Information Geometry and Its Applications: an Overview

Frank Critchley[1] ⋆ and Paul Marriott[2] ⋆⋆

[1] The Open University, Walton Hall, Milton Keynes, Buckinghamshire, UK MK7 6AA
f.critchley@open.ac.uk
[2] University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada
pmarriot@uwaterloo.ca

Abstract. We give a personal view of what Information Geometry is, and what it is becoming, by exploring a number of key topics: dual affine families, boundaries, divergences, tensorial structures, and dimensionality. For each, (A) we start with a graphical illustrative example, (B) give an overview of the relevant theory and key references, and (C) finish with a number of applications of the theory. We treat 'Information Geometry' as an evolutionary term, deliberately not attempting a comprehensive definition. Rather, we illustrate how both the geometries used and application areas are rapidly developing.

Introduction

This paper is *an* overview of information geometry (IG) and it is important to emphasize that ours is one of many possible approaches that could have been taken. It is, necessarily, a somewhat personal view, with a focus on the authors' own expertise. We, the authors, both have our key interest in statistical theory and practice, and were both strongly influenced, just after its publication, by Professor Amari's monograph, Amari (1985). Recently we, and co-workers, have focused our attention on what we call *computational information geometry* (CIG). This, in all its forms – see, for example, Liu et al. (2012), Nielsen and Nock (2013), Nielsen and Nock (2014), Anaya-Izquierdo et al. (2013a), and Critchley and Marriott (2014a) – has been a significant recent development, and this paper includes further contribution to it. In our conception, CIG provides novel approaches to outstanding, major problems in analysing statistical data. In particular, its (uniquely) operational universal space enables new, computable ways to handle model uncertainty and estimate mixture distributions. For reasons of space, we will be forced to make limited reference to a number of exciting areas in, and related to, IG. In particular: (a) quantum information geometry, where the interested reader could look at Nielsen and Barbaresco (2014) and references therein, (b) Hessian geometries, Shima (2007), and (c) what might be called *sample space information geometry*, including manifold learning, Lee and Verleysen (2007) and statistics on manifolds, Bhattacharya (2008).

This paper is not intended to be an introduction to the area for the complete novice, rather it was written as a keynote address for the workshop 'Computational information geometry for image and signal processing' (*ICMS*, Edinburgh, September 2015), where the audience included many experts in IG with different perspectives. It has always been a problem for us when asked: 'what is the best book to read as an introduction to IG?'. The answer depends very much on what

⋆ This work has been partly funded by EPSRC grant EP/L010429/1

⋆⋆ This work has been partly funded by NSERC discovery grant 'Computational Information Geometry and Model Uncertainty'

the questioner already knows, of course. For example we, the authors, represented two extremes when we started working together: one with no statistical background and one with no differential geometry. One aim of the paper is to point to what we, at least, regard as key references in each of the subject areas. We note, to start with, that there are now a number of volumes in the area of IG, for example the early work in Chentsov (1972) that developed the concept of a statistical manifold, Barndorff-Nielsen (1978), Amari et al. (1987), Dodson (1987), Murray and Rice (1993), Marriott and Salmon (2000), Amari and Nagaoka (2007), Arwini and Dodson (2008), Kass and Vos (2011), Nielsen and Bhatia (2013) and Nielsen (2014).

In this paper, we deliberately do not try to give a formal definition of exactly what information geometry is. Rather, we treat it as an evolutionary term. While IG started as the application of differential geometry to statistical theory, it has – and continues to develop – both with the types of geometry used and in its application areas. Early work was based on, what Amari and Nagaoka (2007) call *dualistic differential geometry* but more recently, wider classes of geometry have been used in IG. For example, links between convex geometry and exponential families are well known, Barndorff-Nielsen (1978); Brown (1986), and their geometric closures have been recognised in IG, see Csiszár and Matus (2005). The importance of affine geometry is explored in this paper in Section 1. We will not have space to explore the exciting links with algebraic geometry but point the interested reader to Pistone et al. (2000), Watanabe (2009) and Gibilisco et al. (2010). Symplectic geometry also plays an important role, Barndorff-Nielsen and Jupp (1997) and recent advances in Markov chain Monte Carlo theory, arising from the seminal paper Girolami and Calderhead (2011), has led to the development of applications of Hamiltonian geometry, see Betancourt (2013) and Betancourt et al. (2014). Of recent interest has been Wasserstein geometry and its links with IG, Takatsu (2013). The geometry of functional analysis also has important applications in non-parametric statistics, for an excellent review see Pistone (2013). In this paper we emphasize how the key geometric objects are not always smooth Riemannian manifolds, but that boundaries, changes in dimension, singularity and unboundedness in tensor fields will all play important roles. We also follow a non-traditional route for defining IG structures; starting with embedding spaces in Section 1, rather than directly with manifolds. See Section 6 for a discussion of this approach.

The conference that motivated writing this paper focused on the applications of IG to image and signal processing, giving examples of applications areas moving away from just statistical theory. Other areas where IG has made an impact include quantum systems, neuronal networks (both biological and artificial), image analysis, and optimization problems.

Throughout this paper we always start each section with (A) a simple – potentially 'toy' – motivating example, which we try and make as visual as possible, returning to this example repeatedly as a concrete illustration. One of the appeals, at least to us, of geometry is its visual aspect and we feel that this can often be lost when ideas become formalised. We follow up this motivating example with (B) a discussion of general theory and point to key references for details and proofs. Each section ends with (C) important examples of the application of the theory.

1 Dual affine families

1.A Illustrative example

Example. For fixed integers m_1, m_2, consider the set of $m_1 \times m_2$ arrays of binary valued pixels. Figure 1 illustrates elements of this state space with a realisation for $m_1 = m_2 = 10$ in Panel (a), while Panel (b) shows the complete state space for the $m_1 = 1, m_2 = 2$ case. Let $(\pi_0, \ldots, \pi_k)$ be a

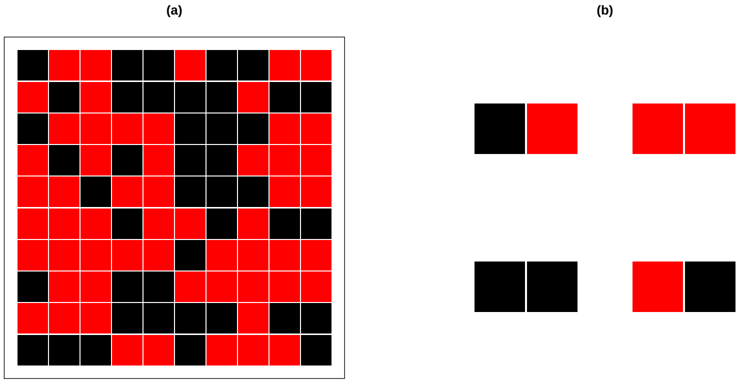

Fig. 1. (a) realisation for 100 binary pixels (b) sample space for 2 binary pixels

probability vector on such a state space, where $k = 2^{m_1 m_2} - 1$. Here, and throughout the paper, we use the weak inequality $\pi_i \geq 0$. The set of all possible probability models is geometrically a closed k-dimensional simplex:

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

Statistically (1) is an extended multinomial family, (Critchley and Marriott, 2014a), which is an example of the closure of an exponential family, studied by Barndorff-Nielsen (1978), Brown (1986), Lauritzen (1996) and Csiszár and Matus (2005).

The sample space for n independent realisations from an extended multinomial distribution is represented by the set of counts $(n_0, \ldots, n_k)$ where $n_i \geq 0$ and $n = \sum_{i=0}^{k} n_i$, and there is the natural correspondence between the sample and model spaces given by the maximum likelihood estimate

$$(\widehat{\pi}_0, \ldots, \widehat{\pi}_k) := \left(\frac{n_0}{n}, \ldots, \frac{n_k}{n} \right) . \tag{2}$$

Why do we insist here on allowing probabilities to be zero? – after all this prevents the geometric objects being manifolds and contradicts the first regularity condition of Amari (1985, p. 16) of distributions having common support. One of the key ideas behind IG is to exploit the link between sample and model spaces – a duality which gives IG its own special flavour – and we want this relationship to be as clean as possible. Since counts in the identification equation (2) can be zero we also want to allow probabilities to have that value. We will also see, later in this paper, how the geometry of the boundary dominates the global IG in the relative interior. Hence explicitly including the boundary makes for a much cleaner analysis.

Example (1.A revisited). For the $m_1 = 1, m_2 = 2$ example both the sample space and the model space can be represented in terms of the 3-simplex, see Fig. 2. The left panel shows the sample space for $n = 3$ with dots representing attainable values. The right panel shows the corresponding parameter space. The red surface in this panel is the set of models where the colour values of the pixels are independent of each other.

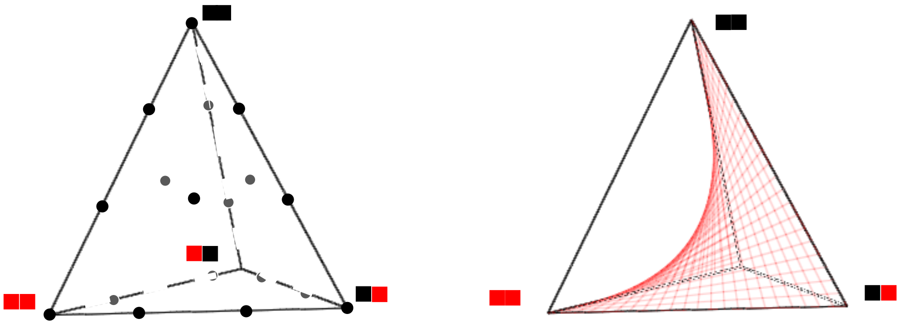

Fig. 2. The 3-simplex: sample, model space and independence subspace. The left plot shows the sample space embedded in the simplex for $n = 3$ by showing with circles the subset of achievable points. The right plot is the model space – the simplex – with the subset of independence models which is a ruled surface. In each plot, the element of the sample space shown in Fig. 1 (b) is shown by the corresponding pair of pixels.

