Binary information geometry: some theory and applications

Radka Sabolová[1], P. Marriott[2], G. Van Bever[1] & F. Critchley[1]

[1] The Open University, United Kingdom and

[2] University of Waterloo, Canada

Contact Information:
Department of Mathematics and Statistics
The Open University,
Walton Hall, Milton Keynes,
Buckinghamshire, MK7 6AA.
UK. Email: radka.sabolova@open.ac.uk

Abstract

Binary information geometry (BIG) is a particular case of computational information geometry (CIG) which (i) provides and exploits a universal space of models for binary random vectors, this being an exponentially high-dimensional *extended* multinomial family, and (ii) finds natural, fruitful application in a range of important areas, including: (A) binary graphical models, notably Boltzmann machines and, (B) logistic regression.

Key features of CIG here

- Want to exploit the key duality of classical IG between sample and model spaces as cleanly as possible. In classical IG counts can be zero but probabilities can not. In CIG we allow both to be zero.
- Thus, the underlying dimensions are not kept fixed as we are working on closed simplexes rather than open subsets of manifolds.
- CIG has a distinctive geometric approach to the closure of exponential families, exploiting the polar dual of the boundary.
- Boundaries play a key role in CIG – *the geometry of the boundary dominates the global IG in the relative interior.*
- Geometric features include: dually ruled spaces, the polar dual, the structure of convex hulls and especially their boundaries.
- Computational issues include: exploiting the ruling, computing the polar dual in high dimensional spaces, using the boundary structure for model selection, and calculating the effect of the polar dual on inference.

Example 1: Binary graphical models

We look at models of joint distributions of binary vectors where the dependence is determined by the values of hidden binary nodes. Conditionally these models are independent, full exponential families, so that unconditionally they are mixture models of a fixed order. Examples include Boltzmann machines where we have N_I input, N_H hidden and N_O output nodes.

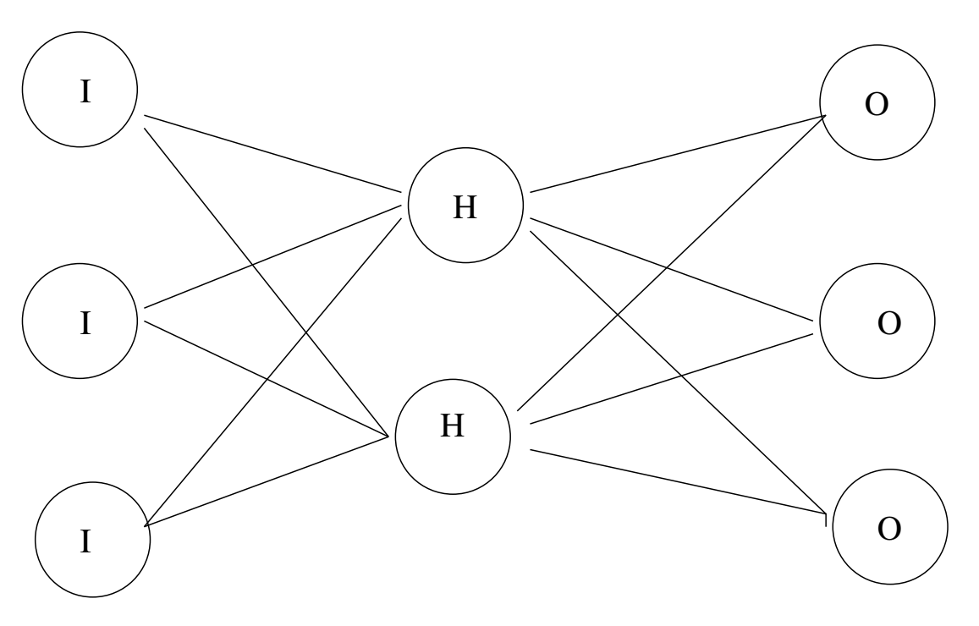

Figure 1: Structure of models

The universal space of joint models for binary vectors is an extended multinomial model whose boundary dominates the global behaviour. Figure 2 corresponds to the case with $N_I = N_H = N_O = 1$. The set of independent models, shown in Panel (a), is a full exponential family defined by unions of (-1)-affine subsets, i.e. is a ruled space.

