entropy

MDPI

Article

On the Limiting Behaviour of the Fundamental Geodesics of Information Geometry

Frank Critchley [1] and Paul Marriott [2,*]

[1] Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, Buckinghamshire MK7 6AA, UK; f.critchley@open.ac.uk

[2] Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 2G1, Canada

* Correspondence: pmarriot@uwaterloo.ca; Tel.: +1-519-888-4567

Received: 13 July 2017; Accepted: 28 September 2017; Published: 30 September 2017

Abstract: The Information Geometry of extended exponential families has received much recent attention in a variety of important applications, notably categorical data analysis, graphical modelling and, more specifically, log-linear modelling. The essential geometry here comes from the closure of an exponential family in a high-dimensional simplex. In parallel, there has been a great deal of interest in the purely Fisher Riemannian structure of (extended) exponential families, most especially in the Markov chain Monte Carlo literature. These parallel developments raise challenges, addressed here, at a variety of levels: both theoretical and practical—relatedly, conceptual and methodological. Centrally to this endeavour, this paper makes explicit the underlying geometry of these two areas via an analysis of the limiting behaviour of the fundamental geodesics of Information Geometry, these being Amari's (+1) and (0)-geodesics, respectively. Overall, a substantially more complete account of the Information Geometry of extended exponential families is provided than has hitherto been the case. We illustrate the importance and benefits of this novel formulation through applications.

Keywords: extended exponential families; information geometry; Riemannian Markov Chain Monte Carlo

1. Introduction

Information Geometry has developed enormously, both theoretically and in its range of applications, since the seminal works of [1–3]. Excellent summaries of this approach, which we shall call classical, can be found in [4], and recently [5]. This approach has the property that the fundamental geometric objects are smooth manifolds. In particular, they are open sets, of constant dimension. However, there has been recent interest in studying the Information Geometry of closures of exponential families, as defined in [6]: these closures typically being unions of manifolds of varying dimension. As discussed in [7], this development gives a more exact duality between sample and model space, which is the key to the intrinsic duality of Information Geometry. From an applications' point of view, studying closures of statistical manifolds is very natural in categorical data analysis [8,9] and graphical [10], random graph [11], and log-linear [12] models. A strongly related approach, which gives an excellent treatment of the closure of statistical models, uses algebraic geometry. See, for example, [13,14].

This paper focuses on extending the manifold-based approach of classical Information Geometry by looking at the limiting behaviour of key objects: (α)-geodesics, where we follow the standard notation in information geometry where $\alpha = +1$ is the exponential representation and $\alpha = 0$ is the Fisher/spherical representation of the manifold. To be precise, we note that we use the term "limiting" here to denote the behaviour of a geodesic as it approaches the boundary of the closure of an exponential family. This may mean that a natural parameter tends to infinity, in the case of

a $(+1)$-geodesic, or a path-length parameter tends to a finite value, in the case of the (0)-geodesics. We look specifically at the finite, discrete case and study two key types of geodesics $\alpha = +1$ and 0. In particular, we show how these two types of geodesics have fundamentally different boundary behaviours. By studying the first of these, we construct an explicit representation of the limiting behaviour of finite dimensional exponential families. The behaviour of the second was introduced in our recent paper [15], which studied applications involving Markov chain Monte Carlo (MCMC) methods. This paper gives the theoretical foundations to the MCMC applications found in the early work. It also extends the results on (0)-geodesics found there to show both the asymptotic and limiting behaviour of $(+1)$-geodesics. We begin with an insightful example.

Example 1. *The extended trinomial model.*

Figure 1 shows, in the simple case of the extended trinomial model, the behaviour of key geodesics. The model is plotted in a mean, or (-1)-affine, parameterisation

$$\left\{ (\pi_0, \pi_1, \pi_2) | \sum_{i=0}^{2} \pi_i = 1, \pi_i \geq 0 \right\},$$

where boundaries are completely explicit, being the points where at least one $\pi_i = 0$. In this figure, three geodesics, passing through the same point and having the same initial tangent vector, have been computed. The (-1)-geodesic in this parametrisation is, of course, a straight line and this cuts the boundary of the extended family (see panel (a)). The (0)-geodesic, panel (b), smoothly touches the boundary. We show that this is generic behaviour. We note here that the closed loop nature of the (0)-geodesic is not generic extended exponential family behaviour. Instead, as we explain in Section 3, it reflects something quite specific about the multinomial distribution. The $(+1)$-geodesic, panel (c), reaches the boundary at a vertex which, as we also show, is generic behaviour. Furthermore, it approaches the vertex close to one of the edges of the simplex. In fact, it does this exponentially fast. Again, we show that this behaviour is quite general.

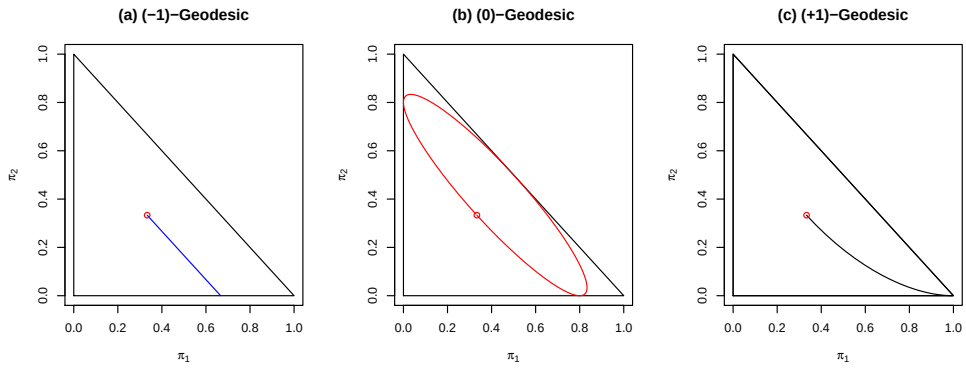

Figure 1. Key geodesics in the extended trinomial model.

The rest of the paper is organised as follows. Section 2 looks at the limiting behaviour of $(+1)$-geodesics. This allows us to explicitly characterise the—sometimes subtle and surprising—boundary behaviour of general discrete exponential families. The results clearly illustrate the differences between the open-set, manifold-based classical information geometry and the geometry required to take into account the boundaries that naturally occur in categorical data analysis. Section 3 looks at the limiting behaviour of Fisher or (0)-geodesics. We show how the boundary behaviour of these geodesics allows them to be used as tools that have important applications. These include designing efficient MCMC algorithms and solving optimisation problems on the closures of exponential families. Throughout, we illustrate our results visually with simple but representative examples.

2. Limits of (+1)-Geodesics

This paper looks at general finite exponential families, as used in categorical data analysis, graphical modelling, random discrete graph models, and log-linear modelling. Each of these models can be embedded in a sufficiently high dimensional closed simplex.

The key intuition behind the behaviour of $(+1)$-geodesics is that they are normalised exponentials of linear functions (see Definition 1). Hence, in the limit, their behaviour is determined by the maximisers of these linear functions. The structure of these maximizers is further determined by the polar duality of the support set of the exponential model. For illustrative examples, see Figures 2 and 3, and, in Theorem 1, we give explicit asymptotic expressions for this behaviour.

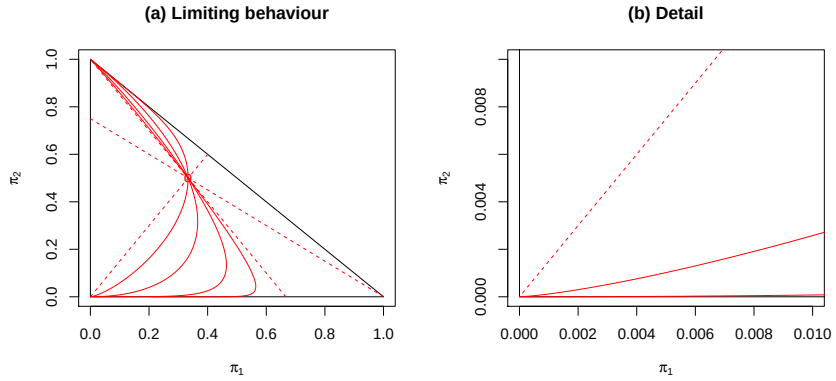

Figure 2. A set of $(+1)$-geodesics in the extended trinomial model.

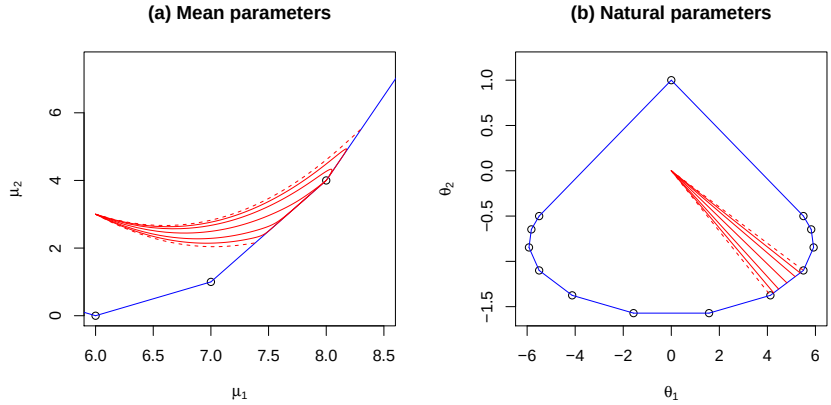

Figure 3. A set of $(+1)$-geodesics in Altham's model.

