# Computationally efficient likelihood inference in exponential families when the maximum likelihood estimator does not exist

Daniel J. Eck
*Department of Biostatistics, Yale School of Public Health.*
*daniel.eck@yale.edu*

Charles J. Geyer
*Department of Statistics, University of Minnesota*
*geyer@umn.edu*

### Abstract

In a regular full exponential family, the maximum likelihood estimator (MLE) need not exist in the traditional sense, but the MLE may exist in the Barndorff-Nielsen completion of the family. Existing algorithms for finding the MLE in the Barndorff-Nielsen completion solve many linear programs; they are slow in small problems and too slow for large problems. We provide new, fast, and scalable methodology for finding the MLE in the Barndorff-Nielsen completion based on approximate null eigenvectors of the Fisher information matrix. Convergence of Fisher information follows from cumulant generating function convergence, conditions for which are given.

Key Words: Barndorff-Nielsen completion of exponential families; Convergence of moments; Cumulant generating function convergence; Generalized affine functions

## 1 Introduction

We develop an inferential framework for regular full discrete exponential families when the observed value of the canonical statistic lies on the boundary of its convex support. Then the maximum likelihood estimator (MLE) for the canonical parameter cannot exist [3, Theorem 9.13]. But the MLE may exist in a completion of the exponential family. Completions for exponential families have been described (in order of increasing generality) by Barndorff-Nielsen [3, pp. 154–156], Brown [4, pp. 191–201], Csiszár and Matúš [5], and Geyer [12, unpublished PhD thesis, Chapter 4]. The latter two are equivalent for full exponential families, but the last, most general, has much stronger algebraic properties that help with theory, so we use it. It also is the only completion that is a completion under no regularity conditions whatsoever (other than exponential family). It works for curved exponential families and other non-full exponential families ([5] works for non-full but closed convex families). Following Geyer [8] we will call all of these completions the Barndorff-Nielsen completion without fuss about the technical details differentiating them.

Geyer [8] developed ways to do hypothesis tests and confidence intervals when the MLE in an exponential family does not exist in the conventional sense. The hypothesis test scheme was credited to Fienberg (personal communication — an answer he gave to a question at the end of a talk). The confidence interval scheme generates one-sided non-asymptotic confidence intervals, because the MLE fails to exist in the conventional sense when the canonical statistic is on the boundary of its convex support, and this is an inherently one-sided situation, and conventional asymptotics do not work near the boundary.

To simplify both explanation and computation, Geyer [8] assumed the regularity conditions that Brown [4] assumed for his completion. These conditions of Brown hold for nearly all applications known to us (applications for which a more general completion are required include aster models [10] and Markov spatial point processes [7]). We will also need to use Brown's conditions to guarantee our methods work.

The issue of when the MLE exists in the conventional sense and what to do when it doesn't is very important because of the wide use of generalized linear models for discrete data and log-linear models for categorical data. In every application of these, existing statistical software gives completely invalid results when the MLE does not exist in the conventional sense, but most such software either does not check for this problem or does very weak checks that have high probability of both false positives and false negatives. Moreover, even if these checks correctly detect nonexistence of the MLE in the conventional sense, conventional software implements no valid procedures for statistical inference when this happens. When this issue is detected most users will go to smaller statistical models for which the MLE seems to exist, even though such models may neither fit the data nor address the questions of scientific interest. Geyer [8] gives examples with valid inference. Authoritative textbooks, such as Agresti [1, Section 6.5], discuss the issue but provide no solutions.

Thus a solution to this issue that is efficiently computable would be very important. The algorithms of Geyer [12, 8] and Albert and Anderson [2] are based on doing many linear programs. The algorithm of Geyer [8] is the most efficient; it is mostly due to Fukuda, who provided the underlying C code for the computational geometry functions of R package `rcdd` [11], which this algorithm uses. This algorithm does at most $n$ linear programs, where $n$ is the number of cases of a generalized linear model (GLM) or the number of cells in a contingency table, in order to determine the existence of the MLE in the conventional sense. Each of these linear programs has $p$ variables, where $p$ is the number of parameters of the model, and up to $n$ inequality constraints. Since linear programming can take time exponential in $n$ when pivoting algorithms are used, and since such algorithms are necessary in computational geometry to get correct answers despite inaccuracy of computer arithmetic (see the warnings about the need to use rational arithmetic in the documentation for R package `rcdd` [11]), these algorithms can be very slow. Typically, they take several minutes of computer time for toy problems and can take longer than users are willing to wait for real applications. These algorithms do have the virtue that if they use infinite-precision rational arithmetic, then their calculations are exact, as good as a mathematical proof.

Previous theoretical discussions of these issues that do not provide algorithms [3, 4, 5] use the notions of faces of convex sets or tangent cones or normal cones and all of these are much harder to compute than the algorithm of Geyer [8]. So they provide no direction toward efficient computing.

Because computational geometry is so slow and does not scale to large problems, we abandon it and return to calculations using the inexact computer arithmetic provided by computer hardware. Conventional maximum likelihood computations come close, in a sense, to finding the MLE in the Barndorff-Nielsen completion. They go uphill on the likelihood function until they meet their convergence criteria and stop. At this point, the canonical parameter estimates are still infinitely far away from their analogs for the MLE in the completion, but the corresponding probability distributions are close in total variation norm. Here we show that they are also close in the sense of moment generating function convergence (Theorem 7 below) and consequently moments of all orders are also close. The MLE in the completion is not only a limit of distributions in the original family but also a distribution in the original family conditioned on the affine hull of a face of the effective domain of the log likelihood supremum function [12, Theorem 4.3, special cases of which were known to other authors]. To do valid statistical inference when the MLE does not exist in the conventional sense, we need to know this affine hull.

This affine hull is the support of the canonical statistic under the MLE distribution (in the completion). Hence it is a translate of the null space of the Fisher information matrix, which (for an exponential family) is the variance-covariance matrix of the canonical statistic. This affine hull must contain the mean vector of the canonical statistic under the MLE distribution. Hence knowing the mean vector and variance-covariance matrix of the canonical statistic under the MLE distribution allows us to do valid statistical inference, and our conventional maximum likelihood calculation (go uphill until things don't change much in an iteration) will give us good approximations of them (relative to the inexactness of computer arithmetic).

We will get nearly the correct affine hull if we can guess the correct null space of the Fisher information matrix from its eigenvalues and eigenvectors computed using inexact computer arithmetic. We will not be able to do this when the statistical model has an ill-conditioned model matrix (the model matrix for categorical data analysis being the model matrix when it is recast as a Poisson regression). Ill-conditioning will add spurious nearly zero eigenvalues that arise from the ill-conditioning rather than the concentration of the MLE distribution on the correct affine hull. We will suppose that the model matrix is not ill-conditioned. If a sequence of parameter estimates maximizes the likelihood, then the corresponding sequence of probability density functions (PDFs) has subsequences converging to PDFs of MLE distributions in the Barndorff-Nielsen completion [12, Theorem 4.1]. If the MLE distribution is unique, as it always is for a full exponential family [8, Section 3.8], then all of these MLE PDFs will correspond to the same probability distribution. For a curved exponential family, the MLE need not be unique, even when it exists in the conventional sense.

## 2 Motivating example

Consider the case of complete separation in the logistic regression model as an example of a discrete exponential family with data on the boundary of the convex support of the canonical statistic. Suppose that we have one predictor vector $x$ having values 10, 20, 30, 40, 60, 70, 80, 90, and suppose the components of the response vector $y$ are 0, 0, 0, 0, 1, 1, 1, 1. Then the simple logistic regression model that has linear predictor $\eta = \beta_0 + \beta_1 x$ exhibits failure of the MLE to exist in the conventional sense. This example is the same as that of Agresti [1, Section 6.5.1].

For an exponential family, the submodel canonical statistic is $M^T y$, where $M$ is the model matrix [8, Section 3.9]. Figure 1 shows the observed value of the canonical statistic vector and the support (all possible values) of this vector. As is obvious from the figure, the observed value of the canonical statistic is on the boundary of the convex support, in which case the MLE does not exist in the conventional sense [8, Theorem 4]. In general, this figure is too computationally intensive and too high-dimensional to draw. So our methods do not use such figures. It is here to develop intuition. In our methodology, this degeneracy follows from the Fisher information matrix at the apparent MLE being nearly the zero matrix.

In this example, like in Example 1 of Geyer [8], the MLE in the Barndorff-Nielsen completion corresponds to a completely degenerate distribution. This MLE distribution says no other data than what was observed could be observed. But the sample is not the population and estimates are not parameters. So this degeneracy is not a problem. To illustrate the uncertainty of estimation we follow Figure 2 of Geyer [8], which shows confidence intervals (necessarily one-sided) for the saturated model mean value parameters. Our Figure 2 shows that, as would be expected from so little data, the confidence intervals are very wide. The MLE in the completion says the probability of observing a response equal to one jumps from zero to one somewhere between 40 and 60. The confidence intervals show that we are fairly sure that this probability goes from near zero at $x = 10$ to near one at $x = 90$ but we are very unsure where jumps are if there are any. These intervals were constructed using the theory of Geyer [8, Section 3.16]. The actual computations follow some later course notes [9].

Our theory allows for inference in not only the complete separation example but also in any

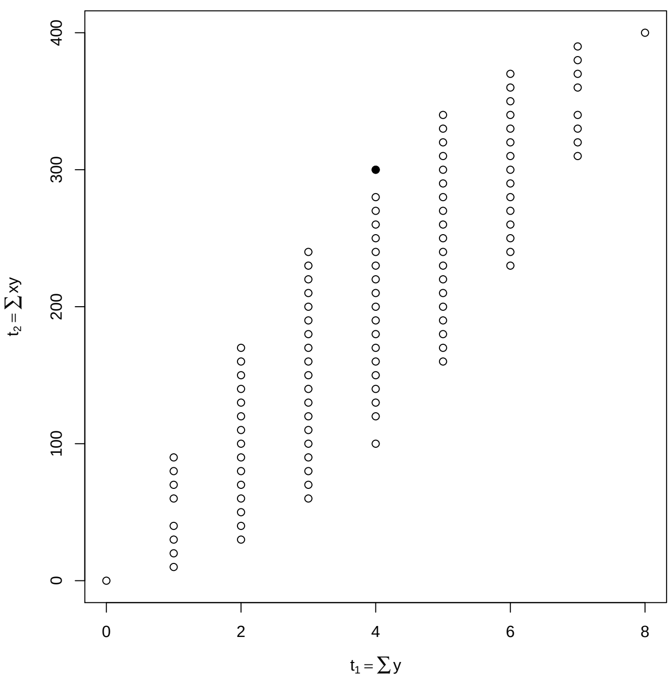

Figure 1: Observed value and support of the submodel canonical statistic vector $M^T y$ for the example of Section 2. Solid dot is the observed value of this statistic.

