Computational Information Geometry:

Theory and Practice

Paul Marriott

Department of Statistics and Actuarial Science, University of Waterloo

December 7, 2010

Overview

- Introduction to Computational Information Geometry
- Encompasses and extends both Amari's information geometry and Lindsay's mixture geometry
- Aims to unlock the power of information geometry to mainstream users by being computational
- Illustrate talk through examples
- Joint work with Karim Anaya-Izquierdo, Frank Critchley and Paul Vos
- Thanks to EPSRC Grant Number EP/E017878/1

Big Picture

- The way that parametric statistical models lie in a ‘space of all models’ is important
- We will use high-dimensional (extended) multinomial space as a proxy for the ‘space of all models’
- Show that this computational approach encompasses both Amari’s information geometry and Lindsay’s mixture geometry
- The geometry of the (extended) multinomial space is highly tractable and mostly explict so very good for building a computational theory
- Long term aim is to build software which releases to power of these geometric theories to mainstream

Information Geometry

- Developed by Efron [10], Amari [4], Barndorff-Nielsen [6], [7] and others, see the book by Kass and Vos [13]
- Used in understanding asymptotic analysis, information loss, the properties of estimators . . .
- How to connect two density functions $f(x)$ and $g(x)$ in the space of all models?

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

  +1: $\frac{f(x)^{\rho} g(x)^{1-\rho}}{C(\rho)}$
- These define two different affine geometries.
- Duality: non-linear relationship between them given by Fisher information.

Example: censored exponential family

- Censored exponential example, [13, 17], with observed R.V. $y = \min(z, t)$ and x the censoring indicator has model $p\left(y|\lambda_1(\theta), \lambda_2(\theta)\right)$ where $(\lambda_1(\theta), \lambda_2(\theta)) = (-\log\theta, -\theta)$

$$\exp\left[\lambda_1 x + \lambda_2 y - \log\left[\frac{1}{\lambda_2}\left(e^{\lambda_2 t} - 1\right) + e^{\lambda_1 + \lambda_2 t}\right]\right]$$

  this is curved exponential family
- Bias of MLE is given by information geometric formula

$$-\frac{1}{2n}\left\{\Gamma^{(-1)\,a}_{cd} g^{cd} + h^{(-1)\,a}_{\kappa\lambda} g^{\kappa\lambda}\right\}$$

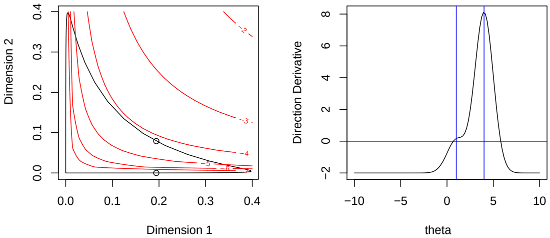
- This formula is 'not difficult' in the sense only uses sums and partial derivatives, but not used in practice
- Can this be computed numerically? treat formula as pseudo-code

Mixture Geometry

- Inference in the general class of mixture models has many hard problems:
  - singularities and multimodality in the likelihood
  - parameterisation issues
  - boundary problems
  - identification problems
- Lindsay [16] has shown how to compute Non-Parametric Maximum Likelihood Estimate using convex and affine geometry
- Mixtures are very open to geometric analysis for example local mixture models, [18] & [3]
- Other common approaches: EM and MCMC

Lindsay's geometry

- Embeds problem in finite dimensional affine space determined by sample size [14]
- For data $x_1, \ldots, x_n$ look at convex hull of curve $(f(x_1 : \theta), \ldots, f(x_n : \theta)) \subset R^n$.



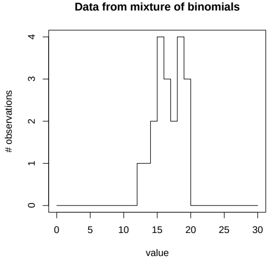
- The directional derivative in embedding space key to finding MLE

Lead by examples

- Show by examples how to build a CIG computational framework
- Start from finite discrete models and lead to general continuous models
- Show how to make information geometry tractable for mainstream users
- Show how to extend Lindsay's mixture geometry
- Open questions concerning foundations of inference and modelling

Mixtures of binomials

- Consider data from a mixture of binomials of size 30:

