Computational Information Geometry:

Geometry of Model Choice

Paul Marriott

University of Waterloo

April 6, 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

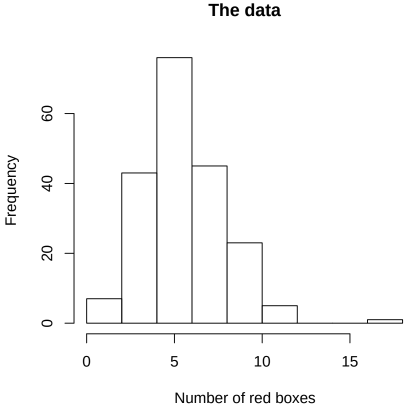
Big Picture

- 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
- Joint work with Karim Anaya-Izquierdo, Frank Critchley and Paul Vos

Big Picture

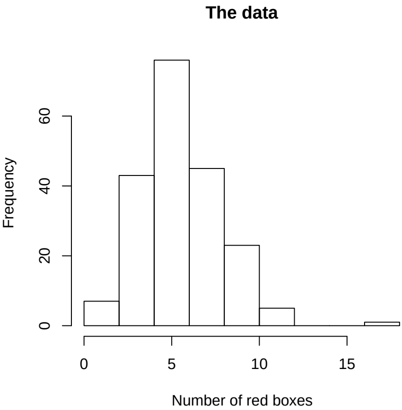
- Introduction to Computational Information Geometry
- The overall objective is to construct diagnostic tools to help understand sensitivity to model choice
- Targeted at applications where Generalised Linear Models are used
- Joint work with Karim Anaya-Izquierdo, Frank Critchley and Paul Vos
- Thanks to EPSRC Grant Number EP/E017878/1

Problem of Interest



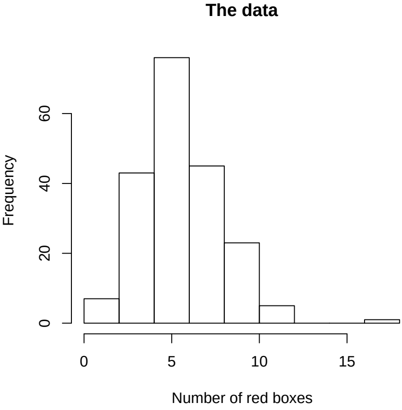
- Question: what is the population mean?
- How do modelling assumptions effect inference about mean?
- Can geometry of ‘space of all models’ give a framework for discussion?

Problem of Interest



- Question: what is the population mean?
- How do modelling assumptions effect 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 effect 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
  - neighbourhood structures

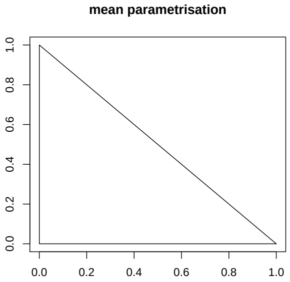
Structured Extended Multinomials

- 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
  - neighbourhood structures
- Structured Extended Multinomials (SEM) include this structure

Structured Extended Multinomials

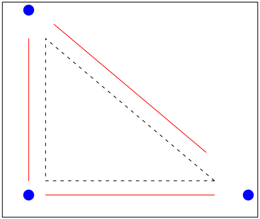
- Extended multinomials are multinomial but allow cell probabilities to be zero.
- Discretizing continuous data gives categorical models with structure on the cells
- Examples of structure:
  - numerical labels
  - ordering
  - neighbourhood structures
- Structured Extended Multinomials (SEM) include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

-1 -simplical structure



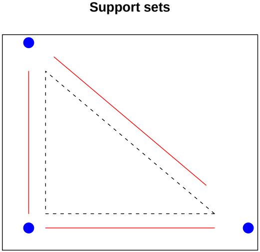
- Mean (-1) parameters can be on boundary
- Different support sets
- Union of exponential families each with corresponding natural $(+1)$ parameters

