Electronic Journal of Statistics
Vol. 15 (2021) 2105–2156
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1815

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

Daniel J. Eck

Department of Statistics, University of Illinois
Illini Hall, 101, 725 S Wright St, Champaign, Illinois 61820
e-mail: dje13@illinois.ed

Charles J. Geyer

Department of Statistics, University of Minnesota
Ford Hall, 313, 224 Church St SE, Minneapolis, MN 55455
e-mail: charlie@stat.umn.edu

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 sequence 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 an accompanying technical report. 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.

MSC2020 subject classifications: Primary 62F10, 62F12; secondary 60F15, 14R10, 54D35.

Keywords and phrases: Completion of exponential families, complete separation, logistic regression, generalized linear models.

Received January 2020.

1. Introduction

In a regular full discrete exponential family, the MLE for the canonical parameter does not exist when the observed value of the canonical statistic lies on the boundary of its convex support [4, 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 [4, pp. 154–156], Brown [7, pp. 191–201], Csiszár and Matúš [11, 12], and Geyer [18, unpublished PhD thesis, Chapter 4]. The completion that we discuss here will consist of the limit of densities under the the topology of pointwise convergence. The properties of this closure are similar to those in Geyer [18, Chapter 4] with conditions similar to those in [7]. 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 [27, 2, 3, 36, 38, 16, 17], the development of an extended or generalized MLE when the traditional MLE does not exist through convex cores of measures [9, 10, 11, 12] and through geometric properties of exponential families and log-linear models [4, 7, 18, 40, 21, 17, 31, 41]. The issue of nonexistence also arises in exponential families for spatial lattice processes [19, 24], spatial point processes [23, 20], aster models [25], aster models with dependency groups [15], and random graphs [29, 30, 32, 37]. In every application of these (with the exception of aster models), 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 [1, Section 6.5] discuss the issue but provide no solutions.

Geyer [21] developed methodology for constructing hypothesis tests and confidence intervals when the MLE in an exponential family does not exist in the traditional sense. The algorithm in Geyer [21], implemented in the `rcdd` R package [26], 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 [4, 7, 11, 12, 17, 31, 41] 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 [21]. 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 develop methodology for constructing hypothesis tests and confidence intervals when the MLE is in the completion. 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 [18, 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. In this paper, we make the following contributions:

- We provide a computationally efficient solution that has its origins with conventional maximum likelihood computations and avoids the computationally slow linear programming algorithms in [21]. Our 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 evaluated along a likelihood maximizing sequence of canonical parameter vectors are close in the sense of moment generating function convergence (Theorems 6 and 7 below) and consequently moments of all orders are also close. Specifically, under the conditions needed for the closure in [7], Theorem 7 restores the convergence of moments that were a consequence of the original [4] theory which was appropriate for logistic and multinomial regression. The conditions of [7] hold for infinite state space models such as Poisson regression and other interesting exponential family models. Our convergence of moments results follow from a dominated convergence argument for generalized affine functions (limits of affine functions), a convex geometry argument for gen-

eralized affine functions, and a Painlevé-Kuratowski set convergence argument which implies that null spaces of the Fisher matrix evaluated along likelihood maximizing sequence of canonical parameter vectors converge.
- We develop the theoretical foundations of generalized affine functions which are the pointwise limits of sequences affine functions. Densities of exponential families are affine functions in the data. Thus, generalized affine functions represent limiting densities along sequences of canonical parameter vectors. This theory is relevant for the closure of exponential family under study and it is essential for the convergence of moments along likelihood maximizing sequences results mentioned in the preceding bullet point.

In a recent paper, Candes and Sur [8] 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 specific setting, even when phase transition arguments say that the MLE is unlikely to exist apriori. This alleviates the concern made in Section 1.2 of [8] that the geometric characterization of exponential families does not tell us when we can expect an MLE to exist and when we cannot.

Our methodology is implemented in the R package `glmdr` [22]. We demonstrate the performance of our methodology on several extensive didactic examples. These include complete separation in logistic regression and Poisson regression. Computational efficiency of our methodology is illustrated in Section 5.3. Quasi-complete separation examples in logistic regression and Bradley-Terry models are investigated in [14]. Detailed R code corresponding to these examples is also provided in [14].

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 [1, 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

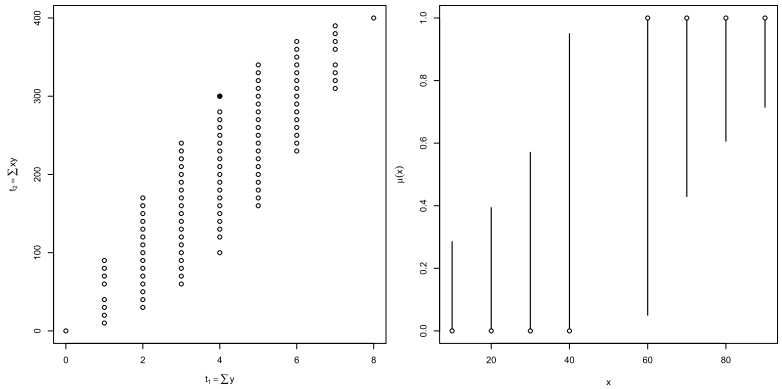

FIG 1. **Left panel**: *Observed value and support of the submodel canonical statistic vector* $M^T y$ *for the example of Section 2. Solid dot is the observed value of this statistic.* **Right panel**: *One-sided 95% confidence intervals for saturated model mean value parameters. Bars are the intervals;*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

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 4.3. The idea is to first find all canonical parameter values such that the probability of observing the realized degenerate data is greater than some testing level α. We then map those canonical parameter values to the mean value parameterization. The degeneracy follows from the estimated Fisher information matrix (for the saturated model canonical parameter vector, also called the linear predictor) at the MLE being singular which it is within the accuracy of computer arithmetic. In this motivating example, the Fisher information matrix is the zero matrix. In this case the MLE of all 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 distribution constrains some components of the response vector to be equal to their observed values, but not all of them. The remaining unconstrained components can be estimated using traditional methods. This is explained in Sections 4.2.

The methodology that we develop is applicable for any discrete regular full exponential family where the MLE does not exist in the traditional sense. We redo Example 2.3 of [21] 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. We also provide an analysis on a big data set (too large for the methods of Geyer [21] to run in an acceptable amount of time) to show the (relative) quickness of our implementation.

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 [34, 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) [18, 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 [18, 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 [18, 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

We first define the completion of the exponential family.

Definition 1. *Let θ_n, $n = 1, \ldots$, be a sequence of canonical parameter vectors for a standard exponential family having log likelihood (3). Let $h_n(x) = l_x(\theta_n)$, and suppose that $h_n(x) \to h(x)$ pointwise as $n \to \infty$ where limits $-\infty$ and $+\infty$ are allowed. The limiting functions h form the closure of the exponential family.*

In the above definition h_n is a sequence of affine functions and the limiting function h is a *generalized affine function*. Generalized affine functions and their properties are defined and discussed in Section 6.1.

4.1. Assumptions

So far everything has been for general exponential families. Our implementation requires that the conditions of Brown [7] hold, and those conditions hold for logistic and log-linear models for categorical data analysis. Now, we restrict our attention to discrete GLMs. This, in effect, includes log-linear models for contingency tables because we can always assume Poisson sampling, which makes them equivalent to multinomial sampling [1, Section 8.6.7; 21, Section 3.17].

The conditions of Brown that are required for our theory to hold are from Brown [7, 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 $h_n(x) = l_x(\theta_n)$ as in Definition 1. 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

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

[18, 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. We collect this in the definition below.

Definition 2. *Let Y be the n-dimensional vector with iid entries from a discrete regular full exponential family. Let $M \in \mathbb{R}^{n \times p}$ be a known model matrix and let $j \le p$ be the dimension of the null space of Fisher information. Then the limiting conditional model (LCM) is the original model conditioned on the event*

$$\langle M^T(Y - y), \eta_i \rangle = \langle Y - y, M\eta_i \rangle = 0, \qquad i = 1, \ldots, j, \tag{5}$$

where y is the observed value of the response vector Y and $\eta_1, \ldots, \eta_j$ spans the null space of the Fisher information matrix.

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 components of $M\eta_i$ is zero (or, taking account of the inexactness of computer arithmetic, nearly zero) for all $i = 1, \ldots, j$.

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 [21].

4.3. 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 [21] 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

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

where $\hat{\beta}$ is an MLE of the submodel canonical parameter vector in the LCM and Γ_{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). Steps for obtaining inferences are outlined in Algorithm 1. Note that confidence intervals computed from this procedure are not as we understand them in classical statistics, and are a prescription for a confidence interval routine that is otherwise inappropriate when the canonical statistic is observed to be on the boundary of its support.

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 $k \in I$. The optimization problem in (7) will be more computational

Algorithm 1 Inference when canonical statistical is on the boundary of its support

1. Declare tolerance ϵ.
2. Fit GLM model and obtain estimated Fisher information matrix.
3. Perform eigenvalue decomposition of estimated Fisher information matrix and assign null eigenvectors as those whose eigenvalues are less than ϵ.
4. Obtain the LCM using estimates of the null eigenvectors computed in the previous step, as in (5). Determine I, the index set of the components of the response vector on which we condition the OM to get the LCM.
5. Obtain inference for mean value parameters in the LCM corresponding to the components of $M\eta_i$ which are 0 for all $i = 1, \ldots, j$.
6. Obtain estimate of $\hat{\beta}$ from the LCM.
7. Obtain one-sided estimates of the mean value parameters as in (6).

stable written 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 log can be used to avoid overflow and underflow. More details are included in Section F.1 of the Appendix.

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 $\text{pr}_\beta(Y_I = y_I) = \exp\bigl(-\sum_{i \in I} \mu_i\bigr)$. 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). Computational details are given in Section F.2 of the Appendix.

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 4.3. The right panel of Figure 1 in Section 2 displays these one-sided intervals.

