Computational information geometry in statistics: foundations

Karim Anaya-Izquierdo[1], Frank Critchley[2], Paul Marriott[3], and Paul Vos[4]

[1] London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK
[2] The Open University, Milton Keynes, MK7 6AA, UK
[3] University of Waterloo, Waterloo, Ontario, Canada N2L 3G1
[4] East Carolina University, Greenville, NC 27858-4353 USA

Abstract. This paper lays the foundations for a new framework for numerically and computationally applying information geometric methods to statistical modelling.

1 Introduction

The power and elegance of information geometry have yet to be fully exploited in statistical practice. To this end, computational information geometry aims to provide operational tools to help resolve important, long-standing problems. For reasons of implementation, all random variables considered take a finite number of values. The key idea of this paper is to represent statistical models – sample spaces, together with probability distributions on them – and associated inference problems, inside adequately large but finite dimensional spaces. In these embedding spaces the building blocks of information geometry in statistics can be numerically computed explicitly and the results used for algorithm development. Accordingly, after a possible initial discretisation, the space of all distributions for the random variable of interest can be identified with the simplex,

$$\Delta^k := \left\{ \pi = (\pi_{\mathbf{0}}, \pi_{\mathbf{1}}, \ldots, \pi_{\mathbf{k}})^\top \, : \, \pi_{\mathbf{i}} \geq \mathbf{0} \, , \, \sum_{\mathbf{i=0}}^{\mathbf{k}} \pi_{\mathbf{i}} = \mathbf{1} \right\}, \tag{1}$$

together with a unique label for each vertex, representing the random variable. Modulo discretisation, this structure therefore acts as a universal model. Clearly, the multinomial family on $k+1$ categories can be identified with the relative interior of this space, $int(\Delta^k)$, while the extended family, (1), allows the possibility of distributions with different support sets.

The starting point for much of statistical inference is a working model for observed data comprising a set of distributions on a sample space. A working model $\mathcal{M}$ can be represented by a subset of Δ^k and may be specified by an explicit parameterisation. Computational information geometry explicitly uses the information geometry of Δ^k to numerically compute statistically important features of $\mathcal{M}$. These features include: properties of the likelihood, which can be nontrivial in many of the examples considered here; the adequacy of first

order asymptotic methods – notably, via higher order asymptotic expansions; curvature based dimension reduction; and inference in mixture models, [3].

A fuller version of this paper, which also outlines further developments in computational information geometry in statistics, is available as [2]. For brevity, all formal proofs are given there.

2 Discretisation

The approach taken in this paper is inherently discrete and finite. Sometimes, of course, this can be with zero loss. In general, though, suitable finite partitions of the sample space can be used, for which an appropriate theory is developed. While this is clearly not the most general case mathematically speaking it does provide an excellent foundation on which to construct a computational theory. Furthermore, since real world measurements can only be made to a fixed precision all models can – arguably, should – be thought of as fundamentally categorical. The relevant question for a computational theory is then: what is the effect on the inferential objects of interest of a particular selection of such categories? This key question is addressed in Theorems 1 and 2.

Example 1. An example in [10] concerns survival times Z for leukaemia patients measured in days from the time of diagnosis. Originally from [6], there are 43 observations. Here the data, while being treated as continuous, is only recorded at integer number of days. Thus, as far as any statistical analysis that can be carried out is concerned, there is literally zero loss in treating it as sparse categorical.

For illustrative purposes, a *further* level of coarseness is added here, by selecting bins of size 4 days. The parameter of interest is the mean lifetime, μ. In panel (a) of Fig. 1 it is shown that there is effectively no inferential loss in such a choice. The solid line is the likelihood function based on binning the data to bins of width four days, using a multinomial approximation. The dots in this panel are the log-likelihood for the raw data based on the continuous censored exponential model. As can be clearly seen there is no real inferential loss in the binning and discretisation process.

In order to use the high dimensional simplex models with continuous random variables it is necessary to truncate and discretise the sample space into a finite number of bins. The following theorems show that the information loss in doing this is arbitrarily small for a fine enough discretisation and that the key to understanding the information in general is controlling the conditional moments in each bin of the random variables of interest, uniformly in the parameters of the model.

