Information Geometry:

an overview

Paul Marriott

Department of Statistics and Actuarial Science,
University of Waterloo

WOGAS III
5th April 2011

Overview

- Overview Information Geometry (IG) in broad sense.
- Models: full, curved, extended exponential and 'universal' families
- Geometries: expected/observed dual affine spaces, information and entropy, mixture geometry
- Methods: higher order asymptotics, tensor calculus, curvature, dimension reduction and spectral techniques
- Notation: [1] refers to References while [S:7] refers to Wogas III session number

Information Geometry

- How to connect two probability density or mass functions $f(x)$ and $g(x)$ in some space of models?

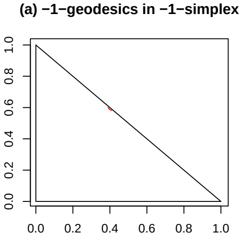
  -1: $\rho f(x) + (1-\rho) g(x)$

  +1: $\frac{f(x)^{\rho} g(x)^{1-\rho}}{C(\rho)}$
- Two different affine structures used simultaneously

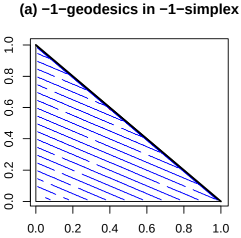
  -1: Mixture affine geometry on unit measures

  +1: Exponential affine geometry on positive measures
- Fisher Information's roles
  - measures angles and lengths
  - maps between $+1$ and -1 representations of tangent vectors, [3], [4], [18]

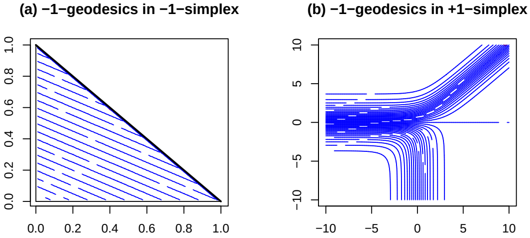
Visualising IG: extended trinomial example



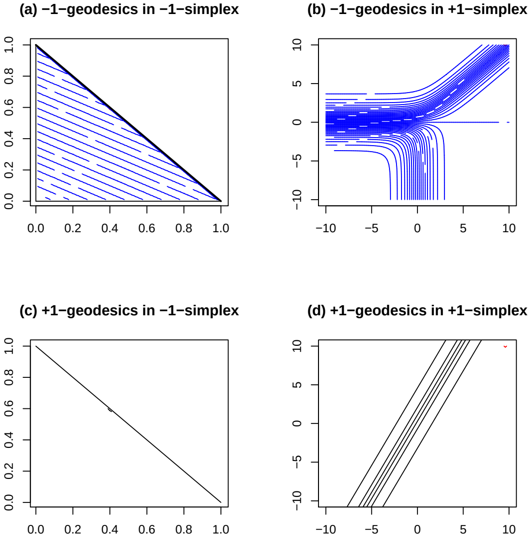
Visualising IG: extended trinomial example



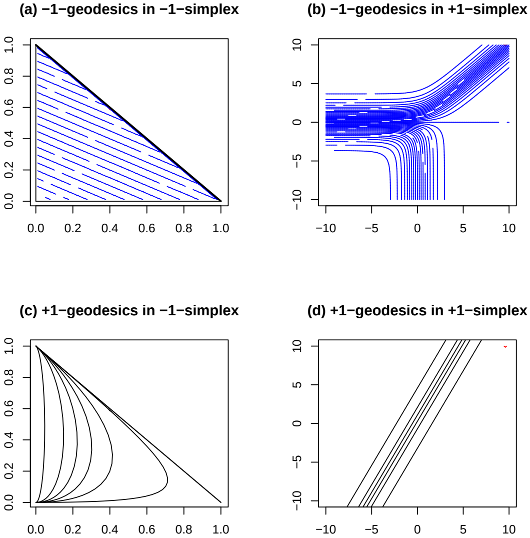
Visualising IG: extended trinomial example



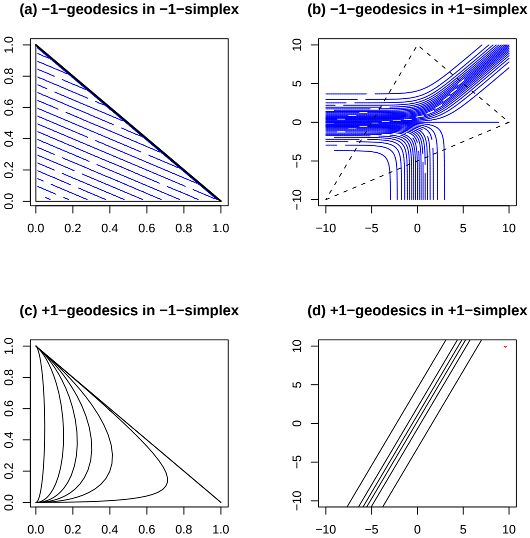
Visualising IG: extended trinomial example



