Testing as estimation: the demise of the Bayes factors

Christian P. Robert
Université Paris-Dauphine and University of Warwick

arXiv:1412.2044
with K. Kamary, K. Mengersen, and J. Rousseau

Outline

Testing hypotheses

Hypothesis testing

- central problem of statistical inference
- 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)]

Besting 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
- posterior predictive tools successfully advocated in Gelman et al. (2013) even though they involve double use of data

Besting 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
- posterior predictive tools successfully advocated in Gelman et al. (2013) even though they involve double use of data

Bayesian modelling

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

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

and associate with each model a prior distribution,

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

[Jeffreys, 1939]

Bayesian modelling

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

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

in order to compare the marginal likelihoods

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

[Jeffreys, 1939]

Bayesian modelling

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

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

either through Bayes factor or posterior probability,

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

the latter depends on the prior weights ω_i

[Jeffreys, 1939]

Bayesian modelling

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

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

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 α

[Jeffreys, 1939]

Some difficulties

- tension between (i) posterior probabilities justified by binary loss but depending on unnatural prior weights, and (ii) Bayes factors that eliminate dependence but lose direct connection with posterior, unless prior weights are integrated within loss
- delicate interpretation (or calibration) of the strength of the Bayes factor towards supporting a given hypothesis or model: not a Bayesian decision rule!
- difficulty with posterior probabilities: tendency to interpret them as p-values while they only report respective strengths of fitting to both models

Some further difficulties

- long-lasting impact of the prior modeling, i.e., choice of prior distributions on both parameter spaces under comparison, despite overall consistency for Bayes factor
- major discontinuity in use of improper priors, not justified in most testing situations, leading to *ad hoc* solutions (zoo), where data is either used twice or split artificially
- binary (*accept* vs. *reject*) outcome more suited for immediate decision (if any) than for model evaluation, connected with rudimentary loss function

Lindley's paradox

In a normal mean testing problem,

$$\bar{x}_n \sim \mathcal{N}(\theta, \sigma^2/n)\,, \quad H_0 : \theta = \theta_0\,,$$

under Jeffreys prior, $\theta \sim \mathcal{N}(\theta_0, \sigma^2)$, the Bayes factor

$$\mathfrak{B}_{01}(t_n) = (1+n)^{1/2} \exp\left(-nt_n^2/2[1+n]\right)\,,$$

where $t_n = \sqrt{n}|\bar{x}_n - \theta_0|/\sigma$, satisfies

$$\mathfrak{B}_{01}(t_n) \overset{n\longrightarrow\infty}{\longrightarrow} \infty$$

[assuming a fixed t_n]

[Lindley, 1957]

A strong impropriety

Improper priors not allowed in Bayes factors:

If

$$\int_{\Theta_1} \pi_1(\mathrm{d}\theta_1) = \infty \quad \text{or} \quad \int_{\Theta_2} \pi_2(\mathrm{d}\theta_2) = \infty$$

then π_1 or π_2 cannot be coherently normalised while the normalisation matters in the Bayes factor $\mathfrak{B}_{12}$

Lack of mathematical justification for "common nuisance parameter" [and prior of]

[Berger, Pericchi, and Varshavsky, 1998; Marin and Robert, 2013]

A strong impropriety

Improper priors not allowed in Bayes factors:

If

$$\int_{\Theta_1} \pi_1(\mathrm{d}\theta_1) = \infty \quad \text{or} \quad \int_{\Theta_2} \pi_2(\mathrm{d}\theta_2) = \infty$$

then π_1 or π_2 cannot be coherently normalised while the normalisation matters in the Bayes factor $\mathfrak{B}_{12}$

Lack of mathematical justification for "common nuisance parameter" [and prior of]

[Berger, Pericchi, and Varshavsky, 1998; Marin and Robert, 2013]

Paradigm shift

New proposal as paradigm shift in Bayesian processing of hypothesis testing and of model selection

- convergent and naturally interpretable solution
- extended use of improper priors
- abandonment of the Neyman-Pearson decision framework
- natural strenght of evidence

