Annals of the Institute of Statistical Mathematics manuscript No.
(will be inserted by the editor)

Towards information geometry on the space of all distributions

K. Anaya-Izquierdo · F. Critchley · P. Marriott · P. W. Vos

Received: date / Revised: date

Abstract A new geometric structure for information geometry is developed with a simplicial, rather than manifold, basis. This structure is reflected in the $+1$, -1 and 0 geometries of Amari and in the duality relationship central to information geometry and corresponds to the support and moment structures of the distributions. This underlying geometric structure gives a foundation on which to build a complete theory of *computational information geometry.*

Keywords Information geometry · Multinomial distribution · Duality · Affine geometry · Exponential family

1 Introduction

This paper develops a new geometric structure for information geometry which, in particular, forms the foundation for what we term *computational information geometry*. While the information geometric properties of regular parametric families of distributions are well understood it has been surprising difficult to extend this concept to the 'space of all distributions', see for example Amari (1990), Pistone and Sempi (1995), Gibilisco and Pistone (1998), Pistone and Rogantin (1999), Amari and Nagaoka (2000), Cena (2003), Fukumizu (2005),

EPSRC Grant Number EP/E017878/1

Frank Critchley and Karim Anaya-Izquierdo
Department of Mathematics and Statistics,
The Open University

Paul Marriott
Department of Statistics and Actuarial Science,
University of Waterloo

Paul W. Vos
Department of Biostatistics,
East Carolina University

Gzyl and Recht (2006b), Gzyl and Recht (2007) and Grasselli (2008). Key assumptions in all these approaches are that all distributions have a common support and that a manifold structure is appropriate. This paper proposes that simplicial structures are, by their nature, more appropriate than manifold based ones. Specifically, they are considerably more tractable while automatically accommodating distributions with different supports. The hierarchical structure of a simplicial complex Lundell and Weingram (1969) is determined by the support and moment structure of the set of distributions. Furthermore, they arise naturally under suitable discretisation of the sample space. While this is clearly not the most general case (an obvious equivalence relation being thereby induced), it does provide an excellent foundation on which to construct a theory of computational information geometry. Indeed, in many practical applications, it can be argued (see, for example, Pitman (1979)) that, since continuous data can only be measured to finite accuracy, this discretisation is sufficient for a complete analysis.

After briefly reviewing some of the fundamentals of information geometry, this paper looks, in §3, at the apparently simple case of distributions on a finite number of categories where the extended multinomial family provides an exhaustive model underlying the corresponding information geometry. The information geometry of the multinomial distribution has been studied by a number of authors, see for example section 3.7 of Amari (1985), section 7 of Kass and Vos (1997), section 2.5 of Amari and Nagaoka (2000), and Gzyl and Recht (2006a). Following these authors, this paper shows how the appropriate simplicial geometry allows the required generalisation to extended multinomial models to be made. These extended exponential families can be viewed as the closure of general exponential families in an appropriate topology. Such closures have been studied by Barndorff-Nielsen (1978), Brown (1986), Lauritzen (1996) and Rinaldo (2006) in the finite dimensional case, and by Csiszar and Matus (2005) in the infinite dimensional case. This paper looks at the complete information geometry of such closures, examining in particular their ± 1-affine structures and duality.

In §4 the paper looks at how the simplicial structure defined for finite dimensions can be extended to the infinite dimensional case. The paper also emphasises how the simplicial structures introduced here are the building blocks of a theory of computational information geometry. The simplicial complex is useful both theoretically and computationally because its definition being completely constructive.

Since the authors feel that the tools of differential geometry are now commonplace in theoretical statistics the following notations shall be used without definition: affine geometries, α-connection, α-geodesic, duality, and Riemannian manifold. Definitions can be found in Marriott (2002), Kass and Vos (1997), Amari and Nagaoka (2000) or Murray and Rice (1993).

1.1 Motivating Example

Throughout this paper the following motivating example will be used to illustrate the geometric and statistical issues. Here the treatment focuses on theoretical issues but there are also a number of computational issues involved, discussion of which will appear in forthcoming work.

Consider a distribution on three bins, B_0, B_1 and B_2. The space of all such distributions is determined by the triple of probabilities (π_0, π_1, π_2) with the constraints,

$$\pi_0 \geq 0, \pi_1 \geq 0, \pi_2 \geq 0, \sum_{i=0}^{2} \pi_i = 1.$$

In other words it is determined by elements of the closed 2-simplex. This space can be viewed as the union of seven different natural exponential families, corresponding to the different possible choices of support set. For each natural exponential family the information geometry was clearly defined in Amari (1990), consisting of the $+1$-affine parameters, which are the natural parameters of the exponential family, the -1-affine or mean parameters, and the 0-geometry which is defined by the Fisher information and the duality which connects these together. All other α-geometries can be constructed from these key elements.

One technical issue which immediately arises is the way that the triple (π_0, π_1, π_2) 'parameterises' the space. The standard definition of a parameterisation of an open subset of a manifold requires a diffeomorphism to an open set of Euclidean space. It is immediate that this triple does not satisfy such a condition due to the fact that the boundary of the closed simplex is part of the simplex.

A union of exponential families of this form is called an extended or generalised exponential family and can be seen as a closure of a natural exponential family. Intuitively it seems that such spaces are geometrically simplexes rather than manifolds, and this paper looks to see how this intuition can be formalised.

2 Affine and information geometries

2.1 Discretisation

Although the geometric definitions in §2.2 are given explicitly only for categorical random variables or random variables defined via a countable set of categories, this section shows that the underlying geometric issues that this paper discusses are in fact much more general. Random variables are typically continuous or discrete albeit that continuous can be arbitrarily approximated by discrete ones. We focus on the discrete case with little loss.

Definition 1 A discretisation of the sample space $(\mathcal{X}, \mathcal{A})$ is defined to be a measurable partition of $\mathcal{X}$ by bins $\mathcal{B} = \{B_i\}_{i \in A}$ for $A \subseteq \mathbb{Z}$.

Example 1 Consider the space of all distributions defined on a real valued random variable X. Discretising the real line with a finite number of bins of the form

$$(-\infty, c_1), [c_1, c_2), [c_2, c_3), \ldots, [c_{k-1}, c_k), [c_k, \infty),$$

leads to a discretisation of the space of all distributions.

This can be contrasted with the following infinite set of bins

$$\{[n\epsilon, (n+1)\epsilon) \,|n \in \mathbb{Z}\}$$

for some fixed $\epsilon > 0$. In Section 4 the underlying geometry of these two choices of discretisation is compared.

Denote by $\mathcal{P}_\infty$ the space of all probability measures on $(\mathcal{X}, \mathcal{A})$. For any $P \in \mathcal{P}_\infty$ the discretising process maps P to a discrete distribution on $\mathcal{B}$ $\mathcal{P}_\mathcal{B}$ via the map which is defined by the bin probabilities

$$\pi_i := P[B = B_i] = \int_{B_i} dP \geq 0\,, \quad i \in A.$$

2.2 Affine geometries

This paper constructs a theory of information geometry following that introduced by Amari (1985) via the affine space construction introduced by Murray and Rice (1993) and extended by Marriott (2002). It concentrates on what Amari (1985) calls the $+1$, -1 and 0-geometries and on the issue of duality. These are the four essential pillars of information geometry.

Since this paper concentrates on categorical random variables, the following definitions are appropriate. Consider now (cf. Defn. 1) a countable set of disjoint categories or bins $\mathcal{B} = \{B_i\}_{i \in A}$ for $A \subseteq \mathbb{Z}$. Any distribution over this countable set of categories is defined by a sequence $\{\pi_i\}_{i \in A}$ which defines the corresponding probabilities.