-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



- 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

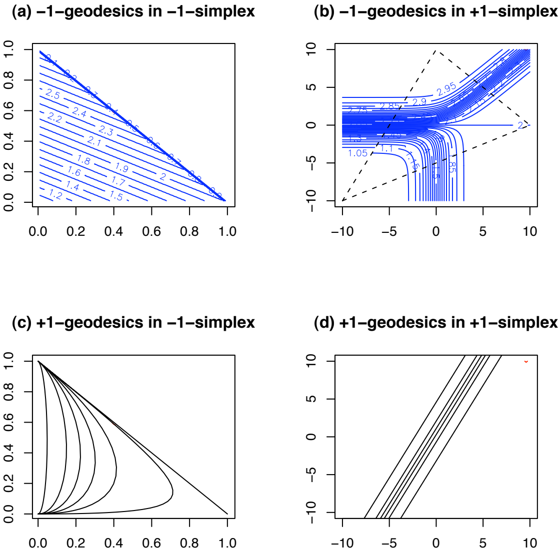
+1 simplex structure

- 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

Dual Parameterisations



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 not 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 not manifolds
- Almost all of the information geometry on SEM is numerically easy

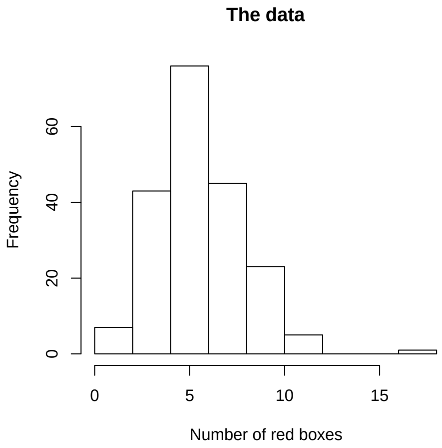
Computational Information Geometry

- 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 not 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 not manifolds
- Almost all of the information geometry on SEM is numerically easy
- Hard part: the mixed parameterisation
- This is our computational framework

The example



The example

- Data is 200 integers

The example

- Data is 200 integers
- Told each is total number of red squares out of 25 red or blue

The example

- Data is 200 integers
- Told each is total number of red squares out of 25 red or blue
- Told each comes from unrelated experiments

The example

- Data is 200 integers
- Told each is total number of red squares out of 25 red or blue
- Told each comes from unrelated experiments
- Statistician A: Binomial model as working problem formulation

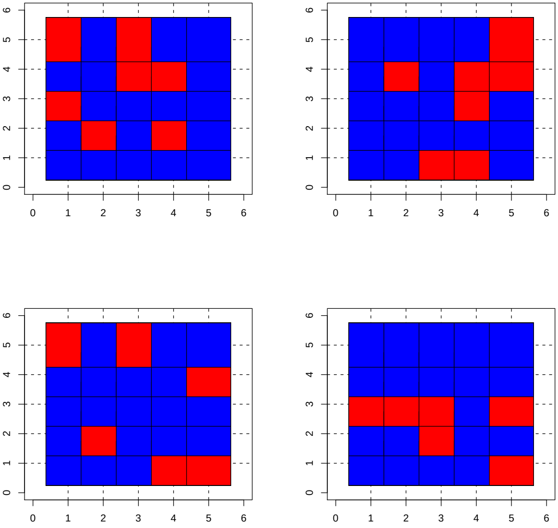
The example

- Data is 200 integers
- Told each is total number of red squares out of 25 red or blue
- Told each comes from unrelated experiments
- Statistician A: Binomial model as working problem formulation
- Passes Goodness of Fit tests but for ‘outlier’

The example

- Data is 200 integers
- Told each is total number of red squares out of 25 red or blue
- Told each comes from unrelated experiments
- Statistician A: Binomial model as working problem formulation
- Passes Goodness of Fit tests but for ‘outlier’
- Statistician B: Looks at ‘raw’ data and talks to scientists

The Example



The example

- Statistician B: uses theory which has equilibrium distribution from a spatial Markov model

