Computational Information Geometry:

Geometry of Model Choice

Paul Marriott

University of Waterloo

Information Geometry and its Applications III
Leipzig August 2010

Big Picture

- Introduction to Computational Information Geometry

Big Picture

- Introduction to Computational Information Geometry
- The overall objective is to construct diagnostic tools to help understand sensitivity to model choice

Big Picture

- Introduction to Computational Information Geometry
- The overall objective is to construct diagnostic tools to help understand sensitivity to model choice
- Targeted at applications where Generalised Linear Models are used

Big Picture

- Introduction to Computational Information Geometry
- The overall objective is to construct diagnostic tools to help understand sensitivity to model choice
- Targeted at applications where Generalised Linear Models are used
- Our geometry is affine and convex, not manifold based: non-constant support and moment structure

Big Picture

- Introduction to Computational Information Geometry
- The overall objective is to construct diagnostic tools to help understand sensitivity to model choice
- Targeted at applications where Generalised Linear Models are used
- Our geometry is affine and convex, not manifold based: non-constant support and moment structure
- Topology defined via duality

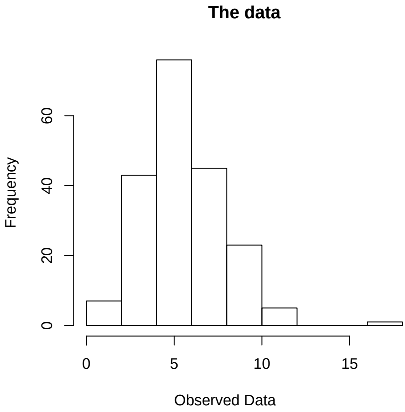
Big Picture

- Introduction to Computational Information Geometry
- The overall objective is to construct diagnostic tools to help understand sensitivity to model choice
- Targeted at applications where Generalised Linear Models are used
- Our geometry is affine and convex, not manifold based: non-constant support and moment structure
- Topology defined via duality
- Joint work with Karim Anaya-Izquierdo, Frank Critchley and Paul Vos

Big Picture

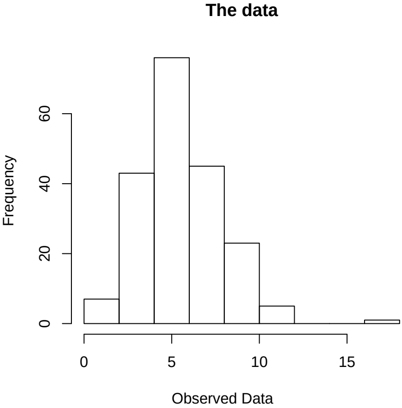
- Introduction to Computational Information Geometry
- The overall objective is to construct diagnostic tools to help understand sensitivity to model choice
- Targeted at applications where Generalised Linear Models are used
- Our geometry is affine and convex, not manifold based: non-constant support and moment structure
- Topology defined via duality
- Joint work with Karim Anaya-Izquierdo, Frank Critchley and Paul Vos
- Thanks to EPSRC Grant Number EP/E017878/1

Problem of Interest



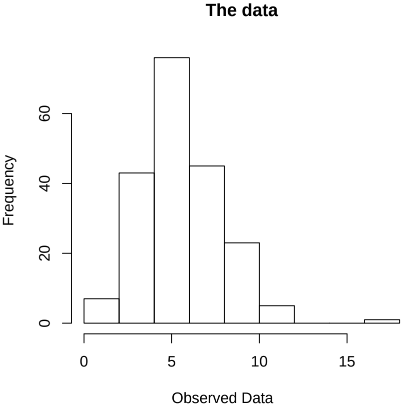
- Question: what is the population mean?
- How do modelling assumptions affect inference about mean?
- Can geometry of ‘space of all models’ give a framework for discussion?

Problem of Interest



- Question: what is the population mean?
- How do modelling assumptions affect inference about mean?
- Can geometry of ‘space of all models’ give a framework for discussion?

Problem of Interest



- Question: what is the population mean?
- How do modelling assumptions affect inference about mean?
- Can geometry of ‘space of all models’ give a framework for discussion?

Structured Extended Multinomials

- *Extended* multinomials are multinomial but allow cell probabilities to be zero.

Structured Extended Multinomials

- *Extended* multinomials are multinomial but allow cell probabilities to be zero.
- Discretizing continuous data gives categorical models with structure on the cells

Structured Extended Multinomials

- *Extended* multinomials are multinomial but allow cell probabilities to be zero.
- Discretizing continuous data gives categorical models with structure on the cells
- Examples of structure:

Structured Extended Multinomials

- *Extended* multinomials are multinomial but allow cell probabilities to be zero.
- Discretizing continuous data gives categorical models with structure on the cells
- Examples of structure:
  - numerical labels
  - ordering

Structured Extended Multinomials

- *Extended* multinomials are multinomial but allow cell probabilities to be zero.
- Discretizing continuous data gives categorical models with structure on the cells
- Examples of structure:
  - numerical labels
  - ordering
- *Structured* Extended Multinomials (SEM) include this *sample space* structure