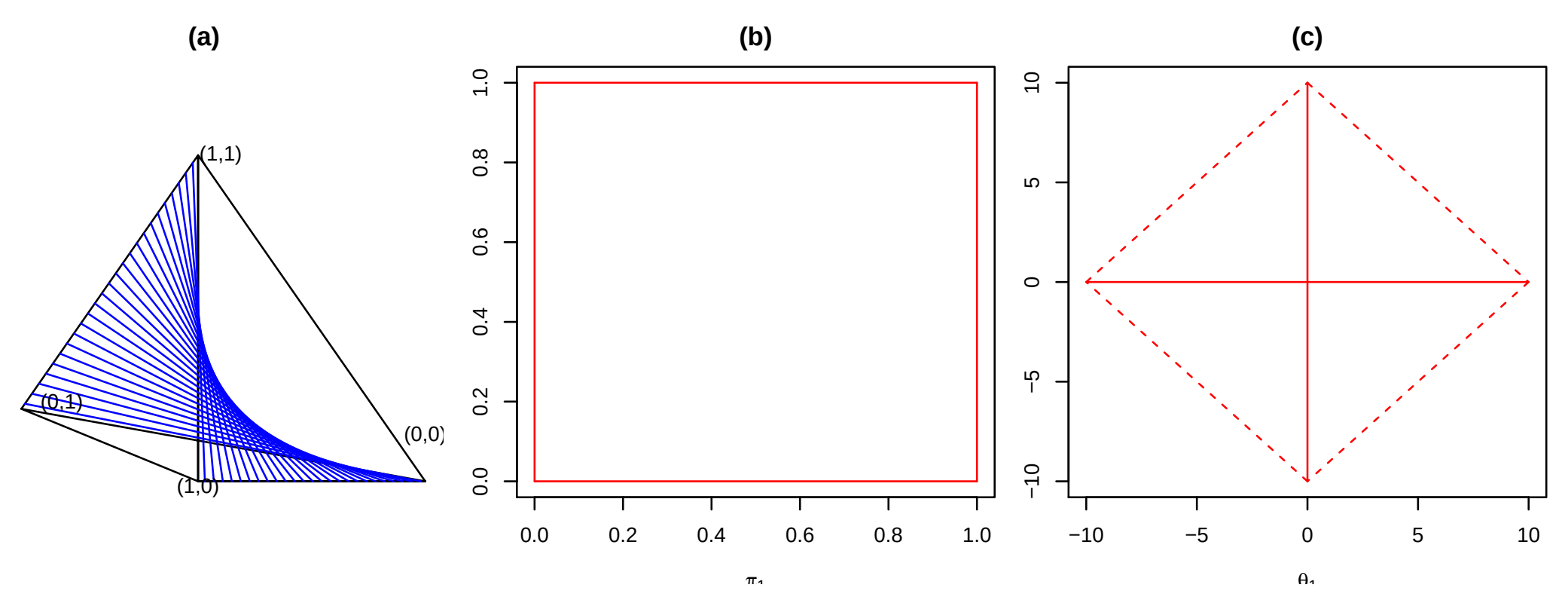

Figure 2: (a) independence space (blue) in universal space, (b) mean parameter space and boundary, (c) polar dual which defines directions of recession

Figure Fig. 2(b) shows its boundary in the (-1)-affine parameters. The limit points in the $(+1)-$affine structure of the relative interior are given by the polar dual to the boundary in (b) and determine the directions of recession.