The example

- Statistician B: uses theory which has equilibrium distribution from a spatial Markov model
- Statistician C: is non parametric

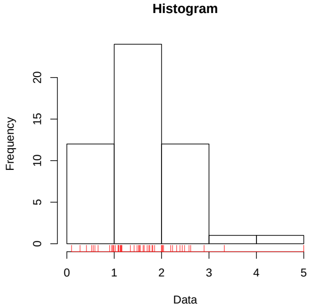
The example

- Statistician B: uses theory which has equilibrium distribution from a spatial Markov model
- Statistician C: is non parametric
- Statistician D: uses robust methods

The example

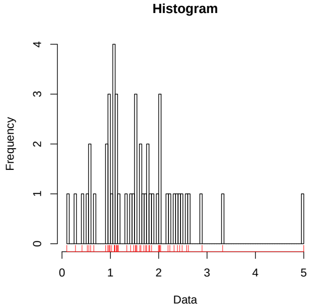
- Statistician B: uses theory which has equilibrium distribution from a spatial Markov model
- Statistician C: is non parametric
- Statistician D: uses robust methods
- Main Question: can the universality and geometry the SEM give a framework in which A, B, C and D can communicate their differing views about the mean number of red boxes?

Shape of likelihood in SEM



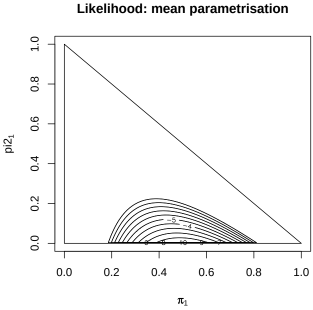
- Inference on population mean
- Working in high dimensional multinomial with many zero counts
- Likelihood in mean parameters, $\mu = N\pi$
- Likelihood in natural parameters, η: no MLE
- Likelihood in natural parameters η: regular case

Shape of likelihood in SEM



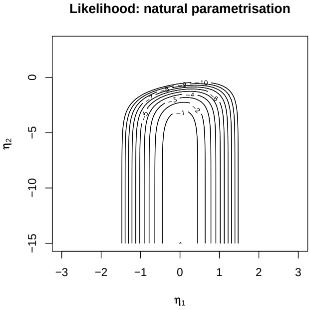
- Inference on population mean
- Working in high dimensional multinomial with many zero counts
- Likelihood in mean parameters, $\mu = N\pi$
- Likelihood in natural parameters, η: no MLE
- Likelihood in natural parameters η: regular case

Shape of likelihood in SEM



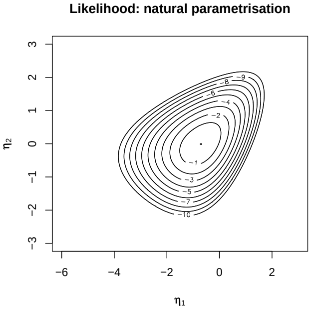
- Inference on population mean
- Working in high dimensional multinomial with many zero counts
- Likelihood in mean parameters, $\mu = N\pi$
- Likelihood in natural parameters, η: no MLE
- Likelihood in natural parameters η: regular case

Shape of likelihood in SEM



- Inference on population mean
- Working in high dimensional multinomial with many zero counts
- Likelihood in mean parameters, $\mu = N\pi$
- Likelihood in natural parameters, η: no MLE
- Likelihood in natural parameters η: regular case

Shape of likelihood in SEM



- Inference on population mean
- Working in high dimensional multinomial with many zero counts
- Likelihood in mean parameters, $\mu = N\pi$
- Likelihood in natural parameters, η: no MLE
- Likelihood in natural parameters η: regular case

Region of interest

- The universality of the SEM is too rich to be the desired framework for communication

Region of interest

- The universality of the SEM is too rich to be the desired framework for communication
- Only want to look at models which are data-supported

