CIG: model sensitivity and approximate cuts

(CIG := Computational Information Geometry)

Frank Critchley (OU)

Joint work with Karim Anaya-Izquierdo, Paul Marriott and Paul Vos

IGAIA3, Leipzig, August 2010

- thanks!
- *work-in-progress*
- focus = ideas/pictures

- thanks!
- *work-in-progress*
- focus = ideas/pictures - not details/formulae

- Overall objective:
  provide tools to help understand *sensitivity to model choice*

- Overall objective:
  provide tools to help understand *sensitivity to model choice*
- Target:
  applications of Generalised Linear Models

- Overall objective:
  provide tools to help understand *sensitivity to model choice*
- Target:
  applications of Generalised Linear Models
- Delivered via ...
  Computational Information Geometry

- Overall objective:
  provide tools to help understand *sensitivity to model choice*
- Target:
  applications of Generalised Linear Models
- Delivered via ...
  Computational Information Geometry

- **Overall objective:**
  provide tools to help understand *sensitivity to model choice*
- **Target:**
  applications of Generalised Linear Models
- **Delivered via ...**
  Computational Information Geometry
  (hidden from the user; available in R)

- Overall objective:
  provide tools to help understand *sensitivity to model choice*
- Target:
  applications of Generalised Linear Models
- Delivered via ...
  Computational Information Geometry
  (hidden from the user; available in R)
- Geometric features:

- Overall objective:
  provide tools to help understand *sensitivity to model choice*
- Target:
  applications of Generalised Linear Models
- Delivered via ...
  Computational Information Geometry
  (hidden from the user; available in R)
- Geometric features:
  1. discretisation gives *operational* proxy to universal space

- Overall objective:
  provide tools to help understand *sensitivity to model choice*
- Target:
  applications of Generalised Linear Models
- Delivered via ...
  Computational Information Geometry
  (hidden from the user; available in R)
- Geometric features:
  1. discretisation gives *operational* proxy to universal space
  2. our SEM geometry is affine and convex, *not* manifold based

- Overall objective:
  provide tools to help understand *sensitivity to model choice*
- Target:
  applications of Generalised Linear Models
- Delivered via ...
  Computational Information Geometry
  (hidden from the user; available in R)
- Geometric features:
  1. discretisation gives *operational* proxy to universal space
  2. our SEM geometry is affine and convex, *not* manifold based

- Overall objective:
  provide tools to help understand *sensitivity to model choice*
- Target:
  applications of Generalised Linear Models
- Delivered via ...
  Computational Information Geometry
  (hidden from the user; available in R)
- Geometric features:
  1. discretisation gives *operational* proxy to universal space
  2. our SEM geometry is affine and convex, *not* manifold based (non-constant: support (dimension) and moment structure)

- Overall objective:
  provide tools to help understand *sensitivity to model choice*
- Target:
  applications of Generalised Linear Models
- Delivered via ...
  Computational Information Geometry
  (hidden from the user; available in R)
- Geometric features:
  1. discretisation gives *operational* proxy to universal space
  2. our SEM geometry is affine and convex, *not* manifold based (non-constant: support (dimension) and moment structure)
  3. topology defined via duality

- Overall objective:
  provide tools to help understand *sensitivity to model choice*
- Target:
  applications of Generalised Linear Models
- Delivered via ...
  Computational Information Geometry
  (hidden from the user; available in R)
- Geometric features:
  1. discretisation gives *operational* proxy to universal space
  2. our SEM geometry is affine and convex, *not* manifold based (non-constant: support (dimension) and moment structure)
  3. topology defined via duality

- Overall objective:
  provide tools to help understand *sensitivity to model choice*
- Target:
  applications of Generalised Linear Models
- Delivered via ...
  Computational Information Geometry
  (hidden from the user; available in R)
- Geometric features:
  1. discretisation gives *operational* proxy to universal space
  2. our SEM geometry is affine and convex, *not* manifold based (non-constant: support (dimension) and moment structure)
  3. topology defined via duality

Work supported by EPSRC grant EP/E017878/1

- One iteration of Statistical Science:

  WPF = (Q, p.c., model, data, inference) $\Rightarrow$ A

- One iteration of Statistical Science:

  WPF = (Q, p.c., model, data, inference) $\Rightarrow$ A

- *perturbations of problem formulation are pertinent*, thus ...

- One iteration of Statistical Science:

  WPF = (Q, p.c., model, data, inference) $\Rightarrow$ A

- *perturbations of problem formulation are pertinent*, thus ...
- *sensitivity analyses are sensible*, three motivations being ...

- One iteration of Statistical Science:

  WPF = (Q, p.c., model, data, inference) $\Rightarrow$ A

- *perturbations of problem formulation are pertinent*, thus ...
- *sensitivity analyses are sensible*, three motivations being ...
  1. stability: small perturbation $\Rightarrow$ small effect

- One iteration of Statistical Science:

  WPF = (Q, p.c., model, data, inference) $\Rightarrow$ A

- *perturbations of problem formulation are pertinent*, thus ...
- *sensitivity analyses are sensible*, three motivations being ...
  1. stability: small perturbation $\Rightarrow$ small effect
  2. robustness: big perturbation $\Rightarrow$ small effect

- One iteration of Statistical Science:

  WPF = (Q, p.c., model, data, inference) $\Rightarrow$ A

- *perturbations of problem formulation are pertinent*, thus ...
- *sensitivity analyses are sensible*, three motivations being ...
  1. stability: small perturbation $\Rightarrow$ small effect
  2. robustness: big perturbation $\Rightarrow$ small effect
  3. warning!: small perturbation $\Rightarrow$ big effect! ... *due to* ...

- One iteration of Statistical Science:

  WPF = (Q, p.c., model, data, inference) $\Rightarrow$ A

- *perturbations of problem formulation are pertinent*, thus ...
- *sensitivity analyses are sensible*, three motivations being ...
  1. stability: small perturbation $\Rightarrow$ small effect
  2. robustness: big perturbation $\Rightarrow$ small effect
  3. warning!: small perturbation $\Rightarrow$ big effect! ... *due to* ...
- wrong data generation process:
  $\varepsilon \neq 0$ in $F = (1-\varepsilon)F_\theta + \varepsilon G$ $\Rightarrow$ need data-*robustness*

- One iteration of Statistical Science:

  WPF = (Q, p.c., model, data, inference) $\Rightarrow$ A

- *perturbations of problem formulation are pertinent*, thus ...
- *sensitivity analyses are sensible*, three motivations being ...

  1. stability: small perturbation $\Rightarrow$ small effect
  2. robustness: big perturbation $\Rightarrow$ small effect
  3. warning!: small perturbation $\Rightarrow$ big effect! ... *due to* ...

- wrong data generation process:
  $\varepsilon \neq 0$ in $F = (1 - \varepsilon)F_\theta + \varepsilon G \Rightarrow$ need data-*robustness*
- wrong model:
  F_θ misspecified in $F = F_\theta \Rightarrow$ need model-*robustness*

- One iteration of Statistical Science:

  WPF = (Q, p.c., model, data, inference) $\Rightarrow$ A

- *perturbations of problem formulation are pertinent*, thus ...
- *sensitivity analyses are sensible*, three motivations being ...
  1. stability: small perturbation $\Rightarrow$ small effect
  2. robustness: big perturbation $\Rightarrow$ small effect
  3. warning!: small perturbation $\Rightarrow$ big effect! ... *due to* ...
- wrong data generation process:
  $\varepsilon \neq 0$ in $F = (1-\varepsilon)F_\theta + \varepsilon G$ $\Rightarrow$ need data-*robustness*
- wrong model:
  F_θ misspecified in $F = F_\theta$ $\Rightarrow$ need model-*robustness*
- or both!
  $\Rightarrow$ need full-blown *Computational Information Geometry*

WPF = (Q, p.c., model, data, inference) $\Rightarrow$ A

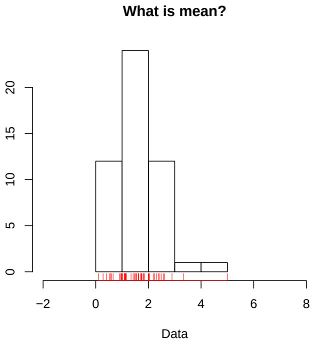
WPF = (Q, p.c., model, data, inference) $\Rightarrow$ A

KEY POINT:

Q takes the form: 'what is $\theta_Q \equiv \theta_Q[F]$?',

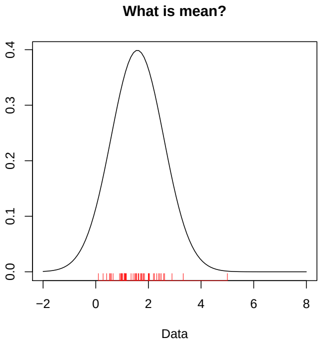
so that θ_Q has same (= population) meaning in all models

What is the population mean?



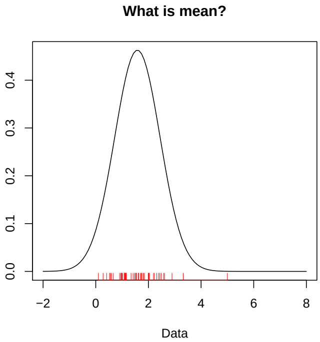
- Data
- Model $N(\mu, 1)$
- Model $N(\mu, \sigma^2)$
- Model log-Normal(μ', σ^2)
- Non-parametric
- Model $N(\mu, 1)$ remove outlier

What is the population mean?



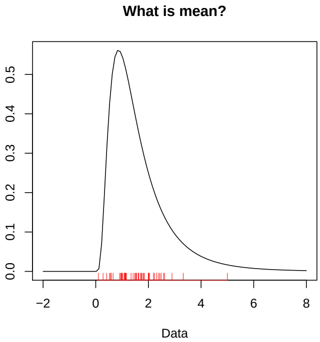
- Data
- Model $N(\mu, 1)$
- Model $N(\mu, \sigma^2)$
- Model log-Normal(μ', σ^2)
- Non-parametric
- Model $N(\mu, 1)$ remove outlier

What is the population mean?



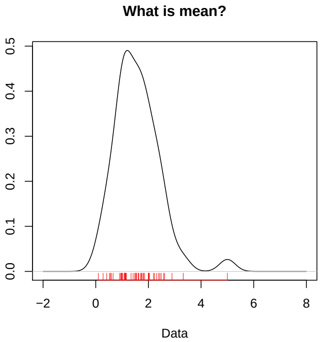
- Data
- Model $N(\mu, 1)$
- Model $N(\mu, \sigma^2)$
- Model log-Normal(μ', σ^2)
- Non-parametric
- Model $N(\mu, 1)$ remove outlier

What is the population mean?



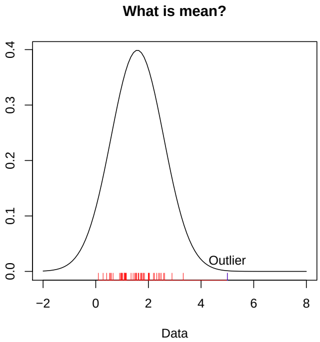
- Data
- Model $N(\mu, 1)$
- Model $N(\mu, \sigma^2)$
- Model log-Normal(μ', σ^2)
- Non-parametric
- Model $N(\mu, 1)$ remove outlier

What is the population mean?



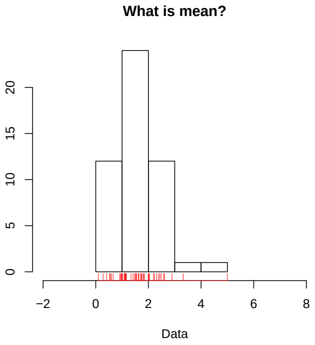
- Data
- Model $N(\mu, 1)$
- Model $N(\mu, \sigma^2)$
- Model log-Normal(μ', σ^2)
- Non-parametric
- Model $N(\mu, 1)$ remove outlier

What is the population mean?



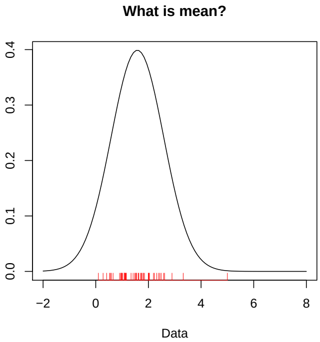
- Data
- Model $N(\mu, 1)$
- Model $N(\mu, \sigma^2)$
- Model log-Normal(μ', σ^2)
- Non-parametric
- Model $N(\mu, 1)$ remove outlier

What is the population mean?



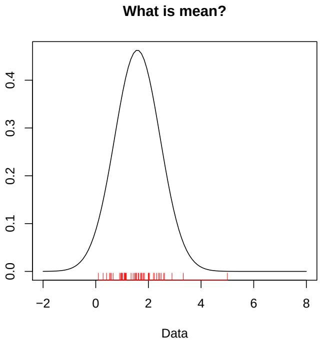
- Data
- Parameter of interest is sole parameter in model
- Extra ‘nuisance’ parameters increase flexibility of model
- Different parametric family
- No (or lots of) parameters
- Robustness/Diagnostics

What is the population mean?



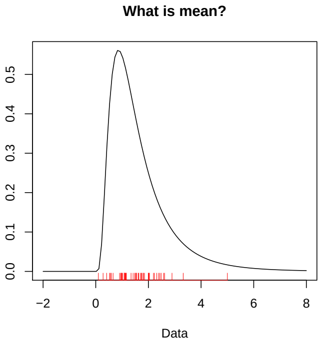
- Data
- Parameter of interest is sole parameter in model
- Extra 'nuisance' parameters increase flexibility of model
- Different parametric family
- No (or lots of) parameters
- Robustness/Diagnostics

What is the population mean?



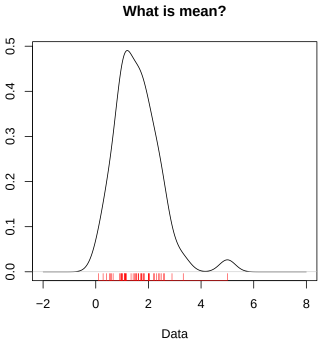
- Data
- Parameter of interest
  is sole parameter in model
- Extra ‘nuisance’ parameters
  increase flexibility of model
- Different parametric family
- No (or lots of) parameters
- Robustness/Diagnostics

What is the population mean?



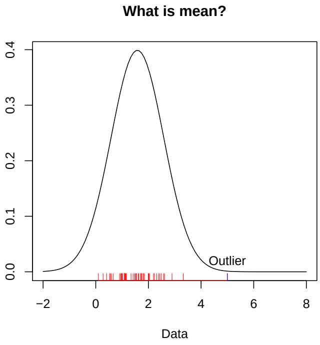
- Data
- Parameter of interest is sole parameter in model
- Extra ‘nuisance’ parameters increase flexibility of model
- Different parametric family
- No (or lots of) parameters
- Robustness/Diagnostics

What is the population mean?



- Data
- Parameter of interest is sole parameter in model
- Extra ‘nuisance’ parameters increase flexibility of model
- Different parametric family
- No (or lots of) parameters
- Robustness/Diagnostics

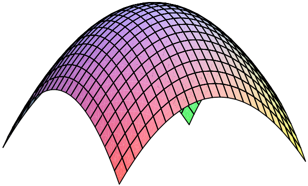
What is the population mean?



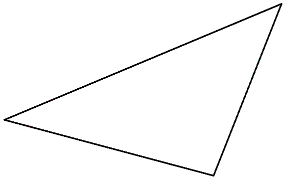
- Data
- Parameter of interest
  is sole parameter in model
- Extra ‘nuisance’ parameters
  increase flexibility of model
- Different parametric family
- No (or lots of) parameters
- Robustness/Diagnostics

What is the underlying Geometry?

- Manifolds

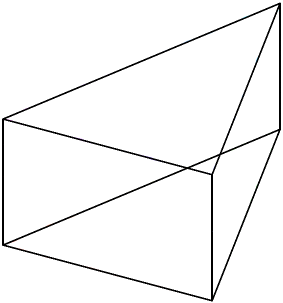


What is the underlying Geometry?



- Manifolds
- Random sample: topologically a simplex

What is the underlying Geometry?



- Manifolds
- Random sample: topologically a simplex
- Two independent samples: topologically a polytope

Distinctive features

- with a choice of cells, can discretise data

Distinctive features

- with a choice of cells, can discretise data
- after discretising, multinomials are universal models

Distinctive features

- with a choice of cells, can discretise data
- after discretising, multinomials are universal models
- do not assume a common support for probability measures

Distinctive features

- with a choice of cells, can discretise data
- after discretising, multinomials are universal models
- do not assume a common support for probability measures
  - *i.e.* allow cell probabilities to be zero

Distinctive features

- with a choice of cells, can discretise data
- after discretising, multinomials are universal models
- do not assume a common support for probability measures
  - *i.e.* allow cell probabilities to be zero
  - so, working in *extended* multinomial models

Distinctive features

- with a choice of cells, can discretise data
- after discretising, multinomials are universal models
- do not assume a common support for probability measures
  - *i.e.* allow cell probabilities to be zero
  - so, working in *extended* multinomial models
- resulting methodology is ...