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

New proposal as paradigm shift in Bayesian processing of hypothesis testing and of model selection

- convergent and naturally interpretable solution
- extended use of improper priors
- abandonment of the Neyman-Pearson decision framework
- natural strenght of evidence

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 equivalent to use of a posterior probability
- More natural approach to testing, while sparse in parameters
- Calibration of the posterior distribution of mixture weight, while moving away from artificial notion of the posterior probability of a model

Encompassing mixture model

Idea: Given two statistical models,

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

embed both within an encompassing mixture

$$\mathfrak{M}_\alpha : \; x \sim \alpha f_1(x|\theta_1) + (1-\alpha) f_2(x|\theta_2)\,, \;\; 0 \leqslant \alpha \leqslant 1 \qquad (1)$$

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

Encompassing mixture model

Idea: Given two statistical models,

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

embed both within an encompassing mixture

$$\mathfrak{M}_\alpha : \; x \sim \alpha f_1(x|\theta_1) + (1-\alpha) f_2(x|\theta_2)\,, \;\; 0 \leqslant \alpha \leqslant 1 \qquad (1)$$

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

Encompassing mixture model

Idea: Given two statistical models,

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

embed both within an encompassing mixture

$$\mathfrak{M}_\alpha : \; x \sim \alpha f_1(x|\theta_1) + (1-\alpha) f_2(x|\theta_2)\,, \;\; 0 \leqslant \alpha \leqslant 1 \qquad (1)$$

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 problem, but definitive advantages to 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, and $0-1$ loss setting
- posterior on α provides measure of proximity to 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 like WAIC

Computational motivations

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

Noninformative motivations

- novel Bayesian feature: 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 processed in this setting, but non-informative (improper) priors also are manageable, provided both models first reparameterised into shared parameters, e.g. location and scale parameters
- in special case when all parameters are common

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

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
- common parameters are nuisance parameters, easily integrated
- prior model weights ω_i rarely discussed in classical Bayesian approach, with linear impact on posterior probabilities
- prior modeling only involves selecting a prior on α, e.g., $\alpha \sim \mathcal{B}(a_0, a_0)$
- 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 example

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

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

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%)

Poisson/Geometric example

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

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

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%)

Poisson/Geometric example

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

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

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

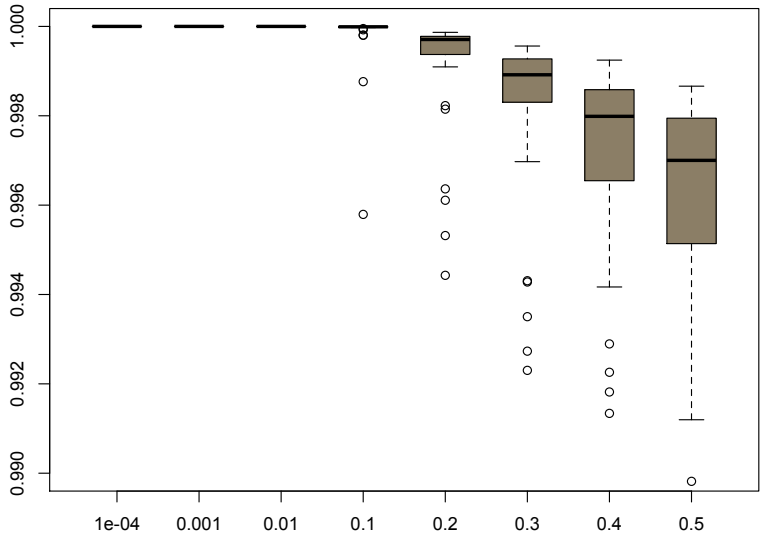
$$\alpha \sim \mathcal{B}e(n_1(\zeta) + a_0, n_2(\zeta) + a_0)$$