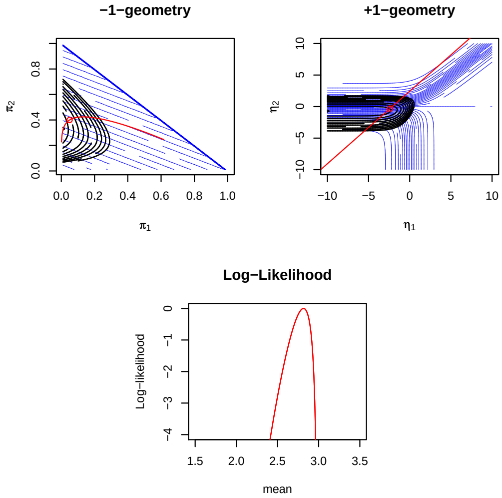
Region of interest

- The universality of the SEM is too rich to be the desired framework for communication
- Only want to look at models which are data-supported
- There are many types of goodness-of-fit tests on simplex

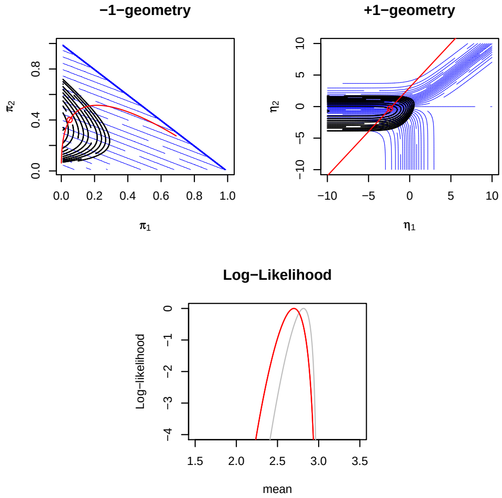
Region of interest

- The universality of the SEM is too rich to be the desired framework for communication
- Only want to look at models which are data-supported
- There are many types of goodness-of-fit tests on simplex
- Such tests are necessary but not sufficient $\cdots$