2.1. Notation

We start with some notational issues. Define

$$\Delta^k := \left\{ \pi | \sum_{i=0}^{k} \pi_i = 1, \pi_i \geq 0 \right\}, \text{ and } S_+^k := \left\{ \xi | \sum_{i=0}^{k} \xi_i^2 = 1, \xi_i \geq 0 \right\},$$

where $k \geq 1$ is the dimension of the simplex. Let $\mathcal{I} := \{0, ..., k\}$ be the labels that we associate with the vertices of Δ^k. In a convenient mild abuse of notation, identify any proper—i.e., not the relative interior (r.i.) of Δ^k—face of Δ^k with the set $\varnothing \subset \mathcal{F} \subset \mathcal{I}$ of vertices spanning it—i.e., $\{i \in \mathcal{I} : \pi_i > 0\}$, or, equivalently, with the *complementary* set $\varnothing \subset \mathcal{F}^C \subset \mathcal{I}$—i.e., $\{i \in \mathcal{I} : \pi_i = 0\}$.

For any $m \geq 1$, let $\underline{1}_m$ denote the vector of m 1s, $\mathbb{C}_m := \{\underline{c} \in \mathbb{R}^m : \underline{1}_m^T \underline{c} = 0\}$ the $(m-1)$-dimensional subspace of all centred (i.e., zero sum) vectors, $C_m := I_m - m^{-1} \underline{1}_m \underline{1}_m^T$ the (Euclidean) orthogonal projector of $\mathbb{R}^m$ onto $\mathbb{C}_m$, and put $\mathcal{Q} := \{\text{the unit Euclidean sphere in } \mathbb{R}^p\} \equiv \{q \in \mathbb{R}^p : q^T q = 1\}$.

Using this notation, we can define a p-dimensional full exponential family in $r.i.(\Delta^k)$ as follows.

Definition 1. *For some $\underline{\pi}^0 = (\pi_i^0) \in r.i.(\Delta^k)$ and V a $(k+1) \times p$ matrix defined by*

$$V = (\underline{v}^{(1)}|...|\underline{v}^{(p)}) = (\underline{v}_0|...|\underline{v}_k)^T$$

with linearly independent columns, where $\underline{1} \notin \text{Range}(V)$, the (-1)-image of an exponential family $\mathcal{M} = \mathcal{M}(\underline{\pi}^0, V)$ comprises $\{\underline{\pi}(\underline{\theta}) = (\pi_i(\underline{\theta})) : \underline{\theta} \in \mathbb{R}^p\}$ where $\underline{\pi}(\underline{\theta}) = \underline{\pi}_{(\underline{\pi}^0, V)}(\underline{\theta})$ has general element:

$$\pi_i(\underline{\theta}) = \pi_i^0 \exp\{(V\underline{\theta})_i - \psi(\underline{\theta})\} \equiv \pi_i^0 \exp\{\underline{v}_i^T\underline{\theta} - \psi(\underline{\theta})\}, \tag{1}$$

$i \in \mathcal{I}$, where

$$\exp(\psi(\underline{\theta})) \equiv \textstyle\sum_{i\in\mathcal{I}} \pi_i^0 \exp((V\underline{\theta})_i) = \sum_{i\in\mathcal{I}} \pi_i^0 \exp(\underline{v}_i^T\underline{\theta}).$$

Without loss of generality, we take each column of V to be centred and take the columns of V to be (Euclidean) orthonormal—i.e., $V^TV = I_p$.

Since these exponential families are $(+1)$-affine sets, the geometry of all one dimensional affine subsets—i.e., $(+1)$-geodesics—determines the underlying geometry. Thus, for each $q \in \mathcal{Q}$, define

$$v(q) = (v_i(q)) := Vq,$$

a centred unit vector in $\mathbb{R}^{k+1}$. The set $\mathcal{M}_q$, comprising all $\pi(\theta_q) = (\pi_i(\theta_q))$, $\theta_q \in \mathbb{R}$, with

$$\pi_i(\theta_q) = \pi_i^0 \exp\{v_i(q)\theta_q - \psi(\theta_q)\}, \tag{2}$$

$i \in \mathcal{I}$, is a one-dimensional exponential sub-family of $\mathcal{M}$. Indeed, $\mathcal{M}_q$ is a $(+1)$-geodesic in both $\mathcal{M}$ and $r.i.(\Delta^k)$. As q varies over $\mathcal{Q}$, we get all such $(+1)$-geodesics in $\mathcal{M}$, and the strategy of this section is to carefully analyse the boundary behaviour of each $\mathcal{M}_q$.

2.2. Limits at the Boundary

In [15], we gave an explicit representation of how the p-dimensional model (1) is attached to the boundary of Δ^k. The key idea was to analyse the polar dual of the convex hull of the columns of V. This convex hull defines the extremal points of the mean parameters, and its polar dual determines the directions of recession [12]. These are the directions in the natural parameter space that attain these extreme points. Here, we look much more explicitly at the way that these limits are attained.

Dropping the q notationally from (2), consider now the $(+1)$-geodesic $\mathcal{M}_{+1} = \mathcal{M}_{+1}(\underline{\pi}^0, \underline{v})$ defined, for given $\underline{\pi}^0 = (\pi_i^0) \in r.i.(\Delta^k)$ and centred unit vector $\underline{v} = (v_i) \in \mathbb{R}^{k+1}$, by:

$$\pi_i(\theta) = \pi_i^0 \exp\{v_i\theta - \psi(\theta)\}, \tag{3}$$

$i \in \mathcal{I}$, $\theta \in \mathbb{R}$. We seek the limit points of $\mathcal{M}_{+1}$—and other related quantities—as $\theta \to \pm\infty$. Since $-\underline{v}$ is also a centred unit vector, while $\theta\underline{v} = (-\theta)(-\underline{v})$, it is enough to consider the case $\theta \to +\infty$.

Definition 2. *Let $\{v_i\}_{i\in\mathcal{I}}$ take distinct values $v_{(0)} > \cdots > v_{(g)}$, with $v_{(j)}$ having multiplicity $m_j \geq 1$, so that $\sum_{j=0}^{g} m_j = k+1$. Without loss, relabel the bins so that: the elements of $\underline{v}$ are in non-increasing order and, for each j, the m_j corresponding values of π_i^0 are also in non-increasing order. Then, we may replace the single index "i" by a double index "(j,r)", thus:*

$$\underline{v} = (v_i) = \begin{pmatrix} \vdots \\ v_{(j)}\underline{1}_{m_j} \\ \vdots \end{pmatrix} \text{ and, correspondingly, } \underline{\pi}^0 = (\pi_i^0) = \begin{pmatrix} \vdots \\ \underline{\pi}_{(j)}^0 \\ \vdots \end{pmatrix},$$

where $\underline{\pi}^0_{(j)} = (\pi^0_{(j,r)})$, $j = 0, ..., g$, $r = 1, ..., m_j$.

Definition 3. *Using this double index, we can define some key notation, the requirement* $j > 0$ *being implicit in all terms but the first:*

$$\begin{aligned}
\pi^0_{j\cdot} &:= \textstyle\sum_{r=1}^{m_j} \pi^0_{(j,r)}, \quad \kappa_j := \pi^0_{j\cdot}/\pi^0_{0\cdot} > 0, \quad \rho_j := v_{(j)}/v_{(0)} \neq 1, \\
\delta_j &:= v_{(j-1)} - v_{(j)} > 0, \quad \delta_{j\cdot} := \delta_1 + ... + \delta_j = v_{(0)} - v_{(j)} > 0, \\
\epsilon(\theta) &:= \exp(-\theta\delta_1), \quad \epsilon_j(\theta) := \exp(-\theta\delta_j), \quad \epsilon_{j\cdot}(\theta) := \exp(-\theta\delta_{j\cdot}) = \textstyle\prod_{h=1}^{j} \epsilon_h(\theta).
\end{aligned}$$

Before giving the main results of this section, we make some comments on these terms. Since all terms $\delta_j > 0$, each $\epsilon_j(\theta)$—in particular, $\epsilon(\theta)$—tends to zero exponentially fast as $\theta \to \infty$, and we will compute first order expansions in these terms. While these are "first order", we emphasise the exponentially fast convergence noted in Example 1.

