Entropy **2014**, *16*, 4088-4100; doi:10.3390/e16074088

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

Using Geometry to Select One Dimensional Exponential Families That Are Monotone Likelihood Ratio in the Sample Space, Are Weakly Unimodal and Can Be Parametrized by a Measure of Central Tendency

Paul Vos [1] and Karim Anaya-Izquierdo [2,*]

[1] Department of Biostatistics, East Carolina University, Greenville, NC 27858, USA; E-Mail: vosp@ecu.edu

[2] Department of Mathematical Sciences, University of Bath, Bath BA27AY, UK

* Author to whom correspondence should be addressed; E-Mail: kai21@bath.ac.uk; Tel: +44-1225-384644

Received: 30 April 2014; in revised form: 30 June 2014 / Accepted: 14 July 2014 / Published: 18 July 2014

Abstract: One dimensional exponential families on finite sample spaces are studied using the geometry of the simplex Δ°_{n-1} and that of a transformation V_{n-1} of its interior. This transformation is the natural parameter space associated with the family of multinomial distributions. The space V_{n-1} is partitioned into cones that are used to find one dimensional families with desirable properties for modeling and inference. These properties include the availability of uniformly most powerful tests and estimators that exhibit optimal properties in terms of variability and unbiasedness.

Keywords: simplex; cone; exponential family; monotone likelihood ratio; unimodal; duality

1. Introduction

The motivation for the constructions in this paper begins with a sample from a one dimensional space that is discrete. We allow for a continuous sample space but assume that this has been suitably discretized into n bins. The simplest underlying structure for the probability assigned to these bins is given by the multinomial distribution. The collection of all multinomial distributions can be identified

with the $n-1$ simplex Δ_{n-1}. We use the geometry of the simplex along with a transformation of its interior Δ_{n-1}° to search for one dimensional subspaces that have good properties for modeling and for inference. In particular, we want families that can be parameterized by the mean, have only unimodal distributions, have desirable test characteristics (such as providing uniformly most powerful unbiased tests) and estimation properties (such as unbiasedness and small variability).

The boundary of the $(n-1)$ dimensional simplex Δ_{n-1} can be written as the union of simplexes of dimension $(n-2)$. This process can be repeated on the simplexes of lower dimension until the boundary consists of the vertices of the original simplex. This construction has statistical relevance to the possible supports for the probability distributions considered on the n bins. We obtain a dual decomposition for a transformation V_{n-1} (defined in Equation (5) in Section 5) of Δ_{n-1}°; it is dual in that the result can be obtained by replacing simplexes with cones. The statistical relevance of the conical decomposition is to the possible modes for all the distributions on the n bins. Since V_{n-1} is the natural parameter space for the distributions in Δ_{n-1}°, one dimensional exponential families are lines in V_{n-1} and these can be related to the cones that partition V_{n-1}. One result is that the limiting distribution for any one dimensional exponential family in Δ_{n-1}° is the uniform distribution whose support is determined by the cone that contains the limiting values of the line corresponding to the exponential family.

While one parameter exponential families can be defined quite generally by choosing a sufficient statistic, it can be useful to start with the sufficient statistics from well-known families such as the binomial, Poisson, negative binomial, normal, inverse Gaussian, and Gamma distribution. These exponential families have good modeling and inferential properties that we try to maintain by limiting the extent to which the sufficient statistic is modified. These restrictions lead to considering vectors in V_{n-1} that lie in a cone. Examples of how to construct these cones are given.

2. Motivating Examples

One dimensional exponential families such as the binomial or Poisson are the workhorse of parametric inference because of their excellent statistical properties. However, being one dimensional means they do not always fit data very well so an extension to a two (or higher) dimensional exponential family can be pursued in order to preserve the nice inferential structure. An issue with such extension is that, for each extra natural parameter added, we need to choose a new sufficient statistic and this choice can substantially change the shape of the corresponding density functions. For example densities can pass from being unimodal to have multiple modes for some parameter values. To see this, consider the following examples.

Example 1. *Altham [1] considered the so-called multiplicative generalization of the binomial distribution with corresponding density*

$$f(x;p,\phi)=\binom{n}{x}p^{x}(1-p)^{n-x}\phi^{x(x-n)}/C(p,\phi) \tag{1}$$

where C is the normalizing constant and where clearly the binomial is recovered when $\phi=1$.