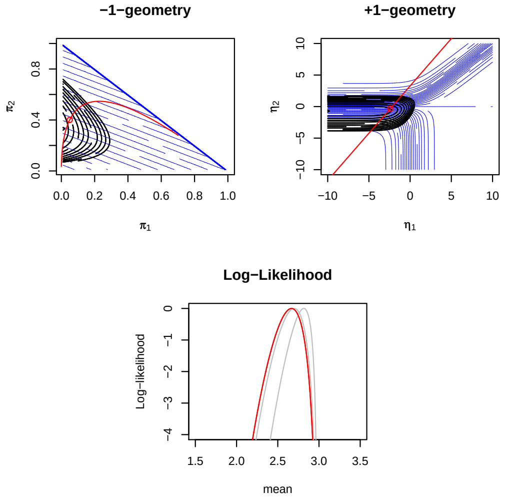
Thought experiment



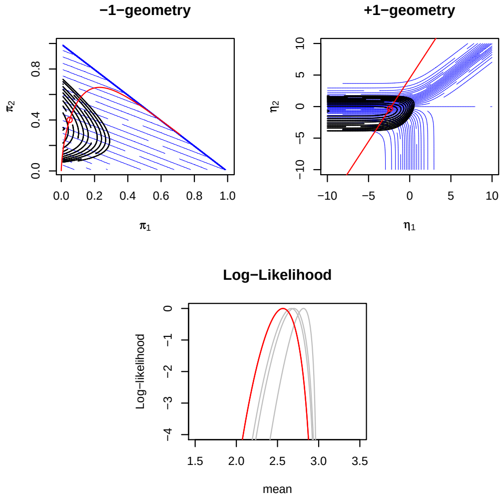
Thought experiment



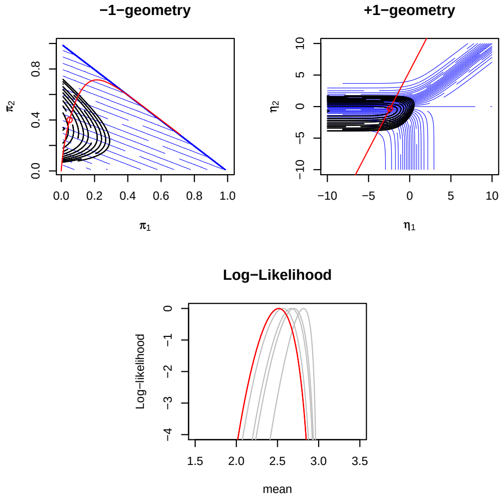
Thought experiment



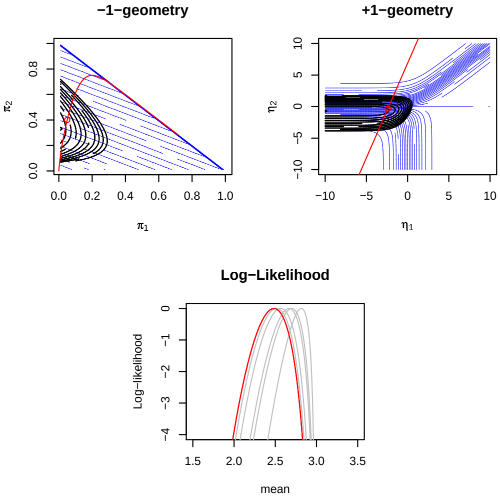
Thought experiment



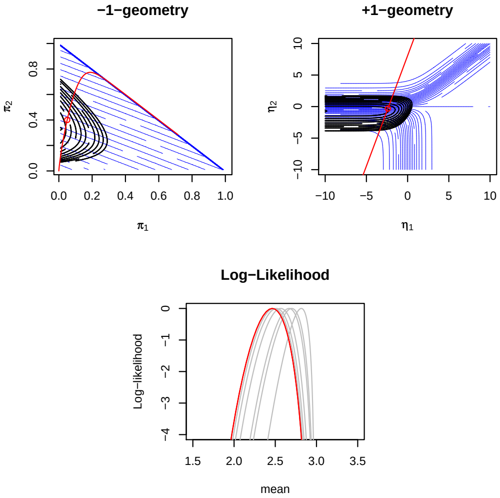
Thought experiment



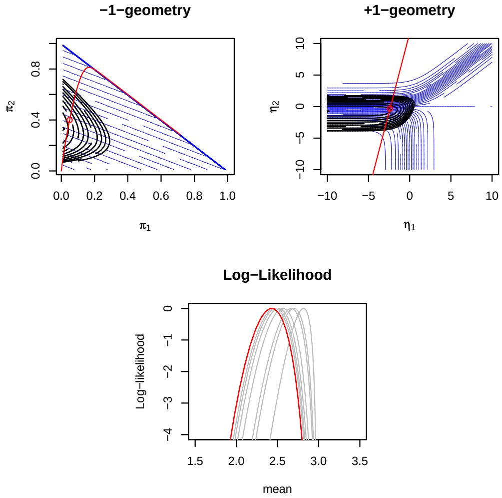
Thought experiment



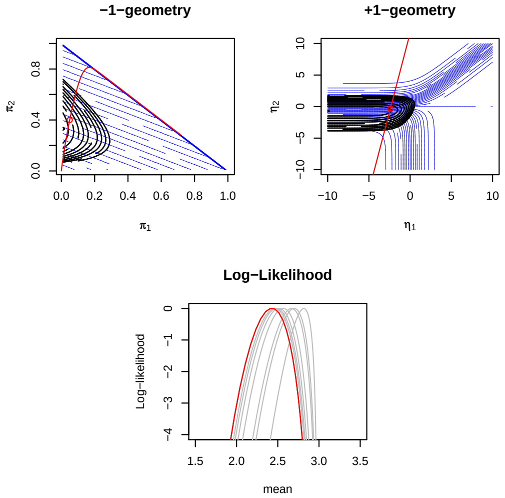
Thought experiment



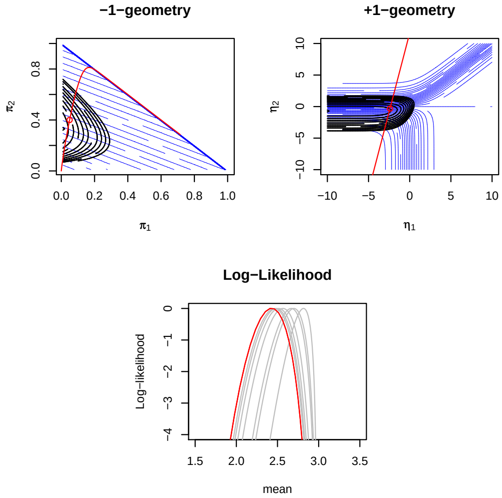
Thought experiment



Thought experiment



Thought experiment



Least informative model

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

Least informative model

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

Least informative model

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

