Towards the Geometry of Model Sensitivity: an Illustration

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

[1] Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK kai21@bath.ac.uk
[2] The Open University, Walton Hall, Milton Keynes, Buckinghamshire, UK MK7 6AA
f.critchley@open.ac.uk
[3] University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada
pmarriot@uwaterloo.ca
[4] East Carolina University, Greenville, NC 27858-4353 USA VOSP@ecu.edu

Abstract. In statistical practice model building, sensitivity and uncertainty are major concerns of the analyst. This paper looks at these issues from an information geometric point of view. Here, we define sensitivity to mean understanding how inference about a problem of interest changes with perturbations of the model. In particular it is an example of what we call computational information geometry. The embedding of simple models in much larger information geometric spaces is shown to illuminate these critically important issues.

1 Introduction

This paper is an introduction to a new approach to ubiquitous problems of modelling - in particular model building, sensitivity and uncertainty. By exploring simple, but illustrative, examples we demonstrate that Computational Information Geometry (CIG) delivers both concrete and unexpected results. The key idea is to construct, in a geometrical way, universal operational spaces which allow perturbations of parametric models to be explored and also throws light on the relationship between parametric and non-parametric approaches to inference. We deliberately restrict attention to a particular class of models and type of associated inference problems, see Definition 1, and use them to illustrate a much wider theory which will be explored in related papers.

We positively agree with Box's view of science, put forward in his landmark paper 'Science and Statistics' Box (1976) and developed in Box (1980). In these papers, scientific knowledge is seen as advancing by 'a motivated *iteration* between theory and practice' (his italics), 'efficient scientific iteration evidently requiring unhampered feedback', adding that: 'since all models are wrong the scientist must be alert to what is importantly wrong.' We are therefore developing operational tools to implement these powerful ideas.

Let us assume that we are starting with a well-defined question of interest which we are trying to answer using a set of data, for example wanting to learn about a population mean. We have a working problem formulation – our current statistical model – which has been constructed by using prior knowledge about the experiment and also diagnostic testing to evaluate the adequacy of the model. This model, of course is just one of many that could have been used and we construct an operational universal space (using the tools of high-dimensional extended sparse multinomial models, Anaya-Izquierdo et al. (2013)) which allows us to define the geometry of the 'space of all models'. Within

this space we make extensive use of the tools of CIG (Critchley and Marriott, 2014a) and the inferential ideas of orthogonal, mixed parametrizations, Barndorff-Nielsen and Blaesild (1983), Cox and Reid (1987), and the related idea of an (approximate) cut, Barndorff-Nielsen and Koudou (1995). All these ideas benefit from a direct computational implementation. In particular since the dimension of the operational space could be very large we need computational tools that are adapted thereto, such as linear programming. Within this operational space we iteratively construct a – as it turns out surprisingly simple – 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. For example, observed data may contain a great deal of information about a population mean, but almost none about a high quantile value. In this we follow the principle spelt out in Critchley and Marriott (2004) of 'learn what you can, explore what you can't'. Aspects which must be putative exploration can then inform future scientific experiments.

1.1 A cartoon of modelling

In parametric statistics the question of model specification is a critical one about which there has been a great deal of research. Here we take a new, information geometric approach to the problem. To illustrate ideas we deliberately select simple models and focus on the independent, identically distributed (i.i.d.) case and one particular form of question of interest. Despite this apparent simplicity we show non-trivial geometric results. Other questions of interest, and the much more important for practitioners, case of models with covariates and/or dependent data, will be studied in following papers.

We look at the very general question of where do statistical models come from? To start the exploration Examples 1 and 2 describes two extremes. In the first, information about model specification comes from theoretical considerations.
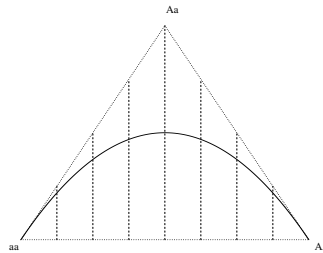

Fig. 1. The Hardy-Weinberg model embedded in the simplex

Example 1. The Hardy-Weinberg model in genetics states that allele frequencies in a population in equilibrium follows a parametric model. The three classes, coded by aa, Aa, and AA, are assumed

to follow a one-dimensional model

$$P(AA) = p^2, P(Aa) = 2pq, P(aa) = q^2,$$

where the marginal probabilities are $P(A) = p, P(a) = q$ and $p = 1 - q$. If we observed frequencies (n_0, n_1, n_2), with fixed $n := n_0 + n_1 + n_2$, the number of independent realisations, then we get a one dimensional exponential family $P((n_0, n_1, n_2)|p)$ with a sufficient statistic $n_0 - n_2$ and natural parameter $\phi := \log(p/(1 - q))$. The mean of the sufficient statistic is $\mu = n(2p - 1)$, and we will consider the case where this, or equivalently p, is of inferential interest. We illustrate this in the space of trinomial models in Fig. 1.

Often, though, models are derived from purely empirical considerations without an underlying theoretical model, thus leaving the problem of finding exactly what the model has contributed to the inference problem.

Example 2. Here we have the question: what is the population mean? To answer this we consider the (simulated) dataset shown in Fig. 2, Panel (a). It is count data with a known support of $[0, 40]$ and an analyst considered that it might be plausibly modelled by a binomial distribution. However, after fitting it is noted that there is some over-dispersion relative to the binomial.

Figure 2 (b) shows the fitted distribution corresponding to a binomial asumption. Panel (c) shows one way of undertaking the inference problem, given we accept the model, using the asymptotic distribution of the deviance – i.e. twice the log-likelihood function. For this example the sample size and simplicity of the model mean first order asymptotic arguments are appropriate. In this panel we contrast the model based inference with a model free ones, such as the t-test, justified here on asymptotic grounds, or the empirical likelihood (Owen, 1988), which is shown in the panel with the dashed line. Here the empirical likelihood is computed by profiling over the whole multinomial; see Appendix 2 for details. We can clearly see how much the choice of model is contributing to the inference question by comparing these two methods. We clarify that we are not, here, judging which one of these methods is "best", rather just noting that there are significant differences which the analyst needs to be aware of. In this toy example we see the model choice is affecting the uncertainty associated with the estimate by a moderate amount. It is one of the aims of this paper to show how the geometry of the model can be used to understand the effect of model choice.

Above we have discussed so-called 'model free methods' whose justification is through asymptotic analysis. We would like to make the point here that most asymptotic arguments are not uniform across the simplex, (Anaya-Izquierdo et al., 2014). That is for a given sample size the quality of the asymptotic approximation depends on where we are in the simplex. Thus, strictly speaking, these methods are not truly 'model free', nevertheless they do provide a sensible and practical base line for comparison.

Suppose we have a working, putative model and we want to check that the model is concordant with the data. One general approach is to perform an appropriate goodness of fit test such as Kolmogorov-Smirnov or Cramer Von Mises. An alternative general approach, the one we follow here, is to build a larger model, or to perturb the original. This is a common approach and we highlight in particular, Box (1980), Cook (1986) and Critchley and Marriott (2004). Of particular interest is the observation in Cox (1986) who points out the importance of assessing that a parameter in a larger model has a meaning which is consistent with that in the smaller model. We want to make sure we are always comparing 'apples' with 'apples'. For that reason we will focus on attention on inference about quantities, such as population means, which have a 'model free' meaning.
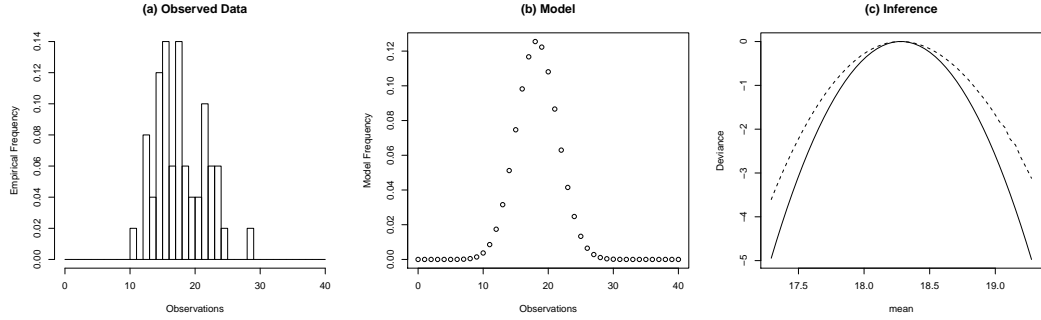
4

