Testing with mixtures

joint work with
K. Kamari, K. Mengersen and C. Robert
april 2016, Toulouse

Testing issues

Hypothesis testing

- central problem of statistical inference
- witness the recent ASA's statement on *p*-values (Wasserstein, 2016)
- dramatically differentiating feature between classical and Bayesian paradigms
- wide open to controversy and divergent opinions, includ. within the Bayesian community
- non-informative Bayesian testing case mostly unresolved, witness the Jeffreys–Lindley paradox

Berger (2003), Mayo & Cox (2006), Gelman (2008)

𝔅

Bayesian modelling

- **Standard Bayesian approach to testing :** consider two families of models, one for each of the hypotheses under comparison,

$$\mathfrak{M}_1 : \; x \sim f_1(x|\theta_1)\,, \; \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : \; x \sim f_2(x|\theta_2)\,, \; \theta_2 \in \Theta_2\,,$$

$$\textbf{Priors} \;\; \theta_1 \sim \pi_1(\theta_1) \quad \text{and} \quad \theta_2 \sim \pi_2(\theta_2)\,,$$

$$m_1(x) = \int_{\Theta_1} f_1(x|\theta_1)\, \pi_1(\theta_1)\, d\theta_1 \quad \text{and} \quad m_2(x) = \int_{\Theta_2} f_2(x|\theta_2)\, \pi_1(\theta_2)\, d\theta_2$$

▶ **Standard Bayesian approach to testing :** consider two families of models, one for each of the hypotheses under comparison,

$$\mathfrak{M}_1 : \; x \sim f_1(x|\theta_1)\,, \; \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : \; x \sim f_2(x|\theta_2)\,, \; \theta_2 \in \Theta_2\,,$$

$$\textbf{Priors} \;\; \theta_1 \sim \pi_1(\theta_1) \quad \text{and} \quad \theta_2 \sim \pi_2(\theta_2)\,,$$

$$m_1(x) = \int_{\Theta_1} f_1(x|\theta_1)\,\pi_1(\theta_1)\,d\theta_1 \quad \text{and} \quad m_2(x) = \int_{\Theta_2} f_2(x|\theta_2)\,\pi_1(\theta_2)\,d\theta_2$$

Bayesian modelling

▶ **Standard Bayesian approach to testing :** consider two families of models, one for each of the hypotheses under comparison,

$$\mathfrak{M}_1 : \; x \sim f_1(x|\theta_1)\,,\; \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : \; x \sim f_2(x|\theta_2)\,,\; \theta_2 \in \Theta_2\,,$$

$$\textbf{Priors} \;\; \theta_1 \sim \pi_1(\theta_1) \quad \text{and} \quad \theta_2 \sim \pi_2(\theta_2)\,,$$

$$m_1(x) = \int_{\Theta_1} f_1(x|\theta_1)\,\pi_1(\theta_1)\,d\theta_1 \quad \text{and} \quad m_2(x) = \int_{\Theta_2} f_2(x|\theta_2)\,\pi_1(\theta_2)\,d\theta_2$$

either through *Bayes factor* or posterior probability, respectively :

$$\mathfrak{B}_{12} = \frac{m_1(x)}{m_2(x)}, \quad \mathbb{P}(\mathfrak{M}_1|x) = \frac{\omega_1 m_1(x)}{\omega_1 m_1(x) + \omega_2 m_2(x)};$$

the latter depends on the prior weights $\omega_i = \pi(\Theta_i)$

Bayesian decision

Bayesian decision step

- comparing Bayes factor $\mathfrak{B}_{12}$ with threshold value of one or
- comparing posterior probability $\mathbb{P}(\mathfrak{M}_1|x)$ with bound 1/2

Bayesian decision

Bayesian decision step

- comparing Bayes factor $\mathfrak{B}_{12}$ with threshold value of one or
- comparing posterior probability $\mathbb{P}(\mathfrak{M}_1|x)$ with bound $1/2$

When comparing more than two models, model with *highest posterior probability* is the one selected, but highly dependent on the prior modelling.

Bayes factor : self-contained concept

Outside decision-theoretic environment :

- eliminates choice of $\pi(\Theta_0)$

Bayes factor : self-contained concept

Outside decision-theoretic environment :

- eliminates choice of $\pi(\Theta_0)$
- but depends on the choice of (π_0, π_1)

Bayes factor : self-contained concept

Outside decision-theoretic environment :

- eliminates choice of $\pi(\Theta_0)$
- but depends on the choice of (π_0, π_1)
- Bayesian/marginal equivalent to the likelihood ratio

Bayes factor : self-contained concept

Outside decision-theoretic environment :

- eliminates choice of $\pi(\Theta_0)$
- but depends on the choice of (π_0, π_1)
- Bayesian/marginal equivalent to the likelihood ratio
- Jeffreys' scale of evidence :
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ between 0 and 0.5, evidence against H_0 *weak*,
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ 0.5 and 1, evidence *substantial*,
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ 1 and 2, evidence *strong* and
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ above 2, evidence *decisive*

