

On the Geometric Interplay Between Goodness-of-Fit and Estimation: Illustrative Examples

Karim Anaya-Izquierdo, Frank Critchley, Paul Marriott and Paul Vos

1 Introduction

In statistical analysis, it is common practice to end the model building phase when one, or more, goodness-of-fit tests no longer reject the hypothesis that the data generation process lies in a given parametric model. This model is, often, then treated as known, and parametric inference theory, within it, is assumed to be sufficient to describe the uncertainty in the problem. As a corollary of this, only information captured by the sufficient statistics for the final model is used in the inference. The excellent papers Eguchi and Copas (2005), Copas and Eguchi (2010), summarised below, take a first order geometric approach, which defines the envelope likelihood and a ‘double the variance’ rule, which are designed to capture the actual model uncertainty.

F. Critchley—This work has been partly funded by EPSRC grant EP/L010429/1 and code which generates the figures in this paper is available at <http://users.mct.open.ac.uk/rs23854/EGSS/software.php>.

P. Marriott—This work has been partly funded by NSERC discovery grant ‘Computational Information Geometry and Model Uncertainty’.

K. Anaya-Izquierdo

Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK

e-mail: kai21@bath.ac.uk; K.Anaya-Izquierdo@bath.ac.uk

F. Critchley

The Open University, Walton Hall, Milton Keynes, Buckinghamshire MK7 6AA, UK

e-mail: f.critchley@open.ac.uk

P. Marriott (✉)

University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada

e-mail: pmarriot@uwaterloo.ca

P. Vos

East Carolina University, Greenville, NC 27858-4353, USA

e-mail: VOSP@ecu.edu

© Springer International Publishing AG 2017

F. Nielsen et al. (eds.), *Computational Information Geometry*,

Signals and Communication Technology, DOI 10.1007/978-3-319-47058-0_3

This paper examines the same problem, but uses a global, rather than local, geometric approach. We show how ‘rotations’ and ‘translations’ of working parametric models – which we define using Information Geometric ideas – affect estimation results in ways analogous to those shown by Copas and Eguchi. Further, through a form of bias-variance trade-off, see Hastie et al. (2001), we define, what we call, least-informative families, these being families which, in some sense, add the least amount of information to the estimation problem. These, we show, are connected to ideas from Maximum Entropy theory, Jaynes (1978), Skilling (2013), Schennach (2005), and non-parametric inference methods, with Efron (1981) and Owen (2001) being important references.

Copas and Eguchi (2010) note that in practice the choice of statistical model, made by an analyst, can be rather arbitrary. There may well be other models which fit the data equally well, but give substantially different inferences. We concur with this conclusion which can be summarised as the main theme of the paper: namely, that goodness-of-fit is necessary but not sufficient for model selection. Of course, this is not a new conclusion. In the extreme case, over-fitting of sample data, giving a poor representation of the population, is an extremely well documented phenomenon, Hastie et al. (2001). Rather, this paper points to new, geometrically-based, methodologies to deal with the consequences of this conclusion.

Copas and Eguchi (2010) define *statistically equivalent* models, f and g , to mean that hypothesis tests that the data were sampled from g rather than f would result in no significant evidence one way or another. So, if one model passed a goodness-of-fit test, the other would too. They define the class of statistically equivalent models using first-order asymptotic statistical theory, and hence local linear geometry. They then, building on earlier results in Eguchi and Copas (2005) and similar ideas in Kent (1986), build an envelope of likelihood functions, which gives a conservative inferential framework across the set of statistically equivalent models. A similar idea, in Eguchi and Copas (2005), again using an elegant first order asymptotic and geometric argument, results in the idea of doubling the Fisher information from a single model before calculating confidence intervals to correct for the existence of statistically equivalent models.

One reason that two different analysts may select two different models, both close to the data but giving different inferences about the same parameter, may come from the fact that the models are built with different a priori information. One of the ideas that this paper starts to explore is to what extent can different a priori assumptions be encoded geometrically in some ‘space of all models’. We can then think of the least-informative model as one which, in some sense, uses a minimal amount of extra-data information.

The paper takes an illustrative approach throughout by using low-dimensional models which are simple enough that figures can convey, without misleading, general truths. In particular, we discuss simple *thought experiments*, see Sect. 2, which, while undoubtedly ‘toy’, illustrate clearly the essential points we wish to make. The paper is organised as follows: in Sect. 2 we use some very basic geometric transformation in the ‘space of models’ to explore the relationship between goodness-of-fit tests

and parametric inference. In Sect. 3 we formalise these ideas and define the least informative family. We conclude with a discussion which links the least informative idea with similar ideas in non-parametric inference.