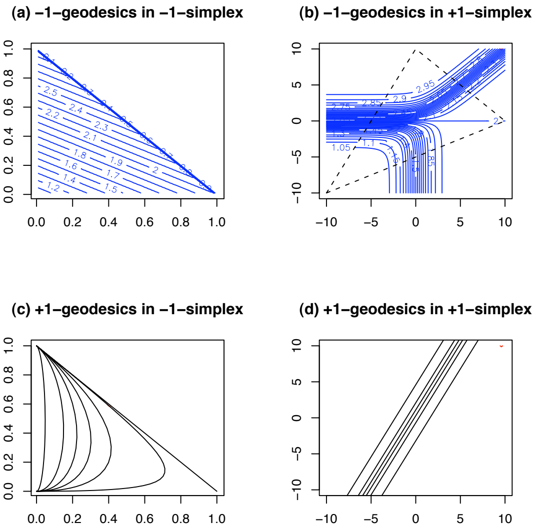


- If size is k the space of models is the simplex

$$\left\{ (\pi_0, \pi_1, \cdots, \pi_k) | \pi_i \geq 0, \text{ and } \sum \pi_i = 1 \right\}$$

- Note that include zero probabilities

Information Geometry of Simplex



Geometry of Simplex

- Simplicial models are extended exponential families since boundaries are included
- ± 1-geometries individually explict and have closed form
- hard computational tasks mixed parameterisation, see [7]
- Fisher information explicit; rank varies with dimensional of face

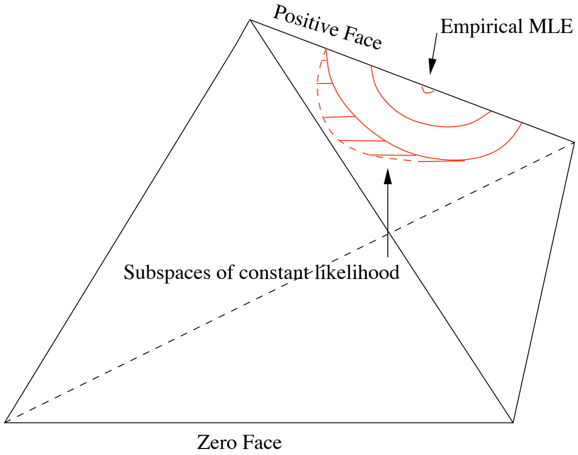
Geometry of Simplex

- Working in high dimensional simplex with large number of cells
- Typically the sample size is (much) smaller than dimension
- Sparse high dimensional simplical geometry
- The information geometry is explicit–mostly in closed form
- Normal n-asymptotics can't work
- **THEOREM:** there is a k-asymptotic theory for distribution of Deviance, see [2]

Shape of likelihood

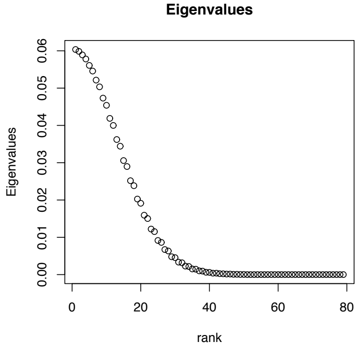
- Working in high dimensional sparse spaces much of our statistical folk-law needs to be reconsidered
- Log-likelihood not approximately quadratic
- **THEOREM:**
  - Log-likelihood concave but not strictly concave
  - There are many directions (in fact -1-affine spaces) where likelihood is flat- data can tell us nothing in these directions
  - Empirical MLE lies on face of simplex, not an interior point

Shape of likelihood



Fisher information

- Fisher information at $\pi = (\pi_1, \ldots, \pi_k)$ is $Diag(\pi) - \pi\pi^T$
- Can be arbitrarily close to singular in interior of simplex
- It is singular as take limit on faces
- **THEOREM:** The singular value decomposition of Fisher information very well understood.



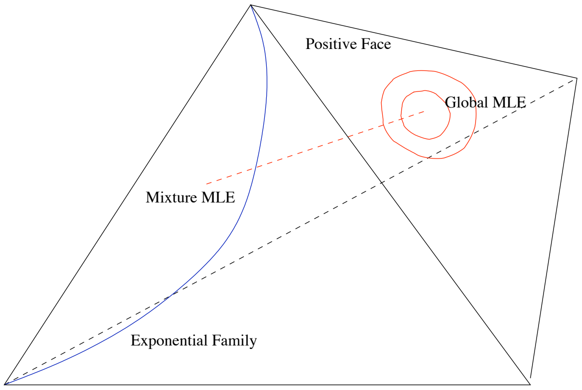
Mixture inference

- In simplex mixtures $\sum \rho_i \pi(\theta_i)$ are fundamentally not identified
- Consider finding MLE in convex hull of curve $\pi(\theta)$ in simplex
- **THEOREM:** If $\pi(\theta)$ is exponential family then convex hull has maximal dimension in simplex
- **THEOREM:** There are very good low dimensional approximations to convex hull (local mixtures)

