.260 0.207732
v1:v4:v7    -0.342583   0.372009  -0.921 0.357103
v1:v5:v6     0.044968   0.399958   0.112 0.910481
v1:v5:v7     0.447641   0.404364   1.107 0.268283
v1:v6:v7     0.218868   0.371499   0.589 0.555763
v2:v3:v4    -0.325914   0.404392  -0.806 0.420280
v2:v3:v5           NA         NA      NA       NA
v2:v3:v6    -0.247853   0.405621  -0.611 0.541168
v2:v3:v7     0.028322   0.414520   0.068 0.945527
v2:v4:v5     0.004655   0.394418   0.012 0.990583
v2:v4:v6    -0.111152   0.373713  -0.297 0.766141
v2:v4:v7    -0.148061   0.376692  -0.393 0.694279
v2:v5:v6    -0.766051   0.394925  -1.940 0.052412 .
v2:v5:v7     0.075213   0.399004   0.189 0.850482
v2:v6:v7     0.460826   0.381109   1.209 0.226597
v3:v4:v5    -0.063494   0.423318  -0.150 0.880771
v3:v4:v6     0.357746   0.366298   0.977 0.328741
v3:v4:v7    -0.106368   0.371567  -0.286 0.774672
v3:v5:v6    -0.234816   0.422424  -0.556 0.578295
v3:v5:v7     0.804923   0.423843   1.899 0.057550 .
v3:v6:v7    -0.659090   0.371085  -1.776 0.075714 .
v4:v5:v6    -0.427957   0.375755  -1.139 0.254734
v4:v5:v7     0.125167   0.377356   0.332 0.740119
v4:v6:v7     0.014192   0.370131   0.038 0.969413
v5:v6:v7    -0.811516   0.377098  -2.152 0.031397 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 156.215  on 111  degrees of freedom
Residual deviance:  31.291  on  49  degrees of freedom
AIC: 526.46

Number of Fisher Scoring iterations: 5

> beta.hat <- coefficients(out3.cond)
> beta.hat[is.na(beta.hat)] <- 0
```

We see that, as expected, the model is not identifiable. However, the glm function decides which predictor to drop or, equivalently, which parameter to constrain to zero, which it reports as NA.

Since exactly one parameter need be set to zero, we see that the dimension of the constancy space of the limiting conditional model is one. Thus δ is a basis for the constancy space Γ_{lim}, and we see that a single one-sided confidence interval will do the job.

Since confidence intervals for mean values for components of the response vector that are not constrained to be zero in the limiting conditional model are entirely conventional and calculated by `predict.glm` applied to the object `out3.cond`, we will omit this step, assuming readers will know how to do it or at least figure out how to do it from reading the help for the function `predict.glm`.

On to one-sided intervals.

```
> eta.hat <- as.numeric(out3$x %*% beta.hat)
> eta.gdor <- q2d(eta.gdor)
> prob.face <- function(s) {
+     moo <- eta.hat + s * eta.gdor
+     exp(-sum(exp(moo[!linear])))
+ }
> foo <- function(s) prob.face(s) - alpha
> fred <- uniroot(foo, lower = -5, upper = 5)
> lowbnd <- fred$root
> c(lowbnd, Inf)
```

```
[1] 3.40117     Inf
```

```
> fred$estim.prec
```

```
[1] 6.103516e-05
```

This is our one-sided confidence interval for s, which we map to one-sided confidence intervals for the mean values of cells that are constrained to be zero in the limiting conditional model.

```
> moo <- exp(eta.hat + lowbnd * eta.gdor)
> foo <- cbind(dat, moo)
```

Table 1 shows these intervals.