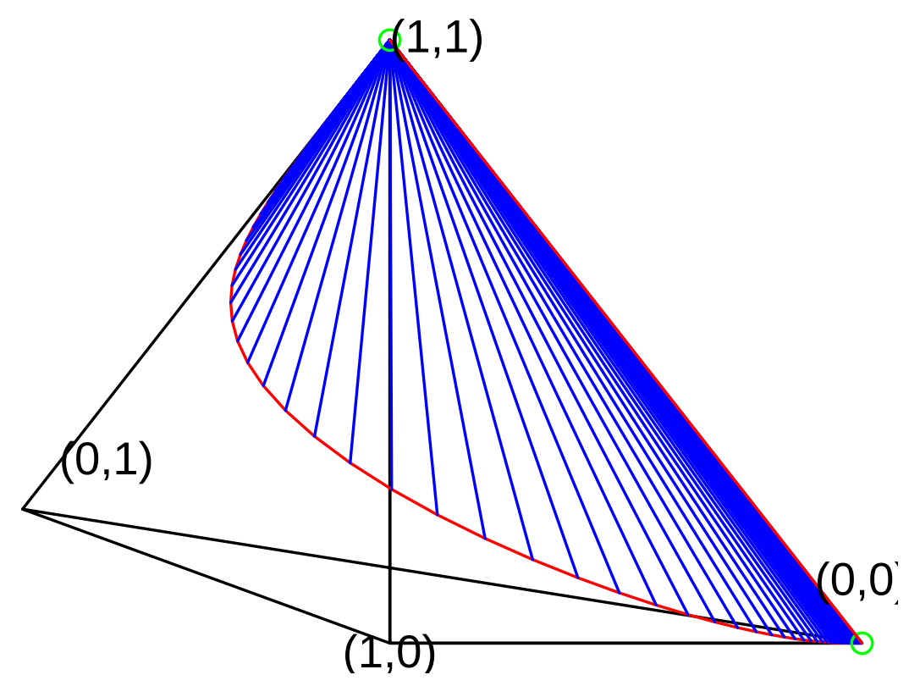

Figure 3: Convex hull and boundary of one-dimensional exponential family (red). 0.5-hull is green, 1-hull is red, 1.5-hull is blue.

The convex hull of a one dimensional exponential family is a union of smooth manifolds of differing dimensions, called N-hulls, where N is the number of components. Note when a vertex is a component it is counted with a value of $\frac{1}{2}$.

In Fig. 3, the $\frac{1}{2}$-hull (green) is the union of two vertices, the 1-hull (red) is the union of the r.i. of the exponential family and the (-1)-geodesic joining the vertices, the $1\frac{1}{2}$-hull (blue) is the union of two ruled surfaces (one for each vertex), and the 2-hull is the r.i. of the full convex hull. In general, unless of full rank, these components are not convex (or even connected), and thus *the likelihood can have multiple modes*.

Take home message: The index-based decomposition of the boundary of a closed convex hull throws light upon the structure of the binary graphical model, [2], distinguishing between multi-modal, unimodal and over-parameterised likelihoods.

Example 2: Logistic regression

The logistic regression model is widely used for a binary response given a set of binary covariates.