  Quite arbitrary really ! : consequence of 0-1 loss function

Bayesian testing of hypotheses

- Bayesian model selection as comparison of k potential statistical models towards the selection of model that fits the data “best"
- mostly accepted perspective : it does not primarily seek to identify which model is “true", but compares fits

Bayesian testing of hypotheses

- Bayesian model selection as comparison of k potential statistical models towards the selection of model that fits the data “best"
- mostly accepted perspective : it does not primarily seek to identify which model is “true", but compares fits
- tools like Bayes factor naturally include a penalisation addressing model complexity, mimicked by Bayes Information (BIC) and Deviance Information (DIC) criteria .
- Under quite genreal conditions : consistent criterion for testing or model selection

Some difficulties

- long-lasting impact of prior modeling, i.e., choice of prior distributions on parameters of both models, despite overall consistency proof for Bayes factor
- discontinuity in valid use of improper priors since they are not justified in most testing situations, leading to many alternative and *ad hoc* solutions, where data is either used twice or split in artificial ways [or further tortured into confession]
- binary (*accept* vs. *reject*) outcome more suited for immediate decision (if any) than for model evaluation, in connection with rudimentary loss function $0-1$ [atavistic remain of Neyman-Pearson formalism]

Some additional difficulties

- related impossibility to ascertain simultaneous misfit or to detect outliers
- no assessment of uncertainty associated with decision itself besides posterior probability
- difficult computation of marginal likelihoods in most settings with further controversies about which algorithm to adopt
- time for a paradigm shift ?

Paradigm shift

New proposal for a paradigm shift (!) in the Bayesian processing of hypothesis testing and of model selection

- convergent and naturally interpretable solution
- more extended use of improper priors

Paradigm shift

New proposal for a paradigm shift (!) in the Bayesian processing of hypothesis testing and of model selection

- convergent and naturally interpretable solution
- more extended use of improper priors

Simple representation of the testing problem as a two-component mixture estimation problem where the weights are formally equal to 0 *or* 1

Paradigm shift

Simple representation of the testing problem as a two-component mixture estimation problem where the weights are formally equal to 0 *or* 1

- Approach inspired from consistency result of Rousseau and Mengersen (2011) on estimated overfitting mixtures
- Mixture representation not directly equivalent to the use of a posterior probability
- Calibration of posterior distribution of the weight of a model, moving from the notion of posterior probability of a model

Encompassing mixture model

Idea : Given two statistical models,

$$\mathfrak{M}_1 : x_i \stackrel{ind.}{\sim} f_1(x_i|\theta_1)\,,\ \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : x_i \stackrel{ind.}{\sim} f_2(x|\theta_2)\,,\ \theta_2 \in \Theta_2\,, i \leq n$$

Encompassing mixture model

Idea : Given two statistical models,

$$\mathfrak{M}_1 : x_i \overset{ind.}{\sim} f_1(x_i|\theta_1)\,,\ \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : x_i \overset{ind.}{\sim} f_2(x|\theta_2)\,,\ \theta_2 \in \Theta_2\,, i \leq n$$

embed both within an encompassing mixture

$$\mathfrak{M}_\alpha : x_i \overset{ind.}{\sim} \alpha f_1(x|\theta_1) + (1-\alpha) f_2(x|\theta_2)\,,\ 0 \leq \alpha \leq 1, \quad i \leq n \qquad (1)$$

Encompassing mixture model

Idea : Given two statistical models,

$$\mathfrak{M}_1 : x_i \overset{ind.}{\sim} f_1(x_i|\theta_1)\,,\ \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : x_i \overset{ind.}{\sim} f_2(x|\theta_2)\,,\ \theta_2 \in \Theta_2\,, i \leq n$$

embed both within an encompassing mixture

$$\mathfrak{M}_\alpha : \ x_i \overset{ind.}{\sim} \alpha f_1(x|\theta_1) + (1-\alpha) f_2(x|\theta_2)\,,\ 0 \leq \alpha \leq 1, \quad i \leq n \qquad (1)$$

Note : Both models correspond to special cases of (1), one for $\alpha = 1$ and one for $\alpha = 0$
Draw inference on mixture representation (1), as if each observation was individually and independently produced by the mixture model

Inferential motivations

Sounds like approximation to the real model, but several definitive advantages to this paradigm shift :

- Bayes estimate of the weight α replaces posterior probability of model $\mathfrak{M}_1$, equally convergent indicator of which model is "true", while avoiding artificial prior probabilities on model indices, ω_1 and ω_2
- interpretation of estimator of α at least as natural as handling the posterior probability, while avoiding zero-one loss setting : proportion of individuals from each model
- α and its posterior distribution provide measure of proximity to the models, while being interpretable as data propensity to stand within one model
- further allows for alternative perspectives on testing and model choice, like predictive tools, cross-validation, and information indices