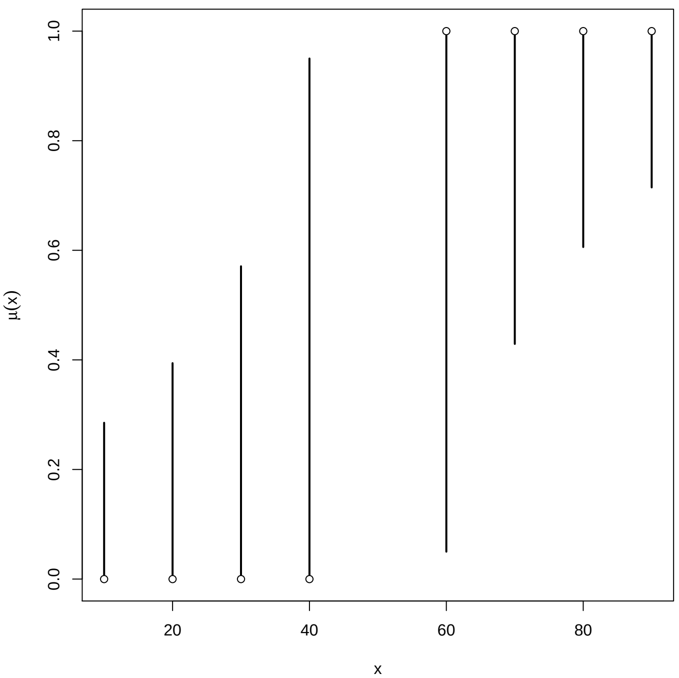

Figure 2: One-sided 95% confidence intervals for saturated model mean value parameters. Bars are the intervals; $\mu(x)$ is the probability of observing response value one when the predictor value is $x$. Solid dots are the observed data.

discrete regular full exponential family where the MLE does not exist in the traditional sense. For further motivation, see the examples in Section 2 of [8]. We redo Example 2.3 of [8] in our Section 7.2, and we find that our methodology produces the same inferences as theirs in a fraction of the time.

## 3 Laplace transforms and standard exponential families

Let $\lambda$ be a positive Borel measure on a finite-dimensional vector space $E$. The *log Laplace transform* of $\lambda$ is the function $c : E^* \to \overline{\mathbb{R}}$ defined by

$$c(\theta) = \log \int e^{\langle x, \theta \rangle} \, \lambda(dx), \qquad \theta \in E^*, \tag{1}$$

where $E^*$ is the dual space of $E$, where $\langle \cdot \, , \, \cdot \rangle$ is the canonical bilinear form placing $E$ and $E^*$ in duality, and where $\overline{\mathbb{R}}$ is the extended real number system, which adds the values $-\infty$ and $+\infty$ to the real numbers with the obvious extensions to the arithmetic and topology [15, Section 1.E].

If one prefers, one can take $E = E^* = \mathbb{R}^p$ for some $p$, and define

$$\langle x, \theta \rangle = \sum_{i=1}^{p} x_i \theta_i, \qquad x \in \mathbb{R}^p \text{ and } \theta \in \mathbb{R}^p,$$

but the coordinate-free view of vector spaces offers more generality and more elegance. Also, as we are about to see, if $E$ is the sample space of a standard exponential family, then a subset of $E^*$ is the canonical parameter space, and the distinction between $E$ and $E^*$ helps remind us that we should not consider these two spaces to be the same space.

A log Laplace transform is a lower semicontinuous convex function that nowhere takes the value $-\infty$ (the value $+\infty$ is allowed and occurs where the integral in (1) does not exist) [12, Theorem 2.1]. The *effective domain* of an extended-real-valued convex function $c$ on $E^*$ is

$$\operatorname{dom} c = \{ \, \theta \in E^* : c(\theta) < +\infty \, \}.$$

For every $\theta \in \operatorname{dom} c$, the function $f_\theta : E \to \mathbb{R}$ defined by

$$f_\theta(x) = e^{\langle x, \theta \rangle - c(\theta)}, \qquad x \in E, \tag{2}$$

is a probability density with respect to $\lambda$. The set $\mathcal{F} = \{ \, f_\theta : \theta \in \Theta \, \}$, where $\Theta$ is any nonempty subset of $\operatorname{dom} c$, is called a *standard exponential family of densities with respect to* $\lambda$. This family is *full* if $\Theta = \operatorname{dom} c$. We also say $\mathcal{F}$ is the standard exponential family *generated by* $\lambda$ having canonical parameter space $\Theta$, and $\lambda$ is the *generating measure* of $\mathcal{F}$.

The log likelihood of this family is (2) have log likelihood

$$l(\theta) = \langle x, \theta \rangle - c(\theta). \tag{3}$$

A general exponential family [12, Chapter 1] is a family of probability distributions having a sufficient statistic $X$ taking values in a finite-dimensional vector space $E$ that induces a family of distributions on $E$ that have a standard exponential family of densities with respect to some generating measure. Reduction by sufficiency loses no statistical information, so the theory of standard exponential families tells us everything about general exponential families [12, Section 1.2].

In the context of general exponential families $X$ is called the *canonical statistic* and $\theta$ the *canonical parameter* (the terms *natural statistic* and *natural parameter* are also used). The set $\Theta$

is the canonical parameter space of the family, the set $\operatorname{dom} c$ is the canonical parameter space of the full family having the same generating measure. A full exponential family is said to be *regular* if its canonical parameter space $\operatorname{dom} c$ is an open subset of $E^*$.

The cumulant generating function (CGF) of the distribution of the canonical statistic for parameter value $\theta$ is the function $k_\theta$ defined by

$$\begin{aligned} k_\theta(t) &= \log \int e^{\langle x, t \rangle} f_\theta(x) \, \lambda(dx) \\ &= c(\theta + t) - c(\theta) \end{aligned} \tag{4}$$

provided this distribution has a CGF, which it does if and only if $k_\theta$ is finite on a neighborhood of zero, that is, if and only if $\theta \in \operatorname{int}(\operatorname{dom} c)$. Thus every distribution in a full family has a CGF if and only if the family is regular. Derivatives of $k_\theta$ evaluated at zero are the cumulants of the distribution for $\theta$. These are the same as derivatives of $c$ evaluated at $\theta$.

# 4 Generalized affine functions

## 4.1 Characterization on affine spaces

Exponential families defined on affine spaces instead of vector spaces are in many ways more elegant [12, Sections 1.4 and 1.5 and Chapter 4]. To start, a family of densities with respect to a positive Borel measure on an affine space is a *standard exponential family* if the log densities are affine functions. Following Geyer [12, Chapter 4], we complete the exponential family by taking pointwise limits of densities, allowing $+\infty$ and $-\infty$ as limits.

We call these limits *generalized affine functions.* Real-valued affine functions on an affine space are functions that are are both convex and concave. *Generalized affine functions* on an affine space are extended-real-valued functions that are are both concave and convex [12, Chapter 4]. (For a definition of extended-real-valued convex functions see Rockafellar [14, Chapter 4].)

We thus have two characterizations of generalized affine functions: functions that are both convex and concave and functions that are limits of sequences of affine functions. Further characterizations will be given below.

Let $h_n$ denote a sequence of affine functions that are log densities in a standard exponential family with respect to $\lambda$, that is, $\int e^{h_n} \, d\lambda = 1$ for all $n$. Since $e^{h_n} \to e^h$ pointwise if and only if $h_n \to h$ pointwise, the idea of completing an exponential family naturally leads to the the study of generalized affine functions.

If $h : E \to \overline{\mathbb{R}}$ is a generalized affine function, we use the notation

$$\begin{aligned} h^{-1}(\mathbb{R}) &= \{ x \in E : h(x) \in \mathbb{R} \} \\ h^{-1}(\infty) &= \{ x \in E : h(x) = \infty \} \\ h^{-1}(-\infty) &= \{ x \in E : h(x) = -\infty \} \end{aligned}$$

**Theorem 1.** *An extended-real-valued function $h$ on a finite-dimensional affine space $E$ is generalized affine if and only if one of the following cases holds*

(a) $h^{-1}(\infty) = E$,

(b) $h^{-1}(-\infty) = E$,

(c) $h^{-1}(\mathbb{R}) = E$ *and $h$ is an affine function, or*

(d) *there is a hyperplane $H$ such that $h(x) = \infty$ for all points on one side of $H$, $h(x) = -\infty$ for all points on the other side of $H$, and $h$ restricted to $H$ is a generalized affine function.*

All theorems for which a proof does not follow the theorem statement are proved in either the appendix or the supplementary material.

The intention is that this theorem is applied recursively. If we are in case (d), then the restriction of $h$ to $H$ is another generalized affine function to which the theorem applies. Since a nested sequence of hyperplanes can have length at most the dimension of $E$, the recursion always terminates.

## 4.2 Topology

Let $G(E)$ denote the space of generalized affine functions on a finite-dimensional affine space $E$ with the topology of pointwise convergence.

**Theorem 2.** *$G(E)$ is a compact Hausdorff space.*

**Theorem 3.** *$G(E)$ is a first countable topological space.*

**Corollary 1.** *$G(E)$ is sequentially compact.*

Sequentially compact means every sequence has a (pointwise) convergent subsequence. That this follows from the two preceding theorems is well known [16, p. 22, gives a proof].

The space $G(E)$ is not metrizable, unless $E$ is zero-dimensional [12, penultimate paragraph of Section 3.3]. So we cannot use $\delta$-$\varepsilon$ arguments, but we can use arguments involving sequences, using sequential compactness.

Let $\lambda$ be a positive Borel measure on $E$, and let $\mathcal{H}$ be a nonempty subset of $G(E)$ such that

$$\int e^h \, d\lambda = 1, \qquad h \in \mathcal{H}. \tag{5}$$

Then, following Geyer [12, Chapter 4], we call $\mathcal{H}$ a *standard generalized exponential family* of log densities with respect to $\lambda$. Let $\overline{\mathcal{H}}$ denote the closure of $\mathcal{H}$ in $G(E)$.