2 Thought Experiments

2.1 Introduction

This section looks at simple geometric concepts, such as translation and rotation, of models in a, in practice, high-dimensional, space of models. We show how these geometric ideas are related to inferential concepts, such as efficiency, information about a parameter, and bias-variance trade-off. The ideas also illustrates relationships between parametric and non-parametric approaches to inference.

We are interested in the global relationship between mean and natural parameters, which we denote by (-1) and $(+1)$ -affine parameters following Amari (1985) and described in Critchley and Marriott (2014). This global approach, explicitly using affine geometries and convex sets, complements that of Copas and Eguchi which is local and first order asymptotic.

To start the discussion, since we want to understand the geometry of sets of models, we first give a formal definition of what we mean by the space of all models, at least in the finite, discrete case. We consider, purely for illustrative reasons, a very simple example where we have a discrete sample space of 3 values: $\{t_0, t_1, t_2\}$. It might be considered natural to consider the space of models as being represented by the multinomial distribution parameterized in the mean parameters by the simplex

$$\Delta_{int} = \left\{ (\pi_0, \pi_1, \pi_2) \mid \pi_i > 0, \sum_{i=0}^2 \pi_i = 1 \right\},$$

and in the natural parameters by

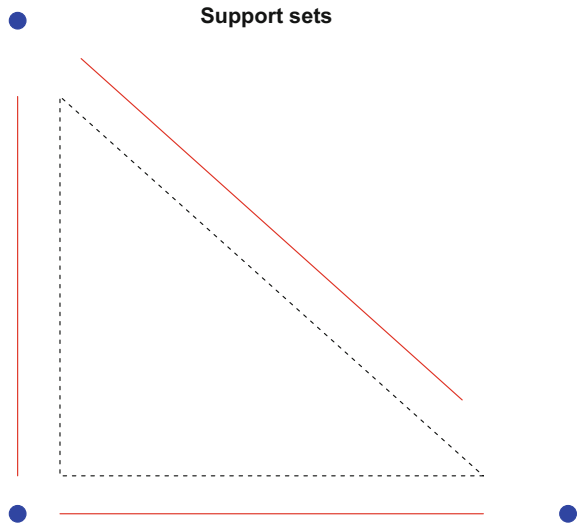
$$\Delta_{int}^* = \left\{ (\eta_1, \eta_2) \mid \eta_i = \log(\pi_i / \pi_0) \right\}.$$

In fact, both geometrically and statistically, it is far neater to work on the closure of the simplex. The global relationship between the (± 1) -parameters is much easier to understand in the closure. Furthermore, when considering non-parametric approaches such as the empirical likelihood, Owen (2001), it is natural to consider the boundary of the closure.

The closure of the multinomial is called the extended multinomial distribution, Critchley and Marriott (2014), with mean parameter space

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

Fig. 1 The extended exponential family as a union of exponential families with different support sets



To define the structure of the ‘natural parameters’ of the closure, Δ^* , we need to use the concept of the polar dual of the boundary of the simplex to define the limiting behaviour, again see Critchley and Marriott (2014). The boundary is a union of exponential families each with corresponding natural (+1)-parameters. The different support sets are illustrated in Fig. 1. To sum up, the space of all models we call a structured extended multinomial (SEM), denoted by $\{\Delta, \Delta^*, (t_0, t_1, t_2)\}$, where the t_i are numerical labels associated with the categories of the extended multinomial. Without loss of generality we assume $t_0 \leq t_1 \leq t_2$. For illustration, in this paper, we take $(t_0, t_1, t_2) = (1, 2, 3)$.

2.2 First Thought Experiment

In our first thought experiment, suppose we are trying to estimate the mean, μ_T , of the random variable T which takes values t_i with probability π_i , $i = 0, 1, 2$. In the simplex, shown in Fig. 2a, sets of distributions with the same mean are (−1)-geodesics and are straight lines in this parameterization. The same sets, in the (+1)-affine parameters of the relative interior of the simplex, are shown in Panel (b), and are clearly non-linear. The global structure of the (−1)-geodesics in the (+1)-affine parameterization is determined by the limit sets defined by the closure and given by the polar dual of the simplex. These can be expressed in terms of the, so-called, directions of recession, Geyer (2009), Feinberg and Rinaldo (2011). The directions associated with the polar dual are illustrated with dashed lines in the panel, again see Critchley and Marriott (2014) for details.