Definition 2 The -1-affine space structure over a categorical random variable on $\mathcal{B} := \{B_i\}_{i \in A}$ is $(X_{mix}, V_{mix}, +)$ where

$$X_{mix} = \left\{ \{x_i\}_{i \in A} | \sum_{i \in A} x_i = 1 \right\}, V_{mix} = \left\{ \{v_i\}_{i \in A} | \sum_{i \in A} v_i = 0 \right\}$$

and the addition operator $+$ is the usual addition of sequences.

In Definition 2 the space of (discretised) distributions is a -1-convex subspace of the affine space $(X_{mix}, V_{mix}, +)$. In this convex subset the elements of X_{mix} are non-negative so there the sequences are absolutely convergent sequences and therefore the order of any potentially infinite sums does not matter. Note however in the general affine space it is necessary to be more careful about the definition of convergence. This is considered in §4.3 where the condition on the absolute convergence in V_{mix} is discussed.

The topology and geometry of the +1-affine space is a little more subtle due to the fact that the space of all distributions can be partitioned into sets of distributions with different support. It is convenient to use the following notation.

Definition 3 A space of discrete strictly positive measures is defined by

$$M = \{(x, \mathcal{P}) | \emptyset \neq \mathcal{P} \subseteq \mathcal{B}, x = \{x_i | i \in \mathcal{P}\} \text{ and } \forall i \in \mathcal{P},\ x_i > 0\},$$

and an equivalence relation $\sim$

$$(x, \mathcal{P}) \sim (x', \mathcal{P}') \iff \mathcal{P} = \mathcal{P}' \text{ and } \exists \lambda > 0, \text{ such that } x_i = \lambda x'_i, \forall i \in \mathcal{P}.$$

The quotient space is defined as $X_{exp} := M/\sim$ following Murray and Rice (1993). In order to define the vector space needed for the +1-affine structure first define the following set

$$V_{exp} = \{(v, \mathcal{P}) | \emptyset \neq \mathcal{P} \subseteq \mathcal{B}, v = \{v_i \in R | i \in \mathcal{P}\}\}$$

and then look at its partition into subsets containing all elements with the same index set $\mathcal{P}$, denoted by $V_{exp,\mathcal{P}}$, which is a vector space for each $\mathcal{P}$. Note the slight abuse of notation where by $\mathcal{P}$ refers to both a subset of bins and a set of indexes.

Define the set $X_{exp,\mathcal{P}}$ to be all equivalence classes with the same support $\mathcal{P}$. The +1-affine space for distributions with support $\mathcal{P}$ is then given by $(X_{exp,\mathcal{P}}, V_{exp,\mathcal{P}}, \oplus)$ where

$$< (x, \mathcal{P}) > \oplus (v, \mathcal{P}) =< (xe^v, \mathcal{P}) > .$$

Definitions 2 and 3 are simple generalisations of the definitions of the +1 and −1 connections via an embedding affine space that are found in Murray and Rice (1993) and Marriott (2002). The extension in Definition 3 to allow arbitrary support sets on a countable number of categories is the main extension.

2.3 Affine coordinates

On a manifold with a connection an affine coordinate system is one in which all corresponding geodesics are affine functions of the coordinates and there is a natural identification, via the exponential map, between the tangent space and the manifold.

For information geometry for the ±1-connections there are many examples of families where such affine coordinates exist. For example the natural parameters in an exponential family are +1-affine coordinates. This paper is primarily interested in the topology and geometry of the space of affine coordinates when the underlying space is not a manifold. It shows that ±1-affine coordinates are naturally defined on simplicial objects which are not open subsets of Euclidean space, which obviously is a variant on the standard definition of a parameterisation in both statistics and manifold theory.

3 Finite Simplicial Geometry

In order to illustrate the geometric issues associated with the geometry of the space of all distributions let us look first at a simple yet highly illuminating example, that of distributions on a finite number of categories.

3.1 -1-geometry

This section looks at the most natural structure underlying the -1-geometry. This is given geometrically by the simplex and statistically via the concept of the extended exponential family and its mean parameterisation.

If the partition $\mathcal{B}$ has $k+1$ bins then an element in the space of all distributions on $\mathcal{B}$, denoted by $\mathcal{P}_k$, can be identified with a point in the k-simplex Δ^k defined as

$$\Delta^k := \left\{ \boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_k)^\top \,:\, \pi_i \geq 0\,, \sum_{i=0}^{k} \pi_i = 1 \right\}.$$

Clearly, the multinomial family on $k+1$ categories, denoted by $\mathcal{P}_k^0$, can be identified with $int(\Delta^k)$. However, the complete -1-geometry of $\mathcal{P}_k$ allows the possibility of distributions with different support sets. In particular, ones where some of the bin probabilities π_i can be zero. To allow for this, the multinomial family is extended by adding all lower dimensional multinomial families defined on subsets of $\mathcal{B}$ in such a way that certain bin probabilities are zero. The simplex has a clear affine structure as it is a closed convex subset of an affine space where π are the affine coordinates in the sense described above.

Definition 4 The family of probability distributions $\mathcal{P}_k$ will be called the extended multinomial family for the set of bins $\mathcal{B} = \{B_i\}_{i=0}^k$.

Example 2 A straightforward example is the extended Binomial distribution with known parameter N and probability $\pi \in [0,1]$. The subset $\pi \in (0,1)$ defines a regular binomial family whose members have support $\{0,1,\ldots,N\}$. The extended family is obtained by adding the two degenerate distributions which correspond to $\pi = 0$, with support $\{0\}$ and to $\pi = 1$, with support $\{N\}$.

In the running trinomial example where $k = 2$, there are six lower dimensional subfamilies: three binomials which correspond to the case where exactly one of the bins has probability zero, and the three distributions degenerate at a given bin.

The closure of exponential families has been studied by Barndorff-Nielsen (1978), Brown (1986), Lauritzen (1996) and Rinaldo (2006) in the finite dimensional case and by Csiszar and Matus (2005) in the infinite dimensional case.

In summary, the -1-affine parameter, naturally exists on the closure of the k dimensional exponential family and its extension to the faces of the

simplex is a continuous function. It follows that the domain of the -1-affine parameter, or equivalently the mean parameter, is a simplex and not an open subset of Euclidean space. It is this simplicial structure, on the affine parameter space, which does not result in a manifold structure that the rest of the paper investigates.

3.2 +1-geometry

This section looks at the case of +1-affine parameters. Note that depending on the support set there are a number of distinct +1 affine spaces, while in the -1 case above there was a single simplex inside a single affine space. First consider the following example to motivate the definition of the +1-affine parameters

Example 3 In the case of the multinomial family with $\mathcal{B} = \{B_i\}_{i=0}^k$, a non-minimal representation of an exponential family is given by the $\sim$-equivalence class

$$\langle(\pi_0 \exp(v_0), \ldots, \pi_k \exp(v_k))\rangle$$

which is 'parameterised' by $v = (v_0, \ldots, v_k)$. By definition of the equivalence relation, multiplying the vector $(\exp(v_0), \ldots, \exp(v_k))$ by a positive constant does not change the equivalence class and hence the distribution. As an example the family can be viewed as

$$\langle(\pi_0, \pi_1 \exp(v_1 - v_0), \ldots, \pi_k \exp(v_k - v_0))\rangle$$

which is parameterised by the k-dimensional vector $(v_1 - v_0, \ldots, v_k - v_0)$. In terms of the standard theory of exponential families, this k-dimensional vector is a minimal representation while that given by $v = (v_0, \ldots, v_k)$ is non-minimal. In this example it is the mimimal representation, i.e. the natural parameters, which is a +1-affine parameterisation.

The following definition of the space of +1-affine parameters should be compared to the definition of the +1-affine space in Definition 3.