This example is reproduced in Section F of the Appendix in [14]. 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 [21]

This example consists of a $2\times2\times\cdots\times2$ contingency table with seven dimensions hence $2^7 = 128$ cells. These data now have a permanent location[13]. There is one response variable y that gives the cell counts and seven categorical predictors $v_1, \ldots, v_7$ 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 throughout the Appendix in [14]. 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 zero. The results are depicted in Table 1, this table is the same as Table 2 in Geyer [21] and it is reproduced in Section J of the Appendix in [14].

TABLE 1
One-sided confidence intervals for cells with MLE equal to zero.

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

The only material difference between our implementation and the linear programming in [21] is computational time. Our implementation provided 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 Section K of the Appendix in [14].

5.3. Big data example

This example uses the other dataset at [13]. It shows our methods are much faster than the linear programming method of [21]. 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 task using the `rcdd` package took 3 days, 4 hours, 0 minutes, and 40.937 seconds. (This was on `oak.stat.umn.edu`, which is an Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz.) 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 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 zero are displayed in Table 3. The full table is included in Section K.5 of the Appendix in [14]. Some of the intervals in Table 3 are relatively wide, this represents non-trivial uncertainty about the observed MLE being zero. This example is completely reproduced in Section K of the Appendix in [14].

6. Mathematical details

In this Section we provide the mathematical justification for our inferential procedure. We develop the theory of generalized affine functions [18] and then show

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

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

that this theory, combined with conditions for the exponential family closure of Brown [7], 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 [18, 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 complete the exponential family by taking pointwise limits of densities, allowing $+\infty$ and $-\infty$ as limits [18, 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 [18, Chapter 4]. (For a definition of extended-real-valued convex functions see Rockafellar [33, 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 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 Sections A-C in the Appendix. 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 [39, p. 22, gives a proof]. The space $G(E)$ is not metrizable, unless E is zero-dimensional [18, 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 [18, Chapter 2] and also by Csiszár and Matúš [11]. 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 [18, Chapter 4] show that improper probability densities cannot be avoided in curved exponential families.

Geyer [18, 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 [7] 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 $0 < i \le j$*

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

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 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 [21] calls the 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 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 Section D of the Appendix 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 quantities in (15) and (16) are technical in nature and are an artifact of the proof technique. Without these conditions, there exists circumstances in which CGF convergence fails to hold. The next remarks elaborate these quantities. The next Theorem shows that (16) is satisfied under the more intuitive conditions of [7].
2. The sets (H_i, C_i^-, C_i^+), $i = 1, \ldots, j$ arise from the characterization of a generalized affine function h given in Theorem 5 which is a pointwise limit of the densities of log densities h_n. The assumption that the exponential family is discrete and full implies that $\int e^{h(y)} \lambda(dy) = 1$ [18, Theorem 2.7] which in turn implies that $\lambda(C_i^+) = 0$ for all $i = 1, \ldots, j$. We now focus on sets of points y such that $\lambda(\{y\}) > 0$. The first iteration of the recursive structure in the Theorem 5 characterization of a generalized affine function gives $E = H_1 \cup C_1^- \cup C_1^+$. Now consider a point $y \in C_1^-$, it is possible in a full regular discrete exponential family for $\langle y, \eta_k \rangle > \delta_k$, $k > 1$ where the pair (η_k, δ_k) form the hyperplane H_k. Such points y form the set D_1, and the sets D_i, $i > 1$, are similarly motivated. Our proof technique requires bounding of the CGF restricted to $\cup_{i=1}^{j-1} D_i$ evaluated along θ_n by the quantities in (16), see (26) in the proof of Theorem 6.
3. The conditions in (16) rule out pathological examples for which CGF convergence does not hold. In Section E of the Appendix we provide an example for which (16) does not hold and a lack of CGF convergence follows. Moreover, this example demonstrates a lack of convergence of second moments and our approach for statistical inference fails as a result. More general closures of exponential families in Csiszár and Matúš [11, 12] and Geyer [18, unpublished PhD thesis, Chapter 4] do not assume condition (16) and therefore rule out CGF convergence in full generality.
4. 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,

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

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. The next Theorem provides CGF convergence for Poisson sampling under the regularity conditions of [7].

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 $K=\{y:\langle y,\alpha_i\rangle\le a_i,\ \text{for } i=1,...,m\}$, as in [34, 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 3. *The normal cone of a convex set K in the finite dimensional vector space E at a point $x\in K$ is*

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

Definition 4. *The tangent cone of a convex set K in the finite dimensional vector space E at a point $x\in K$ is*

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

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 [34, Theorem 6.46]. These formulas are

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

Definition 5. *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 [33, 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 zero.

6.3. Extensions of CGF convergence

Theorems 6 and 7 both verify CGF convergence along likelihood maximizing sequences (4) on neighborhoods of zero. The next theorems show that CGF convergence on neighborhoods of zero 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 [28, 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 6. *Painlevé-Kuratowski set convergence [34, Section 4.A] can be defined as follows (Rockafellar and Wets [34] 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

The R package `glmdr` accompanies this submission [22]. Detailed R code which reproduces the examples in this manuscript can be seen in [14].

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.

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.

Acknowledgments

We are gratefully appreciative of two anonymous referees, an anonymous AE, and the editor for very insightful comments that have led to an improved exposition. We would like to thank Forrest W. Crawford, his comments led to an improved and more interesting version of this paper.

Appendix

Appendix B: 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 $\theta^* + W \subset \text{int}(\text{dom}\, c)$. 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 [18, 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 $\text{int}(\text{dom}\ c_1)$ is nonempty, which holds whenever $\text{int}(\text{dom}\ c)$ is nonempty, then the convergence in (20) is uniform on compact subsets of $\text{int}(\text{dom}\ c_1)$ [34, Theorem 7.17]. Also [34, 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 $\theta^* + W \subset \operatorname{int}(\operatorname{dom} c)$. 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 $\mathcal{N}_{0,\varepsilon}(\theta^*), \mathcal{N}_{t,\varepsilon}(\theta^*) \subset \operatorname{int}(\operatorname{dom} c)$. 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 $i = 1, \ldots, j-1$. 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

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

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 $\int e^h(y)\lambda(dy) = 1$ [18, 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

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

for all $t \in W$, where the last equality is $\kappa(t)$. 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 $\int e^h(y)\lambda(dy) = 1$ [18, 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 $\langle a - x, \eta_m^* \rangle \leq 0$ and $\langle b - x, \eta_m^* \rangle \leq 0$ because $\eta_m^* \in N_K(x)$ by H(m). Now,

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

and

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

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 [34, 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 [34, Theorem 6.42]. The normal cone $N_{H_m}(x)$ has the form

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

Therefore, we can write

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

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 $\eta_1^*, \ldots, \eta_j^* \in N_K(x)$, 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 [42], 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 [34, 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 $j_n(\gamma) = |\{\lambda_{k,n} : \lambda_{k,n} \le \gamma\}|$ 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,

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

for all $n \ge N_3$. 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 $0 < \gamma \le \varepsilon$. 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 $n \ge N_4$, 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 $|x_n^T A_n x_n - x^T A x| \le \gamma$ for all $n \ge N_5$. From the assumption that $x_n \in V_n(\gamma)$ we have

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

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

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

and (31) implies $|x^T A x| \le 2\gamma$ for all $n \ge \max\{N_4, N_5\}$. 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 $V_n \to V$. □

Appendix C: Proofs of the properties of generalized affine functions

We first prove Theorem 2.

Proof. Let $F(E)$ denote the space of all functions $E \to \overline{\mathbb{R}}$ with the topology of pointwise convergence. This makes $F(E) = \overline{\mathbb{R}}^E$, an infinite product. Then $F(E)$ is compact by Tychonoff's theorem. We now show that $G(E)$ is closed in $F(E)$ hence compact.

Let g be any point in the closure of $G(E)$. Then there is a net $\{g_\alpha\}$ in $G(E)$ that converges to g. For any x and y in E such that $g(x) < \infty$ and $g(y) < \infty$ and any $t \in (0, 1)$, write $z = x + t(y - x)$.

Then

$$g_\alpha(z) \le (1 - t) g_\alpha(x) + t g_\alpha(y)$$

whenever the right hand side makes sense (is not $\infty - \infty$), which happens eventually, since $g_\alpha(x)$ and $g_\alpha(y)$ both converge to limits that are not ∞. Hence

$$g(z) \le \lambda g(x) + (1 - \lambda) g(y)$$

and g is convex. By symmetry it is also concave and hence is generalized affine. Thus $G(E)$ contains its closure and is closed.

$F(E)$ is Hausdorff because the product of Hausdorff spaces is Hausdorff. $G(E)$ is Hausdorff because subspaces of Hausdorff spaces are Hausdorff. □

In order to prove Theorem 1, an intermediate Theorem is first stated and its proof is provided.

Theorem 12. *12 An extended-real-valued function h on a finite-dimensional affine space E is generalized affine if and only if $h^{-1}(\infty)$ and $h^{-1}(-\infty)$ are convex sets, $h^{-1}(\mathbb{R})$ is an affine set, and h is affine on $h^{-1}(\mathbb{R})$.*

Proof. To simplify notation, define

$$A = h^{-1}(\mathbb{R}) \tag{32a}$$

$$B = h^{-1}(\infty) \tag{32b}$$

$$C = h^{-1}(-\infty) \tag{32c}$$

First assume h is generalized affine. Then C is convex because h is convex, and B is convex because h is concave. For any two distinct points $x, y \in A$ and any $s \in \mathbb{R}$, The points x, y, and $z = x + s(y - x)$ lie on a straight line. The convexity and concavity inequalities together imply

$$h\big(x + s(y - x)\big) = (1 - s)h(x) + sh(y).$$

It follows that A is an affine set and h restricted to A is an affine function.

Conversely, assume B and C are convex sets, A is an affine set, and h is affine on A. We must show that h is convex and concave. We just prove convexity because the other proof just the same proof applied to $-h$. So consider two distinct points $x, y \in A \cup C$ and $0 < t < 1$ (the convexity inequality is vacuous when either of x or y is in B). Write $z = x + t(y - x)$.