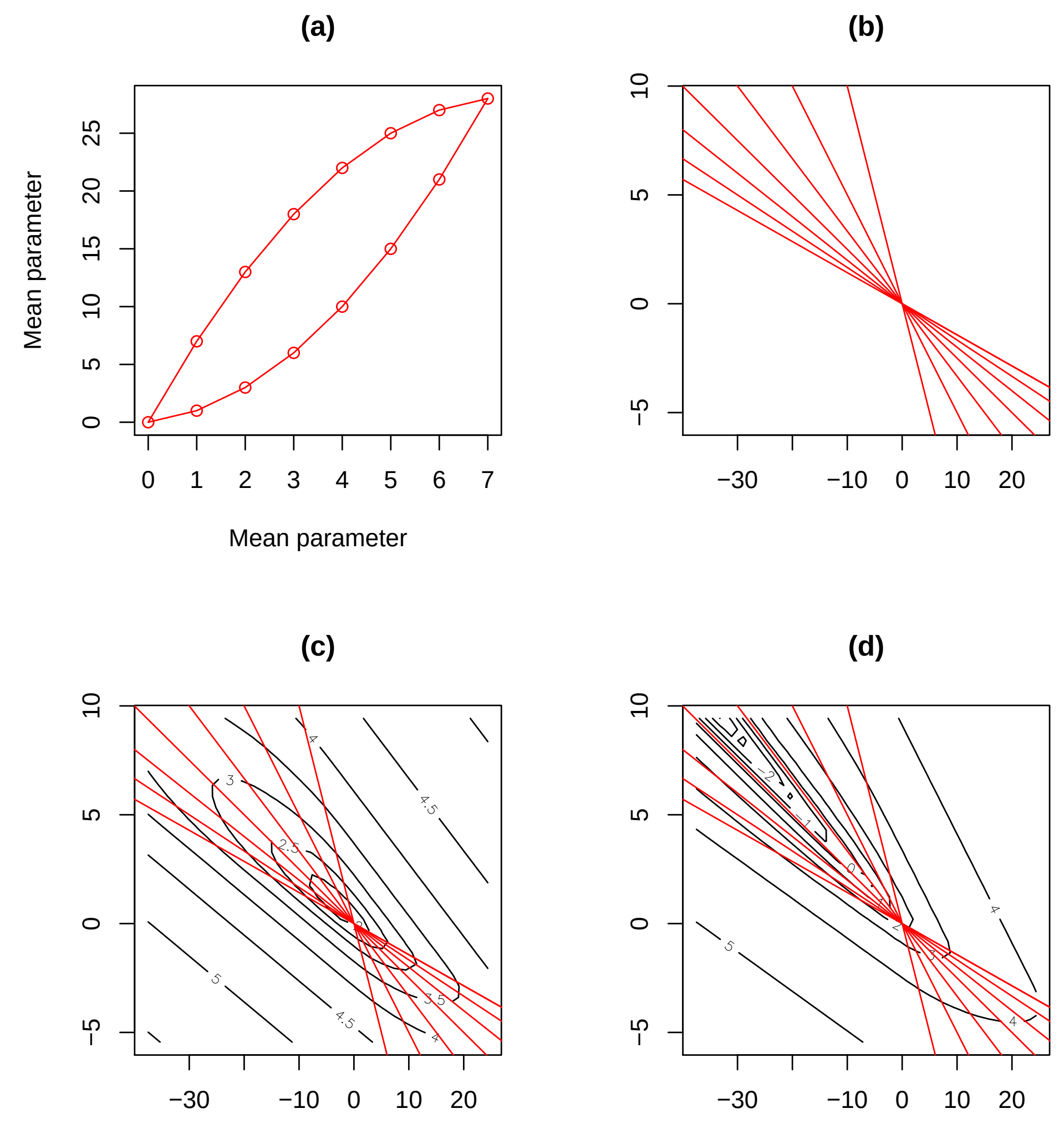

Figure 4: Logistic regression example: (a) mean parameters and boundary, (b) directions of recession in $(+1)$-parameters, (c) polytopal shape of likelihood near boundary, (d) the directions of recession determine the extremes of the likelihood.

Logistic regression has a full exponential structure. The panels (a) and (b) of Fig. 4 show the boundary, and the corresponding directions of recession in the (±1)-affine parameters respectively, for a simple logistic example. For different data sets, the panels (c) and (d) show how the global shape of the likelihood is predominantly determined by the boundary.

Take home message: Logistic regression models are low-dimensional within an exponentially high-dimensional universal space. This high-dimensionality is reflected in the complicated mean parameter polytope. This boundary polytope determines key statistical properties of the model:

- the likelihood has polytope behaviour at infinity, Fig 4 (c).
- The extremal behaviour of the likelihood is determined by the directions of recession, Fig 4 (d).

Figure 5 shows the effect of the boundary on sampling distributions for a more complex model, described in [1], for a subset of Fisher's iris data. In Panel (a), we see the sampling distribution of the MLE, whose level of discreteness is determined by the number of directions of recession. The higher order asymptotic theory, reflected in Edgeworth expansion, is shown in blue, showing appreciable non-normality. In (b) the effect of the directions of recession are more obvious. The non-normality is very clear in the marginal structure shown in (c).