Definition 5 Define on V_{exp} (Defn. 3) the equivalence relation $\sim_v$ such that

$$\begin{aligned}(v, \mathcal{P}) \sim_v (v', \mathcal{P}') &\iff \mathcal{P} = \mathcal{P}' \text{ and } \exists \lambda > 0, \text{ s.t. } \exp(v_i) = \lambda \exp(v_i'), \forall i \in \mathcal{P} \\ &\iff \mathcal{P} = \mathcal{P}' \text{ and } \exists \mu, \text{ s.t. } v_i = v_i' + \mu, \forall i \in \mathcal{P}\end{aligned}$$

It is clear that if two points of V_{exp} are $\sim_v$ equivalent then they define the same distribution. So it is natural to define $V_{class} = V_{exp}/\sim_v$, the set of equivalence classes of $\sim_v$, as the space of +1-affine parameters.

The rest of this section shows that V_{class} is a cell complex, Lundell and Weingram (1969), which is homeomorphic to the standard simplex and so homeomorphic to the -1-geometry. A simplicial complex is a generalisation of a polyhedron and is constructed from basic building blocks called cells which are joined together according to certain rules which are given by a collection

of functions that encode the way the cells are assembled. A simplicial complex is a special case of cell complex so we will use interchangeably the two words. For completeness the definition of a cell complex is given explicitly here, taken from Lundell and Weingram (1969) page 6.

Definition 6 The pair (V, Φ) is a cell complex if V is a set and Φ a set of functions such that each element is a mapping of a closed ball in a Euclidean space to V which satisfies the following conditions

1. If $\phi \in \Phi$ and ϕ has domain S^n, the closed ball in $\mathbb{R}^n$, then ϕ is injective on $S^n - \partial S^n$.
2. The images $\{\phi(S^n - \partial S^n) | \phi \in \Phi\}$ partition V.
3. If $\phi \in \Phi$ has domain S^n then

$$\phi(\partial S^n) \subseteq \bigcup \{\psi(S^m - \partial S^m) | \psi \in \Phi \text{ with domain } S^m, m \leq n-1)\}.$$

Definition 7 A standard $k-1$-simplex in Euclidean space is

$$\Delta^{|\mathcal{S}|-1} := \left\{ \{\rho_i\}_{i \in \mathcal{S} \subseteq \mathcal{B}} | \sum_{i \in \mathcal{S}} \rho_i = 1, \rho_i \geq 0 \right\}.$$

Now define the map $\phi_{\mathcal{S}} : \Delta^{|\mathcal{S}|-1} \to V_{class}$ by

$$(\pi_i)_{i \in \mathcal{S}} \to \langle (\{\log \pi_i\}, \mathcal{Q}) \rangle_V$$

where $\mathcal{Q} \subseteq \mathcal{S} \subseteq \mathcal{B}$ is defined by $i \in \mathcal{Q}$ if and only if $\pi_i > 0$ for $i \in \mathcal{S}$, and $\langle x \rangle_V$ denotes the $\sim_v$-equivalence class containing x, noting that the support set $\mathcal{Q}$ is common in an equivalence class.

The following result shows that the space of +1-affine parameters is a cell complex.

Theorem 1 *For a given finite partition $\mathcal{B}$ of the real line with $k+1$ elements then*

$$(V_{class}, \{\phi_{\mathcal{S}} \mid \emptyset \neq \mathcal{S} \subseteq B\})$$

is a cell complex. Furthermore, restriction of each map to the relative interior of its domain is continuous and bijective, having well-defined inverse.

Proof See Appendix.

On this cell complex it is natural to define the weak topology, which 'glues' the lower dimensional sub-simplices in the natural way. The cells are then defined as a subset which is the image of $S^n - \partial S^n$ under one of the functions ϕ.

Definition 8 (Lundell and Weingram (1969) page 41) The weak topology on (V_{class}, Φ) is obtained by
(a) Giving each cell $\sigma \in V_{class}$ its quotient topology, that is a set is open in σ, if its preimage is open in the corresponding simplex.
(b) Giving V_{class} the weak topology with respect to all the subsets (cells) σ i.e., a set $F \subset V_{class}$ is closed if and only if $F \cap \sigma$ is closed in σ for each cell σ.

From the construction above this means that we give V_{class} the natural topology of the $|\mathcal{B}|$-simplex which defines the largest cell.

In the rest of the paper the the -1 cell complex of distributions on k bins will be denoted by S^k_{-1} and the $+1$ cell complex structure by S^k_{+1}. In (Lundell and Weingram (1969) page 27) is defined an isomorphism between two cell complexes. Following this definition the next result is immediate.

Theorem 2 *(i) The cell complexes S^k_{-1} and S^k_{+1} are isomorphic.*
(ii) The likelihood function is a continuous function on (V_{class}, Φ).

Simplicial complexes are attractive from the computational point of view because they always have a triangulation which gives them a combinatorial nature. This is the basis of the computational information geometry mentioned in the introduction.

3.3 Duality

One of the key aspects of Amari's information geometry is the relationship between the $+1$, -1 and 0-geometric structures via the concept called duality. Following Amari (1985) when the underlying geometric object is a manifold the relationship between the $+1$ and -1 connections, denoted by ∇^{+1} and ∇^{-1} and the Fisher information is captured in the duality relationship which can be written in terms of the inner product at θ, $\langle , \rangle_\theta$, and any vector fields X, Y, Z via the equation

$$X \langle Y, Z \rangle = \left\langle \nabla^{+1}_X Y, Z \right\rangle + \left\langle Y, \nabla^{-1}_X Z \right\rangle . \tag{1}$$

One consequence of this is relationship is the existence on exponential families of the so-called mixed parameterisation of the form (θ, μ), where θ is $+1$-affine, μ is -1-affine, their level sets being Fisher orthogonal across the manifold: see Barndorff-Nielsen and Blaesild (1983).

This section shows that, in fact, this relationship extends in a natural way across the simplex structure and, indeed, gives considerable insight into the nature of the $+1$-simplicial structure just defined (§3.2).

Definition 9 Let $\pi^0 = \left(\pi^0_0, \ldots, \pi^0_k\right)$ be a probability vector, $a_1, \ldots, a_d$ be a set of linearly independent vectors in $\mathbb{R}^{k+1}$, whose images in V_{exp} are linearly independent, and $b_1, \ldots, b_{k-d}$ be a set of linearly independent vectors in V_{mix} such that $a_i^T b_j = 0$ for $i = 1, \ldots, d$ and $j = 1, \ldots, k-d$. Furthermore, define

$$\bar{P}_{\pi^0} := \left\{ (\lambda, \sigma) : \left(p_{\pi^0}(\lambda, \sigma)\right)_h \geq 0 \text{ for all } h = 0, \ldots, k \right\},$$

where $\lambda \in \mathbb{R}^d$, $\sigma \in \mathbb{R}^{k-d}$ and

$$\left(p_{\pi^0}(\lambda, \sigma)\right)_h := \frac{\left(\pi^0_h + \sum_{j=1}^{k-d} (\sigma_j b_j)_h\right) \exp\{\sum_{i=1}^d (\lambda_i a_i)_h\}}{\sum_{h^*=0}^k \left\{\left(\pi^0_{h^*} + \sum_{j=1}^{k-d} (\sigma_j b_j)_{h^*}\right) \exp\{\sum_{i=1}^d (\lambda_i a_i)_{h^*}\}\right\}}. \tag{2}$$

Such a mapping $p_{\pi^0}(\cdot, \cdot)$ will be called a *simplicial parameterisation through* π^0.

The entire simplex, including the boundary, can be covered by using different simplicial parameterisations through different choices of π^0. Note therefore the distinction from a parameterisation of an open set of a manifold which is a diffeomorphism between open sets. Just like a manifold, a simplex can be covered by the union of images of parameterisations, the important difference being that, unlike the manifold case, the dimension may change from one parameterisation to another.

