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

Daniel J. Eck[1] and Charles J. Geyer[2]

1. Department of Statistics, University of Illinois Urbana-Champaign
2. Department of Statistics, University of Minnesota

January 6, 2020

Abstract

In a regular full exponential family, the maximum likelihood estimator (MLE) need not exist in the traditional sense. However, the MLE may exist in the completion of the exponential family. Existing algorithms for finding the MLE in the 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 completion of the exponential family. This methodology is based on conventional maximum likelihood computations which come close, in a sense, to finding the MLE in the completion of the exponential family. These conventional computations construct a likelihood maximizing sequence of canonical parameter values which goes uphill on the likelihood function until they meet a convergence criteria. Nonexistence of the MLE in this context results from a degeneracy of the canonical statistic of the exponential family, the canonical statistic is on the boundary of its support. There is a correspondance between this boundary and the null eigenvectors of the Fisher information matrix. Convergence of Fisher information along a likelihood maximizing sequences follows from cumulant generating function (CGF) convergence along a likelihood maximizing sequence, conditions for which are given. This allows for the construction of necessarily one-sided confidence intervals for mean value parameters when the MLE exists in the completion. We demonstrate our methodology on three examples in the main text and three additional examples in the supplementary materials. We show that when the MLE exists in the completion of the exponential family, our methodology provides statistical inference that is much faster than existing techniques.

Keywords: Completion of exponential families; Convergence of moments; Moment generating function; Complete separation; Logistic regression; Generalized linear models

1 Introduction

In a regular full discrete exponential family, the MLE for the canonical parameter does not exist when when the observed value of the canonical statistic lies on the boundary of its convex support [Barndorff-Nielsen, 1978, Theorem 9.13], but the MLE does exist in a completion of the exponential family. Completions for exponential families have been described by Barndorff-Nielsen [1978, pp. 154–156], Brown [1986, pp. 191–201], Csiszár and Matúš [2005, 2008], and Geyer [1990, unpublished PhD thesis, Chapter 4]. The issue of when the MLE exists in the conventional sense and what to do when it does not is very important because of the wide use of generalized linear models (GLMs) for discrete data and log-linear models for categorical data.

Nonexistence of the MLE in these contexts is a widely studied problem. Advances have been made in establishing necessary and sufficient conditions for existence of the MLE [Haberman, 1974, Aickin, 1979, Eriksson et al., 2006, Fienberg and Rinaldo, 2012], the development of an extended or generalized MLE when the traditional MLE does not exist through convex cores of measures [Csiszár and Matúš, 2001, 2005, 2003, 2008] and through geometric properties of exponential families and log-linear models [Barndorff-Nielsen, 1978, Brown, 1986, Geyer, 1990, Verbeek, 1992, Geyer, 2009, Fienberg and Rinaldo, 2012, Matúš, 2015,

Wang et al., 2019]. The issue of nonexistence also arises in exponential families for spatial lattice processes [Geyer, 1991, Geyer and Thompson, 1992], spatial point processes [Geyer and Møller, 1994, Geyer, 1999], and random graphs [Handcock et al., 2018, Hunter et al., 2008, Rinaldo et al., 2009, Schweinberger, 2011]. In every application of these, existing statistical software gives completely invalid results when the MLE does not exist in the traditional sense, and such software either does not check for this problem or does weak checks that can emit both false positives and false negatives. Moreover, even if these checks correctly detect the nonexistence of the MLE, conventional software implements no valid inferential method in this setting. Authoritative textbooks [Agresti, 2013, Section 6.5] discuss the issue but provide no solutions.

Geyer [2009] developed methodology for constructing hypothesis tests and confidence intervals when the MLE in an exponential family does not exist in the traditional sense. The algorithms in Geyer [2009], implemented in the `rcdd` R package [Geyer et al., 2017], are based on doing many linear programs. This algorithm does at most n linear programs, where n is the number of cases of a GLM or the number of cells in a contingency table, in order to determine the existence of the MLE in the traditional 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`), 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. Previous theoretical discussions [Barndorff-Nielsen, 1978, Brown, 1986, Csiszár and Matúš, 2005, 2008, Fienberg and Rinaldo, 2012, Matúš, 2015, Wang et al., 2019] of these issues do not provide algorithms, use the notions of faces of convex sets or convex core of measure, are specific to particular discrete exponential families, or are all much harder to compute than the algorithm of Geyer [2009]. Therefore they provide no explicit direction toward efficient computing. Thus a valid appropriate solution to this issue that is efficiently computable would be very important.

In this paper, we provide a computationally efficient solution that has is origins with conventional maximum likelihood computations. These computations come close, in a sense, to finding the MLE in the completion of the exponential family. Informally our approach is to first consider a likelihood maximizing sequence of canonical parameter estimates that goes uphill on the likelihood function until a convergence criteria is satisfied. At this point, canonical parameter estimates are still infinitely far away from the MLE in the completion, but mean value parameter estimates are close to the MLE in the completion, and the corresponding probability distributions are close in total variation norm to the MLE probability distribution in the completion. We show that probability distributions are close in the sense of moment generating function convergence (Theorems 6 and 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 [Geyer, 1990, Theorem 4.3]. Valid statistical inference when the MLE does not exist in the conventional sense requires knowledge of this affine hull. This affine hull is a 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 is the variance-covariance matrix of the canonical statistic for an exponential family. 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 conduct valid statistical inference, and the MLE will give us good approximations of these quantities. We will estimate the correct affine hull from the null space of the estimated Fisher information matrix.

Our methodology is implemented in the R package `glmdr` [Geyer and Eck, 2016]. In the main text and the supplementary materials, we demonstrate the performance of our methodology on several extensive didactic examples. These include complete and quasi-complete separation examples in logistic regression, Poisson regression, and Bradley-Terry models. Computational efficiency of our methodology is illustrated in Section 5.3. Our supplementary materials reproduce all of the analyses in this paper.