Theorem 1. *Let $f(x;\theta)$, $\theta \in \Theta$, be a parametric family of density functions with common support $\mathcal{X} \subset \mathbf{R}^d$ each being continuously differentiable on the relative interior of $\mathcal{X}$, assumed non-empty. Further, let $\mathcal{X}$ be compact, while*

$$\left\{ \left\| \frac{\partial}{\partial x} f(x;\theta) \right\| \, | x \in \mathcal{X} \right\}$$

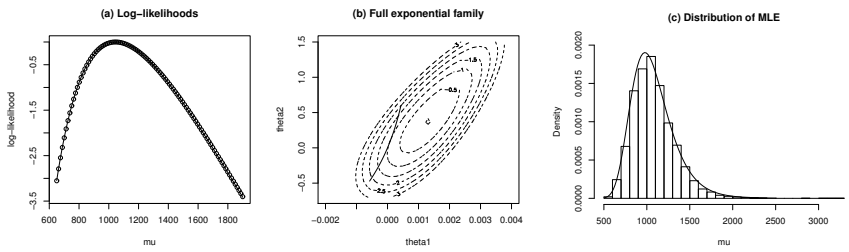


Fig. 1. Computational information geometry: likelihood approximation and dimension reduction

is uniformly bounded in $\theta \in \Theta$ by M, say.

Then, for any $\epsilon > 0$ and for any sample size $N > 0$, there exists a finite, measurable partition $\{B_k\}_{k=0}^{K(\epsilon,N)}$ of $\mathcal{X}$ such that: for all $(x_1, \ldots, x_N) \in \mathcal{X}^N$, and for all $(\theta_0, \theta) \in \Theta^2$

$$\left| \log \left\{ \frac{Lik_d(\theta)}{Lik_d(\theta_0)} \right\} - \log \left\{ \frac{Lik_c(\theta)}{Lik_c(\theta_0)} \right\} \right| \leq \epsilon, \tag{2}$$

where Lik_d and Lik_c are the likelihood functions from the discretised and continuous distributions respectively.

The following result looks at the case where the family that is discretised is itself an exponential family and so the tools of classical information geometry can be applied. In general, after discretisation a full exponential family does not remain full exponential and there is information loss. However, the following results show that this loss can be made small enough to be unimportant for inference and that all information geometric results on the two families can be made arbitrarily close.

Theorem 2. *Let $f(x;\theta) = \nu(x) \exp\left\{\theta^T s(x) - \psi(\theta)\right\}$, $x \in \mathcal{X}, \theta \in \Theta$, be an exponential family which satisfies the regularity conditions of [1], p. 16. Further, assume that $s(x)$ is uniformly continuous and $s(\mathcal{X})$ is compact.*

Then, for any $\epsilon > 0$, there exists a finite measurable partition $\{B_k\}_{k=0}^{K(\epsilon)}$ of $\mathcal{X}$ such that, for all choices of bin labels $s_k \in s(B_k)$, all terms of Amari's information geometry for $f(x;\theta)$ can be approximated to $O(\epsilon)$ by the corresponding terms for the family

$$\left\{ (\pi_i(\theta), s_i) | \pi_i(\theta) := \int_{B_i} f(x;\theta) dx, s_i \in s(B_i) \right\}.$$

In particular:

(a) For all θ, and any norm,

$$\|\mu_d(\theta) - \mu_c(\theta)\| = O(\epsilon)$$

where $\mu_d(\theta) = \sum_{k=0}^{K(\epsilon)} s_k \pi_k(\theta)$ and $\mu_c(\theta) = \int_{\mathcal{X}} x f(x;\theta) dx$.

(b) The expected Fisher information for θ of $f(x;\theta)$, $I_c(\theta)$, and the expected Fisher information for $\{\pi_k(\theta)\}$, $I_d(\theta)$, satisfy

$$\|I_d(\theta) - I_c(\theta)\|_\infty = O(\epsilon^2).$$

(c) The skewness tensors $T_c(\theta)$, see [1], p. 105, of $f(x;\theta)$ and $T_d(\theta)$ for $\{\pi_k(\theta)\}$ satisfy

$$\|T_d(\theta) - T_c(\theta)\|_\infty = O(\epsilon^3).$$

In continuous examples, like Example 1, a compactness condition is used to keep the underlying geometry finite. A following paper will look at the case where the compactness condition is not needed. In this case, infinite dimensional simplexes, and their closures, are used as the 'space of all distributions', the extension of classical information geometry here requiring careful consideration of convergence.