Note that for fixed $\sigma = \sigma^0$ the image of $p_{\pi^0}(\cdot, \sigma^0)$ is a d-dimensional exponential family. As σ^0 changes these exponential families are $+1$-parallel. However for fixed $\lambda = \lambda^0$, the image of $p_{\pi^0}(\lambda^0, \cdot)$ is not in general -1-affine, but is for the special case when $\lambda^0 = 0$. A similar structure is obtained using the mixed parameterisation (see Brown (1986), §3.10).

The simplicial parameterisation defined in Def. 9 is explicit, which is a property which will prove illuminating in what follows. Furthermore, the simplicial parameterisation describes the whole simplex including the boundary. Locally to a given π^0 in the interior of the simplex, the simplicial parameterisation is equivalent to the mixed parameterisation. Although the mixed parameterisation does not require to make reference to a point in the simplex the advantage of the simplicial over the mixed parameterisation is that it can be calculated explicitly, which is also a property which will prove illuminating in what follows.

The $+1$ simplicial structure is, clearly, much harder to visualise than the -1 since it has been constructed as a disjoint union of objects of differing dimensions, while the complete -1 object is just a subset of Euclidean space. The theorem below shows explicitly how the lower dimensional components arise as limits of families of co-dimension 1 in the interior of the simplex. The example which follows illustrates this in the trinomial case.

Theorem 3 *In the notation of Definition 9 choose $b \neq 0 \in V_{mix}$ and π^0 in the interior of the k-simplex. A foliation of the interior is defined by co-dimension 1 $+1$-affine subsets which are Fisher orthogonal to the -1-geodesic which passes through π^0 in the direction b. For any given choice of $\pi^0 \in \delta^k$, these fibres, the images of $p_{\pi^0}(\cdot; \sigma)$ are labelled by the scalar σ. In the topology of S^k_{+1} the limit of these fibres as σ increases (or decreases) exists and is an extended exponential family lying in a strictly lower dimensional components of S^k_{+1}.*

Proof See Appendix.

Example 4 An explicit example of the relationship between the different α-geometries is shown in Fig. 1, which illustrates the duality structure for the trinomial model. Panels (a) and (c) of Fig. 1 illustrate the -1-geometry of S^2_{-1} by showing a 2-simplex embedded in the plane such that straight lines segments in the simplex are exactly -1-geodesics. In panel (a) a set of these -1-geodesics is shown. In this figure, the full -1-simplicial structure can be seen, as all sub-simplices are part of it. The same is not true of the representation of

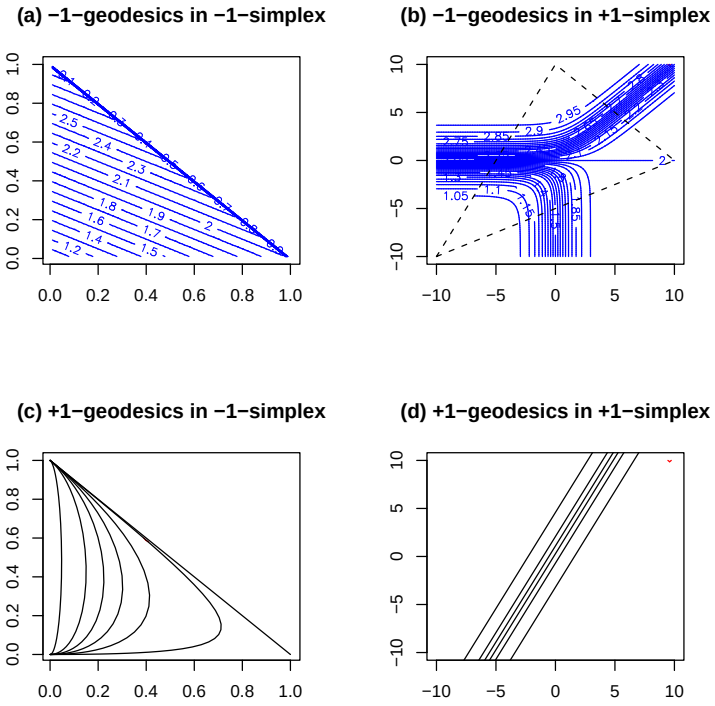

Fig. 1 Duality in the ±1-simplex structures. Panels (a) and (c) show the simplex in −1-affine parameters, while (b) and (d) shows it in +1-affine parameters

the +1-geometry shown in panels (b) and (d). Here straight lines are exactly +1-geodesics, but now the lower dimensional sub-simplices of the +1-simplicial structure are not represented in this figure, since they are "at infinity".

To see the way these +1-sub-simplices are connected "at infinity" requires an understanding of the dual structure. Using Fig. 1 this can be seen explictly, these results being shown in general in Theorem 3.

In this example, the vector $b^T = (1, 2, 3)$ was chosen and the parallel lines in panel (a) are the corresponding −1-parallel −1 geodesics, which are the level sets of the mean of $b^T n$, where $n = (n_0, n_1, n_2)$ is the trinomial random variable. These same lines are shown in panel (b), but now in the +1-parameterisation, and the symmetry of the simplex can be seen. The single line, labelled by the mean value 2, corresponds to the −1 geodesic which passes through the vertex in panel (a) at $(1, 0)$.

The +1-geodesics in panels (c) and (d) are defined by the foliation described in Theorem 3.3, and they are +1-parallel to each other, as can be seen visually in panel (d). What is not clear in that panel is what happens as these parallel lines go to infinity. This is made clear in panel (c). The limits of these +1-

geodesics are subsimplices. As the lines in panel (d) move to the left the corresponding lines in panel (c) converge to the vertical sub-simplex. On the other-hand as they move in (d) to the right the corresponding curve in panel (c) converges to the union of the exponential families given by the diagonal sub-simplex, the vertex at $(1, 0)$ and the horizontal subsimplex. Note that this is a continuous limit in the topology of the cell-complex, this being portayed by the dotted lines superimposed in panel (c). Specifically, the (0,0), (0,1) and (1,0) vertices in panel (a) correspond, respectively, to the lower left, mid-top and mid-right vertices of the dotted triangle in panel (b)

3.4 Summary of finite case

The information geometry of the finite dimensional case has the following properties, which will be reflected in the infinite case discussed below.

Firstly, the two affine coordinate systems correspond to unions of convex cells from a cell complex. Secondly, the corresponding tangent spaces form convex cones, and so the 'tangent space' is a union of convex cones. Finally, the way the cells are connected is defined implicitly by the dual relationship between the $+1$ and -1-affine structures.

4 Infinite dimensional spaces

This section examines what is the effect on the geometry of the space of distributions when there are an infinite number of bins in the discretisation. In the finite case the sub-simplices were defined by changing support sets. In the infinite case this still holds but the moment structure of the distributions, which was trivial in the finite case, now plays a dominant role in defining the geometry.

As described in Lundell and Weingram (1969), an infinite cell complex can be constructed inductively starting from a set of points and building by adding cells of progressively higher dimension, all of which are connected in a continuous way.