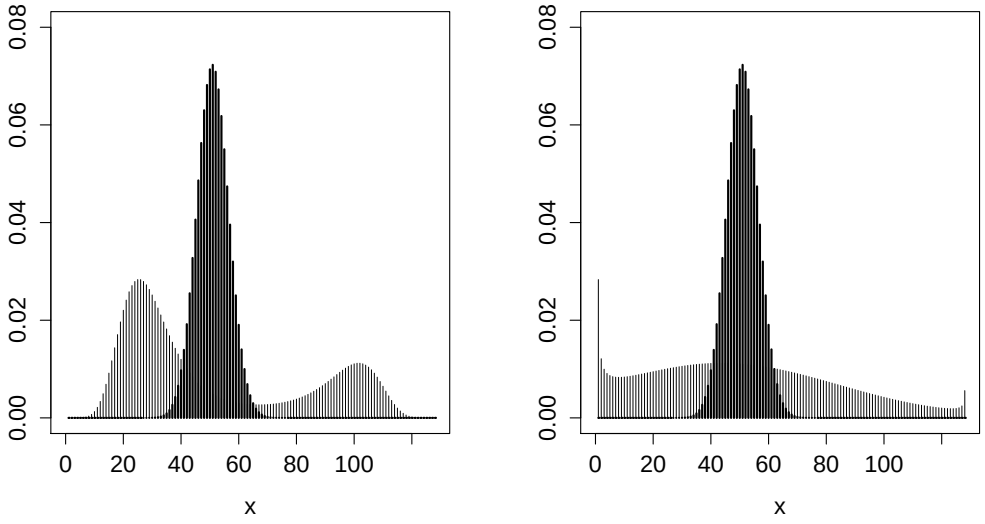
By reparametrizing using $\theta_1=\log(p/(1-p))$ and $\theta_2=\log(\phi)$ this density can be expressed in exponential form as

$$f(x;\theta_1,\theta_2)=h(x)\exp(\theta_1\,x+\theta_2\,T(x)-K(\theta_1,\theta_2)) \tag{2}$$

where $T(x) = x(x-n)$ *is the added sufficient statistic and* $h(x) = \binom{n}{x}$ *where dependence on* n *has been ignored. Note that the same family is obtained if* $T(x) = x^2$ *is added as a sufficient statistic instead of* $x(x-n)$.

If $n = 127$ *and* $(\theta_1, \theta_2) = (-0.0122, 0.018)$ *then density (2) is bimodal as shown in the left panel of Figure 1. The mean* μ *of this distribution is* 50. *Also plotted is the corresponding binomial density with the same mean or equivalently with* $\theta_1 = \log(50/(127-50)) = -0.4318$ *and* $\theta_2 = 0$.

Figure 1. Binomial density (thick in both panels). Multiplicative binomial density (left panel and thin) and double binomial density (right panel and thin). All densities have the same mean $\mu = 50$ and $n = 127$. Variance of the multiplicative and double binomial densities is equal.



As explained by Lovison [2], this distribution has the feature of being under- or over-dispersed with respect to the binomial depending on θ_2 *being negative or positive, respectively. Furthermore, using the mixed parametrization* (μ, θ_2) *(see [3] for details) it is easy to see that this distribution can be parametrized so that one parameter controls dispersion independently of the mean. In fact, for a fixed mean* μ, *as* $\theta_2 \to -\infty$ $f(x; \theta_1, \theta_2)$ *tends to a two point distribution (with support points at the extremes* $x = 0$ *and* $x = n$*) or to a degenerate distribution on* $x = \mu$ *when* $\theta_2 \to \infty$.

Example 2. *Double exponential families [4] are two parameter exponential families that extend standard unidimensional exponential families such as the binomial and the Poisson. Similar to the multiplicative binomial in Example 1, the extra parameter involved in double exponential families controls the variance independently of the mean. The density for the so-called double binomial family can be written in the form (2) with*

$$T(x) = x \log\left(\frac{x}{n}\right) + (n-x)\log\left(1 - \frac{x}{n}\right)$$

$h(x) = \binom{n}{x}$ *and with the particular restriction that* $\theta_2 < 1$ *(see [4] for details). The range* $\theta_2 < 0$ *generates underdispersion and* $\theta_2 \in [0, 1)$ *generates overdispersion with respect to the binomial. As shown on the right panel of Figure 1, the double binomial density can also be multimodal where the double binomial density shown has the same mean and variance as the multiplicative binomial shown in the left panel.*

These examples show that while extending exponential families can lead to useful modeling properties such as overdispersion, the extension can also result in distributions that are not suitable for modeling. We are interested in the relationship between geometric properties of one dimensional families and the modeling properties of their distributions.

3. Sample Space and Distribution-valued Random Variables

We consider first the general case where the sample space for a single observation X_1 consists of n bins

$$S_n = \{B_1, B_2, \ldots, B_{n-1}, B_n\}\,.$$

We consider the space of all probability distributions $\mathcal{P}$ on this sample space S_n. Each probability distribution in $\mathcal{P}$ is defined by the n-tuple p whose i^{th} component is

$$p^i = \Pr(B_i)$$

so that $\mathcal{P}$ can be identified with the $n-1$ simplex

$$\Delta_{n-1} = \{p \in \mathbb{R}^n : p^i \geq 0 \;\; \forall i, 1'p = 1\}$$

where 1 in $1'p$ is the vector $1 \in \mathbb{R}^n$ each of whose components is 1. We will slightly abuse the notation by using p to name a point in Δ_{n-1}, and hence in $\mathbb{R}^n$, as well as the corresponding distribution in $\mathcal{P}$.