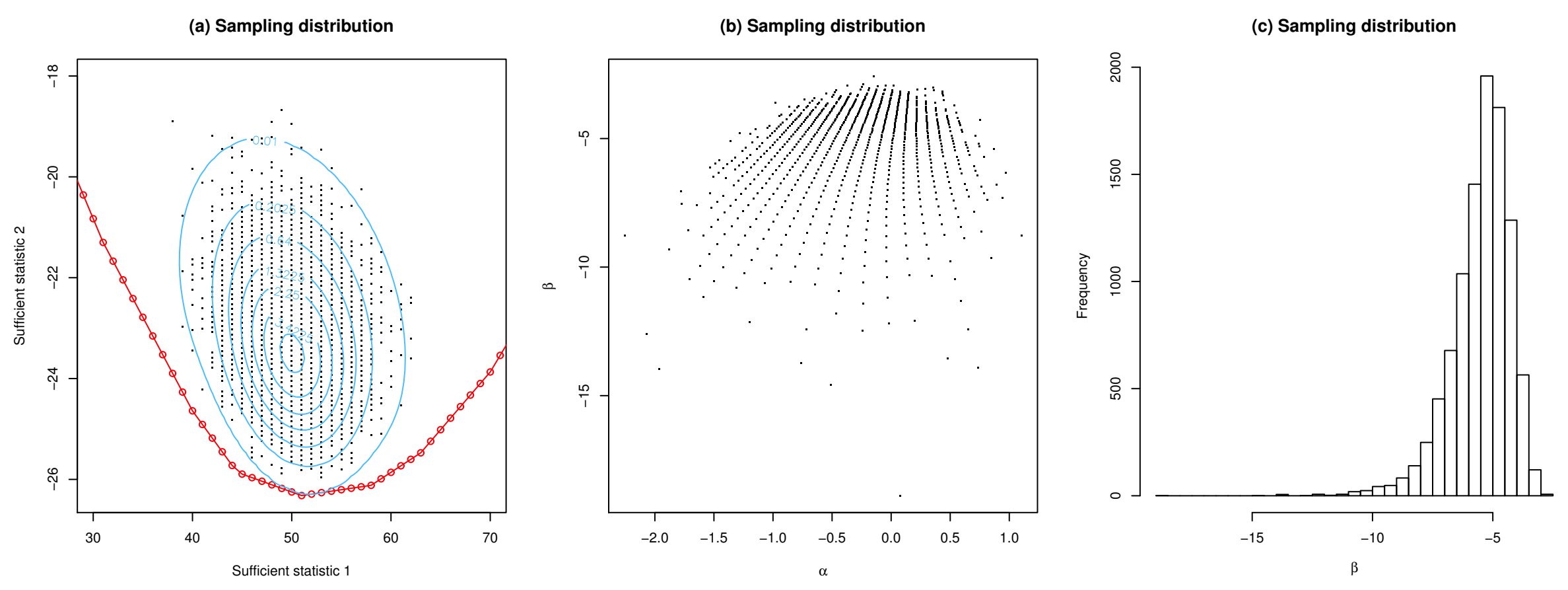

Figure 5: (a) mean parameter space, complex boundary (red), sample from distribution of MLE (black), Edgeworth approximation (blue). (b) sampling distribution of MLE in $(+1)$-parameters, (c) marginal distribution of one component of MLE

Take home message: Both likelihood (Fig. 4) and sampling distributions (Fig. 5) have, simultaneously, *continuous* and *discrete* features:

- Classical IG method accommodate the continuous features well.
- Novel CIG methods accommodate the discrete features, which dominate in the boundary limits.

References

[1] Karim Anaya-Izquierdo, Frank Critchley, and Paul Marriott. When are first–order asymptotics adequate? a diagnostic. *Stat*, 3(1):17–22, 2014.

[2] Samuel Karlin and William J Studden. *Tchebycheff systems: With applications in analysis and statistics*, volume 376. Interscience Publishers New York, 1966.

Acknowledgements

We gratefully acknowledge the support of the UK EPSRC grant EP/L010429/1
Website: **http://users.mct.open.ac.uk/rs23854/EGSS**