Fig. 2. The data for Example 2: Panel (a) shows the empirical distribution of observed data, (b) shows the fitted binomial model, (c) show the model based deviance (solid line) and the empirical deviance (dashed line)

In this initial, and exploratory, paper we will only look at the following class of models. We fully understand that, outside the classroom, it would be unusual for all of the regularity conditions to hold but it is common in practice that a substantial number will and hence we can learn a lot by exploring this basic class of models. Examples include maximum entropy and random graph models

Definition 1. *All models in this paper satisfy the following regularity conditions: (a) all models are for discrete and finite random variables, (b) the observed data is independently and identically distributed, (c) the putative working model is a regular exponential family, (d) the parameter of inferential interest is the mean of a statistic, s, and (e) this statistic is part of the sufficient statistic.*

Simple examples of such families include the distribution of a random vector X, where the probability vector $(P(X = x_i))_{i=0}^k$ is given by

$$\left(\pi_i^0 \exp(\phi s(x_i) - M(\phi))\right)_{i=0}^k \tag{1}$$

in which the inferential question of interest concerns $\mu = \mu(\phi) := E_\phi[s(X)]$.

We look at the sensitivity of the inferential answer to perturbations of Model (1). In particular we might, for example, perturb π_i^0 via

$$\pi_i^0 \to \pi_i^0 + \delta\omega_i =: \pi_i^0(\delta) \tag{2}$$

where $\sum_{i=0}^k \omega_i = 0$ and ω has unit length with respect to the Fisher information at π^0. We also look at extending the sufficient statistic via a larger model of the kind

$$\left(\pi_i^0 \exp(\phi_1 s(x_i) + \phi_2 s_2(x_i) - M(\phi_1, \phi_2))\right)_{i=0}^k \tag{3}$$

while keeping the inferential question about $E(s(X))$ fixed in both cases.

One property of perturbations (2) and (3) is that the maximum likelihood estimate of μ is always the sample mean $\frac{\sum_{j=1}^N s(x_j)}{N}$, where N is the sample size. Thus, for purely pointwise estimation, these perturbations play no role. We study them to investigate other aspects of inference, such as quantifying the uncertainty in the estimate and understanding the role of 'outliers'.

2 The geometry of model sensitivity

In Appendix 1 we review some key concepts that are used in the analysis below, and give references for the interested reader. Nothing in there is new and those familiar with extended exponential families can move on without loss.

The idea of an inferential cut (Barndorff-Nielsen, 1976; Barndorff-Nielsen and Koudou, 1995) is key motivation for what follows. These are studied when we want to undertake inference on an interest parameter in the presence of nuisance parameters. Outside of the Bayesian inference approach this is a difficult question but important for understanding our perturbation approach to sensitivity analysis. Starting with a baseline model we will often be extending it, making it more flexible, at the cost of adding 'nuisance' parameters.

2.1 Approximate cuts

Let

$$\mathcal{F} = \left\{ f_s(s; \phi) = \exp\left(s^T \phi - M(\phi)\right) f_s(s; 0) \,:\, \phi \in \mathcal{P} \right\}$$

be a regular natural exponential family with respect to some fixed σ-finite measure ν on $\mathbb{R}^k$. The mean parameter function will be denoted by $\mu(\phi) := D_\phi M(\phi) = E[s; \phi]$. We will use the following notation

$$s = \left(s_1, s_{(1)}\right)^T, \qquad \mu = \left(\mu_1, \mu_{(1)}\right)^T, \qquad \phi = \left(\phi_1, \phi_{(1)}\right)^T,$$

where s_1, μ_1, ϕ_1 are of dimension r, the $\phi_{(1)}$ notation means exclude the elements in ϕ_1, so that $s_{(1)}, \mu_{(1)}, \phi_{(1)}$ are of dimension $k - r$. The following definition, and much more detail, can be found in Barndorff-Nielsen and Blaesild (1983).

Definition 2. *(Mixed Parameterisation) For a regular exponential family $\mathcal{F}$, the map*

$$\phi \mapsto \begin{pmatrix} \mu_1(\phi) \\ \phi_{(1)}(\phi) \end{pmatrix}$$

is a diffeomorphism on $\mathcal{P}$ with range $\mu_1(\mathcal{P}) \times \phi_{(1)}(\mathcal{P})$. The parameterisation $(\mu_1, \phi_{(1)})$ is called the mixed parameterisation of $\mathcal{F}$.

Definition 3. *(Cut) The statistic s_1 is said to be a cut for the regular exponential family $\mathcal{F}$ if an only if*

$$f_s(s; \mu_1, \phi_{(1)}) = f_{s_1}(s_1; \mu_1) f_{s_{(1)}|s_1}(s_{(1)} \mid s_1; \phi_{(1)})$$

for all $s, \mu_1, \phi_{(1)}$.

In the presence of a cut, if μ_1 is of interest, we can make inferences about it (without knowledge of $\phi_{(1)}$) using only the marginal distribution of s_1. Analogously, if $\phi_{(1)}$ is of interest, we can make inferences about it (without knowledge of μ_1) using only the conditional distribution of $s_{(1)}$ given s_1.

The following structural result about the existence of a cut can be found in Barndorff-Nielsen and Koudou (1995).

Theorem 1. *Let $\mathcal{F}$ be a regular exponential family. The following are equivalent:*

1. s_1 is a cut for $\mathcal{F}$

2. *The variance of s_1 depends only on μ_1*
3. *s_1 follows a natural exponential family model on $\mathbb{R}^r$ with natural parameter given by $\phi_1^*(\mu_1)$*
4. *For some functions $\phi_1^* : \mathbb{R}^r \to \mathbb{R}^r$ and $H : \mathbb{R}^{k-r} \to \mathbb{R}^r$*

$$\phi_1(\mu_1, \phi_{(1)}) = \phi_1^*(\mu_1) + H(\phi_{(1)}) \tag{4}$$

5. *For some functions $k : \mathbb{R}^{k-r} \to \mathbb{R}^{k-r}$ and $h : \mathbb{R}^{k-r} \to \mathbb{R}^{(k-r)\times r}$*

$$\mu_{(1)}(\mu_1, \phi_{(1)}) = k(\phi_{(1)}) - h(\phi_{(1)})\,\mu_1 \tag{5}$$

6. *For some functions $M_1 : \mathbb{R}^r \to \mathbb{R}$ and $K : \mathbb{R}^{(k-r)} \to \mathbb{R}$*

$$M(\phi_1(\mu_1, \phi_{(1)}), \phi_{(1)}) = M_1(\phi_1^*(\mu_1)) + K(\phi_{(1)}) \tag{6}$$

An inferential cut, as defined in Definition 3, is a very useful tool allowing exact likelihood inference about a mean, say, independent of nuisance parameters. However, the existence of an exact cut is rather rare. Instead we might look to loosen its definition somewhat. Theorem 1 gives us a number of equivalent choices, any of which could be relaxed. In this paper we choose to focus on Condition (2) and define an approximate cut in the following, if rather informal, way. We also note related ideas in Christensen and Kiefer (1994, 2000).

Definition 4. *(Approximate cut) In the notation of Theorem 1, the dependence of the variance of s_1 on nuisance parameters is a measure of the sensitivity of the model for inference about μ_1. When this dependence is small we say that we have an approximate cut and the corresponding nuisance parameter is called insensitive.*

The motivation here is that we expect a small inferential change concerning μ_1 when we perturb the model in directions in which the conditions of Theorem 1 hold to a reasonable approximation. We have selected the variance characterisation to focus on due to computational advantages described below. This is clearly not the only choice from this theorem and other characterisations can be explored.

2.2 Directional approach

We wish to investigate the sensitivity of inference on a mean with respect to the specification of a working, putative model. We first take a directional approach to perturbing the working model. That is, we define perturbations of the model using directions in the model space. Since the model space is affine the direction can be considered either as a tangent vector or as a vector in the tangent space. We shall see that many perturbation directions have effectively zero effect on the model performance – which and we call the *insensitive directions* – while other can have a large effect on inference. This second class is called the *sensitive directions*.