The 0-cells are given by the probability distributions degenerate on the bin indexed by $k \in \mathbb{Z}$ and the 1-cells are all (extended) binomial distributions on bins $i, j \in \mathbb{Z}$ with $i < j$. These distributions are parameterised by $[0, 1]$ to be 1-simplices. The 2-cells are added in the obvious way as (extended) trinomial distributions corresponding to the bins at $i, j, k \in \mathbb{Z}$ with $i < j < k$. The infinite cell complex is simply defined by induction giving the set

$$\Delta^{\infty} = \left\{ x \in \mathbb{R}^{\infty} | \sum x_i = 1, 0 \leq x_i \text{ and only a finitely many } 0 < x_i \right\},$$

i.e. the union of all finite supported distributions. This is, by definition, a cell complex and a -1-convex subset of the -1-affine space given in §2. Following Rizzolo and Su (2007), the set Δ^{∞} is not closed in the product topology of

countably many copies of $\mathbb{R}$, and hence in $(X_{mix}, V_{mix}, +)$. In fact its closure is

$$\Delta_0^\infty = \left\{ x \in \mathbb{R}^\infty | \sum x_i = 1, 0 \leq x_i \right\},$$

which is called the standard infinite dimensional simplex.

The issue of how the directions in the vector space can be extended in the space of distributions is explored in much more detail in the +1 case in the following section.

4.1 +1 Geometry: fixed support

Initially to understand the +1-structure consider the case where all distributions have a common support, i.e., assume for $\pi_i > 0$ all i.

Definition 10 Define the function $S(\cdot)$ by

$$S(\{v_i\}, \{\pi_i\}) := \sup_\theta \left\{ \theta | \sum \pi_i \exp(\theta v_i) < \infty \right\},$$

the function value being set to ∞ when the set is unbounded. Furthermore, for a given $\{\pi_i\}$ define the sets

$$V^{+1}(\{\pi_i\}) := \{\{v_i\} | S(\{v_i\}, \{\pi_i\}) > 0\},$$

and

$$L^{+1}(\{\pi_i\}) := \left\{ \{v_i\} | \pm \{v_i\} \in V^{+1}(\{\pi_i\}) \right\}.$$

Finally define the equivalence relation, $\sim_+$ on the set of all distributions by

$$\{\pi_i\} \sim_+ \{\pi_i'\} \iff V^{+1}(\{\pi_i\}) = V^{+1}(\{\pi_i'\}).$$

For each distribution, the set $V^{+1}(\{\pi_i\})$ plays the role of the tangent cone in the case of a cell complex in the finite dimensional case, with L^{+1} being its linear part.

Example 5 Consider the finite simplex defined by the probabilities $\pi_i, i = 0, 1, 2$. The tangent cone at $\pi_0 = 0, \pi_1 = 0.5, \pi_2 = 0.5$ is $\mathbb{R} \times \mathbb{R}^+$, with $L^{+1}(\{\pi_i\})$ being the subspace $\mathbb{R}$ which lies in the direction parallel to the edge.

The following theorems show that there are strong parallels between the deconstruction of the cell simplex structure of the space of +1-affine parameters discussed above and the space of equivalence classes of $\sim_+$.

Theorem 4 *For $\{\pi_i\}$ define a Hilbert space by*

$$H(\{\pi_i\}) := \left\{ \{v_i\} | \sum v_i^2 \pi_i < \infty \right\}$$

with inner product

$$\langle \{v_i\}, \{w_i\} \rangle_{\{\pi_i\}} = \sum v_i w_i \pi_i,$$

and corresponding norm $|| \cdot ||_{\{\pi_i\}}$. Under these conditions
(i) the set $V^{+1}(\{\pi_i\})$ is a convex cone, and
(ii) $L^{+1}(\{\pi_i\})$ is a subspace of $H(\{\pi_i\})$.

Proof See Appendix.

It is easy to show that not all vectors in +1-affine spaces that connect $\{\pi_i\}$ with any other distributions lies in $H(\{\pi_i\})$. However it is also of interest that vectors in the Hilbert space do come very close, in an ℓ^1 sense to all distributions.

The set $L^{+1}(\{\pi_i\})$ can be naturally mapped to the space of distributions via the following exponential map.

Definition 11 Let

$$\Sigma^+_{\{\pi_i\}} = \left\{ \{v_i\} \in L^{+1}(\{\pi_i\}) | \sum \pi_i \exp(v_i) < \infty \right\},$$

and

$$\partial\Sigma^+_{\{\pi_i\}} = \left\{ \{v_i\} \in L^{+1}(\{\pi_i\}) | \sum \pi_i \exp((1+\epsilon)v_i) = \infty \; \forall \epsilon > 0 \right\}.$$

Define the function $Exp : \Sigma^+_{\{\pi_i\}} \rightarrow (X_{exp}, V_{exp}, \oplus)$ via

$$Exp(\{v_i\}) = \langle \{\pi_i \exp(v_i)\} \rangle .$$

Let the image of this function be $Exp(\Sigma^+_{\{\pi_i\}}) := \tilde{V}^{+1}(\pi)$ and $Exp(\partial\Sigma^+_{\{\pi_i\}}) := \partial\tilde{V}^{+1}(\pi)$.

Theorem 5 *The set $\tilde{V}^{+1}(\{\pi_i\}) - \partial\tilde{V}^{+1}(\{\pi_i\}) \subset \Delta_0^\infty$ is the $\sim_+$ equivalence class containing $\{\pi_i\}$.*

Proof See Appendix.

Using this Theorem, it follows that the set of equivalence classes have a similar structure to that of a finite simplicial complex. Each of the classes has a structure which can be through of as an infinite dimensional exponential family, see Fukumizu (2005), and so the union is an infinite dimensional extended exponential family. Recall that the definition of a finite cell complex is given in Definition 9. The following theorem shows in part how the exponential map to the +1-affine space defines a structure which can be thought of as an infinite dimensional equivalent version of the simplicial complex.

Theorem 6 *(a)The exponential map is a function*

$$Exp : \Sigma_{\{\pi_i\}} \rightarrow \tilde{V}^{+1}(\{\pi_i\}),$$

and it is injective on the equivalence class $\tilde{V}^{+1}(\{\pi_i\}) - \partial\tilde{V}^{+1}(\{\pi_i\})$.
(b) These equivalence classes partition Δ_0^∞.

Proof See Appendix.

4.2 +1 Geometry: change in support

In order to understand the way that the images of these cones glue together consider first the possible limit points of a +1-geodesic , that is points of the form

$$\lim_{\theta \to S(\{v_i\},\{\pi_i\})} \langle \{\pi_i \exp(\theta v_i)\} \rangle .$$

There are the following cases to consider.

1. $S(\{v_i\},\{\pi_i\})$ and $\lim \sum \pi_i \exp(\theta v_i)$ are both finite. In this case the distribution
$$\{\bar{\pi}_i\} := \left\{ \frac{\pi_i \exp(\theta v_i)}{\sum_j \pi_j \exp(\theta v_j)} \right\}$$
has a different tangent cone structure structure since $\{v_i\} \notin V^{+1}(\{\bar{\pi}_i\})$. It follows from a similar argument to Theorem 5 (a) that if $\{\pi_i^*\} \in \partial \tilde{V}^{+1}(\{\pi_i\})$ then $V^{+1}(\{\pi_i^*\}) \subseteq V^{+1}(\{\pi_i\})$ but this cannot be an equality of sets since $\log(\pi_i^*) - \log(\pi_i) \notin V^{+1}(\{\pi_i^*\})$.
2. $S(\{v_i\},\{\pi_i\})$ is infinite but $\lim \sum \pi_i \exp(\theta v_i)$ is finite. This corresponds to the distribution $\{\bar{\pi}_i\}$ having a different support set than $\{\pi_i\}$ as long as at least two of the elements of v_i differ.
3. $S(\{v_i\},\{\pi_i\})$ is finite but $\lim \sum \pi_i \exp(\theta v_i)$ is infinite. Here the limit point lies in the same $(X_e, V_e, \oplus)$ affine space, but is no longer a finite measure. Such limit points do not lie inside any of the sets $\tilde{V}^{+1}(\{\pi_i\})$, by definition.
4. $S(\{v_i\},\{\pi_i\})$ and $\lim \sum \pi_i \exp(\theta v_i)$ are both infinite. Here both a change of support and a non-finite measure is possible, with the limit point lying in different +1 affine space than $\{\pi_i\}$.