Computational motivations

- avoids highly problematic computations of the marginal likelihoods, since standard algorithms are available for Bayesian mixture estimation
- straightforward extension to a finite collection of models, with a larger number of components, which considers all models at once and eliminates least likely models by simulation
- eliminates difficulty of label switching that plagues both Bayesian estimation and Bayesian computation, since components are no longer exchangeable
- posterior distribution of α evaluates more thoroughly strength of support for a given model than the single figure outcome of a posterior probability
- variability of posterior distribution on α allows for a more thorough assessment of the strength of this support

Noninformative motivations

- additional feature missing from traditional Bayesian answers : a mixture model acknowledges possibility that, for a finite dataset, *both* models or *none* could be acceptable
- standard (proper and informative) prior modeling can be reproduced in this setting, but non-informative (improper) priors also are manageable therein, provided both models first reparameterised towards shared parameters, e.g. location and scale parameters
- in special case when all parameters are common

$$\mathfrak{M}_\alpha : \; x \sim \alpha f_1(x|\theta) + (1-\alpha) f_2(x|\theta) \, , 0 \le \alpha \le 1$$

if θ is a location parameter, a flat prior $\pi(\theta) \propto 1$ is available

Weakly informative motivations

- using the *same* parameters or some *identical* parameters on both components highlights that opposition between the two components is not an issue of enjoying different parameters
- those common parameters are nuisance parameters, to be integrated out *[unlike Lindley's paradox]*
- prior model weights ω_i rarely discussed in classical Bayesian approach, even though linear impact on posterior probabilities. Here, prior modeling only involves selecting a prior on α, e.g., $\alpha \sim \mathscr{B}(a_0, a_0)$
- while a_0 impacts posterior on α, it always leads to mass accumulation near 1 or 0, i.e. favours most likely model
- sensitivity analysis straightforward to carry
- approach easily calibrated by parametric boostrap providing reference posterior of α under each model
- natural Metropolis–Hastings alternative

Poisson/Geometric

- choice betwen Poisson $\mathscr{P}(\lambda)$ and Geometric $\mathscr{G}eo(p)$ distribution

Poisson/Geometric

- choice betwen Poisson $\mathscr{P}(\lambda)$ and Geometric $\mathscr{G}eo(p)$ distribution
- mixture with common parameter λ

$$\mathfrak{M}_\alpha : \ \alpha\mathscr{P}(\lambda) + (1-\alpha)\mathscr{G}eo(1/1+\lambda)$$

Allows for Jeffreys prior since resulting posterior is proper

Poisson/Geometric

- choice betwen Poisson $\mathscr{P}(\lambda)$ and Geometric $\mathscr{G}eo(p)$ distribution
- mixture with common parameter λ

$$\mathfrak{M}_\alpha : \ \alpha\mathscr{P}(\lambda) + (1-\alpha)\mathscr{G}eo({}^1/_{1+\lambda})$$

  Allows for Jeffreys prior since resulting posterior is proper
- independent Metropolis–within–Gibbs with proposal distribution on λ equal to Poisson posterior (with acceptance rate larger than 75%)

Beta prior

When $\alpha \sim \mathscr{B}e(a_0, a_0)$ prior, full conditional posterior

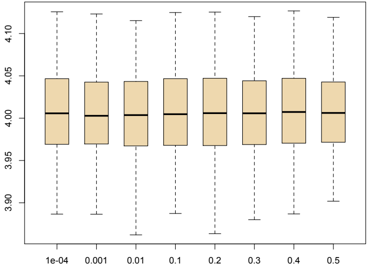
$$\alpha \sim \mathscr{B}e(n_1(\zeta) + a_0, n_2(\zeta) + a_0)$$

Exact Bayes factor opposing Poisson and Geometric

$$\mathfrak{B}_{12} = n^{n\bar{x}_n} \prod_{i=1}^{n} x_i! \, \Gamma\left(n + 2 + \sum_{i=1}^{n} x_i\right) \Big/ \Gamma(n+2)$$

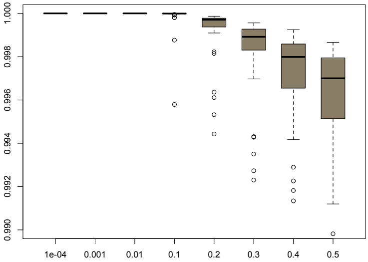
although arbitrary from a purely mathematical viewpoint

Parameter estimation : λ then α



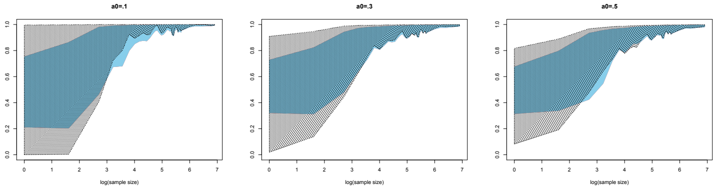
Posterior means of λ and medians of α for 100 Poisson $\mathscr{P}(4)$ datasets of size $n = 1000$, for $a_0 = .0001, .001, .01, .1, .2, .3, .4, .5$. Each posterior approximation is based on 10^4 Metropolis-Hastings iterations.