3 Information geometry of extended multinomial model

3.1 Affine geometries

Information geometry is constructed from two different affine geometries related in a non-linear way via duality and the Fisher information, see [1] or [13]. In the full exponential family context, one affine structure (the so-called +1 structure) is defined by the natural parameterization, the second (the −1 structure) by the mean parameterization. The closure of exponential families has been studied by [4], [5], [14] and [19] in the finite dimensional case and by [7] in the infinite dimensional case. One important difference in the approach taken here is that limits of families of distributions, rather than pointwise limits, are central.

This paper constructs a theory of information geometry following that introduced by [1] via the affine space construction introduced by [18] and extended by [15]. Since this paper concentrates on categorical random variables, the following definitions are appropriate. Consider a finite set of disjoint categories or bins $\mathcal{B} = \{B_i\}_{i\in A}$. Any distribution over this finite set of categories is defined by a set $\{\pi_i\}_{i\in A}$ which defines the corresponding probabilities. Note in a mild abuse of notation we identify a bin B_i with its label i.

Definition 1. *The -1-affine space structure over distributions 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 1, the space of (discretised) distributions is a -1-convex subspace of the affine space $(X_{mix}, V_{mix}, +)$. A similar affine structure for the $+1$-geometry, once the support has been fixed, can be derived from the definitions in [18] pages 9 – 13, or as described in [15] page 82.

3.2 Geometry of extended trinomial distribution

To illustrate the information geometry of the extended multinomial distribution, the trinomial case is now described explicitly. The general case follows by obvious extensions, see [2]. The case when the dimension is so large that numerically evaluating sums becomes impractical is considered in [9].

Example 2. An explicit example of the information geometry of the extended trinomial model is shown in Fig. 2. The closed simplex in panel (a) represents the set of multinomial distributions with bin probabilities (π_0, π_1, π_2) where $\pi_i \geq 0$.

In this example, a vector $b^T = (1, 2, 3)$ was chosen, and the parallel lines in panel (a) are level sets of the mean of $b^T X$, where X is the trinomial random variable. These are -1-geodesics, and it is immediate that they extend to the boundary in a very natural way. These lines are also shown in panel (b), but now in the $+1$ (or natural) parameterization and so are non-linear.

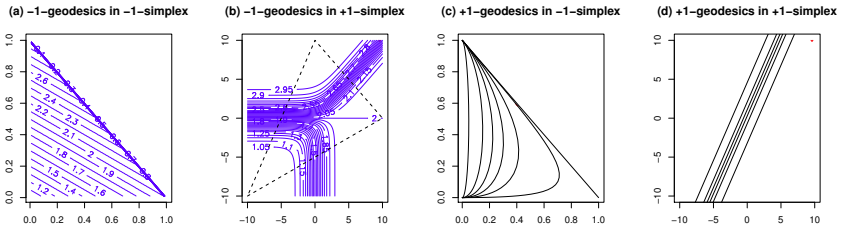


Fig. 2. The information geometry of the extended trinomial model

Panel (d) shows the relative interior of the extended trinomial in the $+1$-affine parameterization. The straight lines represent one dimensional full exponential families with probabilities of the form $\left\{ \frac{\pi_i \exp(\theta b_i)}{\sum_{k=0}^{2} \pi_k \exp(\theta b_k)} \right\}_{i=0}^{2}$, each $\pi_k > 0$. These are $+1$ -geodesics in the direction b through the base-point (π_0, π_1, π_2). It is a standard result that these $+1$ parallel lines are everywhere orthogonal, with respect to the metric defined by the Fisher information matrix, to the -1-parallel lines shown in panels (a) and (b).

The key step in understanding the simplicial nature of the $+1$-geometry is to see how the limits of the $+1$-parallel lines are connected to the boundary of the simplex. This is made clear in panel (c), where the $+1$-geodesics are plotted in the -1-affine parameters as curves. The limits of the curves lie on the boundary of the simplex. The closure of the $+1$-representation multinomial is defined to make these continuous limits defined "at infinity" in the $+1$-parameters and is shown schematically as the dotted triangle in panel b.