The relative interior of the simplex, $r.i.(\Delta^k)$, is commonly parametrized by $(\pi_1, \ldots, \pi_k)$ – which are (-1)-affine, or expectation, parameters in the terminology of Amari (1985) – or by

$$(\theta_1, \ldots, \theta_k) := \left(\log\left(\frac{\pi_1}{\pi_0}\right), \ldots, \log\left(\frac{\pi_k}{\pi_0}\right)\right),$$

the natural, canonical or $(+1)$-affine parameters.

Key Issue 1 *(Fisher information as change of basis) The matrix of partial derivatives between these smooth parameterisations, of the relative interior, is*

$$\left(\frac{\partial \theta_j}{\partial \pi_i}\right) = \left(\frac{\delta_{i,j}}{\pi_i} + \frac{1}{\pi_0}\right), \tag{3}$$

where $\delta_{i,j} = 1$ if $i = j$, and 0 otherwise. This matrix will be a key tool for moving between representations of geometric objects in the two parameterisations, and we note that it is the Fisher information. Its inverse matrix gives the corresponding inverse transformation.

A parametric statistical model of the set of images can be thought of as a subset of Δ^k, typically selected to have 'nice' mathematical properties. Examples might be that the family is a low dimensional affine subset with respect to the $(+1)$ or (-1)-parameters. For example, the red surface shown in Fig. 2 is the set of independence models, which is an affine subset of the $(+1)$-parameters.

1.B Dual affine parameters

The two types of parameters illustrated above are familiar from the theory of exponential families of the form

$$f(x;\theta) := \nu(x)\exp\left(\langle\theta, S(x)\rangle - \psi(\theta)\right), \tag{4}$$

where $\nu(x)$ is a positive measure, $\theta := (\theta_1, \ldots, \theta_p)^T$ are the natural $(+1)$ parameters, $S(x) := (S_1(x), \ldots, S_p(x))^T$ are the sufficient statistics and $\mu := (E_\theta(S_1), \ldots, E_\theta(S_p))^T$ are the expectation (-1) parameters and $\psi(\theta)$ is the normalising term. The natural parameter space requires definition and is the set

$$\varTheta := \{\theta \mid \psi(\theta) < \infty\}.$$

The boundary behaviour of ψ on this set will play an important role in what follows.

These affine structures are, in fact, much more general than their role in finite dimensional exponential families might suggest.

Key Issue 2 *(Existence of affine structures) There is a natural* $(+1)$*-affine structure on the space of positive measures and a* (-1)*-affine structure on the space of unit measures on a given set. The set of probability measures inherits both structures.*

Murray and Rice (1993) first described the $(+1)$-affine structure in Key Issue 2, while Marriott (2002) shows the existence of a (-1)-affine structure in unit measure space. The intersection of positive and unit measures is, of course, the set of probability measures, thus this space inherits both affine structures. However, we note that the ±1-boundaries, where either positivity (-1) or finiteness (+1) fails, will be important in understanding the underlying geometry of 'distribution space'.

The affine structures defined in Key Issue 2 are particularly important when we look at finite dimensional subsets. For example, Murray and Rice (1993, §1.5.1) show that being a finite dimensional affine subspace of the $(+1)$-affine structure characterises exponential families, while Anaya-Izquierdo and Marriott (2007) show how understanding finite dimensional affine subsets of the (-1)-affine structure explains important identification issues in mixture modelling. An example of a finite dimension subset of $(+1)$-affine space is the independence space plotted in Fig. 2. In the plot it looks 'curved' since the (-1)-affine geometry is used for the illustration.

Key Issue 3 *(Inner product form) Perhaps the crux of understanding duality ideas in IG is the geometric interpretation of the term*

$$\langle\theta, S(x)\rangle := \sum_{i=1}^{p} \theta_i S_i(x), \tag{5}$$

which appears in (4). We have intentionally chosen a suggestive notation which looks like an inner product but, while it is bilinear, the arguments of $\langle\cdot,\cdot\rangle$ *lie in different spaces. The first argument lies in the parameter, or model, space and the second lies in the sample space. Of course, as we have seen these spaces can be closely connected. The* $(+1)$*-affine structure is most 'natural' for the first of these, while the* (-1)*-affine is most 'natural' for the second.*

As illustrated by Example 1.A, these spaces are typically only convex subsets of affine spaces, not affine spaces themselves. However, as also illustrated by Example 1.A, these two spaces have strong links and this gives rise to the principal duality of IG.

There is one instance where all these spaces agree and $\langle \cdot, \cdot \rangle$ is indeed an inner product. This is the statistically very important case of normal linear regression. We can view the structure of classical information geometry as a way of extending the geometric foundation of regression to much more general contexts, see Vos and Marriott (2010).

To give Expression (5) an inner product interpretation we need to make some changes of perspective. Firstly, since we need affine spaces, we work with best linear approximations – tangent spaces – giving each the affine structure described in Key Issue 2. Secondly, we need to be able to map between the $(+1)$-representation of the tangent space and the (-1)-representation. This is the classical change of basis formula from differential geometry, instanced by equation (3) in the multinomial case. In general the change of basis between $(+1)$ and (-1)-coordinates, for exponential families, is the Fisher information matrix, see Section 4. Thus by searching for an inner product interpretation of $\langle \cdot, \cdot \rangle$, the Fisher metric structure has naturally arisen. We denote the Fisher information based inner product at a tangent space by $\langle \cdot, \cdot \rangle_F$.

We have therefore, at least where the underlying models are smooth manifolds, arrived at the classical IG structure described in Amari (1985). We have sets of distributions with enough smooth structure to be manifolds, different but related affine structures, and a change of basis formula which has the properties of being a metric tensor.

Before we briefly review the elegant mathematical structures associated with this structure, we make some observations. Historically an important paper was Lauritzen (1987), which described the structure (M, g, ∇^α) of manifold, metric and family of connections which characterise the affine structures. This united the 'expected' IG of Amari and the 'observed' IG as described in Barndorff-Nielsen (1987). These differ in the choice of metric associated with using unconditional or conditional sample spaces.

Secondly, while we always use the term manifold, much of IG only uses the local geometric structures – that is the tangent space. At least in our experience in statistics, most parameterisations are global and the powerful geometric structure associated with the term manifold – non-trivial topology, local charts, atlas etc – are rarely used. This has been a drawback for practitioners since it appears that there is a bigger overhead of mathematical structure required than is really needed.

Thirdly, there are very simple, but practically important, models in statistics – two component mixtures of exponential distributions for example, Li et al. (2009) – where the Fisher information does not exist and yet there is still a very interesting geometry structure, see Section 4.

Finally, as we saw in Example 1.A – but also in the important classes of mixture, graphical and conditional independence models – boundaries and singularities play a critical role and so these models are not manifolds but do, again, have very interesting geometry.

Key Issue 4 *(The pillars of IG) We can now review the keys pillars of IG. First, we note that we use Fisher information to define a Riemannian structure on the statistical manifold. The affine structures can be characterized by the differential geometric tool of an affine connection ∇, (Amari and Nagaoka, 2007, p. 17). There is a one dimensional family of such connections defined by*

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla^{(+1)} + \frac{1-\alpha}{2}\nabla^{(-1)} \tag{6}$$

(Amari and Nagaoka, 2007, p. 33) for $\alpha \in \mathbb{R}$. Here the $\alpha = \pm 1$ connections agree with the affine structures defined in Key Issue 2. The $\alpha = 0$ connection is also of interest since it is the Levi–Civita connection (Murray and Rice, 1993, p. 115) associated with the Fisher information, see Section 4. The relationship between dual connections and the metric is encoded in the duality relationship

$$X\langle Y, Z\rangle_F = \langle \nabla_X^{(\alpha)} Y, Z\rangle_F + \langle Y, \nabla_X^{(-\alpha)} Z\rangle_F, \tag{7}$$

where X, Y and Z are smooth vector fields, (Amari and Nagaoka, 2007, p. 51). From this relationship we have two fundamental results: the dual flatness theorem, (Amari, 1985, Thm 3.2, p. 72), and the Pythagoras theorem, (Amari, 1985, Thm 3.9, p. 91).

The first of these fundamental results says that if a statistical manifold is α-flat (i.e. there exists a parameterisation in which α-geodesics are defined by affine functions of the parameters) then it is also $-\alpha$-flat. The classic example is the exponential family defined in equation (4), which has θ as $(+1)$-affine parameters and μ as (-1)-affine parameters. This result is very powerful since affine parameters are typically hard to find but very useful; they reduce much of the geometry to that of a Euclidean space. To get a 'free' set of affine parameters is thus excellent news. The dual nature of these affine parameters and the relationship with the metric is also exploited in Section 5. The second result is the Pythagoras theorem and this is discussed in Section 3 once we have introduced the concept of a divergence function.

1.C Application areas

Application Area 1. (Exponential families in Statistics) The primary application of finite dimensional dual affine structures in statistics is, of course, the full exponential family, (Brown, 1986; Barndorff-Nielsen, 1978). The finite dimensional $(+1)$-structure induced by (4) has the property that under i.i.d. sampling the dimension of the sufficient statistic does not change as the sample size increases, meaning that information about the parameters of the model can accumulate with increasing sample size. Closely related are exponential dispersion models (Jorgensen, 1987) which form the probabilistic backbone of generalised linear models, (McCullagh and Nelder, 1989). These are the workhorses of much applied statistical modelling. The generalisation from the standard normal linear model – where $(+1)$ and (-1) structures are indistinguishable – is through the separation of the ± 1-affine structures of exponential dispersion models, Vos and Marriott (2010).

Application Area 2. (Maximum entropy models) Exponential families are also naturally generated through the maximum entropy principle, Jaynes (1978, 1982), Skilling (1989), Buck and Macaulay (1991). The principle of maximum entropy here has strong links with the material on divergences in §3 of this paper, and was motivated by notions of entropy as a measure of uncertainty in both statistical physics and information theory.