```
> rownames(foo) <- NULL
> colnames(foo)[colnames(foo) == "y"] <- "lower"
> colnames(foo)[colnames(foo) == "moo"] <- "upper"
> library(xtable)
> print(xtable(foo[!linear, ], digits = c(rep(0, 9),
+     4), align = "ccccccccc", caption = paste("One-sided 95\\% Confidence Intervals for M
+     "Columns v1 to v7 give the cell.", "Lower and Upper give end points of the interval."
+     "Based on Poisson sampling, compare Table~\\ref{tab:two}."),
+     label = "tab:one"), caption.placement = "top",
+     table.placement = "tbp", include.rownames = FALSE)
```

Table 1: One-sided 95% Confidence Intervals for Mean Values. Columns v1 to v7 give the cell. Lower and Upper give end points of the interval. Based on Poisson sampling, compare Table 2.

v1	v2	v3	v4	v5	v6	v7	lower	upper
0	0	0	0	0	0	0	0	0.2863
0	0	0	1	0	0	0	0	0.1408
1	1	0	0	1	0	0	0	0.2200
1	1	0	1	1	0	0	0	0.4210
0	0	0	0	0	1	0	0	0.0895
0	0	0	1	0	1	0	0	0.0938
1	1	0	0	1	1	0	0	0.1930
1	1	0	1	1	1	0	0	0.2887
0	0	0	0	0	0	1	0	0.1063
0	0	0	1	0	0	1	0	0.1141
1	1	0	0	1	0	1	0	0.0913
1	1	0	1	1	0	1	0	0.2646
0	0	0	0	0	1	1	0	0.0667
0	0	0	1	0	1	1	0	0.1548
1	1	0	0	1	1	1	0	0.1410
1	1	0	1	1	1	1	0	0.3239

We would now like to illustrate a hypothesis test in which the MLE in the null hypothesis does not exist in the conventional sense. We will use the model with all three-way interactions and no higher interactions for the null hypothesis and the model with all four-way interactions and no higher interactions for the alternative hypothesis.

```
> out4.cond <- glm(y ~ (v1 + v2 + v3 + v4 + v5 + v6 +
+     v7)^4, family = poisson, data = dat.cond)
> anova(out3.cond, out4.cond, test = "Chisq")

Analysis of Deviance Table

Model 1: y ~ (v1 + v2 + v3 + v4 + v5 + v6 + v7)^3
Model 2: y ~ (v1 + v2 + v3 + v4 + v5 + v6 + v7)^4
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        49    31.2913
2        18    16.0669 31  15.2244    0.9921
```

Simple. The functions `glm` and `anova.glm` do the right thing if they are provided the data `dat.cond` for the limiting conditional model for the null hypothesis.

4.2.2 Multinomial Sampling

Consider one contingency table and one vector of observed data, but two models: Poisson sampling and multinomial sampling. As is well known, the maximum likelihood estimates for the mean value parameters are the same for both sampling schemes. But much more is the same.

Suppose we consider the natural statistic to be the vector of cell counts for both models, so both have the same natural statistic and natural parameter. For Poisson sampling, there are no directions of constancy and the MLE for the natural parameter is unique. For multinomial sampling, the vector $\gamma = (1, 1, \dots, 1)$ is a direction of constancy, and the MLE for the natural parameter is nonunique. However, it is easy to see that the unique MLE for the Poisson model is also an MLE for the multinomial model (when we use the parameterizations just defined).

Moreover, when we use the same natural statistic for both models, the computational geometry is similar. If subscripts P and M refer to the Poisson and multinomial models, respectively, then

$$C_{\text{sat},M} = \{\, v \in C_{\text{sat},P} : \langle v, \gamma \rangle = n \,\}$$

where n is the sample size. Also note that γ is the first column of M, the "intercept" column. From this it follows that

$$T_{C_{\text{sub},M}}(M^T y) = \{\, v \in T_{C_{\text{sub},P}}(M^T y) : \langle v, e_1 \rangle = 0 \,\}$$

where $e_1 = (1, 0, \dots, 0)$. And from this it follows by Theorem 6.42 in Rockafellar and Wets (2004) that

$$N_{C_{\text{sub},M}}(M^T y) \supset \{\, v + s e_1 : v \in N_{C_{\text{sub},P}}(M^T y),\ s \in \mathbb{R} \,\}.$$

Hence every direction of recession in the Poisson model is also one in the multinomial model. Similarly, every GDOR in the Poisson model is also one in the multinomial model. Hence the GDOR we have already calculated is correct for the multinomial model. Hence the support of the limiting conditional model is also correct. Hence also the parameter estimates for the limiting conditional model already obtained are correct. Hence (asymptotically) so is the hypothesis test comparing the 4-way interactions model to the 3-way interactions model.

The only thing we need to change for multinomial sampling is our one-sided confidence intervals, because they are based on exact probabilities that differ between the two models. We proceed to redo that