Distinctive features

- with a choice of cells, can discretise data
- after discretising, multinomials are universal models
- do not assume a common support for probability measures
  - *i.e.* allow cell probabilities to be zero
  - so, working in *extended* multinomial models
- resulting methodology is ...
  - geometrically and computationally tractable

Distinctive features

- with a choice of cells, can discretise data
- after discretising, multinomials are universal models
- do not assume a common support for probability measures
  - *i.e.* allow cell probabilities to be zero
  - so, working in *extended* multinomial models
- resulting methodology is ...
  - geometrically and computationally tractable
  - approximate, not asymptotic

Distinctive features

- with a choice of cells, can discretise data
- after discretising, multinomials are universal models
- do not assume a common support for probability measures
  - *i.e.* allow cell probabilities to be zero
  - so, working in *extended* multinomial models
- resulting methodology is ...
  - geometrically and computationally tractable
  - approximate, not asymptotic
  - seamlessly integrated from local-to-global (affine geometry)

Points, curves and surfaces

- after discretisation to $(k+1)$ cells,
  a probability distribution =
  a point $\pi \equiv (\pi_r)$ in the (probability) *simplex* $\mathbb{P}^{k+1}$

Points, curves and surfaces

- after discretisation to $(k+1)$ cells,
  a probability distribution =
  a point $\pi \equiv (\pi_r)$ in the (probability) *simplex* $\mathbb{P}^{k+1}$
- a 1-parameter model is, then, a curve in $\mathbb{P}^{k+1}$

Points, curves and surfaces

- after discretisation to $(k+1)$ cells,
  a probability distribution =
  a point $\pi \equiv (\pi_r)$ in the (probability) *simplex* $\mathbb{P}^{k+1}$
- a 1-parameter model is, then, a curve in $\mathbb{P}^{k+1}$
- a p-parameter model is a p-dimensional surface in $\mathbb{P}^{k+1}$

Mixture and exponential families; Fisher information

- the mean $(\mu_r) \equiv (n\pi_r)$ and natural (η_r) parameterisations provide *affine coordinates* for, respectively, ...

Mixture and exponential families; Fisher information

- the mean $(\mu_r) \equiv (n\pi_r)$ and natural (η_r) parameterisations provide *affine coordinates* for, respectively, ...
- the -1 and $+1$ *affine geometries* on $\mathbb{P}^{k+1}$, in which ...

Mixture and exponential families; Fisher information

- the mean $(\mu_r) \equiv (n\pi_r)$ and natural (η_r) parameterisations provide *affine coordinates* for, respectively, ...
- the -1 and $+1$ *affine geometries* on $\mathbb{P}^{k+1}$, in which ...
- *geodesics*:
  'going straight between two distributions' corresponds to:–

Mixture and exponential families; Fisher information

- the mean $(\mu_r) \equiv (n\pi_r)$ and natural (η_r) parameterisations provide *affine coordinates* for, respectively, ...
- the -1 and $+1$ *affine geometries* on $\mathbb{P}^{k+1}$, in which ...
- *geodesics*:
  'going straight between two distributions' corresponds to:–
  - -1: convex combinations of $\mu \leftrightarrow$ mix in the density

    and, for any given support,

Mixture and exponential families; Fisher information

- the mean $(\mu_r) \equiv (n\pi_r)$ and natural (η_r) parameterisations provide *affine coordinates* for, respectively, ...
- the -1 and $+1$ *affine geometries* on $\mathbb{P}^{k+1}$, in which ...
- *geodesics*:
  'going straight between two distributions' corresponds to:–
  - -1: convex combinations of $\mu \leftrightarrow$ mix in the density

    and, for any given support,
  - $+1$: convex combinations of $\eta \leftrightarrow$ mix in the exponent

Mixture and exponential families; Fisher information

- the mean $(\mu_r) \equiv (n\pi_r)$ and natural (η_r) parameterisations

  provide *affine coordinates* for, respectively, ...
- the -1 and $+1$ *affine geometries* on $\mathbb{P}^{k+1}$, in which ...
- *geodesics*:
  'going straight between two distributions' corresponds to:–
  - -1: convex combinations of $\mu \leftrightarrow$ mix in the density

    and, for any given support,
  - $+1$: convex combinations of $\eta \leftrightarrow$ mix in the exponent
- lengths and angles in $\mathbb{P}^{k+1}$ – in particular, *orthogonality* – can be measured via the Fisher information matrix (0-geometry)

Duality

Sufficient statistic: $\overline{n}^T := (n_0, n_1, ..., n_k)$
Random variable: $S := a^T \overline{n} \ (a \neq 0)$
Interest parameter: $\theta = \mathbb{E}(S)$

- *the* $+1$-*geodesics* defined by the 1-dimensional exponential families with sufficient statistics S

Duality

Sufficient statistic: $\overline{n}^T := (n_0, n_1, ..., n_k)$
Random variable: $S := a^T \overline{n}$ $(a \neq 0)$
Interest parameter: $\theta = \mathbb{E}(S)$

- *the* $+1$-*geodesics* defined by the 1-dimensional exponential families with sufficient statistics S
- ... *are everywhere Fisher-orthogonal to* ...

Duality

Sufficient statistic: $\overline{n}^T := (n_0, n_1, ..., n_k)$
Random variable: $S := a^T \overline{n}$ $(a \neq 0)$
Interest parameter: $\theta = \mathbb{E}(S)$

- *the* $+1$-*geodesics* defined by the 1-dimensional exponential families with sufficient statistics S
- ... *are everywhere Fisher-orthogonal to* ...
- *the* -1-*geodesics* defined by the $(k-1)$-dimensional level sets of $\theta = \mathbb{E}(S)$

Mixed Parameterisation

- By duality, there is a *mixed parameterisation* $\begin{pmatrix} \theta \\ \eta \end{pmatrix}$ in which:

Mixed Parameterisation

- By duality, there is a *mixed parameterisation* $\begin{pmatrix} \theta \\ \eta \end{pmatrix}$ in which:
  - θ is -1-affine

Mixed Parameterisation

- By duality, there is a *mixed parameterisation* $\begin{pmatrix} \theta \\ \eta \end{pmatrix}$ in which:
  - θ is -1-affine
  - η is $+1$-affine

Mixed Parameterisation

- By duality, there is a *mixed parameterisation* $\begin{pmatrix} \theta \\ \eta \end{pmatrix}$ in which:
  - θ is -1-affine
  - η is $+1$-affine
  - the level sets of θ and η are everywhere Fisher-orthogonal

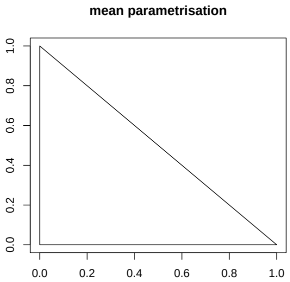
Mixed Parameterisation

- By duality, there is a *mixed parameterisation* $\begin{pmatrix} \theta \\ \eta \end{pmatrix}$ in which:
  - θ is -1-affine
  - η is $+1$-affine
  - the level sets of θ and η are everywhere Fisher-orthogonal
- In practice, this parameterisation can be computed (a key part of our computational information geometry)

Mixed Parameterisation

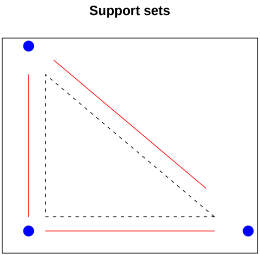
- By duality, there is a *mixed parameterisation* $\begin{pmatrix} \theta \\ \eta \end{pmatrix}$ in which:
  - θ is -1-affine
  - η is $+1$-affine
  - the level sets of θ and η are everywhere Fisher-orthogonal
- In practice, this parameterisation can be computed (a key part of our computational information geometry)
- [SKIP next 2 frames???]

Simplicial structure



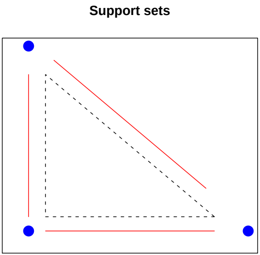
- Mean (-1) parameters can be on boundary
- Different support sets
- Union of exponential families each with corresponding natural $(+1)$ parameters

Simplicial structure



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

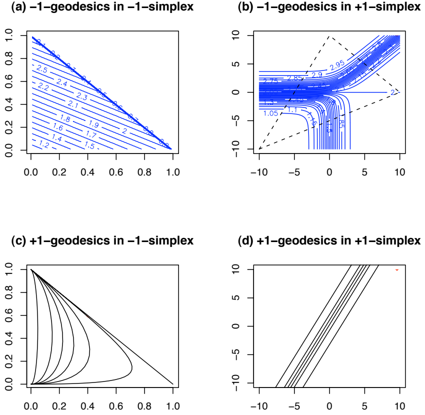
Simplicial structure



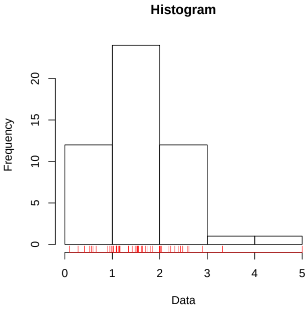
- Mean (-1) parameters can be on boundary
- Different support sets
- Union of exponential families each with corresponding natural $(+1)$ parameters

Dual Parameterisations

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

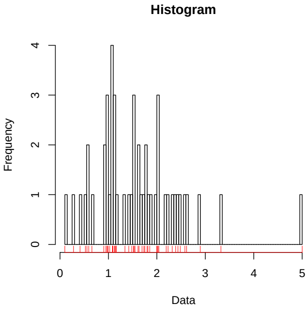


High-dimensional, sparse multinomials



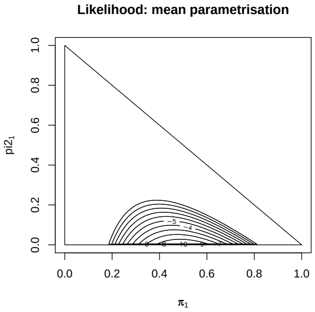
- Inference on population mean
- Working in high dimensional multinomial with many zero counts
- Likelihood in mean parameters, $\mu = N\pi$
- Likelihood in natural parameters, η: no MLE
- Likelihood in natural parameters η: regular case

High-dimensional, sparse multinomials



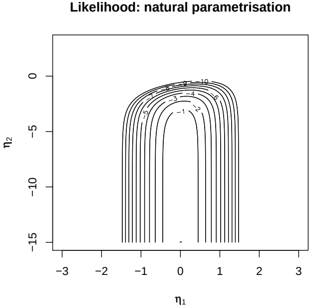
- Inference on population mean
- Working in high dimensional multinomial with many zero counts
- Likelihood in mean parameters, $\mu = N\pi$
- Likelihood in natural parameters, η: no MLE
- Likelihood in natural parameters η: regular case

High-dimensional, sparse multinomials



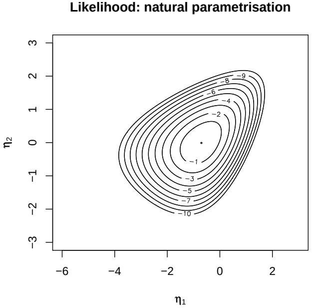
- Inference on population mean
- Working in high dimensional multinomial with many zero counts
- Likelihood in mean parameters, $\mu = N\pi$
- Likelihood in natural parameters, η: no MLE
- Likelihood in natural parameters η: regular case

High-dimensional, sparse multinomials



- 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

High-dimensional, sparse multinomials



- 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

Model sensitivity

- Key ingredients:

Model sensitivity

- Key ingredients:
  - $\theta \equiv \theta_Q$ quantity/parameter of interest

Model sensitivity

- Key ingredients:
  - $\theta \equiv \theta_Q$ quantity/parameter of interest
  - x data

Model sensitivity

- Key ingredients:
  - $\theta \equiv \theta_Q$ quantity/parameter of interest
  - x data
  - $f(x;\theta)$ base/working model

Model sensitivity

- Key ingredients:
  - $\theta \equiv \theta_Q$ quantity/parameter of interest
  - x data
  - $f(x;\theta)$ base/working model
  - inference about θ based on $S(x)$

Model sensitivity

- Key ingredients:
  - $\theta \equiv \theta_Q$ quantity/parameter of interest
  - x data
  - $f(x;\theta)$ base/working model
  - inference about θ based on $S(x)$
- *What is the effect on inference about θ when considering other models apart from $f(x;\theta)$?*

Model sensitivity

- Perturb base model $f(x;\theta)$

Model sensitivity

- Perturb base model $f(x;\theta)$
  - Pick a 'direction' η

Model sensitivity

- Perturb base model $f(x; \theta)$
  - Pick a 'direction' η
  - Perturb $f(x; \theta) \to g(x; \theta, \eta)$ such that:
    1. *θ keeps its (population) meaning*

Model sensitivity

- Perturb base model $f(x;\theta)$
  - Pick a 'direction' η
  - Perturb $f(x;\theta) \longrightarrow g(x;\theta,\eta)$ such that:
    1. θ *keeps its (population) meaning*
    2. $g(x;\theta,\eta_0) = f(x;\theta)$ for some η_0

Model sensitivity

- Perturb base model $f(x;\theta)$
  - Pick a 'direction' η
  - Perturb $f(x;\theta) \to g(x;\theta,\eta)$ such that:
    1. *θ keeps its (population) meaning*
    2. $g(x;\theta,\eta_0) = f(x;\theta)$ for some η_0
- Examples of 'directions' η

Model sensitivity

- Perturb base model $f(x;\theta)$
  - Pick a 'direction' η
  - Perturb $f(x;\theta) \longrightarrow g(x;\theta,\eta)$ such that:
    1. *θ keeps its (population) meaning*
    2. $g(x;\theta,\eta_0) = f(x;\theta)$ for some η_0
- Examples of 'directions' η
  - introduce extra variability (mixture model)

Model sensitivity

- Perturb base model $f(x;\theta)$
  - Pick a 'direction' η
  - Perturb $f(x;\theta) \to g(x;\theta,\eta)$ such that:
    1. θ *keeps its (population) meaning*
    2. $g(x;\theta,\eta_0) = f(x;\theta)$ for some η_0
- Examples of 'directions' η
  - introduce extra variability (mixture model)
  - introduce skewness

Model sensitivity

- Perturb base model $f(x;\theta)$
  - Pick a 'direction' η
  - Perturb $f(x;\theta) \to g(x;\theta,\eta)$ such that:
    1. θ *keeps its (population) meaning*
    2. $g(x;\theta,\eta_0) = f(x;\theta)$ for some η_0
- Examples of 'directions' η
  - introduce extra variability (mixture model)
  - introduce skewness
  - introduce heavier tails

Model sensitivity

- Perturb base model $f(x;\theta)$
  - Pick a 'direction' η
  - Perturb $f(x;\theta) \to g(x;\theta,\eta)$ such that:
    1. *θ keeps its (population) meaning*
    2. $g(x;\theta,\eta_0) = f(x;\theta)$ for some η_0
- Examples of 'directions' η
  - introduce extra variability (mixture model)
  - introduce skewness
  - introduce heavier tails
  - introduce dependence

Model sensitivity

- Perturb base model $f(x;\theta)$
  - Pick a 'direction' η
  - Perturb $f(x;\theta) \to g(x;\theta,\eta)$ such that:
    1. *θ keeps its (population) meaning*
    2. $g(x;\theta,\eta_0) = f(x;\theta)$ for some η_0
- Examples of 'directions' η
  - introduce extra variability (mixture model)
  - introduce skewness
  - introduce heavier tails
  - introduce dependence
- *What is the effect on inference about θ when perturbing in the direction η?*

Cuts

- Model $g(x;\theta,\eta)$, θ = interest and η = nuisance parameter

Cuts

- Model $g(x;\theta,\eta)$, θ = interest and η = nuisance parameter
- A cut is a statistic allowing θ-inference, independently of η

Cuts

- Model $g(x;\theta,\eta)$, θ = interest and η = nuisance parameter
- A cut is a statistic allowing θ-inference, independently of η
- Key idea: factorisation of the likelihood

Cuts

- Model $g(x;\theta,\eta)$, θ = interest and η = nuisance parameter
- A cut is a statistic allowing θ-inference, independently of η
- Key idea: factorisation of the likelihood
- If (a) $(S(x), U(x)) \overset{1-1}{\leftrightarrow} x$,

Cuts

- Model $g(x;\theta,\eta)$, θ = interest and η = nuisance parameter
- A cut is a statistic allowing θ-inference, independently of η
- Key idea: factorisation of the likelihood
- If (a) $(S(x), U(x)) \overset{1-1}{\leftrightarrow} x$, (b) $(\theta,\eta) \in \Theta \times \Lambda$

Cuts

- Model $g(x; \theta, \eta)$, θ = interest and η = nuisance parameter
- A cut is a statistic allowing θ-inference, independently of η
- Key idea: factorisation of the likelihood
- If (a) $(S(x), U(x)) \overset{1-1}{\leftrightarrow} x$, (b) $(\theta, \eta) \in \Theta \times \Lambda$ and (c):

$$\forall\, \theta, \eta, x, \;\; L(\theta, \eta; x) =$$

Cuts

- Model $g(x;\theta,\eta)$, θ = interest and η = nuisance parameter
- A cut is a statistic allowing θ-inference, independently of η
- Key idea: factorisation of the likelihood
- If (a) $(S(x), U(x)) \overset{1-1}{\leftrightarrow} x$, (b) $(\theta,\eta) \in \Theta \times \Lambda$ and (c):

$$\forall\, \theta,\eta,x, \quad L(\theta,\eta;x) = L_1(\theta;S(x)) \times$$

Cuts

- Model $g(x;\theta,\eta)$, θ = interest and η = nuisance parameter
- A cut is a statistic allowing θ-inference, independently of η
- Key idea: factorisation of the likelihood
- If (a) $(S(x),U(x)) \overset{1-1}{\leftrightarrow} x$, (b) $(\theta,\eta) \in \Theta \times \Lambda$ and (c):

$$\forall\,\theta,\eta,x,\quad L(\theta,\eta;x) = L_1(\theta;S(x)) \times L_2(\eta;U(x)),$$

Cuts

- Model $g(x;\theta,\eta)$, θ = interest and η = nuisance parameter
- A cut is a statistic allowing θ-inference, independently of η
- Key idea: factorisation of the likelihood
- If (a) $(S(x), U(x)) \overset{1-1}{\leftrightarrow} x$, (b) $(\theta, \eta) \in \Theta \times \Lambda$ and (c):

$$\forall\, \theta, \eta, x, \quad L(\theta, \eta; x) = L_1(\theta; S(x)) \times L_2(\eta; U(x)),$$

then: S is a cut (for θ)

Cuts

- Model $g(x;\theta,\eta)$, θ = interest and η = nuisance parameter
- A cut is a statistic allowing θ-inference, independently of η
- Key idea: factorisation of the likelihood
- If (a) $(S(x), U(x)) \overset{1-1}{\leftrightarrow} x$, (b) $(\theta,\eta) \in \Theta \times \Lambda$ and (c):

$$\forall\, \theta, \eta, x, \quad L(\theta,\eta;x) = L_1(\theta; S(x)) \times L_2(\eta; U(x)),$$

  then: S is a cut (for θ)
- Very tight definition valid in very special cases:

Cuts

- Model $g(x;\theta,\eta)$, θ = interest and η = nuisance parameter
- A cut is a statistic allowing θ-inference, independently of η
- Key idea: factorisation of the likelihood
- If (a) $(S(x), U(x)) \overset{1-1}{\leftrightarrow} x$, (b) $(\theta,\eta) \in \Theta \times \Lambda$ and (c):

$$\forall\, \theta, \eta, x, \quad L(\theta,\eta;x) = L_1(\theta; S(x)) \times L_2(\eta; U(x)),$$

  then: S is a cut (for θ)
- Very tight definition valid in very special cases:
  Natural Exponential Families (NEF)
  – with so-called mixed parameterisation

SEMs

- Use Structured Extended Multinomial models (SEMs)

SEMs

- Use Structured Extended Multinomial models (SEMs)
- Discretising continuous data gives multinomials, with structure on the cells

SEMs

- Use Structured Extended Multinomial models (SEMs)
- Discretising continuous data gives multinomials,
  with structure on the cells
- Extended multinomials ...
  are multinomials, which allow cell probabilities to be zero

SEMs

- Use Structured Extended Multinomial models (SEMs)
- Discretising continuous data gives multinomials,
  with structure on the cells
- Extended multinomials ...
  are multinomials, which allow cell probabilities to be zero
- SEMs $\Rightarrow$ a *proxy* for the universal space of all distributions

SEMs and Cuts

- data $\leftrightarrow$ $k+1$ cells with
  - labels $\{t_0, t_1, \ldots, t_k\}$ ... as well as ...