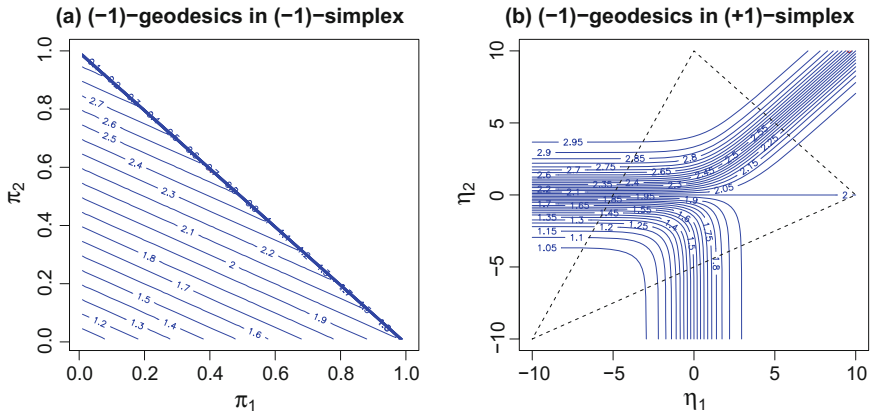


Fig. 2 **a** The extended multinomial in the (-1) -affine parameters, **b** the relative interior of the extended multinomial in the $(+1)$ -affine parameters. The *dash lines* represents the boundary of the closure ‘at infinity’

Consider, then, Fig. 3. In this thought experiment there is a set of models, all one dimensional exponential families, which intersect at the true data generation process, shown as the red circle in panels (a) and (b). These models are shown in panel (b) as straight lines in the $(+1)$ -affine parameters, becoming curves in the (-1) -affine parameters of panel (a). We can, then, think of these models as a set of lines rotating around the data generation process. The red line is the $(+1)$ -geodesic which is Fisher orthogonal to the (-1) -geodesics of interest. In Sect. 3, we will define this as the least-informative model. In our figure we show a set of models, with the most extreme plotted in green for clarity.

For each model the corresponding deviance (twice the normalised log-likelihood) for μ_T , corresponding to each model, and the counts (50, 10, 40), is shown in Fig. 3c. The colour coding here is the same as in panel (b). It is clear that there are considerable differences in inference across this range of models. The vertical scale in panel (c) is selected to show the part of the parameter space of reasonable inferential interest. Since the data generation process lies in each of the models, each should pass any reasonable goodness-of-fit test. Hence, the thought experiment establishes the main theme of the paper: namely, that goodness-of-fit is necessary but not sufficient for model selection.

As discussed above, Fig. 3c shows the set of deviances for our set of models. This is similar to the set of deviances defined in the papers Eguchi and Copas (2005), Copas and Eguchi (2010), where they recommend to use a conservative ‘envelope’ approach. The likelihood which gives the most conservative inference is shown in red and corresponds to the *least informative model* defined and discussed in Sect. 3, and shown by the red curves in (a) and (b). We might extend our thought experiment and imagine the case where one scientist has clear extra-data information which informs the model choice, while a second does not have such information so makes the most conservative choice possible.