Definition 5. *(Directional perturbation) Consider the perturbation given by Expression (2), where we treat $\delta\omega$ as the perturbation parameter. Defining the i^{rmth} component of s_2 by*

$$s_{2i}(\delta) := \log\left(1 + \frac{\pi_i^0(\delta) - \pi_i^0}{\pi_i^0}\right) = \delta\frac{w_i}{\pi_i^0} + O(\delta^2)$$

we can write

$$\pi_i^0(\delta)\exp(\phi s_{1i} - M(\phi)) = \pi_i^0\exp(\phi s_{1i} + s_{2i}(\delta) - M(\phi,\delta)) = \pi_i^0\exp(\phi s_{1i} + \delta s_{2i} - M(\phi,\delta)) + O(\delta^2)$$

where $s_{2i} := \frac{w_i}{\pi_i}$. Thus, to first order in δ, the two perturbations schemes (2) and (3) are equivalent.

By combining Definitions 4 and 5 we look for the most sensitive directional perturbation by solving an optimisation problem. We are looking for the directional perturbation which gives the largest local effect on the variance. First, we note that if, as we do, we want to preserved the meaning of the parameter of interest under such a perturbation it will be convenient, although not essential, to perturb the distribution π^0 in ways which preserved the mean, i.e. satisfying

$$\sum_{i=0}^{k} s_{1i}\pi_i^0 = \sum_{i=0}^{k} s_{1i}(\pi_i^0 + \delta\omega_i) = \mu \Rightarrow \sum_{i=0}^{k} s_{1i}\omega_i = 0.$$

These ideas give rise to the following.

Theorem 2. *Consider the following optimisation problem:*

$$\max_{\omega \in \Omega} \sum_{i=0}^{k} s_{1i}^2\omega_i,$$

where

$$\Omega := \left\{ \omega \mid \sum_{i=0}^{k}\omega_i = 0, \sum_{i=0}^{k} s_{1i}\omega_i = 0, \omega^T\Sigma_{\pi^0}\omega = \epsilon^2 \right\},$$

Σ_{π^0} is defined by the Fisher information at π^0, and ϵ is a small, user-selected tuning parameter.

This has the solution that $\Sigma^{-1}\omega$ is proportional to the Σ^{-1}-orthogonal projection of $s_2^{(2)} := (s_{2i}^2)_{i=0}^{k}$ on the space orthogonal to the space spanned by 1, and s. This problem could also be solved by standard linear programming methods.

Proof. This follows from a direct Lagrangian analysis of the problem. For specific details about the analytic calculation of this infinitesimal direction see Appendix 3.

We note here that we have set up the optimization problem to find the direction which maximise the effect on the variance. This is motivated by the definition of the approximate cut. This results in a computationally tractable problem, even in high dimensional model spaces. This is, of course, not the only possible optimization problem that could be studied. There are many other measures – information theoretic or geometric – which could be used, and this is to be studied in future work.

Example 2 (Revisited) *Returning to our running example we can solve the optimisation problem in Definition 2. Figure 3 shows the results of this based on the binomial working model. In Panel (a) we plot the statistic that we have added to the sufficient statistic, and can see that it is non-linear – and further analysis shows it is well approximated by a quadratic function. Panel (b) shows the corresponding ω-vector, while in (c) we show the effect on inference on the interest parameter μ. The solid line shows the deviance for μ from the unperturbed model, while the dash line is that from the profile likelihood for μ associated with Model in displayed equation (3). This gives us a way of making inference on μ in the presence of the new nuisance parameter ϕ. It can be shown*
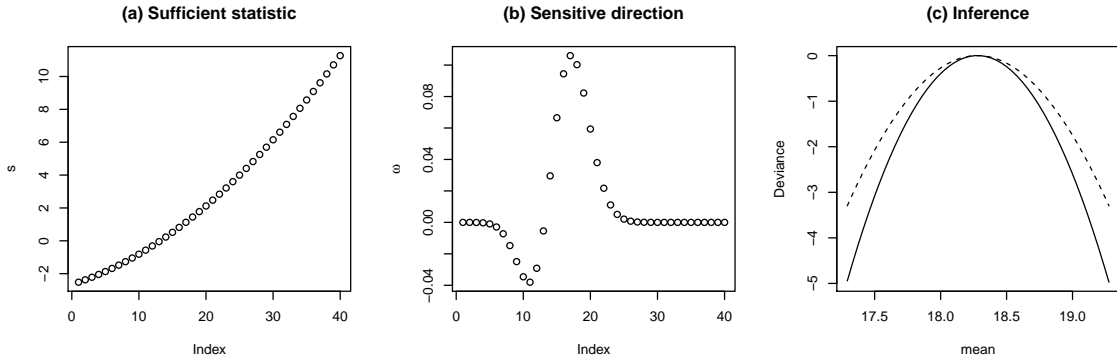
Fig. 3. Computing the direction using the Fisher matrix

that the profile likelihood is very close to the, 'model free', empirical likelihood in Fig. 2 (c). The main difference is that there is a small amount of skewness. It can be shown that adding one further element to the sufficient statistic, corresponding to a cubic – skewness – term gives almost exact agreement between the model-based and the model-free estimation.

For clarity, though, we note that we are not claiming that the model-free approach discussed above is the 'correct' inferential procedure. For instance, in Example 1, where there is a well-established theoretical model, it seems natural that the analyst use this model.

The existence of a sensitive direction means a particular modelling choice has had a substantial impact on inference although, as discussed in Definition 1, not on the value of the point estimate. For example, we might measure that impact on inference by the difference in the plotted solid and dashed lines in Fig. 3 (c). If we have good reason to believe the model (e.g. Example 1) then that is not a problem, but if we are not sure about the model (Example 2) then we might want to use proportionally more information from the data.

It is illuminating to compare the results of this infinitesimal analysis with standard ways described in the literature to generalize the Binomial to a two-dimensional exponential family of the form

Definition 6. *The following are extensions of the binomial model. They are all exponential families of the form*

$$f(x; \phi_1, \phi_2) = \binom{K}{x} \exp(\phi_1 \, x + \phi_2 \, g(x) - M(\phi_1, \phi_2)), \quad x = 0, 1, \ldots, K.$$

where g is a function which the modeller selects. Important examples include (a) letting $g(x) = I(x = 0)$ giving the zero-inflated binomial Lambert (1992), (b) letting $g(x) = x(x - K)$ giving Altham's multiplicative binomial model Altham (1978), and (c) letting $g(x) = x \log\left(\frac{x}{K}\right) + (K - x) \log\left(1 - \frac{x}{K}\right)$ giving one of Efron's double binomial model Efron (1986).

For Example 2 we are adding a quadratic term to the sufficient statistic, this gives a model which is equivalent to Model (b) in Definition 6. If the sample size is large enough for first order

asymptotic inference to be plausible then any model which is flexible enough to fit the sample mean and variance – that is the mode and the hessian of the log-likelihood – would essentially give the same inference.

The infinitesimal perturbation indicates that particular low dimensional exponential families in Δ^k might be of interest. To study these families, as is clear from the definition of an exact cut, Definition 3, the shape of the log-likelihood function in the mixed parameterisation, Def. 2, is critical in understanding the effect on μ-inference. Further, we need to understand the embedding of the families in the simplex and, in particular, the way that they meet the boundary.