SEMs and Cuts

- data $\leftrightarrow k+1$ cells with
  - labels $\{t_0, t_1, \ldots, t_k\}$ ... as well as ...
  - counts $\{n_0, n_1, \ldots, n_k\}$ (summing to N)

SEMs and Cuts

- data $\leftrightarrow$ $k+1$ cells with
  - labels $\{t_0, t_1, \ldots, t_k\}$ ... as well as ...
  - counts $\{n_0, n_1, \ldots, n_k\}$ (summing to N)
- model $\leftrightarrow$ parameterised set of probs. $\{\pi_0, \ldots, \pi_k\}$

SEMs and Cuts

- data $\leftrightarrow$ $k+1$ cells with
  - labels $\{t_0, t_1, \ldots, t_k\}$ ... as well as ...
  - counts $\{n_0, n_1, \ldots, n_k\}$ (summing to N)
- model $\leftrightarrow$ parameterised set of probs. $\{\pi_0, \ldots, \pi_k\}$
- assume interest parameter is the mean: $\theta = \sum_{r=0}^{k} t_r \pi_r$

SEMs and Cuts

- data $\leftrightarrow$ $k+1$ cells with
  - labels $\{t_0, t_1, \ldots, t_k\}$ ... as well as ...
  - counts $\{n_0, n_1, \ldots, n_k\}$ (summing to N)
- model $\leftrightarrow$ parameterised set of probs. $\{\pi_0, \ldots, \pi_k\}$
- assume interest parameter is the mean: $\theta = \sum_{r=0}^{k} t_r \pi_r$
- multinomial: $n := (n_1, \ldots, n_k)^T \sim \text{Multi}(N, \pi)$

SEMs and Cuts

- data $\leftrightarrow$ $k+1$ cells with
  - labels $\{t_0, t_1, \ldots, t_k\}$ ... as well as ...
  - counts $\{n_0, n_1, \ldots, n_k\}$ (summing to N)
- model $\leftrightarrow$ parameterised set of probs. $\{\pi_0, \ldots, \pi_k\}$
- assume interest parameter is the mean: $\theta = \sum_{r=0}^{k} t_r \pi_r$
- multinomial: $n := (n_1, \ldots, n_k)^T \sim \text{Multi}(N, \pi)$ – a NEF with density $\exp(n^T\gamma - \psi(\gamma))$,

SEMs and Cuts

- data $\leftrightarrow$ $k+1$ cells with
  - labels $\{t_0, t_1, \ldots, t_k\}$ ... as well as ...
  - counts $\{n_0, n_1, \ldots, n_k\}$ (summing to N)
- model $\leftrightarrow$ parameterised set of probs. $\{\pi_0, \ldots, \pi_k\}$
- assume interest parameter is the mean: $\theta = \sum_{r=0}^{k} t_r \pi_r$
- multinomial: $n := (n_1, \ldots, n_k)^T \sim \text{Multi}(N, \pi)$ – a NEF with density $\exp(n^T\gamma - \psi(\gamma))$, $\gamma_r := \log(\pi_r / \pi_0)$

SEMs and Cuts

- data $\leftrightarrow$ $k+1$ cells with
  - labels $\{t_0, t_1, \ldots, t_k\}$ ... as well as ...
  - counts $\{n_0, n_1, \ldots, n_k\}$ (summing to N)
- model $\leftrightarrow$ parameterised set of probs. $\{\pi_0, \ldots, \pi_k\}$
- assume interest parameter is the mean: $\theta = \sum_{r=0}^{k} t_r \pi_r$
- multinomial: $n := (n_1, \ldots, n_k)^T \sim \text{Multi}(N, \pi)$ – a NEF with density $\exp(n^T\gamma - \psi(\gamma))$, $\gamma_r := \log(\pi_r/\pi_0)$
- extended multinomial: support sets vary:

SEMs and Cuts

- data $\leftrightarrow k+1$ cells with
  - labels $\{t_0, t_1, \ldots, t_k\}$ ... as well as ...
  - counts $\{n_0, n_1, \ldots, n_k\}$ (summing to N)
- model $\leftrightarrow$ parameterised set of probs. $\{\pi_0, \ldots, \pi_k\}$
- assume interest parameter is the mean: $\theta = \sum_{r=0}^{k} t_r \pi_r$
- multinomial: $n := (n_1, \ldots, n_k)^T \sim \text{Multi}(N, \pi)$ – a NEF
  with density $\exp(n^T\gamma - \psi(\gamma))$, $\gamma_r := \log(\pi_r/\pi_0)$
- extended multinomial: support sets vary:
  for each $\varnothing \neq K \subseteq \{0, 1, ..., k\}$, π_K has $\pi_r > 0 \Leftrightarrow r \in K$,

SEMs and Cuts

- data $\leftrightarrow$ $k+1$ cells with
  - labels $\{t_0, t_1, \ldots, t_k\}$ ... as well as ...
  - counts $\{n_0, n_1, \ldots, n_k\}$ (summing to N)
- model $\leftrightarrow$ parameterised set of probs. $\{\pi_0, \ldots, \pi_k\}$
- assume interest parameter is the mean: $\theta = \sum_{r=0}^{k} t_r \pi_r$
- multinomial: $n := (n_1, \ldots, n_k)^T \sim \mathrm{Multi}(N, \pi)$ – a NEF
  with density $\exp(n^T\gamma - \psi(\gamma))$, $\gamma_r := \log(\pi_r/\pi_0)$
- extended multinomial: support sets vary:
  for each $\varnothing \neq K \subseteq \{0, 1, ..., k\}$, π_K has $\pi_r > 0 \Leftrightarrow r \in K$,
  while $n_K := (n_r)_{r \in K \setminus \{r_0\}} \sim \mathrm{Multi}(N, \pi_K)$

SEMs and Cuts

- data $\leftrightarrow$ $k+1$ cells with
  - labels $\{t_0, t_1, \ldots, t_k\}$ ... as well as ...
  - counts $\{n_0, n_1, \ldots, n_k\}$ (summing to N)
- model $\leftrightarrow$ parameterised set of probs. $\{\pi_0, \ldots, \pi_k\}$
- assume interest parameter is the mean: $\theta = \sum_{r=0}^{k} t_r \pi_r$
- multinomial: $n := (n_1, \ldots, n_k)^T \sim \text{Multi}(N, \pi)$ – a NEF with density $\exp(n^T \gamma - \psi(\gamma))$, $\gamma_r := \log(\pi_r / \pi_0)$
- extended multinomial: support sets vary: for each $\varnothing \neq K \subseteq \{0, 1, ..., k\}$, π_K has $\pi_r > 0 \Leftrightarrow r \in K$, while $n_K := (n_r)_{r \in K \setminus \{r_0\}} \sim \text{Multi}(N, \pi_K)$ – a NEF with density $\exp(n_K^T \gamma_K - \psi(\gamma_K))$,

SEMs and Cuts

- data $\leftrightarrow k+1$ cells with
  - labels $\{t_0, t_1, \ldots, t_k\}$ ... as well as ...
  - counts $\{n_0, n_1, \ldots, n_k\}$ (summing to N)
- model $\leftrightarrow$ parameterised set of probs. $\{\pi_0, \ldots, \pi_k\}$
- assume interest parameter is the mean: $\theta = \sum_{r=0}^{k} t_r \pi_r$
- multinomial: $n := (n_1, \ldots, n_k)^T \sim \text{Multi}(N, \pi)$ – a NEF
  with density $\exp(n^T \gamma - \psi(\gamma))$, $\gamma_r := \log(\pi_r / \pi_0)$
- extended multinomial: support sets vary:
  for each $\varnothing \neq K \subseteq \{0, 1, ..., k\}$, π_K has $\pi_r > 0 \Leftrightarrow r \in K$,
  while $n_K := (n_r)_{r \in K \setminus \{r_0\}} \sim \text{Multi}(N, \pi_K)$ – a NEF
  with density $\exp(n_K^T \gamma_K - \psi(\gamma_K))$, $\gamma_r := \log(\pi_r / \pi_{r_0})$

SEMs and Cuts: mixed parameterisation ...

- Let $A := (d, v_1, \ldots, v_{k-1})$ be nonsingular, with each v_r (Euclidean) orthogonal to $d := N^{-1}(t_1 - t_0, \ldots, t_k - t_0)^T$, ...

SEMs and Cuts: mixed parameterisation ...

- Let $A := (d, v_1, \ldots, v_{k-1})$ be nonsingular, with each v_r (Euclidean) orthogonal to $d := N^{-1}(t_1 - t_0, \ldots, t_k - t_0)^T$, ...
- ... and put $s := A^T n + c$ where $c := (t_0, 0, \ldots, 0)^T$,

  so that $s_1 := t_0 + d^T n$ has $E[s_1] = \theta$

SEMs and Cuts: mixed parameterisation ...

- Let $A := (d, v_1, \ldots, v_{k-1})$ be nonsingular, with each v_r (Euclidean) orthogonal to $d := N^{-1}(t_1 - t_0, \ldots, t_k - t_0)^T$, ...
- ... and put $s := A^T n + c$ where $c := (t_0, 0, \ldots, 0)^T$,

  so that $s_1 := t_0 + d^T n$ has $E[s_1] = \theta$
- Then, s follows a NEF

SEMs and Cuts: mixed parameterisation ...

- Let $A := (d, v_1, \dots, v_{k-1})$ be nonsingular, with each v_r (Euclidean) orthogonal to $d := N^{-1}(t_1 - t_0, \dots, t_k - t_0)^T$, ...
- ... and put $s := A^T n + c$ where $c := (t_0, 0, \dots, 0)^T$,
  so that $s_1 := t_0 + d^T n$ has $E[s_1] = \theta$
- Then, s follows a NEF – having density $\exp(s^T\phi - M(\phi))$, with mean & natural params. $\mu = A^T\pi + c$ & $\phi = A^{-1}\gamma$

SEMs and Cuts: mixed parameterisation ...

- Let $A := (d, v_1, \ldots, v_{k-1})$ be nonsingular, with each v_r (Euclidean) orthogonal to $d := N^{-1}(t_1 - t_0, \ldots, t_k - t_0)^T$, ...
- ... and put $s := A^T n + c$ where $c := (t_0, 0, \ldots, 0)^T$,

  so that $s_1 := t_0 + d^T n$ has $E[s_1] = \theta$
- Then, s follows a NEF – having density $\exp(s^T \phi - M(\phi))$, with mean & natural params. $\mu = A^T \pi + c$ & $\phi = A^{-1}\gamma$
- Splitting these as: $\mu = \begin{pmatrix} \theta \\ \mu_2 \end{pmatrix}$ & $\phi = \begin{pmatrix} \phi_1 \\ \eta \end{pmatrix}$,

SEMs and Cuts: mixed parameterisation ...

- Let $A := (d, v_1, \ldots, v_{k-1})$ be nonsingular, with each v_r (Euclidean) orthogonal to $d := N^{-1}(t_1 - t_0, \ldots, t_k - t_0)^T$, ...
- ... and put $s := A^T n + c$ where $c := (t_0, 0, \ldots, 0)^T$,
  so that $s_1 := t_0 + d^T n$ has $E[s_1] = \theta$
- Then, s follows a NEF – having density $\exp(s^T \phi - M(\phi))$, with mean & natural params. $\mu = A^T \pi + c$ & $\phi = A^{-1}\gamma$
- Splitting these as: $\mu = \begin{pmatrix} \theta \\ \mu_2 \end{pmatrix}$ & $\phi = \begin{pmatrix} \phi_1 \\ \eta \end{pmatrix}$, arrive at ...
- the part mean/part natural mixed parameters $\begin{pmatrix} \theta \\ \eta \end{pmatrix}$

SEMs and Cuts: mixed parameterisation ...

- Let $A := (d, v_1, \ldots, v_{k-1})$ be nonsingular, with each v_r (Euclidean) orthogonal to $d := N^{-1}(t_1 - t_0, \ldots, t_k - t_0)^T$, ...
- ... and put $s := A^T n + c$ where $c := (t_0, 0, \ldots, 0)^T$,
  so that $s_1 := t_0 + d^T n$ has $E[s_1] = \theta$
- Then, s follows a NEF – having density $\exp(s^T \phi - M(\phi))$, with mean & natural params. $\mu = A^T \pi + c$ & $\phi = A^{-1}\gamma$
- Splitting these as: $\mu = \begin{pmatrix} \theta \\ \mu_2 \end{pmatrix}$ & $\phi = \begin{pmatrix} \phi_1 \\ \eta \end{pmatrix}$, arrive at ...
- the part mean/part natural mixed parameters $\begin{pmatrix} \theta \\ \eta \end{pmatrix}$ with:
- θ and η Fisher orthogonal everywhere,

SEMs and Cuts: mixed parameterisation ...

- Let $A := (d, v_1, \ldots, v_{k-1})$ be nonsingular, with each v_r (Euclidean) orthogonal to $d := N^{-1}(t_1 - t_0, \ldots, t_k - t_0)^T$, ...
- ... and put $s := A^T n + c$ where $c := (t_0, 0, \ldots, 0)^T$,
  so that $s_1 := t_0 + d^T n$ has $E[s_1] = \theta$
- Then, s follows a NEF – having density $\exp(s^T \phi - M(\phi))$, with mean & natural params. $\mu = A^T \pi + c$ & $\phi = A^{-1}\gamma$
- Splitting these as: $\mu = \begin{pmatrix} \theta \\ \mu_2 \end{pmatrix}$ & $\phi = \begin{pmatrix} \phi_1 \\ \eta \end{pmatrix}$, arrive at ...
- the part mean/part natural mixed parameters $\begin{pmatrix} \theta \\ \eta \end{pmatrix}$ with:
- θ and η Fisher orthogonal everywhere, ... while ...

SEMs and Cuts: mixed parameterisation ...