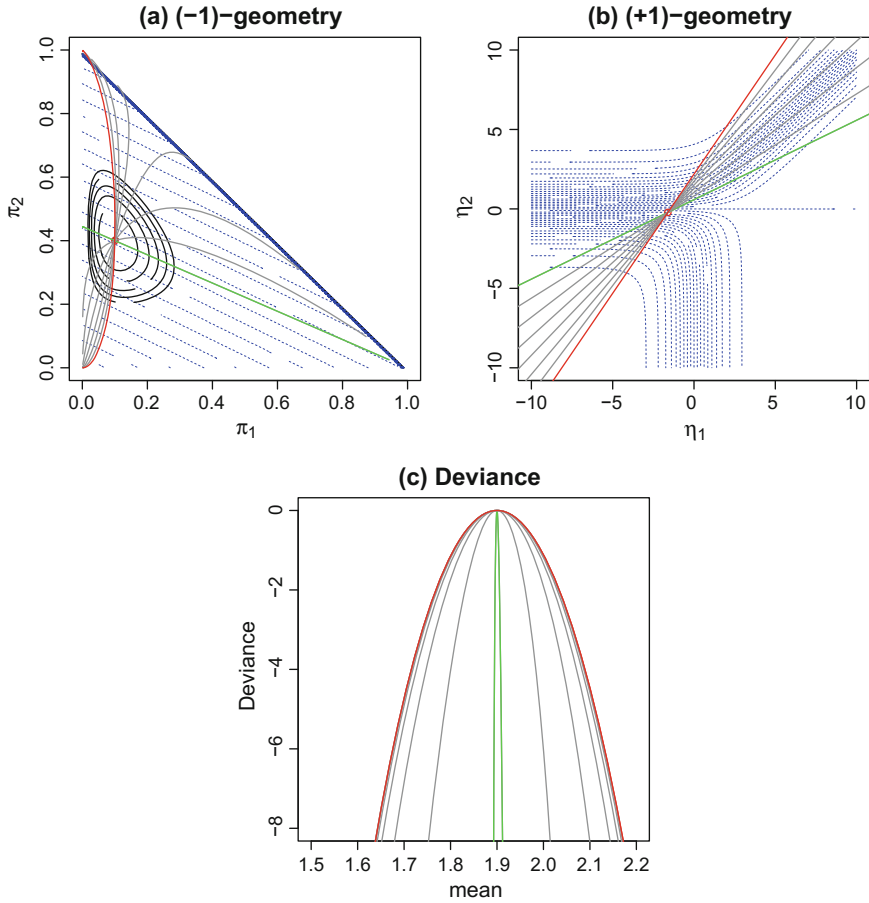


Fig. 3 The *black lines* in panel **a** are the level sets of the log-likelihood for a sample, (50, 10, 40), drawn from the true data generation process, shown as the *red circle* in panels **a** and **b**. **a** The set of models, in the (-1) -affine representation; the *blue lines* are sets of constant parameter values. **b** Same structure in the $(+1)$ -affine representation. **c** The deviance function for the set of models (color figure online)

We also note that adding complexity penalties, as the Akaike information criterion (AIC) or other information criteria do, does not help in the thought experiment since all considered models have the same complexity.

2.3 Second Thought Experiment

The first thought experiment concerns rotations in the $(+1)$ -affine parameters, the second concerns translations, illustrated by Fig. 4. In panel (b) we show a set of $(+1)$ -geodesics. These all have the same sufficient statistic, T , and so their expectation

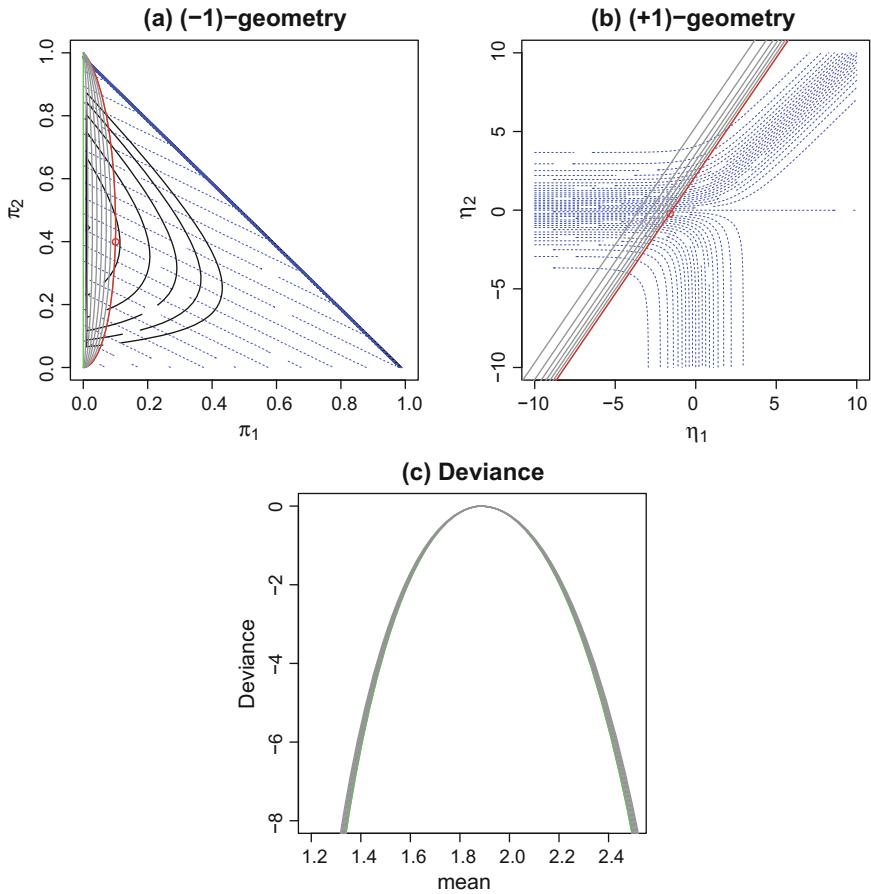


Fig. 4 The data considered here are the counts (10, 0, 8). **a** The set of translated models in the (-1) -affine representation. The *black lines* are the level sets of the log-likelihood, the *blue lines* sets of constant parameter values. **b** The same structure in the $(+1)$ -affine representation. **c** The deviance function for the set of models. Note that the ten different deviance plots are so similar that they are superimposed (color figure online)

parameter is μ_T , the parameter of interest. Since they share a common sufficient statistic, they are all $(+1)$ -parallel in the $(+1)$ -affine parameters. Their corresponding representation in (-1) -affine parameters is shown in panel (a). Here we note that the green line is the limit of this set of translations and lies in the boundary, and so corresponds to a change of support. The data for this example are counts (10, 0, 8) and so the non-parametric maximum likelihood estimate also lies in the boundary. The black curves in Fig. 4a are, as in Fig. 3a, the likelihood contours in the simplex. Panel (c), in Fig. 4, shows the deviance plots for the parameter of interest for this set of models, and it is clear that they all giving essentially the same inference. It can be easily shown that the empirical likelihood for μ_T would also give very similar inference, Owen (2001). In this example we see that the goodness-of-fit does *not*

play an important role in our understanding of the sensitivity of the related inference solution. We call such directions in $(+1)$ -space insensitive, for the inference problem specified. This example illustrates the general fact that the perturbation space of data-supported inferentially sensitive directions may, indeed, be low dimensional. This echoes related results in the companion paper Anaya-Izquierdo et al. (2016): see Sect. 4 for more on links with that paper.