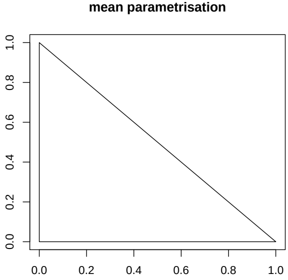
Structured Extended Multinomials

- *Extended* multinomials are multinomial but allow cell probabilities to be zero.
- Discretizing continuous data gives categorical models with structure on the cells
- Examples of structure:
    - numerical labels
    - ordering
- *Structured* Extended Multinomials (SEM) include this *sample space* structure
- SEM proxy for universal space of all distributions

Structured Extended Multinomials

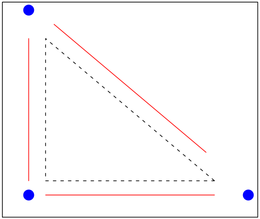
- *Extended* multinomials are multinomial but allow cell probabilities to be zero.
- Discretizing continuous data gives categorical models with structure on the cells
- Examples of structure:
    - numerical labels
    - ordering
- *Structured* Extended Multinomials (SEM) include this *sample space* structure
- SEM proxy for universal space of all distributions
- Theory for infinite dimensions; in practice, finite dimensional.

-1 -simplical structure



- Mean (-1) parameters can be on boundary
- Different support sets
- Union of exponential families each with corresponding natural $(+1)$ parameters

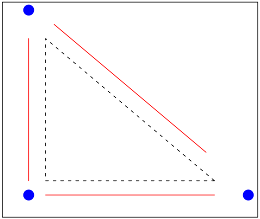
-1 -simplical structure



- Mean (-1) parameters can be on boundary
- Different support sets
- Union of exponential families each with corresponding natural $(+1)$ parameters

-1 -simplical structure

Support sets



- Mean (-1) parameters can be on boundary
- Different support sets
- Union of exponential families each with corresponding natural $(+1)$ parameters

+1 simplex structure

- How can we define $+1$ structure on the extended multinomial?

+1 simplex structure

- How can we define $+1$ structure on the extended multinomial?
- Problem: the support and the moment structure changes across SEM

+1 simplex structure

- How can we define $+1$ structure on the extended multinomial?
- Problem: the support and the moment structure changes across SEM
- Need to glue together different exponential families

+1 simplex structure

- How can we define $+1$ structure on the extended multinomial?
- Problem: the support and the moment structure changes across SEM
- Need to glue together different exponential families
- Use the dual structure of information geometry

Very Simple Example

- Sample space of 3 values: $\{t_0, t_1, t_2\}$

Very Simple Example

- Sample space of 3 values: $\{t_0, t_1, t_2\}$
- Multinomial distribution
$\Delta_{int} = \{(\pi_0, \pi_1, \pi_2) | \pi_i > 0, \ \sum \pi_i = 1\}$

Very Simple Example

- Sample space of 3 values: $\{t_0, t_1, t_2\}$
- Multinomial distribution
  $\Delta_{int} = \{(\pi_0, \pi_1, \pi_2) | \pi_i > 0, \sum \pi_i = 1\}$
  $\Delta^*_{int} = \{(\eta_1, \eta_2) | \eta_i = \log(\pi_i / \pi_0)\}$

Very Simple Example

- Sample space of 3 values: $\{t_0, t_1, t_2\}$
- Multinomial distribution
  $\Delta_{int} = \{(\pi_0, \pi_1, \pi_2) | \pi_i > 0, \sum \pi_i = 1\}$
  $\Delta^*_{int} = \{(\eta_1, \eta_2) | \eta_i = \log(\pi_i / \pi_0)\}$
- Extended Multinomial distribution
  $\Delta = \{(\pi_1, \pi_2, \pi_3) | \pi_i \geq 0, \sum \pi_i = 1\}$

Very Simple Example

- Sample space of 3 values: $\{t_0, t_1, t_2\}$
- Multinomial distribution
  $\Delta_{int} = \{(\pi_0, \pi_1, \pi_2) | \pi_i > 0, \sum \pi_i = 1\}$
  $\Delta^*_{int} = \{(\eta_1, \eta_2) | \eta_i = \log(\pi_i/\pi_0)\}$
- Extended Multinomial distribution
  $\Delta = \{(\pi_1, \pi_2, \pi_3) | \pi_i \geq 0, \sum \pi_i = 1\}$
  $\Delta^* =$ more complicated; but has been done.
- Likelihood and KL divergence defined on Δ_{int} (and Δ)

Very Simple Example

- Sample space of 3 values: $\{t_0, t_1, t_2\}$
- Multinomial distribution
  $\Delta_{int} = \{(\pi_0, \pi_1, \pi_2) | \pi_i > 0, \ \sum \pi_i = 1\}$
  $\Delta^*_{int} = \{(\eta_1, \eta_2) | \eta_i = \log(\pi_i / \pi_0)\}$
- Extended Multinomial distribution
  $\Delta = \{(\pi_1, \pi_2, \pi_3) | \pi_i \geq 0, \ \sum \pi_i = 1\}$
  $\Delta^* =$ more complicated; but has been done.
- Likelihood and KL divergence defined on Δ_{int} (and Δ)
  – These quantities do not depend on $\{t_0, t_1, t_2\}$.