Example 2 (Revisited) *Figure 4 shows the geometry of a two dimensional family inside the simplex Δ^k. Panel (a) shows the function $g(x)$ which has been added, and this corresponds precisely to the term (s_i) in Definition 5. We note that all that matters when studying the sufficient statistics of exponential families is the span of the corresponding linear space. The form of the function Figure 4 (a) only differs from Fig. 3 (a) by a linear function: they both have the same inferential effect.*
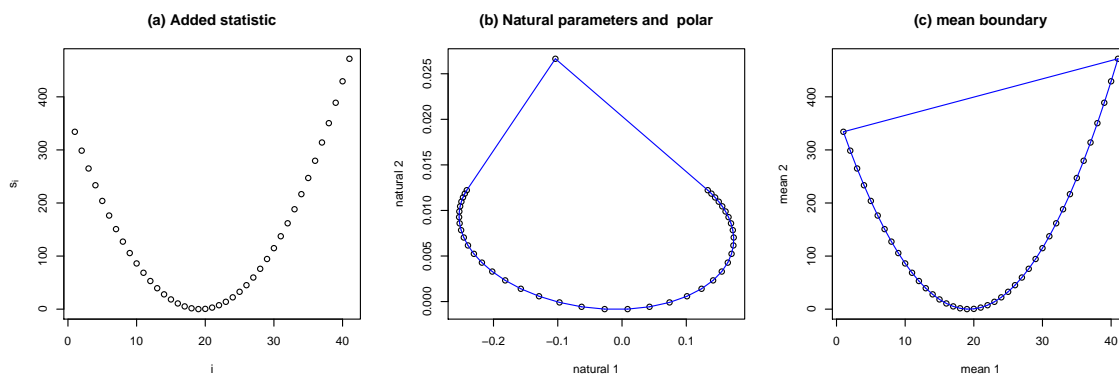

Fig. 4. The geometry of the two dimensional full exponential family

Because the two-dimensional exponential families lies in a closed simplex we need to analyse its boundary. In Panel (c) this is shown in the mean parameters, while the corresponding convex polar – which shows the directions of recession for the natural parameters, see Appendix or Critchley and Marriott (2014b); Geyer (2009); Rinaldo et al. (2009) – is shown in Panel (b). We can extend the analysis by using the theory of approximate cuts as shown in Fig. 5. The fundamental result on exact cuts, Theorem 1, says that in the mixed parameterisation the Fisher information being independent of the nuisance parameter is a sufficient condition for an exact cut.

Figure 5 (a) shows the contours of the log-likelihood in the natural parameters. We have also added the directions of recession, see Fig. 4 (b), for later analysis. We see that the log-likelihood function is close to, but is not exactly, a quadratic function in these parameters. The solid horizontal line through zero is the working model in these parameters. Panel (c) shows the same information, but now in the mean parameters. The dash line corresponds to the boundary of the exponential family shown in detail in Fig. 4 (c). The solid curve is the null model – not straight here since it is an exponential family and these are the mean parameters. Again the log-likelihood is not a quadratic
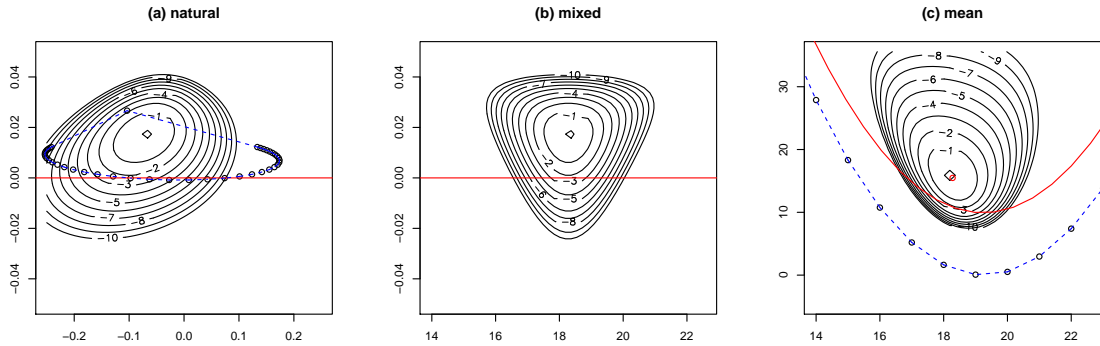
Fig. 5. Using approximate cut theory

function of these parameters. Panel (b) shows the same information in the mixed parameters, noting the vertical axes of (a) and (b) and the horizontal axes of (b) and (c) agree.

If we perturb the model in the nuisance direction – which corresponds in Panel (b) to a translation of the base model vertically – we want to see how this affects inference about the interest parameter – the horizontal axis in (b). We see from the shape the of log-likelihood in (b) that a vertical shift will change the hessian of the log-likelihood– i.e. the Fisher information – thus strongly affects inference.

In more generality we can use the shape of the log-likelihood function in the mixed parameters to assess the effect of a perturbation of the model in a given direction. Theorem 1 essentially states that if the log-likelihood was quadratic – i.e. had a fixed hessian for different horizontal slices – then we would be close to an exact cut. Hence perturbations in that direction would not have much effect on μ-inference. We illustrate this in the example below.

The infinitesimal analysis of of this section indicates that perturbations which add a quadratic, and to a small extent a cubic, sufficient statistic to the model are going to have an effect on inference, but that there will be many perturbations that have no effect. We see this below.

Example 2 (Revisited) *We choose an arbitrarily selected perturbation direction, shown in Fig. 6(a). The only constraint we put on the selection was that it only has weight in cells which have a positive observed count, see Fig. 2 (a) for the data. We will make clearer in the following section why we add this constraint, but intuititively it seems sensible to first focus on directions where the data is informative. Panels (b) and (c) show, as before, the geometry of the natural and mean parameters taking into account their boundaries. We see here that the two dimensional model meets five distinct vertices of the simplex.*

As in the previous example we show the contours of the log-likelihood in three parameterisations (a) natural, (b) mixed and (c) mean. The key plot in Fig. 7 is the middle one where the log-likelihood looks very close to being a quadratic function. This means that perturbations of the based model – illustrated with the solid horizontal line in (b) – have almost no effect on inference on μ. This direction is – as expected – very insensitive as far as μ inference is concerned. It can be shown that this is true of most 'randomly selected' directions for this example.
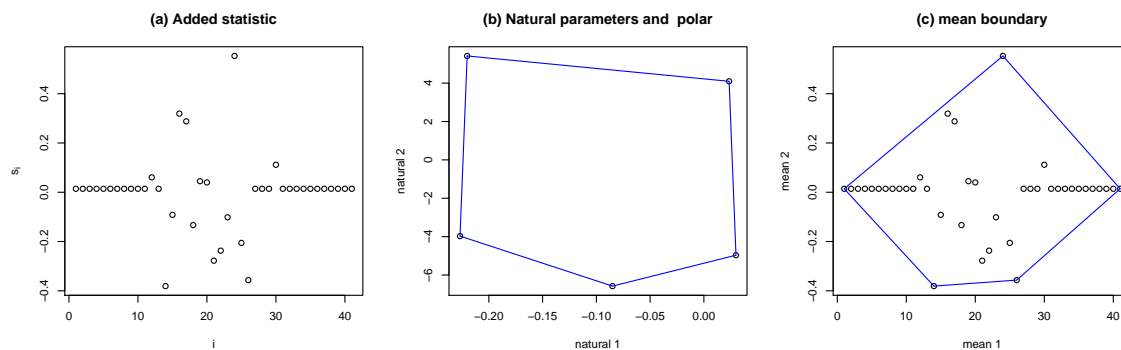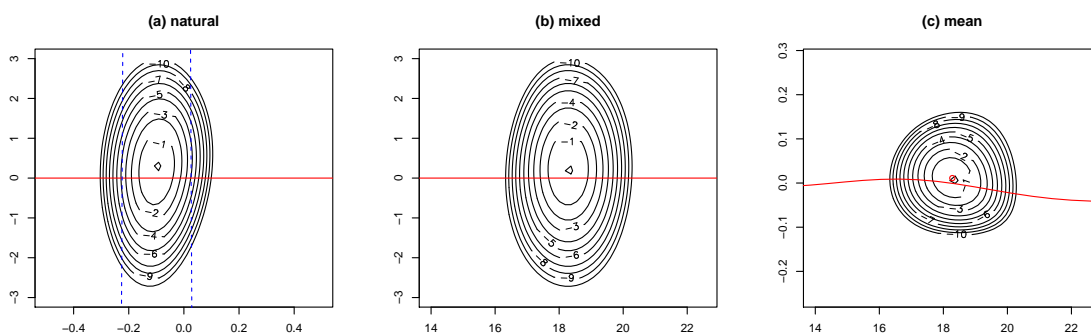
Fig. 6. Using approximate cut theory

Fig. 7. The geometry of the two dimensional full exponential family

At this point in the analysis, of our simple problem, it seems that just a couple of low order polynomials will completely determine the sensitive directions. In fact, as we now show, we can find other interesting directions by considering perturbations of the data rather than the model. This is a form of robustness analysis where some data points might be considered 'outliers' by the analyst so are not representative of the model and can be down-weighted or removed. If we look at the models described in Def. 6 we see that Model (a) allows the changing of the weight in a single cell. In that case the zero cell. Of course we can perturb the weight of other cells in particular ones which we may have identified as containing 'outliers'.