If x and y are both in A, then A being an affine set implies $z \in A$ and the convexity inequality involving x, y, and z follows from h being affine on A. If x and y are both in C, then C being a convex set implies $z \in C$ and the convexity inequality involving x, y, and z follows from $h(z) = -\infty$.

The only case remaining is $x \in A$ and $y \in C$. In this case, there can be no other point on the line determined by x and y that is in A, because A is an affine set. Hence all the points in this line on one side of x must be in B and all the points on the other side must be in C. Thus $z \in C$, and the convexity inequality involving x, y, and z follows from $h(z) = -\infty$. □

We now provide the proof of Theorem 1.

Proof. Again we use the notation in (32a), (32b), and (32c). First we show that all four cases define generalized affine functions. The first three cases obviously satisfy the conditions of Theorem 12.

In case (d), we just prove convexity because the other proof just the same proof applied to $-h$.

If x and y are both in H, then h being generalized affine on H implies the convexity inequality for x and y and any point between them. If x and y are both in C and not both in H, say $x \notin H$, then any point z between x and y is also not in H, and hence is in C because it is on the same side of H as x is. So $h(z) = -\infty$ implies the convexity inequality involving x, y, and z. That completes the proof that all four cases define generalized affine functions.

So we now show that every generalized affine function falls in one of these four cases. Suppose h is generalized affine, and assume that we are not in case (a), (b), or (c). Then at least one of B and C is nonempty. This implies $A \neq E$, hence, A being an affine set, A^c is dense in E. If $B = \emptyset$, then C is dense in E, hence C being a convex set, $C = E$ and we are in case (c) contrary to assumption. Hence $B \neq \emptyset$. The same proof with B and C swapped implies $C \neq \emptyset$.

Hence B and C are disjoint nonempty convex sets, so by the separating hyperplane theorem [33, Theorem 11.3], there is an affine function g on S such that

$$x \notin B, \qquad \text{when } g(x) < 0 \tag{33a}$$

$$x \notin C, \qquad \text{when } g(x) > 0 \tag{33b}$$

and the hyperplane in question is

$$H = \{\, x \in E : g(x) = 0 \,\}.$$

Again we know A^c is dense in E, hence B is dense in the half space on one side of H, and C is dense in the half space on the other side of H. Now convexity of B and C imply

$$x \in C, \qquad \text{when } g(x) < 0 \tag{34a}$$

$$x \in B, \qquad \text{when } g(x) > 0 \tag{34b}$$

That h is generalized affine on H follows from h being generalized affine on E. Thus we are in case (d). □

We now want to show that $G(E)$ is first countable. In aid of that we first prove a lemma.

Lemma 2. *2 Every finite-dimensional affine space E is second countable and metrizable. If D is a countable dense set in E, then every point of E is contained in the interior of the convex hull of some finite subset of D. The same is true of any open convex subset O of E: every point of O is contained in the interior of the convex hull of some finite subset of $D \cap O$.*

Proof. The first assertion is trivial. If the dimension of E is d, then the topology of E is defined to make any invertible affine function $E \to \mathbb{R}^d$ a homeomorphism.

The second assertion is just the case $O = E$ of the third assertion.

Assume to get a contradiction that the third assertion is false. Then there is a point $x \in O$ that is disjoint from the convex hull of $(O \cap D) \setminus \{x\}$. It follows

that there is a strongly separating hyperplane [33, Corollary 11.4.2], hence an affine function g such that

$$\begin{aligned} g(x) &< 0 \\ g(y) &> 0, \qquad y \in O \cap D \text{ and } y \neq x \end{aligned}$$

But this violates x being in O. □

We can now prove Theorem 3.

Proof. We need to show there is a countable local base at h for any $h \in G(E)$. A set is a neighborhood of h if it has the form

$$\{\, g \in G(E) : g(x) \in O_x,\ x \in F \,\}, \tag{35}$$

where F is a finite subset of E and each O_x is a neighborhood of $h(x)$ in $\overline{\mathbb{R}}$.

We prove first countability by induction on the dimension of E using Theorem 1. For the basis of the induction, if $E = \{0\}$, then $G(E)$ is homeomorphic to $\overline{\mathbb{R}}$, hence actually second countable.

We now show that there is a countable local base at h in each of the four cases of Theorem 1. Fix a countable dense set D in E (there is one by Lemma 2).

There is only one h satisfying case (a), the constant function having the value ∞ everywhere. In this case, a general neighborhood (35) contains a neighborhood of the form

$$W = \{\, g \in G(E) : g(x) > m,\ x \in F \,\},$$

where m can be an integer. Also by Lemma 2 there exists a finite subset V of D that contains F in the interior of its convex hull. Then, by concavity of elements of $G(E)$, the neighborhood

$$W_{m,V} = \{\, g \in G(E) : g(x) > m,\ x \in V \,\}$$

is contained in W. Hence the collection

$$\{\, W_{m,V} : m \in \mathbb{N} \text{ and } V \text{ a finite subset of } D \,\} \tag{36}$$

is a countable local base at h.

The proof for case (b) is similar. In case (c) we are considering an affine function h on E. In this case, a general neighborhood (35) contains a neighborhood of the form

$$W = \{\, g \in G(E) : h(x) - \tfrac{1}{m} < g(x) < h(x) + \tfrac{1}{m},\ x \in F \,\},$$

where F is a finite subset of E and m is a positive integer.

Again use Lemma 2 to choose a finite set V containing F in the interior of its convex hull. Then, by convexity and concavity of elements of $G(E)$, the neighborhood

$$W_{m,V} = \{\, g \in G(E) : h(x) - \tfrac{1}{m} < g(x) < h(x) + \tfrac{1}{m},\ x \in V \,\}.$$

is contained in W because any $y \in F$ can be written as a convex combination of the elements of V

$$y = \sum_{x \in V} p_x x,$$

where the p_x are nonnegative and sum to one, so $g \in W_{m,n}$ implies

$$g(y) \le \sum_{x \in V} p_x g(x) < \left(\sum_{x \in V} p_x h(x) \right) + \frac{1}{m} = h(y) + \frac{1}{m}$$

by the convexity inequality, and the same with the inequalities reversed and $1/m$ replaced by $-1/m$ by the concavity inequality. Hence the collection (36) with $W_{m,V}$ as defined in this part is a countable local base at h.

In case (d) we are considering a generalized affine function h that is neither affine nor constant. Then, as the proof of Theorem 1 shows, there is a hyperplane H that is the boundary of $h^{-1}(\infty)$ and $h^{-1}(-\infty)$. The induction hypothesis is that $G(H)$ is first countable, that is, there is a countable family $\mathcal{U}$ of neighborhoods of h in $G(E)$ such that

$$\{ U \cap H : U \in \mathcal{U} \}$$

is a countable local base for $G(H)$ at the restriction of h to H.

Again consider a general neighborhood of h (35); call it W. Let $g|H$ denote the restriction of $g \in G(E)$ to H. For any subset Q of $G(E)$ let $Q|H$ be defined by

$$Q|H = \{ q|H : q \in Q \}.$$

Then the induction hypothesis is that there exists a $U \in \mathcal{U}$ such that $U|H$ is contained in $W|H$.

Also adopt the notation (32a), (32b), and (32c) used in the proofs of Theorems 12 and 1. By Lemma 2 choose a set V_B in $D \cap (B \setminus H)$ that contains $F \cap (B \setminus H)$ in the interior of its convex hull, and choose a set V_C in $D \cap (C \setminus H)$ that contains $F \cap (C \setminus H)$ in the interior of its convex hull,

Then, by convexity and concavity of elements of $G(E)$, the neighborhood

$$W_{m,U,V_B,V_C} = \{ g \in U : h(x) \ge m, \ x \in V_B \text{ and } h(x) \le -m, \ x \in V_C \} \tag{37}$$

is contained in W. To see this, first consider $x \in F \cap H$ (if there are any). Any g in (37) has $g(x) \in O_x$ because of $U|H \subset W|H$. Next consider $x \in F \cap B$ (if there are any). Any g in (37) has $g(x) \in O_x$ because of concavity of g assures $g(x) \ge m$, and we chose m so that $(m, \infty) \subset O_x$. Last consider $x \in F \cap C$ (if there are any). Any g in (37) has $g(x) \in O_x$ because of convexity of g assures $g(x) \le -m$, and we chose m so that $(-\infty, -m) \subset O_x$.

Hence the collection

$$\{ W_{m,U,V \cap (B \setminus H),V \cap (C \setminus H)} : m \in \mathbb{N} \text{ and } U \in \mathcal{U} \text{ and } V \text{ a finite subset of } D \}$$

is a countable local base at h.

We forgot the case where E is empty. Then $G(E)$ is a one-point space whose only element is the empty function (that has no argument-value pairs). It is trivially first countable. □

We now prove Theorem 5.

Proof. First, assume h satisfies the conditions of Theorem 1 on E. We then show that h satisfies the conditions of Theorem 5 by induction on the dimension of E. The induction hypothesis, $\mathrm{H}(p)$, is that the conclusions of Theorem 1 imply that the conclusions of Theorem 5 hold when $\dim(E) = p$. We now show that $\mathrm{H}(0)$ holds. In this setting, $E = \{0\}$. Therefore our result holds with $j = 0$ and h is constant on E. The basis of the induction holds.