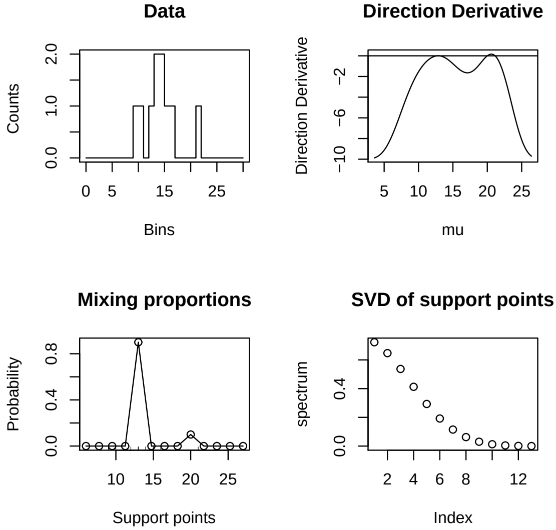
Mixture inference

- Use the geometry of the way that the low dimensional curve is embedded in the high dimensional simplex to get greatly improved algorithms
- **THEOREM:** The spectrum of the SVD of a set of points on the curve determines the quality of an approximation to the MLE in the convex hull
- This approximation method very direct method of computing MLE (and their variability) in the convex hull

Lindsay's geometry and simplex



Binomial mixture application

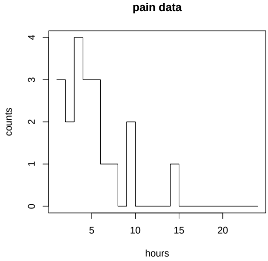


Generalisations?

- Have show that in discrete and finite case have a computational framework for the 'space of all distributions'
- High dimensional sparse simplex- sets of limits important, [9]
- Two types of affine geometry and Fisher information
- Spectral techniques very useful in order to implement numerical methods
- Can we get proxy for space of all distributions in more general settings?
- **Comment:** Computational systems must be finite and that inference is fundamentally a finite process

Pain data

- Pain data: (Wallace 1980). Hours of post-operative pain relief.
- Inference question: is there a difference between types of drug used?



- Measurements only recorded to nearest hour and no recordings after 24 hours
- Could model with censored exponential model mentioned above

Discretisation

- Binomial example naturally discrete... here have discretised a continuous model
- Discretising induces statistical curvature in models
- There are finite number of bins, one of which is semi-infinite
- There are (ordered) values of the random variable to associate with each bin
- **THEOREM:** for finite bins information loss associated with discretisation can made arbitrarily small by controlling conditional variance in bins
- Distinguish between Exponential Families which are discretised and Exponential Families in thesimplex models

Discretisation

- Pitman: [19]

  ”· · · 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.”

More Applications

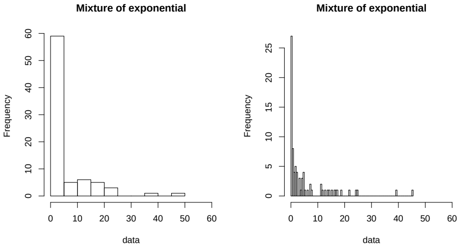
- We saw earlier the information geometric computations for censored exponential family, [17]
- Bias of MLE is given by information geometrical formula

$$-\frac{1}{2n}\left\{\Gamma^{(-1)\,a}_{cd}\,g^{cd}+h^{(-1)\,a}_{\kappa\lambda}\,g^{\kappa\lambda}\right\}$$

- In the application problem is discrete and finite
- Can treat these formulae as pseudo-code for numerical implementation in large sparse simplex
- The resulting code unlocks all results of information geometry to the mainstream user

Infinite to finite: mixture of exponentials

- Consider a problem based on mixing over exponential distributions



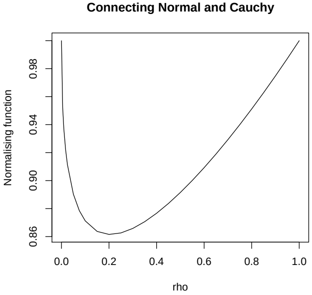
- Can discretise but now have potentially infinite number of bins

Infinite simplex

- There exists geometry of infinite simplex [1]
- Information geometry of infinite dimensional families [12] and [11] uses Hilbert or Banach space structures
- In our approach different 'faces' of the infinite simplex have different support and different moment structures
- There still exist ± 1 geodesics between distributions, but there are boundaries.