The sample space for a random sample of size N from a distribution $p_0 \in \Delta_{n-1}$ is

$$\mathcal{X}_n^N = \{x : x \text{ is an } n \text{ vector of nonnegative integers that sum to } N\}\,.$$

There is simple relationship between $\mathcal{X}_n^N$ and the simplex that we obtain by dividing each component of x by N. Although the sample space $\mathcal{X}_n^N$ can be viewed as formed by compositional data, we will follow a different approach to handle this kind of data compared with the classical approach described by Aitchison [5] because the data we consider have additional structure.

In Figure 2 the sample space for the sample of size $N = 10$ is displayed using open circles. The vertices correspond to the case where all 10 values fall in a single bin. The other points correspond to the less extreme cases. Let p_0 be any point in Δ_{n-1}. By mapping the multinomial random variable of counts X to Δ_{n-1}, we obtain the random distribution $\widehat{P} = X/N$ whose values are multinomial distributions each having number of cases N and probability vector X/N. Identifying $\mathcal{X}_n^N$-valued random variables with distribution-valued random variables provides a natural means for comparing data with probability models using the Kullback–Leibler (KL) divergence.

We can compare distributions in Δ_{n-1} using the KL divergence $D : \mathcal{P} \times \mathcal{P} \mapsto \mathbb{R}$

$$D(p_1, p_2) = \sum p_1 \log\left(p_1/p_2\right) = H(p_1, p_2) - H(p_1)$$

where $H(p_1, p_2) = -\sum p_1 \log(p_2)$ and $H(p_1) = H(p_1, p_1)$ is the entropy of p_1. Note that the arguments to D and H are distributions while the logarithm and ratios are defined on points in $\mathbb{R}^n$. Following Wu and Vos [6], the variance of the random distribution $\widehat{P}$ is defined to be

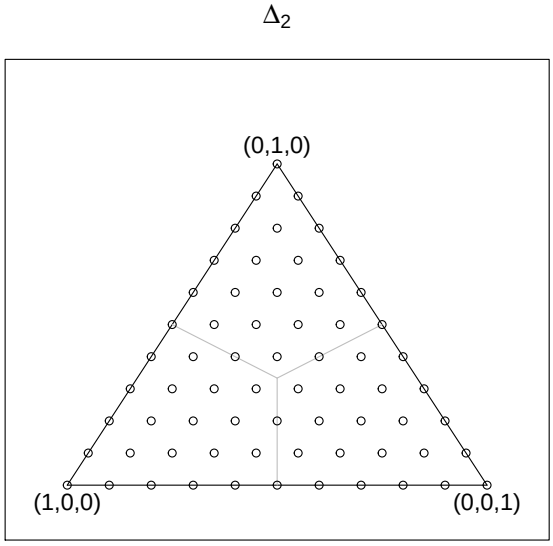
$$\mathrm{Var}_{p_0}(\widehat{P}) = \min_{p \in \Delta_{n-1}} E_{p_0} D(\widehat{P}, p)$$

and its mean is defined to be

$$\mathrm{E}_{p_0}(\widehat{P}) = \arg\min_{p\in\Delta_{n-1}} E_{p_0} D(\widehat{P}, p).$$

Note that the expectation on the right hand side of the equations above are for real-valued random variables while the expectation on the left hand side of the second equation is for a distribution-valued random variable.

Figure 2. Simplex for $n = 3$ bins and sample space for $N = 10$ observations.



It is not difficult to show that $E_{p_0}\widehat{P} = p_0$ so that $\widehat{P}$ can be considered an unbiased estimator for p_0. Details are in [6], which also shows that the KL risk can be decomposed into bias-squared and variance terms:

$$E_{p_0} D(\widehat{P}, q) = D(p_0, q) + \mathrm{Var}_{p_0}(\widehat{P}).$$

The distributional variance is related to the entropy

$$\mathrm{Var}_{p_0}(\widehat{P}) = E_{p_0} D(\widehat{P}, p_0) = H(p_0) - E_{p_0} H(\widehat{P}).$$

Note that for $N = 1$, $H(\widehat{P}) = 0$ so that for a single observation the random distribution $\widehat{P}$ taking values on the vertices of Δ_{n-1} has variance equal to the entropy of p_0.

For inference, p_0 is unknown but we specify a subspace $M \subset \Delta_{n-1}$ that contains p_0, or at least has distributions that are not too different from p_0. Estimates can be obtained by choosing a parameterization for M, say θ, and then considering real-valued functions $\hat{\theta}$ and evaluating these in terms of bias and variance. Bias and variance are useful descriptions when θ describes a feature of the distribution that is of inherent interest. However, if θ is simply a parameterization, or if there are other features that are also of interest, then these quantities are less useful. For inference regarding the distribution p_0 we can use a distribution-valued estimator $\widehat{P}_M$ where the subscript indicates that the estimator is defined to account for the fact that $p_0 \in M$.