```
> n <- sum(dat$y)
> prob.face <- function(s) {
+     moo <- eta.hat + s * eta.gdor
+     mmoo <- max(moo)
+     bark <- exp(moo - mmoo)
+     qqq <- sum(bark[linear])/sum(bark)
+     qqq^n
+ }
> foo <- function(s) prob.face(s) - alpha
> fred <- uniroot(foo, lower = -5, upper = 5)
> lowbnd.multi <- fred$root
> c(lowbnd.multi, Inf)
```

```
[1] 3.398416      Inf
```

```
> fred$estim.prec

[1] 6.103516e-05
```

This is our one-sided confidence interval for s. Our interval for multinomial sampling with lower bound 3.3984 is not that different from the interval for Poisson sampling with lower bound 3.4012, but it is different.

Now we produce the analog of Table 1 for multinomial sampling.

```
> moo <- eta.hat + lowbnd.multi * eta.gdor
> mmoo <- max(moo)
> bark <- exp(moo - mmoo)
> woof <- n * bark/sum(bark)
> foo <- cbind(dat, woof)
```

Table 2 shows these intervals.

```
> rownames(foo) <- NULL
> colnames(foo)[colnames(foo) == "y"] <- "lower"
> colnames(foo)[colnames(foo) == "woof"] <- "upper"
> library(xtable)
> print(xtable(foo[!linear, ], digits = c(rep(0, 9),
+     4), align = "cccccccccc", caption = paste("One-sided 95\\% Confidence Intervals for M
+     "Columns v1 to v7 give the cell.", "Lower and Upper give end points of the interval."
+     "Based on multinomial sampling, compare Table~\\ref{tab:one}."),
+     label = "tab:two"), caption.placement = "top",
+     table.placement = "tbp", include.rownames = FALSE)
```

Again the results are different — Table 1 is different from Table 2 — though not much different.

We note that a section for product-multinomial sampling would look much like this section, so similar that it is left as an exercise for the reader.

5 Alternative Calculational Ideas

Pre-existing theory of Barndorff-Nielsen completion (Barndorff-Nielsen, 1978; Brown, 1986; Geyer, 1990) is based on the family of faces of the convex support C. The R package `rcdd` can be used to calculate all the faces of C, but only in the most toyish of toy problems. In order to calculate C_{sub} we need a V-representation for C_{sat}. For our first example (Section 4.1) $C_{\text{sat}} = [0, 1]^{30}$ has $2^{30} = 1073741824$ generators, which is far too many to deal with in an actual calculation. So the project of calculating all the faces of the convex support is a non-starter in this example. For our second example (Section 4.2) $C_{\text{sat}} = [0, \infty)^{128}$ has 128 generators, so we could attempt to calculate all the faces. Since the function `allfaces` requires H-representation input, we must first use the function `scdd` to convert from the V-representation we have for C_{sub} given by (18) to an H-representation. Unfortunately, this takes a very long time — the process had taken many hours when it was killed for lack of patience, whereas all of the calculations done in this technical report take only a minute

Table 2: One-sided 95% Confidence Intervals for Mean Values. Columns v1 to v7 give the cell. Lower and Upper give end points of the interval. Based on multinomial sampling, compare Table 1.

v1	v2	v3	v4	v5	v6	v7	lower	upper
0	0	0	0	0	0	0	0	0.2855
0	0	0	1	0	0	0	0	0.1404
1	1	0	0	1	0	0	0	0.2194
1	1	0	1	1	0	0	0	0.4198
0	0	0	0	0	1	0	0	0.0892
0	0	0	1	0	1	0	0	0.0935
1	1	0	0	1	1	0	0	0.1925
1	1	0	1	1	1	0	0	0.2879
0	0	0	0	0	0	1	0	0.1060
0	0	0	1	0	0	1	0	0.1138
1	1	0	0	1	0	1	0	0.0910
1	1	0	1	1	0	1	0	0.2639
0	0	0	0	0	1	1	0	0.0665
0	0	0	1	0	1	1	0	0.1543
1	1	0	0	1	1	1	0	0.1406
1	1	0	1	1	1	1	0	0.3230

— so the project of calculating all the faces is a non-finisher for our second example. We conclude that the idea of calculating all the faces of the convex support is not practical, since our examples are not particularly complicated. One would expect most practical applications to be much more complicated.

There is more hope for calculating the entire tangent and normal cones. For our first example (Section 4.1), we recreate the V-representation for the tangent cone of the affine submodel and convert to an H-representation, because an H-representation for the tangent cone is essentially a V-representation for the normal cone and vice versa.