We look at the limiting behaviour of key geometric terms: probabilities, tangent vectors, and the Fisher information as $\theta \to \infty$. We comment that since we are working in the closure of an exponential family, we cannot assume that the usual, open set based, geometric intuition holds. Thus, for example, even the existence of tangent vectors, and their transformation rules, need careful checking.

Theorem 1. *With* $1 \leq r_0 \leq m_0$ *defining a bin* $(0, r_0)$ *and for* $1 \leq r_j \leq m_j$, *we have the following asymptotic expansions as* $\theta \to +\infty$.

(i) For the probabilities, we have:

$$\pi_{(0,r_0)}(\theta) = \frac{\pi^0_{(0,r_0)}}{\pi^0_{0\cdot}} \times \{1 - \kappa_1\epsilon(\theta) + o(\epsilon(\theta))\}, \tag{4}$$

$$\pi_{(1,r_1)}(\theta) = \left[\frac{\pi^0_{(1,r_1)}}{\pi^0_{1\cdot}} \times \kappa_1\epsilon(\theta)\right] + o(\epsilon(\theta)), \tag{5}$$

$$\pi_{(j,r_j)}(\theta) = o(\epsilon(\theta)), \quad j > 1. \tag{6}$$

(ii) The mean parameter has the expansion:

$$\mu(\theta) := \psi'(\theta) = v_{(0)} - \kappa_1[v_{(0)} - v_{(1)}]\epsilon(\theta) + o(\epsilon(\theta)).$$

(iii) The Fisher information has the expansion:

$$\psi''(\theta) = \kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta)). \tag{7}$$

(iv) Finally, for tangent vectors, with respect to μ, *we have the expansions:*

$$\frac{\partial\pi_{(0,r_0)}}{\partial\mu}(\theta) = \frac{\pi^0_{(0,r_0)}}{\pi^0_{0\cdot}} \times \frac{1}{[v_{(0)} - v_{(1)}]} + o(\epsilon(\theta)), \tag{8}$$

$$\frac{\partial\pi_{(1,r_1)}}{\partial\mu}(\theta) = -\frac{\pi^0_{(1,r_1)}}{\pi^0_{1\cdot}} \times \frac{1}{[v_{(0)} - v_{(1)}]} + o(\epsilon(\theta)), \tag{9}$$

$$\frac{\partial\pi_{(j,r_j)}}{\partial\mu}(\theta) = o(\epsilon(\theta)) \quad j > 1. \tag{10}$$

Proof. See Appendix A. □

Corollary 1. *The set of limit points of a p-dimensional exponential family in r.i.*(Δ^k) *is a finite union of exponential families, each lying in its own specific proper face of* Δ^k.

Proof. The limit points in Theorem 1 are functions of an initial point $\underline{\pi}^{(0)}$ and an initial direction $\underline{v}$. However, the support set—denoted by $\mathcal{I}(\underline{v})$—of the limit points is purely a function of $\underline{v}$. Furthermore, since the initial point can be anywhere in the p-dimensional exponential family, it can be written as having general element

$$\pi_i^0 \exp\{\underline{v}_i^T \underline{\phi} - \psi(\underline{\phi})\}.$$

The corresponding limit points have general elements positively proportional to

$$\begin{cases} \pi_i^0 \exp\{\underline{v}_i^T \underline{\phi}\}, & i \in \mathcal{I}(\underline{v}), \\ 0, & i \notin \mathcal{I}(\underline{v}), \end{cases}$$

which is an exponential family with support $\mathcal{I}(\underline{v})$. It is not, of course, necessarily in the minimal form since the columns defining V, once restricted to the subset $\mathcal{I}(\underline{v})$, need not be linearly independent. □

It is important to note that for a fixed $\{\pi_i^0\}$ the set of limit points of $(+1)$-geodesics of the form given by Equation (3) does not form the complete closure of a statistical model (see Example 3 for an illustration of this fact).

Corollary 2. *We have that*

$$\sum_{i=0}^{k} \frac{\partial \pi_i}{\partial \mu}(\theta) \equiv \sum_{r_j=1}^{m_j} \frac{\partial \pi_{(j,r_j)}}{\partial \mu}(\theta) = 0 + o(\epsilon(\theta)), \tag{11}$$

which is consistent with the fact that $\partial \underline{\pi}(\theta)/\partial \mu$ *is a tangent vector in* (-1)*-coordinates and, hence, is centred.*

Proof. By direct calculation. □

Example 2. *Extended Trinomial Model.*

We return to the extended trinomial model in order to visualise and interpret the results of Theorem 1 and its corollaries. In Figure 2a, we select a fixed $\underline{\pi}^0$ *and a number of different unit vectors, q, to define a set of* $(+1)$*-exponential families. For each value of q, we compute the corresponding double index. The "generic" case has* $g = 2$ *and* $m_0 = m_1 = m_2 = 1$*. That is, the vector has no ties, and thus has a unique maximum and minimum value. These cases are plotted with a solid line in the figure. We see that these all converge to a vertex, exponentially approaching one of the edges, as predicted. The process of convergence is emphasised in panel (b), showing the convergence in detail.*

There are two non-generic cases, where there is a tie for the maximum, or the minimum value. Note that, since the vector is centred and non-zero, all three values cannot be the same. In this case, we have $g = 1$ *and either* $m_0 = 2, m_1 = 1$ *or* $m_0 = 1, m_1 = 2$*. As the theorem shows, the limit, in any such case, lies on the face spanned by the two largest (smallest) values, with the other limit point being a vertex. The position of the point on the edge is determined by the expression* $\frac{\pi^0_{(0,r_0)}}{\pi^0_{0\cdot}}$*. These geodesics are plotted with a dashed line in the figure. We note that, in this special case, these* $(+1)$*-geodesics are also* (-1)*-geodesics.*

When we look at the set of tangent vectors, we see behaviour considerably at variance with what would be expected in a manifold-based setting. As mentioned above, since we are not working in open sets, care is needed in checking even standard properties of tangent vectors. First, we note that the set of tangent vectors to $(+1)$*-geodesic, which meet at a vertex, has a conal, rather than a vector space structure. In addition, if we consider the "generic" case where* $g = 2$*, then, from Theorem 1 (iv), we have all limiting tangent vectors that are parallel to (a permutation of)*

$$(1, -1, 0)\,,$$

for all such corresponding $(+1)$*-geodesics. These are plotted with solid lines in the figure, and a close-up of the local behaviour; panel (b) shows clearly that all tangent vectors have the same limit. Note that this means that the exponential map, which maps tangent vectors to points in a manifold, cannot be uniquely defined at a*

boundary point. The fact that the set of limiting directions at a vertex is a (closed) cone comes not, principally, from the "generic" case, but, rather, from the case where $g = 1$. In the figure, one such geodesic converging to $(0, 0)$ is shown with a dashed line. However, for this case, the limiting tangent direction depends on $\underline{\pi}^0$, so all values in the relative interior of the cone can be attained, while it is the boundary directions of the cone that come from the generic case.

Example 3. *Altham's Model.*

A two-dimensional extension of the binomial family, as described by Altham in [16], is given by

$$\binom{k}{y} exp(y\eta + T(y)\phi - \psi(\eta, \phi)) \in \Delta^k, \tag{12}$$

where we take $T(y) = (y - \bar{y})^2$, $y = 0, \cdots, k$. This allows both over and under-dispersion relative to the binomial model and for large k can be thought of as a finite, discrete approximation to the normal model (see [15] for more details).

Figure 3 shows, for $k = 12$, some of the details of the boundary convergence of $(+1)$-geodesics that define this family. Panel (a) is shown in the mean, or (-1)-affine, parameterisation. The solid lines are "generic" $(+1)$-geodesics that converge to a vertex, as predicted by Theorem 1. As can be seen by close inspection, all of these geodesics have a tangent vector that is parallel to the corresponding edge. The dashed lines correspond to the case where there is a tie in the largest component of the initial direction of the geodesic.

Panel (b) of Figure 3 shows the same geodesics in the natural, or $(+1)$, parameters. Here, the dual polytope is shown as the convex hull of a set of vertices, each of which corresponds to a "direction of recession" (for details, see [12] or [15]). The polar duality can be clearly seen, with all $(+1)$-geodesics cutting an edge in (b) intersecting the corresponding vertex in panel (a), while the dashed lines hit vertices in (b), intersecting edges in (a).

3. Fisher Geodesics

We turn now to the Fisher, or (0), geodesic in an exponential family embedded in a finite simplex. For completeness, we recall that the 0-representation maps the simplex to the sphere, and the Fisher metric is the pullback of the standard metric on the sphere. Here, we shall define a new class of geometric object—the extended Fisher geodesic—which lies naturally in the extended exponential family.

In an exponential family, Fisher geodesics are the geodesics of the Levi–Civita connection and have the property of being (local) minimisers of path length and energy [17]. They were one of the first differential geometric objects studied in statistics [1]. In general, they cannot be computed in closed form, except in a few special cases, but can be computed numerically using their defining differential equations. Since these equations need to be defined on open sets, the analysis here is required to understand their limiting behaviour in the closure.