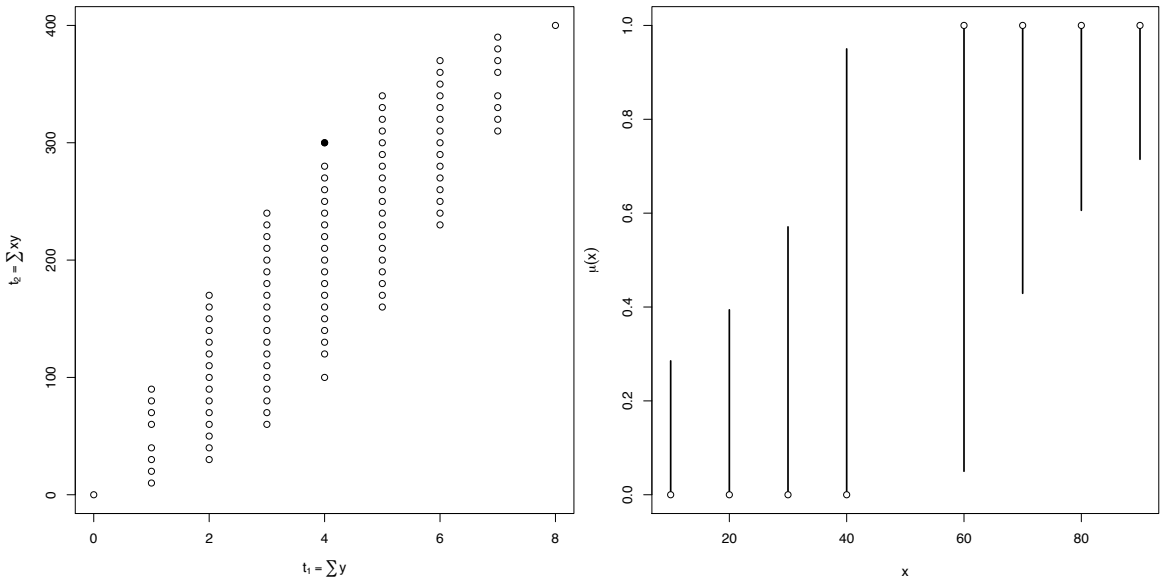

Figure 1: **Left panel**: Observed value and support of the submodel canonical statistic vector $M^T y$ for the example of Section 2. Solid dot is the observed value of this statistic. **Right panel**: One-sided 95% confidence intervals for saturated model mean value parameters. Bars are the intervals; $\mu(x)$ is the probability of observing response value one when the predictor value is x. Solid dots are the observed data.

2 Motivating example

Consider the case of complete separation in the logistic regression model as a motivating example. When perfect separation occurs, the canonical statistic is observed to be on the boundary of its convex support. 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 traditional sense. This example is the same as that of Agresti [2013, Section 6.5.1].

For an exponential family, the submodel canonical statistic is $M^T y$, where M is the model matrix. The left panel of 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 traditional sense. In this example, the MLE in the completion corresponds to a completely degenerate distribution. This MLE distribution says no data other than what was observed could have been observed. But the sample is not the population and estimates are not parameters. Therefore, this degeneracy is not a problem. To illustrate the uncertainty of estimation, we show confidence intervals (necessarily one-sided) for the saturated model mean value parameters. These one-sided confidence intervals are obtained from functionality in the accompanying `glmdr` package.

The right panel of Figure 1 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. We discuss how these intervals are constructed in Section 4.4. The idea is to first find all canonical parameter values such that the probability of observing the realized data, conditional on the degeneracy, is greater than some error tolerance. We then map those canonical parameter values to the mean value parameterization. The degeneracy of this example follows from the estimated Fisher information matrix (for the saturated model canonical parameter vector, also called the linear predictor) at the apparent MLE being the zero matrix, which it is to within the accuracy of computer arithmetic. In this case the MLE

of the saturated model mean value parameters agree with the observed data; they are on the boundary of the set of possible values, either zero or one.

In other examples, such as examples 5.2 and 5.3 below, the MLE distribution is only partially but not completely degenerate. This follows from the estimated Fisher information matrix being singular (to within the accuracy of computer arithmetic) but not the zero matrix. The MLE of some of the saturated model mean value parameters agree with the observed data, but not all. The MLE distribution constrains some components of the response vector to be equal to their observed values, but not all of them. To find the MLE in the completion of the exponential family we need to fit what Geyer [2009] calls the limiting conditional model (LCM), which can be fit using standard methods after it has been identified. This is explained in Sections 4.2 and 4.3.

The methodology that we develop is applicable for any 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 Geyer [2009]. We redo Example 2.3 of Geyer [2009] in Section 5.2 using the methodology developed here, and we find that our methodology produces the inferences in that paper in a fraction of the time, as seen in the supplement. We also provide an analysis on a big data set (too large for the methods of Geyer [2009] to run in an acceptable amount of time) to show how (relatively) quick our implementation is.

3 Standard exponential families

Let λ be a positive Borel measure on a finite-dimensional vector space E. The *log Laplace transform* of λ 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 [Rockafellar and Wets, 1998, 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) [Geyer, 1990, 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 λ. The set $\mathcal{F} = \{ \, f_\theta : \theta \in \Theta \, \}$, where Θ is any nonempty subset of $\operatorname{dom} c$, is called a *standard exponential family of densities with respect to* λ. This family is *full* if $\Theta = \operatorname{dom} c$. We also say $\mathcal{F}$ is the standard exponential family *generated by* λ having canonical parameter space Θ, and λ is the *generating measure* of $\mathcal{F}$.

The log likelihood of this family having densities (2) is

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

A general exponential family [Geyer, 1990, 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 [Geyer, 1990, Section 1.2].

In the context of general exponential families X is called the *canonical statistic* and θ the *canonical parameter* (the terms *natural statistic* and *natural parameter* are also used). The set Θ 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^*.