Very Simple Example

- Sample space of 3 values: $\{t_0, t_1, t_2\}$
- Multinomial distribution
  $\Delta_{int} = \{(\pi_0, \pi_1, \pi_2) | \pi_i > 0, \sum \pi_i = 1\}$
  $\Delta^*_{int} = \{(\eta_1, \eta_2) | \eta_i = \log(\pi_i/\pi_0)\}$
- Extended Multinomial distribution
  $\Delta = \{(\pi_1, \pi_2, \pi_3) | \pi_i \geq 0, \sum \pi_i = 1\}$
  $\Delta^* =$ more complicated; but has been done.
- Likelihood and KL divergence defined on Δ_{int} (and Δ)
  – These quantities do not depend on $\{t_0, t_1, t_2\}$.
  – To use the structure in the sample space we need:

Very Simple Example

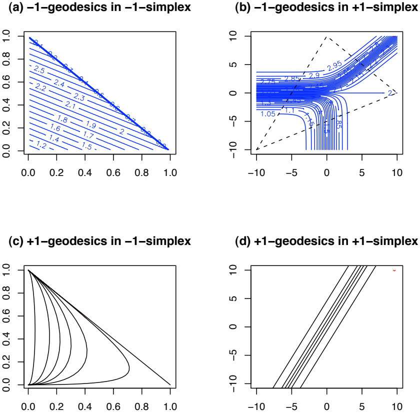
- Sample space of 3 values: $\{t_0, t_1, t_2\}$
- Multinomial distribution
  $\Delta_{int} = \{(\pi_0, \pi_1, \pi_2) | \pi_i > 0, \ \sum \pi_i = 1\}$
  $\Delta^*_{int} = \{(\eta_1, \eta_2) | \eta_i = \log(\pi_i/\pi_0)\}$
- Extended Multinomial distribution
  $\Delta = \{(\pi_1, \pi_2, \pi_3) | \pi_i \geq 0, \ \sum \pi_i = 1\}$
  $\Delta^* =$ more complicated; but has been done.
- Likelihood and KL divergence defined on Δ_{int} (and Δ)
  – These quantities do not depend on $\{t_0, t_1, t_2\}$.
  – To use the structure in the sample space we need:
- Structured Extended Multinomial
  $\{\Delta, \Delta^*, (t_0, t_1, t_2)\}$

Very Simple Example

- Sample space of 3 values: $\{t_0, t_1, t_2\}$
- Multinomial distribution
  $\Delta_{int} = \{(\pi_0, \pi_1, \pi_2) | \pi_i > 0, \ \sum \pi_i = 1\}$
  $\Delta^*_{int} = \{(\eta_1, \eta_2) | \eta_i = \log(\pi_i / \pi_0)\}$
- Extended Multinomial distribution
  $\Delta = \{(\pi_1, \pi_2, \pi_3) | \pi_i \geq 0, \ \sum \pi_i = 1\}$
  $\Delta^* =$ more complicated; but has been done.
- Likelihood and KL divergence defined on Δ_{int} (and Δ)
  – These quantities do not depend on $\{t_0, t_1, t_2\}$.
  – To use the structure in the sample space we need:
- Structured Extended Multinomial
  $\{\Delta, \Delta^*, (t_0, t_1, t_2)\}$
- We'll take $(t_0, t_1, t_2) = (0, 1, 2)$

Dual Parameterisations

Mean Parameter (Δ) | Natural Parameter(Δ^*)



Computational Information Geometry

- Want to be able to numerically compute in high dimensional SEM

Computational Information Geometry

- Want to be able to numerically compute in high dimensional SEM
- Need to get the topology and geometry right

Computational Information Geometry

- Want to be able to numerically compute in high dimensional SEM
- Need to get the topology and geometry right
- Information Geometry sits naturally on these simplicial structures – different from manifolds

Computational Information Geometry

- Want to be able to numerically compute in high dimensional SEM
- Need to get the topology and geometry right
- Information Geometry sits naturally on these simplicial structures – different from manifolds
- Almost all of the information geometry on SEM is numerically easy

Computational Information Geometry

- Want to be able to numerically compute in high dimensional SEM
- Need to get the topology and geometry right
- Information Geometry sits naturally on these simplicial structures – different from manifolds
- Almost all of the information geometry on SEM is numerically easy
- Hard part: the mixed parameterisation

Computational Information Geometry

- Want to be able to numerically compute in high dimensional SEM
- Need to get the topology and geometry right
- Information Geometry sits naturally on these simplicial structures – different from manifolds
- Almost all of the information geometry on SEM is numerically easy
- Hard part: the mixed parameterisation
- This is our computational framework

Pain Relief Example

- Paul Meier *et al*, Reconsideration of Methodology in Studies of Pain Relief, *Biometrics* (1958), pp. 330-342

Pain Relief Example

- Paul Meier *et al*, Reconsideration of Methodology in Studies of Pain Relief, *Biometrics* (1958), pp. 330-342
- Response Y: number of hours for which pain relief was greater than 50%