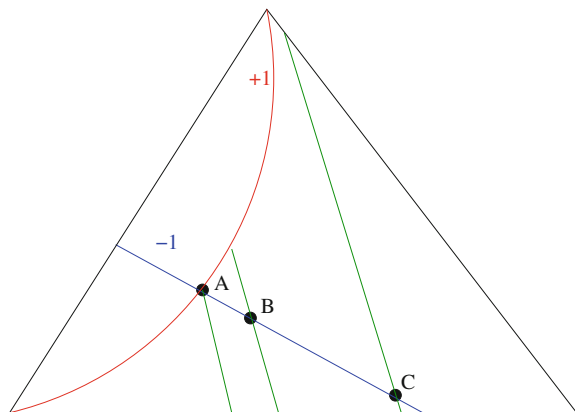
2.4 Third Thought Experiment

The third thought experiment also considered translations shown, this time schematically, in Fig. 5. Assume that we have a *fixed* one dimensional full exponential family, which we plot in Fig. 5 in the (-1) -affine parameters, as a curved line. For example, it might be that an analyst feels a binomial model was appropriate and we denote its sufficient statistic by T . We consider three possible positions for the data generation process, A , B and C . These are selected to lie on the same (-1) -geodesic which is Fisher orthogonal to the working binomial model. This would mean that for each position the pseudo-true value (i.e. the one which minimises the Kullback–Leibler divergence between the model and the data generation process) will be the same, i.e. the point A .

In this example, though, suppose the object of inferential interest was not μ_T but $\mu_S := E(S)$, where S is a different random variable to T . In our thought experiment, the parameter of interest might be, for example, one of the bin probabilities π_i . The sets of distributions which share a common value of $E(S)$ are parallel (-1) -geodesics, and are shown in green. We have three cases to consider.

First, if the data generation process is A then we see that the working model is correctly specified and the value of μ_S from the pseudo-true value and the data generation process are obviously the same. It is the job of the analyst to quantify the variability of the estimate of μ_S and they, naturally, want to use the model that

Fig. 5 A schematic view of bias-variance trade-off



they selected to do this. Given the model we can write μ_S as a, typically non-linear, function of the mean parameter μ_T . By the arguments in the following section, the Fisher information about μ_S is increased by using this approach and the model specification is informative about the inferential question of interest.

Second, consider the case where the data generation process is located at the point B . Above, the model has been actively informative about inference for the parameter of interest. Here, there is a cost. We see that there is considerable bias in the estimate of $E(S)$. Lines of constant values of $E(S)$ are shown in green and the ‘true’ line passes through B . However the pseudo true value is still at A . So, we have a situation where there is a reduction in estimation variance but there is now bias. It is here that goodness-of-fit plays a role. The ‘distance’ between A and B is one of the things that a goodness-of-fit statistic measures and the smaller this is, the smaller the bias.

Third, consider the case where the data generation process is at the point C , considerably further away. This results in an odd situation. We see that the ‘true’ value of μ_S , shown by the green line through C , does not intersect the model at all. Hence, the model is so poorly specified that it cannot estimate the true value of $E(S)$ at all. This is a case where we would expect that goodness-of-fit to play a really important role. Hopefully, it could rule out this working model completely.

3 Least-Informative Models

The three thought experiments have been designed to illustrate, in a visual way the, rather weak, way that goodness-of-fit testing controls the effect of model choice on inference. These models are, of course, toy but the plots represent much more general truths. In this section we move to a much more general statistical analysis of the same problem.

We have explored the effect of different choices of low dimensional, exponential family models in some large model space. We have shown that pure goodness-of-fit tests, and indeed penalty methods, will not give enough information to give unambiguous inference and, in general, methods for taking account of the model uncertainty need to be used. This is in complete agreement with, and gives a global extension to, the results of Eguchi and Copas (2005), Copas and Eguchi (2010). Two questions which naturally arise are, then: why is it that in practical statistical modelling it is very common that low dimensional exponential families are used? and, what sort of information, if not data based, can be used to justify the choice of such low dimensional models?