4 Calculating the MLE in the completion

4.1 Assumptions

So far everything has been for general exponential families. Our implementation requires that the conditions of Brown [1986] hold, and those conditions hold for logistic and log-linear models for categorical data analysis. Now, 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 multinomial sampling [Agresti, 2013, Section 8.6.7; Geyer, 2009, Section 3.17].

The conditions of Brown that are required for our theory to hold are from Brown [1986, 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 λ is the generating measure for the exponential family and $\lambda|F$ is the restriction of λ to the exposed face F.

We let θ_n be a likelihood maximizing sequence of canonical parameter vectors, that is,

$$l_x(\theta_n) \to \sup_{\theta \in \Theta} l_x(\theta), \qquad \text{as } n \to \infty, \tag{4}$$

where the log likelihood l is given by (3), Θ is the canonical parameter space of the family, and $\sup_{\theta \in \Theta} l_x(\theta) < \infty$. Define $h_n(x) = l_x(\theta_n)$. Note that h_n is a sequence of affine functions which are also log densities in a standard exponential family with respect to generating measure λ. Furthermore, let $\lim_{n \to \infty} h_n = h$ where the *generalized affine functions* h form the completion of the exponential family. The limiting density e^h corresponds to the MLE distribution in the completion. The mathematical properties of generalized affine functions and this completion construction are studied in Section 6.

4.2 The form of the MLE in the completion

Suppose we know the *affine support* of the MLE distribution in the completion. This is the smallest affine set (translate of a vector subspace) that contains the canonical statistic with probability one. Denote the affine support by A. 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, we have $A = M^T y + V$ for some vector space V.

Then the limiting conditional model (LCM) in which the MLE in the completion is found is the original model (OM) conditioned on the event

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

[Geyer, 1990, 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. Our theory states that the null space of the Fisher information matrix for the LCM is well approximated by the Fisher information matrix for the OM at parameter values that are close to maximizing the likelihood, see Section 6.4. The vector subspace spanned by the vectors $\eta_1, \ldots, \eta_j$ is called the *constancy space* of the LCM [Geyer, 2009].

4.3 Calculating limiting conditional models

Suppose $\eta_1, \ldots, \eta_j$ and other notation are as in Section 4.2 above. The LCM is the OM conditioned on the event

$$\langle Y, M\eta_i \rangle = \langle y, M\eta_i \rangle, \qquad \text{almost surely for } i \in 1, \ldots, j. \tag{5}$$

The event (5) 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 $i = 1, \ldots, j$.

4.4 Calculating one-sided confidence intervals for mean value parameters

We provide a new method for calculating these one-sided confidence intervals that has not been previously published, but whose concept is found in Geyer [2009] 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 β being the submodel canonical parameter vector. Then endpoints for a $100(1 - \alpha)\%$ confidence interval for a scalar parameter $g(\beta)$ are

$$\min_{\substack{\gamma \in \Gamma_{\lim} \\ \text{pr}_{\hat\beta+\gamma}(Y_I = y_I) \geq \alpha}} g(\hat\beta + \gamma) \qquad \text{and} \qquad \max_{\substack{\gamma \in \Gamma_{\lim} \\ \text{pr}_{\hat\beta+\gamma}(Y_I = y_I) \geq \alpha}} g(\hat\beta + \gamma) \tag{6}$$

where $\Gamma_{\lim}$ is the null space of the Fisher information matrix. At least one of (6) is at the end of the range of this parameter (otherwise we can use conventional two-sided intervals).

For logistic and binomial regression, let $p = \text{logit}^{-1}(\theta)$ denote the mean value parameter vector (here logit^{-1} operates componentwise). Then, $\text{pr}_\beta(Y_I = y_I) = \prod_{i \in I} p_i^{y_i}(1 - p_i)^{n_i - y_i}$ where the n_i are the binomial sample sizes. In logistic regression we have $n_i = 1$ for all i, but in binomial regression we have $n_i \geq 1$ for all i. We could take the confidence interval problem to be

$$\text{maximize} \quad p_k, \qquad \text{subject to} \quad \prod_{i \in I} p_i^{y_i}(1 - p_i)^{n_i - y_i} \geq \alpha, \tag{7}$$

where p is taken to be the function of γ described above. And this can be done for any $k \in I$. However, the problem will be more computationally stable if we state it as

$$\begin{aligned} &\text{maximize} && \theta_k \\ &\text{subject to} && \sum_{i \in I} \bigl[y_i \log(p_i) + (n_i - y_i) \log(1 - p_i) \bigr] \geq \log(\alpha). \end{aligned} \tag{8}$$

Since $\theta_k = \text{logit}(p_k)$ is a monotone transformation and log is a monotone transformation, the two problems (7) and (8) are equivalent. We maximize canonical rather than mean value parameters to avoid extreme

inexactness of computer arithmetic in calculating mean value parameters near zero and one. We take logs in the constraint for the same reasons we take logs of likelihoods.

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 confidence interval problem to be

$$\text{maximize} \quad \mu_k, \qquad \text{subject to} \quad -\sum_{i \in I} \mu_i \geq \log(\alpha) \tag{9}$$

where μ is taken to be the function of γ described in (6). The optimization in (9) can be done for any $k \in I$.

The `inference` function in the R package `glmdr` determines one-sided confidence intervals for mean value parameters corresponding to response values y_I for logistic and binomial regression as in (8) and Poisson regression as in (9). Further implementation details are included the supplementary materials.

5 Examples

5.1 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 only the observed value of the canonical statistic of this exponential family. One-sided confidence intervals for mean value parameters (success probability considered as a function of the predictor x) are computed as in Section 4.4. The right panel of Figure 1 in Section 2 displays these one-sided intervals.

This example is reproduced in the supplement. The functionality in `glmdr` was used to calculate the one-sided confidence intervals for mean value parameters (`inference` function) and determine that the LCM is completely degenerate (`glmdr` function).