Application Area 3. (Curved exponential families) One of the most influential papers in the development of IG was Efron (1975) which first demonstrated that notations of curvature have application in statistical theory. The immediate applications in that paper were to information loss and asymptotic efficiency in inference for a curved exponential family – a submanifold in an exponential family. This class of curved models has important applications in applied statistics including, among many others, Poisson regression, auto-regressive models in time series analysis and common factor models in Econometrics, (Marriott and Salmon, 2000).

Application Area 4. (Graphical models and exponential families) In signal, image and speech processing, one area where the dual affine structure of exponential families has found many applications is through their representation of graphical models. We highlight the paper Wainwright and Jordan (2008) and references in Jordan et al. (2010). Models in these areas can be very high dimensional and direct computation of the normalising constant in Expression (4) – which encodes the full IG structure of such families – can be intractable. The paper points to variational methods in this context, see also Zhao and Marriott (2014) for links with IG.

Application Area 5. (Models in neuroscience) Exponential random graph models (ERGMs) have found important applications in connectivity research in neuroscience, Simpson et al. (2011). The geometry of such models is explored in Rinaldo et al. (2009). Related ideas in belief propagation - a universal method of stochastic reasoning – can be found in Ikeda et al. (2004), while Amari (2015) reviews the IG of, so-called, neural spike data. For related models in neuroscience see Tatsuno and Okada (2003); Tatsuno et al. (2009).

2 Boundaries in Information Geometry

2.A Illustrative example

Example. In the example of modelling sets of binary pixels, consider again the set of independence models, illustrated in Fig. 2. In the case $k = 2$ we can show this space in both its (-1)-affine (Fig. 3(a)) and $(+1)$-affine (Fig. 3(b)) parameters. For the independence model, the expectation parameters are the marginal probabilities of being a colour, π^M. The boundaries for this space are shown in Panel (a) with solid lines.

The relative interior of this space, which is an exponential family, can be parameterised by its natural parameters – the marginal log-odds. We can ask the question of how to represent the boundary in the natural parameters. In Panel (b) we represent this with the red dashed lines 'at infinity'. They represent the 'directions of recession' for this model, Geyer (2009). There is a duality between the two forms of the boundary, with vertices in one representation corresponding to edges in the other, and *vice versa*. To formalize the correspondence between the two we need to understand the closure of the exponential family, Barndorff-Nielsen (1978). That is, what happens to $\theta(\pi^M)$ as at least one component of π^M tends to zero?

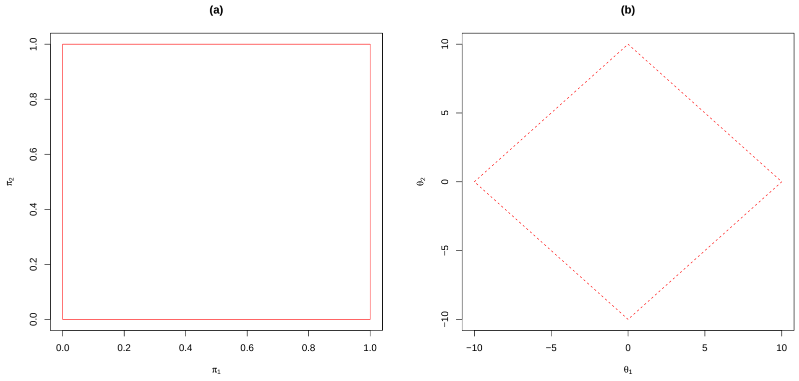

Fig. 3. (a) expectation parameters (b) natural parameters

In our running example, from Section 1.A, a boundary point in model space corresponds to a degenerate distribution. So in the independence model, shown in Fig. 2, boundary points correspond to particular pixels being always the same colour.

2.B Boundaries and polar duals

Key Issue 5 *(Polar duals) We can understand boundary behaviour in extended exponential families by considering the polar dual (Critchley and Marriott, 2014b) or, alternatively, the directions of recession, Geyer (2009); Rinaldo et al. (2009).*

For simplicity we consider discrete p-dimensional exponential families, given by (4), which are subsets of Δ^k described by equation (1). For more general results on closures of exponential families see Barndorff-Nielsen (1978), Brown (1986), Lauritzen (1996) and Csiszár and Matus (2005).

We want to consider the limit points of the p-dimensional exponential family, so we consider the limiting behaviour of the path $\theta(\lambda) := \lambda q$ as $\lambda \to \infty$, where $q \in \mathbb{R}^p$ and $\|q\| = 1$. The support of the limiting distribution is determined by the maximal elements of the set

$$\left\{s_0^T q, \ldots, s_k^T q\right\}$$

where $s_i := (S_0(i), \ldots, S_p(i))^T$. Let $\mathcal{F}_q$ be the set of indices of these maximal elements, so that $1 \leq |\mathcal{F}_q| \leq k+1$. Consider the convex hull, $\mathcal{C}$, of the set

$$\{s_0, \ldots, s_k\} \subset \mathbb{R}^p.$$

The maximum principle for convex functions tells us that $s^T q$ is maximised over the face of $\mathcal{C}$ defined by the vertices $\{s_i | i \in \mathcal{F}_q\}$ and, as Critchley and Marriott (2014b) easily show, q is the normal to the support plane which defines this face. So we have a correspondence between the limiting behaviour of exponential families in a certain direction – the direction of recession – and the set of normals to faces of a convex polygon. The set of outward pointing normals to a polygon is called its polar dual, Tuy (1998).

Example (2.A revisited). In Fig. 3 the polygons in Panels (a) and (b) are polar duals of one another. As the point approaches the boundary in Panel (a) its (+1)-parameters will go to infinity in the direction indicated by the corresponding point on its polar dual.

Often the computation of the boundary polytopes are completely straightforward and there are many cases where the key step, computing the convex hull of a finite number of points in $\mathbb{R}^p$, can be done with standard software. We note however, as the number of parameters and the sample size grows, complete enumeration of the boundary becomes computationally infeasible, see Fukuda (2004).

Key Issue 6 *(Convex geometry) We see here that the key to understanding the closures of exponential families is convex, rather than differential, geometry, and the important geometric objects are convex hulls rather than manifolds. We will also see the important role that convex geometry plays in §3.*

Another place where the dominant geometric tools come from convex geometry is in the analysis of mixture models. A major highlight is found in Lindsay (1995), where convex geometry is shown to give great insight into the fundamental problems of inference in these models and helps in the design of corresponding algorithms. Other differential geometric approaches for mixture models in image analysis can be found in Mio et al. (2005a). Explicit links between this literature and

IG can be found in Anaya-Izquierdo et al. (2013b). The boundaries in this geometry are natural generalisations of the simplest mixture model,

$$\rho f(x) + (1-\rho)g(x),$$

where $\rho \in [0,1]$ with boundaries at $\rho = 0, 1$. Example 7 of Critchley and Marriott (2014a) gives an example of very different statistical behaviour at each boundary point when mixing is between a normal and a Cauchy distribution.

2.C Application areas

Application Area 6. (The finite moment problem) A classical topic in statistics is the moment problem; which distributions can be represented by a finite set of moments? Very early applications of convex geometry in statistical theory can be found in Karlin and Shapley (1953). This work uses convex sets and their conjugate duals to show how moment spaces – sets of achievable moments – are convex bodies whose extreme points can be characterized, often by algebraic means.

Application Area 7. (Boundaries in ERGMs) We have already discussed applications of exponential family random graph models in Application Area 5. The geometry of ERGMs has a number of very interesting features. As pointed out in Geyer (2009) the existence of the maximum likelihood estimate, and corresponding inferences, depends on the boundary behaviour of the closures of the corresponding exponential families. This boundary geometry also dominates the shape of the likelihood and hence also is important in Bayesian inference. Key references here include Rinaldo et al. (2009) and the recent Critchley and Marriott (2014a).

Application Area 8. (Logistic regression) The classical workhorse of statistical modelling with binary data, logistic regression, relies on standard first order asymptotic inference methods using the likelihood. The paper Anaya-Izquierdo et al. (2014) looks at the way that analysing the boundary behaviour of these models generates a simple diagnostic which gives a necessary condition that these first order methods are justified.

Application Area 9. (Marginal polytopes) Connected with these ideas of convex boundaries of exponential families is the idea of a marginal polytope. These are geometric objects associated with any undirected graphical model. They are defined as the set of all marginal probabilities that are realizable under the dependency structure defined by the graphical model. Applications of these geometric ideas can be found in the analysis of Markov Random Fields, which are important in image analysis and many other places. References for this topic include Wainwright and Jordan (2003), Sontag and Jaakkola (2007), and Kahle et al. (2010).

3 Divergences

3.A Illustrative example

The two previous sections looked at basic geometric issues of affineness (i.e. what is a straight line?), convexity, and what happens at boundaries. Section 4 will look at how to measure angles and orthogonality. One major geometric issue not so far mentioned concerns measuring 'distance' in IG and then how to minimize such 'distances'. These questions have been a major driving force in the development of IG, with the following as a key example.

Example. If $f(x;\xi_1)$ and $f(x;\xi_2)$ are two density functions in a parametric model, then we define the Kullback-Leibler divergence, from $f(x;\xi_1)$ to $f(x;\xi_2)$, as

$$K(\xi_1;\xi_2) := E_{f(x;\xi_1)}\left[\log\left(\frac{f(X;\xi_1)}{f(X;\xi_2)}\right)\right], \tag{8}$$

when the expectation exists. Of course, this is not a metric distance, as there is no corresponding triangle inequality and symmetry also fails, Kass and Vos (2011, p. 51). It does, however, have the distance like properties of being greater than, or equal, to zero, with equality if and only if $\xi_1 = \xi_2$.