We will not pursue the details of distribution-valued estimators here; we mention these only because all the subspaces we consider will be exponential families and in this case the maximum likelihood estimator has important properties in terms of distribution variance and distribution bias: when M

is an exponential family, the maximum likelihood estimator is distribution unbiased, and it uniquely minimizes the distribution variance among the class of all distribution unbiased estimators. Furthermore, when $p_0 \notin M$ then the maximum likelihood estimator is the unique unbiased minimum distribution variance estimator of the distribution in M that is closest (in terms of KL) to p_0. Extensions of one dimensional exponential families that do not result in exponential families will not enjoy these properties of maximum likelihood estimation. Details of these results that hold for sample spaces more general than S_n are in [7].

4. Simplices Δ_s

One dimensional exponential families on S_n are curves in Δ_{n-1} whose properties will depend on their location within various subspaces of Δ_{n-1}. An important collection of subspaces will be indexed by the subsets of S_n. For notational convenience we take B_i to the integer i. Using integers is suggestive of an ordering and a scale structure but at this point these are only being used to indicate distinct bins.

For each $s \subset S_n$,

$$\Delta_s = \left\{ p \in \mathbb{R}^n : p^i \geq 0 \ \ \forall i \in s, \ p^i = 0 \ \ \forall i \in s^c, \ 1'p = 1 \right\}$$

where $s^c = \{ i \in S_n : i \notin s \}$. Note that $\Delta_{S_n} = \Delta_{n-1}$. The interior of Δ_s is

$$\Delta_s^\circ = \left\{ p \in \Delta_s : p^i > 0 \ \ \forall i \in s \right\}.$$

As probability distributions in $\mathcal{P}$, Δ_s° corresponds to the set of all distributions having support s. There is a simple and obvious relationship between the dimension of Δ_s, $|\Delta_s|$, and the cardinality of s, $|s|$, which holds for all nonempty $s \subset S_n$

$$|\Delta_s| + 1 = |s|.$$

The boundary of Δ_s is defined as

$$\partial \Delta_s = \{ p \in \Delta_s : p \notin \Delta_s^\circ \}$$

so that

$$\Delta_s = \Delta_s^\circ \uplus \partial \Delta_s$$

where $\uplus$ indicates the sets in the union are disjoint. The boundary $\partial \Delta_s$ can be written as the union of all simplices of dimension one less than that Δ_s

$$\partial \Delta_s = \bigcup_{s' : s' \subset s, \ |s'| = |s| - 1} \Delta_{s'} \tag{3}$$

This boundary property for Δ_s holds because the simplex $\mathcal{S}_n$ consists of all possible subsets. Each nonempty $s \in \mathcal{S}_n$ specifies one of the possible supports for distribution $P \in \mathcal{P}_n$

$$\Delta_s = \biguplus_{s' : s' \subset s} \Delta_{s'}^\circ \tag{4}$$

where we set $\Delta_\emptyset = \emptyset$.

5. Cones Λ_s

The set of all nonempty subsets of the sample space provides a partition of Δ_{n-1} based on the support of the distributions in $\mathcal{P}$. The elements in the partition are simplices whose dimension is one less than the cardinality of the indexing set. In most cases we will consider models having support S_n, that is, models corresponding to Δ°_{n-1}. If we use subsets s to define the mode rather than support, we obtain a partition of $\mathcal{P}^\circ$, the distributions in $\mathcal{P}$ having support S_n. This partition can be expressed using convex cones in an $n-1$ dimensional plane V_{n-1}. The dimension of the cones are n minus the cardinality of the indexing set and the relationship between interiors of cones and their boundaries is analogous to that for simplices expressed in Equations (3) and (4).

Let

$$\mathrm{V}_{n-1} = \{v \in \mathbb{R}^n : 1'v = 0\} \tag{5}$$

be the subspace of $\mathbb{R}^n$ of dimension $n-1$ of all vectors that sum to zero. For each nonempty $s \in \mathcal{S}_n$ define

$$\Lambda_s = \left\{v \in \mathrm{V}_{n-1} : v^i \geq v^j \ \ \forall i \in s, \ \ \forall j \in S_n\right\}.$$

It is easily checked that Λ_s is a convex cone

$$v_1, v_2 \in \Lambda_s \implies a_1 v_1 + a_2 v_2 \in \Lambda_s \ \ \forall a_1, a_2 \in [0, \infty).$$

The dimension of Λ_s is $|\Lambda_s| = n - |s|$ since each point in $j \in s^c$ provides a basis vector b_j whose i^{th} component is 1 if $i \in s$ or $i = j$ and is zero otherwise and $|s^c| = n - |s|$. The interior of Λ_s is

$$\Lambda^\circ_s = \left\{v \in \Lambda_s : v^i > v^j \ \ \forall i \in s, \ \ \forall j \in s^c\right\},$$

the boundary is

$$\partial\Lambda_s = \{v \in \Lambda_s : v \notin \Lambda^\circ_s\},$$

so that

$$\Lambda_s = \Lambda^\circ_s \uplus \partial\Lambda_s$$

by definition. Note $\Lambda_{S_n} = \Lambda^\circ_{S_n} = 0 \in \mathrm{V}_{n-1} \subset \mathbb{R}^n$ where the first equality holds because the conditions in the definition of Λ°_s hold vacuously since $i \in S_n^c = \emptyset$ adds no restriction. Likewise, we can extend the definition of Λ_s to include $s = \emptyset$ and since $i \in \emptyset$ adds no restriction

$$\Lambda_\emptyset = \Lambda^\circ_\emptyset = \mathrm{V}_{n-1}.$$

Note that $\Lambda_\emptyset$ depends on the cardinality of the set S_n. Since we are considering n fixed, we will not show this dependence in the notation.