Let $\dim(E) = p + 1$. We now show that $\mathrm{H}(p)$ implies that $\mathrm{H}(p + 1)$ holds. In the event that h is characterized by case (a) or (b) of Theorem 1 then our result holds with $j = 0$. If case (c) of Theorem 1 characterizes h then there is an affine function f_1 defined by $f_1(x) = \langle x, \eta_1 \rangle - \delta_1$, $x \in E$, such that $h(x) = +\infty$ for x such that $f_1(x) > 0$, $h(x) = -\infty$ for x such that $f_1(x) < 0$, and h is generalized affine on the hyperplane $H_1 = \{x : f_1(x) = 0\}$. The hyperplane H_1 is p-dimensional affine subspace of E. Now, for some arbitrary $\zeta_1 \in H_1$, define

$$\begin{aligned} V_1 &= \{x - \zeta_1 : x \in H_1\} \\ &= \{y \in E : \langle y, \eta_1 \rangle = \delta_1 - \langle \zeta_1, \eta_1 \rangle\} \\ &= \{y \in E : \langle y, \eta_1 \rangle = 0\} \end{aligned}$$

where the last equality follows from $\zeta_1 \in H_1$. The space V_1 is a p-dimensional vector subspace of E since every affine space containing the origin is a vector subspace [33, Theorem 1.1] and because every translate of an affine space is another affine space [33, pp. 4]. Let

$$h_1(y) = h(y + \zeta_1), \qquad y \in V_1. \tag{38}$$

The function h_1 is convex since the composition of a convex function with an affine function is convex. To see this, let $0 < \lambda < 1$, pick $y_1, y_2 \in V_1$ and observe that

$$\begin{aligned} h_1(\lambda y_1 + (1 - \lambda) y_2) &= h(\lambda y_1 + (1 - \lambda) y_2 + \zeta_1) \\ &\leq \lambda h(y_1 + \zeta_1) + (1 - \lambda) h(y_2 + \zeta_1) \\ &= \lambda h_1(y_1) + (1 - \lambda) h_1(y_2). \end{aligned}$$

A similar argument shows that h_1 is concave. Therefore h_1 is generalized affine. From our induction hypothesis, the conclusions of Theorem 1 imply that our result holds for the generalized affine function h_1. These conditions are that there exist finite sequences of vectors $\tilde{\eta}_2, \ldots, \tilde{\eta}_j$ being a linearly independent subset of V_1^*, the dual space of V_1, and scalars $\tilde{\delta}_2, \ldots, \tilde{\delta}_j$ such that h_1 has the following form. Define $\widetilde{H}_1 = V_1$ and, inductively, for integers i such that $2 < i \leq j$

$$\begin{aligned} \widetilde{H}_i &= \{\, x \in \widetilde{H}_{i-1} : \langle x, \tilde{\eta}_i \rangle = \tilde{\delta}_i \,\} \\ \widetilde{C}_i^+ &= \{\, x \in \widetilde{H}_{i-1} : \langle x, \tilde{\eta}_i \rangle > \tilde{\delta}_i \,\} \\ \widetilde{C}_i^- &= \{\, x \in \widetilde{H}_{i-1} : \langle x, \tilde{\eta}_i \rangle < \tilde{\delta}_i \,\} \end{aligned} \tag{39}$$

all of these sets (if any) being nonempty. Then $h_1(x) = +\infty$ whenever $x \in \widetilde{C}_i^+$ for any i, $h_1(x) = -\infty$ whenever $x \in \widetilde{C}_i^-$ for any i, and h_1 is either affine or constant on $\tilde{H}_j$, where $+\infty$ and $-\infty$ are allowed for constant values.

It remains to show that the conditions of Theorem 5 hold with respect to h. The vectors $\tilde{\eta}_i$, $i = 2, ..., j$ can be extended to form a set of vectors η_i, $i = 2, ..., j$ in E^* by the Hahn-Banach Theorem [35, Theorem 3.6]. The vectors η_i, $i = 2, ..., j$, form a linearly independent subset of E^*. To see this, let $\sum_{k=2}^{j} a_k \eta_k = 0$ on E for scalars a_k, $k = 2, ..., j$. Then $\sum_{k=2}^{j} a_k \eta_k = 0$ on V_1 which implies that $a_k = 0$ for $k = 2, ..., j$ by the definition of linearly independent. Let $H_0 = E$, and, for $i = 2, ..., j$, define

$$\begin{aligned} H_i &= \{\, x \in H_{i-1} : \langle x, \eta_i \rangle = \delta_i \,\} \\ C_i^+ &= \{\, x \in H_{i-1} : \langle x, \eta_i \rangle > \delta_i \,\} \\ C_i^- &= \{\, x \in H_{i-1} : \langle x, \eta_i \rangle < \delta_i \,\} \end{aligned} \tag{40}$$

where $\delta_i = \tilde{\delta}_i - \langle \zeta_1, \eta_i \rangle$ for $i = 2, ..., j$ and $\widetilde{H}_i = H_i + \zeta_1$ as a result. We see that $h(x) = h_1(x - \zeta_1) = +\infty$ whenever $\langle x + \zeta_1, \eta_i \rangle > \tilde{\delta}_i$. Therefore $h(x) = +\infty$ for all $x \in C_i^+$ for any i. The same derivation shows that $h(x) = -\infty$ whenever $x \in C_i^-$ for any i. The generalized affine function h is either affine or constant on H_j, where $+\infty$ and $-\infty$ are allowed for constant values since the composition of an affine function with an affine function is affine.

We now show that the vectors $\eta_1, ..., \eta_j$ are linearly independent. Assume that $\sum_{k=1}^{j} a_k \eta_k = 0$ on E for scalars a_k, $k = 1, ..., j$. This assumption implies that $\sum_{k=1}^{j} a_k \tilde{\eta} = 0$ on V_1^* where $\tilde{\eta}_1$ is the restriction of η_1 to V_1. Thus $\tilde{\eta}_1$ is an element of V_1^* and $\tilde{\eta}_1 = 0$ on V_1 since $\langle y, \tilde{\eta}_1 \rangle = \langle y, \eta_1 \rangle = 0$ on V_1. Therefore $\sum_{k=2}^{j} a_k \tilde{\eta}_k = 0$ where $a_k = 0$ for $k = 2, ..., j$ from what has already been shown. In the event that $a_1 = 0$, we can conclude that $\eta_1, \ldots, \eta_j$ are linearly independent. Now consider $a_1 \neq 0$. In this case, $\sum_{k=1}^{j} a_k \eta_k = 0$ implies that $\eta_1 = \sum_{k=2}^{j} b_k \eta_k$ where $b_k = -a_k / a_1$. This states that $\sum_{k=2}^{j} b_k \tilde{\eta}_k = 0$ on V_1. Therefore, $b_k = 0$ for all $k = 2, ..., j$ which implies that η_1 is the zero vector, which is a contradiction. Thus $a_1 = 0$ and we can conclude that $\eta_1, ..., \eta_j$ are linearly independent. This completes one direction of the proof.

Now assume that h satisfies the conclusions of Theorem 5 and show that these conclusions imply that Theorem 1 holds by induction on j. The induction hypothesis, H(j), is that the conclusions of Theorem 5 imply that the conclusions of Theorem 1 hold for sequences of length j. For the basis of the induction let $j = 0$. We now show that H(0) holds. The generalized affine function h is either affine or constant on E where $+\infty$ and $-\infty$ are allowed for constant values. This characterization of h is the same as cases (a) of (b) of Theorem 1. The basis of the induction holds.

We now show that H(j) implies that H($j + 1$) holds. When the length of sequences is $j + 1$, there exist vectors $\eta_1, ..., \eta_{j+1}$ and scalars $\delta_1, ..., \delta_{j+1}$ such that h has the following form. Define $H_0 = E$ and, inductively, for integers i, $0 < i \leq j + 1$, such that the sets in (40) are all 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+1}, where $+\infty$ and $-\infty$ are allowed for constant values. From the definition of the sets H_1, C_1^+, and C_1^-, there is an affine function f_1 defined by $f_1(x) = \langle x, \eta_1 \rangle - \delta_1$, $x \in E$, such that $h(x) = +\infty$ for all $x \in E$ such that $f_1(x) > 0$ and $h(x) = -\infty$ for all $x \in E$ such that $f_1(x) < 0$. This is equivalent to the case (c) characterization of h in Theorem 1, provided we show that the restriction of h to H_1 is a generalized affine function.

Define $V_1 = H_1 - \zeta_1$ for some arbitrary $\zeta_1 \in H_1$. Let $\dim(E) = p$. The space V_1 is a $(p-1)$-dimensional vector subspace of E. Define h_1 as in (38). Let $\tilde{\eta}_i$ be the restriction of η_i to V_1 so that $\tilde{\eta}_i$ is an element of V_1^* for $1 < i \leq j+1$. Now let $\widetilde{H}_1 = V_1$ and, for $1 < i \leq j+1$, we can define the sets as in (39) where $\tilde{\delta}_i = \delta_i - \langle \zeta_1, \tilde{\eta}_i \rangle$. We see that $h_1(x) = h(x+\zeta_1) = +\infty$ whenever $\langle x+\zeta_1, \eta_i \rangle > \tilde{\delta}_i$. Therefore $h_1(x) = +\infty$ for all $x \in \widetilde{C}_i^+$ for any i. The same derivation shows that $h_1(x) = -\infty$ whenever $x \in \widetilde{C}_i^-$ for any i. The generalized affine function h_1 is either affine or constant on H_{j+1}, where $+\infty$ and $-\infty$ are allowed for constant values. Therefore h_1 meets the conditions of Theorem 5 with sequences of length j. From H(j), we know that the conclusions of Theorem 1 hold with respect to h_1. This completes the proof. □

We now prove Lemma 1 using the characterization of generalized affine functions on finite-dimensional vector spaces given by Theorem 5.