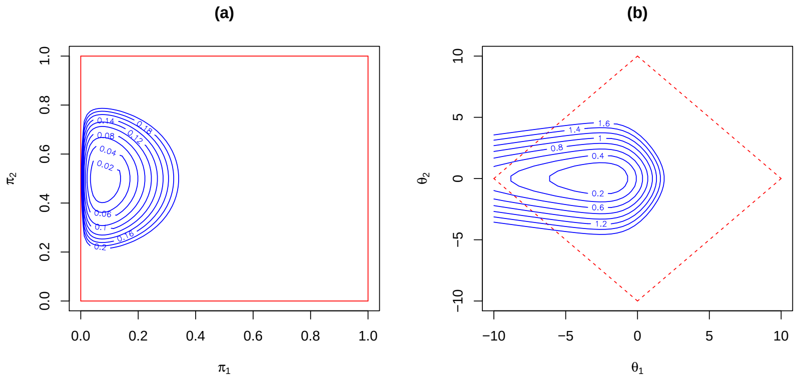

Fig. 4. (a) KL divergence expectation parameters (b) KL divergence natural parameters

Fig. 4 shows concentric KL-spheres in the independence model from the running pixel-based Example 1.A. The level sets are measuring the divergence between two models for the distribution of the pixels in the array. When one of the distributions is degenerate then this distance can be unbounded. As would be expected, from general principles, divergence locally behave qualitatively like the Fisher information spheres of Example 4.A below. This is expected since, locally, this divergence is well approximated by a quadratic form based on the Fisher information.

Further, we see how the boundaries in each model determine the global behaviour of the spheres. In Panel (b) the K-L sphere are stretched ‘to infinity’ in the direction of recession determined by a vertex of the boundary. This vertex is dually equivalent to the edge in Panel (a) which are ‘distorting’ the shape of the spheres.

Key Issue 7 *(Convexity) If a function is going to have distance-like properties then how to minimize it over subsets is a natural question. It is therefore very convenient if the function has nice convexity properties, but since convexity is not invariant to all reparameterisations the link between choice of divergence and the parametrisation used is critical.*

3.B Divergences in IG

While the KL-divergence is very popular, for a number of reasons, it is far from the only possibility. In fact the opposite is true, there is a bewildering number of possible choices which could have been made, depending on what conditions are needed. To help the novice, a useful reference is the annotated bibliography, Basseville (2013) while other important reviews include Kass and Vos (2011, Ch. 9), Cichocki et al. (2009, Ch. 2) and references therein.

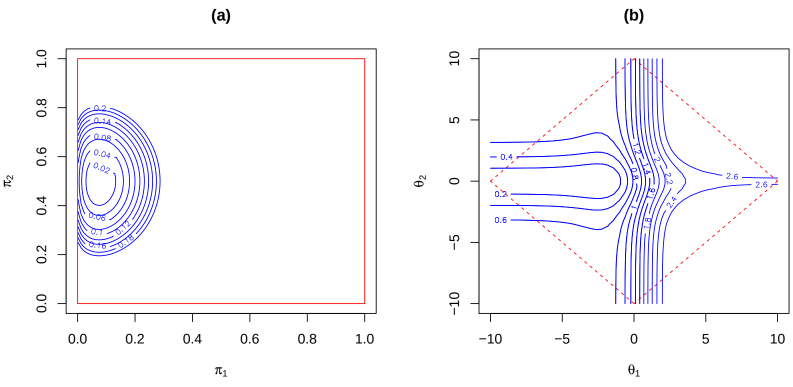

Fig. 5. (a) Dual KL divergence expectation parameters (b) Dual KL divergence natural parameters

One of the most influential developments in basing IG around distance/divergence ideas came from Eguchi et al. (1985), which looked at constructing IG from the point of view of a contrast (divergence) function. Related work can be found in Eguchi et al. (1992); Eguchi (2009); Eguchi et al. (2014). Other important streams of related concepts include: very early work by Csiszár et al. (1967); Csiszár (1975, 1995); asymptotic analysis of related estimators, Pfanzagl (1973); metric based ideas, Rao (1987); the concept of a yoke, Barndorff-Nielsen et al. (1989); Barndorff-Nielsen and Jupp (1997); Blaesild (1991); Barndorff-Nielsen et al. (1994) – which has similar structure to a divergence and also generates IG structures; the relationship with preferred point geometry, Critchley et al. (1994, 1996); and also Zhang (2004), which looks at convexity properties of divergences, f-divergences for affine exponential families Nielsen and Nock (2013), and Belavkin (2013) which looks at optimization problem for measures. A stream of related ideas which was developed rather independently of IG can be found in Cressie and Read (1984); Read and Cressie (2012).

In this paper, for reasons purely of space, we will focus on only one part of this development. In the definition of Bregman (1967), a (Bregman) divergence is a function $D : S \times S \to \mathbb{R}$ where S is a convex set in a linear topological space satisfying certain positivity, projection, convexity and smoothness conditions while, to be precise, the second argument of D should belong to the relative interior of S. Under the conditions of the paper, the function can be expressed using a strictly convex smooth function τ as

$$D_\tau(\xi_1; \xi_2) = \tau(\xi_1) - \tau(\xi_2) - \langle \tau'(\xi_2), \xi_1 - \xi_2 \rangle. \tag{9}$$

for $\xi_i \in S$. Under certain conditions, this can be expressed as

$$D_\tau(\xi_1; \xi_2) = \tau(\xi_1) + \tau^*(\xi_2) - \xi_1^T \xi_2^* \tag{10}$$

where a dual parameter system is defined to ξ by $\xi^*(\xi) := \tau'(\xi)$ and $\tau^*(\xi) := \xi^T \xi^* - \tau(\xi)$ is the Legendre transform when it exists, Rockafellar (1997). We note here that, appropriately interpreted,

(10) is a ‘dualistic form’ of the cosine law. Further, again appropriately dualistically interpreted, (11) below shows that divergence behave like half a squared distance.

Example (3.A revisited). We note that the expression of a divergence in form (10) requires a parameter system ξ and a function which is strictly convex in this parameter system. Since convexity is not invariant to non-linear reparametrisations, each Bregman divergence is associated with particular classes of parameters, called by Kass and Vos (2011, p. 242) the divergence parameter. For the KL divergence in Example 3.A in an exponential family, (4), the expectation parameter is the divergence parameter, since we have

$$K(\mu_1;\mu_2)=\tau(\mu_1)-\tau(\mu_2)-\langle\tau'(\mu_2),(\mu_1-\mu_2)\rangle,$$

where $\tau(\mu):=\theta(\mu)^T\mu-\psi(\theta(\mu))$. The ‘reverse’ KL-divergence, $K^*(\xi_1;\xi_2):=K(\xi_2;\xi_1)$, can be written as

$$K^*(\theta_1;\theta_2)=\tau^*(\theta_1)-\tau^*(\theta_2)-\langle{\tau^*}'(\theta_2),(\theta_1-\theta_2)\rangle$$

where $\tau^*(\theta)=\psi(\theta)$. So, cf. (9), we see the dual affine parameters have corresponding dual divergences.

The divergence spheres for these dual divergences are shown in Fig. 5. The boundaries in each panel are determining the shape of the contours. The vertices in (b) – which correspond to the edges in (a) – are controlling the global shapes associated with the level sets. We also note the lack of convexity in Panel (b) since here the level sets are not being plotted in the affine parameters associated with the Bregmann divergence, Kass and Vos (2011).

For a Bregman divergence in its corresponding affine parametrisation we have the formula

$$D_\tau(\xi_1;\xi_2)+D^*_\tau(\xi_1;\xi_2)=(\xi_1-\xi_2)^T\left(\xi_1^*-\xi_2^*\right). \tag{11}$$

We note the ‘doubly dualistic’ structure of equation 11 where, on the right, we have the dual version of the ‘inner product’ – it might be helpful to refer again to our key equation (5) – and we also have the pair of dual divergences on the left. We can, in fact, build the IG structure of §1 by starting with a pair of dual Bregman divergences and their corresponding dual divergence parameters, see Kass and Vos (2011, §9.3).

3.C Application areas

Application Area 10. (Statistical pattern recognition) The paper Eguchi (2006) looks at ways to apply IG, through a divergence function representation, to statistical pattern recognition. In particular, it looks at boosting algorithms. Boosting is a way of combining the results from simple models, so called weak learners, into a combined result which is much stronger. The paper uses divergences, in this case U-divergences, and their projection properties to construct new boosting algorithms and to give insight into the popular AdaBoost algorithm (Freund and Schapire, 1995). See also Collins et al. (2002) for more links between boosting and divergence functions. Other, more recent, applications to machine learning and signal processing can be found in Takenouchi et al. (2008), Kawakita and Eguchi (2008), and Takenouchi et al. (2012, 2015).

Application Area 11. (Audio stream processing) The paper Cont et al. (2011) applies IG methods, in particular, using Bregman divergences, to build a framework for the analysis of audio signals and shows concrete applications for online audio structure discovery and audio matching.

Application Area 12. (Non-negative matrix factorisation) The book Cichocki et al. (2009) looks at the area of non-negative matrix and tensor factorisation. This is a technique with applications in computer vision, signal processing and many other areas. The mathematical problem is to factorize a 'large' (non-negative) matrix into the product of two 'smaller' (non-negative) matrices. This is often not always possible exactly and so approximation methods are used and measures are needed to measure the size of the error. The geometry found in Cichocki et al. (2009) uses gradient algorithms, often based on different types of divergence to measure the quality of approximation.

We note that the divergences are here not defined on probability spaces, but, rather, on positive measure spaces. This is a good example of how IG has moved beyond the area of probability and statistics.

Application Area 13. (Tsallis entropy) The link between divergence functions and entropy is clear in Example 3.A. The concept of entropy itself has one of its roots in equilibrium statistical mechanics, another being in information theory. Tsallis entropy is a non-additive entropy, which differs from the classical Boltzmann-Gibbs entropy, and has applications in non-extensive statistical mechanics, see Tsallis (1988, 2009) and with a focus on IG issues, Amari and Cichocki (2010) and Amari and Ohara (2011).