Exact Bayes factor opposing Poisson and Geometric

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

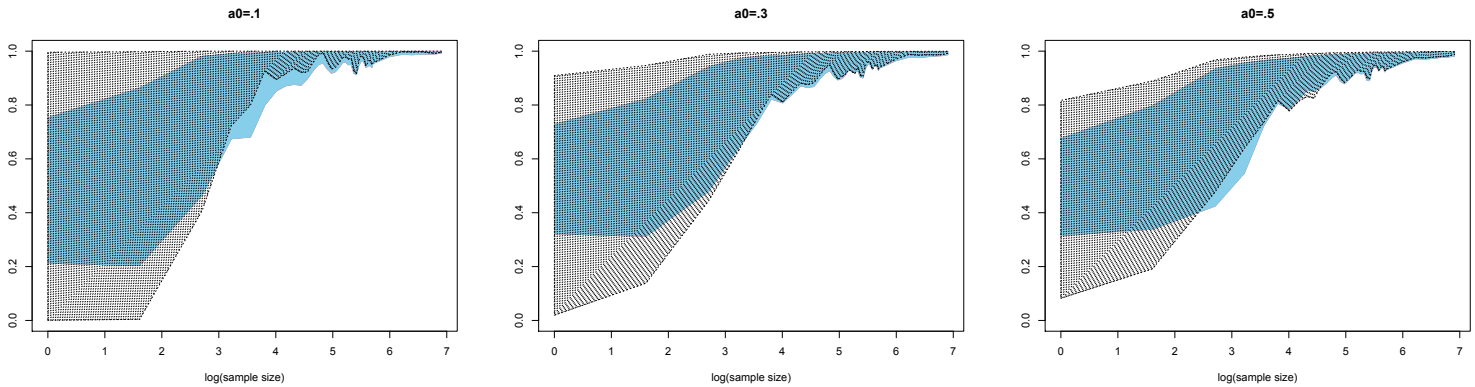
although undefined from a purely mathematical viewpoint

Weight estimation



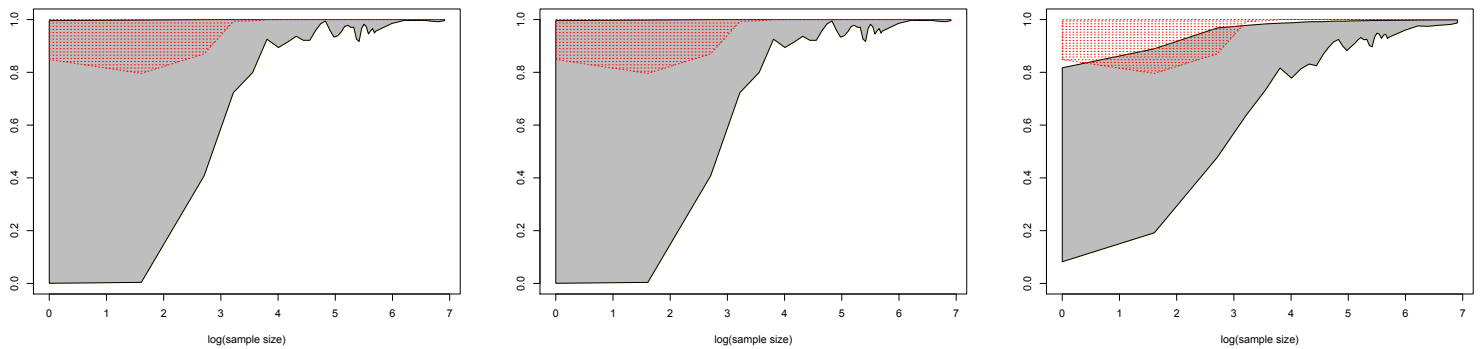
Posterior medians of α for 100 Poisson $\mathcal{P}(4)$ datasets of size $n = 1000$, for $a_0 = .0001, .001, .01, .1, .2, .3, .4, .5$. Each posterior approximation is based on 10^4 Metropolis-Hastings iterations.