5.2 Example in Section 2.3 of Geyer [2009]

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 [Eck and Geyer]. 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. The software in the `glmdr` package reproduces the original analysis, as seen in the supplement. The `inference` function computed the one-sided confidence intervals for mean value parameters that are on the boundary of their support, in this case equal to 0. The results are depicted in Table 1, this table is the same as Table 2 in Geyer [2009] and it is reproduced in the supplement.

The only material difference between our implementation and the linear programming in Geyer [2009] is computational time. Our implementation provides one-sided confidence intervals for those responses that are on the boundary of their support in 1.253 seconds, while the functions in the `rcdd` package take 4.84 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.

5.3 Big data example

This example uses the other dataset at [Eck and Geyer]. It shows our methods are much faster than the linear programming method of Geyer [2009]. The functionality in the `glmdr` determined the LCM and computed one-sided confidence intervals for mean value parameters that are on the boundary of their support in about a minute. The same tasks took over three days using the `rcdd` package. Both methods yielded the same conclusions.

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 χ^2 tests select this model over simpler models, see Table 2.

Table 1: One-sided confidence intervals for cells with MLE equal to 0.

v_1	v_2	v_3	v_4	v_5	v_6	v_7	lower	upper
0	0	0	0	0	0	0	0	0.28631
0	0	0	1	0	0	0	0	0.14083
1	1	0	0	1	0	0	0	0.21997
1	1	0	1	1	0	0	0	0.42096
0	0	0	0	0	1	0	0	0.08946
0	0	0	1	0	1	0	0	0.09377
1	1	0	0	1	1	0	0	0.19302
1	1	0	1	1	1	0	0	0.28870
0	0	0	0	0	0	1	0	0.10631
0	0	0	1	0	0	1	0	0.11415
1	1	0	0	1	0	1	0	0.09129
1	1	0	1	1	0	1	0	0.26461
0	0	0	0	0	1	1	0	0.06669
0	0	0	1	0	1	1	0	0.15478
1	1	0	0	1	1	1	0	0.14097
1	1	0	1	1	1	1	0	0.32392

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

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, this represents non-trivial uncertainty about the observed MLE being 0. This example is reproduced in the supplement.

6 Mathematical details

In this Section we provide the mathematical justification for our inferential procedure. We develop the theory of generalized affine functions [Geyer, 1990] and then show that this theory, combined with conditions for the exponential family closure of Brown [1986], facilitates the convergence of moments of all orders along a sequence of maximum likelihood iterates. We close this Section by establishing that our mathematical technique can estimate the correct null space of the Fisher information matrix, and this allows for valid statistical inference when the MLE does not exist in the conventional sense.

6.1 Generalized affine functions

6.1.1 Characterization on affine spaces

Exponential families defined on affine spaces instead of vector spaces are in many ways more elegant [Geyer, 1990, 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. We

Table 3: One-sided 95% confidence intervals for 6 out of 82 mean-valued parameters whose MLE is equal to 0.

X1	X2	X3	X4	X5	lower bound	upper bound
a	a	b	a	a	0	0.1695
a	b	b	a	a	0	0.1354
a	c	b	a	a	0	0.2292
a	d	b	a	a	0	2.4616
d	d	c	a	a	0	0.0002
a	c	d	a	a	0	0.0133

complete the exponential family by taking pointwise limits of densities, allowing $+\infty$ and $-\infty$ as limits [Geyer, 1990, Chapter 4].

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 [Geyer, 1990, Chapter 4]. (For a definition of extended-real-valued convex functions see Rockafellar [1970, 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 λ, 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.

6.1.2 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 [Steen and Seebach, 1978, p. 22, gives a proof].

The space $G(E)$ is not metrizable, unless E is zero-dimensional [Geyer, 1990, penultimate paragraph of Section 3.3]. So we cannot use δ-ε arguments, but we can use arguments involving sequences, using sequential compactness.

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

$$\int e^h \, d\lambda = 1, \qquad h \in \mathcal{H}. \tag{10}$$

We call $\mathcal{H}$ a *standard generalized exponential family* of log densities with respect to λ. 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. □

For full exponential families or even closed convex exponential families the closure only contains *proper* log probability densities (h that satisfy the equation in (10)). This is shown by Geyer [1990, Chapter 2] and also by Csiszár and Matúš [2005]. We claim that the closure $\overline{\mathcal{H}}$ is the right way to think about completion of the exponential families, as it is explicitly constructed to facilitate useful statistical inference for practitioners. For curved exponential families and for general non-full exponential families, applying Fatou's lemma to pointwise convergence in $G(E)$ gives only

$$0 \le \int e^h \, d\lambda \le 1, \qquad h \in \overline{\mathcal{H}}. \tag{11}$$

When the integral in (11) is strictly less than one we say h is an *improper* log probability density. Examples in Geyer [1990, Chapter 4] show that improper probability densities cannot be avoided in curved exponential families.

Geyer [1990, 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 notion in Brown [1986] 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.

6.1.3 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 $0 < i \le j$*

$$H_i = \{\, x \in H_{i-1} : \langle x, \eta_i \rangle = \delta_i \,\}$$

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

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 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 η vectors that characterize the generalized affine function h according to its Theorem 5 representation. Let θ_n be a likelihood maximizing sequence of canonical parameter vectors as in (4). To make connection with the preceding section, define $h_\theta(x) = l_x(\theta) = \langle x, \theta \rangle - c(\theta)$. Then h_{θ_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:

$$h(x) = \sup_{\theta \in \Theta} l_x(\theta). \tag{12}$$

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 $\eta_1, \ldots, \eta_j$ to be a basis $\eta_1, \ldots, \eta_p$ 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*

$$h_n(y) = a_n + \sum_{i=1}^{j} b_{i,n} \left(\langle y, \eta_i \rangle - \delta_i \right) + \sum_{i=j+1}^{p} b_{i,n} \langle y, \eta_i \rangle, \qquad y \in E, \tag{13}$$

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

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

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

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

(d) *a_n converges.*

In (13) 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 θ_n is a likelihood maximizing sequence satisfying* (4) *with log densities $h_n = h_{\theta_n}$ defined by* (12) *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 θ^* is the MLE of the exponential family conditioned on the event 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 [2009] 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 θ^* is the MLE of the LCM restricted to H_j follows from our assumption that θ_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.