4 Tangent spaces and tensors

4.A Illustrative example

Example. The most familiar object which is a tensor in IG is the Fisher information matrix, already discussed in §1. In that section we highlighted its role defining changes of coordinates on tangent spaces as we change parameterisations. It, of course, has an alternative statistical role. If $\ell(\eta; D)$ is the log-likelihood function in some arbitrary parameterisation, when D is the observed data, then the Fisher information matrix for η is

$$Cov_\eta\left[\frac{\partial \ell}{\partial \eta_i}(\eta; D), \frac{\partial \ell}{\partial \eta_j}(\eta; D)\right] \equiv -E_\eta\left[\frac{\partial^2 \ell}{\partial \eta_i, \partial \eta_j}(\eta; D)\right], \tag{12}$$

where Cov denotes the covariance operator. The form of the matrix obviously depends on the choice of parameters, and it is convenient that it has a tensorial transformation rule. We say 'convenient' because it makes it easy to check when objects constructed using tensors have invariant meanings.

In statistics the Fisher Information is familiar since its inverse determines the variance-covariance matrix for the first order asymptotic distribution of the maximum likelihood estimate, Cox and Hinkley (1979). Figure 6 shows for our running example the $p = 2$ dimensional extended exponential family in its expectation and natural parameters. The red line in (a) is the boundary, a polygon, and the corresponding line in (b) is its polar dual. The blue ellipses represent the variability of the maximum likelihood estimates for different data generation distributions across the model. The different 'shapes' and 'scales' in the different parameterisations are given by the tensorial rules of transformation.

Again we note the way that the dual boundaries determine the global behaviour of the shapes of these contours. In Panel (b) the direction of recession is pulling the boundary to infinity, and this vertex corresponds to the edge in Panel (a) which the contours are cutting.

Key Issue 8 *(Two roles of Fisher information) We have seen that the Fisher information has two distinct roles in IG: first, as the key change of basis matrix between expectation and natural parameters and, second in its role in the Cramér-Rao theorem and asymptotic theory.*

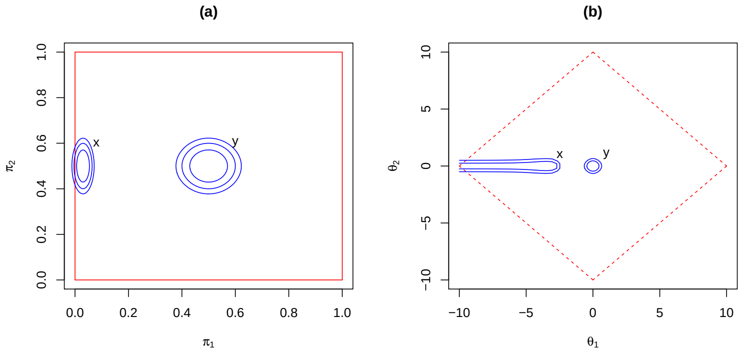

Fig. 6. (a) expectation parameters (b) natural parameters

The Fisher information was recognised to be a Riemannian metric by Rao (1945) and in §4.C we will discuss some aspects to its corresponding geodesics. In statistical theory, outside asymptotic analysis, its key role comes from the famous Cramer-Rao theorem, Cox and Hinkley (1979, p. 254), which gives a bound on the accuracy of estimation of a parameter. Its role in defining the importance of orthogonality in statistical theory was explored in a very influential paper, Cox and Reid (1987).

In the independence case of our running example, all models can be parameterised by the marginal probability of each pixel being a single colour. In this example the Fisher information is diagonal. We could also parameterize by the marginal log-odds of each pixel's colour, and the Fisher information would change by an appropriate tensorial transformation.

Also of interest is the behaviour of the Fisher information near the boundary. This is explored in Anaya-Izquierdo et al. (2014) which shows how first order asymptotic analysis can break down when the boundary is 'close' as measured by the Fisher information. Furthermore, in Critchley and Marriott (2014a), the limiting behaviour of the Fisher information, as it approaches a boundary, is studied by analysing its spectrum.

4.B Tensorial objects

Key Issue 9 *(Invariance) In differential geometry a great deal of attention is paid to understanding the problem of invariance to reparameterisation. The idea here is simply that, at least as far as a geometer is concerned, parameters are just constructs, the manifold is the object of interest, and no results on the manifolds should depend on arbitrary choices. We feel that it is not completely clear that these ideas should be taken without some thought directly into IG in all cases. In statistics it is common that a parameter, such as a mean or probability, has real world meaning in its own right. Indeed this meaning can exist independently of the model selected. In this case we have – what might seem a paradoxical situation to a geometer's eyes – that the parameter is the object of interest while the manifold (model) is the arbitrary construct.*

Nevertheless, the study of invariance has played an important role in the development of IG. Example 4.A has two aspects which are key. Firstly the tensorial nature of the Fisher information and secondly its role in sample size asymptotic expansions.

Good references for the general structure of tensors are Dodson and Poston (2013), which focuses on the geometric aspects of tensorial analysis, and McCullagh (1987), which emphasises their statistical importance. In particular, for a reference to the tensorial properties associated to cumulants, see McCullagh (1987, pp. 57 – 62). The introduction to McCullagh's book is also a good way of learning about the algebraic structure of tensor spaces.

To study the role of asymptotic analysis, in particular its geometrical aspects, good references are Barndorff-Nielsen et al. (1986), Barndorff-Nielsen and Cox (1989), Cox and Barndorff-Nielsen (1994), Barndorff-Nielsen et al. (1994) as well as McCullagh (1987), Murray and Rice (1993, Ch. 9) and Kass and Vos (2011, Ch. 3). This last reference also has material on asymptotic expansions in Bayesian theory.

4.C Application areas

Application Area 14. (Asymptotic expansions) The classical application of IG in statistics is, of course, the asymptotic analysis found in Amari (1985). A representative example is the bias correction of a first-order efficient estimator $\hat{\beta}$ which is defined by

$$b^a(\beta) = -\frac{1}{2n} g^{aa'} \left\{ g^{bc} \Gamma^{(-1)}_{a'bc} + g^{\kappa\lambda} h^{(-1)}_{\kappa\lambda a'} \right\},$$

and has the property that if $\hat{\beta}^* := \hat{\beta} - b(\beta)$ then

$$E_\beta(\hat{\beta}^* - \beta) = O(n^{-3/2}).$$

All terms in this expansion have a direct IG interpretation, see Amari (1985), and their dependence on the choice of parametrisation is made clear. Other important work on the geometry of asymptotic expansions includes Kass (1989), and the books Barndorff-Nielsen and Cox (1989), Cox and Barndorff-Nielsen (1994), and Kass and Vos (2011).

Application Area 15. (Laplace expansions) A related set of work concerns the geometry of the Laplace expansion, which has important applications in Bayesian analysis, Kass et al. (1988), Tierney et al. (1989), Kass et al. (1991), and Wong and Li (1992). Other related work exploiting information geometric properties of the Laplace expansion in mixture models includes Marriott (2002) and Anaya-Izquierdo and Marriott (2007). Other work looking at the local geometry of the likelihood includes Eguchi and Copas (1998).

Application Area 16. (Image analysis) The, so-called, Fisher-Rao geometry which is based on the 0-geodesics of the Fisher metric, has found application in image analysis. We point, in particular to Mio et al. (2005b), Mio and Liu (2006), Lenglet et al. (2006) and Peter and Rangarajan (2006).

Application Area 17. (Model uncertainty) Model uncertainty is a critical problem in applied statistics. The paper Copas and Eguchi (2005) provides an intriguing solution by proposing the 'double the variance' method for addressing the possibility of undetectably small departures from the model. The paper builds local neighbourhoods, using essentially metric based first-order geometric methods, of observationally equivalent models and then studies the inferential effects of working inside this set, which is geometrically a tubular neighbourhood. Much more detail on this area can be found in Anaya-Izquierdo et al. (2016).

Application Area 18. (Infinite Fisher Information) The tensorial structure of IG outside the familiar exponential family can have surprises. The paper Li et al. (2009) shows very simple examples of mixture models – such as a two component mixtures of Poisson or exponential distributions – where the Fisher information does not exist. This means that a great deal of standard statistical methodology does not hold. Nevertheless geometry has a great deal to say about these problems, see for example Morozova and Chentsov (1991), Lindsay (1995) or Anaya-Izquierdo et al. (2013b).

5 Dimensionality and dual parameters

5.A Illustrative example

Example (1.A revisited). Let us return to our running example. We might want to model a binary array of pixels with an independence model, but we may have other modelling assumptions which further reduce the dimension. Accordingly, in Fig. 7 (left hand panel) we illustrate this with a one dimensional exponential family lying in the independence space. As an aside we note the way that such a family, typically, starts and ends at a vertex. Suppose we are interested in a more general model and in the spirit of random effects modelling allow mixing over the one-dimensional family. We show the resulting (-1)-convex hull in the right hand panel. This convex hull is, generically, of full dimension, Critchley and Marriott (2014a). Thus, we have here an example where very low dimensional $(+1)$-objects have very large, indeed maximal, dimensional (-1)-convex hull.

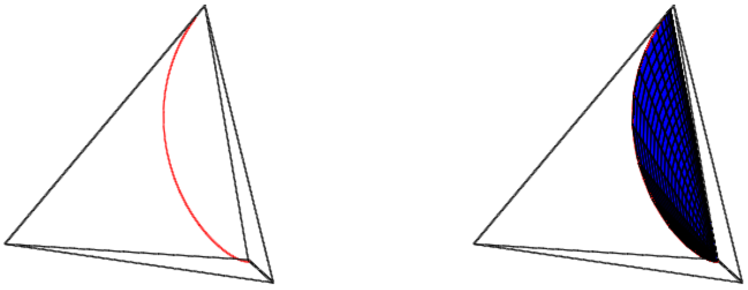

Fig. 7. The 3-simplex: (left) one dimensional exponential family in simplex (right) (-1)-convex hull which represents mixtures over the $(+1)$-family. The convex hull here is a three dimensional subset of the simplex.