Example 2 (Revisited) *Let us start with a perturbation of a cell which is clearly not an outlier and lies right at the centre of the observed data. The geometry of the family is shown in Fig. 8 as before. Panel (a) shows the perturbation vector, which is is reweighing of cell* 18*, the sample mean of the data. The boundary and corresponding polar dual are shown in (b) and (c).*

We can see the inferential effect of this perturbation in Fig. 9. From Panel (c) we see that the boundary of the family is having an effect on the shape of the log-likelihood in the mean parameterisation, with the relevant part of the boundary being the horizontal dashed line and the working
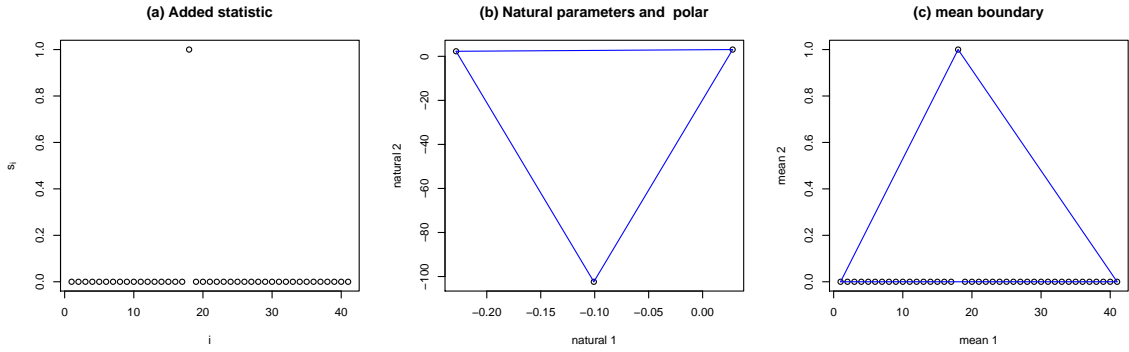
Fig. 8. The geometry of the two dimensional full exponential family

model the curved solid line. In Panel (a) we see some distortion in the corresponding direction of recession. In the mixed paremeters, Panel (b), we see that the log-likelihood is close to quadratic indicating a small effect on µ-inference.
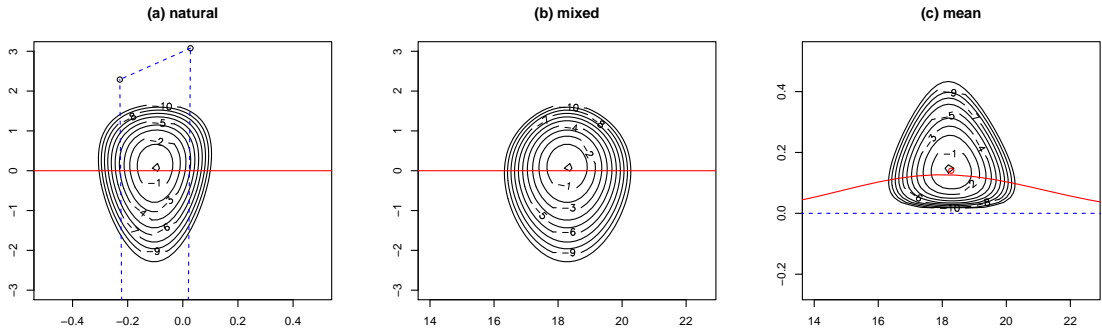

Fig. 9. The geometry of the two dimensional full exponential family

We can contrast this with the following case illustrated in Fig 10 and 11. Here the perturbation is on the cell which contains the largest observed value. This might, in our example, be a candidate for being considered an 'outlier'. The perturbation vector is shown Fig 10 (a) and the corresponding boundary geometry in Panels (b) and (c).

Figure 11 shows the effect on µ-inference. Panel (c) shows that the boundary and the model are very close – on a scale defined by the size of log-likelihood based inference – indeed the solid and dash lines are almost on top of one another in the panel. This effect is mirrored in Panel (a) where the contours of the log-likelihood function are being pulled in the direction of the corresponding direction of recession. In Panel (b) the mixed parameterisation shows the effect on µ-inference is very strong.
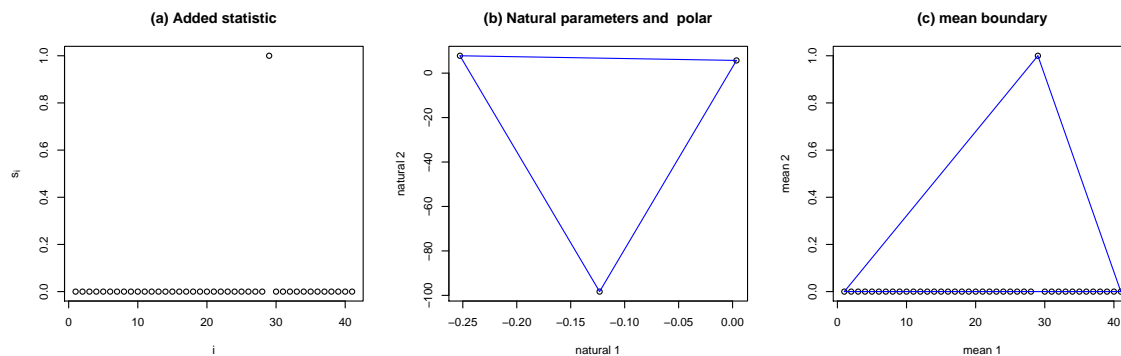
Fig. 10. The geometry of the two dimensional full exponential family

The log-likelihood is very far from being a quadratic function and so keeping or removing this single point can strongly change our inferential conclusions.
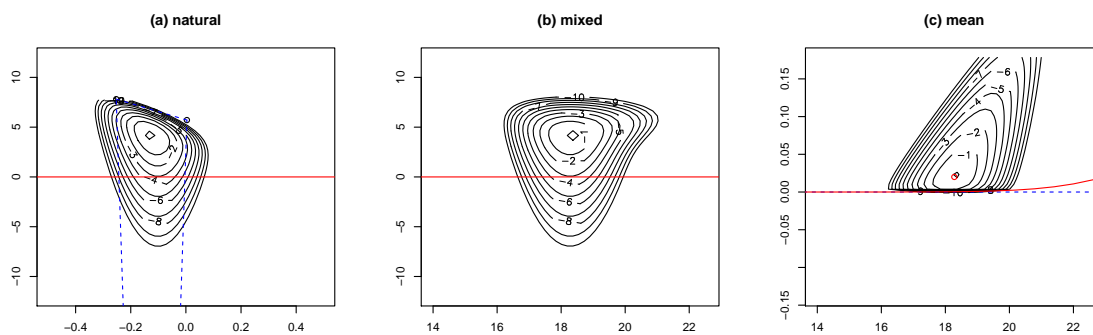

Fig. 11. The geometry of the two dimensional full exponential family

If the distance to the boundary was small we would expect that the shape of the likelihood would be distorted and so it would be unlikely that there would be even approximately a cut in that direction. We therefore, as part of our search for sensitivity directions should look for directions where the distance is small.

Definition 7. *We can define the distance – as measured by the Fisher information – between π^0 and π where π lies on the face defined by the set of indexes $\mathcal{I}$ i.e.*

$$\{\pi|\pi_i = 0 \iff i \in \mathcal{I}\}.$$

We look at the squared -1-distance, in the notation of Amari (1985), with a fixed metric at π^0 which is

$$Q(\pi) := \sum_{i=0}^{k} \frac{(\pi_i - \pi_i^0)^2}{\pi_i^0}.$$

Theorem 3. *The minimum squared distance $Q(\pi)$ is $\frac{\pi_{\mathcal{I}}^0}{1-\pi_{\mathcal{I}}^0}$. where $\pi_{\mathcal{I}}^0 := \sum_{i \in \mathcal{I}} \pi_i$. Further, the set of directions which are close to the boundary form a union of cones.*

Proof. This follows from direct calculation.