In Figure 1b, we see a Fisher geodesic in the extended trinomial model. This is a case where there is a closed form (see [1], p. 32). It can be directly calculated in the (0)-representation of the simplex, given by $\xi_i = \sqrt{\pi_i}$, $i = 0, 1, \ldots, k$. The image of Fisher geodesic connecting ξ and $\tilde{\xi}$ is the set of points of the form

$$\xi_i(t) = c(t) \left(\xi_i + t(\tilde{\xi}_i - \xi_i)\right), \tag{13}$$

where $i = 0, \cdots, k$, and $c(t)$ is the positive normalising constant, which ensures $\sum_{i=0}^{k} \xi_i(t)^2 = 1$. Since we are working in the extended multinomial model, there is no constraint on the positivity of $\xi_i(t)$ and the figure shows the image of the full great circle in the sphere, which is the Fisher geodesic in the (0)-representation. We can alternatively think of it as the union of Fisher geodesics in the relative interior, which is smoothly connected at the boundary. It is this smooth touching of the boundary that motivated the results in this section. In fact, the curve in Figure 1b was computed by solving the underlying differential equation numerically using the methods of Section 3.2 below. The local solution is guaranteed to exist in open neighbourhoods, but the numerical solution was extended

smoothly into, and out of, the boundary. The main result of this section shows, both theoretically and numerically, that this a general property of extended Fisher geodesics in extended exponential families in the simplex.

We note that the way that the (0)-geodesic smoothly intersects with the boundary in Figure 1b is generic in all exponential families, as shown in Theorem 6. However, for clarity, we also note that the closed nature of the (0)-geodesic, seen in the figure, is a special property of multinomial models. This follows since they are equivalent to standard spheres under their Fisher Reimannian structure. Hence, in this special case, the extended geodesics are the images of great circles and hence closed. In general, as illustrated in Example 3, the geodesics do not form closed loops.

3.1. The Fisher Geodesic and the Boundary

In order to define extended Fisher geodesics, we need to consider how to measure the energy of a curve in an extended model. In particular, we need to understand the energy of a path whose limit lies in the boundary, as seen in Figure 1b. The following result on how the Fisher information behaves near the boundary follows from results in [18] and was stated in [15]. It shows the singularity of the metric, in both the mean and natural parameters at the boundary. The importance of this result is to emphasise that standard Riemannian geometry does not extend directly to the boundary of the extended exponential family.

Theorem 2. *(a) Let $\{\mu_i\}$ be a sequence of points in the mean parameter space of an exponential family, lying in r.i.(Δ^k), which converge to $\overline{\mu}$, which lies on a face of the boundary polytope, defined by the half space, characterised by an equation of the form $\langle a, \mu \rangle \leq 1$, for a unit normal vector a.*

Let $I(\mu)$ be the Fisher information, $\lambda_{\min}(\mu)$ its minimum eigenvalue, assumed simple, and $e_{\min}(\mu)$ a corresponding unit eigenvector, and unique up to overall sign. Then,

$$\lim_{i \to \infty} \lambda_{\min}(\mu_i) = 0$$

and $\lim_{i\to\infty} e_{\min}(\mu_i) = a$.

(b) Let $\{\theta_i\}$ be the corresponding sequence to μ_i in the natural parameters, $I(\theta) := I(\mu(\theta))^{-1}$ the Fisher information, with $\lambda_{\max}(\theta)$ its maximum eigenvalue, assumed simple, $e_{\max}(\mu)$ a corresponding unit eigenvector, unique up to overall sign. Then,

$$\lim_{i \to \infty} \lambda_{\max}(\theta_i) = 0$$

and $\lim_{i\to\infty} e_{\max}(\theta) = a$, which is the vertex in the polar which corresponds to the face in (a).

From the proof of Corollary 1, we have that the closure, $\overline{\mathcal{M}}$, of an exponential family $\mathcal{M} \subset r.i.(\Delta^k)$ can be written explicitly as a finite union of exponential families each lying in its own, proper, face. We first define what it means for a curve to be smooth in the closure.

Definition 4. *A (-1)-representation of a curve in the closure, $\overline{\mathcal{M}}$, of an exponential family in $\mathcal{M} \subset r.i.(\Delta^k)$ is*

$$\gamma : [0, 1) \to \overline{\mathcal{M}} \subset \Delta^k.$$

It is defined to be S-smooth in the closure when it can be partitioned into the union of smooth subpaths each in an exponential family,

$$\gamma_j : [a_j, a_{j+1}) \to \mathcal{M}_j \subset \overline{\mathcal{M}} \subset \Delta^k,$$

for $j = 1, ..., J-1$, where:

(i) $0 = a_1 < a_2 < \cdots < a_J = 1$.

(ii) *For each $j = 2, \ldots, J$,*

$$\lim_{t \to a_j^-} \gamma(t) = \gamma(a_j) \in \Delta^k.$$

(iii) For each $j = 2, \ldots, J$, and $s = 1, \ldots, S$,

$$\lim_{t \to a_j^-} \frac{d^s \gamma}{dt^s}(t) = \frac{d^s \gamma}{dt^s}(a_j).$$

We denote the set of S-smooth curves as $\mathcal{C}^S$.

A curve in $\mathcal{C}^1$ has a finite arc length if

$$\lim_{L \to 1} \int_0^L \sqrt{\langle \gamma'(s), \gamma'(s) \rangle_{\gamma(s)}} ds < \infty,$$

where $\langle \cdot, \cdot \rangle_\pi$ is the Fisher information in Δ^k. Furthermore, the curve has finite energy if

$$\lim_{L \to 1} \int_0^L \langle \gamma'(s), \gamma'(s) \rangle_{\gamma(s)} ds < \infty.$$

It is common, and convenient, in Riemannian geometry [17] (Theorem 13, p. 128), to characterise Levi–Civita geodesics as being local minimisers of the energy functional, since these are the same paths that are local minimisers of the length functional. We follow this approach here when extending the definition of a (0)-geodesic to the extended exponential family.

We can now define an extended geodesic on an extended exponential family.

Definition 5. *Let $\overline{\mathcal{M}} \subseteq \Delta^k$ be an extended exponential family, and let $\pi, \tilde{\pi} \in \overline{\mathcal{M}}$. Define the set of finite energy paths by*

$$\mathcal{D}(\pi, \tilde{\pi}) = \left\{ \gamma \mid \gamma : [0,1] \to \overline{\mathcal{M}}, \gamma(0) = \pi, \gamma(1) = \tilde{\pi}, \int_0^1 < \gamma'(s), \gamma'(s) >_{\gamma(s)} ds < \infty \right\}. \tag{14}$$

Definition 6. *If $\gamma \in \mathcal{D}(\pi, \tilde{\pi})$, we call γ an extended Fisher geodesic if it (locally) minimises the energy functional.*

Theorem 3. *(a) Consider a curve $\gamma \in \mathcal{C}^1$, where $\gamma(t) \in r.i.(\Delta^k)$ for $t \in [0,1)$ and $\lim_{t \to 1} \gamma(t)$ lies in the proper face defined by a support set $\mathcal{I}_* \subset \mathcal{I}$, i.e., if $\gamma(1) = 0$, then $\gamma'(1) = 0$.*

Then, the curve has finite energy implies that $\lim_{t \to 1} \frac{d\gamma}{dt}$ lies in the tangent space to the proper face defined by the support set $\mathcal{I}_*$.

(b) Let $\mathcal{M} \subset r.i.(\Delta^k)$ be a p-dimensional exponential family, whose closure is $\overline{\mathcal{M}}$. Let $\pi \in \mathcal{M}$ and let $\tilde{\pi} \in \overline{\mathcal{M}}$ lie in a proper face of Δ^k defined by the index set $\mathcal{F}_1$. If $\gamma \in \mathcal{D}(\pi, \tilde{\pi})$ has the property that $\gamma(t) \in \mathcal{M}$ for $t \in [0,1)$ and is an extended Fisher geodesic, then we have: in the relative interior, after writing $\gamma|_{\mathcal{M}}$ in terms of the mean parameters of $\mathcal{M}$ as $(\mu_1(t), \cdots, \mu_p(t))$

$$\frac{d^2 \mu_i}{dt^2}(t) + \sum_{l,m=1}^{p} \Gamma^i_{lm}(\mu(t)) \frac{d\mu_l}{dt}(t) \frac{d\mu_m}{dt}(t) = 0, \tag{15}$$

where $\Gamma^k_{ij}(\mu)$ are the Christoffel symbols for the Levi–Civita connection of the Fisher metric.

On the boundary, we have that the curve γ has the property that $\lim_{t \to 1} \frac{d\pi}{dt}(t)$ is tangent to the face containing $\tilde{\pi}$.