Parameter estimation : λ then α



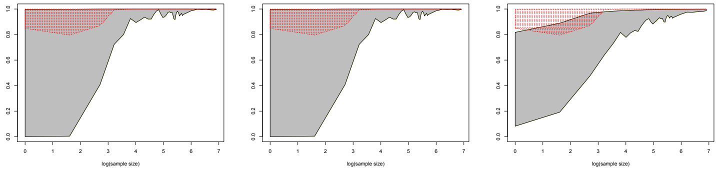
Posterior means of λ and medians of α for 100 Poisson $\mathscr{P}(4)$ datasets of size $n = 1000$, for $a_0 = .0001, .001, .01, .1, .2, .3, .4, .5$. Each posterior approximation is based on 10^4 Metropolis-Hastings iterations.

Consistency



Posterior means *(sky-blue)* and medians *(grey-dotted)* of α, over 100 Poisson $\mathscr{P}(4)$ datasets for sample sizes from 1 to 1000.

Behaviour of Bayes factor



Comparison between $\mathbb{P}(\mathfrak{M}_1|x)$ *(red dotted area)* and posterior medians of α *(grey zone)* for 100 Poisson $\mathscr{P}(4)$ datasets with sample sizes n between 1 and 1000, for $a_0 = .001, .1, .5$

Normal-normal comparison

- comparison of a normal $\mathscr{N}(\theta_1, 1)$ with a normal $\mathscr{N}(\theta_2, 2)$ distribution

Normal-normal comparison

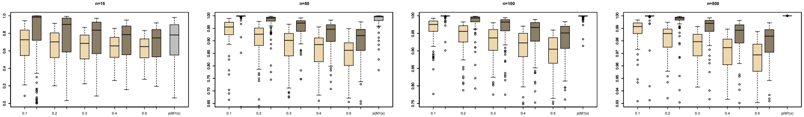
- comparison of a normal $\mathscr{N}(\theta_1, 1)$ with a normal $\mathscr{N}(\theta_2, 2)$ distribution
- mixture with identical location parameter θ
  $\alpha\mathscr{N}(\theta, 1) + (1-\alpha)\mathscr{N}(\theta, 2)$
- Jeffreys prior $\pi(\theta) = 1$ can be used, since posterior is proper

Normal-normal comparison

- comparison of a normal $\mathscr{N}(\theta_1, 1)$ with a normal $\mathscr{N}(\theta_2, 2)$ distribution
- mixture with identical location parameter θ
  $\alpha\mathscr{N}(\theta, 1) + (1-\alpha)\mathscr{N}(\theta, 2)$
- Jeffreys prior $\pi(\theta) = 1$ can be used, since posterior is proper
- Reference (improper) Bayes factor

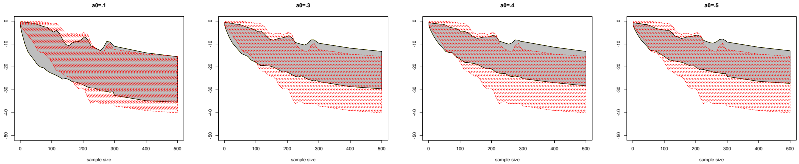
$$\mathfrak{B}_{12} = 2^{n-1/2} \Big/ \exp {}^{1}/_{4} \sum_{i=1}^{n} (x_i - \bar{x})^2 \,,$$

Consistency



Posterior means *(wheat)* and medians of α *(dark wheat)*, compared with posterior probabilities of $\mathfrak{M}_0$ *(gray)* for a $\mathscr{N}(0,1)$ sample, derived from 100 datasets for sample sizes equal to $15, 50, 100, 500$. Each posterior approximation is based on 10^4 MCMC iterations.

Comparison with posterior probability



Plots of ranges of $\log(n)\log(1-\mathbb{E}[\alpha|x])$ *(gray color)* and $\log(1-p(\mathfrak{M}_1|x))$ *(red dotted)* over 100 $\mathscr{N}(0,1)$ samples as sample size n grows from 1 to 500. and α is the weight of $\mathscr{N}(0,1)$ in the mixture model. The shaded areas indicate the range of the estimations and each plot is based on a Beta prior with $a_0 = .1, .2, .3, .4, .5, 1$ and each posterior approximation is based on 10^4 iterations.

Comments

- convergence to one boundary value as sample size n grows
- impact of hyperarameter a_0 slowly vanishes as n increases, but present for moderate sample sizes
- when simulated sample is neither from $\mathscr{N}(\theta_1, 1)$ nor from $\mathscr{N}(\theta_2, 2)$, behaviour of posterior varies, depending on which distribution is closest

Logit or Probit ?