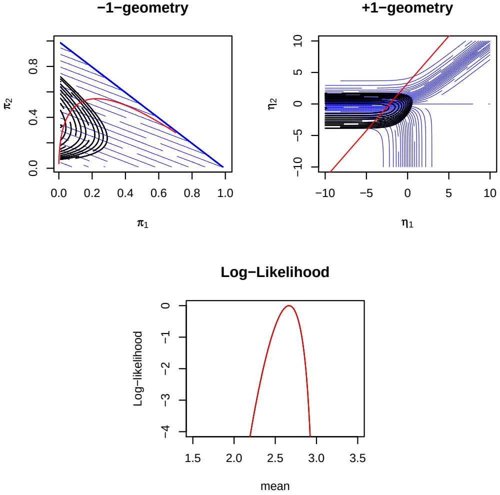
Least informative model

- Since rotation is through data generation process (DGP) can't use goodness-of-fit tests to distinguish between models
- 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
- Models with this orthogonality property we call *least informative models*

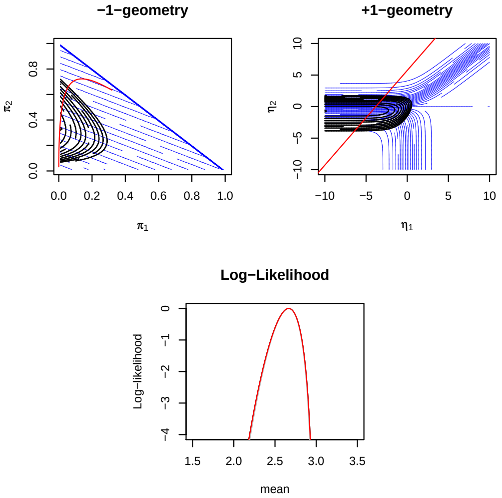
Least informative model

- Since rotation is through data generation process (DGP) can't use goodness-of-fit tests to distinguish between models
- 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
- Models with this orthogonality property we call *least informative models*
- Information in inference comes from two sources: (i) data and (ii) modelling assumptions. To be conservative minimise (ii) relative to (i)