Pain Relief Example

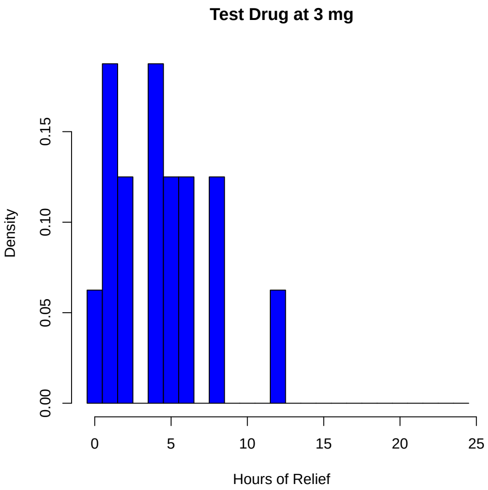
- Paul Meier *et al*, Reconsideration of Methodology in Studies of Pain Relief, *Biometrics* (1958), pp. 330-342
- Response Y: number of hours for which pain relief was greater than 50%
- Explanatory var X: Drug (Demerol, T 1mg, T 3mg)

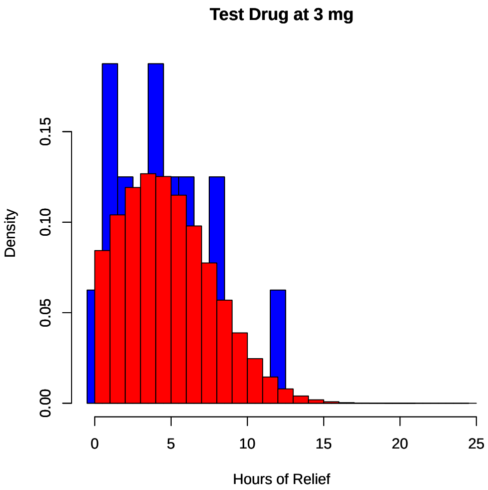
Pain Relief Example

- Paul Meier *et al*, Reconsideration of Methodology in Studies of Pain Relief, *Biometrics* (1958), pp. 330-342
- Response Y: number of hours for which pain relief was greater than 50%
- Explanatory var X: Drug (Demerol, T 1mg, T 3mg)
- 43 patients randomly assigned; double blind

Pain Relief Example

- Paul Meier *et al*, Reconsideration of Methodology in Studies of Pain Relief, *Biometrics* (1958), pp. 330-342
- Response Y: number of hours for which pain relief was greater than 50%
- Explanatory var X: Drug (Demerol, T 1mg, T 3mg)
- 43 patients randomly assigned; double blind
- Inference question concerns differences of means across groups


Test Drug at 3 mg
Density
0.00
0.05
0.10
0.15
0
5
10
15
20
25
Hours of Relief


Test Drug at 3 mg
Density
0.00
0.05
0.10
0.15
0
5
10
15
20
25
Hours of Relief

Model Family Restriction

- The universality of the SEM is very rich, it is desirable to have restrictions

Model Family Restriction

- The universality of the SEM is very rich, it is desirable to have restrictions
- Goodness-of-fit tests can reveal poor model familes; many diverse model families remain

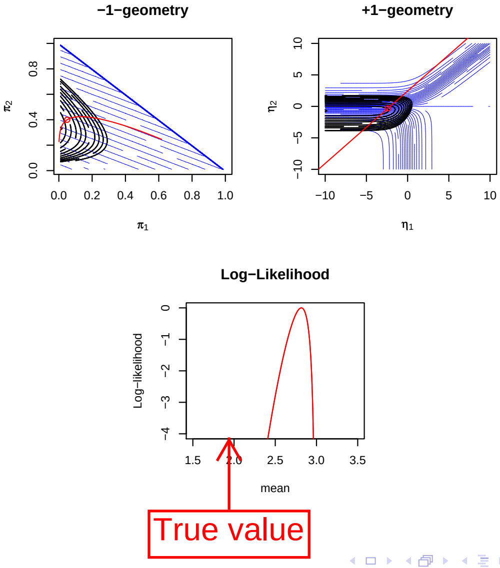
Model Family Restriction

- The universality of the SEM is very rich, it is desirable to have restrictions
- Goodness-of-fit tests can reveal poor model familes; many diverse model families remain
- Question: Among data-supported model families, how does inference depend on the choice of model family?

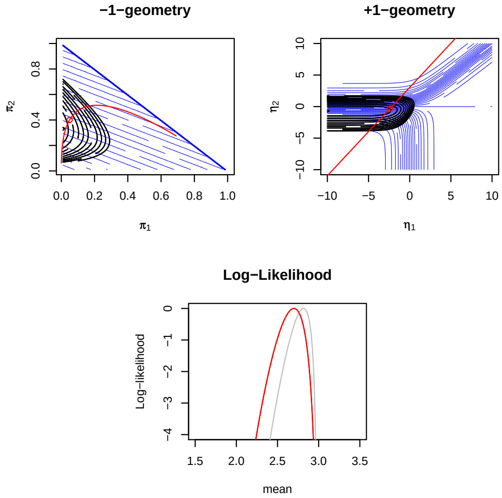
Model Family Restriction

- The universality of the SEM is very rich, it is desirable to have restrictions
- Goodness-of-fit tests can reveal poor model familes; many diverse model families remain
- Question: Among data-supported model families, how does inference depend on the choice of model family?
- C.f. Results of Copas and Eguchi

