Computational information geometry in statistics: mixture modelling

Computational information geometry in statistics: mixture modelling

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 applies the tools of computation information geometry [3] – in particular, high dimensional extended multinomial families as proxies for the 'space of all distributions' – in the inferentially demanding area of statistical mixture modelling. A range of resultant benefits are noted.

1 Introduction

The application of geometry to statistical theory and practice has produced a variety of different approaches and this paper will involve two of these. The first is the application of differential geometry to statistics, which is often called information geometry. It largely focuses on, typically multivariate, invariant and higher-order asymptotic results in full and curved exponential families through the use of differential geometry and tensor analysis; key references include [1], [5], [6], [16] and [9]. Also included in this approach are consideration of curvature, dimension reduction and information loss, see [8] and [14]. The second important, but completely separate, approach is in the inferentially demanding area of mixture modelling. A mixture model is a discrete or continuous convex combination of distributions from a pre-specified (often, exponential) family: $f(y;Q) = \int h(y|\theta)dQ(\theta)$. Such models arise naturally when certain variables are measured, but other important ones are missing. They are also widely used *per se* in statistical practice, because of their flexibility and interpretability. A major highlight is found in Lindsay's work [12], where convex geometry is shown to give great insight into the fundamental problems of inference in these models and to help in the design of corresponding algorithms. Other differential geometric approaches for mixture models can be found in [15].

This paper aims to show how computational information geometry [3] can provide a link between these two approaches. This brings a range of resultant benefits. In particular, Lindsay's structure is extended in a way which affords considerable advantages, improving computation and avoiding the label-switching problem, while offering better understanding of the variability of the nonparametric maximum likelihood estimate of the mixing distribution. 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.

1.1 Examples

We use the following examples of mixture models throughout for illustration.

Example 1. Mixture of binomial distributions This example comes from [10] where the authors state that 'simple one-parameter binomial and Poisson models generally provide poor fits to this type of binary data', and therefore it is of interest to look in a 'neighbourhood' of these models. The extended multinomial space is a natural place to define such a 'neighbourhood' and a new computational algorithm defined in §2 is used for inference.

Example 2. Tripod model The tripod example is discussed in [17] and [18]. The directed graph is shown in Fig. 1, where there are binary variables X_i, $i = 1, 2, 3$, on each of the terminal nodes, these being assumed independent given the binary variable at the internal node H. In the model, it is assumed H is hidden (i.e. not observed) so the model is a mixture of members of an exponential family. Despite the model's apparent simplicity, the mixture structure can generate multiple modes in the likelihood, illustrating difficult identification issues.

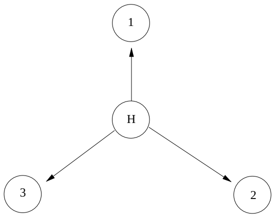

Fig. 1. Graph for Tripod model

2 Inference on Mixtures

2.1 Lindsay's geometry and the simplex

Lindsay's geometry lies in an affine space which is determined by the observed data. In particular, it is always finite dimensional, and the dimension is determined by the number of distinct observations. Following the notation of [11], which looks at mixtures of the model $h(y|\theta)$, i.e. models of the form $f(y;Q) = \int h(y|\theta)dQ(\theta)$, let $L_\theta = (L_1(\theta), \ldots, L_{N^*}(\theta))$ represent the N^* distinct likelihood values of $h(y_i|\theta)$ arising from the data, $\{y_1, \ldots y_n\}$. The likelihood on the space of mixtures is defined on the convex hull of the image of the map

$$\theta \to (L_1(\theta), \ldots, L_{N^*}(\theta)) \subset \mathrm{R}^{N^*}.$$

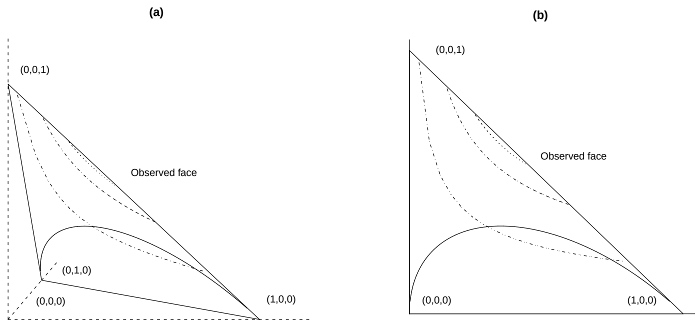

Fig. 2. (a) The simplex with a one-dimensional full exponential family (solid) and likelihood contours (dashed) (b) The image of the simplex under the map $\varPi_L$

Then the problem of finding the non-parametric likelihood estimate, $f(y;\widehat{Q})$, is found by maximising a concave function over this convex set.