- Let $A := (d, v_1, \ldots, v_{k-1})$ be nonsingular, with each v_r (Euclidean) orthogonal to $d := N^{-1}(t_1 - t_0, \ldots, t_k - t_0)^T$, ...
- ... and put $s := A^T n + c$ where $c := (t_0, 0, \ldots, 0)^T$,
  so that $s_1 := t_0 + d^T n$ has $E[s_1] = \theta$
- Then, s follows a NEF – having density $\exp(s^T \phi - M(\phi))$, with mean & natural params. $\mu = A^T \pi + c$ & $\phi = A^{-1}\gamma$
- Splitting these as: $\mu = \begin{pmatrix} \theta \\ \mu_2 \end{pmatrix}$ & $\phi = \begin{pmatrix} \phi_1 \\ \eta \end{pmatrix}$, arrive at ...
- the part mean/part natural mixed parameters $\begin{pmatrix} \theta \\ \eta \end{pmatrix}$ with:
- θ and η Fisher orthogonal everywhere, ... while ...
- s_1 is a cut iff $Var(s_1)$ depends only on θ

SEMs and Cuts for model (in)sensitivity

- Inferences on θ can be made using $S(x)$

SEMs and Cuts for model (in)sensitivity

- Inferences on θ can be made using $S(x)$
  – e.g. an interval estimate for θ

SEMs and Cuts for model (in)sensitivity

- Inferences on θ can be made using $S(x)$
  – e.g. an interval estimate for θ
- *Ideal case*:

SEMs and Cuts for model (in)sensitivity

- Inferences on θ can be made using $S(x)$
  – e.g. an interval estimate for θ
- *Ideal case*:
  $\exists$ an enlarged model $g(x;\theta,\eta)$ with cut $(S(x),U(x))$

SEMs and Cuts for model (in)sensitivity

- Inferences on θ can be made using $S(x)$
  – e.g. an interval estimate for θ
- *Ideal case*:
  $\exists$ an enlarged model $g(x;\theta,\eta)$ with cut $(S(x), U(x))$ $\Rightarrow$
  perturbation $\eta_0 \to \eta$ does not affect inferences about θ

SEMs and Cuts for model (in)sensitivity

- Inferences on θ can be made using $S(x)$
  – e.g. an interval estimate for θ
- *Ideal case*:
  $\exists$ an enlarged model $g(x; \theta, \eta)$ with cut $(S(x), U(x)) \quad \Rightarrow$ perturbation $\eta_0 \to \eta$ does not affect inferences about θ
- In practice, *relax* conditions to define an *approximate* cut:

SEMs and Cuts for model (in)sensitivity

- Inferences on θ can be made using $S(x)$
  – e.g. an interval estimate for θ
- *Ideal case*:
  $\exists$ an enlarged model $g(x; \theta, \eta)$ with cut $(S(x), U(x)) \Rightarrow$ perturbation $\eta_0 \rightarrow \eta$ does not affect inferences about θ
- In practice, *relax* conditions to define an *approximate* cut:
  - inferences about θ based on $S(x)$ 'do not change much', when perturbing in the direction η

SEMs and Cuts for model (in)sensitivity

- Inferences on θ can be made using $S(x)$
  – e.g. an interval estimate for θ
- *Ideal case*:
  $\exists$ an enlarged model $g(x;\theta,\eta)$ with cut $(S(x), U(x)) \Rightarrow$ perturbation $\eta_0 \longrightarrow \eta$ does not affect inferences about θ
- In practice, *relax* conditions to define an *approximate* cut:
  - inferences about θ based on $S(x)$ 'do not change much', when perturbing in the direction η
  - for fixed data x,

SEMs and Cuts for model (in)sensitivity

- Inferences on θ can be made using $S(x)$
  – e.g. an interval estimate for θ
- *Ideal case*:
  $\exists$ an enlarged model $g(x;\theta,\eta)$ with cut $(S(x),U(x))$ $\Rightarrow$
  perturbation $\eta_0 \to \eta$ does not affect inferences about θ
- In practice, *relax* conditions to define an *approximate* cut:
  - inferences about θ based on $S(x)$ 'do not change much', when perturbing in the direction η
  - for fixed data x, focus on:
    $\{(\theta,\eta) : g(x;\theta,\eta)$ is data-supported$\}$

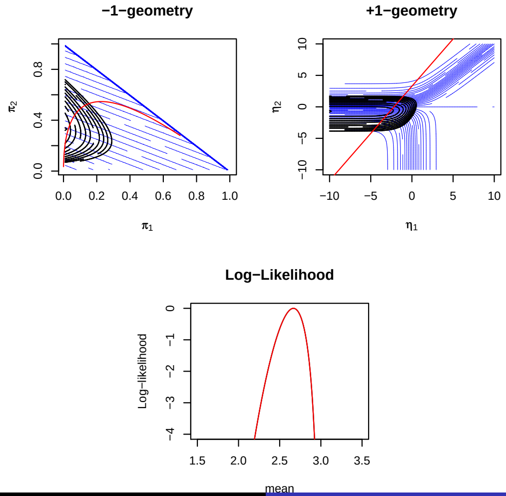
SEMs and Cuts for model (in)sensitivity

- Inferences on θ can be made using $S(x)$
  – e.g. an interval estimate for θ
- *Ideal case*:
  $\exists$ an enlarged model $g(x; \theta, \eta)$ with cut $(S(x), U(x)) \Rightarrow$ perturbation $\eta_0 \to \eta$ does not affect inferences about θ
- In practice, *relax* conditions to define an *approximate* cut:
  - inferences about θ based on $S(x)$ 'do not change much', when perturbing in the direction η
  - for fixed data x, focus on:
    $\{(\theta, \eta) : g(x; \theta, \eta)$ is data-supported$\}$
- Conversely,
  we can find directions η with a large effect on θ-inference

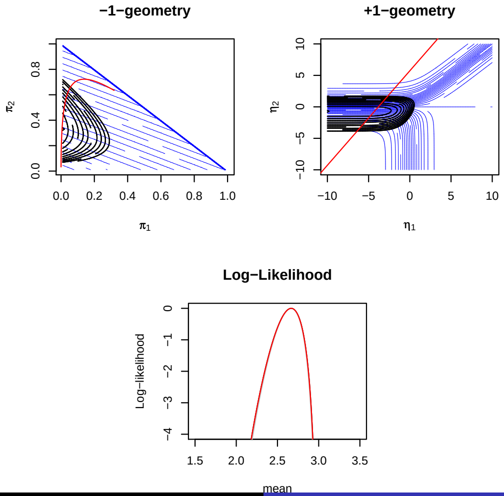
SEMs and Cuts for model (in)sensitivity

- Inferences on θ can be made using $S(x)$
  – e.g. an interval estimate for θ
- *Ideal case*:
  $\exists$ an enlarged model $g(x;\theta,\eta)$ with cut $(S(x), U(x))$ $\Rightarrow$ perturbation $\eta_0 \to \eta$ does not affect inferences about θ
- In practice, *relax* conditions to define an *approximate* cut:
  - inferences about θ based on $S(x)$ 'do not change much', when perturbing in the direction η
  - for fixed data x, focus on:
    $\{(\theta,\eta) : g(x;\theta,\eta)$ is data-supported$\}$
- Conversely,
  we can find directions η with a large effect on θ-inference
  – e.g. length of confidence interval increases substantially

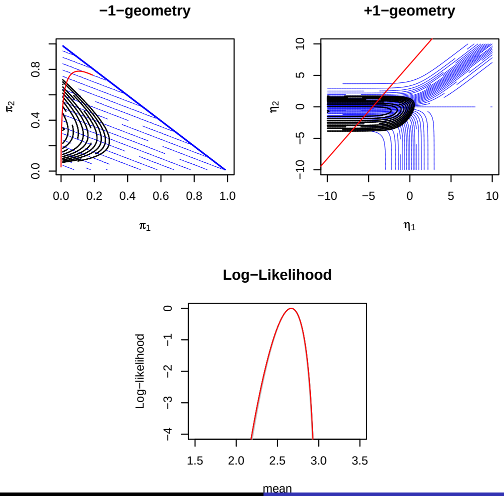
Effect of translation



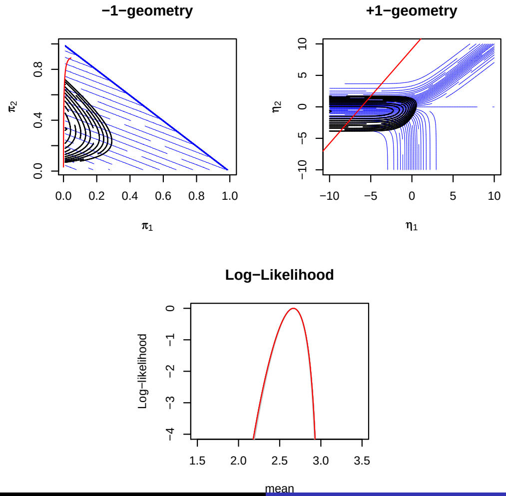
Effect of translation



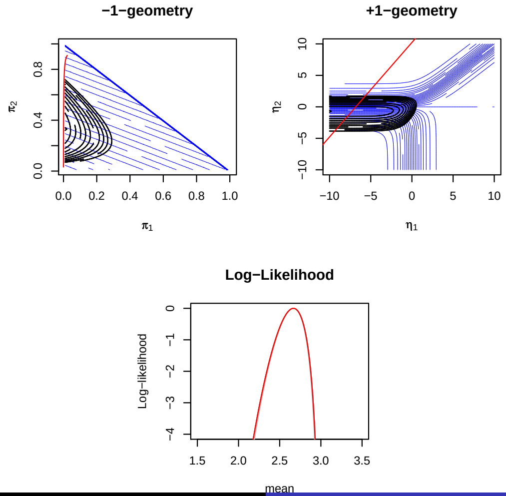
Effect of translation



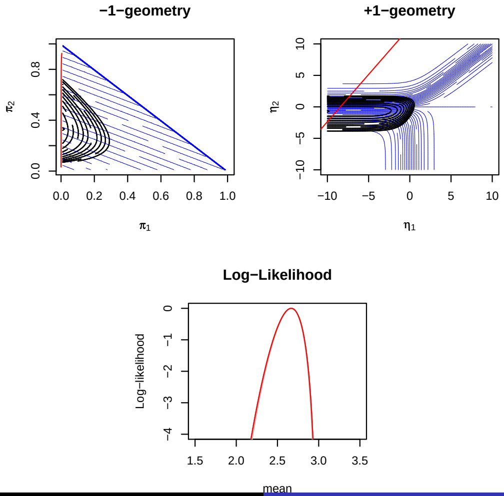
Effect of translation



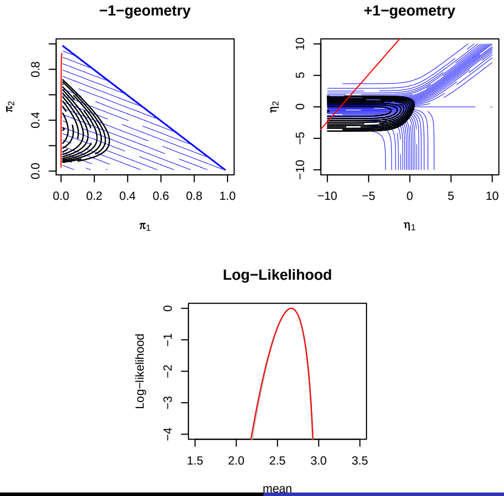
Effect of translation



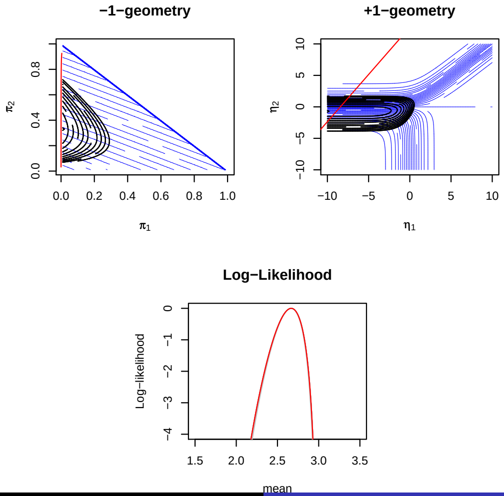
Effect of translation



Effect of translation



Effect of translation



Sensitive perturbations

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

- developing an appropriate sensitivity analysis for a given WPF requires seeing how its model sits inside the universal class of (extended) multinomial models

- developing an appropriate sensitivity analysis for a given WPF requires seeing how its model sits inside the universal class of (extended) multinomial models
- natural correspondences between geometric & statistical ideas allow us to formalise this by *building an appropriate* perturbation space which:

- developing an appropriate sensitivity analysis for a given WPF requires seeing how its model sits inside the universal class of (extended) multinomial models
- natural correspondences between geometric & statistical ideas allow us to formalise this by *building an appropriate* perturbation space which:
  - enriches the working model ...

- developing an appropriate sensitivity analysis for a given WPF requires seeing how its model sits inside the universal class of (extended) multinomial models
- natural correspondences between geometric & statistical ideas allow us to formalise this by *building an appropriate* perturbation space which:
  - enriches the working model ...
  - in perturbation directions having a strong effect on ...

- developing an appropriate sensitivity analysis for a given WPF requires seeing how its model sits inside the universal class of (extended) multinomial models
- natural correspondences between geometric & statistical ideas allow us to formalise this by *building an appropriate* perturbation space which:
  - enriches the working model ...
  - in perturbation directions having a strong effect on ...
  - A := inference for θ_Q

- developing an appropriate sensitivity analysis for a given WPF requires seeing how its model sits inside the universal class of (extended) multinomial models
- natural correspondences between geometric & statistical ideas allow us to formalise this by *building an appropriate* perturbation space which:
  - enriches the working model ...
  - in perturbation directions having a strong effect on ...
  - A := inference for θ_Q
- ... in general, as follows ...

- developing an appropriate sensitivity analysis for a given WPF requires seeing how its model sits inside the universal class of (extended) multinomial models
- natural correspondences between geometric & statistical ideas allow us to formalise this by *building an appropriate* perturbation space which:
  - enriches the working model ...
  - in perturbation directions having a strong effect on ...
  - A := inference for θ_Q
- ... in general, as follows ...
- ??? OR focus solely on:

- developing an appropriate sensitivity analysis for a given WPF requires seeing how its model sits inside the universal class of (extended) multinomial models
- natural correspondences between geometric & statistical ideas allow us to formalise this by *building an appropriate* perturbation space which:
  - enriches the working model ...
  - in perturbation directions having a strong effect on ...
  - A := inference for θ_Q
- ... in general, as follows ...
- ??? OR focus solely on:
  - statistically: change (baseline) measure = ...

- developing an appropriate sensitivity analysis for a given WPF requires seeing how its model sits inside the universal class of (extended) multinomial models
- natural correspondences between geometric & statistical ideas allow us to formalise this by *building an appropriate* perturbation space which:
  - enriches the working model ...
  - in perturbation directions having a strong effect on ...
  - A := inference for θ_Q
- ... in general, as follows ...
- ??? OR focus solely on:
  - statistically: change (baseline) measure = ...
  - geometrically: translation (+1 parallel transport)

- developing an appropriate sensitivity analysis for a given WPF requires seeing how its model sits inside the universal class of (extended) multinomial models
- natural correspondences between geometric & statistical ideas allow us to formalise this by *building an appropriate* perturbation space which:
  - enriches the working model ...
  - in perturbation directions having a strong effect on ...
  - A := inference for θ_Q
- ... in general, as follows ...
- ??? OR focus solely on:
  - statistically: change (baseline) measure = ...
  - geometrically: translation (+1 parallel transport)
  - [jump to ... Information Directions]

change number of parameters = change dimension

Statistical Idea	$\leftrightarrow$	Geometric Idea
add sufficient statistics ($\nearrow$ model flexibility)		embed model in richer (E) family

change number of parameters = change dimension

Statistical Idea	$\leftrightarrow$	Geometric Idea
add sufficient statistics ($\nearrow$ model flexibility)		embed model in richer (E) family

- Examples

change number of parameters = change dimension

Statistical Idea	$\leftrightarrow$	Geometric Idea
add sufficient statistics ($\nearrow$ model flexibility)		embed model in richer (E) family

- Examples
  - $N(\mu, 1) \to N(\mu, \sigma^2)$

change number of parameters = change dimension

Statistical Idea	$\leftrightarrow$	Geometric Idea
add sufficient statistics ($\nearrow$ model flexibility)		embed model in richer (E) family

- Examples
  - $N(\mu, 1) \rightarrow N(\mu, \sigma^2)$
  - outlier accommodation (add term(s), as reqd.)

change number of parameters = change dimension

Statistical Idea	$\leftrightarrow$	Geometric Idea
■ $\sim$ no empirical info. in some directions		$l(\cdot)$ is $\sim$ flat on a (-1) subspace

change number of parameters = change dimension

Statistical Idea	$\leftrightarrow$	Geometric Idea
$\sim$ no empirical info. in some directions		$I(\cdot)$ is $\sim$ flat on a (-1) subspace

- **N.B.** corresponding pertns. are *putative*

change number of parameters = change dimension

Statistical Idea	$\leftrightarrow$	Geometric Idea
$\sim$ no empirical info. in some directions		$l(\cdot)$ is $\sim$ flat on a (-1) subspace

- **N.B.** corresponding pertns. are *putative*
- Examples

change number of parameters = change dimension

Statistical Idea	$\leftrightarrow$	Geometric Idea
$\sim$ no empirical info. in some directions		$I(\cdot)$ is $\sim$ flat on a (-1) subspace

- **N.B.** corresponding pertns. are *putative*
- Examples
  - extreme tail behaviour

change number of parameters = change dimension

Statistical Idea	$\leftrightarrow$	Geometric Idea
$\sim$ no empirical info. in some directions		$l(\cdot)$ is $\sim$ flat on a (-1) subspace

- $\sim$ no empirical info. in some directions
- **N.B.** corresponding pertns. are *putative*
- Examples
  - extreme tail behaviour
  - $\phi \to \phi/\omega_i$ (non-identifiability)

keep same no. of parameters = keep same dimension

Statistical Idea	$\longleftrightarrow$	Geometric Idea
change {sufficient statistics} (bias/variance trade-off)		rotate (about fixed base point) ($\exists$ LID: see end of 'Info. Dirns.')

- Note: In A. I. literature, 'select different features'

keep same no. of parameters = keep same dimension

Statistical Idea $\leftrightarrow$ **Geometric Idea**

Statistical Idea		Geometric Idea
change {sufficient statistics}	$\leftrightarrow$	rotate (about fixed base point)
(bias/variance trade-off)		($\exists$ LID: see end of 'Info. Dirns.')

- Note: In A. I. literature, 'select different features'
- Example: $N(\mu, \sigma^2) \rightarrow \log N(\mu, \sigma^2)$

keep same no. of parameters = keep same dimension

Statistical Idea	$\leftrightarrow$	Geometric Idea
change {sufficient statistics}		rotate (about fixed base point)
(bias/variance trade-off)		($\exists$ LID: see end of 'Info. Dirns.')