- binary dataset, R dataset about diabetes in 200 Pima Indian women with body mass index as explanatory variable
- comparison of logit and probit fits could be suitable. We are thus comparing both fits via our method

$$\mathfrak{M}_1 : y_i \mid \mathbf{x}^i, \theta_1 \sim \mathscr{B}(1, p_i) \quad \text{where} \quad p_i = \frac{\exp(\mathbf{x}^i\theta_1)}{1+\exp(\mathbf{x}^i\theta_1)}$$

$$\mathfrak{M}_2 : y_i \mid \mathbf{x}^i, \theta_2 \sim \mathscr{B}(1, q_i) \quad \text{where} \quad q_i = \Phi(\mathbf{x}^i\theta_2)$$

Common parameterisation

Local reparameterisation strategy that rescales parameters of the probit model $\mathfrak{M}_2$ so that the MLE's of both models coincide. **Choudhuty et al., 2007**

$$\Phi(\mathbf{x}^i\theta_2) \approx \frac{\exp(k\mathbf{x}^i\theta_2)}{1+\exp(k\mathbf{x}^i\theta_2)}$$

and use best estimate of k to bring both parameters into coherency

$$(k_0, k_1) = (\widehat{\theta_{01}}/\widehat{\theta_{02}}, \widehat{\theta_{11}}/\widehat{\theta_{12}})\,,$$

reparameterise $\mathfrak{M}_1$ and $\mathfrak{M}_2$ as

$$\mathfrak{M}_1 : y_i \mid \mathbf{x}^i, \theta \sim \mathscr{B}(1, p_i) \quad \text{where} \quad p_i = \frac{\exp(\mathbf{x}^i\theta)}{1+\exp(\mathbf{x}^i\theta)}$$

$$\mathfrak{M}_2 : y_i \mid \mathbf{x}^i, \theta \sim \mathscr{B}(1, q_i) \quad \text{where} \quad q_i = \Phi(\mathbf{x}^i(\kappa^{-1}\theta))\,,$$

with $\kappa^{-1}\theta = (\theta_0/k_0, \theta_1/k_1)$.

Prior modelling

Under default g-prior

$$\theta \sim \mathscr{N}_2(0, n(X^\mathsf{T} X)^{-1})$$

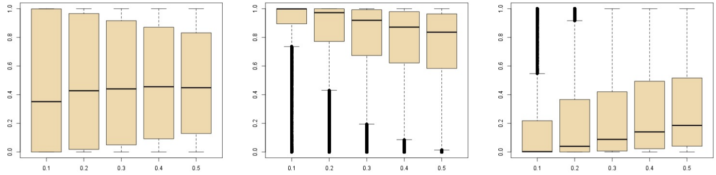
full conditional posterior distributions given allocations

$$\pi(\theta \mid \mathbf{y}, X, \zeta) \propto \frac{\exp\left\{\sum_i \mathbb{I}_{\zeta_i=1} y_i \mathbf{x}^i \theta\right\}}{\prod_{i;\zeta_i=1}[1+\exp(\mathbf{x}^i\theta)]} \exp\left\{-\theta^T(X^T X)\theta/2n\right\}$$
$$\times \prod_{i;\zeta_i=2} \Phi(\mathbf{x}^i(\kappa^{-1}\theta))^{y_i}(1-\Phi(\mathbf{x}^i(\kappa^{-1}\theta)))^{(1-y_i)}$$

hence posterior distribution clearly defined

Results

		Logistic		Probit	
a_0	α	θ_0	θ_1	$\frac{\theta_0}{k_0}$	$\frac{\theta_1}{k_1}$
.1	.352	-4.06	.103	-2.51	.064
.2	.427	-4.03	.103	-2.49	.064
.3	.440	-4.02	.102	-2.49	.063
.4	.456	-4.01	.102	-2.48	.063
.5	.449	-4.05	.103	-2.51	.064



Histograms of posteriors of α in favour of logistic model where $a_0 = .1, .2, .3, .4, .5$ for (a) Pima dataset, (b) Data from logistic model, (c) Data from probit model

Survival analysis

Testing hypothesis that data comes from a

1. log-Normal(ϕ, κ^2),
2. Weibull(α, λ), or
3. log-Logistic(γ, δ)

distribution

Testing hypothesis that data comes from a

1. log-Normal(ϕ, κ^2),
2. Weibull(α, λ), or
3. log-Logistic(γ, δ)

distribution

Corresponding mixture given by the density

$$\begin{aligned} \alpha_1 \exp\{-(\log x - \phi)^2/2\kappa^2\}/\sqrt{2\pi} x\kappa + \\ \alpha_2 \frac{\alpha}{\lambda} \exp\{-(x/\lambda)^\alpha\}((x/\lambda)^{\alpha-1} + \\ \alpha_3(\delta/\gamma)(x/\gamma)^{\delta-1}/(1+(x/\gamma)^\delta)^2 \end{aligned}$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$

Reparameterisation

Looking for common parameter(s) :

$$\phi = \mu + \gamma\beta = \xi$$
$$\sigma^2 = \pi^2\beta^2/6 = \zeta^2\pi^2/3$$

where $\gamma \approx 0.5772$ is Euler-Mascheroni constant.