There are clear parallels between the convex geometry of Lindsay and the embedding of a model in the -1-simplex, defined by

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

see [3]. Lindsay's geometry is designed for working with the likelihood, so only concerns the observed data rather than the full sample space. For simplicity, consider discrete models where the distinct likelihood components are represented by probabilities π_i where, by definition, i lies in the observed face $\mathcal{P}$ defined in Theorem 2.1 of [2] via the index set of the strictly positive observed counts. The affine structure of Lindsay is thus determined by the vertices of $\mathcal{P}$, see Fig. 2.

Definition 1. *Define $\varPi_L$ to be the Euclidean orthogonal projection from the simplex $\varDelta^k$ to the smallest vector space containing the vertices indexed by $\mathcal{P}$.*

The following result is also strongly connected to Theorem 2.1 of [2]. In it, the level sets of the likelihood are now characterised as the pre-images of the mapping $\varPi_L$. It also shows that searching for the maximum likelihood in the convex hull in the simplex is the same as in Lindsay's geometry.

Theorem 1. *a) The likelihood on the simplex is completely determined by the likelihood on the image of $\varPi_L$. In particular, all elements of the pre-image of $\varPi_L$ have the same likelihood value.*
(b) $\varPi_L$ maps -1 convex hulls in the -1-simplex to the convex hull of Lindsay's geometry.

Given this result, it is natural to study the likelihood of the convex hull of a family in the simplex rather than in Lindsay's space. There are some definite advantages to this, some of which will be explored in this paper, while others will only be briefly mentioned. In Sections 2.2 and 2.3 a new search algorithm is proposed which exploits the information geometry of the full simplex. In particular, it exploits dimension reduction directly in the simplex to give a direct way of computing the non-parametric maximum likelihood estimate. This direct working with mixture distributions has the considerable additional advantage of finessing the label-switching problem encountered by many other methods.

A further advantage of working in the simplex is that while Theorem 1 shows that Lindsay's geometry captures the -1 and likelihood structure, it does not capture the full information geometry. For example, the expected Fisher information cannot be represented, since it is defined using the full sample space, and hence analysis of the variability of the non-parametric maximum likelihood estimate is more natural in the full simplex, rather than in the data-dependent space proposed by Lindsay. This will be looked at in future work.

2.2 Total positivity and local mixing

In order to consider dimension reduction in the -1 simplex, and the corresponding dimension of the convex hull, this paper concentrates on the case where the mixture is over an exponential family. At first sight, Theorem 2 and the following comments may appear contradictory.

Theorem 2 shows that -1-convex hulls of (generic) full exponential families have maximal dimension in the simplex, whereas the concept of local mixing, and its extension to polytope approximation in Theorem 3, shows that there exist very good low dimensional approximations to these convex hulls. It is the existence of these low dimensional approximations which is exploited by the proposed algorithm. Using results on total positivity, we have

Theorem 2. *The -1-convex hull of an open subset of a generic one dimensional exponential family $\pi(\theta)$ is of full dimension.*

In this result "generic" means that the $+1$ tangent vector which defines the exponential family has components which are all distinct.

Nevertheless, Theorem 2 can be contrasted with the results of [13] or [4] which state, under regularity and for many applications, mixtures of exponential families have accurate low dimensional representations.

The essential resolution of this apparent contradiction is that if the segment of the curve $\pi(\theta)$ for $\theta \in \Theta$ lies 'close' to a low dimensional -1-affine subspace, then all mixtures over Θ also lie 'close' to this space. The following discussion is then concerned with the appropriate definition of 'close' for modelling purposes.

Motivated by the idea of a local mixture, consider how well a full exponential family $\pi(\theta)$ can be approximated by a -1 polygonal path which vertices $\pi(\theta_i)$, $i = 1, \ldots, M$. Any point on this polygonal path will have the form

$$\rho\pi(\theta_i) + (1 - \rho)\pi(\theta_{i+1}) \tag{2}$$

with $\rho \in [0,1]$. Define the segment $S_i := \{\rho\pi(\theta_i) + (1-\rho)\pi(\theta_{i+1}) | \rho \in [0,1]\}$. So, we have the identification problem induced by

$$\int \{\rho\pi(\theta_i) + (1-\rho)\pi(\theta_{i+1})\} \, dQ(\rho) = \int \{\rho\pi(\theta_i) + (1-\rho)\pi(\theta_{i+1})\} \, dQ'(\rho) \quad (3)$$

when $E_Q(\rho) = E_{Q'}(\rho)$. While lack of identification is usually considered a statistical problem, computationally it restricts the space the likelihood needs to be optimised over. It will be shown that restricting attention to this space has considerable computational advantages.