Proof. See Appendix B. □

3.2. Computing the Extended Fisher Geodesic

Figure 1b shows an example of a Fisher geodesic's limiting behaviour. In particular, it is smooth on the boundary. However, it shows more, as we see a smooth curve in the extended multinomial

model that has three points lying on the edges and three disconnected, Fisher geodesics in the relative interior. These geodesics are smoothly connected in the extended family. This example motivated our investigation of the properties of extended Fisher geodesics. In this section, we investigate if it is possible to numerically find extended Fisher geodesics for arbitrary exponential families and numerically investigate their limiting properties. We have, from the consideration of the limiting properties of $(+1)$-geodesics, that, near the boundary, the exponential family lies almost parallel to a low-dimensional face of the simplex. Thus, locally (0)-geodesics in general exponential families will behave rather like (0)-geodesics in multinomial families, shown in Figure 1b, and reflect back into the the interior. At least locally, the geodesic behaves like the projection of the continuation of geodesics on the sphere. Example 4 below shows an explicit example of such a solution, where we have added the so-called *reflection principle*, Definition 7, in order to ensure both uniqueness and numerical stability in the solution.

The characterisation of the exponential family via Equation (1) is the familiar explicit representation in terms of the natural parameters, but is problematic numerically in representing the limiting distributions since the natural parameter needs to be unbounded to attain the boundaries. Thus, we replace this with a (-1)-representation, and then, invoking the reflection principle, we will work numerically in a (0)-representation.

From above, a p-dimensional exponential familiy is defined by $v^{(1)}, \dots v^{(p)}$, an orthonormal set of centred vectors. This set can be extended to form a k-dimensional orthonormal set of centred vectors, by selecting $u^{(p+1)}, \dots, u^{(k)}$. Within $r.i.(\Delta^k)$, a (0)-geodesic within this exponential family can be characterised by a set of differential equations of the form: for $j = 1, \cdots, p$ and $m = p+1, \cdots, k$,

$$\sum_{i=0}^{k} 1 \frac{d^2\pi_i(s)}{ds^2} = 0, \tag{16}$$

$$\sum_{i=0}^{k} \frac{u_i^{(m)}}{\pi_i} \frac{d^2\pi_i(s)}{ds^2} = \sum_{i=0}^{k} \frac{d\pi_i(s)}{ds} \frac{d\pi_i(s)}{ds} u_i^{(m)} \frac{1}{\pi_i^2(s)}, \tag{17}$$

$$\sum_{i=0}^{k} v_i^{(j)} \frac{d^2\pi_i(s)}{ds^2} = \frac{1}{2} \sum_{i=0}^{k} \frac{d\pi_i(s)}{ds} \frac{d\pi_i(s)}{ds} (v_i^{(j)} - E_\pi(V^{(k)})) \frac{1}{\pi_i(s)}, \tag{18}$$

where Equation (16) constrains the curve to lie in the space of unit measures, Equation (17) forces the solution to lie in the p-dimensional exponential family, and finally Equation (18) constrains the solution to be a Fisher geodesic inside that family.

In order to solve these equations numerically, we discretise in the standard way, but, near the boundary, we recognise that there is numerical instability in Equations (16)–(18) for small values of π_i. From the analysis of Section 2, we know that the small values are a fundamental part of the limit process. To illustrate such a solution, consider the following example.

Example 4. *Altham's Model.*

We return to Altham's two-dimensional extension of the binomial family. We take the Equations (16)–(18) and solve them numerically, in order to get an extended Fisher geodesic. This numerical solution is shown in Figure 4, with panel (a) showing the complete extended exponential family and panel (b) showing detail of the boundary behaviour.

As can be seen, in this example, the extended Fisher geodesic smoothly touches the boundary in two places. In fact, we can think of the path as the smooth union of a set of extended geodesics.

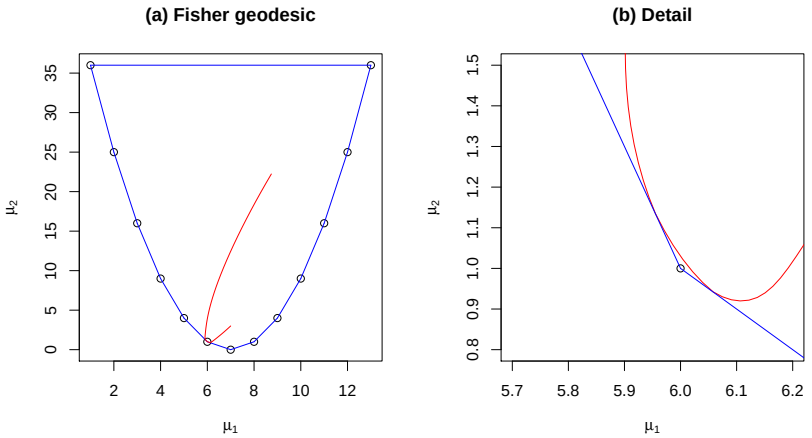

Figure 4. Extended Fisher geodesic in Altham's model.

As the previous example shows, we can think about the smooth union of extended geodesics. To ensure smoothness, we employ the following idea, which ensures the paths 'reflect' at a boundary and join in a smooth way.

Definition 7. *The reflection principle.*

To the conditions of Theorem 3, we add the condition that the limit of the second derivative, $\frac{d^2\pi}{dt^2}$, is finite and continuous on the path and use this to smoothly connect extended Fisher geodesics at the boundary.

In Appendix C, we discuss how we implement code that computes smooth unions of extended Fisher geodesics numerically using this principle by working in the $\xi = \sqrt{\pi}$ parameters.

3.3. Applications

While Fisher geodesics were studied very early in the Information Geometry literature, the importance of their applied utility is still an open question. The geodesic and the corresponding geodesic distance seem to be a natural thing to study and has, for example, found applications in image analysis (see [19,20], for example).

The seminal paper [21] illustrates a very important way that Fisher Riemannian geometry can have an impact on statistical practice. It considers, under regularity, parameter spaces of statistical models as smooth manifolds, and designs highly efficient Markov chain Monte Carlo algorithms by using Langevin methods on Riemannian geometric structures. In our recent paper [18], we showed the way that Fisher geodesics can smoothly attach to the boundaries of exponential families, and how this is one of the ways that the MCMC method achieves its efficiency. The results of this paper give the details of the results announced there.

In the paper [18], it was also shown how the boundary effects in extended exponential families mean that the log-likelihood can be very far from approximately quadratic in the mean and natural parameters.

Example 5. *Altham's model.*

Returning to Altham's model, Figure 5 shows an example of the shape of the log-likelihood function, in the mean parameters, when the maximum likelihood estimate is near the boundary. We see that the log-likelihood is very far from being approximately quadratic.

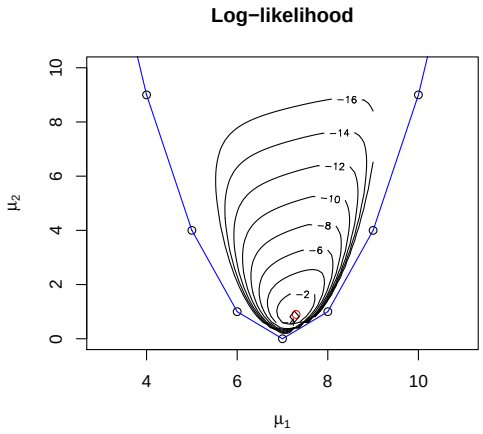

Figure 5. Contours of the log-likelihood in Altham's model.

Example 5 illustrates, in a simple visual way, how the log-likelihood can be far from approximately quadratic when near the boundary. The condition that the maximum likielihood estimate is on, or close to, the boundary is very common in categorical data analysis [9] and other discrete models [11]. This means that that standard iterative gradient based method, such as Newton's method, can fail in the mean or natural parameters. This was explored in [22] where it was shown that the boundary effects mean that commonly used first order asymptotic analyses, in, say logistic regression, can also fail.

We propose here that the smooth way that the Fisher Riemannian geometry deals with the boundary can be a useful tool to help deal with these problems. If we explore the model space using its Riemannian geodesics structure, we can smoothly reach the boundary. We see that this is in sharp contrast to working in the mean parameters, which, using Newton's method, would jump outside the boundary, or the natural parameters, which can never reach the boundary in finite time using a gradient approach. We note in fact Amari's highly efficient natural gradient method [23], while often motivated by divergence ideas, exactly uses the Fisher Riemannian geometry.

Acknowledgments: The authors would like to thank the Engineering and Physical Sciences Research Council (EPSRC) for the support of Grant No. EP/E017878/. We would also like to thank the referees for their very helpful comments.