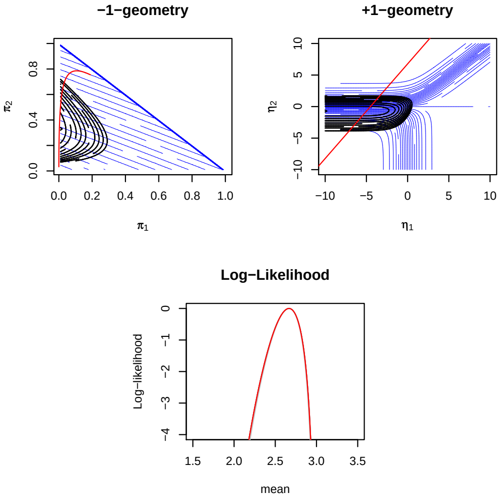
Effect of translation



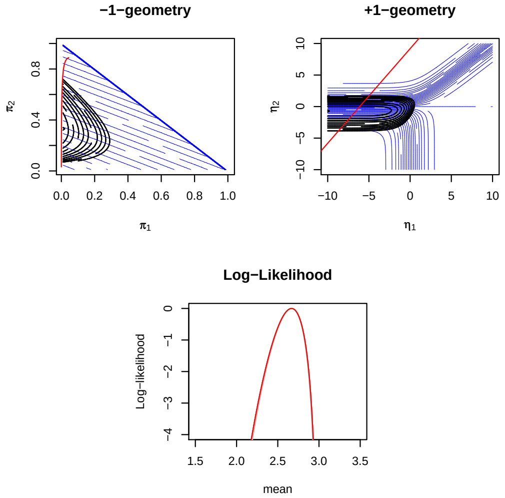
Effect of translation



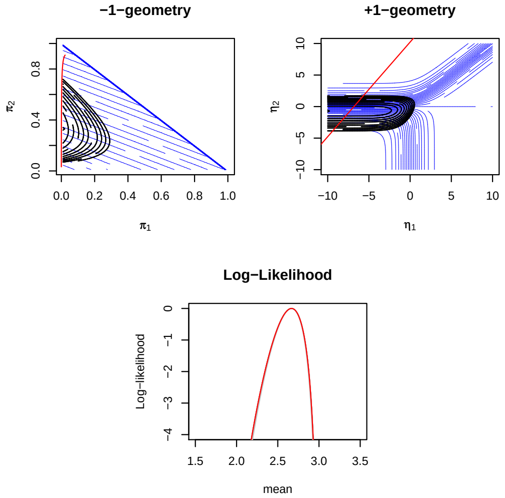
Effect of translation



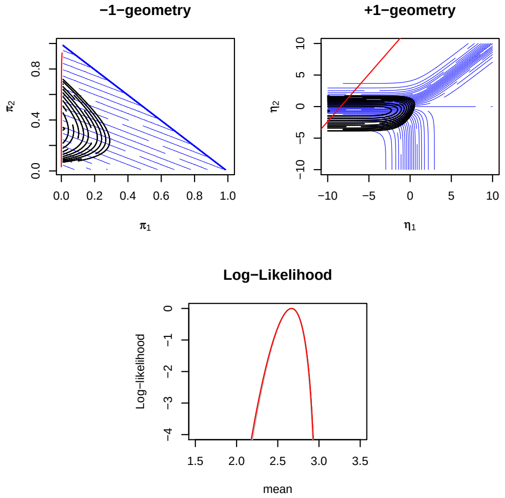
Effect of translation



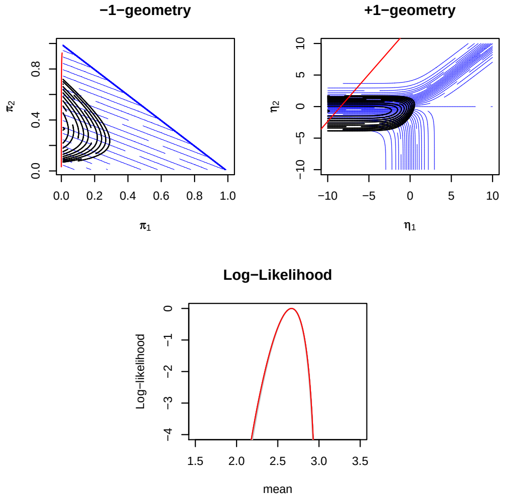
Effect of translation



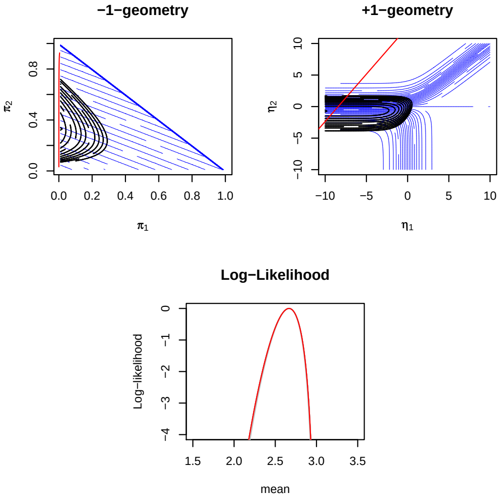
Effect of translation



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 Profile 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 Profile 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 Profile Likelihood
- Shows link between least informative parametric inference and non-parametric inference
- So SEM captures both Statistician A and C views