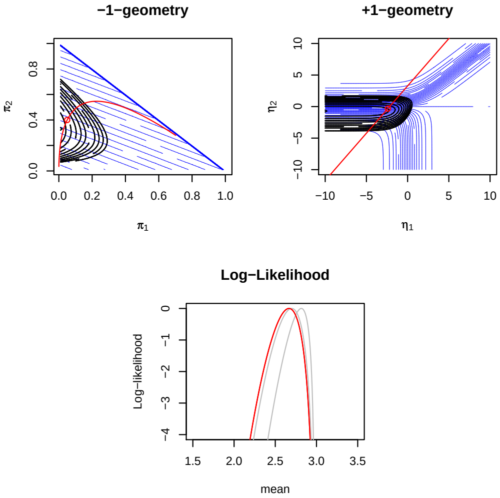
Thought experiment



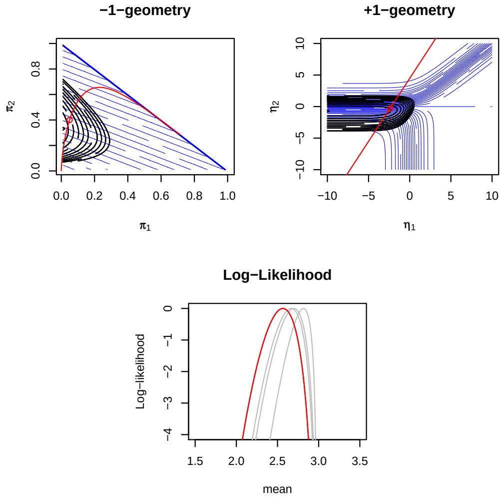
Thought experiment



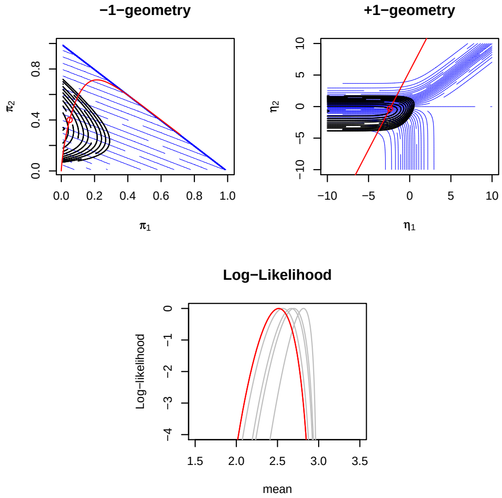
Thought experiment



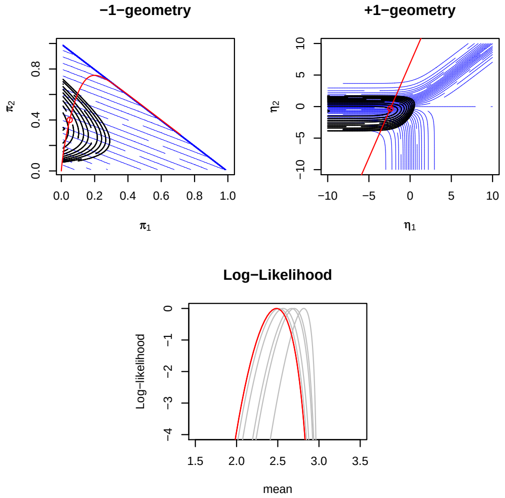
Thought experiment



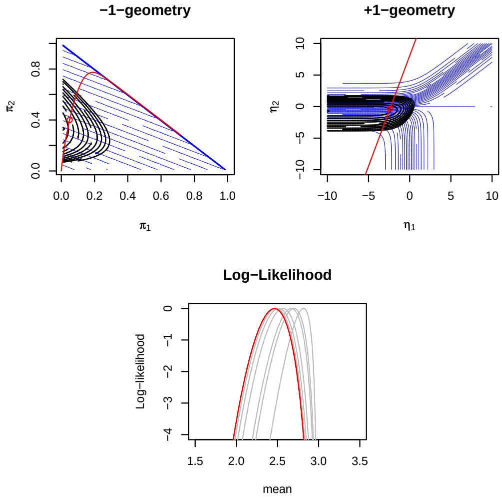
Thought experiment



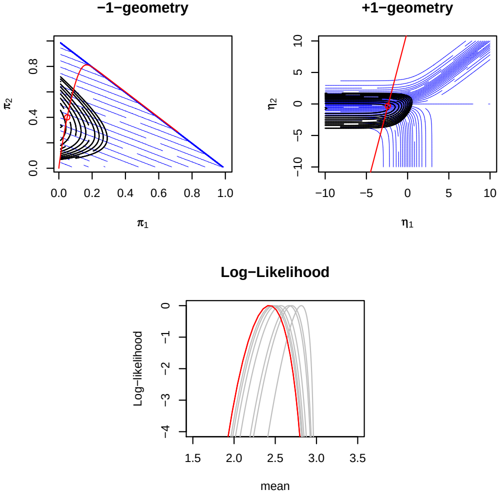
Thought experiment



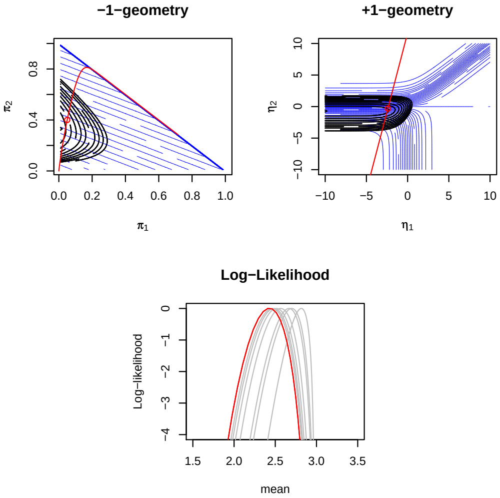
Thought experiment



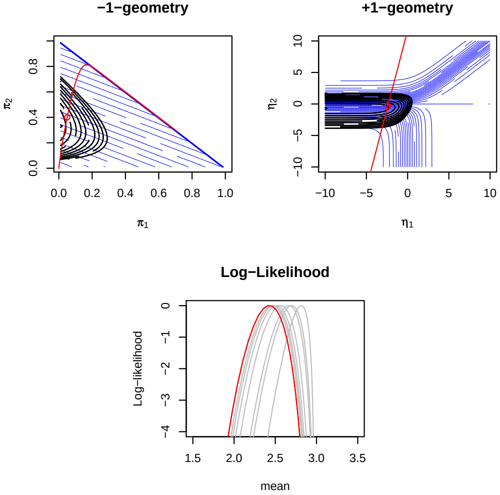
Thought experiment



Thought experiment



Thought experiment



Least informative model

- Since rotation is through data generation process (DGP)we can't use goodness-of-fit tests to distinguish between model families

Least informative model

- Since rotation is through data generation process (DGP)we can't use goodness-of-fit tests to distinguish between model families
- Rotation changes the mode and the shape of the likelihood

Least informative model

- Since rotation is through data generation process (DGP)we can't use goodness-of-fit tests to distinguish between model families
- Rotation changes the mode and the shape of the likelihood
- Smallest Expected Fisher information at DGP is when exponential model is orthogonal to level sets of mean

Least informative model

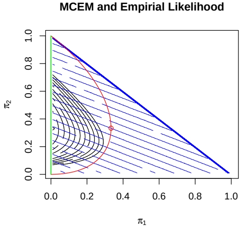
- Since rotation is through data generation process (DGP)we can't use goodness-of-fit tests to distinguish between model families
- Rotation changes the mode and the shape of the likelihood
- Smallest Expected Fisher information at DGP is when exponential model is orthogonal to level sets of mean