**Theorem 4.** *Maximum likelihood estimates always exist in the closure $\overline{\mathcal{H}}$.*

*Proof.* Suppose $x$ is the observed value of the canonical statistic. Then there exists a sequence $h_n$ in $\mathcal{H}$ such that

$$h_n(x) \to \sup_{h \in \mathcal{H}} h(x).$$

This sequence has a convergent subsequence $h_{n_k} \to h$ in $G(E)$. This limit $h$ is in $\overline{\mathcal{H}}$ and maximizes the likelihood. □

We claim this is the right way to think about completion of exponential families. For full exponential families or even closed convex exponential families the closure only contains *proper* log probability densities ($h$ that satisfy the equation in (5)). This is shown by Geyer [12, Chapter 2] and also by Csiszár and Matúš [5].

For curved exponential families and for general non-full exponential families, applying Fatou's lemma to pointwise convergence in $G(E)$ gives only

$$0 \le \int e^h \, d\lambda \le 1, \qquad h \in \overline{\mathcal{H}}. \tag{6}$$

When the integral in (6) is strictly less than one we say $h$ is an *improper* log probability density. The examples in Geyer [12, Chapter 4] show that improper probability densities cannot be avoided in curved exponential families.

Geyer [12, Theorem 4.3] shows that this closure of an exponential family can be thought of as a union of exponential families, so this generalizes the conception of Brown [4] of the closure as an *aggregate exponential family.* Thus our method generalizes all previous methods of completing exponential families.

Admittedly, this characterization of the completion of an exponential family is very different from any other in its ignoring of parameters. Only log densities appear. Unless one wants to call them parameters — and that conflicts with the usual definition of parameters as real-valued — parameters just do not appear.

So in the next section, we bring parameters back.

### 4.3 Characterization on vector spaces

In this section we take sample space $E$ to be vector space (which, of course, is also an affine space, so the results of the preceding section continue to hold). Recall from Section 3 above, that $E^*$ denotes the dual space of $E$, which contains the canonical parameter space of the exponential family.

**Theorem 5.** *An extended-real-valued function $h$ on a finite-dimensional vector space $E$ is generalized affine if and only if there exist finite sequences (perhaps of length zero) of vectors $\eta_1, \ldots, \eta_j$ in in $E^*$ and scalars $\delta_1, \ldots, \delta_j$ such that $\eta_1, \ldots, \eta_j$ are linearly independent and $h$ has the following form. Define $H_0 = E$ and, inductively, for integers $i$ such that $0 < i \leq j$*

$$\begin{aligned} H_i &= \{ x \in H_{i-1} : \langle x, \eta_i \rangle = \delta_i \} \\ C_i^+ &= \{ x \in H_{i-1} : \langle x, \eta_i \rangle > \delta_i \} \\ C_i^- &= \{ x \in H_{i-1} : \langle x, \eta_i \rangle < \delta_i \} \end{aligned}$$

*all of these sets (if any) being nonempty. Then $h(x) = +\infty$ whenever $x \in C_i^+$ for any $i$, $h(x) = -\infty$ whenever $x \in C_i^-$ for any $i$, and $h$ is either affine or constant on $H_j$, where $+\infty$ and $-\infty$ are allowed for constant values.*

The "if any" refers to the case where the sequences have length zero, in which case the theorem asserts that $h$ that $h$ is affine on $E$ or constant on $E$.

As we saw in the preceding section, we are interested in likelihood maximizing sequences. Here we represent the likelihood maximizing sequence in the coordinates of the linearly independent $\eta$ vectors that characterize the generalized affine function $h$ according to its Theorem 5 representation. Let $\theta_n$ be a likelihood maximizing sequence of canonical parameter vectors, that is,

$$l(\theta_n) \to \sup_{\theta \in \Theta} l(\theta), \qquad \text{as } n \to \infty, \tag{7}$$

where the log likelihood $l$ is given by (3) and where $\Theta$ is the canonical parameter space of the family. To make connection with the preceding section, define

$$h_\theta(x) = l(\theta) = \langle x, \theta \rangle - c(\theta).$$

Then $h_{\theta_n}$ is a sequence of affine functions, which has a subsequence that converges (in $G(E)$) to some generalized affine function $h \in \overline{\mathcal{H}}$, which maximizes the likelihood:

$$h(x) = \sup_{\theta \in \Theta} l(\theta). \tag{8}$$

The following lemma gives us a better understanding of the convergence $h_{\theta_n} \to h$.

**Lemma 1.** *Suppose that a generalized affine function $h$ on a finite dimensional vector space $E$ is finite at at least one point. Represent $h$ as in Theorem 5, and extend $\eta_1$, ..., $\eta_j$ to be a basis $\eta_1$, ..., $\eta_p$ for $E^*$. Suppose $h_n$ is a sequence of affine functions converging to $h$ in $G(E)$. Then there are sequences of scalars $a_n$ and $b_{i,n}$ such that*

$$h_n(y) = a_n + \sum_{i=1}^{j} b_{i,n} \left( \langle y, \eta_i \rangle - \delta_i \right) + \sum_{i=j+1}^{p} b_{i,n} \langle y, \eta_i \rangle, \qquad y \in E, \tag{9}$$

*and, as $n \to \infty$, we have*

(a) $b_{i,n} \to \infty$, *for* $1 \leq i \leq j$,

(b) $b_{i,n}/b_{i-1,n} \to 0$, *for* $2 \leq i \leq j$,

(c) $b_{i,n}$ *converges, for* $i > j$, *and*

(d) $a_n$ *converges.*

In (9) the first sum is empty when $j = 0$ and the second sum is empty when $j = p$. Such empty sums are zero by convention.

The results given in Lemma 1 are applicable to generalized affine functions in full generality. The case of interest to us, however, is when $h_n = h_{\theta_n}$ is the likelihood maximizing sequence constructed above.

**Corollary 2.** *For data $x$ from a regular full exponential family defined on a vector space $E$, suppose $\theta_n$ is a likelihood maximizing sequence satisfying* (7) *with log densities $h_n = h_{\theta_n}$ defined by* (8) *converging pointwise to a generalized affine function $h$. Characterize $h$ and $h_n$ as in Theorem 5 and Lemma 1. Define $\psi_n = \sum_{i=j+1}^{p} b_{i,n} \langle x, \eta_i \rangle$. Then conclusions* (a) *and* (b) *of Lemma 1 hold in this setting and*

$$\psi_n \to \theta^*, \quad \text{as } n \to \infty,$$

*where $\theta^*$ is the MLE of the exponential family conditioned to $H_j$.*

In case $j = p$ the conclusion $\psi_n \to \theta^*$ is the trivial zero converges to zero. The original exponential family conditioned on the event $H_j$ is what Geyer [8] calls the limiting conditional model (LCM).

*Proof.* The conditions of Lemma 1 are satisfied by our assumptions so all conclusions of Lemma 1 are satisfied. As a consequence, $\psi_n \to \theta^*$ as $n \to \infty$. The fact that $\theta^*$ is the MLE of the LCM restricted to $H_j$ follows from our assumption that $\theta_n$ is a likelihood maximizing sequence. □

Taken together, Theorem 5, Lemma 1, and Corollary 2 provide a theory of maximum likelihood estimation in the completions of exponential families that is the theory of the preceding section with canonical parameters brought back.

# 5 Convergence theorems

## 5.1 Cumulant generating function convergence

We now show CGF convergence along likelihood maximizing sequences (7). This implies convergence in distribution and convergence of moments of all orders.

Theorems 6 and 7 in this section say when CGF convergence occurs. Their conditions are somewhat unnatural (especially those of Theorem 6). However, the example in Section 4 of the supplementary material shows not only that some conditions are necessary to obtain CGF convergence (it does not occur for all full discrete exponential families) but also that the conditions of Theorem 6 are sharp, being just what is needed to rule out that example.

The CGF of the distribution having log density that is the generalized affine function $h$ is defined by

$$\kappa(t) = \log \int e^{\langle y,t \rangle} e^{h(y)} \, \lambda(dy),$$

and similarly

$$\kappa_n(t) = \log \int e^{\langle y,t \rangle} e^{h_n(y)} \, \lambda(dy)$$

where we assume $h_n$ are the log densities for a likelihood maximizing sequence such that $h_n \to h$ pointwise. The next theorem characterizes when $\kappa_n \to \kappa$ pointwise.

Let $c_A$ denote the log Laplace transform of the restriction of $\lambda$ to the set $A$, that is,

$$c_A(\theta) = \log \int_A e^{\langle y,\theta \rangle} \, \lambda(dy),$$

where, as usual, the value of the integral is taken to be $+\infty$ when the integral does not exist (a convention that will hold for the rest of this section).

**Theorem 6.** *Let $E$ be a finite-dimensional vector space of dimension $p$. For data $x \in E$ from a regular full exponential family with natural parameter space $\Theta \subseteq E^*$ and generating measure $\lambda$, assume that all LCMs are regular exponential families. Suppose that $\theta_n$ is a likelihood maximizing sequence satisfying* (7) *with log densities $h_n$ converging pointwise to a generalized affine function $h$. Characterize $h$ as in Theorem 5. When $j \geq 2$, and for $i = 1, ..., j-1$, define*

$$\begin{aligned} D_i &= \{y \in C_i^- : \langle y, \eta_k \rangle > \delta_k, \text{ some } k > i\}, \\ F &= E \setminus \cup_{i=1}^{j-1} D_i = \{y : \langle y, \eta_i \rangle \leq \delta_i, \ 1 \leq i \leq j\}, \end{aligned} \tag{10}$$

*and assume that*

$$\sup_{\theta \in \Theta} \sup_{y \in \cup_{i=1}^{j-1} D_i} e^{\langle y,\theta \rangle - c_{\cup_{i=1}^{j-1} D_i}(\theta)} < \infty \quad \text{or} \quad \lambda\left(\cup_{i=1}^{j-1} D_i\right) = 0. \tag{11}$$

*Then $\kappa_n(t)$ converges to $\kappa(t)$ pointwise for all $t$ in a neighborhood of 0.*

Discrete exponential families automatically satisfy (11) when

$$\inf_{y \in \cup_{i=1}^{j-1} D_i} \lambda(\{y\}) > 0.$$

In this setting, $e^{\langle y,\theta \rangle - c_{\cup_{i=1}^{j-1} D_i}(\theta)}$ corresponds to the probability mass function for the random variable conditional on the occurrence of $\cup_{i=1}^{j-1} D_i$. Thus,

$$\begin{aligned} &\sup_{\theta \in \Theta} \sup_{y \in \cup_{i=1}^{j-1} D_i} \left( e^{\langle y,\theta \rangle - c_{\cup_{i=1}^{j-1} D_i}(\theta)} \right) \\ &\quad = \sup_{\theta \in \Theta} \sup_{y \in \cup_{i=1}^{j-1} D_i} \left( \frac{e^{\langle y,\theta \rangle} \lambda(\{y\})}{\lambda(\{y\}) \sum_{x \in \cup_{i=1}^{j-1} D_i} e^{\langle x,\theta \rangle} \lambda(\{x\})} \right) \end{aligned}$$

$$\leq \sup_{y \in \cup_{i=1}^{j-1} D_i} (1/\lambda(\{y\})) < \infty.$$

Therefore, Theorem 6 is applicable for the non-existence of the maximum likelihood estimator that may arise in logistic and multinomial regression.