Example (1.A revisited). We can also consider the one dimensional family from Example 5.A in another way. Figure 8 shows the one dimensional family considered above, in the two affine parameterisations. In both panels the family is shown by the solid line. The fact that it is an exponential family in its own right from its linearity is clear in from Panel (b). The duality relation,

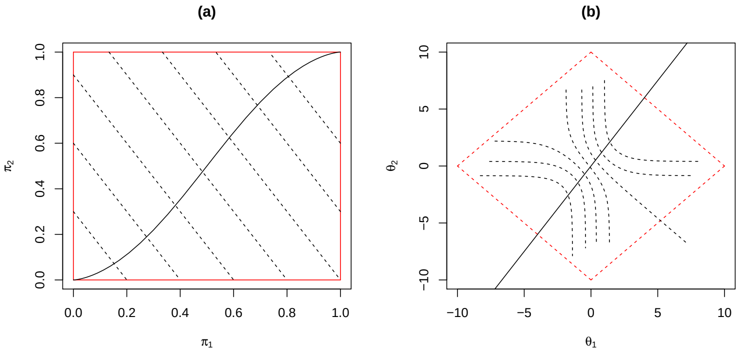

Fig. 8. Mixed parameters in the independence model: (a) expectation parameters and (b) natural parameters

given by equation (7), allows us to define a set of (-1)-flat families which cut the model (Fisher) orthogonally. These are plotted in both plots by the dashed lines.

In the figure again we note the way that the dual boundaries are determining the global structure of the IG. In Panel (b), the (-1)-parallel set of (-1)-geodesics are pulled in the (recession) directions determined by the vertices of the boundary – which are dually equivalent to the edges in Panel (a). The one (-1)-geodesic which passes through a vertex in (a) corresponds to the one cutting an edge in (b).

In terms of our running example, a one-dimensional family of the form shown in Fig. 7 could come from a logistic regression model. This would be a low dimensional $(+1)$-affine subset of the independence space. Mixtures of such families can be derived from random effects models over such logistic regression models, Agresti (2013)

5.B Dual dimensionality

As shown in Critchley and Marriott (2014a), the results in Example 5.A are general. From that paper we have that the (-1)-convex hull of an open subset of a generic one-dimensional exponential family in Δ^k is of full dimension, where generic here means that the one dimensional sufficient statistic for the model has no ties.

Key Issue 10 *(Dimensional duality) We can summarise this by saying that in general low dimensional* $(+1)$*-objects have maximal dimensional* (-1)*-convex hulls in the simplex of distributions. Results such as this follow from total positivity properties of exponential families, Karlin (1968). Such results, despite being classical, probably have not been sufficiently explored in IG.*

The mixed parameterisation of Example 5.A is also very general, see Barndorff-Nielsen and Blaesild (1983), as is the related idea of an inferential cut, (Barndorff-Nielsen and Koudou, 1996), which gives geometric conditions on when inference on subparameters – often called interest parameters – can be achieved independently of the remaining 'nuisance' parameters. See also Pistone et al. (1999). In these constructions, we have a duality relationship between the ±1-affine parts of the construction with the sum of the dimensions being constant. Thus, if one is 'small' the other will be 'big'.

The IG theory which is found in Amari (1985) is based on the differential geometry of finite dimensional manifolds. It is natural to ask if it can be extended to 'infinite dimensional' models, applications to non-parametric statistics being the stand-out motivation, Pistone (2013); see also Morozova and Chentsov (1991). We note that, at least in statistical applications, some thought is required as to what 'infinite dimensional' should mean. For example, in his elegant geometric theory, Lindsay (1995) defines a non-parametric maximum likelihood estimate (NPMLE) in a finite, but data dependent, geometric construction. In applied statistics, at least, the sample size is always finite despite useful tools coming from infinite dimensional ideas, Small and McLeish (2011). Accordingly, a potentially fruitful concept is to think of 'infinite dimensional' as being the case where the dimension is not fixed *a priori*, rather is a function of the data.

Nevertheless, we can still think about the truly unbounded dimensional case, but this needs care. For example, Amari notes the problem of finding an 'adequate topology', Amari (1985, p. 93). There has been work following up this topological challenge.

Key Issue 11 *(Infinite dimensional affine structures) We note that the affine structures defined in Issue 2 are naturally infinite dimensional. Of course, to link them in a standard IG way we need the Fisher information which does not always exist, see Li et al. (2009).*

To try and construct a more complete infinite dimensional IG, Pistone et al. (1999) use the geometry of a Banach manifold and Orlicz spaces – where local patches on the manifold are modelled by Banach spaces. This generates a form of infinite dimensional exponential family, with expectation, natural and mixed parameterisations. Interestingly, as pointed out in Fukumizu (2005), the likelihood function with finite samples is not continuous on the manifold with this Banach structure. He points out that a reproducing kernel Hilbert space structure has a stronger topology and can be usefully employed. Another approach to the infinite dimensional case can be found in Newton (2012). More discussion on infinite versions of the simplex geometry used here as a running example can be found in Critchley and Marriott (2014a); see also Zhang (2013).

5.C Application areas

Application Area 19. (Neural networks) The papers Amari (1995, 1998) look at the way divergences can be used to efficiently fit neural network models. It uses a dual geometric form of the EM algorithm to estimate hidden layers in a neural network. In particular, it exploits the idea of a mixed parameterisation and Fisher orthogonality. Applications in this paper include stochastic multilayer perceptron models, mixtures of experts, and the normal mixture model. Related applications in this area include Amari et al. (1992) and Amari (1997).

Application Area 20. (Image segmentation) Image segmentation is a key step in image analysis. The paper Fu et al. (2013) uses entropy methods in the class of Gaussian mixture models to undertake image segmentation. Related work can be found in Zhang et al. (2013).

Application Area 21. (Multi-scale analysis) The spike train analysis described in Application Area 5 can involve the estimation of intensity functions of point processes. The paper Ramezan et al. (2014) analysed the multi-scale properties of these intensity functions in the spike train context. Here, a critical aspect is the concept of an inferential cut, strongly associated with the IG structure of the mixed parametrisation and discussed above in Example 5.A. Inferential cuts are studied when we want to undertake inference on an interest parameter in the presence of nuisance parameters and, outside of the Bayesian inference approach, this is a difficult question. The work of Kolaczyk and Nowak (2004, 2005) gives the foundation for applying the idea of cuts to a multi-scale analysis of intensity functions of point processes, and in other areas.

Application Area 22. (Non- and semi-parametric modelling) Non-parametric and semi-parametric modelling are very popular approaches in statistical practice and they can be viewed from a geometric perspective. The Hilbert space methodology of Small and McLeish (2011) is closely related to the Hilbert bundle approach of Amari and Kumon (1988) and the geometry of the estimating function approach – often called a semi-parametric method – can also be seen in Amari (1997). We also note the work of Gibilisco and Pistone (1998) and Zhang (2013) in this area. A nice applied example of a Hilbert space approach to interest rate modelling can be found in Brody and Hughston (2001).

6 Closing comments

In this paper we have seen a number of ‘dual’ objects and, albeit without a formal definition, this is a characteristic which enables us to recognise an IG object when we see it. In §1 we have the pair: sample and model (parameter) spaces, in §2 we have a polytope and its polar dual, in §3 we have a divergence and its ‘dual’ where arguments are reversed, in §4 we have tangent and cotangent spaces, and in §5 we have pairs of low dimension $(+1)$-affine spaces, and high dimensional (-1)-convex hulls. We also note the work of Zhang (2006, 2015) which looks at the closely related ideas of reference and representation duality in IG.

One point we would like to make is that to give these objects dual structures, which are truly symmetric, often requires stronger regularity conditions than the user might need, or be able to provide. For example, while the sample/parameter space pairing is attractive in some ways, these are very different objects. For any given sample size, n, the sample space, in the running example of this paper, is a lattice inside a convex set and not a convex set itself. The link becomes clear in the ‘asymptotic limit’, but the user might not have large enough n for this to be at all relevant. Another example is the duality in the divergence section. To have the cleanest links between $D(\cdot;\cdot)$ and $D^*(\cdot;\cdot)$ requires regularity conditions on the Legendre transform, Rockafellar (1997), which can fail in simple examples. Another example is the way that the Fisher information allows the duality seen in §1 and §4 but as Li et al. (2009) illustrate, there are very simple, and useful, statistical models where this object does not exist. What can we take away from these examples? We feel that it would be a mistake to aim for a very elegant mathematical theory *per se*, as attractive as that might be, requiring regularity conditions which contextual considerations indicate to be overly restrictive. Rather, we would like IG to be as inclusive as possible, while still remaining a coherent set of theories.

One issue, that has been a focus of this paper, is the importance of boundaries in IG. In this paper, we concentrated on sets where probabilities are allowed to be zero. In fact, there are other boundaries where normalising constants, or moments, fail to exist. There are important and interesting open questions as to the limiting properties of traditional information geometric objects at these boundaries. Some results in this direction already exist. In Critchley and Marriott (2014a), the behaviour of the Fisher information near a boundary is analysed, while in Critchley and Marriott (2014b) it is shown that the (0)-geodesic (i.e. the minimum path length geodesic) smoothly touches the boundary set.

We have deliberately taken a non-traditional approach to building information geometric structures. It is common, in the literature, to start with the manifold structure of statistical models, defining differential geometric structures, such as metrics and connections, on them. This follows a standard approach in differential geometry, where the geometry of a manifold is defined *implicitly* and independent of any embedding space. This has the advantage for the geometer that they would not have to check that any construction depends on the choice of embedding space. Rather, since there are natural (± 1)-affine embedding spaces, defined in Section 1, we deliberately exploit their simplicity, generality and natural duality. Furthermore, the boundaries which we regard as fundamental, occur completely naturally in this approach.

In this paper, we have taken a personal tour through the emerging subject that is Information Geometry. As we are statisticians, we have mostly focused on applications related to modern statistical practice but, as instanced in the introduction, we note that IG has become a broad church and that there are many other places where it has had an important impact. The general notions of geometric dualistic structures and ideas of divergence that we have seen here are, of course, very widely applicable.