Corresponding to Equation (3) we have for all nonempty s that the boundary of the cone Λ_s is the union of all cones having dimension one less than the dimension of Λ_s

$$\partial\Lambda_s = \bigcup_{s' : s \subset s', \ |s'| = |s|+1} \Lambda_{s'}. \tag{6}$$

Corresponding to Equation (4) we have

$$\Lambda_s = \biguplus_{s' : s \subset s'} \Lambda^\circ_{s'} \tag{7}$$

The relationship between the simplices Δ and cones Λ is more easily seen if we suppress the sets that index these objects. Let Δ and Δ_* be any two simplices and let Λ and Λ_* be any two convex cones. We only consider cones and simplices that correspond to a nonempty subset of S_n. Then the Equations (6) and (7) for the convex cones are obtained by simply replacing Δ in Equations (3) and (4) with Λ:

$$\partial\Delta = \bigcup_{\Delta_*:|\Delta_*|=|\Delta|-1} \Delta_*, \quad \partial\Lambda = \bigcup_{\Lambda_*:|\Lambda_*|=|\Lambda|-1} \Lambda_* \tag{8}$$

$$\Delta = \biguplus_{\Delta_*\subset\Delta} \Delta_*^\circ, \quad \Lambda = \biguplus_{\Lambda_*\subset\Lambda} \Lambda_*^\circ \tag{9}$$

Equation (9) also holds for the empty set since $\Delta_\emptyset = \emptyset$ and $\Lambda_\emptyset = \mathrm{V}_{n-1}$.

6. V_{n-1} and $\mathcal{P}^\circ$

There is a natural bijection ϕ between V_{n-1} and Δ_{n-1}° defined by

$$\phi(p) = \log(p) - m(p)1$$

where $\log(p)$ is the vector with i^{th} component $\log(p^i)$ and $m(p)$ is defined so that $1'\phi(p) = 0$. The inverse is

$$\varphi(v) = k^{-1}(v)\exp(v)$$

where $\exp(v)$ is the vector with i^{th} component $\exp(v^i)$ and $k(v)$ is defined so that $1'\exp(v) = 1$.

Each cone Λ_s° in the partition

$$\mathrm{V}_{n-1} = \biguplus \Lambda_s^\circ$$

corresponds to one of the $2^n - 1$ possible modes for any distribution having support S_n since $v^i > v^j$ if and only if $\varphi^i(v) > \varphi^j(v)$.

7. V_{n-1} and Exponential Families in $\mathcal{P}^\circ$

We define a line by a pair of vectors $v_0, v_1 \in \mathrm{V}_{n-1}$ with $v_1 \neq 0$

$$\ell = \ell(t) = \{v \in \mathrm{V}_{n-1} : v = v_0 + tv_1, \;\; t \in \mathbb{R}\}$$

Note that v_0 and v_1 are not unique. Applying the inverse transformation φ to points in ℓ gives probability densities

$$\varphi(v_0 + tv_1) = \frac{\exp(v_0 + tv_1)}{1'\exp(v_0 + tv_1)} \tag{10}$$

which have the exponential family form with t playing the role of the natural parameter. Therefore, the space V_{n-1} is easily recognized as the natural parameter space for the distributions Δ_{n-1}° so that each line ℓ in V_{n-1} corresponds to a one dimensional exponential family.

For each line $\ell(t)$ there is a value t_{max} such that $\{\ell(t) : t \geq t_{max}\}$ is contained in one of the cones Λ_s° where s is the subset of S_n with the property that $v_1^i \geq v_1^j$ for all $i \in s$ for vectors $v_1 \in \Lambda_x^\circ$. For each line $\ell(t)$ there is a value t_{min} such that $\{\ell(t) : t \leq t_{min}\}$ is contained in one of the cones $\Lambda_{s'}^\circ$ where s' is the subset of S_n with the property that $v_1^i \leq v_1^j$ for all $i \in s'$ for vectors $v_1 \in \Lambda_x^\circ$. The cones Λ_s° and $\Lambda_{s'}^\circ$

are disjoint and will be called the *extremal* cones for ℓ. There is at least one other cone $\Lambda^\circ_{s''}$ such that $\ell \cap \Lambda^\circ_{s''} \neq \emptyset$.