Visualising IG: extended trinomial example

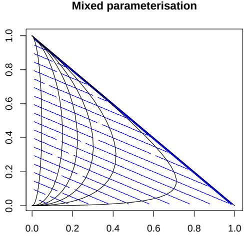


Visualising IG: extended trinomial example



Duality

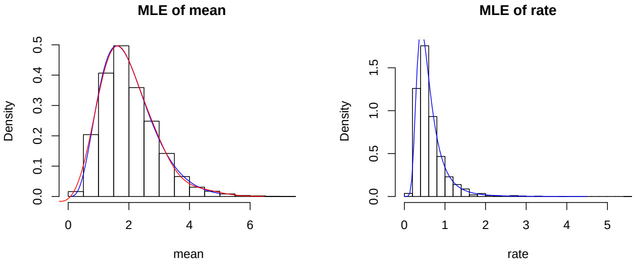
- There exists a mixed parameterisation [6] as solution of differential equation



- -1-geodesics Fisher orthogonal to $+1$-geodesics
- Limit of mixed parameters give extended exponential family
- Key to structural theorem [3] and idea of inferential cuts

Asymptotic expansions

- Strong links between IG and higher order asymptotic expansions [7]
- Can apply Edgeworth, saddlepoint or Laplace expansions [29]



- Flexible, tractable given IG, invariance properties clear [3]

Example: survival times

- Censored survival times in leukaemia patients adapted from [15]
- Censored exponential model, [18, 24]

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

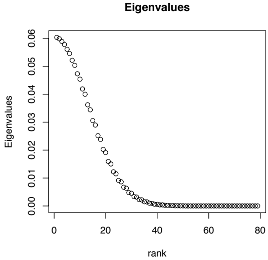
  this is curved exponential family $(\lambda_1(\mu), \lambda_2(\mu))$
- Bias of MLE is given by information geometric formula

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

- This formula is 'not difficult' in the sense only uses sums and partial derivatives.

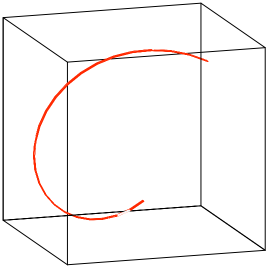
Asymptotic expansions

- High dimensional calculus though tensor analysis, McCullagh [27]
- Many terms need to be computed in high dimensional problems [S:7]
- Language issue
- Singularity of Fisher information matters



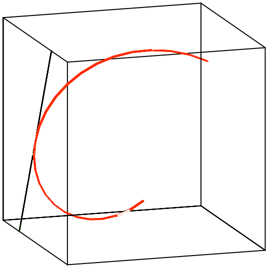
- Fisher information can be singular or infinite [22]

Embedding curvature and affine approximation



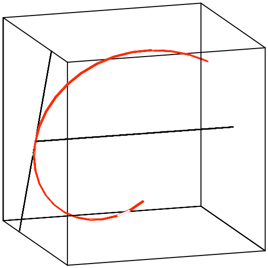
- Curvature(s) key part(s) of differential geometry
- Tangent space gives best linear approximation
- Tangent and curvature gives best two dimensional affine embedding space

Embedding curvature and affine approximation



- Curvature(s) key part(s) of differential geometry
- Tangent space gives best linear approximation
- Tangent and curvature gives best two dimensional affine embedding space

Embedding curvature and affine approximation



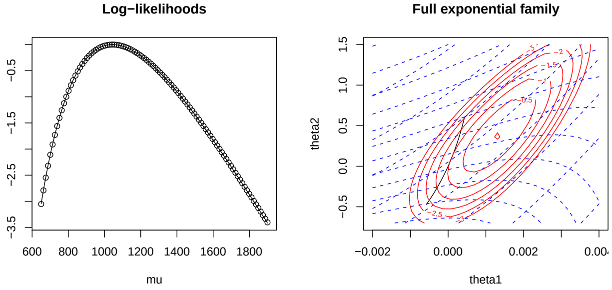
- Curvature(s) key part(s) of differential geometry
- Tangent space gives best linear approximation
- Tangent and curvature gives best two dimensional affine embedding space

Curvature and affine approximation

- Different affine geometries give different approximating spaces
- Low dimensional $+1$-affine spaces give approximate sufficient statistics [26]
- Low dimensional -1 approximations give limits to identification and computation in mixture models [25], [2], [S:3], [S:6]