6.2 Convergence theorems

6.2.1 Cumulant generating function convergence

The cumulant generating function (CGF) of the distribution of the canonical statistic for parameter value θ is the function k_θ defined by

$$k_\theta(t) = \log \int e^{\langle x,t\rangle} f_\theta(x)\, \lambda(dx) = c(\theta + t) - c(\theta) \tag{14}$$

provided this distribution has a CGF, which it does if and only if k_θ 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 exponential family has a CGF if and only if the family is regular. Derivatives of k_θ evaluated at zero are the cumulants of the distribution for θ. These are the same as derivatives of c evaluated at θ.

We now show CGF convergence along likelihood maximizing sequences (4). 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 counterexample in 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 λ 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 λ, assume that every distribution in the family has a cumulant generating function. Suppose that θ_n is a likelihood maximizing sequence satisfying* (4) *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*

$$\begin{aligned} D_i &= \{y \in C_i^- : \langle y, \eta_k\rangle > \delta_k, \text{ some } k > i\}, \\ F &= E \setminus \cup_{i=1}^{j-1} D_i = \{y : \langle y, \eta_i\rangle \leq \delta_i,\ 1 \leq i \leq j\}, \end{aligned} \tag{15}$$

and assume that

$$\sup_{\theta\in\Theta} \sup_{y \in \cup_{i=1}^{j-1} D_i} e^{\langle y,\theta\rangle - c_{\cup_{i=1}^{j-1} D_i}(\theta)} < \infty \quad \text{or} \quad \lambda\left(\cup_{i=1}^{j-1} D_i\right) = 0. \tag{16}$$

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

Remarks:

1. The cumulant function $c_{\cup_{i=1}^{j-1} D_i}(\cdot)$ that appears in (16) is the log Laplace transform of the generating measure λ restricted to $\cup_{i=1}^{j-1} D_i$ where j is the number of recursively defined sets H_i, C_i^+, C_i^- in the Theorem 5 characterization of the generalized affine function h which is the pointwise limit of

the log densities h_n. The assumption that the exponential family is discrete and full implies that $\int e^{h(y)}\lambda(dy) = 1$ [Geyer, 1990, Theorem 2.7] which in turn implies that $\lambda(C_i^+) = 0$ for all $i = 1,\ldots,j$. Thus $\lambda(y)$ implies that y can only belong to C_i^- or H_i until we either arrive at an i such that $y \in C_i^-$ or $i = j$ and $y \in H_j$. However, this requirement does not rule out the possibility that $y \in D_i$. In the Supplementary Materials we provide an example of an exponential family such that $\lambda(\{y : y \in D_i\}) > 0$ and the density bound in (16) fails to hold. This example shows that no neighborhood about zero can contain values of t for which CGF convergence holds in this setting.

2. Discrete exponential families automatically satisfy (16) when the generating measure satisfies $\inf_{y \in \cup_{i=1}^{j-1} D_i} \lambda(\{y\}) > 0$. 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,

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

Therefore, Theorem 6 is applicable for the non-existence of the maximum likelihood estimator that may arise in logistic and multinomial regression or any exponential family with finite support. The same is not necessarily so for Poisson regression.

We show in the next theorem that discrete families with convex polyhedral support K also satisfy (16) 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 [Rockafellar and Wets, 1998, 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*

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

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

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

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 [Rockafellar and Wets, 1998, Theorem 6.46]. These formulas are

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

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 [Rockafellar, 1970, p. 162].*

The four conditions of Brown, stated in Section 4.1 are required for the Theorem to hold. Conditions (i) and (ii) are already assumed in Theorem 6. It is now shown that discrete exponential families satisfy (16) under the above conditions.

Theorem 7. *Assume the conditions of Theorem 6 with the omission of* (16) *when* $j \geq 2$. *Let* K *denote the convex support of the exponential family. Assume that the exponential family satisfies the conditions of Brown:*

(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* λ *is the generating measure for the exponential family.*

Then (16) *holds and we have that* $\kappa_n(t)$ *converges to* $\kappa(t)$ *pointwise for all* t *in a neighborhood of 0.*

6.3 Extensions of CGF convergence

Theorems 6 and 7 both verify CGF convergence along likelihood maximizing sequences (4) 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 (4) 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) φ_X, then $\varphi_X(t) = \varphi_{\langle X,t\rangle}(1)$, for $t \in E^*$, 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,

$$\varphi_{\langle X,t\rangle}(s) = \varphi_X(st), \qquad t \in E^* \text{ and } s \in \mathbb{R}. \tag{17}$$

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 Halmos [1974, Theorem 2 of Section 15] states that there exists a unique dual basis $w_1, \ldots, w_d$ for E^* that satisfies

$$\langle v_i, w_j\rangle = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{18}$$

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*

$$X_n \xrightarrow{d} X, \tag{19}$$

and X *has an MGF* φ_X, *and* $\varphi_{X_n}(t) \to \varphi_X(t)$, *for* $t \in E^*$.

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 completion of the exponential family. 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.

6.4 Convergence of null spaces of Fisher information

Our implementation for finding the MLE in the completion relies on finding the null space of Fisher information matrix. We first define an appropriate notion of convergence of vector subspaces, and then prove that the null spaces corresponding to a sequence of semidefinite matrices converge.