- Note: In A. I. literature, 'select different features'
- Example: $N(\mu, \sigma^2) \to \log N(\mu, \sigma^2)$
- This example also has elements of ...

keep same no. of parameters = keep same dimension

Statistical Idea	$\leftrightarrow$	Geometric Idea
change (baseline) measure		translation ($+1$ parallel transport)

- Example:
  $N(\mu, \sigma^2) \rightarrow \log N(\mu, \sigma^2)$

keep same no. of parameters = keep same dimension

Statistical Idea		Geometric Idea
change (baseline) measure	$\leftrightarrow$	translation ($+1$ parallel transport)

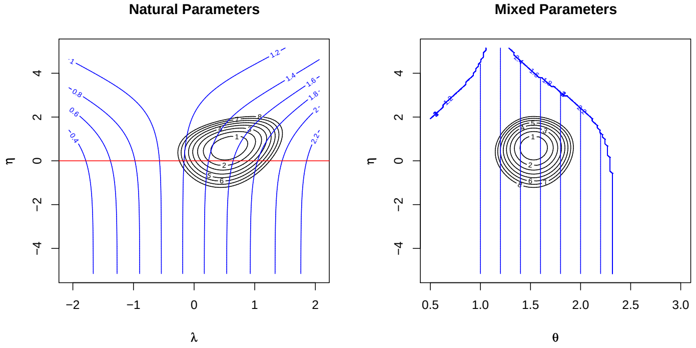
- Example:
  $N(\mu, \sigma^2) \to \log N(\mu, \sigma^2)$
- For this type of perturbation, the (approximate) CUT tool helps build a *data-informed* perturbation space
  [cf. C+M, (2004), Biometrika, Data-Informed IA]

keep same no. of parameters = keep same dimension

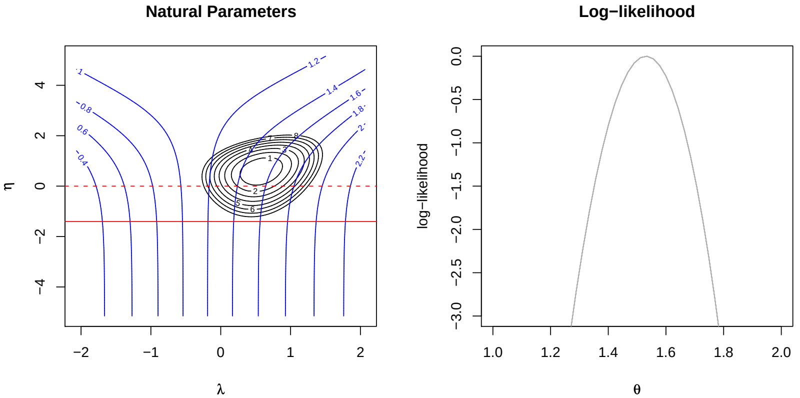
Statistical Idea	$\leftrightarrow$	Geometric Idea
change (baseline) measure		translation ($+1$ parallel transport)

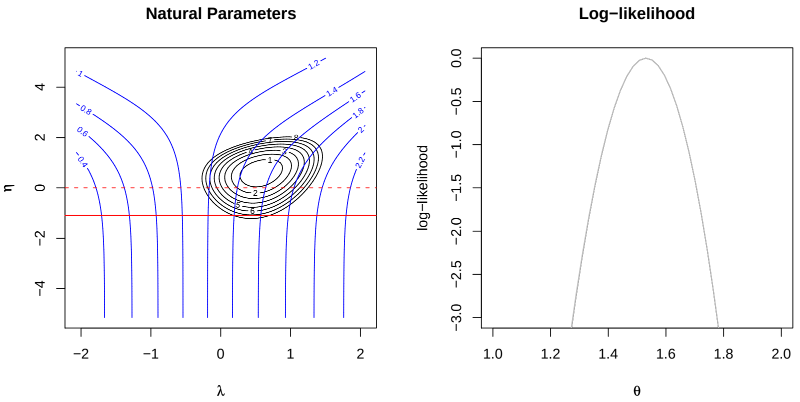
- Example:
  $N(\mu, \sigma^2) \to \log N(\mu, \sigma^2)$
- For this type of perturbation, the (approximate) CUT tool helps build a *data-informed* perturbation space
  [cf. C+M, (2004), Biometrika, Data-Informed IA]
- ... as we now illustrate ...

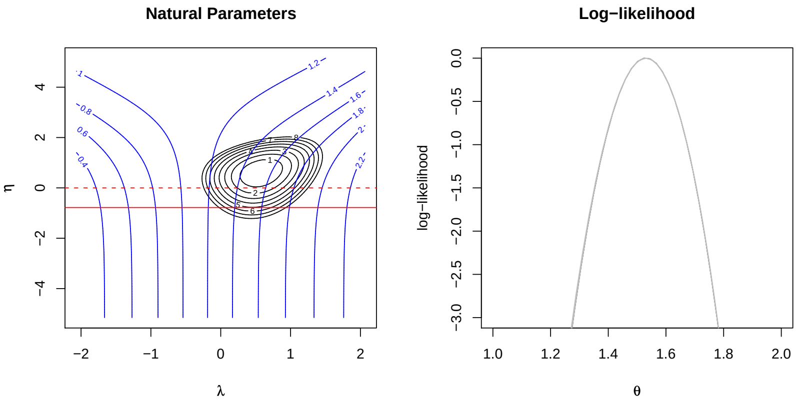
Insensitive direction: Mixed Parameters (Duality) CUT

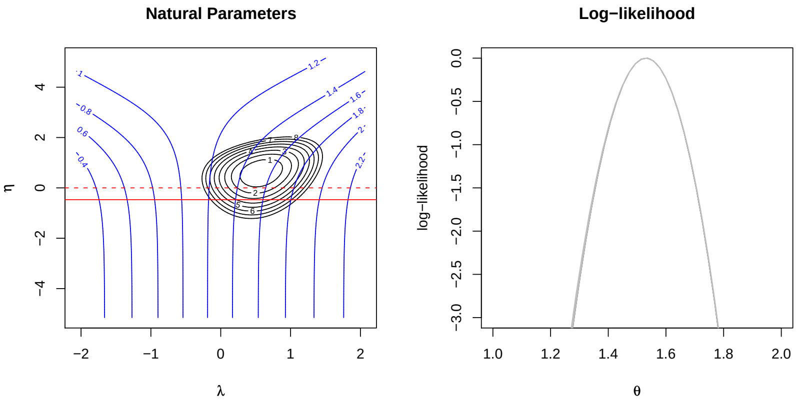


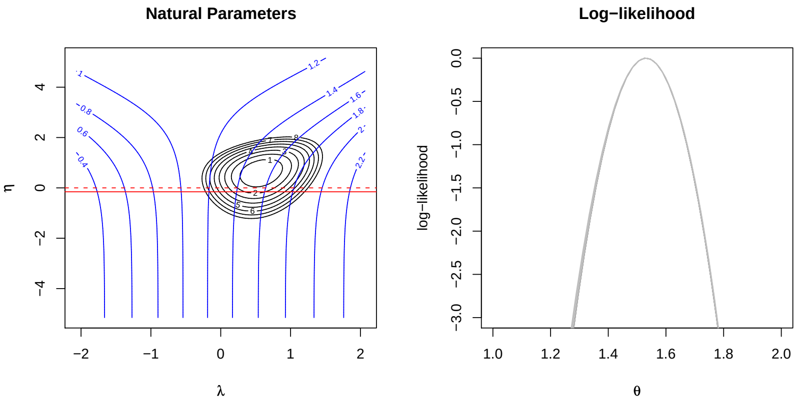
Blue = mean; Black = Likelihood; Red = Base Model [MP remarks] data \ outlier

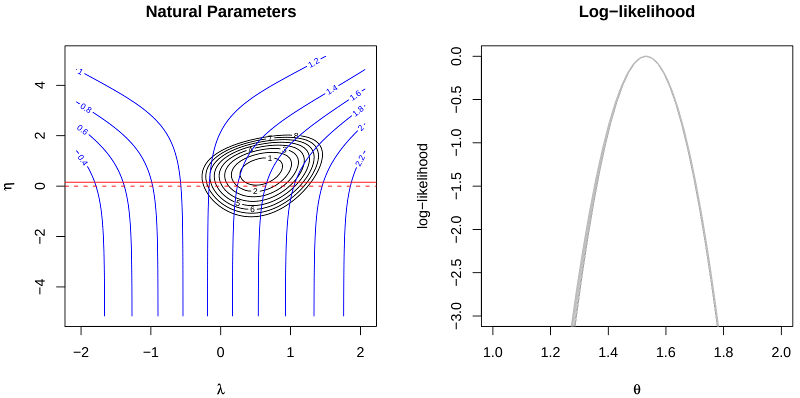

Natural Parameters
η
λ
−2
−1
0
1
2
4
2
0
−2
−4
1
0.8
0.6
0.4
1.2
1.4
1.6
1.8
2
2.2
Log−likelihood
log−likelihood
θ
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
1.0
1.2
1.4
1.6
1.8
2.0

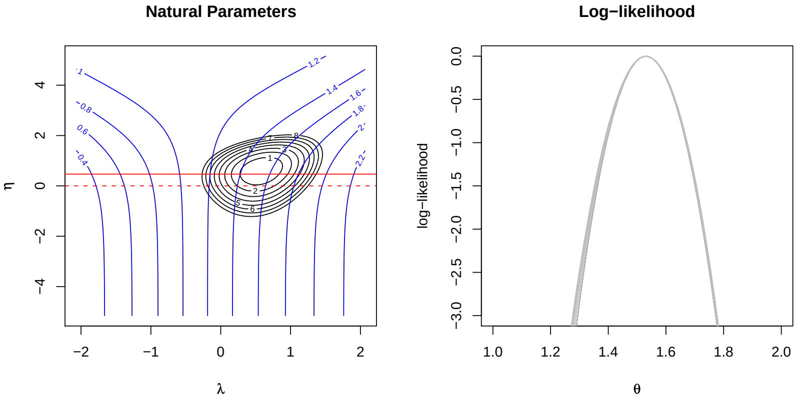

Natural Parameters
η
λ
Log−likelihood
log−likelihood
θ

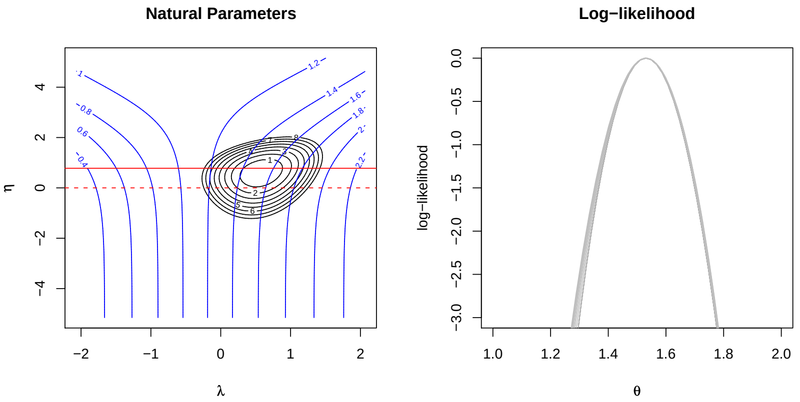

Natural Parameters
η
λ
Log−likelihood
log−likelihood
θ

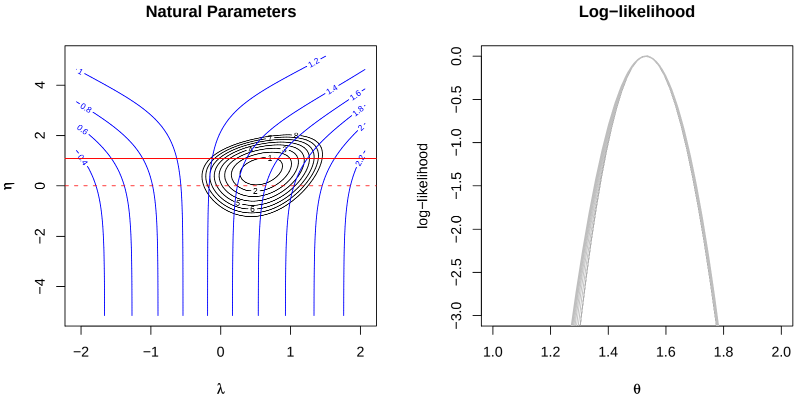

Natural Parameters
η
λ
−2
−1
0
1
2
4
−4
Log−likelihood
log−likelihood
θ
1.0
1.2
1.4
1.6
1.8
2.0
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0

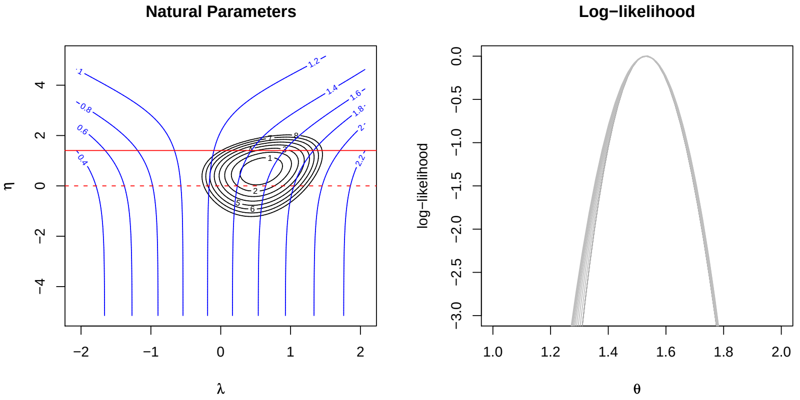

Natural Parameters
η
λ
Log−likelihood
log−likelihood
θ

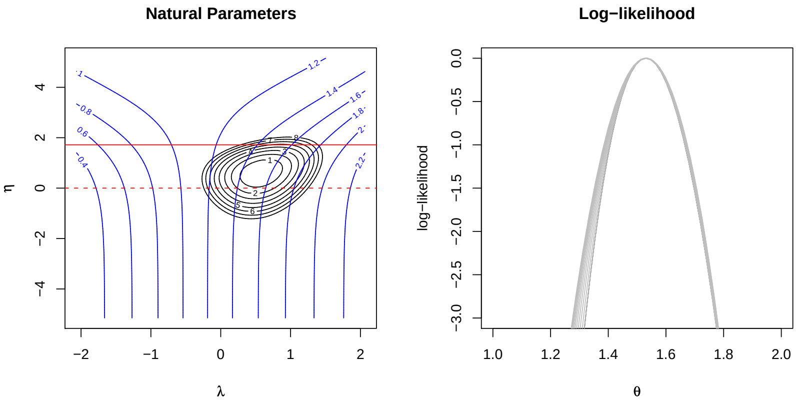

Natural Parameters
η
λ
Log−likelihood
log−likelihood
θ

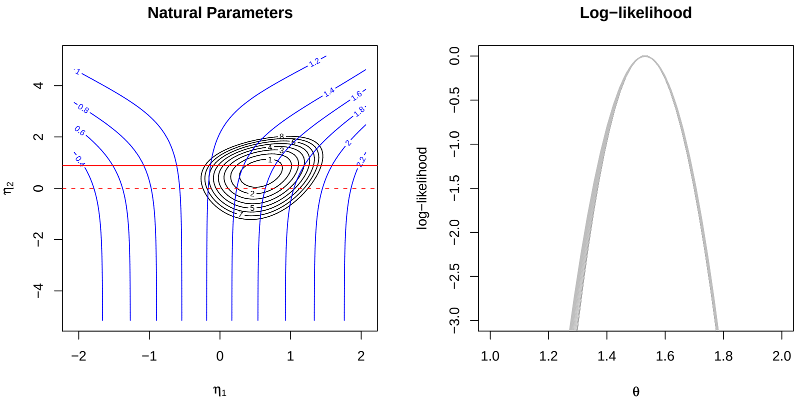

Natural Parameters
η
λ
Log−likelihood
log−likelihood
θ

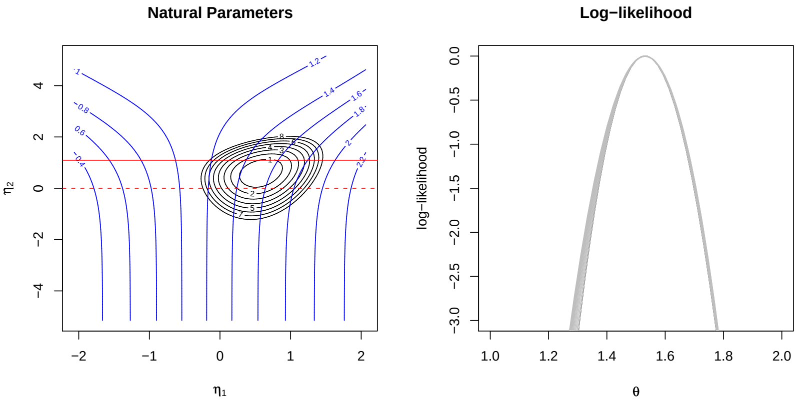

Natural Parameters
η
λ
−2
−1
0
1
2
4
2
0
−2
−4
Log−likelihood
log−likelihood
θ
1.0
1.2
1.4
1.6
1.8
2.0
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0

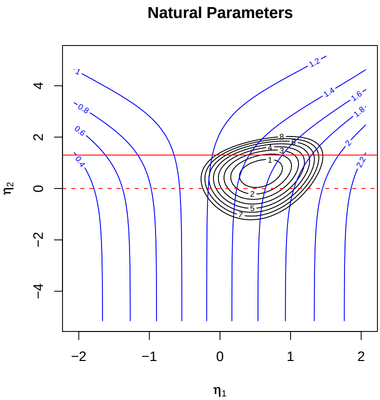

Natural Parameters
η
λ
−2 −1 0 1 2
4 2 0 −2 −4
Log−likelihood
log−likelihood
θ
1.0 1.2 1.4 1.6 1.8 2.0
0.0 −0.5 −1.0 −1.5 −2.0 −2.5 −3.0