To close, we would like to reiterate some of the key ideas that we have tried to emphasize above. First, we note that the fundamental geometric objects of interest are not always going to be smooth manifolds – boundaries and closures matter. Second, we started our tour with the existence of *very general* affine structures. This is not the only way to build the foundations of IG, of course, but we find it a very attractive one. Third, convexity and other ideas from convex geometry are key in understanding IG structures. This relates to our fourth point, that boundaries of convex sets, and in particular their polar duals, give a great deal of information about the global IG of a problem. Fifth, we note a very attractive duality in dimension inherent in IG that has perhaps not had the attention in the literature that it could have. Sixth, and finally, we note that singularities in tensor fields and boundary effects – which again would not be expected for a geometry based on smooth manifolds – do play an important part in understanding IG as a whole and, we feel, understanding them will be important in moving IG forward.

Bibliography

Agresti, A. (2013). *Categorical data analysis.* Wiley.

Amari, S.-I. (1985). *Differential-geometrical methods in statistics*, Volume 28. Springer-Verlag.

Amari, S.-I. (1995). Information geometry of the EM and em algorithms for neural networks. *Neural networks 8*(9), 1379–1408.

Amari, S.-I. (1997). Information geometry of neural networks – an overview. In *Mathematics of Neural Networks*, pp. 15–23. Springer.

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural computation 10*(2), 251–276.

Amari, S.-I. (2015). Information geometry as applied to neural spike data. *Encyclopedia of Computational Neuroscience*, 1431–1433.

Amari, S.-I., O. E. Barndorff-Nielsen, R. Kass, S. Lauritzen, and C. Rao (1987). Differential geometry in statistical inference. *IMS Lecture Notes-Monograph Series*, i–240.

Amari, S.-I. and A. Cichocki (2010). Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences 58*(1), 183–195.

Amari, S.-I. and M. Kumon (1988). Estimation in the presence of infinitely many nuisance parameters–geometry of estimating functions. *The Annals of Statistics*, 1044–1068.

Amari, S.-i., K. Kurata, and H. Nagaoka (1992). Information geometry of Boltzmann machines. *Neural Networks, IEEE Transactions on 3*(2), 260–271.

Amari, S.-I. and H. Nagaoka (2007). *Methods of information geometry*, Volume 191. American Mathematical Soc.

Amari, S.-I. and A. Ohara (2011). Geometry of q-exponential family of probability distributions. *Entropy 13*(6), 1170–1185.

Anaya-Izquierdo, K., F. Critchley, and P. Marriott (2014). When are first–order asymptotics adequate? a diagnostic. *Stat 3*(1), 17–22.

Anaya-Izquierdo, K., F. Critchley, P. Marriott, and P. Vos (2013a). Computational information geometry: foundations. In *Geometric Science of Information*, pp. 311–318. Springer.

Anaya-Izquierdo, K., F. Critchley, P. Marriott, and P. Vos (2013b). Computational information geometry in statistics: Mixture modelling. In *Geometric Science of Information*, pp. 319–326. Springer.

Anaya-Izquierdo, K., F. Critchley, P. Marriott, and P. Vos (2016). The geometry of model sensitivity: an illustration. In *Computational Information Geometry: For Image and Signal Processing.* Springer.

Anaya-Izquierdo, K. and P. Marriott (2007). Local mixture models of exponential families. *Bernoulli*, 623–640.

Arwini, K. A. and C. T. J. Dodson (2008). *Information Geometry: Near Randomness and Near Independence.* Springer.

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

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

Barndorff-Nielsen, O., D. Cox, and N. Reid (1986). The role of differential geometry in statistical theory. *International Statistical Review/Revue Internationale de Statistique*, 83–96.

Barndorff-Nielsen, O. E. (1987). Differential geometry and statistics: some mathematical aspects. *Indian Journal of Mathematics 29*(3), 335–350.

Barndorff-Nielsen, O. E., P. Blaesild, and M. Mora (1989). Generalized higher-order differentiation. *Acta Applicandae Mathematica 16*(3), 243–259.

Barndorff-Nielsen, O. E. and D. R. Cox (1989). *Asymptotic techniques for use in statistics.* Chapman & Hall.

Barndorff-Nielsen, O. E. and P. E. Jupp (1997). Statistics, yokes and symplectic geometry. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, Volume 6, pp. 389–427.

Barndorff-Nielsen, O. E., P. E. Jupp, and W. S. Kendall (1994). Stochastic calculus, statistical asymptotics, Taylor strings and phyla. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, Volume 3, pp. 5–62.

Barndorff-Nielsen, O. E. and A. E. Koudou (1996). Cuts in natural exponential families. *Theory of Probability & Its Applications 40*(2), 220–229.

Basseville, M. (2013). Divergence measures for statistical data processingan annotated bibliography. *Signal Processing 93*(4), 621–633.

Belavkin, R. V. (2013). Optimal measures and Markov transition kernels. *Journal of Global Optimization 55*(2), 387–416.

Betancourt, M. (2013). A general metric for Riemannian manifold Hamiltonian Monte Carlo. In *Geometric science of information*, pp. 327–334. Springer.

Betancourt, M., S. Byrne, S. Livingstone, and M. Girolami (2014). The geometric foundations of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1410.5110*.

Bhattacharya, A. (2008). *Nonparametric statistics on manifolds with applications to shape spaces.* ProQuest.

Blaesild, P. (1991). Yokes and tensors derived from yokes. *Annals of the Institute of Statistical Mathematics 43*(1), 95–113.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics 7*(3), 200–217.

Brody, D. C. and L. P. Hughston (2001). Interest rates and information geometry. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Volume 457, pp. 1343–1363. The Royal Society.

Brown, L. D. (1986). Fundamentals of statistical exponential families with applications in statistical decision theory. *IMS Lecture Notes-monograph series*.

Buck, B. and V. A. Macaulay (1991). *Maximum entropy in action: a collection of expository essays.* Clarendon Press Oxford.

Chentsov, N. N. (1972). *Statistical decision rules and optimal inference.* Number 53. American Mathematical Soc.

Cichocki, A., R. Zdunek, A. H. Phan, and S.-I. Amari (2009). *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation.* John Wiley & Sons.

Collins, M., R. E. Schapire, and Y. Singer (2002). Logistic regression, Adaboost and Bregman distances. *Machine Learning 48*(1-3), 253–285.

Cont, A., S. Dubnov, and G. Assayag (2011). On the information geometry of audio streams with applications to similarity computing. *Audio, Speech, and Language Processing, IEEE Transactions on 19*(4), 837–846.

Copas, J. and S. Eguchi (2005). Local model uncertainty and incomplete-data bias (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(4), 459–513.

Cox, D. and O. Barndorff-Nielsen (1994). *Inference and asymptotics*, Volume 52. CRC Press.

Cox, D. R. and D. V. Hinkley (1979). *Theoretical statistics.* CRC Press.

Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–39.

Cressie, N. and T. R. Read (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 440–464.

Critchley, F. and P. Marriott (2014a). Computational information geometry in statistics: theory and practice. *Entropy 16*, 2454 –2471.

Critchley, F. and P. Marriott (2014b). Computing with Fisher geodesics and extended exponential families. *Statistics and Computing*, 1–8.

Critchley, F., P. Marriott, and M. Salmon (1994). Preferred point geometry and the local differential geometry of the Kullback-Leibler divergence. *The Annals of Statistics*, 1587–1602.

Critchley, F., P. Marriott, and M. Salmon (1996). On the differential geometry of the Wald test with nonlinear restrictions. *Econometrica: Journal of the Econometric Society*, 1213–1222.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 146–158.

Csiszár, I. (1995). Generalized projections for non-negative functions. *Acta Mathematica Hungarica 68*(1-2), 161–186.

Csiszár, I. et al. (1967). On topological properties of f-divergences. *Studia Sci. Math. Hungar. 2*, 329–339.

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

Dodson, C. T. (1987). *Geometrization of statistical theory: proceedings of the GST Workshop, University of Lancaster Department of Mathematics, 28-31 October 1987.* ULDM Publications.

Dodson, C. T. and T. Poston (2013). *Tensor geometry: the geometric viewpoint and its uses*, Volume 130. Springer Science & Business Media.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, 1189–1242.

Eguchi, S. (2006). Information geometry and statistical pattern recognition. *Sugaku Expositions 19*(2), 197–216.

Eguchi, S. (2009). Information divergence geometry and the application to statistical machine learning. In *Information theory and statistical learning*, pp. 309–332. Springer.

Eguchi, S. et al. (1985). A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Math. J 15*(2), 341–391.

Eguchi, S. et al. (1992). Geometry of minimum contrast. *Hiroshima Math. J 22*(3), 631–647.

Eguchi, S. and J. Copas (1998). A class of local likelihood methods and near-parametric asymptotics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*(4), 709–724.

Eguchi, S., O. Komori, and A. Ohara (2014). Duality of maximum entropy and minimum divergence. *Entropy 16*(7), 3552–3572.

Freund, Y. and R. E. Schapire (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pp. 23–37. Springer.

Fu, W., M. Johnston, and M. Zhang (2013). Gaussian mixture models and information entropy for image segmentation using particle swarm optimisation. In *Image and Vision Computing New Zealand (IVCNZ), 2013 28th International Conference of*, pp. 328–333. IEEE.

Fukuda, K. (2004). From the zonotope construction to the Minkowski addition of convex polytopes. *Journal of Symbolic Computation 38*, 1261–1272.

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

Geyer, C. J. (2009). Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics 3*, 259–289.

Gibilisco, P. and G. Pistone (1998). Connections on non-parametric statistical manifolds by Orlicz space geometry. *Infinite Dimensional Analysis, Quantum Probability and Related Topics 1*(02), 325–347.

Gibilisco, P., E. Riccomagno, M. Rogantin, and H. Wynn (2010). *Algebraic and Geometric Methods in Statistics*. New York, NY: Cambridge University Press.

Girolami, M. and B. Calderhead (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(2), 123–214.

Ikeda, S., T. Tanaka, and S.-I. Amari (2004). Stochastic reasoning, free energy, and information geometry. *Neural Computation 16*(9), 1779–1810.

Jaynes, E. T. (1978). Where do we stand on maximum entropy. *The maximum entropy formalism*, 15–118.