One justification is through limit theory. The analyst assumes that enough regularity holds such that central or Poisson limit theorems, or similar, hold and uses these to justify the low dimensional assumptions. Such arguments mean that, in the geometry of model spaces, there will be particular directions which are special and some directions in the space of rotations are preferred. Similar arguments can come through assumption that particular equilibrium distributions are appropriate. An example would be assuming a binomial model in Hardy–Weinberg equilibrium

theory, Hardy (1908). Another argument, which also produces low dimensional models is maximum entropy theory, Jaynes (1978), Skilling (2013), Schennach (2005), where distributions are selected which maximise the entropy subject to a set of moment constraints. Again, this would imply that in the (+1)-affine structure, certain directions are special.

In this section, we present a new approach which may give some guidance for the selection of models. We call this the *least-informative model approach*. We work in the general space of k -dimensional discrete distributions, the k -dimensional extended multinomial models, see Critchley and Marriott (2014).

We consider here the one-dimensional subfamily generated by exponential tilting of a fixed distribution $\{\pi_0^0, \pi_1^0, \dots, \pi_k^0\}$ via a real-valued function g . It is easy to see that, as is necessary for ϕ_1^* to parameterise this subfamily, the map defined on \mathbb{R} by

$$\phi_1^* \mapsto \{\pi_i^0 e^{\phi_1^* g(t_i) - \psi(\phi_1^*)}\}_{i=0}^k \quad \text{where} \quad \psi(\phi_1^*) := \log \left(\sum_{i=0}^k \pi_i^0 e^{\phi_1^* g(t_i)} \right)$$

is one-to-one if and only if the $\{g(t_i)\}_{i=0}^k$ are not all equal, which we now assume. The general member of this subfamily assigns cell i ($i = 0, \dots, k$) the probability denoted by:

$$f_g(t_i; \phi_1^*) = P[T = t_i; \phi_1^*] = \pi_i(\phi_1^*) = e^{\phi_1^* g(t_i) - \psi(\phi_1^*)} f_g(t_i; 0), \quad (1)$$

so that the original distribution corresponds to $\phi_1^* = 0$, in which case $\mu_T = \sum_{i=0}^k t_i \pi_i^0$. This is an exponential family with natural parameter ϕ_1^* , mean parameter $\mu := E[g(T)]$ and sufficient statistic

$$s = \sum_{i=0}^k n_i g(t_i)$$

where n_i is the number of times t_i appears in a sample of size $N = \sum_{i=0}^k n_i$ from T .

Our set of rotations, from the above discussion, corresponds to different choices of the function g , according to what the analyst thinks is important, or, equally importantly, has available as a sufficient statistic. Again, the set of translations is basically the choice of the base-line distribution $\{\pi_0^0, \pi_1^0, \dots, \pi_k^0\}$.

In this context, we are considering the case where the inferential problem of interest concerns not the mean of $g(T)$, but the mean of T . For any member of the subfamily (1), this mean is

$$\mu(\phi_1^*) = \sum_{i=0}^k t_i \pi_i(\phi_1^*). \quad (2)$$

If the function $g(t_i) = a t_i + b$ ($a \neq 0$) is affine and invertible, then the map $\phi_1^* \mapsto \mu(\phi_1^*)$ is one-to-one since $\mu_1(\phi_1^*) = a \mu(\phi_1^*) + b$ and $\phi_1^* \mapsto \mu_1(\phi_1^*)$ is one-to-one.

Suppose now that g is any function such that $\phi_1^* \mapsto \mu(\phi_1^*)$ is one-to-one. Now let $\mu \mapsto \phi_1^*(\mu)$ be the inverse map. Then the expected Fisher information about μ in a sample of size one is given by

$$I_g(\mu) = \psi''(\phi_1^*(\mu))[(\phi_1^*)'(\mu)]^2 = \text{Var}_\mu(g(T))[(\phi_1^*)'(\mu)]^2$$

where the subscript μ in Var_μ means the variance is calculated with respect to (1), and $'$ denotes the derivative. But, differentiating (2) with respect to μ we obtain

$$(\phi_1^*)'(\mu) = \frac{1}{\text{Cov}_\mu(T, g(T))},$$

where we make a further regularity assumption that this is also finite. This gives

$$I_g(\mu) = \frac{\text{Var}_\mu(g(T))}{\text{Cov}_\mu^2(g(T), T)}.$$

Denoting by h any invertible affine function of T , as considered above, then

$$I_h(\mu) = \frac{1}{\text{Var}_\mu(T)}.$$

Thus, Cauchy-Schwarz gives at once

$$I_h(\mu) \leq I_g(\mu)$$

and equality holds if and only if g is of the form h . For this reason we call the family (1) the least-informative family for estimation of μ .