Definition 4. *Painlevé-Kuratowski set convergence [Rockafellar and Wets, 1998, Section 4.A] can be defined as follows (Rockafellar and Wets [1998] 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 ε. Let V denote the null space of A. Then $V_n \to V$ (Painlevé-Kuratowski).*

In our context, the sequence of matrices A_n in Theorem 11 correspond to the Fisher information matrices obtained from a discrete exponential family whose canonical parameters are substituted for those in a likelihood maximizing sequence.

Supplementary Materials

All proofs that do not appear in the main text and all of the code producing our examples can be seen in the supplementary materials. The supplement also includes additional data analyses. All data analyses demonstrate the functionality of the accompanying R package `glmdr` [Geyer and Eck, 2016].

7 Discussion

The theory of generalized affine functions and the geometry of exponential families allow GLM software to provide fast and scalable maximum likelihood estimation when the observed value of the canonical statistic is on the boundary of its support. The limiting probability distribution evaluated along the iterates of a likelihood maximizing sequence has log density that is a generalized affine function with structure given by Theorem 5. Cumulant generating functions converge along this sequence of iterates (Theorems 6 and 7), as do estimates of moments of all orders (Theorem 10), and so do the null spaces of Fisher information matrices (Theorem 11). These results allow one to obtain the MLE in the completion of the exponential family and to construct one-sided confidence intervals for mean-value parameters that are on the boundary of their support.

In a recent paper, Candes and Sur [2019] studied phase transitions for logistic regression models with Gaussian covariates. They showed that one may be able to determine whether or not the MLE is likely to exist before an analysis is conducted. The configuration of n and p in their setting is such that $n/p \to \kappa$ where $\kappa < 1$. Our methodology has the potential to provide useful and computationally inexpensive statistical inferences in this setting, even when phase transition arguments say that the MLE is unlikely to exist apriori. This alleviates the concern that the geometric characterization of exponential families does not tell us when we can expect an MLE to exist and when we cannot [Candes and Sur, 2019, Section 1.2].

The `glmdr` package computes one-sided confidence intervals for mean value parameters that are on the boundary of their support. Parameter estimation in the LCM is conducted in the traditional manner. The costs of computing the support of a LCM using the `glmdr` package are minimal compared to the repeated linear programming in the `rcdd` package. 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,

determine null eigenvectors of the limiting Fisher information matrix, and then compute one-sided confidence intervals than it is to compute the necessary repeated linear programming to achieve the same inferences. Our examples demonstrate that massive time savings are possible using our methodology.

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.

Acknowledgements

We would like to thank Forrest W. Crawford, his comments led to an improved and more interesting version of this paper.

A Technical appendix: Proofs of main results

Proof of Theorem 6. First consider the case when $j = 0$, the sequences of η vectors and scalars δ 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 $\theta^* + W \subset \operatorname{int}(\operatorname{dom} c)$. 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 $\theta \in \operatorname{int}(\operatorname{dom} c_1)$. 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 [Geyer, 1990, Theorem 2.2], we know that

$$c\left(\theta^* + t + s\eta_1\right) - s\delta_1 \to c_1\left(\theta^* + t\right), \quad c\left(\theta^* + s\eta_1\right) - s\delta_1 \to c_1\left(\theta^*\right), \tag{20}$$

as $s \to \infty$ since $\delta_1 \geq \langle y, \eta_1\rangle$ for all $y \in H_1$. The left hand side of both convergence arrows in (20) are convex functions of θ and the right hand side is a proper convex function. If $\operatorname{int}(\operatorname{dom} c_1)$ is nonempty, which holds whenever $\operatorname{int}(\operatorname{dom} c)$ is nonempty, then the convergence in (20) is uniform on compact subsets of $\operatorname{int}(\operatorname{dom} c_1)$ [Rockafellar and Wets, 1998, Theorem 7.17]. Also [Rockafellar and Wets, 1998, 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 θ_n in coordinate form as $\theta_n = \sum_{i=1}^{p} b_{i,n}\eta_i$, 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 $\theta^* + W \subset \text{int}(\text{dom}\, c)$. Pick $t \in W$, fix $\varepsilon > 0$, and construct ε-boxes about θ^* and $\theta^* + t$, denoted by $\mathcal{N}_{0,\varepsilon}(\theta^*)$ and $\mathcal{N}_{t,\varepsilon}(\theta^*)$ respectively, such that both $\mathcal{N}_{0,\varepsilon}(\theta^*), \mathcal{N}_{t,\varepsilon}(\theta^*) \subset \text{int}\,(\text{dom}\ c)$. Let $V_{t,\varepsilon}$ be the set of vertices of $\mathcal{N}_{t,\varepsilon}(\theta^*)$. For all $y \in E$ define

$$M_{t,\varepsilon}(y) = \max_{v \in V_{t,\varepsilon}} \{\langle v, y\rangle\}, \qquad \widetilde{M}_{t,\varepsilon}(y) = \min_{v \in V_{t,\varepsilon}} \{\langle v, y\rangle\}. \tag{21}$$

From the conclusions of Lemma 1 and Corollary 2, we can pick an integer N such that $\langle y, \psi_n + t\rangle \le M_{t,\varepsilon}(y)$ 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

$$\langle y, \theta_n + t\rangle - \sum_{i=1}^{j} b_{i,n}\delta_i = \langle y, \psi_n + t\rangle + \sum_{i=1}^{j} b_{i,n}\left(\langle y, \eta_i\rangle - \delta_i\right) \le M_{t,\varepsilon}(y) \tag{22}$$

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

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

Therefore,

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

which implies that

$$c_F(\theta_n + t) - c_F(\theta_n) \to c_{H_j}(\theta^* + t) - c_{H_j}(\theta^*) \tag{23}$$

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

$$\begin{aligned} c(\theta_n + t) - c(\theta_n) &= c_F(\theta_n + t) - c_F(\theta_n) \\ &\quad + c_{\cup_{i=1}^{j-1} D_i}(\theta_n + t) - c_{\cup_{i=1}^{j-1} D_i}(\theta_n) \\ &\to c_{H_j}(\theta^* + t) - c_{H_j}(\theta^*). \end{aligned} \tag{24}$$