Consistency



Posterior means *(sky-blue)* and medians *(grey-dotted)* of α, over 100 Poisson $\mathcal{P}(4)$ 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 $\mathcal{P}(4)$ datasets with sample sizes n between 1 and 1000, for $a_0 = .001, .1, .5$

Normal-normal comparison

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

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

Normal-normal comparison

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

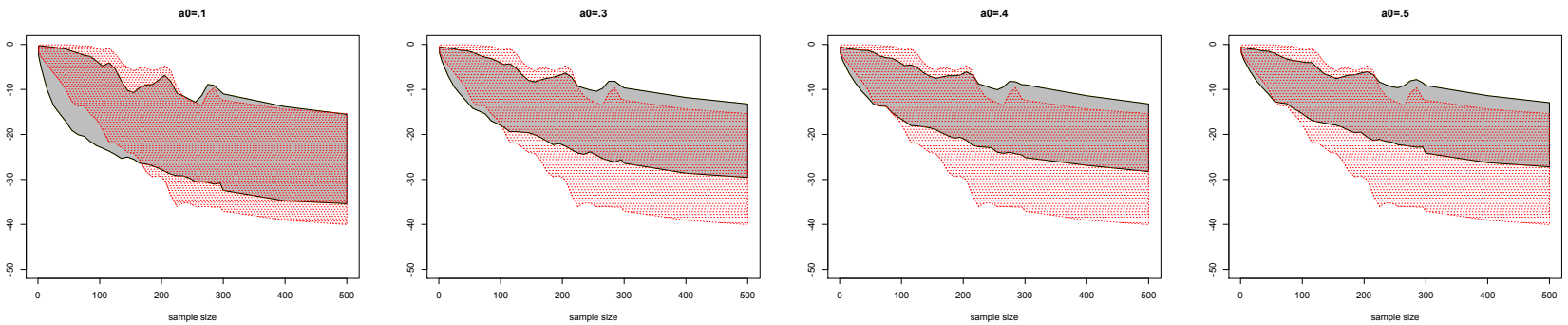
$$\mathfrak{B}_{12} = 2^{n-1/2} \Big/ \exp 1/4 \sum_{i=1}^{n} (x_i - \bar{x})^2 ,$$

Normal-normal comparison

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

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

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 $\mathcal{N}(0,1)$ samples as sample size n grows from 1 to 500. and α is the weight of $\mathcal{N}(0,1)$ 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 $\mathcal{N}(\theta_1, 1)$ nor from $\mathcal{N}(\theta_2, 2)$, 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

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

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

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

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

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

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

Prior modelling

Under default g-prior

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

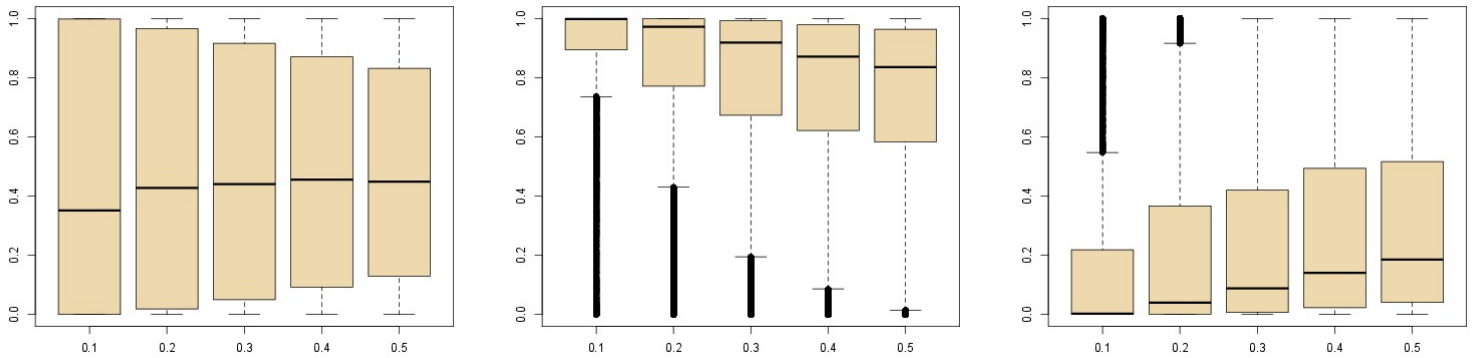
full conditional posterior distributions given allocations

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

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

Survival analysis models

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

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

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

Survival analysis models

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

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

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.