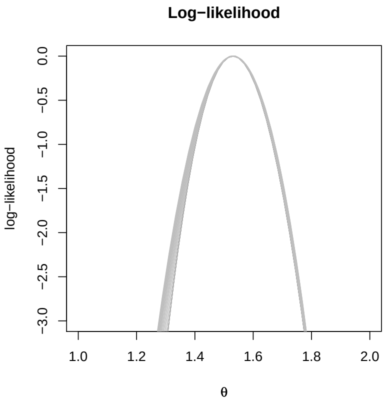

Natural Parameters
η
λ
Log−likelihood
log−likelihood
θ

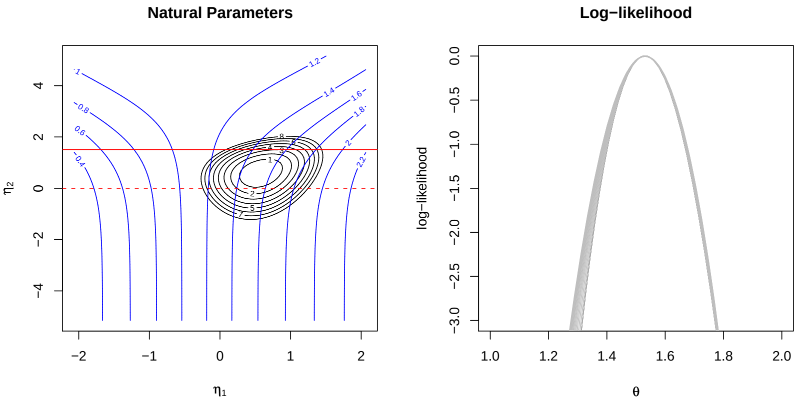

Natural Parameters
η
λ
Log−likelihood
log−likelihood
θ

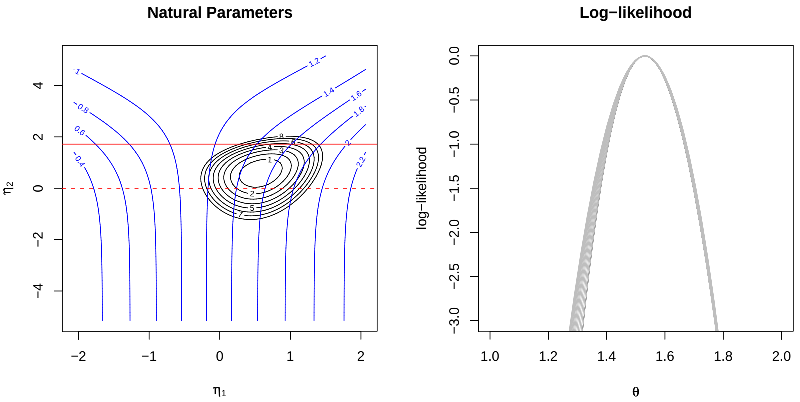

Natural Parameters
η2
η1
Log−likelihood
log−likelihood
θ

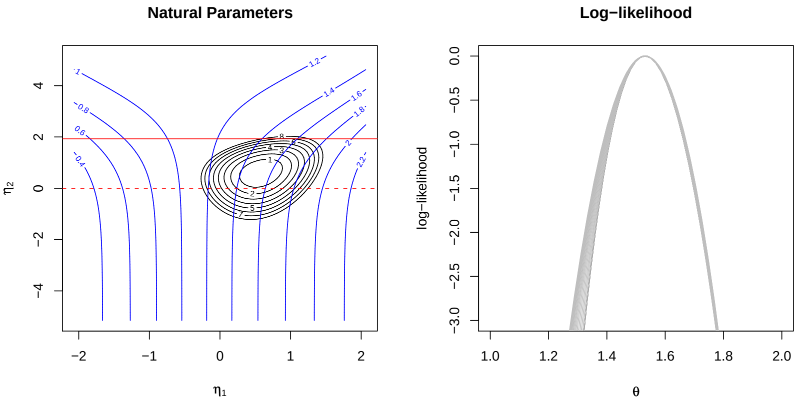

Natural Parameters
Log−likelihood
η2
η1
log−likelihood
θ


Natural Parameters
η2
η1
4
2
0
−2
−4
−2
−1
0
1
2

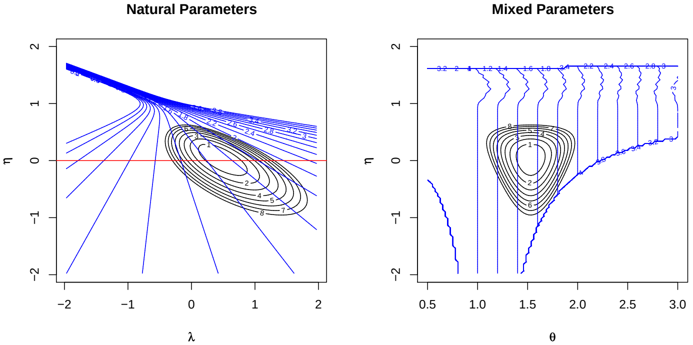

Log−likelihood
log−likelihood
θ
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
1.0
1.2
1.4
1.6
1.8
2.0


Natural Parameters
η_1
η_2
Log−likelihood
log−likelihood
θ

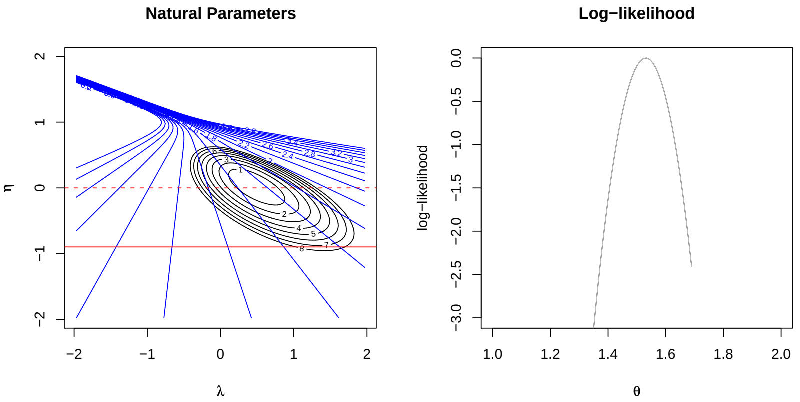

Natural Parameters
Log−likelihood
η1
η2
θ
log−likelihood

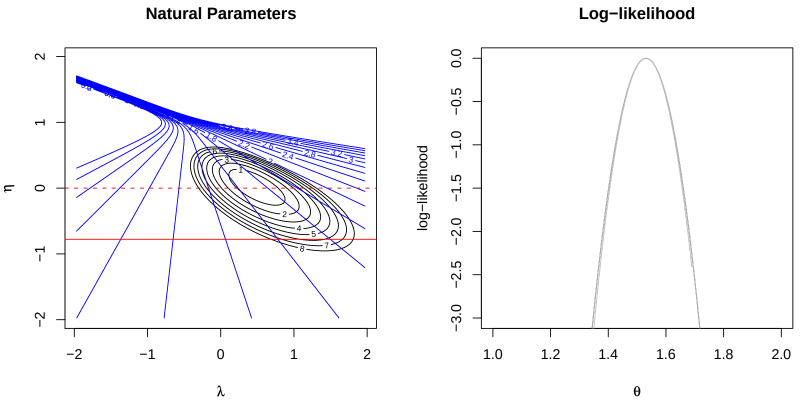

Natural Parameters
η2
η1
Log−likelihood
log−likelihood
θ

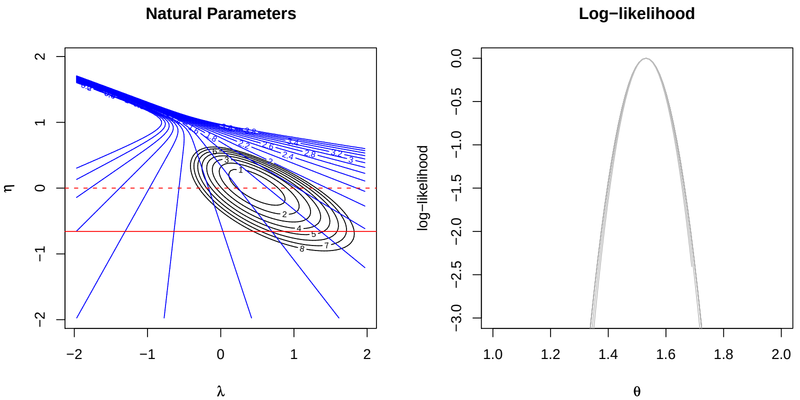
Highly sensitive direction: Mixed Parameters (Duality)

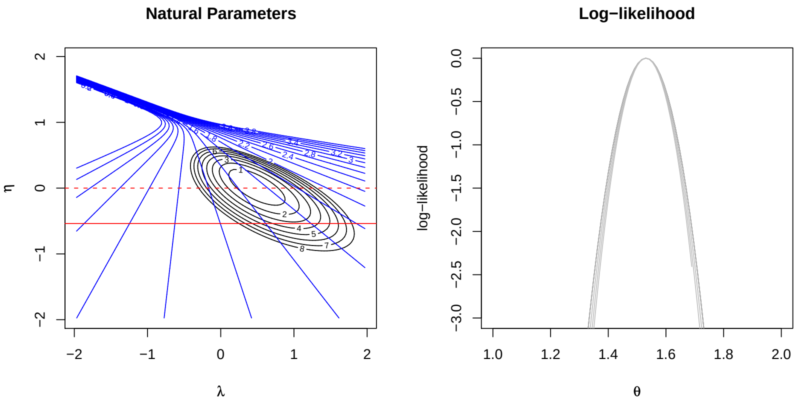


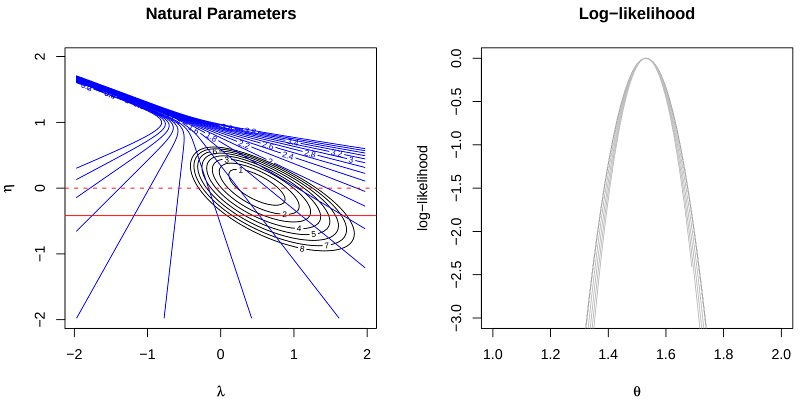
Blue = mean; Black = Likelihood; Red = Base Model [MP remarks] data \ outlier

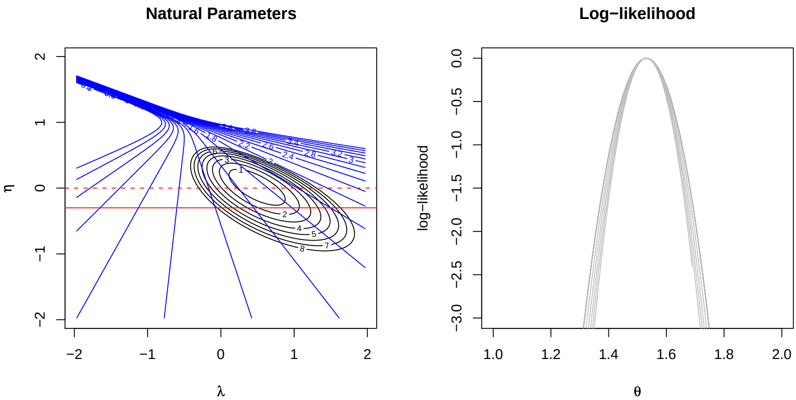

Natural Parameters
η
λ
Log−likelihood
log−likelihood
θ

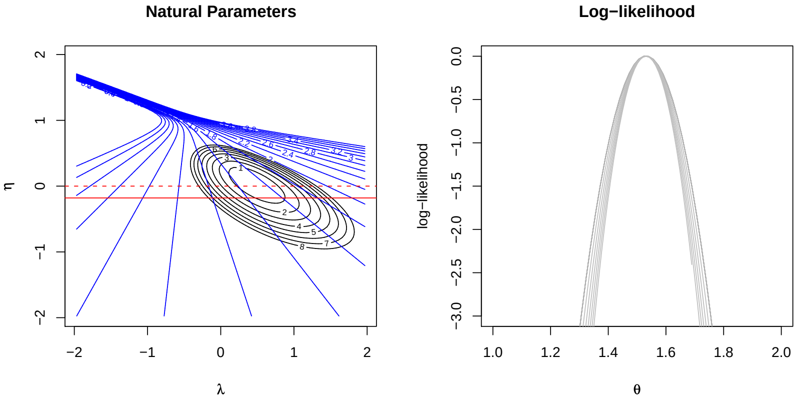

Natural Parameters
η
λ
−2
−1
0
1
2
Log−likelihood
log−likelihood
θ
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
1.0
1.2
1.4
1.6
1.8
2.0

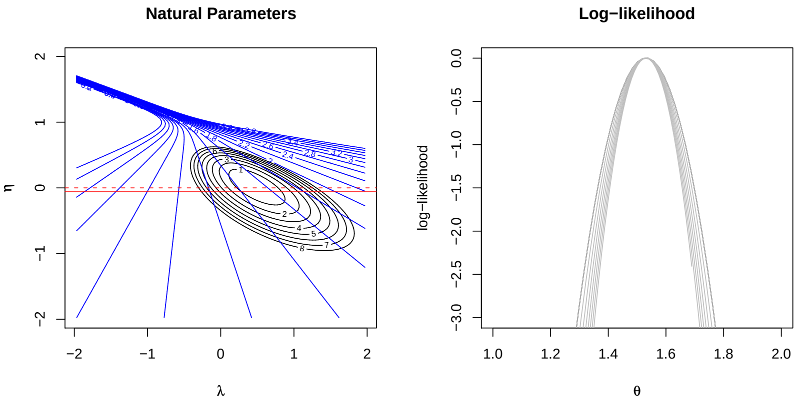

Natural Parameters
η
λ
−2
−1
0
1
2
Log−likelihood
log−likelihood
θ
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
1.0
1.2
1.4
1.6
1.8
2.0

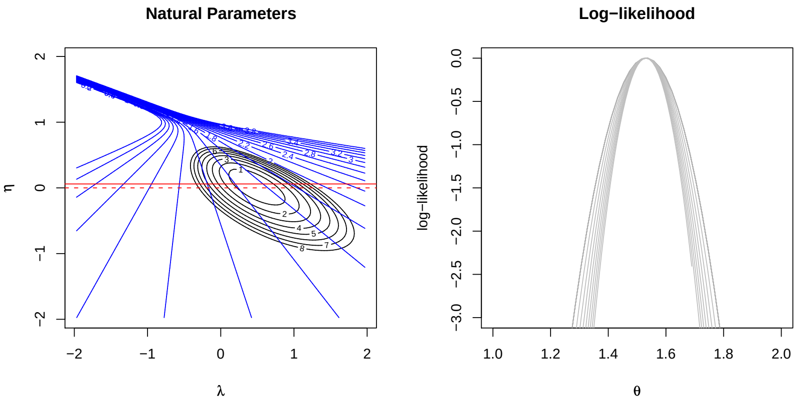

Natural Parameters
η
λ
−2
−1
0
1
2
Log−likelihood
log−likelihood
θ
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
1.0
1.2
1.4
1.6
1.8
2.0

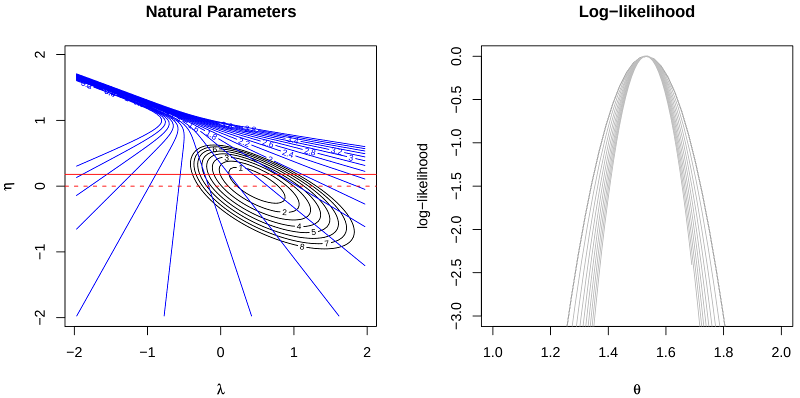

Natural Parameters
η
λ
Log−likelihood
log−likelihood
θ

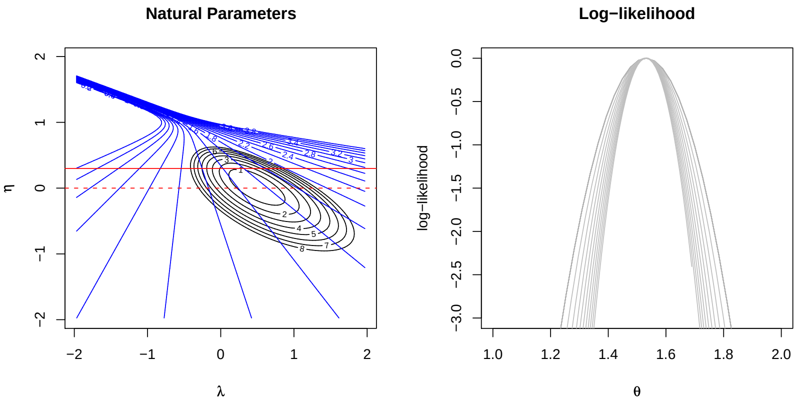

Natural Parameters
η
λ
2
1
0
−1
−2
−2
−1
0
1
2
Log−likelihood
log−likelihood
θ
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
1.0
1.2
1.4
1.6
1.8
2.0