Proof. First suppose that h_n converges to h. The assumption that h is finite at at least one point guarantees that h is affine on H_j from Theorem 5. For all $y \in H_j$ we can write $h(y) = \langle y, \theta^* \rangle + a$ where $\langle y, \theta^* \rangle = \sum_{i=j+1}^{p} d_i \langle y, \eta_i \rangle$ and $s, d_i \in \mathbb{R}$. The convergence $h_n \to h$ implies that $b_{i,n} \to d_i$, $i = j+1, ..., p$ where the set of $b_{i,n}$s is empty when $j = p$ and that $a_n \to a$ as $n \to \infty$. Thus conclusions (c) and (d) hold. To show that conclusions (a) and (b) hold we will suppose that $j > 0$, because these conclusions are vacuous when $j = 0$. Both cases (a) and (b) will be shown by induction with the hypothesis H(m) that $b_{(j-m),n} \to +\infty$ and $b_{(j-m+1),n}/b_{(j-m),n} \to 0$ as $n \to \infty$ for $0 \leq m \leq j-1$. We now show that the basis of this induction holds. Pick $y \in C_j^+$ and observe that

$$h_n(y) = a_n + b_{j,n} \left(\langle y, \eta_j \rangle - \delta_j \right) + \sum_{k=j+1}^{p} b_{k,n} \langle y, \eta_k \rangle \to +\infty.$$

since $h(y) = +\infty$ and $h_n \to h$ pointwise. From this, we see that $b_{j,n} \to +\infty$ as $n \to \infty$ and $b_{j+1,n}/b_{j,n} \to 0$ as $n \to \infty$ from part (c). Therefore H(0) holds. It is now shown that H(m) implies that H($m+1$) holds. There exists a basis $y_1, ..., y_p$ in E^{**}, the dual space of E^*, such that $\langle y_i, \eta_k \rangle = 0$ when $i \neq k$ and $\langle y_i, \eta_k \rangle = 1$ when $i = k$. The set of vectors $y_1, ..., y_p$ is a basis of E since $E = E^{**}$. Arbitrarily choose a $y \in H_{j-m-1}$ such that $y = \sum_{i=1}^{j-m-1} \delta_i y_i + c_1 y_{j-m}$ where $c_1 > \delta_{j-m}$. At this choice of y we see that $h(y) = +\infty$ and

$$h_n(y) = a_n + \sum_{i=1}^{j-m+1} b_{i,n} \left(\langle y, \eta_i \rangle - \delta_i \right)$$

$$\begin{aligned} &= a_n + b_{(j-m),n} \left(\langle y, \eta_{j-m} \rangle - \delta_{j-m} \right) \\ &\to +\infty \end{aligned}$$

as $n \to \infty$. Therefore $b_{(j-m),n} \to +\infty$ as $n \to \infty$. Now arbitrarily choose $y = \sum_{i=1}^{j-m-1} \delta_i y_i + c_1 y_{j-m} + c_2 y_{j-m+1}$ where c_1 is defined as before and $c_2 < \delta_{j-m+1}$. At this choice of y we see that $h(y) = +\infty$ and

$$\begin{aligned} h_n(y) &= a_n + \sum_{i=1}^{j-m+1} b_{i,n} \left(\langle y, \eta_i \rangle - \delta_i \right) \\ &= a_n + b_{(j-m),n} \left(\langle y, \eta_{j-m} \rangle - \delta_{j-m} \right. \\ &\qquad \left. + \frac{b_{(j-m+1),n}}{b_{(j-m),n}} \left(\langle y, \eta_{j-m+1} \rangle - \delta_{j-m+1} \right) \right) \\ &= a_n + b_{(j-m),n} \left(c_1 - \delta_{j-m} - \frac{b_{(j-m+1),n}}{b_{(j-m),n}} \left(c_2 - \delta_{j-m+1} \right) \right) \\ &\to +\infty \end{aligned} \tag{41}$$

as $n \to \infty$. It follows from (41) that

$$\left(c_1 - \delta_{j-m} - \frac{b_{(j-m+1),n}}{b_{(j-m),n}} \left(c_2 - \delta_{j-m+1} \right) \right) \geq 0$$

for sufficiently large n. This implies that

$$\frac{b_{(j-m+1),n}}{b_{(j-m),n}} \leq \frac{c_1 - \delta_{j-m}}{\delta_{j-m-1} - c_2}$$

for sufficiently large n. From the arbitrariness of the constants c_1 and c_2 and (41), we can conclude that $b_{(j-m+1),n}/b_{(j-m),n} \to 0$ as $n \to \infty$. Therefore $\mathrm{H}(m+1)$ holds and this completes one direction of the proof.

We now assume that conditions (a) through (d) and the h_n takes the form in (13). Let $\lim_{n\to\infty} \sum_{i=j+1}^{p} b_{i,n} \eta_i = \theta^*$ and $\lim_{n\to\infty} a_n = a$. Cases (a) through (d) then imply that

$$h_n(y) \to \begin{cases} -\infty, & y \in C_i^- \\ \langle y, \theta^* \rangle + a, & y \in H_j \\ +\infty, & y \in C_i^+ \end{cases} \tag{42}$$

for all $i = 1, \ldots, j$ where the right hand side of (42) is a generalized affine function in its Theorem 5 representation. This completes the proof. □

Appendix D: Proofs of MGF and moment convergence results

We first prove Theorem 8.

Proof. Suppose φ_X is an MGF, hence finite on a neighborhood W of zero.Fix $t \in E^*$. Then by (17) $\varphi_{\langle X,t\rangle}(s)$ is finite whenever $st \in W$. Continuity of scalar multiplication means there exists an $\varepsilon > 0$ such that $st \in W$ whenever $|s| < \varepsilon$. That proves one direction.

Conversely, suppose $\varphi_{\langle X,t\rangle}$ is an MGF for each $t \in E^*$. Suppose $v_1, \ldots, v_d$ is a basis for E and $w_1, \ldots, w_d$ is the dual basis for E^* that satisfies (18). Then there exists $\varepsilon > 0$ such that $\varphi_{\langle X,w_i\rangle}$ is finite on $[-\varepsilon, \varepsilon]$ for each i.

We can write each $t \in E^*$ as a linear combination of basis vectors

$$t = \sum_{i=1}^{d} a_i w_i,$$

where the a_i are scalars that are unique [28, Theorem 1 of Section 15]. Applying (18) we get

$$\langle v_j, t\rangle = a_j,$$

so

$$t = \sum_{i=1}^{d} \langle v_i, t\rangle w_i,$$

and

$$\langle X, t\rangle = \sum_{i=1}^{d} \langle v_i, t\rangle \langle X, w_i\rangle.$$

Suppose

$$|\langle v_i, t\rangle| \le \varepsilon, \qquad i = 1, \ldots, d$$

(the set of all such t is a neighborhood of 0 in E^*). Let sign denote the sign function, which takes values -1, 0, and $+1$ as its argument is negative, zero, or positive, and write

$$s_i = \operatorname{sign}(\langle v_i, t\rangle), \qquad i = 1, \ldots, d.$$

Then we can write $\langle X, t\rangle$ as a convex combination

$$\langle X, t\rangle = \sum_{i=1}^{d} \frac{\langle v_i, t\rangle}{s_i \varepsilon} \cdot s_i \varepsilon \langle X, w_i\rangle + \left(1 - \sum_{i=1}^{d} \frac{\langle v_i, t\rangle}{s_i \varepsilon}\right) \cdot \langle X, 0\rangle.$$

So, by convexity of the exponential function,

$$\varphi_X(t) \le \sum_{i=1}^{d} \frac{\langle v_i, t\rangle}{s_i \varepsilon} \varphi_{\langle X,w_i\rangle}(s_i \varepsilon) + \left(1 - \sum_{i=1}^{d} \frac{\langle v_i, t\rangle}{s_i \varepsilon}\right) < \infty.$$

That proves the other direction. □

We now prove Theorem 9.

Proof. The one-dimensional case of this theorem is proved in [6]. We only need to show the general case follows by Cramér-Wold. It follows from the assumption that $\varphi_{\langle X_n,t\rangle}$ converges on a neighborhood W of zero for each $t \in E^*$. Then (19) follows from the one-dimensional case of this theorem and the Cramér-Wold theorem. And this implies

$$\langle X_n, t\rangle \xrightarrow{d} \langle X, t\rangle, \qquad t \in E^*.$$

By the one-dimensional case of this theorem, this implies $\langle X, t\rangle$ has an MGF for each t, and then Theorem 8 implies X has an MGF φ_X. By the one-dimensional case of this theorem, $\varphi_{\langle X_n,t\rangle}$ converges pointwise to $\varphi_{\langle X,t\rangle}$. So by (17), φ_{X_n} converges pointwise to φ_X. □

We now prove Theorem 10.

Proof. From Theorem 9, we have that $\langle X_n, t_i\rangle \xrightarrow{d} \langle X, , t_i\rangle$. Continuity of the exponential function implies that $e^{\langle X_n,t_i\rangle} \xrightarrow{d} e^{\langle X,t_i\rangle}$. Now, pick an $\varepsilon > 0$ such that both $\varepsilon \sum_{i=1}^k t_i \in W$ and $\varepsilon \sum_{i=1}^k u_i \in W$ where $u_1 = -t_1$ and $u_i = t_i$ for all $i > 1$. This construction gives

$$e^{\langle X_n,\varepsilon \sum_{i=1}^k t_i\rangle} \xrightarrow{d} e^{\langle X,\varepsilon \sum_{i=1}^k t_i\rangle} \tag{43}$$

and

$$\mathrm{E}\left(e^{\langle X_n,\varepsilon \sum_{i=1}^k t_i\rangle}\right) \xrightarrow{d} \mathrm{E}\left(e^{\langle X,\varepsilon \sum_{i=1}^k t_i\rangle}\right). \tag{44}$$

Equations (43) and (44) imply that $e^{\langle X_n,\varepsilon \sum_{i=1}^k t_i\rangle}$ is uniformly integrable by [5, Theorem 3.6]. A similar argument shows that $e^{\langle X_n,\varepsilon \sum_{i=1}^k u_i\rangle}$ is uniformly integrable. We now bound $|\varepsilon^k \prod_{i=1}^k \langle X_n, t_i\rangle|$ to show uniform integrability of $\prod_{i=1}^k \langle X_n, t_i\rangle$. Define

$$A_n = \{X_n : \prod_{i=1}^k \langle X_n, t_i\rangle \geq 0\}.$$

and let I_A be the indicator function. We have,

$$\begin{aligned}
\varepsilon^k \prod_{i=1}^k \langle X_n, t_i\rangle &\leq \prod_{i=1}^k \langle X_n, \varepsilon t_i\rangle I_{A_n} \\
&\leq e^{\langle X_n,\varepsilon \sum_{i=1}^k t_i\rangle} I_{A_n} \\
&\leq e^{\langle X_n,\varepsilon \sum_{i=1}^k t_i\rangle}
\end{aligned}$$

and

$$\begin{aligned}
-\varepsilon^k \prod_{i=1}^k \langle X_n, t_i\rangle &= \prod_{i=1}^k \langle X_n, \varepsilon u_i\rangle \\
&\leq \prod_{i=1}^k \langle X_n, \varepsilon u_i\rangle I_{A_n^c}
\end{aligned}$$

$$\leq e^{\langle X_n, \varepsilon \sum_{i=1}^k u_i \rangle} I_{A_n^c}$$
$$\leq e^{\langle X_n, \varepsilon \sum_{i=1}^k u_i \rangle}.$$

Therefore

$$|\varepsilon^k \prod_{i=1}^k \langle X_n, t_i \rangle| \leq e^{\langle X_n, \varepsilon \sum_{i=1}^k t_i \rangle} + e^{\langle X_n, \varepsilon \sum_{i=1}^k u_i \rangle}$$

The sum of uniformly integrable is uniformly integrable. This implies that $|\varepsilon^k \prod_{i=1}^k \langle X_n, t_i \rangle|$ is uniformly integrable. Scaling of uniformly integrable is also uniformly integrable, which implies $\prod_{i=1}^k \langle X_n, t_i \rangle$ is uniformly integrable. Our result follows from [5, Theorem 3.5] and this completes the proof. □

Appendix E: Counterexample

This section provides a counterexample to the non-theorem which is Theorem 6 with its conditions removed (that is, the assertion that cumulant generating function convergence always occurs). It shows that some conditions like those the theorem requires are needed.