We know that (24) holds when $\lambda(\cup_{i=1}^{j-1} D_i) = 0$ in (16) because of (23). Now suppose that $\lambda(\cup_{i=1}^{j-1} D_i) > 0$. We have,

$$\langle y, \psi_n + t\rangle + \sum_{i=1}^{j} b_{i,n}\left(\langle y, \eta_i\rangle - \delta_i\right) \to -\infty, \qquad y \in \cup_{i=1}^{j-1} D_i, \tag{25}$$

and

$$\begin{aligned} \exp\left(c_{\cup_{i=1}^{j-1} D_i}(\theta_n + t) - c_{\cup_{i=1}^{j-1} D_i}(\theta_n)\right) &= \int_{\cup_{i=1}^{j-1} D_i} e^{\langle y, \theta_n + t\rangle - c_{\cup_{i=1}^{j-1} D_i}(\theta_n)}\lambda(dy) \\ &\le \int_{\cup_{i=1}^{j-1} D_i} e^{M_{t,\varepsilon}(y) - \widetilde{M}_{0,\varepsilon}(y) + \langle y, \theta_n\rangle - c_{\cup_{i=1}^{j-1} D_i}(\theta_n)}\lambda(dy) \\ &\le \sup_{y \in \cup_{i=1}^{j-1} D_i} \left(e^{\langle y, \theta_n\rangle - c_{\cup_{i=1}^{j-1} D_i}(\theta_n)}\right) \lambda\left(\cup_{i=1}^{j-1} D_i\right) \int_{\cup_{i=1}^{j-1} D_i} e^{M_{t,\varepsilon}(y) - \widetilde{M}_{0,\varepsilon}(y)}\lambda(dy) \\ &\le \sup_{\theta \in \Theta} \sup_{y \in \cup_{i=1}^{j-1} D_i} \left(e^{\langle y, \theta\rangle - c_{\cup_{i=1}^{j-1} D_i}(\theta)}\right) \lambda\left(\cup_{i=1}^{j-1} D_i\right) \\ &\quad \times \int_{\cup_{i=1}^{j-1} D_i} e^{M_{t,\varepsilon}(y) - \widetilde{M}_{0,\varepsilon}(y)}\lambda(dy) \ < \ \infty \end{aligned} \tag{26}$$

for all $n > N$ by the assumption given by (16). The assumption that the exponential family is discrete and full implies that $\int e^{h}(y)\lambda(dy) = 1$ [Geyer, 1990, 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 (22), (25), and (26) together we can conclude that (24) holds as $n \to \infty$ by dominated convergence and

$$\begin{aligned} & c_{H_j}(\theta^* + t) - c_{H_j}(\theta^*) \\ & \quad = \log \int_{H_j} e^{\langle y, \theta^* + t \rangle} \lambda(dy) - \log \int_{H_j} e^{\langle y, \theta^* \rangle} \lambda(dy) \\ & \quad = \log \int e^{\langle y, t \rangle + h(y)} \lambda(dy) = \kappa(t). \end{aligned} \tag{27}$$

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

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 δ_i^* and a linearly independent set of vectors $\eta_i^* \in E^*$ can be chosen so that $\eta_i^* \in N_K(x)$, and

$$\begin{aligned} H_i &= \{y \in H_{i-1} : \langle y, \eta_i^* \rangle = \delta_i^*\}, \\ C_i^+ &= \{y \in H_{i-1} : \langle y, \eta_i^* \rangle > \delta_i^*\}, \\ C_i^- &= \{y \in H_{i-1} : \langle y, \eta_i^* \rangle < \delta_i^*\}, \end{aligned} \tag{28}$$

for $i = 1, ..., j$ where $H_0 = E$ so that (16) holds. We will prove this by induction with the hypothesis H(m), $m = 1, ..., j$, that (28) holds for $i \le m$ 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 $\int e^{h}(y)\lambda(dy) = 1$ [Geyer, 1990, Theorem 2.7]. This in turn implies that $\lambda(C_k^+) = 0$ for all $k = 1, ..., j$. This then implies that $K \subseteq \{y \in E : \langle y, \eta_1 \rangle \le \delta_1\} = H_1 \cup C_1^-$. 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 H($m+1$) follows from H(m) for $m = 1, ..., j-1$. We first establish that $K \cap H_m$ is an exposed face of K. This is needed so that (28) 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 $y \in l$ and $y \in K \cap H_m$. 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 [Rockafellar and Wets, 1998, 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 [Rockafellar and Wets, 1998, Theorem 6.42]. The normal cone $N_{H_m}(x)$ has the form

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

$$= \left\{ \sum_{i=1}^{m} a_i \eta_i : \ a_i \in \mathbb{R}, \ i = 1, ..., m \right\}.$$

Therefore, we can write

$$\eta_{m+1} = \eta^*_{m+1} + \sum_{i=1}^{m} a_{m,i} \eta^*_i \tag{29}$$

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

$$\langle y, \eta^*_{m+1} \rangle = \langle y, \eta_{m+1} \rangle - \sum_{i=1}^{m} a_{m,i} \langle y, \eta_i \rangle = \delta_{m+1} - \sum_{i=1}^{m} a_{m,i} \delta_i.$$

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

$$\begin{aligned} C^+_{m+1} &= \{ y \in H_m : \langle y, \eta_{m+1} \rangle > \delta_{m+1} \} \\ &= \left\{ y \in H_m : \langle y, \eta^*_{m+1} \rangle + \sum_{i=1}^{m} a_{m,i} \delta_i > \delta_{m+1} \right\} \\ &= \left\{ y \in H_m : \langle y, \eta^*_{m+1} \rangle > \delta_{m+1} - \sum_{i=1}^{m} a_{m,i} \delta_i \right\} \\ &= \left\{ y \in H_m : \langle y, \eta^*_{m+1} \rangle > \delta^*_{m+1} \right\}. \end{aligned} \tag{30}$$

A similar argument to that of (30) verifies that

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

This confirms that (28) holds for $i = 1, ...m + 1$ and this establishes that H$(m + 1)$ follows from H(m).