In the finite case in order to complete the proof that the +1-geometry is that of a cell complex it would be required to show that each $\partial \tilde{V}(\{\pi_i\})$ lies in a union of images of lower dimensional equivalence classes. Since the equivalence classes are infinite dimensional this is clearly an in appropriate condition. Nevertheless the following result is immediate from the above discussion.

Theorem 7 *The boundary points of the equivalence classes which are distributions satisfy one of the following conditions.*

1. *The tangent cone of a distribution in $\partial \tilde{V}^{+1}(\{\pi_i\})$ is a subset of $V^{+1}(\{\pi_i\})$, or*
2. *The support of a distribution in $\partial \tilde{V}^{+1}(\{\pi_i\})$ is dominated by that of $\{\pi_i\}$.*

Thus the concept of the boundary lying in a lower dimensional component can be replaced by either a smaller support or a small set of random variables with finite moments.

4.3 -1 Geometry

The -1 structure is, as in the finite case, a little simpler than the +1-geometry. For brevity proofs which are very similar to those in the +1 case are omitted.

First consider the absolute convergence of vectors in V_{mix} defined in Def. 2. The following result shows that as far as the -1-structure on the convex set of distributions is concerned absolute convergence of the relevant vectors is automatic since if

$$v_i = \pi_i - \pi_i^* \implies |v_i| \leq \pi_i + \pi_i^*.$$

Hence $\sum |v_i| < \infty$. In the following therefore absolute convergence will be assumed for all vectors in V_{mix}.

First consider the definition of the tangent cone.

Definition 12 Define the functions $s(\cdot)$ by

$$s(\{v_i\}, \{\pi_i\}) := \sup_{\theta} \{\theta > 0 | \forall i, \pi_i + \theta v_i \geq 0\}.$$

Furthermore, for a given $\{\pi_i\}$ define the set

$$V^{-1}(\{\pi_i\}) := \{\{v_i\} | s(\{v_i\}, \{\pi_i\}) > 0\},$$

and

$$L^{-1}(\{\pi_i\}) := \left\{\{v_i\} | \pm \{v_i\} \in V^{-1}(\{\pi_i\})\right\}.$$

Finally define the equivalence relation, $\sim_-$ on the set of all distributions by

$$\{\pi_i\} \sim_- \{\pi_i'\} \iff V^{-1}(\{\pi_i\}) = V^{-1}(\{\pi_i'\}).$$

The following is then the equivalent of Theorem 4.

Theorem 8 *(i) $V^{-1}(\{\pi_i\})$ is a convex cone.*
(ii) If $\{v_i\} \in L^{-1}(\{\pi_i\})$, then the vector has finite Fisher information, i.e.

$$\sum_i \frac{v_i^2}{\pi_i} < \infty.$$

Note that the sum is over strictly positive probabilities since $\pi_i = 0 \implies v_i = 0$ when $\{v_i\} \in L^{-1}(\{\pi_i\})$.

Proof See Appendix.

Note that the fact that there are directions in the space of all distributions which have infinite Fisher information is far from pathological behaviour as the examples in Li et al. (2009) show.

The following definitions and Theorem are then analogous to Definition 11 and Theorem 5.

Definition 13 Let

$$\Sigma^-_{\{\pi_i\}} = \left\{ \{v_i\} \in L^{-1}(\{\pi_i\}) | \pi_i + (1+\epsilon) v_i \geq 0 \text{ for some } \epsilon > 0 \right\},$$

and $\partial\Sigma_{\{\pi_i\}}$ the subset where the bound, $\theta = s(\{v_i\}, \{\pi_i\})$, is attained.

Define the function $Exp : \Sigma^-_{\{\pi_i\}} \to (X_{mix}, V_{mix}, +)$ via

$$Exp(\{v_i\}) = \pi_i + v_i.$$

Let the image of this function be $Exp(\Sigma^-_{\{\pi_i\}}) := \tilde{V}^{-1}(\pi)$ and $Exp(\partial\Sigma^-_{\{\pi_i\}}) := \partial\tilde{V}^{-1}(\pi)$.

Theorem 9 *The set $\tilde{V}^{-1}(\{\pi_i\}) - \partial\tilde{V}^{+1}(\{\pi_i\}) \subset \Delta_0^\infty$ is the $\sim_-$ equivalence class containing $\{\pi_i\}$.*

Proof This is a simple variation of that of Theorem 5 and is omitted.

Thus the infinite dimensional -1 structure is very like that of the $+1$ case, being given by the following theorem which is analogous to the results of Theorems 6 and 7

Theorem 10 *(a)The exponential map is a function*

$$Exp : \Sigma^-_{\{\pi_i\}} \to \tilde{V}^{-1}(\{\pi_i\}),$$

and it is injective on the equivalence class $\tilde{V}^{-1}(\{\pi_i\}) - \partial\tilde{V}^{-1}(\{\pi_i\})$.
(b) The $\sim_-$-equivalence classes partition Δ^∞.
(c) The set $\partial\tilde{V}^{-1}(\{\pi_i\})$ lies in an equivalence class which is either (i) a subset of $V^{-1}(\{\pi_i\})$ or (ii) one where the support is strictly dominated by that of $\{\pi_i\}$.

4.4 Duality in infinite case

It remains to consider the equivalent of the results of §3.3 in the infinite case. The following definition is motivated by Definition 9

Definition 14 Let $\pi^0 = \left(\pi^0_i\right)$ be a probability vector, $b \neq 0$ an element of $L^{-1}(\{\pi^0\})$ and $a_i, i = 1, 2, \ldots$ an orthonormal basis of $(\frac{b_i}{\pi^0_i})^\perp$ in $L^{+1}(\{\pi^0\})$, which is a subset of $H(\{\pi^0\})$, such that $a_i^T b = 0$ for all i. Define the function $p_{\pi^0}(\cdot,\cdot)$ by

$$(p_{\pi^0}(\lambda,\sigma))_h := \frac{\left(\pi^0_h + \sigma b_h\right)\exp\{\sum_{i=1}^\infty (\lambda_i a_i)_h\}}{\sum_{h^*=0}^\infty \{(\pi^0_{h^*} + \sigma b_{h^*})\exp\{\sum_{i=1}^\infty (\lambda_i a_i)_{h^*}\}\}}.$$

Theorem 11 *In the notation of Definition 14 fixing σ defines a foliation of the interior of an equivalence class by co-dimension 1, $+1$-affine subsets which are Fisher orthogonal to b.*

Proof See Appendix.

5 Discussion

This paper looks at the information geometry of the space of all distributions. It is shown that this geometry is much closer to that of a simplex than a manifold. In the finite dimensional case it is shown that there is both a $+1$ and -1 simplicial complex which are connected by the duality which is familiar from the regular information geometric structure which requires a fixed support. In the infinite dimensional cases similar structures are also found. In these cases not only does the support of the set of distributions define the simplicial structure, but also the moment structure. Nevertheless much of the finite dimensional structure is preserved.

The paper looks at the geometry of the space of distributions arising from discretisation of the sample space. While this is clearly not the most general case, it does give an excellent foundation on which to construct a theory of *computational information geometry*. In many practical applications it can be argued that, since continuous data can only be measured to a finite accuracy, then this discretisation is sufficient for a complete analysis.