```
> tanv <- M
> tanv[y == 1, ] <- (-tanv[y == 1, ])
> vrep <- cbind(0, 0, tanv)
> sout <- scdd(d2q(vrep), rep = "V")
> unclass(sout$output)[, -c(1, 2)]
```

```
     [,1]  [,2]  [,3]
[1,] "276" "-35" "1"
[2,] "299" "-36" "1"
[3,] "312" "-37" "1"
[4,] "288" "-36" "1"
```

The four vectors shown generate the normal cone. Any vector that is linear combination of these four vectors with strictly positive coefficients is a GDOR. So this calculation, instead of finding one GDOR, finds every GDOR. But as we have seen, we only need one GDOR to carry out all the statistical inferential procedures one might want to do. Moreover, when we try to apply this scheme to our second example (Section 4.2), the `scdd` operation takes a very long time — again taking many hours before the process was killed for lack of patience. Hence the idea of calculating the whole normal cone is a non-finisher except in some toy problems.

Hence we see that the idea of seeking a single GDOR, is strongly motivated by efficiency concerns. The methods of this technical report, using repeated linear programming, are not the only such methods. Geyer (1990) provided a competing scheme, which, although not aimed at calculating a GDOR (in fact, the GDOR concept does not appear in that thesis) does calculate the linearity space L_{sub}, and hence does allow computation of the MLE in the Barndorff-Nielsen completion. Neither confidence intervals nor hypothesis tests are discussed in Geyer (1990), so the GDOR notion was not needed.

We mention Geyer (1990) merely to show that the scheme introduced here based on one invocation of the function `linearity` and one invocation of the function `lpcdd`, both in the package `rcdd`, is not the only efficient method of calculating the MLE in the Barndorff-Nielsen completion and associated hypothesis tests and confidence intervals. We do conjecture that any efficient method, if general, must be based on repeated invocations of linear programming. Since the `linearity` function is particularly designed for the job it does, calculating L_{sub}, and since this is an essential step in computing the support of the limiting conditional model, it would seem that any method more efficient than what is proposed here must essentially improve on the scheme used in the `linearity` function, which was explained in Section 3.11. Whether or not this algorithm can be improved upon, it does seem that the `linearity` function implements a fairly efficient algorithm.

6 Proofs

Proof of Theorem 1. Clearly, $s \mapsto l(\theta + s\delta)$ fails to be strictly concave if and only if $s \mapsto c(\theta + s\delta)$ fails to be strictly convex, and by Theorem 2.1 in Geyer (1990) this happens if and only if $\langle Y, \delta \rangle$ is concentrated at one point, in which case this point must be $\langle y, \delta \rangle$ so (e) holds. Since all distributions in the family are mutually absolutely continuous by (4), (e) implies (f), which trivially implies (e). If (f) holds, then by (5)

$$\begin{aligned} c(\theta + s\delta) &= c(\psi) + \log E_\psi\bigl(e^{\langle Y, \theta + s\delta - \psi \rangle}\bigr) \\ &= c(\psi) + s\langle y, \delta \rangle + \log E_\psi\bigl(e^{\langle Y, \theta - \psi \rangle}\bigr) \\ &= c(\theta) + s\langle y, \delta \rangle \end{aligned} \tag{27}$$

Hence (b) holds, and (b) clearly implies (a). We have now proved that (a), (b), (e), and (f) are equivalent.

Also (27) implies (d) by (4), so (f) implies (d). Trivially, (d) implies (c). Conversely, if (c) holds, then f_θ and $f_{\theta + s\delta}$ must be equal almost surely, hence by (4)