3.3 The shape of the likelihood

Potentially high dimensional simplicial structures being the natural spaces in which to base computational information geometry, a primary question is to look at the way that the likelihood, or log-likelihood, behaves in them. First note two important issues: in typical applications the sample size will be much smaller than the dimension of the simplex, while the simplex contains sub-simplexes with varying support. These two statements mean that our standard intuition about the shape of the log-likelihood function will not hold. In particular, the standard χ^2-approximation to the distribution of the deviance does not hold.

It will be convenient to call the face of the simplex spanned by the vertices (bins) having strictly positive counts the *observed face*, and the face spanned by the complement of this set the *unobserved face*. In the -1-representation, the log-likelihood is strictly concave on the observed face, strictly decreasing in the normal direction from it to the unobserved face and, otherwise, constant. For more details of the geometry of the observed face see the paper [3].

Theorem 3. *Let the observed counts be* $\{n_i\}_{i=0}^k$ *and define two subsets of the index set* $\{0, \cdots, k\}$ *by* $\mathcal{P} = \{i | n_i > 0\}$ *and* $\mathcal{Z} = \{i | n_i = 0\}$. *Let* $V_{\text{mix}} = \{(v_0, \ldots, v_k) | \sum v_i = 0\}$, *and further define the set* $V^0 \subset V_{\text{mix}}$ *by* $\{v \in V_{\text{mix}} | v_i = 0 \; \forall i \in P\}$.

(a) The set V^0 *is a linear subspace of* V_{mix}. *The log-likelihood is constant on* -1 *affine subspaces of the form* $\pi + V^0$.

(b) Select $k^* \in \mathcal{Z}$ *and consider the vector subspace of* V_{mix} *defined by*

$$V^{k^*} := \{v \in V_{\text{mix}} | v_i = 0 \text{ if } i \in \mathcal{Z} \backslash \{k^*\}\}.$$

Then V_{mix} *can be decomposed as a direct sum of vector spaces* $V_{\text{mix}} = V^0 \oplus V^{k^*}$.

3.4 Spectrum of Fisher information

With $\pi_{(0)}$ denoting the vector of all bin probabilities except π_0, the Fisher information matrix for the $+1$ parameters, when viewed as the variance-covariance matrix for the score [20] Def. 2.79 page 111, can be written explicitly as a function of the probabilities. It is given by the sample size times

$$I(\pi) := diag(\pi_{(0)}) - \pi_{(0)} \pi_{(0)}^T,$$

see [20] page 674. Its explicit spectral decomposition is, in all cases, an example of interlacing eigenvalue results, (see for example [11], Chapter 4). In particular, suppose $\{\pi_i\}_{i=1}^k$ comprises $g > 1$ distinct values $\lambda_1 > \cdots > \lambda_g > 0$, λ_i occurring m_i times, so that $\sum_{i=1}^g m_i = k$. Then, the spectrum of $I(\pi)$ comprises g simple eigenvalues $\{\widetilde{\lambda}_i\}_{i=1}^g$, the roots of an explicit polynomial, satisfying

$$\lambda_1 > \tilde{\lambda}_1 > \cdots > \lambda_g > \tilde{\lambda}_g \geq 0, \tag{3}$$

together, if $g < k$, with $\{\lambda_i : m_i > 1\}$, each such λ_i having multiplicity $m_i - 1$. Further, $\widetilde{\lambda}_g > 0 \Leftrightarrow \pi_0 > 0$, while each $\widetilde{\lambda}_i$ $(i < g)$ is typically (much) closer to λ_i than to λ_{i+1}, making it a near replicate of λ_i.

In this way, the Fisher spectrum mimics key features of the bin probabilities. Of central importance, one or more eigenvalues are exponentially small if and only if the same is true of the bin probabilities, the Fisher information matrix being singular if and only if one or more of the $\{\pi_i\}_{i=0}^k$ vanishes. Again, typically, two or more eigenvalues will be close when two or more corresponding bin probabilities are.

3.5 Closure

Given a full exponential family embedded in the high-dimensional sparse simplex an important question is to identify its limit points – how it is connected to the boundary. It is generally true and, shown in a concrete example in Fig. 2 (c), that one dimensional exponential families limits lie at vertices, and the vertex is determined by the rank order of the components of the tangent vector of the +1-geodesic. In general, see [2], finding the limit points is a problem of finding redundant linear constraints. As shown in [8], this can be converted, via duality, into the problem of finding extremal points in a finite dimensional affine space.