We show in the next theorem that discrete families with convex polyhedral support $K$ also satisfy (11) under additional regularity conditions that hold in practical applications. When $K$ is convex polyhedron, we can write $K = \{y : \langle y, \alpha_i \rangle \leq a_i, \text{ for } i = 1, ..., m\}$, as in [15, Theorem 6.46]. When the MLE does not exist, the data $x \in K$ is on the boundary of $K$. Denote the active set of indices corresponding to the boundary $K$ containing $x$ by $I(x) = \{i : \langle x, \alpha_i \rangle = a_i\}$. In preparation for Theorem 7 we define the normal cone $N_K(x)$, the tangent cone $T_K(x)$, and faces of convex sets and then state conditions required on $K$.

**Definition 1.** *The normal cone of a convex set $K$ in the finite dimensional vector space $E$ at a point $x \in K$ is*

$$N_K(x) = \{\, \eta \in E^* : \langle y - x, \eta \rangle \leq 0 \textit{ for all } y \in K \,\}.$$

**Definition 2.** *The tangent cone of a convex set $K$ in the finite dimensional vector space $E$ at a point $x \in K$ is*

$$T_K(x) = \mathrm{cl}\{\, s(y - x) : y \in K \text{ and } s \geq 0 \,\}$$

*where* cl *denotes the set closure operation.*

When $K$ is a convex polyhedron, $N_K(x)$ and $T_K(x)$ are both convex polyhedron with formulas given in [15, Theorem 6.46]. These formulas are

$$T_K(x) = \{y : \langle y, \alpha_i \rangle \leq 0 \text{ for all } i \in I(x)\},$$
$$N_K(x) = \{c_1\alpha_1 + \cdots + c_m\alpha_m : c_i \geq 0 \text{ for } i \in I(x),\ c_i = 0 \text{ for } i \notin I(x)\}.$$

**Definition 3.** *A* face *of a convex set $K$ is a convex subset $F$ of $K$ such that every (closed) line segment in $K$ with a relative interior point in $F$ has both endpoints in $F$. An* exposed face *of $K$ is a face where a certain linear function achieves its maximum over $K$ [14, p. 162].*

The conditions required on $K$ for our theory to hold are from Brown [4, pp. 193–197]. These conditions are:

(i) The support of the exponential family is a countable set $X$.

(ii) The exponential family is regular.

(iii) Every $x \in X$ is contained in the relative interior of an exposed face $F$ of the convex support $K$.

(iv) The convex support of the measure $\lambda|F$ equals $F$, where $\lambda$ is the generating measure for the exponential family.

Conditions (i) and (ii) are already assumed in Theorem 6. It is now shown that discrete exponential families satisfy (11) under the above conditions.

**Theorem 7.** *Assume the conditions of Theorem 6 with the omission of* (11) *when $j \geq 2$. Let $K$ denote the convex support of the exponential family. Assume that the exponential family satisfies the conditions of Brown. Then* (11) *holds.*

### 5.2 Consequences of CGF convergence

Theorems 6 and 7 both verify CGF convergence along likelihood maximizing sequences (7) on neighborhoods of 0. The next theorems show that CGF convergence on neighborhoods of 0 is enough to imply convergence in distribution and of moments of all orders. Therefore moments of distributions with log densities that are affine functions converge along likelihood maximizing sequences (7) to those of a limiting distributions whose log density is a generalized affine function.

Suppose that $X$ is a random vector in a finite-dimensional vector space $E$ having a moment generating function (MGF) $\varphi_X$, then

$$\varphi_X(t) = \varphi_{\langle X,t \rangle}(1), \qquad t \in E^*,$$

regardless of whether the MGF exist or not. It follows that the MGF of $\langle X, t\rangle$ for all $t$ determine the MGF of $X$ and vice versa, when these MGF exist. More generally,

$$\varphi_{\langle X,t \rangle}(s) = \varphi_X(st), \qquad t \in E^* \text{ and } s \in \mathbb{R}. \tag{12}$$

This observation applied to characteristic functions rather than MGF is called the Cramér-Wold theorem. In that context it is more trivial because characteristic functions always exist.

If $v_1, \ldots, v_d$ is a basis for a vector space $E$, then there exists a unique dual basis $w_1, \ldots, w_d$ for $E^*$ that satisfies

$$\langle v_i, w_j \rangle = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{13}$$

[13, Theorem 2 of Section 15].

**Theorem 8.** *If $X$ is a random vector in $E$ having an MGF, then the random scalar $\langle X, t\rangle$ has an MGF for all $t \in E^*$. Conversely, if $\langle X, t\rangle$ has an MGF for all $t \in E^*$, then $X$ has an MGF.*

**Theorem 9.** *Suppose $X_n$, $n = 1, 2, \ldots$ is a sequence of random vectors, and suppose their moment generating functions converge pointwise on a neighborhood $W$ of zero. Then*

$$X_n \xrightarrow{d} X, \tag{14}$$

*and $X$ has an MGF $\varphi_X$, and*

$$\varphi_{X_n}(t) \to \varphi_X(t), \qquad t \in E^*.$$

**Theorem 10.** *Under the assumptions of Theorem 9, suppose $t_1, t_2, \ldots, t_k$ are vectors defined on $E^*$, the dual space of $E$. Then $\prod_{i=1}^k \langle X_n, t_i \rangle$ is uniformly integrable so*

$$\mathrm{E}\left\{ \prod_{i=1}^k \langle X_n, t_i \rangle \right\} \to \mathrm{E}\left\{ \prod_{i=1}^k \langle X, t_i \rangle \right\}.$$

The combination of Theorems 6-10 provide a methodology for statistical inference along likelihood maximizing sequences when the MLE is in the Barndorff-Nielsen completion. In particular, we have convergence in distribution and convergence of moments of all orders along likelihood maximizing sequence. The limiting distribution in this context is a generalized exponential family with density $e^h$ where $h$ is a generalized affine function.

### 5.3 Convergence of null spaces of Fisher information

Our method for finding the MLE in the Barndorff-Nielsen completion relies on finding the null space of the Fisher information matrix. We need to show that we have convergence for that. In order to prove this we need an appropriate notion of convergence of vector subspaces.

**Definition 4.** *Painlevé-Kuratowski set convergence [15, Section 4.A] can be defined as follows (Rockafellar and Wets [15] give many equivalent characterizations). If $C_n$ is a sequence of sets in $\mathbb{R}^p$ and $C$ is another set in $\mathbb{R}^p$, then we say $C_n \to C$ if*

(i) *For every $x \in C$ there exists a subsequence $n_k$ of the natural numbers and there exist $x_{n_k} \in C_{n_k}$ such that $x_{n_k} \to x$.*

(ii) *For every sequence $x_n \to x$ in $\mathbb{R}^p$ such that there exists a natural number $N$ such that $x_n \in C_n$ whenever $n \geq N$, we have $x \in C$.*

**Theorem 11.** *Suppose that $A_n \in \mathbb{R}^{p \times p}$ is a sequence of positive semidefinite matrices and $A_n \to A$ componentwise. Fix $\varepsilon > 0$ less than half of the least nonzero eigenvalue of $A$ unless $A$ is the zero matrix in which case $\varepsilon > 0$ may be chosen arbitrarily. Let $V_n$ denote the subspace spanned by the eigenvectors of $A_n$ corresponding to eigenvalues that are less than $\varepsilon$. Let $V$ denote the null space of $A$. Then $V_n \to V$ (Painlevé-Kuratowski).*

## 6 Calculating the MLE in the completion

### 6.1 Assumptions

So far everything has been for general exponential families except for Theorems 6 and 7, the later of which assumes the conditions of Brown [4], and those conditions hold for GLM and log-linear models for categorical data analysis.

Now, following Geyer [8] we restrict our attention to discrete GLM. This, in effect, includes log-linear models for contingency tables because we can always assume Poisson sampling, which makes them equivalent to GLM [1, Section 8.6.7; 8, Section 3.17].

### 6.2 The form of the MLE in the completion

#### 6.2.1 First characterization

Suppose we know the *affine support* of the MLE distribution in the completion. This is the smallest affine set that contains the canonical statistic with probability one. Denote the affine support by $A$. An affine set is a translate of a vector subspace. Since the observed value of the canonical statistic is contained in $A$ with probability one, and the canonical statistic for a GLM is $M^T Y$, where $M$ is the model matrix, $Y$ is the response vector, and $y$ its observed value [8, Section 3.9], we have $A = M^T y + V$ for some vector space $V$.

Then the LCM in which the MLE in the completion is found is the OM conditioned on the event

$$M^T(Y - y) \in V, \qquad \text{almost surely}$$

[12, Theorem 4.3]. Suppose we characterize $V$ as the subspace where a finite set of linear equalities are satisfied

$$V = \{\, w \in \mathbb{R}^p : \langle w, \eta_i \rangle = 0,\ i = 1, \ldots, j \,\}.$$

Then the LCM is the OM conditioned on the event

$$\langle M^T(Y - y), \eta_i \rangle = \langle Y - y, M\eta_i \rangle = 0, \qquad i = 1, \ldots, j.$$

From this we see that the vectors $\eta_1, \ldots, \eta_j$ span the null space of the Fisher information matrix for the LCM, which our Theorems 7 and 10 say is well approximated by the Fisher information matrix for the OM at parameter values that are close to maximizing the likelihood.