Reparameterisation

Looking for common parameter(s) :

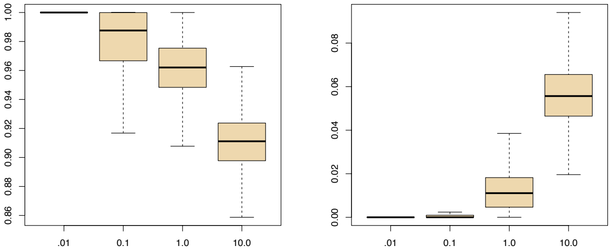
$$\phi = \mu + \gamma\beta = \xi$$
$$\sigma^2 = \pi^2\beta^2/6 = \zeta^2\pi^2/3$$

where $\gamma \approx 0.5772$ is Euler-Mascheroni constant.

Allows for a noninformative prior on the common location scale parameter,

$$\pi(\phi, \sigma^2) = 1/\sigma^2$$

Recovery



Boxplots of the posterior distributions of the Normal weight α_1 under the two scenarii : truth = Normal *(left panel)*, truth = Gumbel *(right panel)*, a_0=0.01, 0.1, 1.0, 10.0 *(from left to right in each panel)* and $n = 10,000$ simulated observations.

Asymptotic consistency

Posterior consistency holds for mixture testing procedure [under minor conditions]

Asymptotic consistency

Posterior consistency holds for mixture testing procedure [under minor conditions]

Two different cases

- the two models, $\mathfrak{M}_1$ and $\mathfrak{M}_2$, are well separated
- model $\mathfrak{M}_1$ is a submodel of $\mathfrak{M}_2$.

Let π be the prior and $\mathbf{x}^n = (x_1, \cdots, x_n)$ a sample with true density f^*

proposition

Assume that, for all $c > 0$, there exist $\Theta_n \subset \Theta_1 \times \Theta_2$ and $B > 0$ such that

$$\pi\left[\Theta_n^c\right] \leq n^{-c}, \quad \Theta_n \subset \{\|\theta_1\| + \|\theta_2\| \leq n^B\}$$

and that there exist $H \geq 0$ and $L, \delta > 0$ such that, for $j = 1, 2$,

$$\sup_{\theta,\theta' \in \Theta_n} \|f_{j,\theta_j} - f_{j,\theta_j'}\|_1 \leq L n^H \|\theta_j - \theta_j'\|, \quad \theta = (\theta_1, \theta_2),\, \theta' = (\theta_1', \theta_2'),$$

$$\forall \|\theta_j - \theta_j^*\| \leq \delta; \quad KL(f_{j,\theta_j}, f_{j,\theta_j^*}) \lesssim \|\theta_j - \theta_j^*\|.$$

Then, when $f^* = f_{\theta^*,\alpha^*}$, with $\alpha^* \in [0,1]$, there exists $M > 0$ such that

$$\pi\left[(\alpha,\theta); \|f_{\theta,\alpha} - f^*\|_1 > M\sqrt{\log n/n} | \mathbf{x}^n\right] = o_p(1).$$

II. Recovery of the parameters : Separated models – $f_{\theta,\alpha} = \alpha f_{1,\theta_1} + (1-\alpha) f_{2,\theta_2}$

Assumption : Models are separated, i.e. identifiability holds :

$$\forall \alpha, \alpha' \in (0,1), \quad \forall \theta_j, \theta_j^{'}, j = 1,2 \quad f_{\theta,\alpha} = f_{\theta^{'},\alpha^{'}} \quad \Rightarrow \alpha = \alpha^{'}, \quad \theta = \theta^{'}$$

Further

$$\inf_{\theta_1 \in \Theta_1} \inf_{\theta_2 \in \Theta_2} \|f_{1,\theta_1} - f_{2,\theta_2}\|_1 > 0$$

and, for $\theta_j^* \in \Theta_j$, if P_{θ_j} weakly converges to $P_{\theta_j^*}$, then

$$\theta_j \longrightarrow \theta_j^*$$

in the Euclidean topology

II. Recovery of the parameters : Separated models – $f_{\theta,\alpha} = \alpha f_{1,\theta_1} + (1-\alpha) f_{2,\theta_2}$

Assumption : Models are separated, i.e. identifiability holds :

$$\forall \alpha, \alpha' \in (0,1), \quad \forall \theta_j, \theta_j^{'}, j = 1,2 \quad f_{\theta,\alpha} = f_{\theta',\alpha'} \quad \Rightarrow \alpha = \alpha^{'}, \quad \theta = \theta^{'}$$

theorem

Under above assumptions, then for all $\epsilon > 0$,

$$\pi\left[|\alpha - \alpha^*| > \epsilon | \mathbf{x}^n\right] = o_p(1)$$

II. Recovery of the parameters : Separated models – $f_{\theta,\alpha} = \alpha f_{1,\theta_1} + (1-\alpha) f_{2,\theta_2}$