3.6 Total positivity and the convex hull

The -1-convex hull of an exponential family is of great interest, mixture models being widely used in many areas of statistical science. In particular they are explored further in [3] in this volume. Here we simply state the main result, a simple consequence of the total positivity of exponential families [12], that, generically, convex hulls are of maximal dimension. In this result, "generic" means that the +1 tangent vector which defines the exponential family has components which are all distinct.



Theorem 4. *The -1-convex hull of an open subset of a generic one dimensional exponential family is of full dimension.*

4 Example

The following example illustrates these results and also shows an application of dimension reduction based on information geometry.

Example 1 (continued). For illustrative purposes, the data is censored at a fixed value such that the censored exponential distribution gives a reasonable, but not perfect, fit. It is assumed the random variable Z has an exponential distribution, but only $Y = \min\{Z, t\}$ is observed. As discussed in [17], this gives a one-dimensional curved exponential family inside a two dimensional regular exponential family.

Figure 1 shows some of the details of the geometry of the curved exponential family which is created after censoring. The censoring value was chosen at 750. The log-likelihood plot, panel (a), shows appreciable skewness, which suggests

that standard first order asymptotics might be improved by the higher order asymptotic methods of classical information geometry. Panel (b) shows the censored exponential (solid curve) embedded in the two-dimensional full exponential family in the +1-parameterization. The dashed contours are the log-likelihood contours in the full exponential family. It is clear, even visually, that there is not much +1 curvature for this family on this inferential scale. So this is an example where the curved exponential family behaves inferentially like a one-dimensional full exponential family. In particular, the dimension reduction techniques found in [16], can be used. Illustrating the effectiveness of this idea, panel (c) shows how well a saddlepoint based approximation does at approximating the distribution of the maximum likelihood estimator of the parameter of interest.

References

1. S.-I. Amari. *Differential-geometrical methods in statistics.* Springer-Verlag, 1990.
2. K. Anaya-Izquierdo, F. Critchley, P. Marriott, and P. Vos. Computational information geometry: theory and practice. *arXiv:1209.1988*, 2012.
3. K. Anaya-Izquierdo, F. Critchley, P. Marriott, and P. Vos. Computational information geometry: mixture modelling. *Proceedings of GSI 2013, LNCS*, 2013.
4. O.E. Barndorff-Nielsen. *Information and exponential families in statistical theory.* John Wiley & Sons, 1978.
5. L.D. Brown. *Fundamentals of statistical exponential families: with applications in statistical decision theory.* Institute of Mathematical Statistics, 1986.
6. M.C. Bryson and M.M. Siddiqui. Survival times: some criteria for aging. *JASA*, 64:1472–1483, 1969.
7. I. Csiszar and F. Matus. Closures of exponential families. *The Annals of Probability*, 33(2):582–600, 2005.
8. H. Edelsbrunner. *Algorithms in combinatorial geometry.* Springer-Verlag: NewYork, 1987.
9. D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families: Graphical models and model selection. *Annals of Statistics*, 29(2):505–529, 2001.
10. D.J. Hand, F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. *A handbook of small data sets.* Chapman and Hall, London, 1994.
11. R.A. Horn and C.R. Johnson. *Matrix Analysis.* CUP, 1985.
12. S. Karlin. *Total Positivity, Vol. I.* Stanford University Press, 1968.
13. R.E. Kass and P.W. Vos. *Geometrical foundations of asymptotic inference.* John Wiley & Sons, 1997.
14. S.L. Lauritzen. *Graphical models.* Oxford University Press, 1996.
15. P. Marriott. On the local geometry of mixture models. *Biometrika*, 89(1):77–93, 2002.
16. P. Marriott and P.W. Vos. On the global geometry of parametric models and information recovery. *Bernoulli*, 10:639–649, 2004.
17. P. Marriott and S. West. On the geometry of censored models. *Calcutta Statistical Association Bulletin*, 52:235–249, 2002.
18. M.K. Murray and J.W. Rice. *Differential geometry and statistics.* Chapman & Hall, 1993.
19. A. Rinaldo. On maximum likelihood estimation in log-linear models. *Tech. Rep. Dep. of Statistics, Carnegie Mellon University*, 2006.
20. M. J. Schervich. *Theory of Statistics.* Springer-Verlag, 1995.