Infinite simplex

- Infinite Fisher information possible, even in mixtures of exponentials [15]
- Look geodesics joining standard normal and Cauchy, [8]
- $+1$- geodesic $f(x)^{\rho} g(x)^{1-\rho} / C(\rho)$



- -1- geodesic $(1-\rho)f(x) + \rho g(x)$, What if $\rho << 1/n$?

Infinite to finite

- To work with finite model need to make modelling assumptions
- **THEOREM:** Need to be able to truncate the Laplace transform
- Asymptotics vs fixed sample size inference: when taking fixed size approach no empirical tests possible to check modelling assumptions
- Limits to empirical knowledge-seen before in flat directions of likelihood in sparse simplex

Application: Weibull example

- Weibull is not in curved exponential family class needed for classical Information Geometry
- After making modelling assumptions can embedded Weibull family in large sparse simplical model with small loss for inference
- Make into a Curved Exponential Family so have extended Amari both theoretically and practically
- The numerical code then makes the results of extended information geometry available to mainstream user

Summary

- The way that parametric statistical models lie in a ‘space of all models’ is important
- We will use high-dimensional (extended) multinomial space as a proxy for the ‘space of all models’
- Show that this computational approach encompasses both Amari’s information geometry and Lindsay’s mixture geometry
- The geometry of the (extended) multinomial space is highly tractable and mostly explict so very good for building a computational theory
- Long term aim is to build software which releases to power of these geometric theories to mainstream

References I

[1] Anaya-Izquierdo, K., Critchley, F, Marriott P. & Vos P. (2010), Towards information geometry on the space of all distributions, submittted to *Annals Inst. Math. Statist.*

[2] Anaya-Izquierdo, K., Critchley, F, Marriott P. & Vos P. (2010), Sparse multinomial goodness of fit testing, in preparation

[3] Anaya-Izquierdo, K and Marriott, P. (2007) Local mixtures of Exponential families, *Bernoulli* Vol. 13, No. 3, 623-640.

[4] Amari, S.-I. (1985). *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics, No. 28, New York: Springer.

References II

[5] Amari, S.-I. and Nagaoka, H. (2000). *Methods of Information Geometry*. Providence, Rhode Island: American Mathematical Society.

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

[7] Barndorff-Nielsen, O. E. and Blaesild, P. (1983). *Exponential models with affine dual foliations*. Annals of Statistics, 11(3):753–769.

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

References III

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

[10] Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency), *The Annals of Statistics*, 3(6):1189–1242

[11] Fukumizu, K. (2005). Infinite dimensional exponential families by reproducing kernel hilbert spaces. *Proceedings of the 2nd International Symposium on Information Geometry and its Applications*,p324-333.

[12] Gibilisco, P. and Pistone, G. (1998). Connections on non-parametric statistical manifolds by orlicz space geometry. *Infinite Dimensional Analysis, Quantum Probability and Related Topics,* 1(2):325-347.

References IV

[13] Kass, R. E. and Vos, P. W. (1997). *Geometrical Foundations of Asymptotic Inference*. New York: Wiley.

[14] Mary L. Lesperance and John D. Kalbfleisch (1992) An Algorithm for Computing the Nonparametric MLE of a Mixing Distribution*JASA* Vol. 87, No. 417 (Mar., 1992), pp. 120-126

[15] Li P., Chen J., & Marriott P., (2009) Non-finite Fisher information and homogeneity: the EM approach, *Biometrika* 96, 2 pp 411-426.

[16] Lindsay, B.G. (1995). *Mixture models: Theory, Geometry, and Applications*, Hayward CA: Institute of Mathematical Sciences.

References V

[17] Marriott P and West S, (2002), On the Geometry of Censored Models, *Calcutta Statistical Association Bulletin* 52, pp 235-250.

[18] Marriott, P (2002), On the local geometry of Mixture Models, *Biometrika*, 89, 1, pp 77-89

[19] Pitman E.J.G, (1979), *Some basic theory for statistical inference*, London: Chapman & Hall