- Families of Models with this orthogonality property we call *least informative model families*

Least informative model

- Since rotation is through data generation process (DGP)we can't use goodness-of-fit tests to distinguish between model families
- Rotation changes the mode and the shape of the likelihood
- Smallest Expected Fisher information at DGP is when exponential model is orthogonal to level sets of mean
- Families of Models with this orthogonality property we call *least informative model families*
- Information in inference comes from two sources: (i) data and (ii) modelling assumptions. To be conservative minimise (ii) relative to (i)

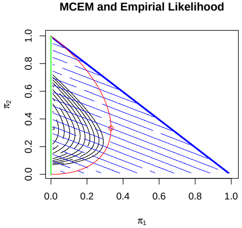
Least informative models

- Moment Constrained Maximum Entropy (MCME) select model which maximizes entropy for fixed moments is a special case of a least informative model



Least informative models

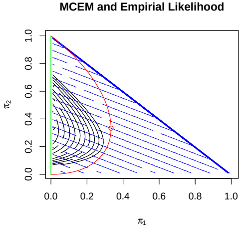
- Moment Constrained Maximum Entropy (MCME) select model which maximizes entropy for fixed moments is a special case of a least informative model



- Bootstrapping “least favourable models” Efron (1981)

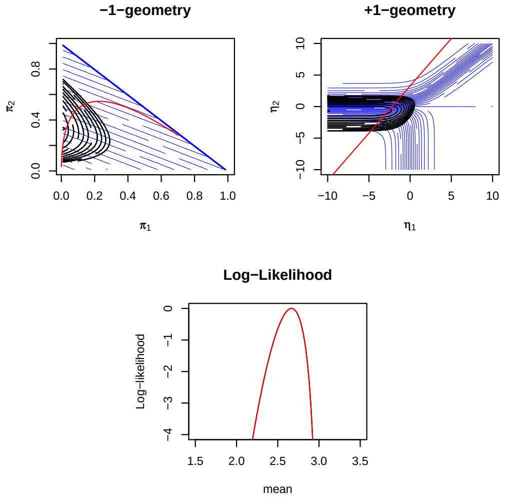
Least informative models

- Moment Constrained Maximum Entropy (MCME) select model which maximizes entropy for fixed moments is a special case of a least informative model

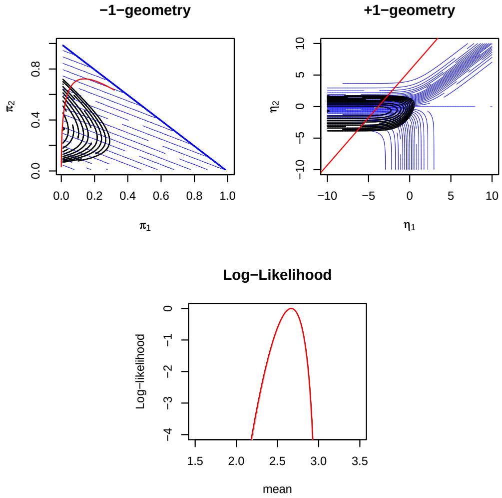


- Bootstrapping “least favourable models” Efron (1981)
- Empirical likelihood: maximize likelihood for fixed moments