We can reconsider the thought experiments in the light of this concept. In Figs. 3 and 4 panels (b) the least informative model corresponds to the red (+1)-geodesic. Under rotations these will have the smallest Fisher information about the parameter of interest, and this can be seen in Fig. 3c. Under translation, as shown in Fig. 4c, there is relative stability in the inferences. Further, there is very good agreement with the empirical likelihood, a model free inference method, Owen (2001).

Each possible choice of one dimensional model introduces information about the parameter of interest that has not come from the data. Therefore one argument would be, if you have no reason to prefer any of one of the set of data supported models select the model which introduces the least amount of extra-data information. In terms of the size of the confidence interval this would be a conservative approach. It is not as conservative as the envelope method, which gives all models in the rotation set equal weight, even if there was no scientific reason for justifying the dimensional reduction for a particular one. We note that the notion of using the most conservative model, rather than averaging inference over sets of models, was advocated by Tukey (1995) in his discussion of Draper (1995).

The third thought experiment, shown in Fig. 5, gives an interesting illustration of what not using least-informative models means. If we have good extra-data reasons for using a non least-informative model – for example a limit result or scientific theory such as Hardy–Weinberg equilibrium – then that information enters the inference problem through an increase in the Fisher information, and thus, smaller confidence intervals and more precise inference. However, there is a cost to this. When the models selected is misspecified then there is a bias introduced into the estimation problem. Hence there is, in this sense a bias-variance trade-off in the model selection choice.

4 Discussion

The concept of least-informative models is related to similar ideas in the literature. For example, moment constrained maximum entropy. A good introduction can be found in the book Buck and Macaulay (1991) and we also note the work in Jaynes (1978), Skilling (2013) and Schennach (2005). In the simplex, we can look for the distribution which maximises entropy, $-\sum_{i=0}^k \pi_i \log \pi_i$, with a given value of the mean of $\mu_T = E(T)$. As the value of μ_T changes the solution set is a least informative family for μ_T which passes through the uniform distribution at the centre of the simplex. In our definition of least informative model, we do not insist that the uniform distribution is part of the model. In practice, goodness-of-fit with the observed data might be an alternative way of selecting which of the (+1)-parallel least informative models to use. Indeed, it would be interesting to explore if this is the actual role that, in some sense, goodness-of-fit should play in model selection. Note that there is also a difference here in that the maximum entropy principle focuses on properties of the underlying distribution, while we are primarily interested in inference about a given interest parameter. It would also be interesting to explore the link between least informative models, which by design cut the level sets of the interest parameter Fisher orthogonally, and the minimum description length approach to parametric model selection, see for example, Balasubramanian (2005).

An obvious question, about selecting low dimensional parametric models, is the following. If you are unsure about the parametric model, why not use non-parametric approaches? In fact, there are close connections between non-parametric approaches and least-informative models. For example, consider Efron (1981), which investigates a set of common nonparametric methods including the bootstrap and jackknife among other methods. It defines the corresponding confidence intervals by basing them on an exponential tilting model. This model, which they called ‘least favourable’ is essentially a least informative model. The least favourable method is also discussed in DiCiccio et al. (1989). Another link with nonparametric methods is through the empirical likelihood function as described in the second thought experiment, see also Murphy and Van der Vaart (2000). The empirical likelihood in this case can be interpreted as a real likelihood on a model which lies in the boundary of the closure of the simplex.

In this paper we have used extensively the (-1) -affine geometry of the extended multinomial model. This, like all affine structures, is global. Our analysis is complementary to the local, and asymptotically based, work of Copas and Eguchi (2010). The choice of the (-1) -structure, aside from its global nature, has some other advantages due to the interpretation of its parameters as expectations, which are model free concepts. As Cox (1986) points out, when we are looking at perturbations of models – for example, as described here in the thought experiments – there are two ways of defining what the ‘same’ parameter means in different models. Firstly, that the parameters have the same real world meaning in different models. We have exploited the fact that (-1) -affine expectation parameters have this property, while $(+1)$ -affine parameters, in general, do not. The second way of connecting parameters in different models, Cox (1961, 1986), regards them merely as model labels and not of intrinsic interest themselves. It operates via minimising some ‘natural measure of distance’, using Cox’s words, between points in the different models. This measure may be naturally suggested by the fitting criterion, as Kullback–Leibler divergence is by maximum likelihood. This would be a natural approach to take if an affine structure other than the (-1) one was used. Further, it might be possible to generalising our results to non-exponential families by using the approximating exponential families of Barndorff-Nielsen and Jupp (1989).

As noted in the second thought experiment (Sect. 2.3), this paper can be viewed as complementary to another paper in this volume, Anaya-Izquierdo et al. (2016). That paper shows, within the space of extended multinomial models, how to iteratively construct a – surprisingly simple (low-dimensional) – space of all important perturbations of the working model, where important is relative to changes in inference for the given question of interest. The iterative search first looks for the directions of most sensitivity. It also carefully distinguishes between possible modelling choices that are empirically answerable and those which must remain purely putative. Unlike the approach taken in this paper, the iterative steps changes the dimension of the model by adding ‘nuisance parameters’ whose role is to inform the inference on the interest parameter.