The vector subspace spanned by the vectors $\eta_1, \ldots, \eta_j$ is called the *constancy space* of the LCM in [8].

### 6.2.2 Second characterization

Any vector $\delta$ in the canonical parameter space of an exponential family is called a *direction of recession* (DOR) if the likelihood function is nondecreasing in that direction or, equivalently, if

$$\langle Y, M\delta \rangle \le \langle y, M\delta \rangle, \qquad \text{almost surely,} \tag{15}$$

where $Y$ denotes the response vector considered as a random vector, $y$ denotes its observed value, and $M$ is the model matrix [12, Section 2.2; 8, Theorem 3 and the following discussion and Section 3.9].

If we can find a DOR, then the MLE in the completion is a distribution in the LCM, which is the family of distributions in the original model (OM) conditioned on the event

$$\langle Y, M\delta \rangle = \langle y, M\delta \rangle, \qquad \text{almost surely,} \tag{16}$$

or a distribution in the completion of the LCM [12, Chapter 2; 8, Section 3].

Define $\zeta = M\delta$. In light of (15) the only way (16) can hold is if $\zeta_i \neq 0$ implies $Y_i = y_i$ almost surely.

Thus the distributions in the LCM are the distributions in the OM conditioned on this event. Moreover the MLE in the LCM is easily found using standard software. We simply change the model by removing the components of the response that are fixed in the LCM. Using R function GLM this is done using the optional argument `subset = zeta == 0`, where `zeta` is the R object corresponding to the vector $\zeta$.

If the MLE for the LCM exists in the conventional sense, then we have solved the problem of finding the MLE in the completion of the OM. If not we have to solve the problem of finding the MLE in the completion of the LCM we found, and that is done as we did before. And so forth. The iteration must terminate because each LCM has smaller dimension than the one before. Geyer [12, Chapter 2] gives details.

### 6.2.3 Third characterization

Geyer [8, Section 3] shows that the recursion in the preceding section can be avoided by use of a *generic direction of recession* (GDOR), which is a DOR in the relative interior of the set of all DOR.

## 6.3 Calculating limiting conditional models

### 6.3.1 Based on the first characterization

We do not need a DOR because we only use that to determine the affine support of the LCM, and we can estimate that by other methods. Suppose $\eta_1, \ldots, \eta_j$ and other notation are as in

Section 6.2.1 above. The LCM is the OM conditioned on the event

$$\langle Y, M\eta_i \rangle = \langle y, M\eta_i \rangle, \qquad \text{almost surely for } i \in 1, \ldots, j. \tag{17}$$

We have no readily available way to fit a GLM subject to (17).

We know, however, (Section 6.2.2 above) that the event (17) fixes some components of the response vector at their observed values and leaves the rest entirely unconstrained. Those components, that are entirely unconstrained are those for which the corresponding component of $M\eta_i$ is zero (or, taking account of the inexactness of computer arithmetic, nearly zero) for all $i = 1$, ..., $j$.

### 6.3.2 Based on the second characterizations

We can find a DOR by minimizing the function $f$ defined by

$$f(a) = \max_{i \in 1, \ldots, m} \sum_{k=1}^{j} a_k \langle v_i, \eta_k \rangle, \tag{18}$$

where $v_1$, ..., $v_m$ are vectors that generate the tangent cone for the GLM, which are defined in Geyer [8, Sections 3.10 and 3.11] and for which R code to calculating them is given in the technical reports accompanying [8], and where $\eta_1$, ..., $\eta_j$ are as in Section 6.2.3 above.

We minimize $f$ over all unit vectors $a$, to avoid unbounded domain (if we minimized over all vectors, the optimal value might be $-\infty$, and no solution would exist) and also to avoid the zero vector being a solution.

If $\bar{a}_1$, ..., $\bar{a}_j$ are the components of the solution, the DOR is

$$\delta = \sum_{k=1}^{j} \bar{a}_k \eta_k,$$

and the LCM is the OM conditioned on the event that every component of the response vector for which the corresponding component of $\zeta = M\delta$ is nonzero is constrained to be equal to its observed value.

We fit the model using the `subset` argument of R function `glm` as explained in Section 6.2.2 above.

We can use ideas from Section 6.2.1 to tell us whether we need to iterate. We already know the dimension $j$ of the constancy space of the MLE in the completion. If the LCM determined by this GDOR has a constancy space of this dimension, then we have the correct LCM and do not need to iterate. R function `glm` will figure out the dimension of the constancy space (how many coefficients it needs to drop to get an identifiable model) on its own.

### 6.3.3 Based on the third characterization

As far as we know, there is no way to calculate a GDOR except by using the very time consuming computational geometry calculations explained by [8].

# 7 Examples

## 7.1 Complete separation example

We return to the motivating example of Section 2. Here we see that the Fisher information matrix has only null eigenvectors. Thus the LCM is completely degenerate at the one point set containing

Table 1: The estimated null eigenvector of the Fisher information matrix (column 2) and the gdor computed by [8] (column 3). Only nonzero components are shown.

| coefficient | $\hat{\eta}$ | $\hat{\eta}_{\text{gdor}}$ |
|---|---|---|
| intercept | -1 | -1 |
| $v1$ | 1 | 1 |
| $v2$ | 1 | 1 |
| $v3$ | 1 | 1 |
| $v5$ | 1 | 1 |
| $v1 : v2$ | -1 | -1 |
| $v1 : v3$ | -1 | -1 |
| $v1 : v5$ | -1 | -1 |
| $v2 : v3$ | -1 | -1 |
| $v2 : v5$ | -1 | -1 |
| $v3 : v5$ | -1 | -1 |
| $v1 : v2 : v3$ | 1 | 1 |
| $v1 : v3 : v5$ | 1 | 1 |
| $v2 : v3 : v5$ | 1 | 1 |

only the observed value of the canonical statistic of this exponential family.

We adopt the techniques of Section 3.16.2 of [8] to make inferences about mean-value parameters (success probability considered as a function of the predictor $x$). This is outlined in Section 2. One-sided confidence intervals are seen in Figure 2. As stated in Section 2, the actual computations follow some later course notes [9].

### 7.2 Example in Section 2.3 of [8]

This example consists of a $2 \times 2 \times \cdots \times 2$ contingency table with seven dimensions hence $2^7 = 128$ cells. These data now have a permanent location [6]. There is one response variable $y$ that gives the cell counts and seven categorical predictors $v_1$, …, $v_7$ that specify the cells of the contingency table. We fit a generalized linear regression model where $y$ is taken to be Poisson distributed. We consider a model with all three-way interactions included but no higher-order terms. Geyer [8] shows the MLE in this example does not exist in the traditional sense, and then computes a generic direction of recession using the repeated linear programming with R package `rcdd` (Section 6.3.3). In this example there is only single null eigenvector of the Fisher information matrix, which consequently must be a generic direction of recession. Therefore all of our methods of determining the support of the LCM in Sections 6.3.1, 6.3.2, and 6.3.3 must do the same thing.

Table 1 displays the comparison between the characterizations in Sections 6.3.2 and 6.3.3. The vector $\hat{\eta}$ is the estimated null eigenvector of the Fisher information matrix using our implementation. The vector $\hat{\eta}_{\text{gdor}}$ is the estimated gdor in [8]. The $\hat{\eta}$ vector is identical to $\hat{\eta}_{\text{gdor}}$ up to six decimal places (the results in Table 1 are rounded). Therefore, the inferences resulting from these two distinct approaches is identical up to rounding. The only material difference between our implementation and the linear programming in [8] is computational time. Our implementation estimates $\eta$ in 0.017 seconds of computer time, while the functions in the `rcdd` package estimates $\hat{\eta}_{\text{gdor}}$ in 4.481 seconds of computer time. This is a big difference for a relatively small amount of data.

Inference for the MLE in the LCM are included in the supplementary materials.

### 7.3 Big data example

This example uses the other dataset at [6]. It shows our methods are much faster than the linear programming method of [8] for recovering directions of recession (Sections 6.2.2 and 6.2.3). The characterization in Section 6.2.1 is even faster since no direction of recession is computed. This dataset consists of five categorical variables with four levels each and a response variable $y$ that is Poisson distributed. A model with all four-way interaction terms is fit to this data. It may seem that the four way interaction model is too large (1024 data points vs 781 parameters) but $\chi^2$ tests select this model over simpler models, see Table 2.

Table 2: Model comparisons for Example 2. The model m1 is the main-effects only model, m2 is the model with all two way interactions, m3 is the model with all three way interactions, and m4 is the model with all four way interactions.

| null model | alternative model | df | Deviance | $\Pr(>\chi^2)$ |
|---|---|---|---|---|
| m1 | m4 | 765 | 904.8 | 0.00034 |
| m2 | m4 | 675 | 799.2 | 0.00066 |
| m3 | m4 | 405 | 534.4 | 0.00002 |

We estimate that the dimension of the null space of the estimated Fisher information matrix is 23. In the Section 6.3.2 characterization we minimize $f$ over $a \in \mathbb{R}^{23}$ in (18), $\|a\| = 1$ to find a DOR. The resulting vector $\hat{\eta}_{\text{gdor}} = \sum_{k=1}^{23} a_k \hat{\eta}_k$ is a GDOR since it satisfies conditions (20a) and (20b) of [8]. Fitting the model, estimating the dimension of the null space of estimated Fisher information, finding $a$, and estimating the support of the LCM took less than 2 seconds of computer time. In the Section 6.3.3 characterization, the functions in the `rcdd` package perform the same tasks in 334701 seconds (roughly 3.8 days) of computer time. The two different methods estimated different GDORs but they estimate the same support for the LCM.

Inferences for the MLE in the LCM are included in the supplementary materials. One-sided 95% confidence intervals for mean-value parameters that correspond to components of the canonical statistic which are on the boundary of their support (MLE equal to 0) are also included in the supplementary materials. We provide a new method for calculating these intervals that has not been previously published, but whose concept is found in Geyer [8] in the penultimate paragraph of Section 3.16.2.