$$\log f_{\theta + s\delta}(\omega) - \log f_\theta(\omega) = s\langle Y(\omega), \delta \rangle - c(\theta + s\delta) + c(\theta)$$

almost surely, hence $\langle Y, \delta \rangle$ is constant almost surely, and the constant must be $\langle y, \delta \rangle$; hence (e) holds. Because all distributions in the family are mutually absolutely continuous by (4), (e) implies (f). We have now proved that (a) through (f) are equivalent.

By definition of normal cone and convex support, (e) and (g) are equivalent, and (g) and (h) are equivalent by the polarity relationship of normal and tangent cones (Rockafellar and Wets, 2004, Theorem 6.9 and Corollary 6.30). □

Proof of Corollary 2. By Theorem 7.1 and p. 140 in Barndorff-Nielsen (1978), l is concave; thus we must have

$$l\bigl(t\hat{\theta}_1 - (1-t)\hat{\theta}_2\bigr) \geq t l(\hat{\theta}_1) + (1-t) l(\hat{\theta}_2), \qquad 0 < t < 1, \tag{28}$$

and since $\hat{\theta}_1$ and $\hat{\theta}_2$ are MLE, (28) must actually hold with equality. Thus by (a) of the theorem $\hat{\theta}_1 - \hat{\theta}_2$ is a direction of constancy. □

For the proof of Theorem 3 we use Corollary 2.4.1 in Geyer (1990), which relies on Theorem 2.3 in Geyer (1990), but the proof of that theorem given in Geyer (1990) is murky at best. So we give a corrected version.

Corrected Proof of Theorem 2.3 in Geyer (1990). First, equation (2.5) in Geyer (1990) contains an obvious typographical error. It should read

$$\begin{aligned}(\operatorname{rc}\log c)(\phi) &= \lim_{s \to \infty} \frac{\log c(\theta + s\phi) - \log c(\theta)}{s} \\ &= \lim_{s \to \infty} \log \left(\left[\frac{c(\theta + s\phi)}{c(\theta) e^{s \sigma_K(\phi)}} \right]^{1/s} e^{\sigma_K(\phi)} \right)\end{aligned}$$

The rest of the proof of the $\lambda(H_\phi) > 0$ case is correct. In the proof of the of the $\lambda(H_\phi) = 0$ case, the last displayed formula of the proof is incorrect. Clearly

$$e^{a - \sigma_K(\phi)} F_\theta(A)^{1/s} \to e^{a - \sigma_K(\phi)}, \qquad \text{as } s \to \infty.$$

However, since $a < \sigma_K(\phi)$ was arbitrary, the limit can be made arbitrarily close to 1, and we see that

$$\left[\frac{c(\theta + s\phi)}{c(\theta) e^{s \sigma_K(\phi)}} \right]^{1/s} \to 1, \qquad \text{as } s \to \infty,$$

as is required for the completion of the proof. □

Proof of Theorem 3. The equivalence of (a) and (b) is Theorem 8.6 in Rockafellar (1970). The equivalence of (a) and (c) is Corollary 2.4.1 in Geyer (1990). The equivalence of (c) and (d) is mutual absolute continuity of the distributions in an exponential family, which follows from (4). The equivalence of (c) and (e) is immediate from our definition (8) of the normal cone. The equivalence of (e) and (f) is the polarity relationship of tangent and normal cones (Rockafellar and Wets, 2004, Theorem 6.9 and Corollary 6.30). □

Proof of Theorem 4. That (a) and (b) are equivalent is Theorem 2.5 in Geyer (1990). That (b) and (c) are equivalent follows from (g) of Theorem 1 and (e) of Theorem 3. That (c) and (d) are equivalent is the polarity relationship of tangent and normal cones. □

Proof of Corollary 5. By assumption $s \mapsto l(\theta + s\delta)$ is a nondecreasing function. Suppose to get a contradiction that

$$l(\theta + s_1\delta) = l(\theta + s_2\delta) \tag{29}$$

for some s_1 and s_2 such that both sides of (29) are finite and $s_1 < s_2$. In order that l be nondecreasing we must have

$$l(\theta + s_1\delta) = l(\theta + s\delta), \qquad s_1 \leq s \leq s_2$$

but then δ is a direction of constancy by Theorem 1 (a). □