Any one dimensional exponential family $\ell(t)$ can be described by an ordered sequence of disjoint cones

$$\left(\Lambda^\circ_{s_1}, \Lambda^\circ_{s_2}, \ldots, \Lambda^\circ_{s_k}\right)$$

where $k = k(\ell)$ will depend on the family. These are simply the cones that are traversed by $\ell(t)$ between its extremal cones. We take $\Lambda^\circ_{s_k}$ to be the cone that contains $\ell(t)$ for all sufficiently large t. Equation (6) for cones means that

$$\partial\Lambda_{s_i} \subset \Lambda_{s_j} \text{ for } j = i+1 \text{ or } j = i-1$$

The ordered sequence of cones provides an ordered sequence of unique subsets of S_n

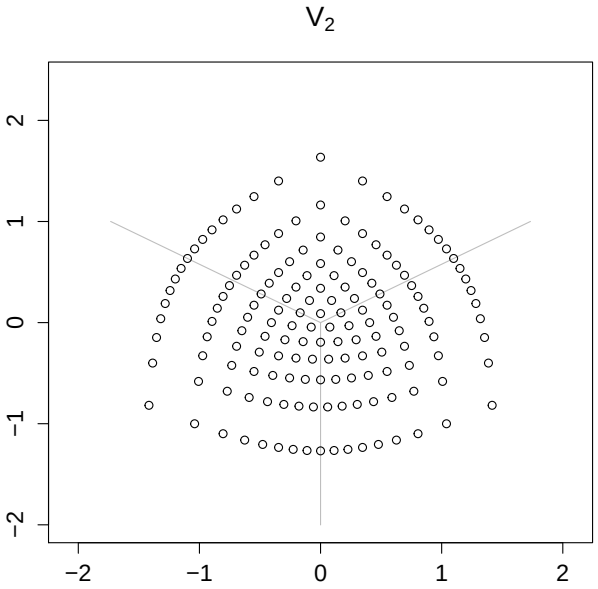
$$(s_1, s_2, \ldots, s_k)$$

that we call the *modal profile* for ℓ as these are the modes realized by the exponential family $\ell(t)$ between its extremal cones that have modes s_1 and s_k.

Each point on a line $\ell(t)$ in V_{n-1} corresponds to a distribution having support S_n. As t goes to $-\infty$ ($+\infty$) $\varphi(\ell(t))$ goes to a distribution having support s_1 (s_k). In fact, these are the uniform distribution on these supports. For every $s \subset S_n$ other than $\emptyset$ and S_n, the uniform distribution on s is a limiting distribution for some one dimensional exponential family in $\mathcal{P}^\circ$.

Figure 3 shows V_{n-1} for the two dimensional simplex shown in Figure 2. The three rays are the one dimensional cones and the spaces between these cones are the two dimensional cones. The origin is the zero dimensional cone. The sample values on the boundary of Δ_2 are not in V_2. Note that the one dimensional cones are line segments in Δ_2.

Figure 3. V_2 for $n = 3$ bins and sample space for $N = 10$ observations that are in the interior of Δ_2.



8. Ordered Bins and the Monotone Likelihood Ratio Property

Let the bins be ordered and assign the first n integers to the bins to reflect this ordering. We seek to define exponential families that have a modal profile of the form

$$(\{1\},\{1,2\},\{2\},\{2,3\},\ldots,\{n-1,n\},\{n\}) \tag{11}$$

or a contiguous sub-collection of this profile. Extensions to three or more contiguous modes are clearly possible but not discussed here.

From the definition of modal profile, it follows that a family with modal profile (11) will have the property that the mode is a non-decreasing function of t. In addition to this property for the mode, we want the likelihood ratio for any two members of the family to provide the same ordering structure as that of the bins. A family that satisfies this condition is said to have the monotone likelihood ratio property with respect to x where x takes the values of the bin labels: $1,2,\ldots,n$. Let p_{θ_1} and p_{θ_2} be two distributions in a one dimensional family parameterized by θ and let $p_{\theta_2}/p_{\theta_1}$ be the n-vector with components $p^j_{\theta_2}/p^j_{\theta_1}$ for $1 \leq j \leq n$. This family has monotone likelihood ratio if for all $\theta_1 < \theta_2$ and $j < j'$

$$\frac{p^j_{\theta_2}}{p^j_{\theta_1}} < \frac{p^{j'}_{\theta_2}}{p^{j'}_{\theta_1}}.$$

A family with this property avoids the problem situation where in general the data in the higher numbered bins are evidence for p_{θ_2} but in going from a particular bin, say j_0 to j_0+1, the likelihood ratio actually decreases. Exponential families such as the binomial and Poisson have this monotone likelihood ratio property for the bin labels. The monotone likelihood ratio property can be extended to allow for likelihood ratios that are monotone in some function of x. An important advantage of families with the monotone likelihood ratio property is the existence of uniformly most powerful tests.