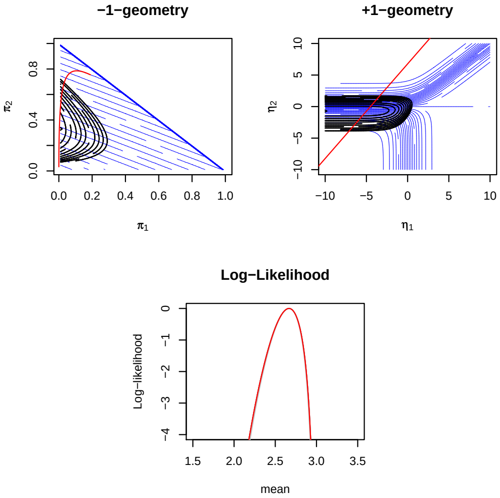
Effect of translation



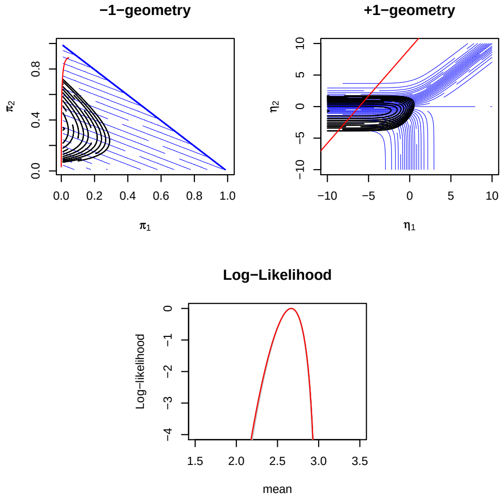
Effect of translation



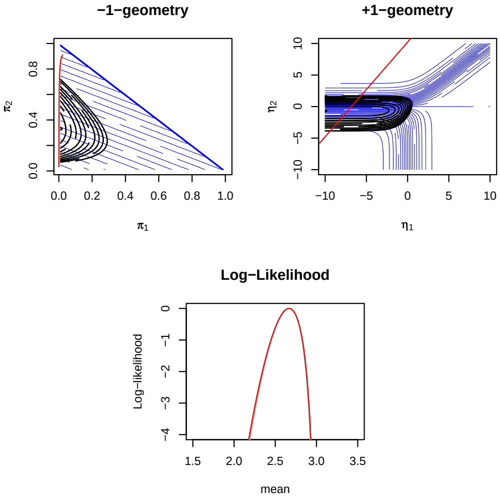
Effect of translation



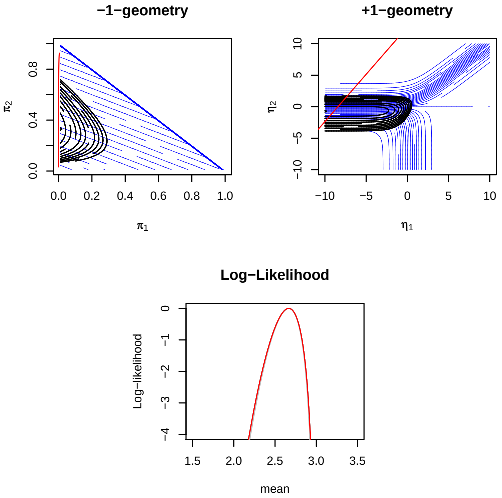
Effect of translation



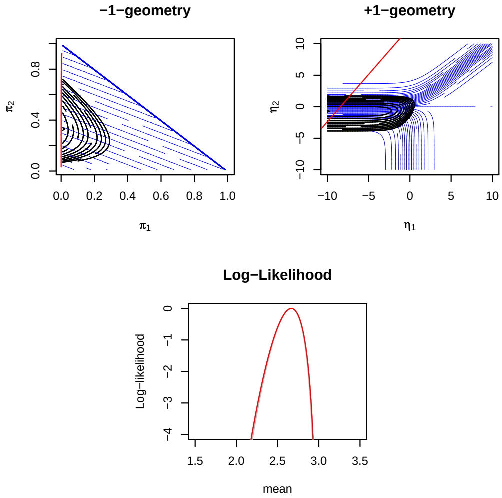
Effect of translation



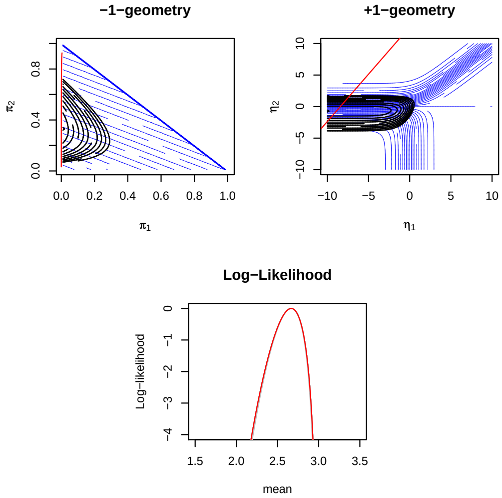
Effect of translation



Effect of translation



Effect of translation



Sensitive perturbations

- There exists large perturbations of models which have ‘no effect’ on inference

Sensitive perturbations

- There exists large perturbations of models which have ‘no effect’ on inference
- Limit of these translation exists-use the correct topology