Proof of Theorem 6. Except for the last sentence, this follows immediately from Theorem 2.2 in Geyer (1990). From (4)

$$\begin{aligned}
\Pr_{\theta+s\delta}(Y \in H) &= E_\psi\bigl\{I_H e^{\langle Y, \theta + s\delta - \psi\rangle - c(\theta+s\delta) + c(\psi)}\bigr\} \\
&= e^{s\langle y,\delta\rangle - c(\theta+s\delta) + c(\theta)} E_\psi\bigl\{I_H e^{\langle Y, \theta - \psi\rangle - c(\theta) + c(\psi)}\bigr\} \\
&= e^{s\langle y,\delta\rangle - c(\theta+s\delta) + c(\theta)} \Pr_\theta(Y \in H)
\end{aligned}$$

where I_H denotes the indicator function of the event $Y \in H$. By Corollary 5, the function $s \mapsto \langle y, \theta + s\delta\rangle - c(\theta + s\delta)$ is strictly increasing, hence so is $s \mapsto \Pr_{\theta+s\delta}(Y \in H)$. That $\Pr_{\theta+s\delta}(Y \in H) \to 1$ as $s \to \infty$ follows from Scheffé's lemma (see the comments following the theorem). The continuity assertion follows from the fact that the moment generating function of the random variable $\langle Y, \delta\rangle$ is

$$\begin{aligned}
E_\theta\bigl\{e^{s\langle Y,\delta\rangle}\bigr\} &= E_\psi\bigl\{e^{\langle Y, \theta + s\delta - \psi\rangle - c(\theta) + c(\psi)}\bigr\} \\
&= e^{c(\theta+s\delta) - c(\theta)}
\end{aligned}$$

Hence $s \mapsto c(\theta + s\delta)$ is actually infinitely differentiable and so is $s \mapsto \Pr_{\theta+s\delta}(Y \in H)$. □

Proof of Theorem 7. Suppose $L = V$. Then $\operatorname{con}(\operatorname{pos} V)$ is the subspace spanned by V, in which case a GDOR does not exist by Theorem 4.

Suppose $L \neq V$. Then by the polarity relationship of normal and tangent cones for each $v \in V \setminus L$ there exists $\delta_v \in N_C(y)$ such that $\langle v, \delta_v\rangle < 0$. Hence $-\delta_v \notin N_C(y)$ and $N_C(y)$ is not a vector subspace. So a GDOR does exist by Theorem 4.

Let $\delta^* = \sum_{v \in V \setminus L} \delta_v$. Then δ^* satisfies (16a) and (16b). Observe that $\delta \in N_C(y)$ if and only if (16a) holds and (16b) holds with $<$ replaced by $\leq$. Then it is clear that for every $\delta \in N_C(y)$ there exists $t > 1$ such that $t\delta^* + (1-t)\delta$ is in $N_C(y)$. Hence $\delta^* \in \operatorname{rint} N_C(y)$ by Theorem 6.4 in Rockafellar (1970). It now follows from Proposition 2.42 in Rockafellar and Wets (2004) that the set of points satisfying (16a) and (16b) is $\operatorname{rint} N_C(y)$. □

Proof of Corollary 8. In the proof of the theorem we saw that if a GDOR exists, then $L \neq V$ and $N_C(y)$ is not a vector subspace. □

Proof of Corollary 9. Since C is polyhedral convex, every tangent vector is of the form $s(w - y)$ for some $w \in C$ and $s \geq 0$, that is, the closure operation in (7) is not necessary. This implies, in particular, that for each $v \in L$ there exist points $w_{v,+}$ and $w_{v,-}$ in C and positive scalars $s_{v,+}$ and $s_{v,-}$ such that $\pm v = s_{v,\pm}(w_{v,\pm} - y)$. Observe that these $w_{v,\pm}$ are also in $C \cap H$, but no $w \in V \setminus L$ is in $C \cap H$. Thus $T_{C\cap H}(y) = \operatorname{con}(\operatorname{pos} L) = \operatorname{span} L$. Since $y + \operatorname{span} L \subset H$, we have $C \cap H \supset C \cap (y + \operatorname{span} L)$. If $C \cap H \not\subset C \cap (y + \operatorname{span} L)$, then we cannot have $T_{C\cap H}(y) = \operatorname{span} L$. □

Proof of Theorem 10. The polar of a convex cone K is

$$K^* = \{\, \delta : \langle w, \delta \rangle \le 0,\ w \in K \,\}$$