Appendix

Proof (Proof of Theorem 1)
Firstly, for a given $\phi_{\mathcal{P}}$ the set $\Delta^n - \partial\Delta^n$, for $n := |\mathcal{P}| - 1$, is given by

$$\left\{ \{\pi_i\}_{i\in\mathcal{P}} | \sum_{i\in\mathcal{P}} \pi_i = 1, \pi_i > 0 \right\}$$

and on this set the function is continuous and bijective, the map having well-defined inverse

$$\langle(\{v_i\},\mathcal{P})\rangle_V \to \exp(v_i)/\sum_{i\in\mathcal{P}} \exp(v_i).$$

Any point in V_{class} has the form $\langle(\{v_i\},\mathcal{P})\rangle_V$ which lies in the image of $\phi_{\mathcal{P}}(\Delta^{|\mathcal{P}|} - \partial\Delta^{|\mathcal{P}|})$, since these are bijections. Furthermore, if $\mathcal{P} \neq \mathcal{P}'$ then it is immediate that

$$\phi_{\mathcal{P}}(\Delta^{|\mathcal{P}|} - \partial\Delta^{|\mathcal{P}|}) \cap \phi_{\mathcal{P}'}(\Delta^{|\mathcal{P}'|} - \partial\Delta^{|\mathcal{P}'|}) = \emptyset,$$

as the corresponding distributions have different support.

Finally, consider $\phi_{\mathcal{P}}(\partial\Delta^n)$. For any point in the boundary define $\mathcal{Q} \subset \mathcal{P}$ to be the set of bins such that $\pi_i > 0$. This point will map to $\langle(\{\log(\pi_i)\},\mathcal{Q})\rangle_V$ which lies in the image of the interior of the cell $\phi_{\mathcal{Q}}(\Delta^{|\mathcal{Q}|} - \partial\Delta^{|\mathcal{Q}|})$, while $|\mathcal{Q}| \leq |\mathcal{P}| - 1$.

Hence all three conditions hold and we have a cell complex.

Proof (Proof of Theorem 3) For any choice of $b(\neq 0)$ and distribution in the interior of the simplex π^0, define

$$\mathcal{Z}(\sigma; b, \pi^0) := \left\{ i \in 0, \ldots, k : \pi_i^0 + \sigma b_i = 0 \right\}$$

and put

$$\bar{\sigma} = \bar{\sigma}(b, \pi^0) := \inf \left\{ \sigma \mid \mathcal{Z}(\sigma, b; \pi^0) \neq \emptyset \right\}.$$

For each fixed (π^0, σ) the image of p_{π^0} is a $k-1$-dimensional exponential family. Consider the limit of this family as $\sigma \to \bar{\sigma}$, in S^k_{-1}. From Equation 2 it is clear this has support in $P(\bar{\sigma})$ the complement of $\mathcal{Z}(\bar{\sigma})$.

Note that in this construct the limiting support set depends on the choice of π^0. However using the topology of the $+1$-simplicial structure defined in Theorem 1, which is isomorphic to that of the -1-structure by Theorem 2, each possible choice of support set is continuously attached. Hence the union of possible support sets is a well defined limit independent of π^0. This limit is clearly an extended exponential family.

The fact that all lower dimensional components of the simplex can be constructed using these limits is immediate via the form of Equation 2.

Proof (Proof of Theorem 4)

(i) It is immediate that $V^{+1}(\{\pi_i\})$ is a cone.

Convexity follows from the Cauchy-Schwartz inequality since for all $\{v_i\}, \{v_i^*\} \in V^{+1}(\{\pi_i\})$ and $\lambda \in [0,1]$ it follows that

$$\begin{aligned}\left\{\sum \pi_i e^{\frac{\theta}{2}(\lambda v_i + (1-\lambda)v_i^*)}\right\}^2 &= \left\{\sum \left(\sqrt{\pi_i} e^{\frac{\theta}{2}\lambda v_i}\right)\left(\sqrt{\pi_i} e^{\frac{\theta}{2}(1-\lambda)v_i^*}\right)\right\}^2 \\ &\leq \left\{\sum \pi_i e^{\theta \lambda v_i}\right\}\left\{\sum \pi_i e^{\theta(1-\lambda)v_i^*}\right\},\end{aligned}$$

and so is finite for a strictly positive value of θ.

(ii) First, if $\{v_i\} \in L^{+1}(\{\pi_i\})$, the moment generating function

$$\sum \exp(\theta v_i)\pi_i,$$

is defined for θ in an open set contain $\theta = 0$. Hence have both

$$\sum v_i \pi_i < \infty, \text{ and } \sum v_i^2 \pi_i < \infty.$$

Thus $\{v_i\} \in H(\{\pi_i\})$. The fact that it is a subspace follows from (i).

Proof (Proof of Theorem 5)

Note that by the definition of $\partial \tilde{V}^{+1}(\{\pi_i\})$ any element of $\tilde{V}^{+1}(\{\pi_i\}) - \partial \tilde{V}^{+1}(\{\pi_i\})$ corresponds to a probability distribution within the set of strictly positive measures.

First, it shall be shown that the $\sim_+$ equivalence class is a subset of $\tilde{V}^{+1}(\{\pi_i\}) - \partial \tilde{V}^{+1}(\{\pi_i\})$.

If $\{\pi_i^*\}$ lies in the same equivalence class as $\{\pi_i\}$ then it is required to show that

$$\{\pi_i^*\} \propto \{\pi_i \exp(v_i)\} \tag{3}$$

for some $\{v_i\} \in L^{+1}(\{\pi_i\})$ such that $1 < S(\{v_i\}, \{\pi\})$.

Consider the vector $v_i = \log(\pi_i^*) - \log(\pi_i)$. Since

$$\sum \exp(v_i)\pi_i = \sum \frac{\pi_i^*}{\pi_i}\pi_i = 1,$$

it follows that $\{v_i\} \in V^{+1}(\{\pi_i\})$. Reversing this argument gives $\{-v_i\} \in V^{+1}(\{\pi_i^*\})$, which means that $\{-v_i\} \in V^{+1}(\{\pi_i\})$ since the two distributions lie in the same equivalence class. This means that $\{v_i\} \in L^{+1}(\{\pi_i\})$ and it satisfies (3). Furthermore, since $\{v_i\} \in V^{+1}(\{\pi_i^*\})$, there exists $\epsilon > 0$ such that $\sum \pi_i^* \exp(\epsilon v_i) < \infty$. Hence $\sum_i \pi_i \exp((1+\epsilon)v_i) = \sum \pi_i^* \exp(\epsilon v_i) < \infty$, so that $1 < S(\{v_i\}, \{\pi_i\})$.

Conversely it is required to show that $\tilde{V}^{+1}(\{\pi_i\}) - \partial\tilde{V}^{+1}(\{\pi_i\})$ is a subset of the equivalence class of $\{\pi_i\}$.