Consider, then, the following definition and lemma.

Definition 2. *Given a norm $\|\cdot\|$, the curve $\pi(\theta)$ and the polygonal path $\cup S_i$, define the distance function by, for each θ,*

$$d(\pi(\theta)) := \inf_{\pi \in \cup S_i} \|\pi(\theta) - \pi\| \, .$$

Lemma 1. *If $d(\pi(\theta)) \leq \epsilon$ for all θ then any point in the convex hull of $\pi(\theta)$ lies within ϵ of the convex hull of the finite set $\pi(\theta_i)$.*

Let $\hat{\pi}^{NP}$ be the non-parametric maximum likelihood estimate for mixtures of the curve $\pi(\theta)$. A consequence of Lemma 1 is that, under the uniform approximation assumption, $\hat{\pi}^{NP}$ lies within ϵ of the convex hull of the polygon. The question is then: which norm is appropriate for measuring the quality of its polygonal approximation?

Definition 3. *Define the inner product*

$$\langle v, w \rangle_\pi := \sum_{i=0}^{k} \frac{v_i w_i}{\pi_i}$$

for $v, w \in V_{mix}$ and π such that $\pi_i > 0$ for all i. This defines a preferred point metric as discussed in [7]. Further, let $\|\cdot\|_\pi$ be the corresponding norm.

As motivation for using such a metric, consider the Taylor expansion for the likelihood around $\hat{\pi}$ when the maximum is defined by turning point conditions, i.e. occurs at a point in the relative interior of the simplex. Under these conditions, to high order, it follows that

$$\ell(\pi) - \ell(\hat{\pi}) \approx -\frac{N}{2} \|\pi - \hat{\pi}\|^2_{\hat{\pi}} \, . \quad (4)$$

So small dispersions, as measured by $\|\cdot\|_{\hat{\pi}}$, correspond to small changes in likelihood values. Note that this is clearly not true under the standard Euclidean norm, where unbounded changes in likelihood values are possible.

Following [12], the maximum of the likelihood in a convex hull is determined by the non-positivity of directional derivatives, rather than turning points. So the following likelihood approximation theorem is appropriate.

Theorem 3. *Let $\pi(\theta)$ be an exponential family, and $\{\theta_i\}$ a finite and fixed set of support points such that $d(\pi(\theta)) \leq \epsilon$ for all θ. Further, denote by $\hat{\pi}^{NP}$ and $\hat{\pi}$ the maximum likelihood estimates in the convex hulls of $\pi(\theta)$ and $\{\pi(\theta_i)|i = 1, \ldots, M\}$ respectively, and by $\hat{\pi}_i^G := \frac{n_i}{N}$ the global maximiser in the simplex. Then,*

$$\ell(\hat{\pi}^{NP}) - \ell(\hat{\pi}) \leq \epsilon N ||(\hat{\pi}^G - \hat{\pi}^{NP})||_{\hat{\pi}} + o(\epsilon) \tag{5}$$

2.3 Implementation of Algorithm

Algorithms using the polygonal approximation technique will be evaluated in detail in future work. Here a general outline is given and our two running examples examined. The fundamental idea is to compute the convex hull of a finite number of points on the curve as an approximation to the convex hull of the curve itself. The positioning of the points can be decided by using singular value decomposition methods to see if the $+1$ line segment joining consecutive points has small enough -1 curvature. From these it is necessary to compute ϵ which bounds the uniform approximation of the curve by the polygon and then apply Theorem 3.

The first example implements the theorem for a mixture of binomials.

Example 1 (continued). Consider the data discussed in [10] and shown in part in Table 1. Mixture models are of interest scientifically since the data concerns frequency of implanted foetuses in laboratory animals, and it could be expected that there is underlying clustering. Simple plots shows over-dispersion relative to the variance of a fitted binomial model, which implies that a mixture approach might be appropriate.

Table 1. Observed frequencies of number of dead implants

Number of dead implants	0	1	2	3	4	5	6	7
Frequency	214	154	83	34	25	9	5	0

Using the polygonal approximation approach allows us to compute easily a good approximation to the mixture. The result is shown in Fig. 3. The crosses show the fitted model with circles the data, here with a mixture over $Bin(\pi, 7)$. We also see the mixing proportions and the directional derivative.

Note in this example the near perfect fit of the data with the mixture model. In terms of the simplex this is easily explained since the maximum likelihood estimate in the simplex, in this case, is close to the convex hull of the binomial model.