E.1. Model

Suppose we have a two-dimensional exponential family with generating measure λ concentrated on the set

$$S = \{(0,0), (0,1)\} \cup \{\, (1,n) : n \in \mathbb{N} \,\},$$

where $\mathbb{N}$ is the set of natural numbers 0, 1, 2, ... And suppose λ takes values

$$\lambda(x) = \frac{1}{x_2!}, \qquad x \in S.$$

The Laplace transform of λ is the function of θ given by

$$1 + e^{\theta_2} + e^{\theta_1} \sum_{x_2=0}^{\infty} \frac{e^{x_2 \theta_2}}{x_2!} = 1 + e^{\theta_2} + e^{\theta_1} e^{e^{\theta_2}}$$

and the cumulant function (log Laplace transform) is

$$c(\theta) = \log \left[1 + e^{\theta_2} + e^{\theta_1 + e^{\theta_2}} \right] \tag{45}$$

E.2. Maximum likelihood

Suppose the observed value of the canonical statistic is $x = (0, 1)$.

From Chapter 2 of Geyer [18] we know that we can find the MLE in the completion of the family by taking limits first in the direction $\eta_1 = (-1, 0)$ (which is

a direction of recession) and second in the direction $\eta_2 = (0, 1)$ (which is a direction of recession for the limiting conditional model resulting from the first limit). Thus the MLE in the completion is the completely degenerate distribution concentrated at the observed data. The Theorem 5 (in the main article) characterization of the corresponding generalized affine function evaluated at data x, $h(x)$, yields set $C_1^- = \{(1, n) : n \in \mathbb{N}\}$ and thus $D_1 = \{(1, n), n\mathbb{N},\ n \geq 1\}$. Clearly $\lambda(D_1) > 0$, and we have

$$\sup_{\theta\in\Theta} \sup_{y\in D_1} e^{\langle y,\theta\rangle - c_{D_1}(\theta)} \geq \sup_{y\in\mathbb{N}} e^{\langle (1,y),(0,1)\rangle - c_{D_1}((0,1))} = \infty.$$

Therefore the bound condition of Theorem 6 in the main article is violated. We now show that CGF convergence along a likelihood maximizing sequence fails for t in a neighborhood of 0.

E.3. Log likelihood

The log likelihood is

$$\begin{aligned} l(\theta) &= x_1\theta_1 + x_2\theta_2 - c(\theta) \\ &= \theta_2 - c(\theta) \\ &= -\log\left[e^{-\theta_2} + 1 + e^{\theta_1 - \theta_2 + e^{\theta_2}}\right] \end{aligned}$$

E.4. Likelihood maximizing sequences

Because the MLE in the completion is completely degenerate and because $\lambda(x) = 1$, the log likelihood must go to $\log(1) = 0$ along any likelihood maximizing sequence.

We know from Lemma 1 in the main article that any likelihood maximizing sequence θ_n must have

(i) $\theta_{1,n} \to -\infty$,
(ii) $\theta_{2,n} \to +\infty$,
(iii) $|\theta_{2,n}/\theta_{1,n}| \to 0$,

but now we see that, in this example, it must also have

(iv) $\theta_{1,n} - \theta_{2,n} + e^{\theta_{2,n}} \to -\infty$.

Thus we see that Lemma 1 doesn't tell us everything about likelihood maximizing sequences (it may do under the conditions of Brown).

E.5. Cumulant generating function convergence

The cumulant generating function for canonical parameter value θ is

$$k_\theta(t) = c(\theta + t) - c(\theta).$$

Thus along a likelihood maximizing sequence we have

$$\begin{aligned} k_{\theta_n}(t) &= \log\left[\frac{1 + e^{\theta_2 + t_2} + e^{\theta_1 + t_1 + e^{\theta_2 + t_2}}}{1 + e^{\theta_2} + e^{\theta_1 + e^{\theta_2}}}\right] \\ &= \log\left[\frac{e^{-\theta_2} + e^{t_2} + e^{\theta_1 - \theta_2 + t_1 + e^{\theta_2 + t_2}}}{e^{-\theta_2} + 1 + e^{\theta_1 - \theta_2 + e^{\theta_2}}}\right] \end{aligned}$$

We know the denominator of the fraction converges to one along any likelihood maximizing sequence. The cumulant generating function of the distribution concentrated at x is the log of

$$e^{0 \cdot t_1 + 1 \cdot t_2}$$

so

$$k_{\text{limit}}(t) = t_2$$

Thus we see that to get the correct limit we need a different condition

(v) $\theta_{1,n} - \theta_{2,n} + e^{\theta_{2,n} + t_2} \to -\infty$.

Since (i) through (iv) do not imply (v) unless $t_2 \leq 0$, we cannot guarantee cumulant generating function convergence on a neighborhood of zero.

Suppose, for concreteness

$$\theta_n = (-n, \log(n)) \tag{46}$$

so the sequence in (v) becomes

$$-n - \log(n) + ne^{t_2}$$

Hence condition (v) is not satisfied unless $t_2 \leq 0$, but conditions (i) through (iv) are satisfied.

E.6. Nonconvergence of first moments

First moments (of the canonical statistic) are given by differentiating the cumulant function (45)

$$\nabla c(\theta) = \begin{pmatrix} \frac{e^{\theta_1 + e^{\theta_2}}}{1 + e^{\theta_2} + e^{\theta_1 + e^{\theta_2}}} \\ \frac{e^{\theta_2} + e^{\theta_1 + e^{\theta_2} + \theta_2}}{1 + e^{\theta_2} + e^{\theta_1 + e^{\theta_2}}} \end{pmatrix}$$

The first moment of the LCM, which is concentrated at x is just x. So the necessary and sufficient condition for convergence of first moments to the first moments of the LCM is

$$\frac{e^{\theta_{1,n} + e^{\theta_{2,n}}}}{1 + e^{\theta_{2,n}} + e^{\theta_{1,n} + e^{\theta_{2,n}}}} \to 0$$

$$\frac{e^{\theta_{2,n}} + e^{\theta_{1,n}+e^{\theta_{2,n}}+\theta_{2,n}}}{1 + e^{\theta_{2,n}} + e^{\theta_{1,n}+e^{\theta_{2,n}}}} \to 1$$

For the specific likelihood maximizing sequence (46) we have

$$\begin{aligned}
\frac{e^{\theta_{1,n}+e^{\theta_{2,n}}}}{1 + e^{\theta_{2,n}} + e^{\theta_{1,n}+e^{\theta_{2,n}}}} &= \frac{e^{-n+n}}{1+n+e^{-n+n}} \\
&= \frac{1}{2+n} \\
\frac{e^{\theta_{2,n}} + e^{\theta_{1,n}+e^{\theta_{2,n}}+\theta_{2,n}}}{1 + e^{\theta_{2,n}} + e^{\theta_{1,n}+e^{\theta_{2,n}}}} &= \frac{n+e^{-n+n+\log(n)}}{1+n+e^{-n+n}} \\
&= \frac{2n}{2+n}
\end{aligned}$$

The first converges to 0 as it must for CGF convergence. The second converges to 2, but it must converge to 1 for CGF convergence. So we do not get convergence of first moments for this model and this likelihood maximizing sequence, hence cannot have CGF convergence.

E.7. Nonconvergence of second moments

Non-convergence of first moments already makes CGF convergence impossible, but since our main interest in CGF convergence is convergence of second moments, which are components of the Fisher information matrix, we compute them too.

For c given by (45) and θ_n given by (46)

$$\nabla^2 c(\theta_n) = \frac{1}{(2+n)^2} \begin{pmatrix} 1+n & n^2 \\ n^2 & n(4+n^2) \end{pmatrix} \to \begin{pmatrix} 0 & 1 \\ 1 & \infty \end{pmatrix}$$

The variance-covariance matrix for the LCM is the zero matrix (the variance-covariance matrix of a completely degenerate distribution). Hence we do not get convergence of Fisher information for this example.

Appendix F: Additional computational materials

In this Section we go into more details about the computational methods implemented in the accompanying R package `glmdr`.