Let $I$ denote the index set of the components of the response vector on which we condition the OM to get the LCM, and let $Y_I$ and $y_I$ denote these components considered as a random vector and as an observed value, respectively. Let $\theta = M\beta$ denote the saturated model canonical parameter (usually called "linear predictor" in GLM theory) with $\beta$ being the submodel canonical parameter vector. Then endpoints for a $100(1-\alpha)\%$ confidence interval for a scalar parameter $g(\beta)$ are

$$\min_{\substack{\gamma \in \Gamma_{\text{lim}} \\ \text{pr}_{\hat{\beta}+\gamma}(Y_I = y_I) \geq \alpha}} g(\hat{\beta} + \gamma) \qquad \text{and} \qquad \max_{\substack{\gamma \in \Gamma_{\text{lim}} \\ \text{pr}_{\hat{\beta}+\gamma}(Y_I = y_I) \geq \alpha}} g(\hat{\beta} + \gamma) \tag{19}$$

where $\Gamma_{\text{lim}}$ is a basis for the null space of Fisher information. At least one of (19) is at the end of the range of this parameter (otherwise we can use conventional two-sided intervals). For Poisson sampling, let $\mu = \exp(\theta)$ denote the mean value parameter (here exp operates componentwise like the R function of the same name does), then $\text{pr}_\beta(Y_I = y_I) = \exp\left(-\sum_{i \in I} \mu_i\right)$. We take the

Table 3: One-sided 95% confidence intervals for 5 out of 82 mean-valued parameters whose MLE is equal to 0.

| X1 | X2 | X3 | X4 | X5 | lower bound | upper bound |
|---|---|---|---|---|---|---|
| b | c | c | b | a | 0 | 0.60 |
| c | c | c | b | a | 0 | 2.28 |
| d | c | c | b | a | 0 | 1.47 |
| d | d | c | b | a | 0 | 2.99 |
| a | c | d | b | a | 0 | 0.02 |

confidence interval problem to be

$$\begin{aligned} &\text{maximize} && \mu_k \\ &\text{subject to} && -\sum_{i \in I} \mu_i \geq \log(\alpha) \end{aligned} \tag{20}$$

where $\mu$ is taken to be the function of $\gamma$ described in (19). The optimization in (20) can be done for any $k \in I$. Implementation details are included in Sections 10.7.1 and 10.7.2 in the supplementary materials.

One-sided 95% confidence intervals for mean-valued parameters whose MLE is equal to 0 are displayed in Table 3. The full table is included in the supplementary materials. Some of the intervals in Table 3 are relatively wide which represents non-trivial uncertainty about the observed MLE being 0.

# 8 Discussion

The chance of observing a canonical statistic on the boundary of its support increases when the dimension of the model increases. Researchers naturally want to include all possibly relevant covariates in an analysis, and this will often result in the MLE not existing in the conventional sense. Our methods provide a computationally inexpensive solution to this problem.

The theory of generalized affine functions and the geometry of exponential families allows GLM software to provide a MLE when the observed value of the canonical statistic is on the boundary of its support. In such settings, the MLE does not exist in the traditional sense and is said to belong to the Barndorff-Nielsen completion of the exponential family [3, 4, 8, 5] when the supremum of the log likelihood is finite. [3, 4, 5] all provided a MLE when it exists in the Barndorff-Nielsen completion of the family and [8] provided estimates of variability under the conditions of [4]. We do the same here using the theory of generalized affine functions.

The limiting distribution evaluated along the iterates of such an optimization is a generalized exponential family taking the form of a generalized affine function with structure given by Theorem 5. Cumulant generating functions converge along this sequence of iterates (Theorems 6 and 7) as well as estimates of moments of all orders (Theorem 10) for distributions taking estimated parameter values along this sequence of iterates. We can then use the null eigenvectors of estimated Fisher information to find a DOR and the support of a LCM. Parameter estimation in the LCM is conducted in the traditional manner using GLM software. One-sided confidence intervals for mean-value and canonical parameters that are observed to be on the boundary can also be provided.

The costs of computing a DOR and the support of a LCM are minimal compared to the repeated linear programming in the `rcdd` package, especially when the dimension of the data is large. This is where the desirability of our approach stems from. It is much faster to let optimization software, such as `glm` in R, simply go uphill on the log likelihood of the exponential family until a convergence tolerance is reached. Our examples show what kind of time saving is possible using our methods on small and large datasets.

# A Technical appendix

*Proof of Theorem 6.* First consider the case when $j = 0$, the sequences of $\eta$ vectors and scalars $\delta$ are both of length zero. There are no sets $C^+$ and $C^-$ in this setting and $h$ is affine on $E$. From Lemma 1 we have $\psi_n = \theta_n$. From Corollary 2, $\theta_n \to \theta^*$ as $n \to \infty$. We observe that $c(\theta_n) \to c(\theta^*)$ from continuity of the cumulant function. The existence of the MLE in this setting implies that there is a neighborhood about 0 denoted by $W$ such that $\theta^* + W \subset \text{int}(\text{dom}\, c)$. Pick $t \in W$ and observe that $c(\theta_n + t) \to c(\theta^* + t)$. Therefore $\kappa_n(t) \to \kappa(t)$ when $j = 0$.

Now consider the case when $j = 1$. Define $c_1(\theta) = \log \int_{H_1} e^{\langle y,\theta\rangle} \lambda(dy)$ for all $\theta \in \text{int}(\text{dom } c_1)$. In this scenario we have

$$\begin{aligned}
\kappa_n(t) &= c\left(\psi_n + t + b_{1,n}\eta_1\right) - c\left(\psi_n + b_{1,n}\eta_1\right) \\
&= c\left(\psi_n + t + b_{1,n}\eta_j\right) - c\left(\psi_n + b_{1,n}\eta_1\right) \pm b_{1,n}\delta_1 \\
&= \left[c\left(\psi_n + t + b_{1,n}\eta_1\right) - b_{1,n}\delta_1\right] - \left[c\left(\psi_n + b_{1,n}\eta_1\right) - b_{1,n}\delta_1\right].
\end{aligned}$$

From [12, Theorem 2.2], we know that

$$\begin{aligned}
c\left(\theta^* + t + s\eta_1\right) - s\delta_1 &\to c_1\left(\theta^* + t\right), \\
c\left(\theta^* + s\eta_1\right) - s\delta_1 &\to c_1\left(\theta^*\right),
\end{aligned} \tag{21}$$

as $s \to \infty$ since $\delta_1 \geq \langle y, \eta_1\rangle$ for all $y \in H_1$. The left hand side of (21) is a convex function of $\theta$ and the right hand side is a proper convex function. If int(dom $c_1$) is nonempty, which holds whenever int(dom $c$) is nonempty, then the convergence in (21) is uniform on compact subsets of int(dom $c_1$) [15, Theorem 7.17]. Also [15, Theorem 7.14], uniform convergence on compact sets is the same as continuous convergence. Using continuous convergence, we have that both

$$\begin{aligned}
c\left(\psi_n + t + b_{1,n}\eta_1\right) - b_{1,n}\delta_1 &\to c_1\left(\theta^* + t\right), \\
c\left(\psi_n + b_{1,n}\eta_1\right) - b_{1,n}\delta_1 &\to c_1\left(\theta^*\right),
\end{aligned}$$

where $b_{1,n} \to \infty$ as $n \to \infty$ by Lemma 1. Thus

$$\begin{aligned}
\kappa_n(t) = c(\theta_n + t) - c(\theta_n) &\to c_1\left(\theta^* + t\right) - c_1\left(\theta^*\right) \\
&= \log \int_{H_1} e^{\langle y+t,\theta^*\rangle - c(\theta^*)} \lambda(dy) = \log \int_{H_1} e^{\langle y,t\rangle + h(y)} \lambda(dy) \\
&= \log \int e^{\langle y,t\rangle + h(y)} \lambda(dy) = \kappa(t).
\end{aligned}$$

This concludes the proof when $j = 1$.

For the rest of the proof we will assume that $1 < j \leq p$ where $\dim(E) = p$. Represent the sequence $\theta_n$ in coordinate form as

$$\theta_n = b_{1,n}\eta_1 + b_{2,n}\eta_2 + \cdots + b_{p,n}\eta_p, \tag{22}$$

with scalars $b_{i,n}$, $i = 1, ..., p$. For $0 < j < p$, we know that $\psi_n \to \theta^*$ as $n \to \infty$ from Corollary 2. The existence of the MLE in this setting implies that there is a neighborhood about 0, denoted by $W$, such that $\theta^* + W \subset \mathrm{int}(\mathrm{dom}\, c)$. Pick $t \in W$, fix $\varepsilon > 0$, and construct $\varepsilon$-boxes about $\theta^*$ and $\theta^* + t$, denoted by $\mathcal{N}_{0,\varepsilon}(\theta^*)$ and $\mathcal{N}_{t,\varepsilon}(\theta^*)$ respectively, such that both $\mathcal{N}_{0,\varepsilon}(\theta^*), \mathcal{N}_{t,\varepsilon}(\theta^*) \subset \mathrm{int}\,(\mathrm{dom}\ c)$. Let $V_{t,\varepsilon}$ be the set of vertices of $\mathcal{N}_{t,\varepsilon}(\theta^*)$. For all $y \in E$ define

$$M_{t,\varepsilon}(y) = \max_{v \in V_{t,\varepsilon}} \{\langle v, y\rangle\}, \qquad \widetilde{M}_{t,\varepsilon}(y) = \min_{v \in V_{t,\varepsilon}} \{\langle v, y\rangle\}. \tag{23}$$

From the conclusions of Lemma 1 and Corollary 2, we can pick an integer $N$ such that $\langle y, \psi_n + t\rangle \leq M_{t,\varepsilon}(y)$ and $b_{(i+1),n}/b_{i,n} < 1$ for all $n > N$ and $i = 1, ..., j-1$. For all $y \in F$, we have

$$\begin{aligned} \langle y, \theta_n + t\rangle - \sum_{i=1}^{j} b_{i,n}\delta_i &= \langle y, \psi_n + t\rangle + \sum_{i=1}^{j} b_{i,n}\left(\langle y, \eta_i\rangle - \delta_i\right) \\ &\leq M_{t,\varepsilon}(y) \end{aligned} \tag{24}$$

for all $n > N$. The integrability of $e^{M_{t,\varepsilon}(y)}$ and $e^{\widetilde{M}_{t,\varepsilon}(y)}$ follows from

$$\begin{aligned} \int e^{\widetilde{M}_{t,\varepsilon}(y)}\lambda(dy) \leq \int e^{M_{t,\varepsilon}(y)}\lambda(dy) &= \sum_{v \in V_{t,\varepsilon}} \int_{\{y\colon\, \langle y,v\rangle = M_{t,\varepsilon}(y)\}} e^{\langle y,v\rangle}\lambda(dy) \\ \leq \sum_{v \in V_{t,\varepsilon}} \int e^{\langle y,v\rangle}\lambda(dy) < \infty. \end{aligned}$$