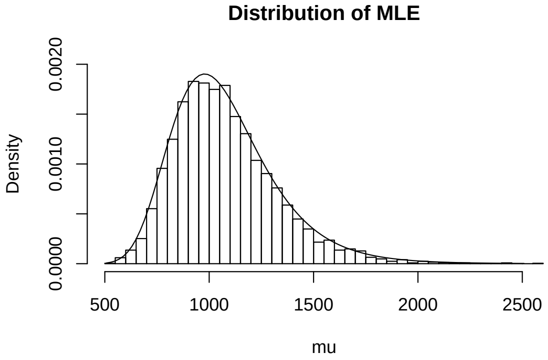
Example: survival times

- Example: censored survival times in leukaemia patients adapted from [15]
- Use censored exponential distribution



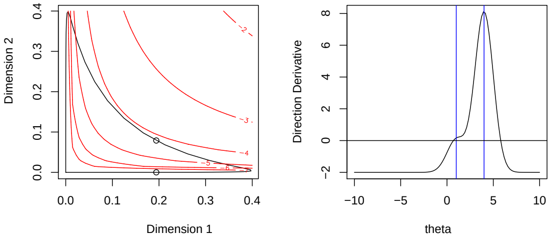
Example: survival times

- Example: censored survival times in leukaemia patients adapted from [15]
- Use censored exponential distribution



Mixture Geometry

- Lindsay [23] embeds problem in finite dimensional affine space determined by sample size [21]
- Enough structure to compute non-parametric maximum likelihood, [21].



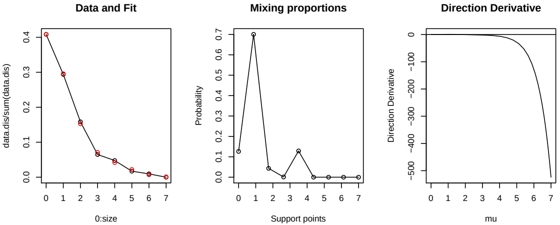
- Directional derivative is key tool to maximise likelihood over a -1-convex hull [S:3]

Mixture geometries

- Lindsay's geometry is finite dimensional version of Amari's -1 geometry
- Need to work in convex hull in -1-dimensional affine space
- Low dimensional -1 approximations give limits to identification and computation in mixture models, [2]
- Information geometry can give efficient approximation of high dimensional convex hulls by polytopes

Example: mixture of binomials

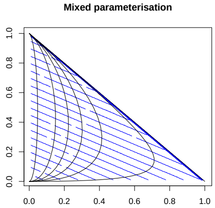
- Toxicological experiment [20] studied frequencies of dead implants in rats
- 'simple one-parameter binomial [...] models generally provide poor fits to this type of binary data'



- IG gives ways to explore convex hull efficiently

Information and entropy

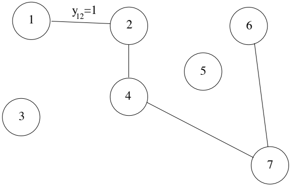
- Given a set of statistics $s_i(x)$ construct model which maximise entropy with fixed $E(s_i(S))$, [S:1], [S:2]
- Models which are orthogonal to level sets of $E(s_i(S))$ called least informative models



- Pythagorean results minimising KL divergence by orthogonal projection, [S:6]
- Links with decision theory [14], non parameteric methods such as bootstrap and empirical likelihood, [S:5]

Network Models

- High dimensional (curved) exponential family examples based on network models, [16], [S:1]
- Binary indicator functions Y_{ij} such that we have 1 if an edge exists from i to j and zero otherwise.



- Build 'least informative model'

$$P_\eta(Y = y) = \frac{\exp\left\{\eta^T g(y, X)\right\}}{\kappa(\eta)}$$

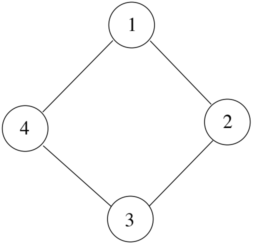
where sufficient statistics are graph statistics

Example: New England Lawyers

- [19] looks at working relations among 36 partners in a New England law firm
- Computing $\kappa(\eta)$ is typically intractable since a sum over 2^{630} terms
- Approaches include [S:8]
  - Pseudo-likelihood [30]
  - simulated moments [28]
  - MCMC [17]

Graphical models: FEF

- Consider the example from [13] of the cyclic graph of order 4 with binary values at each node, [S:1], [S:4].