Assumption : Models are separated, i.e. identifiability holds :

$$\forall \alpha, \alpha' \in (0,1), \quad \forall \theta_j, \theta_j^{'}, j = 1,2 \quad f_{\theta,\alpha} = f_{\theta',\alpha'} \quad \Rightarrow \alpha = \alpha^{'}, \quad \theta = \theta^{'}$$

theorem

If

- $\theta_j \longrightarrow f_{j,\theta_j}$ is $\mathscr{C}^2$ around θ_j^*, $j = 1,2$,
- $f_{1,\theta_1^*} - f_{2,\theta_2^*}, \nabla f_{1,\theta_1^*}, \nabla f_{2,\theta_2^*}$ are linearly independent in y and
- there exists $\delta > 0$ such that

$$\nabla f_{1,\theta_1^*}, \nabla f_{2,\theta_2^*}, \sup_{|\theta_1 - \theta_1^*| < \delta} |D^2 f_{1,\theta_1}|, \sup_{|\theta_2 - \theta_2^*| < \delta} |D^2 f_{2,\theta_2}| \in L_1$$

then

$$\pi \left[|\alpha - \alpha^*| > M \sqrt{\log n / n} \middle| \mathbf{x}^n \right] = o_p(1).$$

II. Recovery of the parameters : Separated models – $f_{\theta,\alpha} = \alpha f_{1,\theta_1} + (1-\alpha) f_{2,\theta_2}$

Assumption : Models are separated, i.e. identifiability holds :

$$\forall \alpha, \alpha' \in (0,1), \quad \forall \theta_j, \theta_j^{'}, j = 1,2 \quad f_{\theta,\alpha} = f_{\theta',\alpha'} \quad \Rightarrow \alpha = \alpha^{'}, \quad \theta = \theta^{'}$$

theorem allows for interpretation of α under the posterior : If data $\mathbf{x}^n$ is generated from model $\mathfrak{M}_1$ then posterior on α concentrates around $\alpha = 1$

Embedded case

Here $\mathfrak{M}_1$ is a submodel of $\mathfrak{M}_2$, i.e.

$$\theta_2 = (\theta_1, \psi) \quad \text{and} \quad \theta_2 = (\theta_1, \psi_0 = 0)$$

corresponds to $f_{2,\theta_2} \in \mathfrak{M}_1$
Same posterior concentration rate

$$\sqrt{\log n/n}$$

for estimating α when $\alpha^* \in (0,1)$ and $\psi^* \neq 0$.

Null case

- Case where $\psi^* = 0$, i.e., f^* is in model $\mathfrak{M}_1$
- Two possible paths to approximate f^* : either α goes to 1 (path 1) or ψ goes to 0 (path 2)
- New identifiability condition : $P_{\theta,\alpha} = P^*$ only if

$$\alpha = 1,\ \theta_1 = \theta_1^*,\ \theta_2 = (\theta_1^*, \psi) \quad \text{or} \quad \alpha \leq 1,\ \theta_1 = \theta_1^*,\ \theta_2 = (\theta_1^*, 0)$$

Null case

- Case where $\psi^* = 0$, i.e., f^* is in model $\mathfrak{M}_1$
- Two possible paths to approximate f^* : either α goes to 1 (path 1) or ψ goes to 0 (path 2)
- New identifiability condition : $P_{\theta,\alpha} = P^*$ only if

$$\alpha = 1,\ \theta_1 = \theta_1^*,\ \theta_2 = (\theta_1^*, \psi) \quad \text{or} \quad \alpha \leq 1,\ \theta_1 = \theta_1^*,\ \theta_2 = (\theta_1^*, 0)$$

Prior

$$\pi(\alpha, \theta) = \pi_\alpha(\alpha)\pi_1(\theta_1)\pi_\psi(\psi), \quad \theta_2 = (\theta_1, \psi)$$

with common (prior on) θ_1

Assumptions

• [B1] *Regularity* : Assume that $\theta_1 \rightarrow f_{1,\theta_\mathbf{1}}$ and $\theta_2 \rightarrow f_{2,\theta_\mathbf{2}}$ are 3 times continuously differentiable and that

$$F^* \left(\frac{\sup_{|\theta_\mathbf{1} - \theta_\mathbf{1}^*| < \delta} |D^r f_{1,\theta_\mathbf{1}^*}|^s}{\underline{f}^s_{1,\theta_\mathbf{1}^*}} \right) < +\infty, \quad r \leq 3$$

• [B2] *Integrability* :
$\exists\, \mathscr{S}_0 \subset \mathscr{S} \cap \{|\psi| > \delta_0 > 0\}$ s.t. Leb$(\mathscr{S}_0) > 0$, and s.t. $\forall \psi \in \mathscr{S}_0$,

$$F^* \left(\frac{\sup_{|\theta_\mathbf{1} - \theta_\mathbf{1}^*| < \delta} f_{2,\theta_\mathbf{1},\psi}}{f^4_{1,\theta_\mathbf{1}^*}} \right) < +\infty, \quad F^* \left(\frac{\sup_{|\theta_\mathbf{1} - \theta_\mathbf{1}^*| < \delta} f^3_{2,\theta_\mathbf{1},\psi}}{\underline{f}^3_{1,\theta 1^*}} \right) < +\infty,$$