Author Contributions: Both authors contributed equally to all aspects of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Theorem 1

We begin with the following preliminaries. Recall: $f(\theta) = o(\epsilon(\theta))$ means that $f(\theta)/\epsilon(\theta) \equiv \exp(\theta\delta_1)f(\theta) \to 0$ as $\theta \to \infty$. It follows that, for $h, j \in \{1, ..., g\}$:

$$\begin{array}{ll} \text{(a) } \forall j \geq 2: & \epsilon_{j\cdot}(\theta) = o(\epsilon(\theta)) \\ \text{(b) } \forall h < j: & f(\theta) = o(\epsilon_{j\cdot}(\theta)) \Rightarrow f(\theta) = o(\epsilon_{h\cdot}(\theta)). \end{array} \tag{A1}$$

In particular, for all $j \geq 2$,

$$f(\theta) = o(\epsilon_{j\cdot}(\theta)) \Rightarrow f(\theta) = o(\epsilon(\theta)). \tag{A2}$$

Again, $x \to x^n$ being continuous at 0 for all $n \geq 1$,

$$f(\theta) = o(\epsilon(\theta)) \Rightarrow [f(\theta)]^n = o(\epsilon(\theta)). \tag{A3}$$

Note that we work throughout to first order (only) in $\epsilon(\theta)$.

Now, (3) gives:

$$\pi_{(j,r)}(\theta) = \pi^0_{(j,r)} \exp\{\theta v_{(j)} - \psi(\theta),\} \tag{A4}$$

where

$$\begin{aligned} \exp\psi(\theta) &= \textstyle\sum_{j=0}^{g} \pi^0_{j\cdot} \exp(\theta v_{(j)}) \\ &= \pi^0_{0\cdot} \exp(\theta v_{(0)}) \times \{1 + \textstyle\sum_{j=1}^{g} \kappa_j \epsilon_{j\cdot}(\theta)\} \\ &= \pi^0_{0\cdot} \exp(\theta v_{(0)}) \times \{1 + \kappa_1 \epsilon(\theta) + o(\epsilon(\theta))\}. \end{aligned} \tag{A5}$$

Differentiating:

$$\begin{aligned} \psi'(\theta) \exp\psi(\theta) &= \textstyle\sum_{j=0}^{g} \pi^0_{j\cdot} v_{(j)} \exp(\theta v_{(j)}) \\ &= \pi^0_{0\cdot} v_{(0)} \exp(\theta v_{(0)}) \times \{1 + \textstyle\sum_{j=1}^{g} \kappa_j \rho_j \epsilon_{j\cdot}(\theta)\} \\ &= \pi^0_{0\cdot} v_{(0)} \exp(\theta v_{(0)}) \times \{1 + \kappa_1 \rho_1 \epsilon(\theta) + o(\epsilon(\theta))\}. \end{aligned} \tag{A6}$$

In addition, again:

$$\begin{aligned} \{\psi''(\theta) + (\psi'(\theta))^2\} \exp\psi(\theta) &= \textstyle\sum_{j=0}^{g} \pi^0_{j\cdot} v^2_{(j)} \exp(\theta v_{(j)}) \\ &= \pi^0_{0\cdot} v^2_{(0)} \exp(\theta v_{(0)}) \times \{1 + \textstyle\sum_{j=1}^{g} \kappa_j \rho_j^2 \epsilon_{j\cdot}(\theta)\} \\ &= \pi^0_{0\cdot} v_{(0)} \exp(\theta v_{(0)}) \times \{1 + \kappa_1 \rho_1^2 \epsilon(\theta) + o(\epsilon(\theta))\}. \end{aligned} \tag{A7}$$

In addition, for a third time:

$$\begin{aligned} \{\psi'''(\theta) + 3\psi'(\theta)\psi''(\theta) + (\psi'(\theta))^3\} \exp\psi(\theta) &= \textstyle\sum_{j=0}^{g} \pi^0_{j\cdot} v^3_{(j)} \exp(\theta v_{(j)}) \\ &= \pi^0_{0\cdot} v^3_{(0)} \exp(\theta v_{(0)}) \times \{1 + \textstyle\sum_{j=1}^{g} \kappa_j \rho_j^3 \epsilon_{j\cdot}(\theta)\} \\ &= \pi^0_{0\cdot} v_{(0)} \exp(\theta v_{(0)}) \times \{1 + \kappa_1 \rho_1^3 \epsilon(\theta) + o(\epsilon(\theta))\}. \end{aligned}$$

Using these preliminary results, we can look at limits of the relevant quantities. Throughout $1 \le r_0 \le m_0$ defines a reference bin $(0, r_0)$, for which the maximum value $v_{(0)}$ is observed.

(i) First look at the bin probabilities.

For j = 0:
Using (A4), (A5) and, twice, (A2) gives:

$$\begin{aligned} \pi_{(0,r_0)}(\theta) &= \frac{\pi^0_{(0,r_0)}}{\pi^0_{0\cdot}} \times \{1 + \textstyle\sum_{j=1}^{g} \kappa_j \epsilon_{j\cdot}(\theta)\}^{-1} \\ &= \frac{\pi^0_{(0,r_0)}}{\pi^0_{0\cdot}} \times \{1 - \textstyle\sum_{j=1}^{g} \kappa_j \epsilon_{j\cdot}(\theta) + o(\epsilon(\theta))\} \\ &= \frac{\pi^0_{(0,r_0)}}{\pi^0_{0\cdot}} \times \{1 - \kappa_1 \epsilon(\theta) + o(\epsilon(\theta))\}, \end{aligned} \tag{A8}$$

where, recall, $\kappa_1 := \pi^0_{1\cdot}/\pi^0_{0\cdot} > 0$, so that

$$\pi_{0\cdot}(\theta) := \textstyle\sum_{r_0=1}^{m_0} \pi_{(0,r_0)}(\theta) = \{1 - \kappa_1 \epsilon(\theta) + o(\epsilon(\theta))\}. \tag{A9}$$

For j > 0:

For each $j > 0$ and $1 \leq r \leq m_j$, (A4) gives

$$\frac{\pi_{(j,r)}(\theta)}{\pi_{(0,r_0)}(\theta)} = \frac{\pi^0_{(j,r)}}{\pi^0_{(0,r_0)}} \times \epsilon_{j\cdot}(\theta) = \frac{\pi^0_{(j,r)}}{\pi^0_{(0,r_0)}} \times \prod_{h=1}^{j} \epsilon_h(\theta). \tag{A10}$$

Thus, (A8) and (A10) give, for $j = 1$:

$$\begin{aligned} \pi_{(1,r_1)}(\theta) &= \left[\frac{\pi^0_{(1,r_1)}}{\pi^0_{(0,r_0)}} \times \epsilon(\theta)\right] \times \left[\frac{\pi^0_{(0,r_0)}}{\pi^0_{0\cdot}} \times \{1 - \kappa_1\epsilon(\theta) + o(\epsilon(\theta))\}\right] \\ &= \left[\frac{\pi^0_{(1,r_1)}}{\pi^0_{1\cdot}} \times \kappa_1 \times \epsilon(\theta)\right] + o(\epsilon(\theta)), \end{aligned} \tag{A11}$$

so that

$$\pi_{1\cdot}(\theta) := \textstyle\sum_{r_1=1}^{m_1} \pi_{(1,r_1)}(\theta) = \kappa_1\epsilon(\theta) + o(\epsilon(\theta)) \tag{A12}$$

and, for $j \geq 2$:

$$\begin{aligned} \pi_{(j,r_j)}(\theta) &= \left[\frac{\pi^0_{(j,r_j)}}{\pi^0_{(0,r_0)}} \times \prod_{h=1}^{j} \epsilon_h(\theta)\right] \times \left[\frac{\pi^0_{(0,r_0)}}{\pi^0_{0\cdot}} \times \{1 - \kappa_1\epsilon(\theta) + o(\epsilon(\theta))\}\right] \\ &= \left[\frac{\pi^0_{(j,r_j)}}{\pi^0_{j\cdot}} \times \kappa_j \times \epsilon(\theta)\right] \times \left[\prod_{h=2}^{j} \epsilon_h(\theta)\right] + o(\epsilon(\theta)) \\ &= o(\epsilon(\theta)), \end{aligned} \tag{A13}$$

so that

$$\pi_{j\cdot}(\theta) := \textstyle\sum_{r_j=1}^{m_j} \pi_{(j,r_j)}(\theta) = o(\epsilon(\theta)). \tag{A14}$$

(ii) We can now look at the mean parameter for the geodesic $\mu(\theta) = \psi'(\theta)$.