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE 70*(9), 939–952.

Jordan, M., E. B. Sudderth, M. Wainwright, A. S. Willsky, et al. (2010). Major advances and emerging developments of graphical models [from the guest editors]. *Signal Processing Magazine, IEEE 27*(6), 17–138.

Jorgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 127–162.

Kahle, T. et al. (2010). Neighborliness of marginal polytopes. *Contributions to Algebra and Geometry 51*(1), 45–56.

Karlin, S. (1968). *Total positivity*, Volume 1. Stanford University Press.

Karlin, S. and L. S. Shapley (1953). Geometry of moment spaces. *Memoirs of the American Mathematical Society 12*.

Kass, R., L. Tierney, and J. Kadane (1988). Asymptotics in Bayesian computation. *Bayesian statistics 3*, 261–278.

Kass, R. E. (1989). The geometry of asymptotic inference. *Statistical Science*, 188–219.

Kass, R. E., L. Tierney, and J. B. Kadane (1991). Laplaces method in Bayesian analysis. *Contemporary Mathematics 115*, 89–99.

Kass, R. E. and P. W. Vos (2011). *Geometrical foundations of asymptotic inference*, Volume 908. John Wiley & Sons.

Kawakita, M. and S. Eguchi (2008). Boosting method for local learning in statistical pattern recognition. *Neural computation 20*(11), 2792–2838.

Kolaczyk, E. D. and R. D. Nowak (2004). Multiscale likelihood analysis and complexity penalized estimation. *Annals of Statistics*, 500–527.

Kolaczyk, E. D. and R. D. Nowak (2005). Multiscale generalised linear models for nonparametric function estimation. *Biometrika 92*(1), 119–133.

Lauritzen, S. L. (1987). Statistical manifolds. In *Differential geometry in Statistical Science*, pp. 163–216. IMS Hayward, CA.

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

Lee, J. A. and M. Verleysen (2007). *Nonlinear dimensionality reduction*. Springer Science & Business Media.

Lenglet, C., M. Rousson, R. Deriche, and O. Faugeras (2006). Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor MRI processing. *Journal of Mathematical Imaging and Vision 25*(3), 423–444.

Li, P., J. Chen, and P. Marriott (2009). Non-finite fisher information and homogeneity: an em approach. *Biometrika 96*(2), 411426.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry, and Applications.* Institute of Mathematical Statistics.

Liu, M., B. Vemuri, S.-I. Amari, and F. Nielsen (2012). Shape retrieval using heirarchical total Bregman soft clustering. *IEEE Transactions on pattern analysis and machine intelligence 34*, 2407–2419.

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

Marriott, P. and M. Salmon (2000). *Applications of differential geometry to econometrics.* Cambridge University Press.

McCullagh, P. (1987). *Tensor methods in statistics*, Volume 161. Chapman and Hall London.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models*, Volume 37. CRC press.

Mio, W., D. Badlyans, and X. Liu (2005a). A computational approach to Fisher information geometry with applications to image analysis. *Proceedings of the EMMCVPR*, 18–33.

Mio, W., D. Badlyans, and X. Liu (2005b). A computational approach to Fisher information geometry with applications to image analysis. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 18–33. Springer.

Mio, W. and X. Liu (2006). Landmark representation of shapes and Fisher-Rao geometry. In *Image Processing, 2006 IEEE International Conference on*, pp. 2113–2116. IEEE.

Morozova, E. A. and N. N. Chentsov (1991). Natural geometry of families of probability laws. *Itogi Nauki i Tekhniki. Seriya" Sovremennye Problemy Matematiki. Fundamental'nye Napravleniya" 83*, 133–265.

Murray, M. K. and J. W. Rice (1993). *Differential geometry and statistics*, Volume 48. CRC Press.

Newton, N. J. (2012). An infinite-dimensional statistical manifold modelled on Hilbert space. *Journal of Functional Analysis 263*(6), 1661–1681.

Nielsen, F. (2014). *Geometric Theory of Information.* Springer.

Nielsen, F. and F. Barbaresco (2014). *Proceedings of the 2nd International Symposium on Information Geometry and its Applications.* Springer.

Nielsen, F. and R. Bhatia (2013). *Matrix information geometry.* Springer.

Nielsen, F. and N. Nock (2014). Optimal interval clustering: Application to Bregman clustering and statistical mixture learning. *IEEE Transactions on pattern analysis and machine intelligence 21*(10), 1289–1292.

Nielsen, F. and R. Nock (2013). On the chi square and higher-order chi distances for approximating f-divergences. *arXiv preprint arXiv:1309.3029*.

Peter, A. and A. Rangarajan (2006). Shape analysis using the Fisher-Rao Riemannian metric: Unifying shape representation and deformation. In *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*, pp. 1164–1167. IEEE.

Pfanzagl, J. (1973). Asymptotic expansions related to minimum contrast estimators. *The Annals of Statistics*, 993–1026.

Pistone, G. (2013). Nonparametric information geometry. In *Geometric Science of Information*, pp. 5–36. Springer.

Pistone, G., E. Riccomagno, and H. Wynn (2000). *Algebraic Statistics: Computational Commutative Algebra in Statistics.* Chapman and Hall.

Pistone, G., M. P. Rogantin, et al. (1999). The exponential statistical manifold: mean parameters, orthogonality and space transformations. *Bernoulli 5*(4), 721–760.

Ramezan, R., P. Marriott, and S. Chenouri (2014). Multiscale analysis of neural spike trains. *Statistics in medicine 33*(2), 238–256.

Rao, C. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society 37*(3), 81–91.

Rao, C. R. (1987). Differential metrics in probability spaces. *Differential geometry in statistical inference 10*, 217–240.

Read, T. R. and N. Cressie (2012). *Goodness-of-fit statistics for discrete multivariate data.* Springer Science & Business Media.

Rinaldo, A., S. Feinberg, and Y. Zhou (2009). On the geometry of discrete exponential families with applications to exponential random graph models. *Electron. J. Statist. 3*, 446–484.

Rockafellar, R. T. (1997). *Convex Analysis. Princeton landmarks in mathematics.* Princeton University Press, Princeton, NJ.

Shima, H. (2007). *The geometry of Hessian structures*, Volume 1. World Scientific.

Simpson, S. L., S. Hayasaka, and P. J. Laurienti (2011). Exponential random graph modeling for complex brain networks. *PLoS One 6*(5), e20039.

Skilling, J. (1989). Classic maximum entropy. In *Maximum Entropy and Bayesian Methods*, pp. 45–52. Springer.

Small, C. G. and D. L. McLeish (2011). *Hilbert space methods in probability and statistical inference*, Volume 920. John Wiley & Sons.

Sontag, D. and T. S. Jaakkola (2007). New outer bounds on the marginal polytope. In *Advances in Neural Information Processing Systems (NIPS) 20*, pp. 1393–1400.

Takatsu, A. (2013). Behaviors of φ-exponential distributions in Wasserstein geometry and an evolution equation. *SIAM Journal on Mathematical Analysis 45*(4), 2546–2556.

Takenouchi, T., S. Eguchi, N. Murata, and T. Kanamori (2008). Robust boosting algorithm against mislabeling in multiclass problems. *Neural computation 20*(6), 1596–1630.

Takenouchi, T., O. Komori, and S. Eguchi (2012). An extension of the receiver operating characteristic curve and AUC-optimal classification. *Neural computation 24*(10), 2789–2824.

Takenouchi, T., O. Komori, and S. Eguchi (2015). A novel boosting algorithm for multi-task learning based on the Itakuda-Saito divergence. In *Bayesian inference and Maximum Entropy methods in science and engineering (MAXENT 2014)*, Volume 1641, pp. 230–237. AIP Publishing.

Tatsuno, M., J.-M. Fellous, and S.-I. Amari (2009). Information-geometric measures as robust estimators of connection strengths and external inputs. *Neural computation 21*(8), 2309–2335.

Tatsuno, M. and M. Okada (2003). How does the information-geometric measure depend on underlying neural mechanisms? *Neurocomputing 52*, 649–654.

Tierney, L., R. E. Kass, and J. B. Kadane (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association 84*(407), 710–716.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics 52*(1-2), 479–487.

Tsallis, C. (2009). *Introduction to nonextensive statistical mechanics.* Springer.

Tuy, H. (1998). *Convex analysis and global optimization.* Klumer academic publishers: London.

Vos, P. W. and P. Marriott (2010). Geometry in statistics. *Wiley Interdisciplinary Reviews: Computational Statistics 2*(6), 686–694.

Wainwright, M. J. and M. I. Jordan (2003). Variational inference in graphical models: The view from the marginal polytope. In *Proceedings of the annal Allerton conference on communication control and computing*, Volume 41, pp. 961–971. Citeseer.

Wainwright, M. J. and M. I. Jordan (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning 1*(1-2), 1–305.

Watanabe, S. (2009). *Algebraic geometry and statistical learning theory*, Volume 25. Cambridge University Press.

Wong, W. H. and B. Li (1992). Laplace expansion for posterior densities of nonlinear functions of parameters. *Biometrika 79*(2), 393–398.

Zhang, H., Q. Wu, and T. M. Nguyen (2013). Image segmentation by a robust modified gaussian mixture model. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 1478–1482. IEEE.

Zhang, J. (2004). Divergence function, duality, and convex analysis. *Neural Computation 16*(1), 159–195.

Zhang, J. (2006). Referential duality and representational duality on statistical manifolds. In *Proceedings of the Second International Symposium on Information Geometry and Its Applications, Tokyo*, pp. 58–67.

Zhang, J. (2013). Nonparametric information geometry: From divergence function to referential-representational biduality on statistical manifolds. *Entropy 15*(12), 5384–5418.

Zhang, J. (2015). Reference duality and representation duality in information geometry. In *Bayesian inference and Maximum Entropy methods in science and engineering (MAXENT 2014)*, Volume 1641, pp. 130–146. AIP Publishing.

Zhao, H. and P. Marriott (2014). Variational Bayes for regime-switching log-normal models. *Entropy 16*(7), 3832–3847.