To ensure that our exponential families have the monotone likelihood ratio property we consider vectors in the cone $\Lambda^{\uparrow} \subset \Lambda_n$

$$\Lambda^{\uparrow} = \left\{v : v^i < v^j, i < j\right\}.$$

From Equation (10), the exponential family indexed by θ is $k(\theta)\exp(v_0+\theta v_1)$

$$\frac{p^j_{\theta_2}}{p^j_{\theta_1}} = \frac{k(\theta_2)}{k(\theta_1)}\exp\left\{(\theta_2-\theta_1)\,v^j_1\right\}$$

so that the likelihood ratio is monotone in j if $v_1 \in \Lambda^{\uparrow}$.

9. Selecting Vectors in $\Lambda^{\uparrow}$

In order to choose n-dimensional vectors $v \in \Lambda^{\uparrow}$ we will consider a set of infinite dimensional vectors f. Let $\bar{f} : \mathbb{R} \mapsto \mathbb{R}$ and consider $f = \bar{f}|_{\mathbb{Z}}$ where $\mathbb{Z}$ is the set of integers. The function f is represented by a doubly infinite sequence

$$f = \ldots, f^{j-1}, f^j, f^{j+1}, \ldots$$

and we denote the set of all such functions as

$$\mathcal{F} = \left\{f : f^j \in \mathbb{R}\ \forall\, j \in \mathbb{Z}\right\}.$$

While it is not necessary to consider functions $\bar{f}$ to define f, these functions are useful to describe properties of f, which can be thought of as a discretized version of $\bar{f}$.

Define the gradient of f as the function ∇ whose j^{th} component is

$$(\nabla f)^j = f^j - f^{j-1}$$

The simplest functions in $\mathcal{F}$ are the constant functions

$$\mathcal{F}_0 = \left\{ f \in \mathcal{F} : f^j = f^{j'} \; \forall j, j' \in \mathbb{Z} \right\}.$$

The next simplest functions are those whose gradient is constant. We call these first order functions and denote the set of these as

$$\mathcal{F}_1 = \{ f \in \mathcal{F} : \nabla f \in \mathcal{F}_0 \}.$$

Functions in $\mathcal{F}_1$ are such that changes from one bin to the next bin is the same for all bins. That is, these functions describe constant change. We can write the functions in $\mathcal{F}_1$ explicitly as

$$\mathcal{F}_1 = \left\{ f \in \mathcal{F} : f^j = aj + b, \;\; a, b \in \mathbb{R} \right\}$$

which shows that each $f \in \mathcal{F}_1$ is the discretized version of a function $\bar{f}$ whose graph is a line in $\mathbb{R} \times \mathbb{R}$. We obtain a vector v from f by defining the j^{th} component of v as

$$v^j = f^j - \sum_1^n f^i$$

. From this definition we see that the intercept b of f does not affect v and that the slope is a scaling factor so that the restriction to first order functions results in a single direction in $\Lambda^{\uparrow}$. This direction defines the one dimensional cone defined by the vector with $v^j = j - (n+1)/2$.

Additional directions can be obtained from the second order functions

$$\mathcal{F}_2 = \{ f \in \mathcal{F} : \nabla f \in \mathcal{F}_1 \}.$$

If $f \in \mathcal{F}_2$ then $(\nabla^2 f)^j = a$ for some $a \in \mathbb{R}$ and for all $j \in \mathbb{Z}$. Using the fact that

$$\begin{aligned}(\nabla^2 f)^j = (\nabla(\nabla f))^j &= (f^j - f^{j-1}) - (f^{j-1} - f^{j-2}) \\ &= f^j + f^{j-2} - 2f^{j-1}\end{aligned}$$

the second order functions can be written explicitly as

$$\mathcal{F}_2 = \left\{ f \in \mathcal{F} : f^j = \frac{a}{2} j(j+1) + bj + c, \;\; a, b, c \in \mathbb{R} \right\}$$

.

In order for the vector v obtained from $f \in \mathcal{F}_2$ to be in $\Lambda^{\uparrow}$ we need $(\nabla f)^j \geq 0$ for $j = 1, 2, \ldots, n$. With $f^j = (a/2)j(j+1) + bj + c$ we have $(\nabla f)^j = aj + b$ so that for $a > 0$ we require $b \geq -a$ and for $a < 0$ we require $b \geq -an$. Since we are concerned with the direction rather than the magnitude we can take $a = \pm 1$ and the value of c is chosen so the sum of the components is zero.