Example 2 (continued). For this example, the distribution of the random variables at all the observed nodes lies in the $2^3 - 1 = 7$ dimensional simplex, parameterized by the joint probabilities for (X_1, X_2, X_3). If H were observed

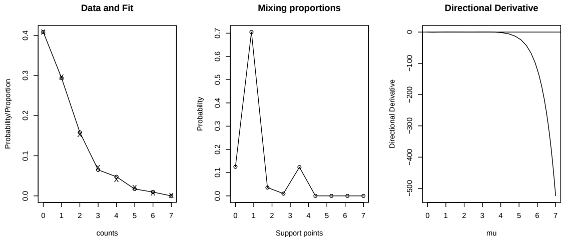

Fig. 3. The mixture fit using polygonal approximation

each node would be independent, so that conditionally on H this space is 3-dimensional, and can be parameterized by the marginal probabilities. It is easy to show that the conditional model includes all 8 vertices of the 7 simplex, intersects six pairs of opposite edges and three pairs of opposite 2-faces. The full tripod model is a two component mixture over the three-dimensional full exponential family. Unlike the full convex hull of Example 1, the two component mixture model need not be convex in the -1-affine space and so can have a complex multimodal likelihood structure. In order to aid visualisation, we also consider here the corresponding bipod model, see Fig. 4

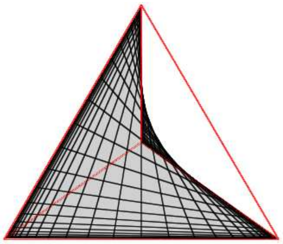

Fig. 4. The bipod model: space of unmixed independent distributions showing the ruled-surface structure.

In the tri- and bi-pod examples, the unmixed model can be approximated with unions of -1-affine polytopes. These can then be used to compute likelihood objects on the two hull – that is the set of all convex combinations of two elements – and on the convex hull very efficiently, just using convex programming. On each polytope, the likelihood has a unique maximum which may, or may not, be on its boundary. To see the whole two hull structure, you just need to glue together this finite number of polytopes and their maxima. Local maxima in the likelihood correspond to internal maxima in the polytopes.

To see how to construct these approximating polytopes, consider Figure 4. The curved surface shown is a, so-called, ruled-surface intersecting the boundary in two pairs of opposite edges. Choose a finite number of support points on each edge of the surface and the same number on the opposed edge. Joining corresponding pairs of points gives a set of -1 convex sets, or slices, close to the surface. Any point in the two hull lies in the convex polytope which is the convex hull of two of these slices.

In summary, this treatment gives a clear computational approach to deal with the complex likelihood structures described in [17] and [18].

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: foundations. *Proceedings of GSI 2013, LNCS*, 2013.
4. K. Anaya-Izquierdo and P. Marriott. Local mixtures models of exponential families. *Bernoulli*, 13(3):623–640, 2007.
5. O.E. Barndorff-Nielsen and D.R. Cox. *Asymptotic techniques for use in statistics.* Chapman & Hall, 1989.
6. O.E. Barndorff-Nielsen and D.R. Cox. *Inference and asymptotics.* Chapman & Hall, 1994.
7. F. Critchley, P. Marriott, and M. Salmon. Preferred point geometry and statistical manifolds. *The Annals of Statistics*, 21:1197–1224, 1993.
8. B. Efron. Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, 3(6):1189–1242, 1975.
9. R.E. Kass and P.W. Vos. *Geometrical foundations of asymptotic inference.* John Wiley & Sons, 1997.
10. L.L. Kupper and J.K. Haseman. The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics*, 34(1):69–76, 1978.
11. M.L. Lesperance and J.D. Kalbfleisch. An algorithm for computing the nonparametric MLE of a mixing distribution. *Journal of the American Statistical Association*, 87:120–126, 1992.
12. B.G. Lindsay. *Mixture Models: Theory, Geometry, and Applications.* Institute of Mathematical Statistics, 1995.
13. P. Marriott. On the local geometry of mixture models. *Biometrika*, 89(1):77–93, 2002.
14. P. Marriott and P.W. Vos. On the global geometry of parametric models and information recovery. *Bernoulli*, 10:639–649, 2004.
15. W. Mio, D. Badlyans, and X. Liu. A computational approach to fisher information geometry with applications to image analysis. *Proceedings of the EMMCVPR*, pages 18–33, 2005.
16. M.K. Murray and J.W. Rice. *Differential geometry and statistics.* Chapman & Hall, 1993.
17. P. Zwiernik and J.Q. Smith. Implicit inequality constraints in a binary tree model. *Electron. J. Statist.*, 5:1276–1312, 2011.
18. P. Zwiernik and J.Q. Smith. Tree-cumulants and the geometry of binary tree models. *Bernoulli*, 18(1):290–321, 2012.