Any distribution $\{\pi_i^*\} \in \tilde{V}^{+1}(\{\pi_i\}) - \partial\tilde{V}^{+1}(\{\pi_i\})$ must be proportional to $\pi_i \exp((1-\epsilon)w_i)$, where $\sum \pi \exp(w_i) < \infty$ and $0 < \epsilon < 1$. From the cone structure, if $\{v_i\} \in V^{+1}(\{\pi_i\})$ then $\epsilon\{v_i\} \in V^{+1}(\{\pi_i\})$, so by Theorem 4 it follows that

$$\{\epsilon v_i + (1-\epsilon)w_i\} \in V(\{\pi_i\}). \tag{4}$$

Hence

$$\sum \pi_i^* \exp(\epsilon v_i) = \sum \pi_i \exp((1-\epsilon)w_i) \exp(\epsilon v_i) = \sum \pi_i \exp((1-\epsilon)w_i + \epsilon v_i) < \infty$$

by Equation 4, so it must follow that

$$V^{+1}(\{\pi_i\}) \subseteq V^{+1}(\{\pi_i^*\}).$$

The reverse inclusion follows from the fact that $(1-\epsilon)(-w_i) \in V^{+1}(\{\pi_i^*\})$. So for any $v_i \in V^{+1}(\{\pi_i^*\})$, convexity gives

$$\sum \pi_i \exp(\epsilon v_i) = \sum \pi_i^* \exp((1-\epsilon)(-w_i)) \exp(\epsilon v_i) = \sum \pi_i^* \exp((1-\epsilon)(-w_i) + \epsilon v_i) < \infty$$

so

$$V^{+1}(\{\pi_i^*\}) \subseteq V^{+1}(\{\pi_i\}).$$

Proof (Proof of Theorem 6) (a) Suppose, for any two sequences $\{v_i\}$ and $\{v_i'\}$ in $\Sigma^+_{\{\pi_i\}}$, that for all i

$$\frac{\pi_i \exp(v_i)}{\sum_j \pi_i \exp(v_i)} = \frac{\pi_i \exp(v_i')}{\sum_j \pi_i \exp(v_i')}.$$

It immediately follows that there exists constants a, b such that for all i

$$v_i = av_i' + b.$$

As discussed above, vectors which differ by an additive constant are equivalent so, without loss of generality, we can set $b = 0$. Since the two vectors lie in the Hilbert space and are both of norm one then $a = 1$. Hence the function Exp is injective.

(b) By Theorem 5 these sets are equivalence classes, so they must partition the space.

Proof (Proof of Theorem 8) (i) Trivial

(ii) Let $\epsilon = \min\{S(\{v_i\},\{\pi_i\}), S(\{-v_i\},\{\pi_i\})\}$. Then since $\{v_i\} \in L^{-1}(\{\pi_i\})$, it follows that $\epsilon > 0$. Hence , have that

$$\pi_i + \epsilon v_i \geq 0 \implies \epsilon \leq \left|\frac{\pi_i}{v_i}\right| \quad \text{for } v_i < 0$$

$$\text{and } \pi_i - \epsilon v_i \geq 0 \implies \epsilon \leq \left|\frac{\pi_i}{v_i}\right| \quad \text{for } v_i > 0$$

Hence for all i where $\pi_i \neq 0$ have that

$$\left|\frac{v_i}{\pi_i}\right| \leq 1/\epsilon.$$

Thus the terms in the sum $\sum_i \frac{v_i^2}{\pi_i}$ are dominated by the absolutely convergent series $(1/\epsilon)v_i$ and so the sum is finite.

Proof (Proof of Theorem 11)

We need to show that $p_{\pi^0}(\sigma,\cdot)$ defines a foliation of the equivalence class. Since the underlying Hilbert space is separable there will exist separating hyperplanes and so it is sufficient to show that these hyperplanes are indexed by the parameter σ.

By Theorem 5 any element of the equivalence class can be written as proportional to

$$\pi_i^0 \exp\left(\tilde{\sigma}\frac{b_i}{\pi_i^0} + \sum \lambda_j a_{ij}\right),$$

since the exponent spans $L^{+1}(\{\pi_i^0\})$. It is then sufficient to show that all such points lie in a hyperplane which contains a distribution of the form $\{\pi_i^0 + \sigma b_i\}$. This can be done if we can solve the following equation up to an element of $(\frac{b_i}{\pi_i^0})^{\perp}$. Working in the $+1$ affine space it is required to find $\tilde{\sigma} = \tilde{\sigma}(\sigma)$ such that the difference

$$\tilde{\sigma}(\sigma)\frac{b_i}{\pi_i^0} - \log\left(1 + \sigma\frac{b_i}{\pi_i^0}\right),$$

is the orthogonal to b must satisfy

$$\tilde{\sigma}(\sigma)\sum \frac{b_i^2}{\pi_i^0} - \sum b_i \log\left(1 + \sigma\frac{b_i}{\pi_i^0}\right) = 0.$$

Differentiating with respect to σ gives the equation

$$\frac{d\tilde{\sigma}(\sigma)}{d\sigma}\sum \frac{b_i^2}{\pi_i^0} - \sum \frac{b_i^2}{\pi_i^0 + \sigma b_i} = 0$$

or

$$\frac{d\tilde{\sigma}(\sigma)}{d\sigma}||b||^2_{\pi_i^0} - ||b_i^2||_{\pi_i^0 + \sigma b_i} = 0$$

where the norms are with respect to the Fisher information, which by Theorem 8 are finite. Hence the function $\tilde{\sigma}(\sigma)$ is found by solving the corresponding differential equation.

References

Amari, S.-I. (1985). *Lecture notes in statistics-28: Differential-geometrical methods in statistics*. Springer-Verlag Inc.

Amari, S.-I. (1990). *Lecture notes in statistics-28: Differential-geometrical methods in statistics*. Springer-Verlag Inc.

Amari, S.-I. and Nagaoka, H. (2000). *Methods of information geometry*. American Mathematical Society.

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory*. John Wiley & Sons.

Barndorff-Nielsen, O. and Blaesild, P. (1983). Exponential models with affine dual foliations. *Annals of Statist.*, 11(3):753–769.

Brown, L. D. (1986). *Fundamentals of statistical exponential families: with applications in statistical decision theory*. Institute of Mathematical Statistics.

Cena, A. (2003). Geometric structures on the non-parametric statistical manifold. *PhD Thesis, University of Milan*.

Csiszar, I. and Matus (2005). Closures of exponential families. *The Annals of Probability*, 33(2):582–600.

Fukumizu, K. (2005). Infinite dimensional exponential families by reproducing kernel hilbert spaces. *Proceedings of the 2nd International Symposium on Information Geometry and its Applications*, pages 324–333.

Gibilisco, P. and Pistone, G. (1998). Connections on non-parametric statistical manifolds by orlicz space geometry. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, 1(2):325–347.

Grasselli, M. R. (2008). Dual connections in nonparametric classical information geometry. *to appear in the Annals of the Institute for Statistical Mathematics*.

Gzyl, H. and Recht, L. (2006a). A geometry on the space of probabilities i. the finite dimensional case. *Revista Matematica Iberoamericana*, 22(2):545–558.

Gzyl, H. and Recht, L. (2006b). A geometry on the space of probabilities ii. projective spaces and exponential families. *Revista Matematica Iberoamericana*, 22(3):833–849.

Gzyl, H. and Recht, L. (2007). Intrinsic geometry on the class of probability densities and exponential families. *Publicacions Matematiques*, 51(2):309–332.

Kass, R. E. and Vos, P. W. (1997). *Geometrical foundations of asymptotic inference*. John Wiley & Sons.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.

Li, P., Chen, J., and Marriott, P. (2009). Non-finite Fisher information and homogeneity: an EM approach. *Biometrika*, 96(2):411–426.

Lundell, A. and Weingram, S. (1969). *The Topology of CW Complexes*. Van Nostrand Reinhold: New York.

Marriott, P. (2002). On the local geometry of mixture models. *Biometrika*, 89(1):77–93.

Murray, M. K. and Rice, J. W. (1993). *Differential geometry and statistics*. Chapman & Hall Ltd.

Pistone, G. and Rogantin, M. P. (1999). The exponential statistical manifold: Mean parameters, orthogonality and space transformations. *Bernoulli*, 5:721–760.

Pistone, G. and Sempi, C. (1995). An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The Annals of Statistics*, 23:1543–1561.

Pitman, E. (1979). *Some basic theory for statistical inference*. Chapman and Hall, London.

Rinaldo, A. (2006). On maximum likelihood estimation in log-linear models. *Tech. Rep. Dep. of Statistics, Carnegie Mellon University*.

Rizzolo, D. and Su, F. (2007). A fixed point theorem for the infinite-dimensional simplex. *J. of Mathematical Analysis and Applications*, 332(2):1063–1070.