F.1. One-sided confidence intervals: Logistic regression

F.1.1. Theory for logistic regression

The math of logistic regression is very tricky for the computer. Unless arranged very carefully, the computer may overflow or underflow causing loss of all sig-

nificant figures. First there is the map from canonical to mean value parameters $p = \text{logit}^{-1}(\theta)$ where this inverse logit function operates componentwise

$$\begin{aligned} p_i &= \frac{e^{\theta_i}}{1 + e^{\theta_i}} = \frac{1}{1 + e^{-\theta_i}} \\ 1 - p_i &= \frac{1}{1 + e^{\theta_i}} = \frac{e^{-\theta_i}}{1 + e^{-\theta_i}} \end{aligned}$$

for all i. We should always choose one of these formulas for which we know we can have neither overflow, nor catastrophic cancellation. We always calculate $1 - p_i$ using the second line, we never calculate p_i and subtract from one because this results in catastrophic cancellation when p_i is near one. If θ_i is large positive, we choose a formula that has $e^{-\theta_i}$ in it, as that cannot overflow. If θ_i is large negative, we choose a formula that has e^{θ_i} in it, as that cannot overflow. If θ_i is not large, it doesn't matter which we choose.

We also never use the log function to take logarithms as this can cause horrible inaccuracy when the argument is near one. R has a function `log1p` that calculates $\log(1 + x)$ accurately for small values of x. Note that the map from canonical to mean value parameters gives

$$\begin{aligned} \log(p_i) &= \theta_i - \log(1 + e^{\theta_i}) = -\log(1 + e^{-\theta_i}) \\ \log(1 - p_i) &= -\log(1 + e^{\theta_i}) \quad\;\; = -\theta_i - \log(1 + e^{-\theta_i}) \end{aligned}$$

so we calculate

$$\begin{aligned} \log(p_i) &= \theta_i - \log(1 + e^{\theta_i}) = -\log(1 + e^{-\theta_i}) \\ \log(1 - p_i) &= -\log(1 + e^{\theta_i}) \quad\;\; = -\theta_i - \log(1 + e^{-\theta_i}) \end{aligned}$$

With this care, we have a hope of getting approximately correct answers out of the computer. Thus the optimization problem in (7) will be more computational stable written as (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. 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.

Because optimizers expect to optimize over $\mathbb{R}^q$ for some q, let N be a matrix whose columns are a basis for Γ_{lim}. Γ_{lim} is the whole parameter space in the complete separation example of Section 2. Thus, N can be the identity matrix. In other problems we take it to be a matrix whose columns are null eigenvectors of the Fisher information matrix. Then every $\gamma \in \Gamma_{\text{lim}}$ can be written as $\gamma = N\xi$ for some $\xi \in \mathbb{R}^q$, where q is the column dimension of N and the dimension of

Γ_{lim}. To an optimizer (the `inference` function in the `glmdr` package will use the R function `auglag` in CRAN package `alabama`) problem (8) has the abstract form

$$\begin{aligned} &\text{minimize} && f(\xi) \\ &\text{subject to} && g(\xi) \geq 0 \end{aligned} \tag{47}$$

and the optimization works better if derivatives of f and g are provided. Because R function `auglag` only does minimization, the objective function must be the negation of what we have in (8). That is

$$\begin{aligned} f(\xi) &= -\theta_k \\ \frac{\partial f(\xi)}{\partial \xi_j} &= -o_{kj} \\ g(\xi) &= \sum_{i \in I} \bigl[y_i \log(p_i) + (n_i - y_i) \log(1 - p_i) \bigr] - \log(\alpha) \\ \frac{\partial g(\xi)}{\partial \xi_j} &= \sum_{i \in I} (y_i - n_i p_i) o_{ij} \end{aligned}$$

where o_{ij} are the components of $O = MN$.

F.1.2. Quick and dirty intervals

As a sanity check and as a quick and dirty conservative (perhaps very conservative) confidence interval, we note that since all the p_i are between zero and one we must have

$$\begin{aligned} p_k^{n_k} &\geq \alpha, && y_k = n_k \\ (1 - p_k)^{n_k} &\geq \alpha, && y_k = 0 \end{aligned}$$

or

$$\begin{aligned} \alpha^{1/n_k} \leq p_k &\leq 1, && y_k = n_k \\ 0 \leq p_k &\leq 1 - \alpha^{1/n_k}, && y_k = 0 \end{aligned}$$

For $\alpha = 0.05$ and $n_k = 1$ we have

$$\begin{aligned} \alpha^{1/n_k} &= 0.05 \\ 1 - \alpha^{1/n_k} &= 0.95 \end{aligned}$$

In Section 2 no upper bound for a one-sided 95% confidence interval for the mean value parameter for a cell for which the MLE in the LCM is zero can be larger than than 0.95 and no lower bound for the analogous confidence interval for which the MLE in the LCM is one can be smaller than 0.05.

F.1.3. Support of the submodel canonical statistic

For GLM the (submodel) canonical statistic is $M^T Y$, where M is the model matrix and y is the response vector [21, Section 3.9]. There are 2^n possible values where n is the dimension of the response vector because each component of y can be either zero or one. The left panel of Figure 1 shows these possible values of the submodel canonical statistic.

F.2. One-sided confidence intervals: Poisson sampling

F.2.1. Theory

Here we modify Section F.1 above, changing what needs to be changed for Poisson regression rather than logistic regression. As in Section F.1 above, let β denote the vector of submodel canonical parameters, let $l(\beta)$ denote the log likelihood, and let $\hat{\beta}$ denote an MLE in the LCM. 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 the corresponding components of the response vector considered as a random vector and as an observed value, respectively. Then endpoints for a $100(1 - \alpha)\%$ confidence interval for a scalar parameter $g(\beta)$ are given by (6), when it does give a one-sided interval.

Since the only boundary of the mean value parameter space of the Poisson distribution is zero, in this section, we will be doing confidence intervals for mean value parameters for cells of the contingency table where the MLE in the LCM is zero. And we know the min is zero, so we only have to calculate the max.

In (6) pr denotes probability with respect to the OM not the LCM. As always in categorical data analysis, we have different possible sampling models: Poisson, multinomial, and product multinomial. So we get different intervals depending on which sampling model we use. In this section we are assuming Poisson. Let M denote the model matrix. Let $\theta = M\beta$ denote the saturated model canonical parameter (usually called "linear predictor" in GLM theory). Let $\mu = \exp(\theta)$ denote the mean value parameter (here exp operates componentwise like the R function of the same name does), then

$$\operatorname{pr}_\beta(Y_I = y_I) = \operatorname{pr}_\beta(Y_I = 0) = \exp\left(-\sum_{i \in I} \mu_i\right).$$

We could take the confidence interval problem to be

$$\begin{aligned} &\text{maximize} && \mu_k \\ &\text{subject to} && \exp\left(-\sum_{i \in I} \mu_i\right) \geq \alpha \end{aligned} \tag{48}$$

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

$$\begin{aligned} &\text{maximize} && \theta_k \\ &\text{subject to} && -\sum_{i \in I} \mu_i \geq \log(\alpha) \end{aligned} \tag{49}$$

Since $\mu_k = \exp(\theta_k)$ is a monotone transformation and log is a monotone transformation, the two problems are equivalent (a solution for one is also a solution for the other). We maximize canonical rather than mean value parameters to avoid extreme inexactness of computer arithmetic in calculating mean value parameters near zero. We take logs in the constraint for the same reasons we take logs of likelihoods.

As in logistic regression, let N be a matrix whose columns are a basis for Γ_{lim}. Then every $\gamma \in \Gamma_{\text{lim}}$ can be written as $\gamma = N\xi$ for some $\xi \in \mathbb{R}^q$, where q is the column dimension of N and the dimension of Γ_{lim}. To an optimizer (we will use R function `auglag` in CRAN package `alabama`) problem (49) has the abstract form (47) and the optimization works better if derivatives of f and g are provided. Because R function `auglag` only does minimization, the objective function must be the negation of what we have in (49). That is

$$\begin{aligned} f(\xi) &= -\theta_k \\ \frac{\partial f(\xi)}{\partial \xi_j} &= -o_{kj} \\ g(\xi) &= -\sum_{i \in I} \mu_i - \log(\alpha) \\ \frac{\partial g(\xi)}{\partial \xi_j} &= -\sum_{i \in I} \mu_i o_{ij} \end{aligned}$$

where o_{ij} are the components of $O = MN$.

F.2.2. Quick and dirty intervals

As a sanity check and as a quick and dirty conservative (perhaps very conservative) confidence interval, we note that since all the μ_i are nonnegative, the only way the constraint in (48) can be satisfied is if $\mu_k \leq -\log(\alpha)$. For $\alpha = 0.05$ this upper bound is -log(0.05) = 2.996. No upper bound for a one-sided 95% confidence interval for the mean value parameter for a cell for which the MLE in the LCM is zero can be larger than that.

F.3. One-sided confidence intervals: Multinomial sampling

F.3.1. Theory

We use the same notation as in Section F.2 above, except where modified here. Since the only boundary of the mean value parameter space of the multinomial

distribution is where one or more components of the state vector are zero, we will be doing confidence intervals for mean value parameters for cells of the contingency table where the MLE in the LCM is zero. And we know the min is zero, so we only have to calculate the max. (If the MLE in the LCM for mean value parameter vector had all but one component equal to zero, so the other was equal to one, then we could make one-sided intervals for all components. For multinomial sampling, contingency table cell probabilities are defined by

$$p_i = \frac{e^{\theta_i}}{\sum_{j \in J} e^{\theta_j}}, \qquad i \in J, \tag{50}$$

where J is the index set for the whole table. Now

$$\operatorname{pr}_\beta(Y_I = y_I) = \operatorname{pr}_\beta(Y_I = 0) = \left(\sum_{i \in J \setminus I} p_i \right)^n$$

where

$$n = \sum_{j \in J} y_j$$

is the multinomial sample size, where I is the index set of the cells that have mean value zero for the MLE in the LCM. So we could take the confidence interval problem to be

$$\begin{aligned} &\text{maximize} && p_k \\ &\text{subject to} && \left(\sum_{i \in J \setminus I} p_i \right)^n \geq \alpha \end{aligned} \tag{51}$$

where p is taken to be the function of γ described above. And this can be done for any $k \in I$. Unlike preceding theory for this problem, we cannot take θ_k to be the objective function because p_k is not a function of θ_k only (much less a monotone function of it). Consequently, to obtain computational stability, we will take logs of both equations obtaining

$$\begin{aligned} &\text{maximize} && \theta_k - \log \left(\sum_{j \in J} e^{\theta_j} \right) \\ &\text{subject to} && n \log \left(\sum_{i \in J \setminus I} e^{\theta_i} \right) - n \log \left(\sum_{j \in J} e^{\theta_j} \right) \geq \log(\alpha) \end{aligned} \tag{52}$$