All examples in this paper are finite discrete models and it is natural to consider extensions to the infinite discrete and continuous cases. The underlying IG for the infinite case is considered in detail in Critchley and Marriott (2014). Section 3 of that paper explores the question of whether the simplex structure, which describes the finite dimensional space of distributions, extends to the infinite dimensional case. Overall the paper examines some of the differences from the finite dimensional case, illustrating them with clear, commonly occurring examples.

The fundamental approach of computational information geometry is, though, inherently discrete and finite, if only for computationally operational reasons. Sometimes, this is with no loss at all, the model used involves only such random variables. In general, suitable finite partitions of the sample space can be used in constructing these computational spaces. While this is clearly not the most general case mathematically speaking (an obvious equivalence relation being thereby induced), it does provide an excellent foundation on which to construct a computational theory. Indeed it has been argued, Pitman (1979)

... statistics being essentially a branch of applied mathematics, we should be guided in our choices of principles and methods by the practical applications. All actual sample spaces are discrete, and all observable random variables have discrete distributions. The continuous distribution is a mathematical construction, suitable for mathematical treatment, but not practically observable.

Since real world measurements can only be made to a fixed precision, all models can, or should, be thought of as fundamentally categorical. The relevant question for a computational theory is then: what is the effect on the inferential objects of interest of a particular selection of such categories?

In summary of this and related papers, it will be of great interest to see how far the potential conceptual, inferential and practical advantages of computational information geometry can be realised.

References

- Amari, S.-I. (1985). *Differential-geometrical methods in statistics*. New York: Springer.
- Anaya-Izquierdo, K., Critchley, F., Marriott, P., & Vos, P. (2016). The geometry of model sensitivity: an illustration. In *Computational information geometry: For image and signal processing*.
- Balasubramanian, V. (2005). Mdl, bayesian inference, and the geometry of the space of probability distributions. In *Advances in minimum description length: Theory and applications* (pp. 81–98).
- Barndorff-Nielsen, O., & Jupp, P. (1989). Approximating exponential models. *Annals of the Institute of Statistical Mathematics*, 41, 247–267.
- Buck, B., & Macaulay, V. A. (1991). *Maximum entropy in action: a collection of expository essays*. Oxford: Clarendon Press.
- Copas, J., & Eguchi, S. (2010). Likelihood for statistically equivalent models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2), 193–217.
- Cox, D. (1986). Comment on ‘Assessment of local influence’ by R. D. Cook. *Journal of the Royal Statistical Society. Series B (Methodological)*, 133–169.
- Cox, D. R. (1961). Tests of separate families of hypotheses. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, 1*, 105–123.
- Critchley, F., & Marriott, P. (2014). Computational information geometry in statistics: theory and practice. *Entropy*, 16(5), 2454–2471.
- DiCiccio, T. J., Hall, P., & Romano, J. P. (1989). Comparison of parametric and empirical likelihood functions. *Biometrika*, 76(3), 465–476.
- Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45–97.
- Efron, B. (1981). Nonparametric standard errors and confidence intervals. *The Canadian Journal of Statistics*, 9(2), 139–158.
- Eguchi, S., & Copas, J. (2005). Local model uncertainty and incomplete-data bias. *Journal of the Royal Statistical Society, Series B, Methodological*, 67, 1–37.
- Feinberg, S., & Rinaldo, A. (2011). Maximum likelihood estimation in log-linear models: Theory and algorithms. [arxiv:1104.3618v1](https://arxiv.org/abs/1104.3618v1).
- Geyer, C. J. (2009). Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, 3, 259–289.
- Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 49–50.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Springer series in statistics (Vol. 1). Berlin: Springer.

- Jaynes, E. T. (1978). Where do we stand on maximum entropy? *The maximum entropy formalism conference*, MIT, 15–118.
- Kent, J. T. (1986). The underlying structure of nonnested hypothesis tests. *Biometrika*, 73(2), 333–343.
- Murphy, S. A., & Van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450), 449–465.
- Owen, A. B. (2001). *Empirical likelihood*. Boca Raton: CRC Press.
- Pitman, E. (1979). *Some basic theory for statistical inference*. London: Chapman and Hall.
- Schennach, S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika*, 92(1), 31–46.
- Skilling, J. (2013). *Maximum entropy and bayesian methods: Cambridge, England, 1988* (Vol. 36). Springer Science & Business Media.
- Tukey, J. W. (1995). Comment on ‘Assessment and propagation of model uncertainty’ by D. Draper. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45–97.