Sensitive perturbations

- There exists large perturbations of models which have ‘no effect’ on inference
- Limit of these translation exists-use the correct topology
- Limit is Empirical Likelihood

Sensitive perturbations

- There exists large perturbations of models which have ‘no effect’ on inference
- Limit of these translation exists-use the correct topology
- Limit is Empirical Likelihood
- Shows link between least informative parametric inference and non-parametric inference

Sensitive perturbations

- There exists large perturbations of models which have ‘no effect’ on inference
- Limit of these translation exists-use the correct topology
- Limit is Empirical Likelihood
- Shows link between least informative parametric inference and non-parametric inference
- The number of perturbations which matter for inference about the mean can be surprising small

Sensitive perturbations

- There exists large perturbations of models which have ‘no effect’ on inference
- Limit of these translation exists-use the correct topology
- Limit is Empirical Likelihood
- Shows link between least informative parametric inference and non-parametric inference
- The number of perturbations which matter for inference about the mean can be surprising small

Exploring Model space

- Our geometry allows us to explore the range of inferences about mean across different data-plausible models

Exploring Model space

- Our geometry allows us to explore the range of inferences about mean across different data-plausible models
- The inferences drawn depend upon (i) data through choice of sufficient statistics (rotations) (ii) other modelling assumptions (translations)

Exploring Model space

- Our geometry allows us to explore the range of inferences about mean across different data-plausible models
- The inferences drawn depend upon (i) data through choice of sufficient statistics (rotations) (ii) other modelling assumptions (translations)
- If number of sufficient statistics is greater than the dimension of interest parameter then need to select way of drawing marginal inference e.g. plug-in, profile likelihood, marginal posterior . . .

Exploring Model space: pain example

- In this example a GLM approach is often taken in practice

Exploring Model space: pain example

- In this example a GLM approach is often taken in practice
- This selects a two dimensional sufficient statistic and corresponding exponential family

Exploring Model space: pain example

- In this example a GLM approach is often taken in practice
- This selects a two dimensional sufficient statistic and corresponding exponential family
    - Gamma family
    - Log-Normal family
    - Moment Constrained Maximum Entropy family

Exploring Model space: pain example

- In this example a GLM approach is often taken in practice
- This selects a two dimensional sufficient statistic and corresponding exponential family
  - Gamma family
  - Log-Normal family
  - Moment Constrained Maximum Entropy family
- Fixes a mean-variance relationship through specification of the variance function $\phi V(\mu)$

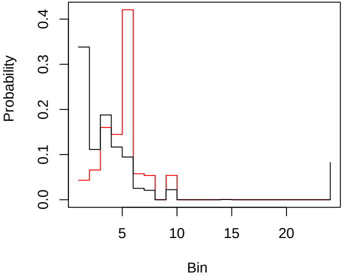
Exploring Model space: pain example

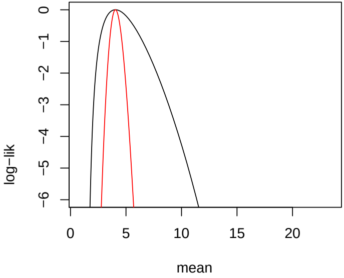
- In this example a GLM approach is often taken in practice
- This selects a two dimensional sufficient statistic and corresponding exponential family
  - Gamma family
  - Log-Normal family
  - Moment Constrained Maximum Entropy family
- Fixes a mean-variance relationship through specification of the variance function $\phi V(\mu)$
- Often using the plug-in, $\hat{\phi}$, for marginal inference

Exploring Model space: pain example

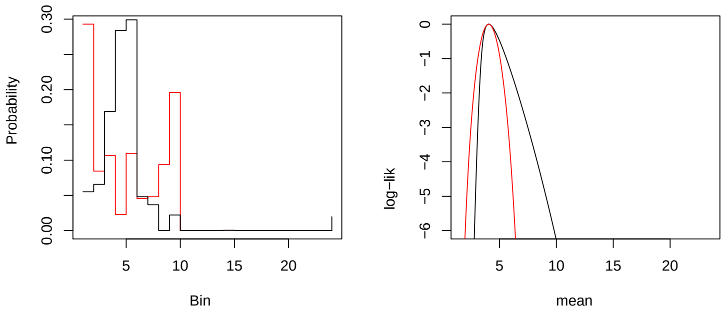
- In this example a GLM approach is often taken in practice
- This selects a two dimensional sufficient statistic and corresponding exponential family
  - Gamma family
  - Log-Normal family
  - Moment Constrained Maximum Entropy family
- Fixes a mean-variance relationship through specification of the variance function $\phi V(\mu)$
- Often using the plug-in, $\hat{\phi}$, for marginal inference
- Our geometric tools allow us to explore the sensitivity to these kinds of assumptions

Range of Inference





Range of Inference



Summary

- Computational Information Geometry
- The overall objective is to construct diagnostic tools to help understand sensitivity to model choice
- Targeted at applications where Generalised Linear Models are used
- SEM geometry is affine and convex, not manifold based: non-constant support and moment structure
- Topology defined via duality