The parameterization (50) introduces a direction of constancy (DOC) [21, Theorem 1 and the following discussion], the vector all of whose components are the same. So perhaps we should redo our null space of the Fisher information matrix calculation using the multinomial distribution. But this is not necessary. Movement along the DOC does not change any of the p_i so does not change any

of the equations in either of our optimization problems. We do not need to add it to the null space we obtained from the Poisson analysis. (Section 3.17 in [21] shows that every DOR for the Poisson model is also a DOR for the multinomial model.) Thus our problem has the abstract form (47) with

$$f(\xi) = -\theta_k + \log\left(\sum_{j\in J} e^{\theta_j}\right) \tag{53}$$

$$\frac{\partial f(\xi)}{\partial \xi_j} = -o_{kj} + \frac{\sum_{i\in J} e^{\theta_i} o_{ij}}{\sum_{i\in J} e^{\theta_i}} \tag{54}$$

where o_{kj} are the components of $O = MN$, and

$$g(\xi) = n\log\left(\sum_{i\in J\setminus I} e^{\theta_i}\right) - n\log\left(\sum_{j\in J} e^{\theta_j}\right) - \log(\alpha) \tag{55}$$

$$\begin{aligned}\frac{\partial g(\xi)}{\partial \xi_j} &= n\frac{\sum_{i\in J\setminus I} e^{\theta_i} o_{ij}}{\sum_{k\in J\setminus I} e^{\theta_k}} - n\frac{\sum_{i\in J} e^{\theta_i} o_{ij}}{\sum_{k\in J} e^{\theta_k}} \\ &= n\sum_{i\in J}(p_i^* - p_i)o_{ij}\end{aligned} \tag{56}$$

where

$$p_i^* = \begin{cases} e^{\theta_i} \Big/ \sum_{j\in J\setminus I} e^{\theta_j}, & i \in J \setminus I \\ 0, & \text{otherwise} \end{cases}$$

(p is the vector of probabilities in the OM, p^* is the vector of probabilities in the LCM).

F.3.2. Quick and dirty intervals

If $p_i > 0$ for some $i \in I$, then

$$\left(\sum_{j\in J\setminus I} p_j\right)^n \le (1 - p_i)^n$$

Introducing $\mu_i = np_i$ we get

$$\alpha \le \left(\sum_{i\in J\setminus I} p_i\right)^n \le \left(1 - \frac{\mu_i}{n}\right)^n \approx \exp(-\mu_i)$$

for large n. Thus this agrees with our analysis in Section F.2.2 when n is large. We get the exact inequality

$$\alpha \le \left(1 - \frac{\mu_i}{n}\right)^n$$

or

$$\alpha^{1/n} \leq 1 - \frac{\mu_i}{n}$$

or

$$\mu_i \leq n(1 - \alpha^{1/n}) = 2.9875$$

when $n = 544$, which is what it is in our contingency table example example, and $\alpha = 0.05$. And this too agrees approximately with our analysis in Section F.2.2 above.

F.3.3. Careful coding

We can modify (53) above as

$$f(\xi) = a - \theta_k + \log\left(\sum_{j \in J} e^{\theta_j - a}\right)$$

where a is any real number. We avoid overflow and catastrophic cancellation if we choose

$$a = \theta_m = \max_{j \in J} \theta_j$$

in which case we have

$$f(\xi) = \theta_m - \theta_k + \log\left(1 + \sum_{j \in J \setminus \{m\}} e^{\theta_j - \theta_m}\right)$$

in which overflow cannot occur and we avoid catastrophic cancellation in $\log(1+x)$ for small x. Using the same definition of θ_m, we modify (54) above as

$$\frac{\partial f(\xi)}{\partial \xi_j} = -o_{kj} + \frac{e^{\theta_k - \theta_m} o_{kj}}{\sum_{i \in J} e^{\theta_i - \theta_m}} = \left[-1 + \frac{e^{\theta_k - \theta_m}}{\sum_{i \in J} e^{\theta_i - \theta_m}}\right] o_{kj}$$

in which overflow cannot occur. We can modify (55) above as

$$g(\xi) = nb + n \log\left(\sum_{i \in J \setminus I} e^{\theta_i - b}\right) - na - n \log\left(\sum_{j \in J} e^{\theta_j - a}\right) - \log(\alpha)$$

where a and b are any real numbers. We avoid overflow and catastrophic cancellation if we choose a as above and

$$b = \theta_{m^*} = \max_{i \in J \setminus I} \theta_i$$

in which case we have

$$g(\xi) = n\left[\theta_{m^*} - \theta_m + \log\left(1 + \sum_{i \in (J \setminus I) \setminus \{m^*\}} e^{\theta_i - \theta_{m^*}}\right)\right.$$

$$- \log \left(1 + \sum_{j \in J \setminus \{m\}} e^{\theta_j - \theta_m} \right) \Bigg] - \log(\alpha)$$

in which overflow cannot occur and we avoid catastrophic cancellation in $\log(1+x)$ for small x. Then using the same definitions of θ_m and θ_{m^*} we modify (56) above as

$$\frac{\partial g(\xi)}{\partial \xi_j} = n \left[\frac{\sum_{i \in J \setminus I} e^{\theta_i - \theta_{m^*}} o_{ij}}{\sum_{k \in J \setminus I} e^{\theta_k - \theta_{m^*}}} - \frac{\sum_{i \in J} e^{\theta_i - \theta_m} o_{ij}}{\sum_{k \in J} e^{\theta_k - \theta_m}} \right]$$

in which overflow cannot occur.

References

[1] A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, Hoboken, NJ, third edition, 2013. MR3087436

[2] M. Aickin. Existence of mles for discrete linear exponential models. *Annals of the Institute of Statistical Mathematics*, 31(1):103–113, 1979. MR0541956

[3] A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71:1–10, 1984. MR0738319

[4] O. Barndorff-Nielsen. *Information and Exponential Families In Statistical Theory*. John Wiley & Sons, Chichester, 1978. MR0489333

[5] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, New York, second edition, 1999. 10.1002/9780470316962.

[6] P. Billingsley. *Probability and Measure*. John Wiley & Sons, Hoboken, NJ, anniversary edition, 2012. MR2893652

[7] L. D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, Hayward, CA, 1986. MR0882001

[8] 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*. MR4065151

[9] I. Csiszár and F. Matúš. Convex cores of measures on r d. *Studia Scientiarum Mathematicarum Hungarica*, 38(1-4):177–190, 2001. MR1877777

[10] I. Csiszár and F. Matúš. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, 2003. MR1984936

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

[12] I. Csiszár and F. Matúš. Generalized maximum likelihood estimates for exponential families. *Probab. Theory Relat. Fields*, 141:213–246, 2008. 10.1007/s00440-007-0084-z. MR2372970

[13] 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.

[14] D. J. Eck and C. J. Geyer. Computationally efficient likelihood inference in exponential families when the maximum likelihood estimator does not exist. *arXiv preprint 1803.11240*, 2020.

[15] D. J. Eck, R. G. Shaw, C. J. Geyer, and J. G. Kingsolver. An integrated analysis of phenotypic selection on insect body size and development time. *Evolution*, 69(9):2525–2532, 2015.

[16] 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. MR2197157

[17] S. E. Fienberg and A. Rinaldo. Maximum likelihood estimation in log-linear models. *Annals of Statistics*, 40(2):996–1023, 2012. MR2985941

[18] C. J. Geyer. *Likelihood and Exponential Families*. PhD thesis, University of Washington, 1990. http://hdl.handle.net/11299/56330. MR2685353

[19] 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.

[20] C. J. Geyer. Likelihood inference for spatial point processes. In *Stochastic Geometry (Toulouse, 1996)*, pages 79–140. Chapman & Hall/CRC, Boca Raton, FL, 1999. MR1673118

[21] C. J. Geyer. Likelihood inference in exponential families and directions of recession. *Electron. J. Stat.*, 3:259–289, 2009. 10.1214/08-EJS349. MR2495839

[22] 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.

[23] C. J. Geyer and J. Møller. Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Statist.*, 21(4):359–373, 1994. MR1310082

[24] 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. MR1185217

[25] C. J. Geyer, S. Wagenius, and R. G. Shaw. Aster models for life history analysis. *Biometrika*, 94(2):415–426, 2007. MR2380569

[26] 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.

[27] S. J. Haberman. *The Analysis of Frequency Data*. Chicago Press, 1974. MR0408098

[28] P. R. Halmos. *Finite-Dimensional Vector Spaces*. Springer-Verlag, New York, second edition, 1974. Reprint of 1958 edition published by Van Nostrand. MR0409503

[29] 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.

[30] 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.

[31] F. Matúš. On limiting towards the boundaries of exponential families. *Kybernetika*, 51(5):725–738, 2015. MR3445980

[32] 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. MR2507456

[33] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970. MR0274683

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

[35] W. Rudin. *Functional Analysis*. McGraw-Hill, New York, second edition, 1991. MR1157815

[36] T. J. Santner and D. E. Duffy. A note on a. albert and ja anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 73(3):755–758, 1986. MR0897873

[37] M. Schweinberger. Instability, sensitivity, and degeneracy of discrete exponential families. *J. Amer. Statist. Assoc.*, 106(496):1361–1370, 2011. MR2896841

[38] M. J. Silvapulle and J. Burridge. Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(1):100–106, 1986. MR0848055

[39] L. A. Steen and J. A. Seebach, Jr. *Counterexamples in Topology*. Springer-Verlag, New York, second edition, 1978. MR0507446

[40] A. Verbeek. The compactification of generalized linear models. *Statistica neerlandica*, 46(2-3):107–142, 1992. MR1178475

[41] 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. MR3911110

[42] H. Weyl. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Mathematische Annalen*, 71:441–479, 1912. MR1511670