Dividing (A6) by (A5), and repeatedly using (A2), gives:

$$\begin{aligned} \mu(\theta) = \psi'(\theta) &= \frac{\pi^0_{0\cdot} v_{(0)} \exp(\theta v_{(0)}) \times \{1 + \kappa_1\rho_1\epsilon(\theta) + o(\epsilon(\theta))\}}{\pi^0_{0\cdot} \exp(\theta v_{(0)}) \times \{1 + \kappa_1\epsilon(\theta) + o(\epsilon(\theta))\}} \\ &= v_{(0)}\{1 + \kappa_1\rho_1\epsilon(\theta) + o(\epsilon(\theta))\}\{1 - \kappa_1\epsilon(\theta) + o(\epsilon(\theta))\} \\ &= v_{(0)}\{1 - \kappa_1(1 - \rho_1)\epsilon(\theta) + o(\epsilon(\theta))\} \\ &= v_{(0)} - \kappa_1[v_{(0)} - v_{(1)}]\epsilon(\theta) + o(\epsilon(\theta)). \end{aligned} \tag{A15}$$

It follows at once from (A15) that:

$$[v_{(0)} - \mu(\theta)] = \kappa_1[v_{(0)} - v_{(1)}]\epsilon(\theta) + o(\epsilon(\theta)), \tag{A16}$$

while, for each $j > 0$,

$$-[v_{(j)} - \mu(\theta)] = [v_{(0)} - v_{(j)}] - \kappa_1[v_{(0)} - v_{(1)}]\epsilon(\theta) + o(\epsilon(\theta)). \tag{A17}$$

Furthermore, using (A3), we have:

$$[v_{(0)} - \mu(\theta)]^2 = o(\epsilon(\theta)), \tag{A18}$$

while, for each $j > 0$,

$$[v_{(j)} - \mu(\theta)]^2 = [v_{(0)} - v_{(j)}]^2 - 2\kappa_1[v_{(0)} - v_{(1)}][v_{(0)} - v_{(j)}]\epsilon(\theta) + o(\epsilon(\theta)). \tag{A19}$$

(iii) Next, we look at the Fisher information $Var(\theta) = \psi''(\theta)$.

Using (A9), (A12) and (A14), together with (A18) and (A19):

$$\begin{aligned} &Var(\theta) = \psi''(\theta) = \textstyle\sum_{j=0}^{g} \pi_{j\cdot}(\theta)[v_{(j)} - \mu(\theta)]^2 \\ &= \pi_{0\cdot}(\theta)[v_{(0)} - \mu(\theta)]^2 + \pi_{1\cdot}(\theta)[v_{(1)} - \mu(\theta)]^2 + \textstyle\sum_{j=2}^{g} \pi_{j\cdot}(\theta)[v_{(j)} - \mu(\theta)]^2 \\ &= \{1 - \kappa_1\epsilon(\theta) + o(\epsilon(\theta))\} \times o(\epsilon(\theta)) \\ &+ \{\kappa_1\epsilon(\theta) + o(\epsilon(\theta))\} \times \{[v_{(0)} - v_{(1)}]^2 - 2\kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta))\} \\ &+ \textstyle\sum_{j=2}^{g} \{o(\epsilon(\theta))\} \times \{[v_{(0)} - v_{(j)}]^2 - 2\kappa_1[v_{(0)} - v_{(1)}][v_{(0)} - v_{(j)}]\epsilon(\theta) + o(\epsilon(\theta))\} \\ &= \kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta)). \end{aligned} \tag{A20}$$

(iv) Finally, we look at limiting tangent vectors in (-1)-coordinates.

For $j = 0$:
Now, using (A8), (A16) and (A20):

$$\begin{aligned} \frac{\partial \pi_{(0,r_0)}(\theta)}{\partial \mu} &= \frac{\partial \pi_{(0,r_0)}(\theta)/\partial\theta}{\partial \mu/\partial\theta} = \frac{[v_{(0)} - \psi'(\theta)]\pi_{(0,r_0)}(\theta)}{\psi''(\theta)} \\ &= \frac{[\kappa_1[v_{(0)} - v_{(1)}]\epsilon(\theta) + o(\epsilon(\theta))] \times \frac{\pi^0_{(0,r_0)}}{\pi^0_{0\cdot}} \times \{1 - \kappa_1\epsilon(\theta) + o(\epsilon(\theta))\}}{\kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta))} \\ &= \frac{\pi^0_{(0,r_0)}}{\pi^0_{0\cdot}} \times \frac{1}{[v_{(0)} - v_{(1)}]} \times \frac{\{\kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta))\}}{\{\kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta))\}} \\ &= \frac{\pi^0_{(0,r_0)}}{\pi^0_{0\cdot}} \times \frac{1}{[v_{(0)} - v_{(1)}]} + o(\epsilon(\theta)), \end{aligned} \tag{A21}$$

so that

$$\textstyle\sum_{r_0=1}^{m_0} \frac{\partial \pi_{(0,r_0)}(\theta)}{\partial \mu} = [v_{(0)} - v_{(1)}]^{-1} + o(\epsilon(\theta)). \tag{A22}$$

For $j = 1$:
Note first that, from (A5) and (A20):

$$\begin{aligned} \psi''(\theta) \times \exp(\psi(\theta)) &= \{\kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta))\} \\ &\times \pi^0_{0\cdot} \exp(\theta v_{(0)}) \times \{1 + \kappa_1\epsilon(\theta) + o(\epsilon(\theta))\} \\ &= \pi^0_{0\cdot} \exp(\theta v_{(0)}) \times \{\kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta))\}. \end{aligned} \tag{A23}$$

Using (A23), together with (A4) and the $j = 1$ version of (A17), we have:

$$\begin{aligned}\frac{\partial \pi_{(1,r_1)}(\theta)}{\partial \mu} &= \frac{\partial \pi_{(1,r_1)}(\theta)/\partial\theta}{\partial \mu/\partial\theta} = \frac{[v_{(1)} - \psi'(\theta)]\pi^0_{(1,r_1)} \times \{\exp(\theta v_{(0)}) \times \epsilon(\theta)\}}{\psi''(\theta) \times \exp(\psi(\theta))} \\ &= -\frac{\{[v_{(0)} - v_{(1)}] + \kappa_1[v_{(0)} - v_{(1)}]\epsilon(\theta) + o(\epsilon(\theta))\} \times \pi^0_{(1,r_1)} \times \epsilon(\theta)}{\pi^0_{0.} \times \{\kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta))\}} \\ &= -\frac{\pi^0_{(1,r_1)}}{\pi^0_{0.}} \times \frac{\{[v_{(0)} - v_{(1)}] + \kappa_1[v_{(0)} - v_{(1)}]\epsilon(\theta) + o(\epsilon(\theta))\} \times \epsilon(\theta)}{\{\kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta))\}} \\ &= -\frac{\pi^0_{(1,r_1)}}{\pi^0_{0.}} \times \frac{\{[v_{(0)} - v_{(1)}]\epsilon(\theta) + o(\epsilon(\theta))\}}{\kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta))} \\ &= -\frac{\pi^0_{(1,r_1)}}{\pi^0_{0.}} \times \frac{1}{\kappa_1[v_{(0)} - v_{(1)}]} \times \frac{\{\kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta))\}}{\{\kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta))\}} \\ &= -\frac{\pi^0_{(1,r_1)}}{\pi^0_{1.}} \times \frac{1}{[v_{(0)} - v_{(1)}]} + o(\epsilon(\theta))\},\end{aligned} \tag{A24}$$

so that

$$\sum_{r_1=1}^{m_1} \frac{\partial \pi_{(1,r_1)}(\theta)}{\partial \mu} = -[v_{(0)} - v_{(1)}]^{-1} + o(\epsilon(\theta)). \tag{A25}$$

For $j \geq 2$:

Finally, consider any $j \geq 2$. Then, using (A17) and (A23), we have:

$$\begin{aligned}&\frac{\partial \pi_{(j,r_j)}(\theta)}{\partial \mu} \\ &= \frac{[v_{(j)} - \psi'(\theta)]\pi_{(j,r_j)}(\theta)}{\psi''(\theta)} = \frac{[v_{(j)} - \psi'(\theta)]\pi^0_{(j,r_j)} \times \{\exp(\theta v_{(0)}) \times \epsilon_{j\cdot}(\theta)\}}{\psi''(\theta) \times \exp(\psi(\theta))} \\ &= -\frac{\pi^0_{(j,r_j)}}{\pi^0_{0.}} \times \frac{\{[v_{(0)} - v_{(j)}] + \kappa_1[v_{(0)} - v_{(1)}]\epsilon(\theta) + o(\epsilon(\theta))\} \times \{\epsilon(\theta)\prod_{h=2}^{j}\epsilon_h(\theta)\}}{\{\kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta))\}} \\ &= -\frac{\pi^0_{(j,r_j)}}{\pi^0_{0.}} \times \frac{\{[v_{(0)} - v_{(j)}]\epsilon(\theta) + o(\epsilon(\theta))\} \times \prod_{h=2}^{j}\epsilon_h(\theta)}{\{\kappa_1[v_{(0)} - v_{(1)}]^2\epsilon(\theta) + o(\epsilon(\theta))\}} \\ &= o(\epsilon(\theta)),\end{aligned} \tag{A26}$$

so that

$$\sum_{r_j=1}^{m_j} \partial \pi_{(j,r_j)}(\theta)/\partial \mu = o(\epsilon(\theta)). \tag{A27}$$

Appendix B. Proof of Theorem 3

(a) It is convenient to work in the (0)-representation of the simplex i.e., $S^k_+ = \left\{\xi | \sum_{i=0}^k \xi_i^2 = 1, \xi_i \geq 0\right\}$. In S^k_+, the Fisher metric on all tangent spaces is induced by the identity matrix in R^{k+1} and this agrees with the Fisher metric for all tangent spaces of Δ^k.