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 Profile Likelihood
- Shows link between least informative parametric inference and non-parametric inference
- So SEM captures both Statistician A and C views
- 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 Profile Likelihood
- Shows link between least informative parametric inference and non-parametric inference
- So SEM captures both Statistician A and C views
- The number of perturbations which matter for inference about the mean can be surprising small
- We can compute these directions: see Karim's talk on approximate cuts

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 Profile Likelihood
- Shows link between least informative parametric inference and non-parametric inference
- So SEM captures both Statistician A and C views
- The number of perturbations which matter for inference about the mean can be surprising small
- We can compute these directions: see Karim's talk on approximate cuts

Information from Model

- Parametric Models: (A) Binomial (B) local Markov model

Information from Model

- Parametric Models: (A) Binomial (B) local Markov model
- Both data consistent

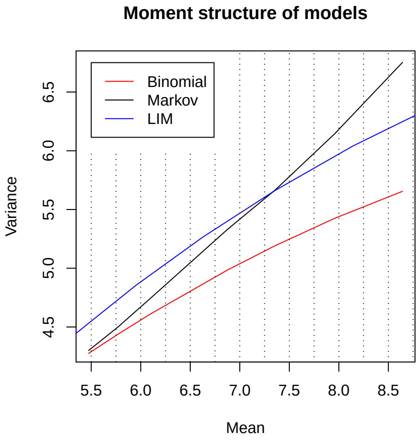
Information from Model

- Parametric Models: (A) Binomial (B) local Markov model
- Both data consistent
- Binomial is a least informative model

Information from Model

- Parametric Models: (A) Binomial (B) local Markov model
- Both data consistent
- Binomial is a least informative model
- Local Markov model is not $\cdots$

Information from Model



Information from Model

- The information from the model is captured using the mixed parametrisation in the universal SEM

Information from Model

- The information from the model is captured using the mixed parametrisation in the universal SEM
- Look at angle between parametric model and level set of mean in SEM

Information from Model

- The information from the model is captured using the mixed parametrisation in the universal SEM
- Look at angle between parametric model and level set of mean in SEM
- The smaller the angle the larger the *model* information about parameter of interest

Information from Model

- The information from the model is captured using the mixed parametrisation in the universal SEM
- Look at angle between parametric model and level set of mean in SEM
- The smaller the angle the larger the *model* information about parameter of interest
- If the models assumptions are correct increase information

Information from Model

- The information from the model is captured using the mixed parametrisation in the universal SEM
- Look at angle between parametric model and level set of mean in SEM
- The smaller the angle the larger the *model* information about parameter of interest
- If the models assumptions are correct increase information
- Errors in model assumptions generate bias

Information from Model

- The information from the model is captured using the mixed parametrisation in the universal SEM
- Look at angle between parametric model and level set of mean in SEM
- The smaller the angle the larger the *model* information about parameter of interest
- If the models assumptions are correct increase information
- Errors in model assumptions generate bias
- The set of sensitive directions defines a framework in which the four Statisticians can communicate, see Karim's talk

Local to Global

- The basic geometries of SEM are affine and convex, rather than differential geometric

Local to Global

- The basic geometries of SEM are affine and convex, rather than differential geometric
- Topology allows limits on boundaries to be taken

Local to Global

- The basic geometries of SEM are affine and convex, rather than differential geometric
- Topology allows limits on boundaries to be taken
- Affine geometry allows downweight/delete outlier c.f.Statistician D

Local to Global

- The basic geometries of SEM are affine and convex, rather than differential geometric
- Topology allows limits on boundaries to be taken
- Affine geometry allows downweight/delete outlier c.f.Statistician D
- Affine geometry allows local and global analysis

Computation Information Geometry

- 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