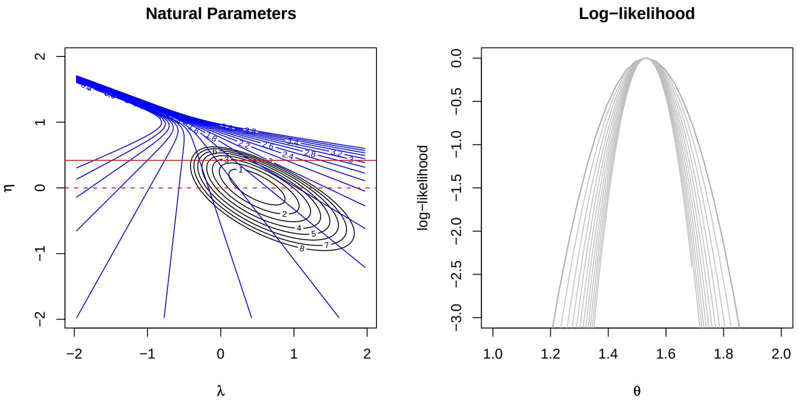

Natural Parameters
η
λ
−2
−1
0
1
2
Log−likelihood
log−likelihood
θ
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
1.0
1.2
1.4
1.6
1.8
2.0

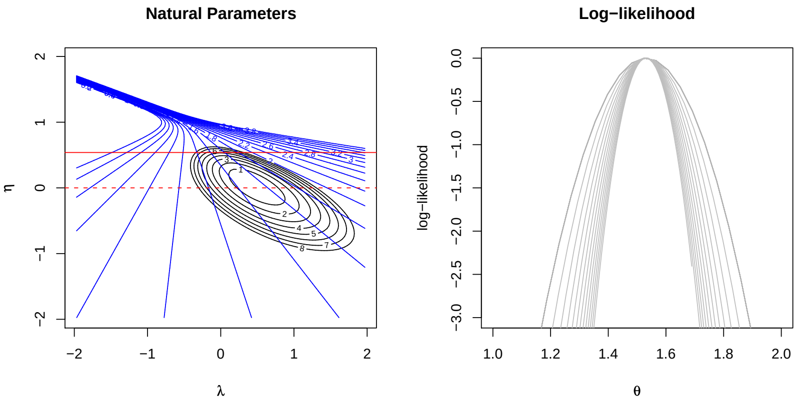

Natural Parameters
η
λ
Log−likelihood
log−likelihood
θ

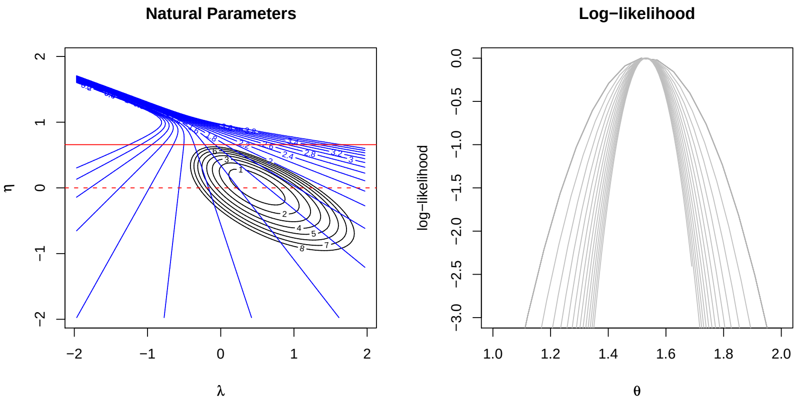

Natural Parameters
η
λ
−2
−1
0
1
2
Log−likelihood
log−likelihood
θ
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
1.0
1.2
1.4
1.6
1.8
2.0

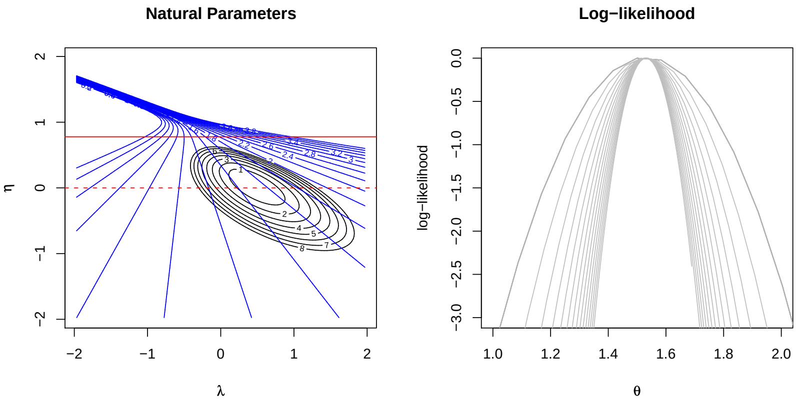

Natural Parameters
η
λ
Log−likelihood
log−likelihood
θ

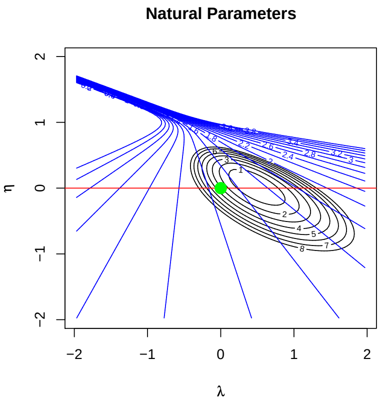

Natural Parameters
η
λ
−2
−1
0
1
2
Log−likelihood
log−likelihood
θ
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
1.0
1.2
1.4
1.6
1.8
2.0

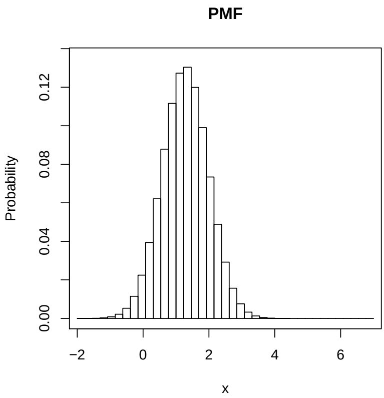

Natural Parameters
η
λ
Log−likelihood
log−likelihood
θ


Natural Parameters
η
λ
2
1
0
−1
−2
Log−likelihood
log−likelihood
θ
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
1.0
1.2
1.4
1.6
1.8
2.0

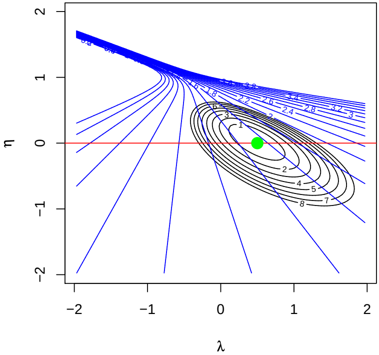

Natural Parameters
Log−likelihood
η
λ
log−likelihood
θ


Natural Parameters
η
λ
Log−likelihood
log−likelihood
θ

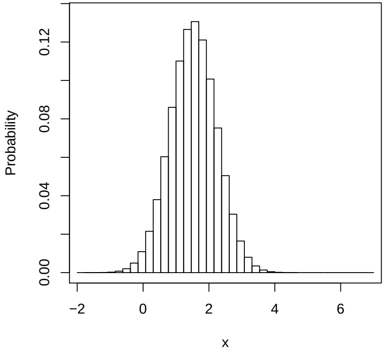

Natural Parameters
η
λ
−2
−1
0
1
2

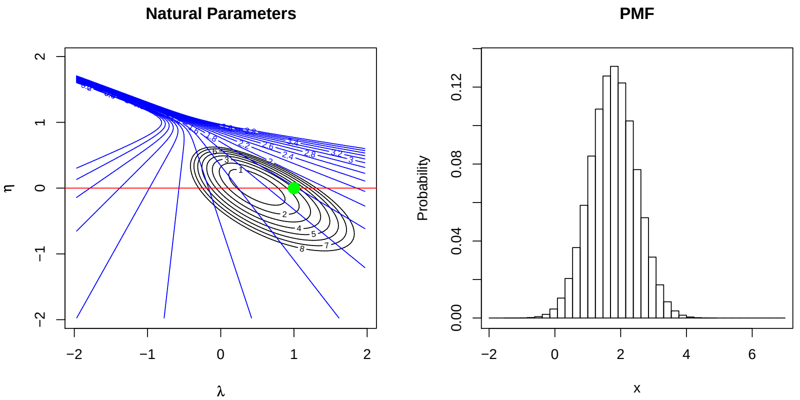

PMF
Probability
x
0.00
0.04
0.08
0.12
−2
0
2
4
6

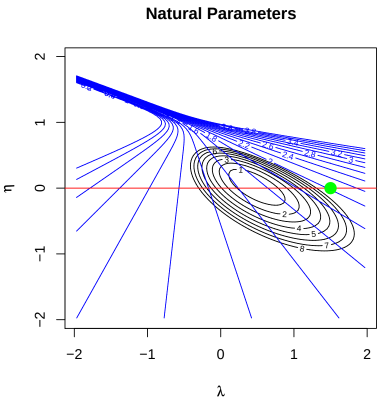
Natural Parameters

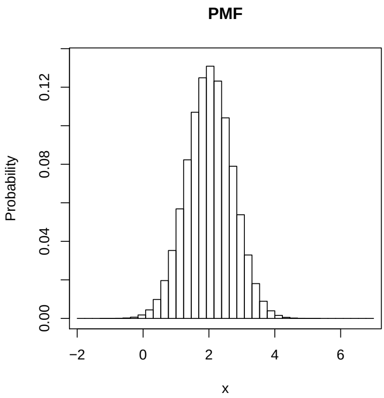


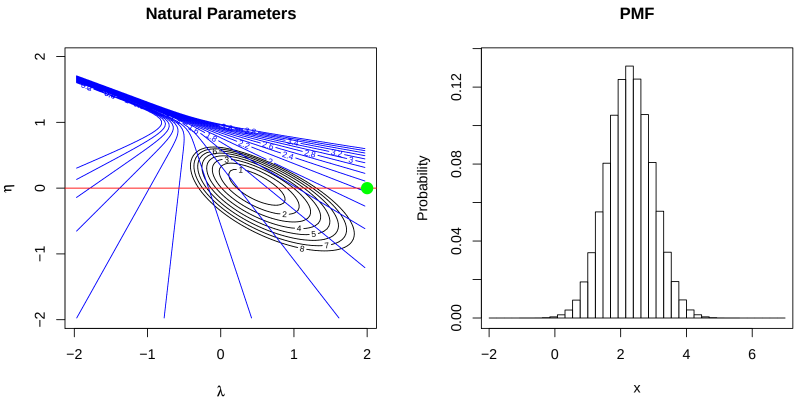
PMF

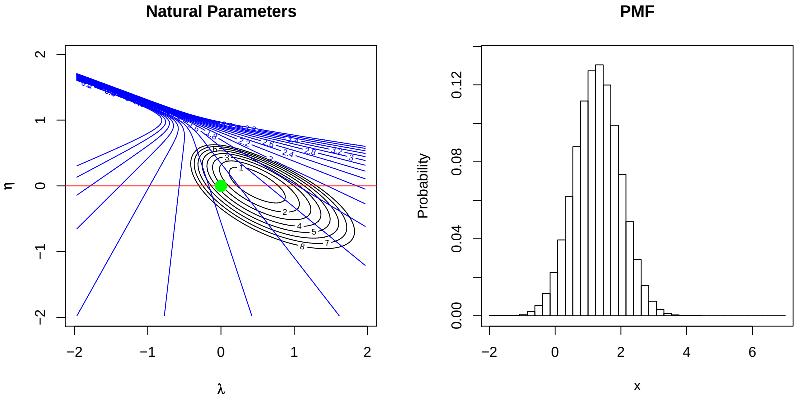


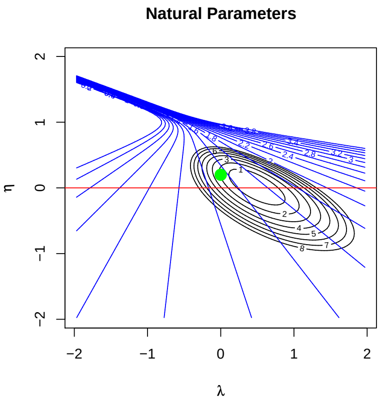

Natural Parameters
PMF
η
λ
Probability
x
−2
−1
0
1
2
0.00
0.04
0.08
0.12
4
6

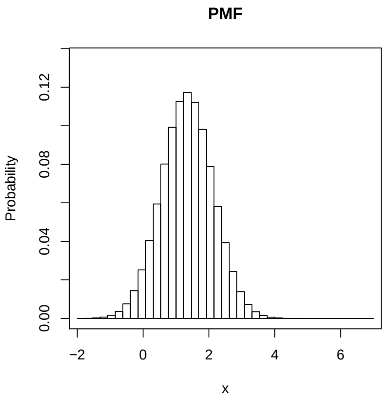

Natural Parameters
η
λ
−2
−1
0
1
2

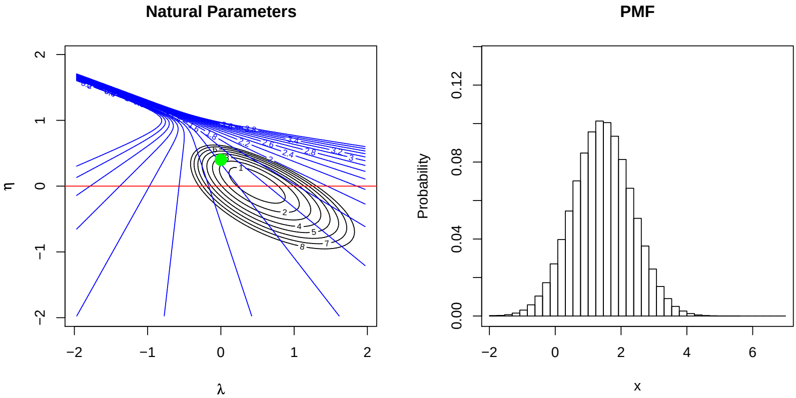

PMF
Probability
x
0.00
0.04
0.08
0.12
−2
0
2
4
6

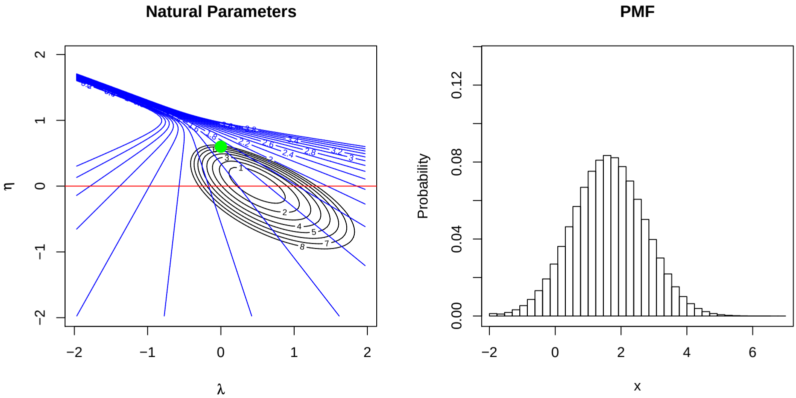

Natural Parameters
PMF
η
λ
Probability
x

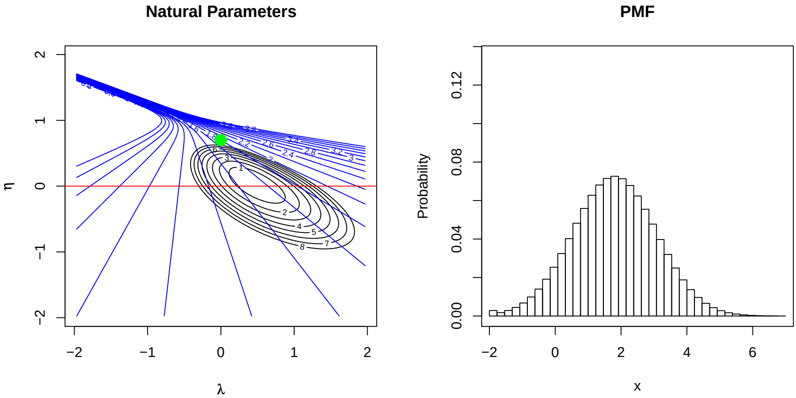

Natural Parameters
η
λ
PMF
Probability
x

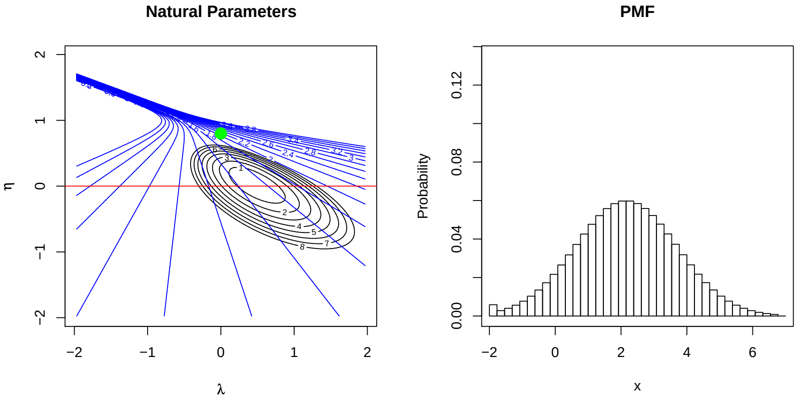

Natural Parameters
η
λ
−2
−1
0
1
2

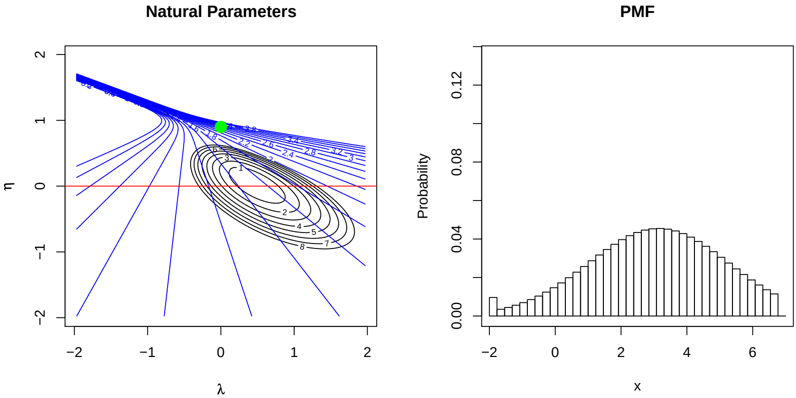

PMF
Probability
x
0.00
0.04
0.08
0.12
−2
0
2
4
6

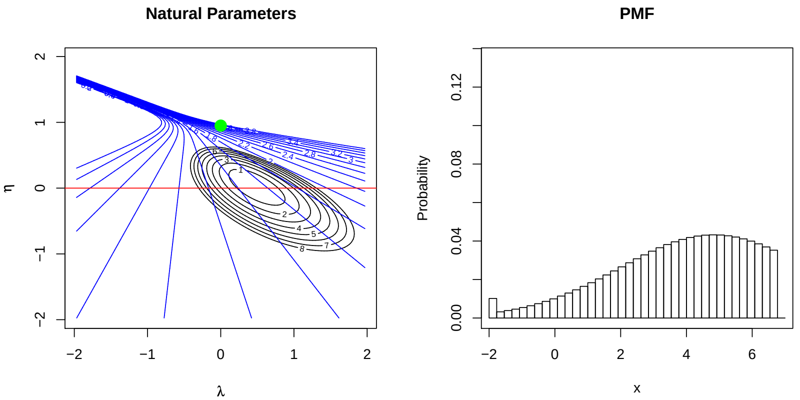

Natural Parameters
η
λ
PMF
Probability
x


Natural Parameters
η
λ
PMF
Probability
x

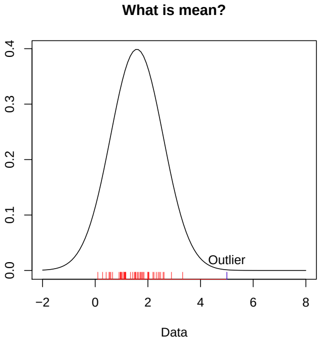

Natural Parameters
η
λ
PMF
Probability
x


Natural Parameters
PMF
η
λ
Probability
x


Natural Parameters
η
λ
−2
−1
0
1
2
PMF
Probability
x
0.00
0.04
0.08
0.12
−2
0
2
4
6