Therefore,

$$\langle y, \psi_n + t\rangle + \sum_{i=1}^{j} b_{i,n}\left(\langle y, \eta_i\rangle - \delta_i\right) \to \left\{ \begin{array}{cc} \langle y, \theta^* + t\rangle, & y \in H_j, \\ -\infty, & y \in F \setminus H_j. \end{array} \right.$$

which implies that

$$c_F(\theta_n + t) - c_F(\theta_n) \to c_{H_j}(\theta^* + t) - c_{H_j}(\theta^*) \tag{25}$$

by dominated convergence. To complete the proof, we need to verify that

$$\begin{aligned} c(\theta_n + t) - c(\theta_n) &= c_F(\theta_n + t) - c_F(\theta_n) \\ &\quad + c_{\cup_{i=1}^{j-1} D_i}(\theta_n + t) - c_{\cup_{i=1}^{j-1} D_i}(\theta_n) \\ &\to c_{H_j}(\theta^* + t) - c_{H_j}(\theta^*). \end{aligned} \tag{26}$$

We know that (26) holds when $\lambda(\cup_{i=1}^{j-1} D_i) = 0$ in (11) because of (25). Now suppose that $\lambda(\cup_{i=1}^{j-1} D_i) > 0$. We have,

$$\langle y, \psi_n + t\rangle + \sum_{i=1}^{j} b_{i,n}\left(\langle y, \eta_i\rangle - \delta_i\right) \to -\infty, \qquad y \in \cup_{i=1}^{j-1} D_i, \tag{27}$$

and

$$
\begin{aligned}
&\exp\left(c_{\cup_{i=1}^{j-1} D_i}(\theta_n + t) - c_{\cup_{i=1}^{j-1} D_i}(\theta_n)\right) \\
&\quad = \int_{\cup_{i=1}^{j-1} D_i} e^{\langle y, \theta_n + t\rangle - c_{\cup_{i=1}^{j-1} D_i}(\theta_n)} \lambda(dy) \\
&\quad \leq \int_{\cup_{i=1}^{j-1} D_i} e^{M_{t,\varepsilon}(y) - \widetilde{M}_{0,\varepsilon}(y) + \langle y, \theta_n\rangle - c_{\cup_{i=1}^{j-1} D_i}(\theta_n)} \lambda(dy) \\
&\quad \leq \sup_{y \in \cup_{i=1}^{j-1} D_i} \left(e^{\langle y, \theta_n\rangle - c_{\cup_{i=1}^{j-1} D_i}(\theta_n)}\right) \lambda\left(\cup_{i=1}^{j-1} D_i\right) \\
&\qquad \times \int_{\cup_{i=1}^{j-1} D_i} e^{M_{t,\varepsilon}(y) - \widetilde{M}_{0,\varepsilon}(y)} \lambda(dy) \\
&\quad \leq \sup_{\theta \in \Theta} \sup_{y \in \cup_{i=1}^{j-1} D_i} \left(e^{\langle y, \theta\rangle - c_{\cup_{i=1}^{j-1} D_i}(\theta)}\right) \lambda\left(\cup_{i=1}^{j-1} D_i\right) \\
&\qquad \times \int_{\cup_{i=1}^{j-1} D_i} e^{M_{t,\varepsilon}(y) - \widetilde{M}_{0,\varepsilon}(y)} \lambda(dy) \ < \ \infty
\end{aligned}
\tag{28}
$$

for all $n > N$ by the assumption given by (11). The assumption that the exponential family is discrete and full implies that $\int e^{h}(y)\lambda(dy) = 1$ [12, Theorem 2.7]. This in turn implies that $\lambda(C_i^+) = 0$ for all $i = 1, ..., j$ which then implies that $c(\theta) = c_F(\theta) + c_{\cup_{i=1}^{j-1} D_i}(\theta)$. Putting (24), (27), and (28) together we can conclude that (26) holds as $n \to \infty$ by dominated convergence and

$$
\begin{aligned}
&c_{H_j}(\theta^* + t) - c_{H_j}(\theta^*) \\
&\quad = \log \int_{H_j} e^{\langle y, \theta^* + t\rangle} \lambda(dy) - \log \int_{H_j} e^{\langle y, \theta^*\rangle} \lambda(dy) \\
&\quad = \log \int e^{\langle y, t\rangle + h(y)} \lambda(dy) = \kappa(t).
\end{aligned}
\tag{29}
$$

for all $t \in W$. This verifies CGF convergence on neighborhoods of 0 which completes the proof. □

*Proof of Theorem 7.* Represent $h$ as in Theorem 5. Denote the normal cone of the convex polyhedron support $K$ at the data $x$ by $N_K(x)$. We show that a sequence of scalars $\delta_i^*$ and a linearly independent set of vectors $\eta_i^* \in E^*$ can be chosen so that $\eta_i^* \in N_K(x)$, and

$$
\begin{aligned}
H_i &= \{y \in H_{i-1} : \langle y, \eta_i^*\rangle = \delta_i^*\}, \\
C_i^+ &= \{y \in H_{i-1} : \langle y, \eta_i^*\rangle > \delta_i^*\}, \\
C_i^- &= \{y \in H_{i-1} : \langle y, \eta_i^*\rangle < \delta_i^*\},
\end{aligned}
\tag{30}
$$

for $i = 1, ..., j$ where $H_0 = E$ so that (11) holds. We will prove this by induction with the hypothesis $\mathrm{H}(m)$, $m = 1, ..., j$, that (30) holds for $i \leq m$ where the vectors $\eta_i^* \in N_K(x)$ $i = 1, ..., m$.

We first verify the basis of the induction. The assumption that the exponential family is discrete and full implies that $\int e^{h}(y)\lambda(dy) = 1$ [12, Theorem 2.7]. This in turn implies that $\lambda(C_k^+) = 0$ for all $k = 1, ..., j$. This then implies that $K \subseteq \{y \in E : \langle y, \eta_1\rangle \leq \delta_1\} = H_1 \cup C_1^-$. Thus $\eta_1 \in N_K(x)$ and the base of the induction holds with $\eta_1 = \eta_1^*$ and $\delta_1 = \delta_1^*$.

We now show that $\mathrm{H}(m+1)$ follows from $\mathrm{H}(m)$ for $m = 1, ..., j-1$. We first establish that $K \cap H_m$ is an exposed face of $K$. This is needed to show that (30) holds for $i = 1, ..., m+1$. Let $L_K$ be the collection of closed line segments with endpoints in $K$. Arbitrarily choose $l \in L_K$ such that an interior point $y \in l$ is such that $y \in K \cap H_m$. We can write $y = \gamma a + (1-\gamma)b$, $0 < \gamma < 1$, where $a$ and $b$ are the endpoints of $l$. Since $a, b \in K$ by construction, we have that $\langle a - x, \eta_m^* \rangle \le 0$ and $\langle b - x, \eta_m^* \rangle \le 0$ because $\eta_m^* \in N_K(x)$ by H($m$). Now,

$$\begin{aligned} 0 &\ge \langle a - x, \eta_m^* \rangle = \langle a - y + y - x, \eta_m^* \rangle \\ &= \langle a - y, \eta_m^* \rangle = \langle a - (\gamma a + (1-\gamma)b), \eta_m^* \rangle \\ &= (1-\gamma)\langle a - b, \eta_m^* \rangle \end{aligned}$$

and

$$\begin{aligned} 0 &\ge \langle b - x, \eta_m^* \rangle = \langle b - y + y - x, \eta_m^* \rangle \\ &= \langle b - y, \eta_m^* \rangle = \langle b - (\gamma a + (1-\gamma)b), \eta_m^* \rangle \\ &= -\gamma\langle a - b, \eta_m^* \rangle. \end{aligned}$$

Therefore $a, b \in K \cap H_m$ and this verifies that $K \cap H_m$ is a face of $K$ since $l$ was chosen arbitrarily. The function $y \mapsto \langle y - x, \eta_m^* \rangle - \delta_m^*$, defined on $K$, is maximized over $K \cap H_m$. Therefore $K \cap H_m$ is an exposed face of $K$ by definition. The exposed face $K \cap H_m = K \cap (H_{m+1} \cup C_{m+1}^-)$ since $\lambda(C_{m+1}^+) = 0$ and the convex support of the measure $\lambda | H_m$ is $H_m$ by assumption. Thus, $\eta_{m+1} \in N_{K \cap H_m}(x)$.