Define the sets D_i in (15) with starred quantities replacing the unstarred quantities. Since the vectors η^*_1, ...,$\eta^*_j \in N_K(x)$, the sets $K \cap D_i$ are all empty for all $i = 1, \ldots, j - 1$. Thus (16) holds with $\lambda \left(\cup_{i=1}^{j-1} D_i \right) = 0$. □

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 [Weyl, 1912], 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 γ. By construction, $V_n(\gamma) \subseteq V_n$.

From [Rockafellar and Wets, 1998, 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 ε and $p - k$ eigenvalues of A_n greater than ε for all $n > N_1$. The same is true with respect to γ 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, $x = \sum_{k=1}^{p} (x^T e_{k,n}) e_{k,n}$, $1 = \|x\|^2 = \sum_{k=1}^{p} (x^T e_{k,n})^2$, and $x^T A_n x = \sum_{k=1}^{p} \lambda_{k,n} (x^T e_{k,n})^2$. 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,

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

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 $x_n^T A_n x_n \to x^T A x$. Without loss of generality, we assume that x is a unit vector and that $|x_n^T A_n x_n - x^T A x| \leq \gamma$ for all $n \geq N_5$. From the assumption that $x_n \in V_n(\gamma)$ we have

$$x_n^T A_n x_n = \sum_{k=1}^{p} \lambda_{k,n} (x_n^T e_{k,n})^2 = \sum_{k=p-j_n(\gamma)+1}^{p} \lambda_{k,n} (x_n^T e_{k,n})^2 \leq \gamma \tag{31}$$

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

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

and (31) implies $|x^T A x| \leq 2\gamma$ for all $n \geq \max\{N_4, N_5\}$. Since this argument holds for all $0 < \gamma < \varepsilon$, we have that $x \in V$. This establishes part (ii) of Painlevé-Kuratowski convergence with respect to $V_n(\gamma)$. Thus $V_n \to V$. □

References

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, Hoboken, NJ, third edition, 2013.

M. Aickin. Existence of mles for discrete linear exponential models. *Annals of the Institute of Statistical Mathematics*, 31(1):103–113, 1979.

O. Barndorff-Nielsen. *Information and Exponential Families In Statistical Theory*. John Wiley & Sons, Chichester, 1978.

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

E. Candes and P. Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Annals of Statistics*, 2019. *To appear*.

I. Csiszár and F. Matúš. Convex cores of measures on r d. *Studia Scientiarum Mathematicarum Hungarica*, 38(1-4):177–190, 2001.

I. Csiszár and F. Matúš. Information projections revisited. *IEEE Transactions on Information Theory*, 49 (6):1474–1490, 2003.

I. Csiszár and F. Matúš. Closures of exponential families. *Ann. Probab.*, 33:582–600, 2005. doi: 10.1214/009117904000000766.

I. Csiszár and F. Matúš. Generalized maximum likelihood estimates for exponential families. *Probab. Theory Relat. Fields*, 141:213–246, 2008. doi: 10.1007/s00440-007-0084-z.

D. J. Eck and C. 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`.

N. Eriksson, S. E. Fienberg, A. Rinaldo, and S. Sullivant. Polyhedral conditions for the nonexistence of the mle for hierarchical log-linear models. *Journal of Symbolic Computation*, 41(2):222–233, 2006.

S. E. Fienberg and A. Rinaldo. Maximum likelihood estimation in log-linear models. *Annals of Statistics*, 40(2):996–1023, 2012.

C. J. Geyer. *Likelihood and Exponential Families*. PhD thesis, University of Washington, 1990. `http://hdl.handle.net/11299/56330`.

C. J. Geyer. Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proc. 23rd Symp. Interface*, pages 156–163, 1991. `http://purl.umn.edu/58440`.

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

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

C. J. Geyer and D. J. Eck. *R package `glmdr`: Exponential Family Generalized Linear Models Done Right, version 0.1*, 2016. `https://github.com/cjgeyer/glmdr/tree/master/package`.

C. J. Geyer and J. Møller. Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Statist.*, 21(4):359–373, 1994.

C. J. Geyer and E. A. Thompson. Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B*, 54(3):657–699, 1992.

C. J. Geyer, G. D. Meeden, and K. Fukuda. *R package `rcdd`: Computational Geometry, version 1.2*, 2017. `https://CRAN.R-project.org/package=rcdd`.

S. J. Haberman. *The Analysis of Frequency Data*. Chicago Press, 1974.

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

M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, P. N. Krivitsky, and M. Morris. *R package `ergm`: Fit, Simulate and Diagnose Exponential-Family Models for Networks, version 3.9.4*, 2018. `https://CRAN.R-project.org/package=ergm`.

D. R. Hunter, M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3):1–29, 2008.

F. Matúš. On limiting towards the boundaries of exponential families. *Kybernetika*, 51(5):725–738, 2015.

A. Rinaldo, S. E. Fienberg, and Y. Zhou. On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Stat.*, 3:446–484, 2009.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

R. T. Rockafellar and R. 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.

M. Schweinberger. Instability, sensitivity, and degeneracy of discrete exponential families. *J. Amer. Statist. Assoc.*, 106(496):1361–1370, 2011.

L. A. Steen and J. A. Seebach, Jr. *Counterexamples in Topology*. Springer-Verlag, New York, second edition, 1978.

A. Verbeek. The compactification of generalized linear models. *Statistica neerlandica*, 46(2-3):107–142, 1992.

N. Wang, J. Rauh, and H. Massam. Approximating faces of marginal polytopes in discrete hierarchical models. *The Annals of Statistics*, 47(3):1203–1233, 2019.

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