(Rockafellar and Wets, 2004, Section 6.E). The double polar theorem (Rockafellar and Wets, 2004, Corollary 6.2.1) says that $K^{**} = \operatorname{cl} K$. When K is closed, in particular when K is polyhedral, then $K^{**} = K$. Here let $K = \operatorname{con}(\operatorname{pos}(V_{\text{sub}} \setminus \{w\}))$. Then the feasible region for the linear program (20) is $-K^*$. Now the optimal value to (20) is nonpositive if and only if $\langle w, \delta \rangle \le 0$ for all $\delta \in -K^*$, which is equivalent by the double polar theorem to $w \in (-K^*)^* = -K$ or to $-w \in K$.

Now w is in (19) if and only if $-w$ is a linear combination of elements of V_{sub} with nonnegative coefficients, that is, if $-w = a \cdot w + \sum_{v \in V_{\text{sub}} \setminus \{w\}} a_v \cdot v$ where a and all the a_v are nonnegative scalars. But this happens if and only if $-w = \sum_{v \in V_{\text{sub}} \setminus \{w\}} (a_v / (1 + a)) \cdot v$, which is equivalent to $-w \in K$. □

Proof of Theorem 11. With probability one

$$Y - y = \sum_{v \in V_{\text{sat}}} b_v(Y) \cdot v$$

where all the coefficients $b_v(Y)$ are nonnegative. From (21a) and (21b) we can derive

$$\begin{aligned} \langle v, M\delta \rangle &= 0, \qquad && v \in L_{\text{sat}} \\ \langle v, M\delta \rangle &< 0, \qquad && v \in V_{\text{sat}} \setminus L_{\text{sat}} \end{aligned}$$

Hence

$$\langle Y - y, M\delta \rangle = \sum_{v \in V_{\text{sat}} \setminus L_{\text{sat}}} b_v(Y) \cdot \langle v, M\delta \rangle \tag{30}$$

and since all of the $\langle v, M\delta \rangle$ in (30) are strictly negative, the sum can only be zero if all the $b_v(Y)$, $v \in V_{\text{sat}} \setminus L_{\text{sat}}$ are zero. Thus the support of the limiting conditional model consists of points of the form $y + \sum_{v \in L_{\text{sat}}} b_v \cdot v$, where the coefficients are arbitrary. Since all such points are in the preimage of H_{sub} under the map $y \mapsto M^T y$, we conclude (22) holds. □

Acknowledgments

Much of the theory of this technical report comes from a thesis (Geyer, 1990) done under the supervision of Elizabeth Thompson.

References

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families.* Chichester: John Wiley.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory.* Hayward, CA: Institute of Mathematical Statistics.

Fukuda, K. (2008) cddlib package, version 094f. `http://www.ifor.math.ethz.ch/~fukuda/cdd_home/cdd.html`. Documentation is the file `cddlibman.pdf`, which is included in the `cddlib` source and also in every `rcdd` installation.

Geyer, C. J. (1990). *Likelihood and Exponential Families.* Unpublished Ph. D. Thesis. University of Washington.

Geyer, C. J. (2008). R package `aster` (Aster Models), version 0.7-4. `http://www.stat.umn.edu/geyer/aster/`.

Geyer, C. J. and Meeden, G. D. (2005). Fuzzy and randomized confidence intervals and P-values (with discussion). *Statistical Science*, **20**, 358–387.

Geyer, C. J. and Meeden, G. D. (2008). R package `rcdd` (C Double Description for R), version 1.1. `http://http://www.stat.umn.edu/geyer/rcdd/`. Incorporates code from Fukuda (2008).

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society, Series B*, **54** 657–99.

Geyer, C. J., Wagenius, S. and Shaw, R. G. (2007). Aster models for life history analysis. *Biometrika*, **94**, 415–426.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses.* New York: Wiley.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`.

Rockafellar, R. T. (1970). *Convex Analysis.* Princeton: Princeton University Press.

Rockafellar, R. T. and Wets, R. J.-B. (2004). *Variational Analysis*, corr. 2nd printing. Berlin: Springer-Verlag.