The sets $K$ and $H_m$ are both convex and are therefore regular at every point [15, Theorem 6.20]. We can write $N_{K \cap H_m}(x) = N_K(x) + N_{H_m}(x)$ since $K$ and $H_m$ are convex sets that cannot be separated where $+$ denotes Minkowski addition in this case [15, Theorem 6.42]. The normal cone $N_{H_m}(x)$ has the form

$$\begin{aligned} N_{H_m}(x) &= \{\eta \in E^* : \langle y - x, \eta \rangle \le 0 \text{ for all } y \in H_m\} \\ &= \{\eta \in E^* : \langle y - x, \eta \rangle \le 0 \text{ for all } y \in E \\ &\qquad \text{such that } \langle y - x, \eta_i \rangle = 0,\ i = 1, ..., m\} \\ &= \left\{ \sum_{i=1}^{m} a_i \eta_i :\ a_i \in \mathbb{R},\ i = 1, ..., m \right\}. \end{aligned}$$

Therefore, we can write

$$\eta_{m+1} = \eta_{m+1}^* + \sum_{i=1}^{m} a_{m,i} \eta_i^* \tag{31}$$

where $\eta_{m+1}^* \in N_K(x)$ and $a_{m,i} \in \mathbb{R}$, $i = 1, ..., m$. For $y \in H_{m+1}$, we have that

$$\begin{aligned} \langle y, \eta_{m+1}^* \rangle &= \langle y, \eta_{m+1} \rangle - \sum_{i=1}^{m} a_{m,i} \langle y, \eta_i \rangle \\ &= \delta_{m+1} - \sum_{i=1}^{m} a_{m,i} \delta_i. \end{aligned}$$

Let $\delta_{m+1}^* = \delta_{m+1} - \sum_{i=1}^{m} a_{m,i} \delta_i$. We can therefore write

$$H_{m+1} = \left\{ y \in H_m : \langle y, \eta_{m+1}^* \rangle = \delta_{m+1}^* \right\}$$

and

$$
\begin{aligned}
C^+_{m+1} &= \{y \in H_m : \langle y, \eta_{m+1}\rangle > \delta_{m+1}\} \\
&= \left\{ y \in H_m : \langle y, \eta^*_{m+1}\rangle + \sum_{i=1}^{m} a_{m,i}\delta_i > \delta_{m+1} \right\} \\
&= \left\{ y \in H_m : \langle y, \eta^*_{m+1}\rangle > \delta_{m+1} - \sum_{i=1}^{m} a_{m,i}\delta_i \right\} \\
&= \left\{ y \in H_m : \langle y, \eta^*_{m+1}\rangle > \delta^*_{m+1} \right\}.
\end{aligned} \tag{32}
$$

A similar argument to that of (32) verifies that

$$
C^-_i = \left\{ y \in H_m : \langle y, \eta^*_{m+1}\rangle < \delta^*_{m+1} \right\}.
$$

This confirms that (30) holds for $i = 1, ...m + 1$ and this establishes that H$(m + 1)$ follows from H$(m)$.

Define the sets $D_i$ in (10) with starred quantities replacing the unstarred quantities. Since the vectors $\eta^*_1, ..., \eta^*_j \in N_K(x)$, the sets $K \cap D_i$ are all empty for all $i = 1, ..., j - 1$. Therefore (11) holds with $\lambda\left(\cup_{i=1}^{j-1} D_i\right) = 0$. This completes the proof. □

*Proof of Theorem 11.* We first consider the case that $A$ is positive definite and $V = \{0\}$. We can write $A_n = A + (A_n - A)$ where $(A_n - A)$ is a perturbation of $A$ for large $n$. From Weyl's inequality [17], we have that all eigenvalues of $A_n$ are bounded above zero for large $n$ and $V_n = \{0\}$ as a result. Therefore, $V_n \to V$ as $n \to \infty$ when $A$ is positive definite.

Now consider the case that $A$ is not strictly positive definite. Without loss of generality, let $x \in V$ be a unit vector. For all $0 < \gamma \leq \varepsilon$, let $V_n(\gamma)$ denote the subspace spanned by the eigenvectors of $A_n$ corresponding to eigenvalues that are less than $\gamma$. By construction, $V_n(\gamma) \subseteq V_n$.

From [15, Example 10.28], if $A$ has $k$ zero eigenvalues, then for sufficiently large $N_1$ there are exactly $k$ eigenvalues of $A_n$ are less than $\varepsilon$ and $p - k$ eigenvalues of $A_n$ greater than $\varepsilon$ for all $n > N_1$. The same is true with respect to $\gamma$ for all $n$ greater than $N_2$. Thus $j_n(\gamma) = j_n(\varepsilon)$ which implies that $V_n(\gamma) = V_n$ for all $n > \max\{N_1, N_2\}$.

We now verify part (i) of Painlevé-Kuratowski set convergence with respect to $V_n(\gamma)$. Let $N_3$ be such that $x^T A_n x < \gamma^2$ for all $n \geq N_3$. Let $\lambda_{k,n}$ and $e_{k,n}$ be the eigenvalues and eigenvectors of $A_n$, with the eigenvalues listed in decreasing orders. Without loss of generality, we assume that the eigenvectors are orthonormal. Then,

$$
\begin{aligned}
x &= \sum_{k=1}^{p} (x^T e_{k,n}) e_{k,n}, \qquad 1 = \|x\|^2 = \sum_{k=1}^{p} (x^T e_{k,n})^2, \\
x^T A_n x &= \sum_{k=1}^{p} \lambda_{k,n} (x^T e_{k,n})^2.
\end{aligned}
$$

There have to be eigenvectors $e_{k,n}$ such that $x^T e_{k,n} \geq 1/\sqrt{p}$ with corresponding eigenvalues $\lambda_{k,n}$ that are very small since $\lambda_{k,n}(x^T e_{k,n})^2 < \gamma$. But conversely, any eigenvalues $\lambda_{k,n}$ such that $\lambda_{k,n} \geq \gamma$ must have

$$
\lambda_{k,n}(x^T e_{k,n})^2 < \gamma^2 \implies (x^T e_{k,n})^2 < \gamma^2/\lambda_{k,n} \leq \gamma.
$$

Define $j_n(\gamma) = |\{\lambda_{k,n} : \lambda_{k,n} \leq \gamma\}|$ and $x_n = \sum_{k=p-j_n(\gamma)+1}^{p}(x^T e_{k,n})e_{k,n}$ where $x_n \in V_n(\gamma)$ by construction. Now,

$$\begin{aligned}
\|x - x_n\| &= \|\sum_{k=1}^{p}(x^T e_{k,n})e_{k,n} - \sum_{k=p-j_n(\gamma)+1}^{p}(x^T e_{k,n})e_{k,n}\| \\
&= \|\sum_{k=1}^{p-j_n(\gamma)}(x^T e_{k,n})e_{k,n}\| \\
&\leq \sum_{k=1}^{p-j_n(\gamma)}|x^T e_{k,n}| \\
&\leq (p - j_n)\sqrt{\gamma} \\
&\leq p\sqrt{\gamma}
\end{aligned}$$

for all $n \geq N_3$. Therefore, for every $x \in V$, there exists a sequence $x_n \in V_n(\gamma) \subseteq V_n$ such that $x_n \to x$ since this argument holds for all $0 < \gamma \leq \varepsilon$. This establishes part (i) of Painlevé-Kuratowski set convergence.

We now show part (ii) of Painlevé-Kuratowski set convergence. Suppose that $x_n \to x \in \mathbb{R}^p$ and there exists a natural number $N_4$ such that $x_n \in V_n(\gamma)$ whenever $n \geq N_4$, and we will establish that $x \in V$. From hypothesis, we have that $x_n^T A_n x_n \to x^T Ax$. Without loss of generality, we assume that $x$ is a unit vector and that $|x_n^T A_n x_n - x^T Ax| \leq \gamma$ for all $n \geq N_5$. From the assumption that $x_n \in V_n(\gamma)$ we have

$$x_n^T A_n x_n = \sum_{k=1}^{p}\lambda_{k,n}(x_n^T e_{k,n})^2 = \sum_{k=p-j_n(\gamma)+1}^{p}\lambda_{k,n}(x_n^T e_{k,n})^2 \leq \gamma \tag{33}$$

for all $n \geq N_4$. The reverse triangle inequality gives

$$||x_n^T A_n x_n| - |x^T Ax|| \leq |x_n^T A_n x_n - x^T Ax| \leq \gamma$$

and (33) implies $|x^T Ax| \leq 2\gamma$ for all $n \geq \max\{N_4, N_5\}$. Since this argument holds for all $0 < \gamma < \varepsilon$, we can conclude that $x \in V$. This establishes part (ii) of Painlevé-Kuratowski set convergence with respect to $V_n(\gamma)$. Therefore $V_n \to V$ and this completes the proof. □

## Supplementary materials

The supplement to "Computationally efficient likelihood inference in exponential families when the maximum likelihood estimator does not exist" is available upon request. The proofs of Theorems 1-3, Theorem 5, Lemma 1 and Theorems 8-10 in the main text and all of the code producing our examples can be seen in the supplementary materials.

## References

[1] Alan Agresti. *Categorical Data Analysis.* John Wiley & Sons, Hoboken, NJ, third edition, 2013.

[2] A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71:1–10, 1984.

[3] Ole Barndorff-Nielsen. *Information and Exponential Families In Statistical Theory.* John Wiley & Sons, Chichester, 1978.

[4] Lawrence D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics, Hayward, CA, 1986.

[5] Imre Csiszár and František Matúš. Closures of exponential families. *Ann. Probab.*, 33:582–600, 2005. doi: 10.1214/009117904000000766.

[6] Daniel J. Eck and Charles J. Geyer. Two data sets that are examples for an article titled "computationally efficient likelihood inference in exponential families when the maximum likelihood estimator does not exist". `http://hdl.handle.net/11299/197369`.

[7] Charles J. Geyer. Likelihood inference for spatial point processes. In *Stochastic geometry (Toulouse, 1996)*, pages 79–140. Chapman & Hall/CRC, Boca Raton, FL, 1999.

[8] Charles J. Geyer. Likelihood inference in exponential families and directions of recession. *Electron. J. Stat.*, 3:259–289, 2009. doi: 10.1214/08-EJS349.

[9] Charles J. Geyer. Two examples of agresti, 2016. `http://www.stat.umn.edu/geyer/8931expfam/infinity.pdf`, knitr source `http://www.stat.umn.edu/geyer/8931expfam/infinity.Rnw`.

[10] Charles J. Geyer, Stuart Wagenius, and Ruth G. Shaw. Aster models for life history analysis. *Biometrika*, 94:415–426, 2007.

[11] Charles J. Geyer, Glen D. Meeden, and Komei Fukuda. *R package `rcdd`: Computational Geometry, version 1.2*, 2017. `https://CRAN.R-project.org/package=rcdd`.

[12] Charles James Geyer. *Likelihood and Exponential Families.* PhD thesis, University of Washington, 1990. `http://hdl.handle.net/11299/56330`.

[13] Paul R. Halmos. *Finite-Dimensional Vector Spaces.* Springer-Verlag, New York, second edition, 1974. Reprint of 1958 edition published by Van Nostrand.

[14] R. Tyrrell Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, NJ, 1970.

[15] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis.* Springer-Verlag, Berlin, 1998. doi: 10.1007/978-3-642-02431-3. Corrected printings contain extensive changes. We used the third corrected printing, 2010.

[16] Lynn Arthur Steen and J. Arthur Seebach, Jr. *Counterexamples in topology.* Springer-Verlag, New York, second edition, 1978.

[17] Hermann Weyl. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Mathematische Annalen*, 71:441–479, 1912.