The second order vectors in $\Lambda^{\uparrow}$ consists of the cone defined by the vectors v_{20} and v_{21} having components defined by

$$(n-1)(v_{20})^j = \frac{1}{2}j(j+1) - j - c_{20}$$
$$(n-1)(v_{21})^j = -\frac{1}{2}j(j+1) + nj - c_{21}$$

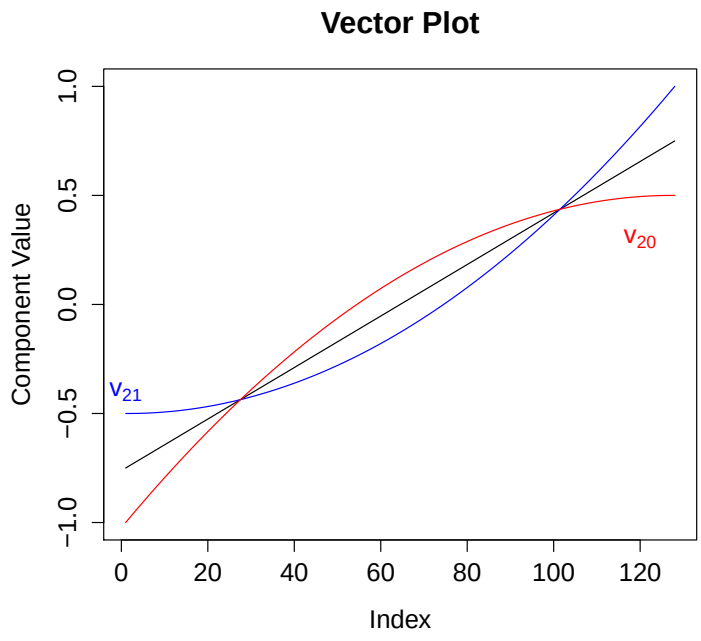
Notice that this cone contains v_1 since v_1 is proportional to $v_{20} + v_{21}$. Many discrete one dimensional exponential families (e.g., binomial, negative binomial, and Poisson) use the vector v_1. Furthermore, many continuous one dimensional exponential families use the continuous function f used to define v_1: normal with σ known, and the gamma and inverse Gaussian distributions with known shape parameter (the shape parameter is the non-scale parameter). The cone defined by v_{20} and v_{21} allows us to perturb the v_1 direction to obtain related exponential families that we would expect to have similar properties. Figure 4 shows v_{20} and v_{21} as well as $v_1 = 0.5v_{20} + 0.5v_{21}$.

Other vectors can be used to define cones around v_1. Looking at common exponential families we see that $\log(x)$ and x^{-1} are sufficient statistics so that these suggest taking $\bar{f}(x) = \log(x)$ or $\bar{f}(x) = 1/x$. These can be further generalized to $\bar{f}(x;\lambda)$, which can be the power family or some other family of transformations. The vectors v_{f0} and v_{f1} are defined using the discretized f with the constraints that $v_{f0}, v_{f1} \in \Lambda^{\uparrow}$ and $0.5v_{f0} + 0.5v_{f1} = v_1$.

An exponential family with sufficient statistic x can be modified by choosing a function $\bar{f}(x)$ and $0 \leq \alpha \leq 1$ where $\alpha = 0.5$ corresponds to the original exponential family and other values perturb this direction. We denote this vector as $v_{f\alpha}$ so that $v_0 + tv_{f\alpha}$ is the natural parameter of the modified family.

Figure 4 shows the components of the vectors v_{20} and v_{21}.

Figure 4. Components of the vectors v_{20} and v_{21} for $n = 128$ bins.



Since v_0 is common to each exponential family with natural parameter $\ell(t) = v_0 + tv_{f\alpha}$, the monotone likelihood ratio property will hold even if $v_0 \notin \Lambda^{\uparrow}$. Initial choices for v_0 are suggested by the Poisson, binomial, and negative binomial distributions:

$$
\begin{aligned}
(v_{\text{Poisson}})^j &= -\log\Gamma(j) + c \quad \notin \Lambda^{\uparrow} \\
(v_{\text{binomial}})^j &= \log\Gamma(n) - \log\Gamma(j) - \log\Gamma(n-j) + c \quad \notin \Lambda^{\uparrow} \\
(v_{\text{neg.bin.}})^j &= \log\Gamma(j+r) - \log\Gamma(j) + c \quad \in \Lambda^{\uparrow}
\end{aligned}
$$

where c is a constant chosen so that the components sum to 1, n is the number of bins, and r is a positive real constant.

Author Contributions

This paper was initiated by the first author but all sections reflect a collaborative effort. Both authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Altham, P.M.E. Two Generalizations of the Binomial Distribution. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1978**, *27*, 162–167.
2. Lovison, G. An alternative representation of Altham's multiplicative-binomial distribution. *Stat. Probab. Lett.* **1998**, *36*, 415–420.
3. Brown, L. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*; IMS Lecture Notes; Institute of Mathematical Statistics: Hayward, CA, USA, 1986.
4. Efron, B. Double Exponential Families and Their Use in Generalized Linear Regression. *J. Am. Stat. Assoc.* **1986**, *81*, 709–721.
5. Aitchison, J. *The Statistical Analysis of Compositional Data*; Chapman and Hall: London, UK, 1986.
6. Wu, Q.; Vos, P. Decomposition of Kullback–Leibler risk and unbiasedness for parameter-free estimators. *J. Stat. Plan. Inference* **2012**, *142*, 1525–1536.
7. Vos, P.; Wu, Q. Maximum Likelihood Estimators Uniformly Minimize Distribution Variance among Distribution Unbiased Estimators in Exponential Families. *Bernoulli* **2014**, submitted.