2.3 Global perturbations: region of interest

One feature of the analyses shown in Figs. 9 and 11 is that the 'distance' to the boundary seems to play an important role when looking for sensitive directions. It is for this reason that the previous analyses – which are fundamentally based on infinitesimal arguments – can be complemented with more global ones, briefly explored in this section.

One major advantage of the infinitesimal approach is that, because of Definition 5 perturbations of the form (2) and (3) are the same. The corresponding direction of perturbation is just a tangent vector and can be represented in the $+1$ or -1 form in the notation of Amari (1985). In this section we take a more global approach and focus on perturbations of the form (2) i.e. $\pi_i^0 \to \pi_i^0 + \omega_i$. Looking for interesting perturbation vectors would then involve the following optimisation problem.

$$\max_{\omega} \sum_{i=0}^{k} t_i^2 w_i \text{ such that } \sum_{i=0}^{k} \omega_i = 0, \sum_{i=0}^{k} t_i \omega_i = 0, \pi_i + \omega_i \geq 0 \tag{7}$$

Assuming that all values of t_i are distinct, all but 2 values of $\pi_i + w_i$ will be zero. Using this, or simply using linear programming to solve the problem numerically, gives the solution shown in Fig. 12. This is simply the distribution which has the empirical mean of the data and the maximum variance. Of course Fig. 12 is not going to be a plausible model for the analyst, not least because it is
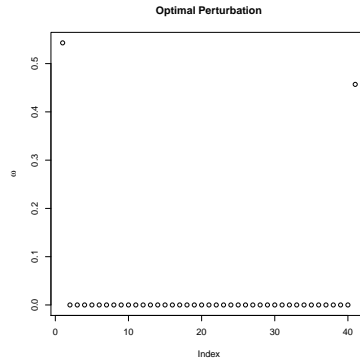

Fig. 12. The linear programming solution to the unconstrained optimisation problem

completely inconsistent with empirical distribution, aside from having the right mean. We propose that we need to restrict the search space of ω in the optimisation problem to a subset of Δ^k of distributions which are 'consistent' with the data. We call such a subset a *region of interest* in Δ^k. There are a number of ways to do this and we explore just one here.

We first note that, for computational reasons, since we are trying to solve an optimisation problem, it would be advantageous to look for convex regions of interest and the simplest choice is to have the convexity in the (-1)-affine structure of Δ^k. We can then still apply the method of linear programming which can work in very high dimensional problems.

The region of interest is designed to represent models which are consistent with the data. For any set of indices, $\mathcal{I}$, we have a corresponding count set $\sum_{i \in \mathcal{I}} n_i$ and we can use these aggregate counts to define linear inequalities on probabilities;

$$l_{\mathcal{I}} \left(\sum_{i \in \mathcal{I}} n_i \right) \leq \sum_{i \in \mathcal{I}} \pi_i \leq u_{\mathcal{I}} \left(\sum_{i \in \mathcal{I}} n_i \right), \tag{8}$$

which would show that the empircal and model based probability masses are consistent after selecting the lower bound function l and the upper bound function u. How to choose the index subsets is a matter of choice, since there are exponentially many in k to select. For simplicity here, and to explore the problem, we only look at contiguous subsets. A simple way to select upper and lower bounds is to treat each count $\sum_{i \in \mathcal{I}} n_i$ as an observation from a binomial distribution and use a corresponding binomially based confidence interval. This will result is a very conservative region of interest.

Figure 13 shows the results of the corresponding linear programming problem which results when constraint (8) is added to (7). The solutions show in Fig. 13 (b), (c) are much closer to the
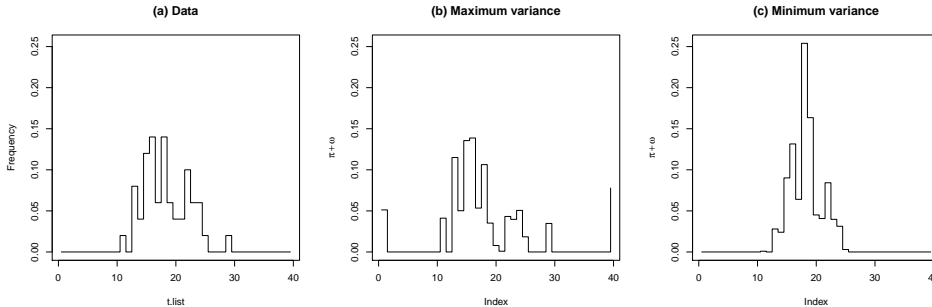

Fig. 13. The data and distribution with maximum and minimum variance in the region of interest

data than that of Fig. 12. We see that in Panel (b) the optimisation is still putting some probability in the tails of the solutions in particular in the extreme bins.

One of the advantages of using the extended exponential family Δ^k, rather than the more common multinomial, is that there is a very close relationship between the space of models and the sample space. In particular we note that in data shown in Fig. 2 for Example 2 there many bins where there are zero counts. We can define the two sets

$$\mathcal{P} = \{i | n_i > 0\}, \ \mathcal{Z} = \{i | n_i = 0\},$$

and we call the subsimplex of Δ^k indexed by $\mathcal{P}$ the observed face. The way that this decomposition effects the shape of the likelihood function across Δ^k is discussed in Critchley and Marriott (2014a), which points out that there are many directions in which the log-likelihood is flat – i.e. we can learn nothing from this particular set of data. Using this decomposition gives a different way to construct a region of interest of Δ^k which looks at points with high likelihood values. That is we add the following constraint

$$\sum_{i \in \mathcal{P}} n_i \log \pi_i \geq C_1 \tag{9}$$

$$\sum_{i \in \mathcal{Z}} \pi_i \leq C_2 \tag{10}$$

for suitably chosen values of C_1, C_2. We do not have space here to describe the solutions except to note that they appear to be computational tractable and can give attractive solutions to simple problems such as Example 2.

3 Discussion

This paper is an example of what we call computational information geometry and gives an illustration of how it can be used in the foundational problem of understanding the way that selecting a statistical model affects a given inference problem. We contrasted Example 1, where the model has been selected by theoretical considerations, with Example 2, where a more empirical approach has been taken. We believe that ideas of this paper could be applied to both examples, but have particular importance in the second.

We note that using the methods of CIG we can find sensitive perturbations directions which generate a range of inferences about μ. These include ones where the model is treated as completely correct to ones where the model has been extended so that the resultant inferences agree with 'model free' inferences. What was perhaps surprising was that the space of sensitive perturbations was very small.

We also showed that there are different types of perturbations – which are based on more global considerations – which explore the robustness aspects of the inference problem. While the 'model free' inference might seem to have less assumptions they do put much more weight on the observed data being exactly as expected, without 'outliers'. The sensitive directions discovered allows the analyst to understand the different choices available to them, balancing belief between the model and the data.

There are a number of computational issues which have naturally arisen in our analysis. These include the potential high dimensionality of Δ^k, so that optimisations based on methods which work in high dimensions have been focused on, such as convex and linear optimisation. Further, the role of boundaries in the mean parameters and the polar dual of such boundaries, turned out to be critical in the analysis. In the examples shown in this paper 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 $\mathbb{R}^k$, can be done with standard software. In general, however, as the number of parameters and the sample size grows, complete enumeration of the boundary becomes computationally infeasible, see Fukuda (2004) and the corresponding computational issues will be the subject of further work. Finally the role of a mixed parameterisation, which has aspects of both the ± 1 geometries of exponential families was highlighted. In general these can only be computed

numerically and further research will be done on efficient ways to do this, partiucally in the case of non-trivial boundaries

Appendix 1: The model space, cuts and closures

Model space

A key concept in building the perturbation space is to first represent statistical models – sample spaces, together with probability distributions on them – and associated inference problems, inside adequately large but finite dimensional spaces, see Critchley and Marriott (2014a) for details. Consider the general k–dimensional extended multinomial model

$$\Delta^k := \left\{ \pi = (\pi_0, \ldots, \pi_k)^T, \pi_i \geq 0, \sum_{i=0}^{k} \pi_i = 1 \right\}. \tag{11}$$

The multinomial family on $k + 1$ categories can be identified with the (relative) interior of this space, $int(\Delta^k)$, while the extended family, (11), allows the possibility of distributions with different support sets. This paper looks at (extended) exponential families embedded in Δ^k and uses the following notation.