Assumptions

• [B1] *Regularity* : Assume that $\theta_1 \to f_{1,\theta_{\mathbf{1}}}$ and $\theta_2 \to f_{2,\theta_{\mathbf{2}}}$ are 3 times continuously differentiable and that

$$F^*\left(\frac{\sup_{|\theta_{\mathbf{1}}-\theta_{\mathbf{1}}^*|<\delta} |D^r f_{1,\theta_{\mathbf{1}}^*}|^s}{\underline{f}^s_{1,\theta_{\mathbf{1}}^*}}\right) < +\infty, \quad r \leq 3$$

• [B2] *Integrability* :
$\exists\, \mathscr{S}_0 \subset \mathscr{S} \cap \{|\psi| > \delta_0 > 0\}$ s.t. Leb$(\mathscr{S}_0) > 0$, and s.t. $\forall \psi \in \mathscr{S}_0$,

$$F^*\left(\frac{\sup_{|\theta_{\mathbf{1}}-\theta_{\mathbf{1}}^*|<\delta} f_{2,\theta_{\mathbf{1}},\psi}}{f^4_{1,\theta_{\mathbf{1}}^*}}\right) < +\infty, \quad F^*\left(\frac{\sup_{|\theta_{\mathbf{1}}-\theta_{\mathbf{1}}^*|<\delta} f^3_{2,\theta_{\mathbf{1}},\psi}}{\underline{f}^3_{1,\theta 1^*}}\right) < +\infty,$$

Assumptions

• [B1] *Regularity* : Assume that $\theta_1 \to f_{1,\theta_1}$ and $\theta_2 \to f_{2,\theta_2}$ are 3 times continuously differentiable and that

$$F^* \left(\frac{\sup_{|\theta_1 - \theta_1^*| < \delta} |D^r f_{1,\theta_1^*}|^s}{\underline{f}^s_{1,\theta_1^*}} \right) < +\infty, \quad r \leq 3$$

• [B2] *Integrability* :
$\exists \mathscr{S}_0 \subset \mathscr{S} \cap \{|\psi| > \delta_0 > 0\}$ s.t. Leb$(\mathscr{S}_0) > 0$, and s.t. $\forall \psi \in \mathscr{S}_0$,

$$F^* \left(\frac{\sup_{|\theta_1 - \theta_1^*| < \delta} f_{2,\theta_1,\psi}}{f^4_{1,\theta_1^*}} \right) < +\infty, \quad F^* \left(\frac{\sup_{|\theta_1 - \theta_1^*| < \delta} f^3_{2,\theta_1,\psi}}{\underline{f}^3_{1,\theta 1^*}} \right) < +\infty,$$

[B3] *Stronger identifiability* : Set

$$\nabla f_{2,\theta_1^*,\psi^*}(x) = \left(\nabla_{\theta_1} f_{2,\theta_1^*,\psi^*}(x)^\mathsf{T}, \nabla_\psi f_{2,\theta_1^*,\psi^*}(x)^\mathsf{T} \right)^\mathsf{T}.$$

Then for all $\psi \in \mathscr{S}$ with $\psi \neq 0$, if $\eta_0 \in \mathbb{R}$, $\eta_1 \in \mathbb{R}^{d_1}$

$$\eta_0 (f_{1,\theta_1^*} - f_{2,\theta_1^*,\psi}) + \eta_1^\mathsf{T} \nabla_{\theta_1} f_{1,\theta_1^*} = 0 \quad \Leftrightarrow \eta_1 = 0, \; \eta_2 = 0$$

Consistency

theorem

In the model : $f_{\theta_1,\psi,\alpha} = \alpha f_{1,\theta_1} + (1-\alpha) f_{2,\theta_1,\psi}$ and $\mathbf{x}^n = (x_1, \cdots, x_n) \overset{i.i.d}{\sim} f_{1,\theta_1^*}$, If $B1 - B3$ hold , then

$$\pi\left[(\alpha,\theta); \|f_{\theta,\alpha} - f^*\|_1 > M\sqrt{\log n/n} | \mathbf{x}^n\right] = o_p(1).$$

If $\alpha \sim \mathscr{B}(a_1, a_2)$, with $a_2 < d_2$, and if the prior dens. $\pi_{\theta_1,\psi}$ is C^o and > 0 at $(\theta_1^*, 0)$, then $M_n \longrightarrow \infty$

$$\pi\left[\alpha < 1 - M_n (\log n)^\gamma/\sqrt{n} | \mathbf{x}^n\right] = o_p(1), \quad \gamma = \max((d_1 + a_2)/(d_2 - a_2), 1)/2,$$

M-open case

When the true model behind the data is neither of the tested models, what happens ?

M-open case

When the true model behind the data is neither of the tested models, what happens ?

- issue mostly bypassed by classical Bayesian procedures
- theoretically produces an α^* away from both 0 and 1
- possible (recommended ?) inclusion of a Bayesian non-parametric model within alternatives

Towards which decision ?

And if we have to make a decision ?

- soft: consider behaviour of posterior under prior predictives
  - or posterior predictive [e.g., prior predictive does not exist]
  - boostrapping behaviour
  - comparison with Bayesian non-parametric solution
- hard: rethink the loss function

Thank You