We write

$$\gamma(t) = (\pi_0(t), \cdots, \pi_k(t)).$$

The image of $\gamma(t)$ in the $r.i.(S^k_+)$ is

$$\left(\sqrt{\pi_0(t)}, \sqrt{\pi_1(t)}, \ldots, \sqrt{\pi_k(t)}\right),$$

and its tangent vector is

$$\frac{1}{2}\left(\frac{\pi_0'(t)}{\sqrt{\pi_0(t)}},\frac{\pi_1'(t)}{\sqrt{\pi_1(t)}},\ldots,\frac{\pi_k'(t)}{\sqrt{\pi_k(t)}}\right).$$

Thus, for $t \in [0,1)$, the squared length of the tangent vector is

$$\frac{1}{4}\sum_{i=0}^{k}\left(\frac{\pi_i'(t)}{\sqrt{\pi_i(t)}}\right)^2 = \frac{1}{4}\sum_{i=0}^{k}\frac{(\pi_i'(t))^2}{\pi_i(t)}.$$

For any index i such that $\pi_i(1)=0$, suppose $\pi_i'(1)\neq 0$. There are two cases. First, $\pi_i'(t)<C<0$ for $t\in(1-\epsilon,1]$, so that

$$\int_{1-\epsilon}^{L}\frac{(\pi_i'(t))^2}{\pi_i(t)}dt = \left|\int_{1-\epsilon}^{L}\frac{(\pi_i'(t))^2}{\pi_i(t)}dt\right| > |C|\left|\int_{1-\epsilon}^{L}\frac{(\pi_i'(t))}{\pi_i(t)}dt\right| = C\left|(\log(\pi_i(1-\epsilon))-\log(\pi_i(L))\right|,$$

which is unbounded above as $L\to 1$. The second case, $\pi_i'(t)>C>0$, can not happen due to the non-negativity of probabilities. Thus, we have

$$\lim_{L\to 1}\int_0^L <\gamma'(s),\gamma'(s)>_{\gamma(s)}ds<\infty \Rightarrow \left\{\gamma_i(1)=0\Rightarrow\gamma_i'(1)=0\right\}. \tag{A28}$$

Appendix C. Numerically Solving for Extended Fisher Geodesic

To numerically solve Equations (16)–(18), we have found that it is convenient to work in the (0)-representation of the simplex, that is, $\xi_i := \sqrt{\pi_i}$. In these parameters, the defining equations, in the relative interior, are

$$\sum_{i=0}^{k}\xi_i\frac{d^2\xi_i(s)}{ds^2} = -\sum_{i=0}^{k}\left(\frac{d\xi_i(s)}{ds}\right)^2, \tag{A29}$$

$$\sum_{i=0}^{k}\frac{u_i^{(m)}}{\xi_i}\frac{d^2\xi_i(s)}{ds^2} = \sum_{i=0}^{k}\left(\frac{d\xi_i(s)}{ds}\right)^2\frac{u_i^{(m)}}{\xi_i^2}, \tag{A30}$$

$$\sum_{i=0}^{k}(v_i^{(j)}-E_{\xi(s)}(V^{(j)}))\frac{d^2\xi_i(s)}{ds^2} = 0. \tag{A31}$$

We use the package `deSolve` in R [24], to numerically solve the Equations (A29)–(A31). This method discretises the equations and solves them iteratively. At iteration s, we have state variables $\xi^{(s)}$ and $v^{(s)} := \frac{d\xi_i}{dt}$. From the standard theory of ODE, in the relative interior, such state variables are enough to determine what happens at the next state by solving the linear equations, which are the discrete version of Equations (A29)–(A31).

At the boundary, where at least one $\xi_i = 0$, the corresponding linear equations are singular and do not define a unique next state. We therefore apply the reflection principle to extend the path through this singularity. To apply the principle numerically, we add to the algorithm the condition that whenever the value of

$$\xi^{(s)}+v^{(s)}<0,$$

we reflect in this boundary by changing this component of the tangent vector to $-v^{(s)}$. This means that the path in ξ-space is locally of the form

$$\xi(t) = a|t-t_0|$$

in a neighbourhood of $\xi(t_0)=0$. In terms of the π-parameters, this gives

$$\pi(t) = a^2(t-t_0)^2, \frac{d\pi}{dt}(t) = 2a^2(t-t_0), \frac{d^2\pi}{dt^2}(t) = 2a^2,$$

which means at $t = t_0$, where the path touches the boundary, we have that the tangent is zero, as required by Theorem 3(b) but also that the second derivative is finite and continuous as required by the reflection principle.

(b) We use the characterisation of a Fisher geodesic as being a local minimiser of the energy functional among smooth curves. We follow the standard argument using the calculus of variations, which give the geodesic Equations (15) within the relative interior. A path that is smooth in the extended exponential family will have finite energy if the tangent vector is parallel to the boundary. Since the geodesic is a local minimiser, it must have finite energy. This gives the final boundary condition.

References

1. Amari, S.-I. Differential-geometrical methods in statistics. In *Lecture Notes in Statistics*; Springer: New York, NY, USA, 1985; Volume 28.
2. Amari, S.-I.; Barndorff-Nielsen, O.E.; Kass, R.E.; Lauritzen, S.L.; Rao, C.R. *Differential Geometry in Statistical Inference*; IMS Lecture Notes-Monograph Series; IMS: Hayward, CA, USA, 1987.
3. Barndorff-Nielsen, O.E. *Information and Exponential Families in Statistical Theory*; John Wiley & Sons, Ltd.: Chichester, UK, 1978.
4. Kass, R.E.; Vos, P.W. *Geometrical Foundations of Asymptotic Inference*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1997.
5. Amari, S.-I. *Information Geometry and Its Applications*; Springer: Tokyo, Japan, 2016.
6. Brown, L.D. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*; Lecture Notes-Monograph Series; IMS: Hayward, CA, USA, 1986; Volume 9.
7. Critchley, F.; Marriott, P. *Information Geometry and Its Applications: An Overview*; Computational Information Geometry; Springer: Cham, Switzerland, 2017; pp. 1–31.
8. Agresti, A. *Categorical Data Analysis*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2013.
9. Geyer, C.J. Likelihood inference in exponential families and directions of recession. *Electron. J. Stat.* **2009**, *3*, 259–289.
10. Lauritzen, S.L. *Graphical Models*; Clarendon Press: Oxford, UK, 1996.
11. Rinaldo, A.; Feinberg, S.; Zhou, Y. On the geometry of discrete exponential families with applications to exponential random graph models. *Electron. J. Stat.* **2009**, *3*, 446–484.
12. Fienberg, S.E.; Rinaldo, A. Maximum likelihood estimation in log-linear models. *Ann. Stat.* **2012**, *40*, 996–1023.
13. Rauh, J.; Kahle, T.; Ay, N. Support sets in exponential families and oriented matroid theory. *Int. J. Approx. Reason.* **2011**, *52*, 613–626.
14. Geiger, D.; Meek, C.; Sturmfels, B. On the toric algebra of graphical models. *Ann. Stat.* **2006**, *34*, 1463–1492.
15. Critchley, F.; Marriott, P. Computing with Fisher geodesics and extended exponential families. *Stat. Comput.* **2016**, *26*, 325–332.
16. Altham, P. Two generalizations of the binomial distribution. *Appl. Stat.* **1978**, *27*, 162–167.
17. Petersen, P. *Riemannian Geometry*, 2nd ed.; Graduate Texts in Mathematics; Springer: New York, NY, USA, 2006; Volume 171.
18. Critchley, F.; Marriott, P. Computational Information Geometry in Statistics: Theory and practice. *Entropy* **2014**, *16*, 2454–2471.
19. Mio, W.; Badlyans, D.; Liu, X. A computational approach to Fisher information geometry with applications to image analysis. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*; Springer: Berlin, Germany, 2005.
20. Peter, A.; Rangarajan, A. Shape analysis using the Fisher-Rao Riemannian metric: Unifying shape representation and deformation. In Proceedings of the 3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, Arlington, VA, USA, 6–9 April 2006.
21. Girolami, M.; Calderhead, B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. B* **2011**, *73*, 123–214.

22. Anaya-Izquierdo, K.; Critchley, F.; Marriott, P. When are first-order asymptotics adequate? A diagnostic. *STAT* **2014**, *3*, 17–22.
23. Amari, S.I. Natural gradient works efficiently in learning. *Neural Comput.* **1998**, *10*, 251–276.
24. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2016.