Definition 8. *Let $\pi^0 = (\pi_i^0) \in int(\Delta^k)$, and V be a $(k+1) \times p$ matrix of the form $(v^{(1)} | \ldots | v^{(p)}) = (v_0 | \ldots | v_k)^T$ with linearly independent columns and chosen such that $1_{k+1} := (1, \ldots, 1)^T \notin \mathrm{Range}(V)$. With these definitions there exists a p-dimensional full exponential family in Δ^k, denoted by $\pi(\phi) = \pi_{(\pi^0, V)}(\phi)$ with general element:*

$$\pi_i(\phi) = \pi_i^0 \exp\{v_i^T \phi - M(\phi)\}, \tag{12}$$

$i = 0, \ldots, k$ with normalising constant

$$\exp\{M(\phi)\} := \sum_{i=0}^{k} \pi_i^0 \exp\{(V\phi)_i\} = \sum_{i=0}^{k} \pi_i^0 \exp\{v_i^T \phi\},$$

for all $\phi \in \mathbb{R}^p$.

Using this formalism selecting a one dimensional model to undertake inference about $\mu = E(V)$, as in Examples (1) and (2), requires selecting a sufficient statistic V and a basepoint π^0. Initially we concentrate on the case where the choice of model contributes the minimal amount of information to the inference problem. We call these least informative models.

Definition 9. *(Least informative model) Let X be the random variable over the $k + 1$ categories of Δ^k which takes values x_i in category i. The model $\pi(\phi) = \pi_{(\pi^0, V)}(\phi)$ is a one dimensional least informative model for the estimation of $E(X)$ when V is $(k + 1) \times 1$ and $v^{(1)} \propto (x_i)$.*

Both the models considered in Examples (1) and (2) are least informative for the parameter of interest. Choices between different least informative models then correspond to selecting different base measures $\pi^0 \in \Delta^k$. We can think of these geometrically as translations of exponential families in the affine geometry defined by the natural parameters.

Closures of exponential families

In this section we consider the closure of discrete p-dimensional exponential families which are subsets of Δ^k For more general results on closures of exponential families see Barndorff-Nielsen (1978), Brown (1986), Lauritzen (1996) and Csiszar and Matus (2005). In the discrete case considered here, 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) and described in detail in Anaya-Izquierdo et al. (2014).

We want to consider the limit points of the p-dimensional exponential family, so we consider the limiting behaviour of the path $\phi(\lambda) := \lambda q$ 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 $s_i := (S_0(i), \ldots, S_p(i))^T$. There exist 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 polar dual, Tuy (1998).

Appendix 2: Empirical likelihood for the mean parameter in a multinomial setting

Let T be a discrete random variable with $k + 1$ values $\{t_0, \ldots, t_k\}$ so that the probability mass function is $P[T = t_i] = \pi_i$ for $i = 0, 1, \ldots, k$, where $\sum_{i=0}^k \pi_i = 1$ and $\pi_i \geq 0$. The distribution of T depends on k free parameters and we are interested in making inferences about the expectation parameter

$$\phi = \sum_{i=0}^k t_i \pi_i = t_0 + \sum_{i=1}^k \pi_i (t_i - t_0)$$

in the presence of the other $k - 1$ nuisance parameters.

Theorem 4. *For a given random sample of size N from T, let t_- be the minimum observed value of T and t_+ be the maximum observed value of T, and we work in the generic case where all t_i's are distinct. Then for any $\phi \in (t_-, t_+)$ the profile likelihood for the mean parameter ϕ is given by*

$$\hat{\pi}_i(\phi) = \frac{n_i}{N + \hat{\delta}_\phi (\phi - t_i)}, \qquad i \in \mathcal{P}$$

Here, n_j is the number of times that t_j appears in the sample so that $N = \sum_{i=0}^k n_i$ and $\hat{\delta}_\phi$ is the unique solution to the equation

$$\sum_{i \in \mathcal{P}} \frac{n_i(t_i - \phi)}{N + \hat{\delta}_\phi (\phi - t_i)} = 0$$

in the interval $\left(\frac{N}{t_- - \phi}, \frac{N}{t_+ - \phi} \right)$.

Proof. The empirical (profile) likelihood for ϕ can be found by solving the following optimization problem

$$\max_{\pi} \sum_{i \in \mathcal{P}} n_i \log \pi_i \text{ s.t.} \sum_{i \in \mathcal{P} \cup \mathcal{Z}} \pi_i = 1, \sum_{i \in \mathcal{P} \cup \mathcal{Z}} t_i \pi_i = \phi$$

where, we recall, $\mathcal{P} = \{i : n_i > 0\}$ and $\mathcal{Z} = \{i : n_i = 0\}$. Since the t_i's are distinct and we can also assume without loss that $\pi_i > 0$ for $i \in \mathcal{P}$ because otherwise $\ell = -\infty$. The Lagrangian is given by

$$\mathcal{L} = \sum_{i \in \mathcal{P}} n_i \log \pi_i + \lambda \left(\sum_{\in \mathcal{P} \cup \mathcal{Z}} \pi_i - 1 \right) + \delta \left(\sum_{\in \mathcal{P} \cup \mathcal{Z}} \pi_i t_i - \phi \right)$$

and the key turning point equations are given by

$$i \in \mathcal{P}, \frac{\partial}{\partial \pi_i} = 0 \Rightarrow n_i + \hat{\lambda}\hat{\pi}_i + \hat{\delta}t_i\hat{\pi}_i = 0$$

$$i \in \mathcal{Z}, \frac{\partial}{\partial \pi_i} = 0 \Rightarrow \hat{\lambda} + \hat{\delta}t_i = 0$$

which give the solutions

$$\hat{\pi}_i = \frac{n_i}{N + \hat{\delta}(\phi - t_i)}, N + \hat{\delta}(\phi - t_i) > 0$$

with $\hat{\delta}_\phi$ defined as the solution $H_\phi(\delta) = 0$ where

$$H_\phi(\delta) := \sum_{i \in \mathcal{P}} \frac{n_i(t_i - \phi)}{N + \delta(\phi - t_i)}.$$

Calculations show that

$$\delta_{min} = \frac{N}{t_- - \phi} < \hat{\delta} < \frac{N}{t_+ - \phi} = \delta_{max}.$$

giving

$$H_\phi'(\delta) = \sum_{i \in \mathcal{P}} \frac{n_i(t_i - \phi)^2}{(N + \delta(\phi - t_i))^2} > 0$$

so that $H_\phi(\delta)$ is a strictly increasing function. Also

$$\lim_{\delta \to \delta_{\min}} H_\phi(\delta) = -\infty, \qquad \lim_{\delta \to \delta_{\max}} H_\phi(\delta) = \infty$$

so that $H_\phi(\delta) = 0$ has a unique solution in the interval $(\delta_{min}, \delta_{max})$

Appendix 3: Sensitive infinitesimal perturbations

We proceed from the minimal exponential family representation of the multinomial for the observed counts $n = (n_1, \dots, n_k)^T$

$$f_n(n; \eta) = \exp\left(n^T\eta - \varphi(\eta)\right) h(n)$$

where the relation with the probability parameter π is given by $\eta_i(\pi) = \log\left(\frac{\pi_i}{1 - \sum_{r=1}^k \pi_r}\right)$, $\pi_i(\eta) = \frac{e^{\eta_i}}{1 + \sum_{i=1}^k e^{\eta_i}}$ for $i = 1, \dots, k$, $\varphi(\eta) = N \log(1 + \sum_{i=1}^k e^{\eta_i})$, and $h(n)$ is the multinomial coefficient.