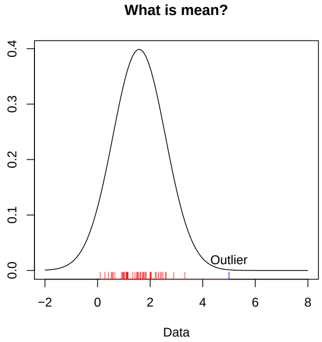

Natural Parameters
η
λ
−2
−1
0
1
2
PMF
Probability
x
0.00
0.04
0.08
0.12
−2
0
2
4
6

Perturbation Space



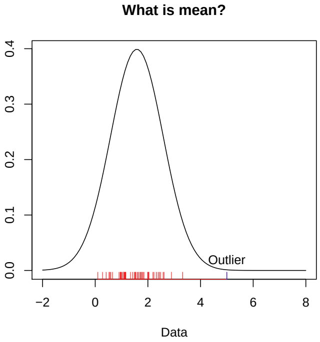
- Data $\setminus$ outlier

Perturbation Space



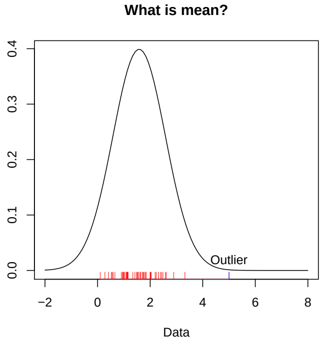
- Data $\setminus$ outlier

Perturbation Space



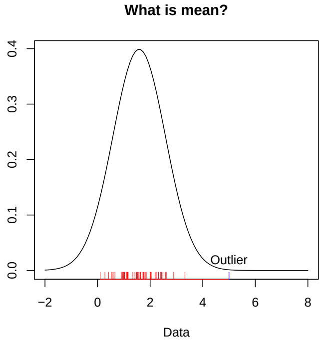
- Data $\setminus$ outlier and base model $N(\mu, 1)$:

Perturbation Space



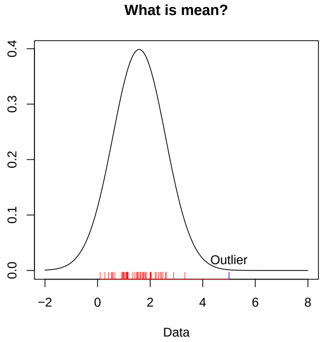
- Data $\setminus$ outlier and base model $N(\mu, 1)$: only one direction is highly sensitive

Perturbation Space



- Data $\setminus$ outlier and base model $N(\mu, 1)$: only one direction is highly sensitive
- This corresponds to changing variance giving model $N(\mu, \sigma^2)$

Perturbation Space



- Data $\setminus$ outlier and base model $N(\mu, 1)$: only one direction is highly sensitive
- This corresponds to changing variance giving model $N(\mu, \sigma^2)$
- If include outlier one more parameter needed

- Overall objective:
  provide tools to help understand *sensitivity to model choice*
- Target:
  applications of Generalised Linear Models
- Delivered via ...
  Computational Information Geometry
  (hidden from the user; available in R)
- Geometric features:
  1. discretisation gives *operational* proxy to universal space
  2. our SEM geometry is affine and convex, *not* manifold based (non-constant: support (dimension) and moment structure)
  3. topology defined via duality

Work supported by EPSRC grant EP/E017878/1

Discretisation $\leftrightarrow$ choice of (1) bins and (2) their labels

Discretisation $\leftrightarrow$ choice of (1) bins and (2) their labels

1. Bins
   - in practice, 'data' $\neq$ 'real numbers' (*cf.* I J Good)

Discretisation $\leftrightarrow$ choice of (1) bins and (2) their labels

1. Bins
   - in practice, ‘data’ $\neq$ ‘real numbers’ (*cf.* I J Good)
   - make bins as fine as possible, *cet. par.*

Discretisation $\leftrightarrow$ choice of (1) bins and (2) their labels

1. Bins
   - in practice, ‘data’ $\neq$ ‘real numbers’ (*cf.* I J Good)
   - make bins as fine as possible, *cet. par.*
     (in the limit, use FMP $:=$ Finite Measurement Precision)

Discretisation $\leftrightarrow$ choice of (1) bins and (2) their labels

1. Bins
   - in practice, 'data' $\neq$ 'real numbers' (*cf*. I J Good)
   - make bins as fine as possible, *cet. par.* (in the limit, use FMP $:=$ Finite Measurement Precision)
2. Labels
   - the finer the bins, the less the choice matters

Discretisation $\leftrightarrow$ choice of (1) bins and (2) their labels

1. Bins
   - in practice, 'data' $\neq$ 'real numbers' (*cf.* I J Good)
   - make bins as fine as possible, *cet. par.*
     (in the limit, use FMP $:=$ Finite Measurement Precision)
2. Labels
   - the finer the bins, the less the choice matters
   - when Q is: 'what is $\theta := E(X)$?',

Discretisation $\leftrightarrow$ choice of (1) bins and (2) their labels

1. Bins
   - in practice, 'data' $\neq$ 'real numbers' (*cf.* I J Good)
   - make bins as fine as possible, *cet. par.*
     (in the limit, use FMP $:=$ Finite Measurement Precision)
2. Labels
   - the finer the bins, the less the choice matters
   - when Q is: 'what is $\theta := E(X)$?',
     key *putative* qu. for a semi-∞ bin $B_\infty = (b, \infty)$, w. loss:

Discretisation $\leftrightarrow$ choice of (1) bins and (2) their labels

1. Bins
   - in practice, 'data' $\neq$ 'real numbers' (*cf.* I J Good)
   - make bins as fine as possible, *cet. par.*
     (in the limit, use FMP := Finite Measurement Precision)
2. Labels
   - the finer the bins, the less the choice matters
   - when Q is: 'what is $\theta := E(X)$?',
     key *putative* qu. for a semi-∞ bin $B_\infty = (b, \infty)$, w. loss:
     'Is $E(X|X \in B_\infty) < \infty$?'

Discretisation $\leftrightarrow$ choice of (1) bins and (2) their labels

1. Bins
   - in practice, 'data' $\neq$ 'real numbers' (*cf.* I J Good)
   - make bins as fine as possible, *cet. par.*
     (in the limit, use FMP $:=$ Finite Measurement Precision)
2. Labels
   - the finer the bins, the less the choice matters
   - when Q is: 'what is $\theta := E(X)$?',
     key *putative* qu. for a semi-∞ bin $B_\infty = (b, \infty)$, w. loss:
     'Is $E(X|X \in B_\infty) < \infty$?'
     - YES $\Rightarrow$ estimate it (under suitable *assumptions*)

Discretisation $\leftrightarrow$ choice of (1) bins and (2) their labels

1. Bins
   - in practice, 'data' $\neq$ 'real numbers' (*cf.* I J Good)
   - make bins as fine as possible, *cet. par.*
     (in the limit, use FMP $:=$ Finite Measurement Precision)
2. Labels
   - the finer the bins, the less the choice matters
   - when Q is: 'what is $\theta := E(X)$?',
     key *putative* qu. for a semi-∞ bin $B_\infty = (b, \infty)$, w. loss:
     'Is $E(X|X \in B_\infty) < \infty$?'
     - YES $\Rightarrow$ estimate it (under suitable *assumptions*)
     - NO $\Rightarrow$ robustness issues! ($\pi_{B_\infty} > 0 \Rightarrow \theta = \infty$)

Effect of model choice

- Problem: Inference for θ, which has a population meaning – e.g. mean, difference in means, regression parameters

Effect of model choice

- Problem: Inference for θ, which has a population meaning – e.g. mean, difference in means, regression parameters
- Note: θ has the *same* meaning in different models

Effect of model choice

- Problem: Inference for θ, which has a population meaning – e.g. mean, difference in means, regression parameters
- Note: θ has the *same* meaning in different models
- What is the effect of model choice on inference for θ?

Effect of model choice

- Problem: Inference for θ, which has a population meaning – e.g. mean, difference in means, regression parameters
- Note: θ has the *same* meaning in different models
- What is the effect of model choice on inference for θ?
- Can we *understand* this by looking at the space of all models?

Affine geometry: mixtures

- Can write simplest mixture $(1-\rho)f(z|\mu_1)+\rho f(z|\mu_2)$ as

$$f(z|\mu_1)+\rho\left\{f(z|\mu_2)-f(z|\mu_1)\right\}=X+V$$

Affine geometry: mixtures

- Can write simplest mixture $(1-\rho)f(z|\mu_1)+\rho f(z|\mu_2)$ as

$$f(z|\mu_1)+\rho\left\{f(z|\mu_2)-f(z|\mu_1)\right\}=X+V$$

- $X \in X_{mix}$ – the set of functions which integrate to 1,
  $V \in V_{mix}$ – the set of functions which integrate to 0

Affine geometry: mixtures

- Can write simplest mixture $(1-\rho)f(z|\mu_1)+\rho f(z|\mu_2)$ as

$$f(z|\mu_1)+\rho\left\{f(z|\mu_2)-f(z|\mu_1)\right\}=X+V$$

- $X \in X_{mix}$ – the set of functions which integrate to 1,
  $V \in V_{mix}$ – the set of functions which integrate to 0
- (X_{mix}, V_{mix}) is an affine space and has an affine geometry

Affine geometry: mixtures

- Can write simplest mixture $(1-\rho)f(z|\mu_1)+\rho f(z|\mu_2)$ as

$$f(z|\mu_1)+\rho\left\{f(z|\mu_2)-f(z|\mu_1)\right\}=X+V$$

- $X \in X_{mix}$ – the set of functions which integrate to 1,
  $V \in V_{mix}$ – the set of functions which integrate to 0
- (X_{mix}, V_{mix}) is an affine space and has an affine geometry
- $f \in X_{mix}$ is not necessary a density,
  since it does not have to be positive

Affine geometry: mixtures

- Can write simplest mixture $(1-\rho)f(z|\mu_1)+\rho f(z|\mu_2)$ as

$$f(z|\mu_1)+\rho\left\{f(z|\mu_2)-f(z|\mu_1)\right\}=X+V$$

- $X \in X_{mix}$ – the set of functions which integrate to 1,
  $V \in V_{mix}$ – the set of functions which integrate to 0
- (X_{mix}, V_{mix}) is an affine space and has an affine geometry
- $f \in X_{mix}$ is not necessary a density,
  since it does not have to be positive
- V_{mix} is vector space
  – and has a translation operator: $X+V \in X_{mix}$

Affine geometry: mixture

- $(X_{mix}, V_{mix}, +)$ is infinite dimensional
  – but there exist natural affine maps to R^n

Affine geometry: mixture

- $(X_{mix}, V_{mix}, +)$ is infinite dimensional
  – but there exist natural affine maps to R^n
- This allows us to draw meaningful pictures in R^n

Affine geometry: mixture

- $(X_{mix}, V_{mix}, +)$ is infinite dimensional
  – but there exist natural affine maps to R^n
- This allows us to draw meaningful pictures in R^n
- The affine line segment has two types of boundaries:
  1. soft boundary where $\rho = 0, 1$

$$(1-\rho)f(z|\mu_1) + \rho f(z|\mu_2)$$

Affine geometry: mixture

- $(X_{mix}, V_{mix}, +)$ is infinite dimensional
  – but there exist natural affine maps to R^n
- This allows us to draw meaningful pictures in R^n
- The affine line segment has two types of boundaries:
  1. soft boundary where $\rho = 0, 1$

$$(1-\rho)f(z|\mu_1) + \rho f(z|\mu_2)$$

  2. hard boundary where

$$f(z|\mu_1) + \rho\left[(f(z|\mu_2) - f(z|\mu_1)\right]$$

  stops being positive for all z

Affine geometry: mixture

- Space of distributions is a convex subset of an affine space $(X_{mix}, V_{mix}, +)$

Affine geometry: mixture

- Space of distributions is a convex subset of an affine space $(X_{mix}, V_{mix}, +)$
- Example would be all distributions on $k+1$ categories

Affine geometry: mixture

- Space of distributions is a convex subset of an affine space $(X_{mix}, V_{mix}, +)$
- Example would be all distributions on $k+1$ categories
- Affine space is R^k

Affine geometry: mixture

- Space of distributions is a convex subset of an affine space $(X_{mix}, V_{mix}, +)$
- Example would be all distributions on $k + 1$ categories
- Affine space is R^k
- Convex subset is a simplex

Affine geometry: mixture

- Space of distributions is a convex subset of an affine space $(X_{mix}, V_{mix}, +)$
- Example would be all distributions on $k+1$ categories
- Affine space is R^k
- Convex subset is a simplex

Affine geometry: exponential

- For any chosen support set $S \subseteq Z$,

Affine geometry: exponential

- For any chosen support set $S \subseteq Z$, there is an affine space $\mathrm{Aff}(S)$ such that

Affine geometry: exponential

- For any chosen support set $S \subseteq Z$,
  there is an affine space $\text{Aff}(S)$ such that
  finite dimnl. affine subspaces are exponential families

Affine geometry: exponential

- For any chosen support set $S \subseteq Z$, there is an affine space $\mathrm{Aff}(S)$ such that finite dimnl. affine subspaces are exponential families
- We need to glue these affine spaces together, if we want the support set to be more general

Affine geometry: exponential

- For any chosen support set $S \subseteq Z$,
  there is an affine space $\mathrm{Aff}(S)$ such that
  finite dimnl. affine subspaces are exponential families
- We need to glue these affine spaces together,
  if we want the support set to be more general
- This gives extended (generalised) exponential families,
  which are the closure of regular exponential families

Perturbations of Models

A Model is a low dimensional subset of the SEM space.

Perturbations of Models

A Model is a low dimensional subset of the SEM space.

We can perturb by:

- Translations $\cdots$ look at via 'cuts'

Perturbations of Models

A Model is a low dimensional subset of the SEM space.

We can perturb by:

- Translations $\cdots$ look at via ‘cuts’
- Rotations another story $\cdots$ bias, variance trade-off $\cdots$

Perturbations of Models

A Model is a low dimensional subset of the SEM space.

We can perturb by:

- Translations $\cdots$ look at via ‘cuts’
- Rotations another story $\cdots$ bias, variance trade-off $\cdots$
- Adding nuisance parameters $\cdots$

Perturbations of Models

A Model is a low dimensional subset of the SEM space.

We can perturb by:

- Translations $\cdots$ look at via 'cuts'
- Rotations another story $\cdots$ bias, variance trade-off $\cdots$
- Adding nuisance parameters $\cdots$
- Down-weight/delete outliers $\cdots$

Perturbations of Models

A Model is a low dimensional subset of the SEM space.

We can perturb by:

- Translations $\cdots$ look at via 'cuts'
- Rotations another story $\cdots$ bias, variance trade-off $\cdots$
- Adding nuisance parameters $\cdots$
- Down-weight/delete outliers $\cdots$

OVERALL:
Explore which choices affect inference for interest parameter

Information directions

- We have information about the cell probabilities where we have data

Information directions

- We have information about the cell probabilities where we have data
- We have 'no' information about the cell probabilities where the counts are zero

Information directions

- We have information about the cell probabilities where we have data
- We have ‘no’ information about the cell probabilities where the counts are zero
- Hard inference problem: inference about Binomial probability when observed count is zero.

Information directions

- We have information about the cell probabilities where we have data
- We have 'no' information about the cell probabilities where the counts are zero
- Hard inference problem: inference about Binomial probability when observed count is zero.
- In model space there are different 'directions' that have different effects on θ-inference