- Models lie in 15-dimensonal simplex, but with constraints imposed by conditional independence
- Constraints linear in $+1$-affine parameters

$$\eta_i + \eta_j = \eta_k + \eta_l$$

- So get full exponential family

Graphical models: CEF

- Give *ordered* set of discrete random variables X_i, $i = 1, 2, 3$ be binary random variables, [S:1], [S:4].
- The simplex which describes their distribution is $2^3 - 1 = 7$ dimensional
- A DAG defines dependences for example the simple graph

$$1 \longrightarrow 2 \longrightarrow 3$$

or $P(X_3|X_2, X_1) = P(X_3|X_2)$

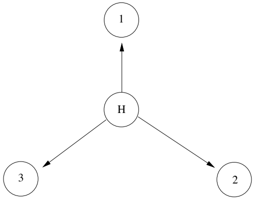
- These constraints give non-linear constraints in $+1$ affine space

$$\eta_{001} = \log\left(\frac{(1-\pi^3_{10})(1-\pi^2_{11})\pi_1}{\pi^3_{11}\pi^2_{11}\pi_1}\right)$$

and curved exponential families [13].

DAG with hidden variables

- In multinomials independence is expressible as a finite set of polynomial equalities, [S:4].
- Add hidden variables



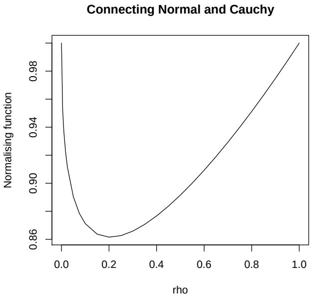
- Example lies in 7 dimensional simplex- mixes over a 3 dimensional CEF
- The model space is not a manifold but a variety- union of different dimensional manifolds- extended exponential family

Infinite dimensional issues

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

Neighbourhood of Model

- Look $+1$-geodesic joining standard normal and Cauchy, [8]
- Given by $f(x)^\rho g(x)^{1-\rho}/C(\rho)$



- Infinite Fisher information possible [22]

Neighbourhood of Model

- Sensitivity of inference to model assumptions by understanding 'neighbourhood of model', [S:3]
- Links to non and semi-parametrics
- Mixture of normal and Cauchy, -1- geodesic $(1-\rho)f(x)+\rho g(x)$
  - If $\rho << 1/n$ models very 'close' by some measures
  - models very different by other measures
- Asymmetry of KL divergence

Summary

- Overview Information geometry in broad sense.
- Models: full, curved and extended exponential, and ‘universal’ families
- Geometries: expected/observed dual flat manifolds, information theory, mixture geometry
- Methods: Higher order asymptotics, tensor calculus, curvature, dimension reduction and spectral techniques

References I

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

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

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

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

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

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

[7] Barndorff-Nielsen, O.E. and Cox, D.R, (1994), *Inference and Asymptotics*, Chapman & Hall:London

References II

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

[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 III

[13] D. Geiger, D.Heckerman, H.King and C. Meek (2001) Stratified Exponential Families: Graphical Models and Model Selection, *Annals of Statistics*, Vol. 29, No. 2, pp 505-529

[14] Grunwald P.D. and Dawid A.P. (2004) Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory, *Annals of Statistics*, 32, 4 1267-1433

[15] Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. and Ostrowski, E. (1994), A handbook of small data sets , Chapman & Hall, London.

[16] D. Hunter, (2007) Curved exponential family models for social networks, *Social Networks*, 29 216–230.

[17] Hunter D, Handcock M. (2006) Inference in curved exponential family models for networks *J. of Computational and Graphical Statistics* 15, 565-583

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

References IV

[19] Lazega, E., Pattison, P.E., (1999) Multiplexity, generalized exchange and cooperation in organizations: a case study. Social Networks 21, 67–90

[20] L. L. Kupper L.L., and Haseman J.K., (1978), The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments, *Biometrics*, Vol. 34, No. 1 (Mar., 1978), pp. 69-76

[21] 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

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

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

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

References V

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

[26] Marriott, P., & Vos, P. (2004), On The Global Geometry of Parametric Models and Information Recovery, *Bernoulli*, **10** (2), 1-11

[27] McCullagh P. (1987) *Tensor Methods in Statistics*, Chapman and Hall, London.

[28] Snijders, T. A. B. (2002), Markov Chain Monte Carlo Estimation of Exponential Random Graph Models, *Journal of Social Structure*, 3.

[29] Small, C.G. (2010) *Expansions and asymptotics for statistics*, Chapman and Hall

[30] Strauss, D., and Ikeda, M. (1990), Pseudolikelihood Estimation for Social Networks,*Journal of the American Statistical Association* , 85, 204–212.