We define the following coordinate system in $\mathcal{N}$, the natural parameter space. Consider a fixed point $\eta_0 \in \mathbb{R}^k$ and $d^T := (t_1 - t_0, \dots, t_k - t_0)/N$. Let $\{v_1, \dots, v_{k-1}\}$ be an orthogonal basis for the orthogonal complement of d. If we take $A = (d, v_1 \dots, v_{k-1})$, then for any $\eta \in \mathbb{R}^k$ we can write

$\eta = \eta_0 + A\phi$ for some $\phi \in \mathbb{R}^k$. So ϕ defines a new parameterisation for the multinomial. By defining $s := A^T n + c$ with $c^T = (t_0, 0, \ldots, 0)$ we have

$$f_s(s; \phi) = \exp\left(s^T \phi - M(\phi)\right) f_n((A^T)^{-1}(s - c); \eta_0)$$
$$= \exp\left(s^T \phi - M(\phi)\right) f_s(s; 0)$$

where

$$M(\phi) = \varphi(\eta_0 + A\phi) - \varphi(\eta_0) - c^T \phi.$$

This is of course, the same regular natural exponential family but now with natural parameter ϕ and expectation parameter

$$\mu(\phi) = D_\phi M(\phi) = E[s; \phi] = A^T E[n] + c.$$

We are interested in making inferences about $\mu_1 = E[s_1] = \sum_{i=0}^{k} t_i \pi_i = \phi$.

According to the variance Condition 2 in Theorem 1: $s_1 = n^T d + t_0$ is an exact cut for the regular exponential family

$$\mathcal{F} = \left\{ f_s(s; \phi) = \exp\left(s^T \phi - M(\phi)\right) f_s(s; 0) \; : \; \phi \in \mathcal{P} \right\}$$

if and only if its variance depends only on μ_1. If such exact cut exists, we can then make exact marginal inferences for μ_1 using the marginal distribution of s_1 given by

$$f_{s_1}(s_1; \mu_1) = \exp(s_1 \phi^*(\mu_1) - \psi(\phi_1^*(\mu_1))) \, h^*(s_1)$$

for some real valued functions h^* and ψ. We define

$$\pi(\mu) = N^{-1}(A^T)^{-1}(\mu - c)$$

and then we have

$$Var(s_1; \mu) = d^T Var(n; \pi(\mu)) d$$
$$= N \, d^T [diag(\pi(\mu)) - \pi(\mu)\pi(\mu)^T] d$$

so we can check how much this vary as a function of $\mu_{(1)}$. For any fixed μ_1^0 we would like to explore the variation of $V(s_1; \mu)$ in the subspace of densities given by $\mu_1 = \mu_1^0$. We would like to find a direction in such space such that $Var(s_1; \mu)$ changes the most.

We define the following inner products for $u, v \in \mathbb{R}^k$

$$\langle u, v \rangle_\mu := u^T I(\mu) v, \langle u, v \rangle_\phi := u^T I(\phi) v, \langle u, v \rangle_\pi := u^T I(\pi) v, \langle u, v \rangle_\eta := u^T I(\eta) v$$

and orthogonal projections matrices are

$$P_\eta^\perp(v; u) := v - \left[\frac{\langle u, v \rangle_{\eta_0}}{\langle u, u \rangle_{\eta_0}} \right] u$$

If ω is such that $\omega_1 = 0$ and $\mu_0 = \mu(\phi)$ with $\phi = 0$ then

$$Var(s_1; \mu_0 + \lambda\omega) = Var(s_1; \mu) + \lambda \langle \omega, I(\phi_0) A^{-1} d^{(2)} \rangle_{\mu_0}$$

so the directional derivative at μ_0 along the vector ω is given by $\langle \omega, I(\phi_0)A^{-1}d^{(2)}\rangle_{\mu_0}$. To explore the variation of $Var(s_1; \mu_0 + \lambda\omega)$ we define the following optimisation problem

$$\max_w \langle \omega, I(\phi_0)A^{-1}d^{(2)}\rangle_{\mu_0} \text{ s.t. } \langle w, w\rangle_{\mu_0} = 1, \langle w, I(\phi_0)e_1\rangle_{\mu_0} = 0$$

where $e_1^T = (1, 0, \ldots, 0)$.

The solution is given by $\hat{\omega} = \hat{u}/\|\hat{u}\|_{\mu_0}$ where

$$\hat{u} = P_{\mu_0}^{\perp}(I(\phi_0)A^{-1}d^{(2)}; I(\phi_0)e_1)$$

that is, the normalised projection of $I(\phi_0)A^{-1}d^{(2)}$ orthogonal to $I(\phi_0)A^{-1}d$ in the metric $I(\mu_0)$. Note that $A^{-1}d = e_1$ and also $\|\hat{u}\|_{\mu_0} = \|P_{\eta_0}^{\perp}(d^{(2)}; d)\|_{\eta_0}$. We can write $\hat{\omega}$ as

$$\hat{\omega} = I(\phi_0)A^{-1} \frac{P_{\eta_0}^{\perp}(d^{(2)}; d)}{\|P_{\eta_0}^{\perp}(d^{(2)}; d)\|_{\eta_0}}$$

The objective function evaluated at the maximum is

$$\langle \hat{\omega}, I(\phi_0)A^{-1}d^{(2)}\rangle_{\mu_0} = \left\| P_{\eta_0}^{\perp}(d^{(2)}; d) \right\|_{\eta_0}$$

This has a nice interpretation. If we take $\eta_0 = \eta(\hat{\pi}_{Global}) = \eta(n/N)$ we have

$$\langle \hat{\omega}, I(\phi_0)A^{-1}d^{(2)}\rangle_{\mu_0} = \|d\|_{\eta_0}^2 \, C_{+1}(\hat{\phi}_{Global})$$

then it can be interpreted as $\|d\|_{\eta_0}^2$ times the $+1$ curvature of the profile likelihood curve for ϕ at $\phi = \hat{\phi}_{Global}$. The profile likelihood curve defines a curved exponential family embedded in the multinomial. We have

$$N \frac{\partial \eta}{\partial \phi}(\hat{\phi}) = \frac{1}{\|d\|_{\eta_0}^2} d$$

$$N^2 \frac{\partial^2 \eta}{\partial \phi^2}(\hat{\phi}) = \frac{1}{\|d\|_{\eta_0}^4} \left[P_{\eta_0}^{\perp}(d^{(2)}; d) \right] - \frac{S}{\|d\|_{\eta_0}^6} d$$

so the $+1$ embedding curvature of the profile likelihood curve at $\phi = \hat{\phi}$ is given by

$$C_{+1}(\hat{\phi}_{Global}) = \frac{\left\| P_{\eta_0}^{\perp}(d^{(2)}; d) \right\|_{\eta_0}}{\|d\|_{\eta_0}^2}$$

The solution $\hat{\omega}$ determines a direction in the -1 space of the exponential family $\mathcal{F}$. If variation in this direction is small we can consider s_1 as an approximate cut.

Bibliography

Altham, P. M. (1978). Two generalizations of the binomial distribution. *Applied Statistics*, 162–167.

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

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 (2013). Computational information geometry in statistics: Foundations. In *Geometric Science of Information*, pp. 311–318. Springer.

Barndorff-Nielsen, O. (1976). Factorization of likelihood functions for full exponential families. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37–44.

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. *Annals of Statist. 11*(3), 753–769.

Barndorff-Nielsen, O. and A. Koudou (1995). Cuts in natural exponential families. *Theory Prob. Its Appl. 40*, 220–229.

Box, G. (1976). Science and statistics. *JASA 71*, 791–799.

Box, G. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *JRSS, B 143*, 383–430.

Brown, L. (1986). *Fundamentals of statistical exponential families: with applications in statistical decision theory*. Institute of Mathematical Statistics.

Christensen, B. J. and N. M. Kiefer (1994). Local cuts and separate inference. *Scandinavian Journal of Statistics*, 389–401.

Christensen, B. J. and N. M. Kiefer (2000). Panel data, local cuts and orthogeodesic models. *Bernoulli*, 667–678.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B, Methodological 48*, 133–155.

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. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B: Methodological 49*, 1–18.

Critchley, F. and P. Marriott (2004). Data-informed influence analysis. *Biometrika 91*, 125–140.

Critchley, F. and P. Marriott (2014a). Computational information geometry in statistics: theory and practice. *Entropy 16*(5), 2454– 2471.

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

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

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association 81*(395), 709–721.

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

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

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics 34*(1), 1–14.

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

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika 75*(2), 237–249.

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

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