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

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

Reparameterisation

Looking for common parameter(s):

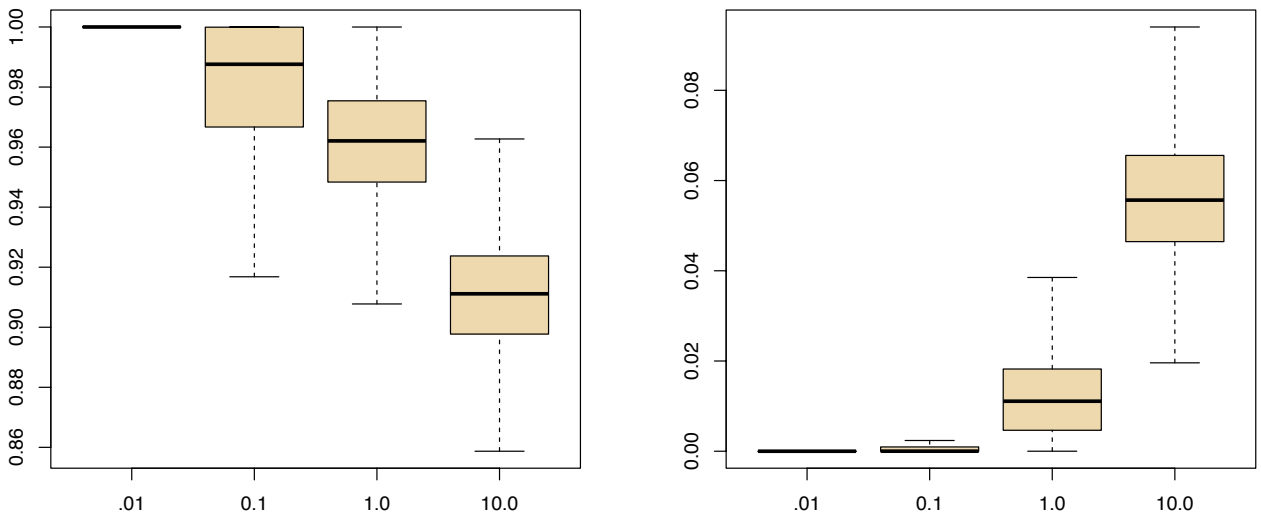
$$\phi = \mu + \gamma\beta = \xi$$
$$\sigma^2 = \pi^2\beta^2/6 = \zeta^2\pi^2/3$$

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]

Two different cases

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

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 $\mathfrak{M}_2$.

Separated models

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

$$\forall \alpha, \alpha' \in [0,1], \quad \forall \theta_j, \theta_j', j = 1,2 \quad P_{\theta,\alpha} = P_{\theta',\alpha'} \quad \Rightarrow \alpha = \alpha', \quad \theta = \theta'$$

theorem

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

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

Separated models

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

$$\forall \alpha, \alpha' \in [0,1], \quad \forall \theta_j, \theta_j', j = 1,2 \quad P_{\theta,\alpha} = P_{\theta',\alpha'} \quad \Rightarrow \alpha = \alpha', \quad \theta = \theta'$$

theorem

If

- $\theta_j \longrightarrow f_{j,\theta_j}$ is $\mathcal{C}^2$ 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

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

Separated models

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

$$\forall \alpha, \alpha' \in [0,1], \quad \forall \theta_j, \theta_j', j = 1,2 \quad P_{\theta,\alpha} = P_{\theta',\alpha'} \quad \Rightarrow \alpha = \alpha', \quad \theta = \theta'$$

theorem allows for interpretation of α under the posterior: If data $\boldsymbol{x}^n$ 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

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

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

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

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

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

Consistency

theorem

Given the mixture $f_{\theta_1,\psi,\alpha} = \alpha f_{1,\theta_1} + (1-\alpha) f_{2,\theta_1,\psi}$ and a sample $\mathbf{x}^n = (x_1, \cdots, x_n)$ issued from f_{1,θ_1^*}, under regularity assumptions, and an $M > 0$ such that

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

If $\alpha \sim \mathcal{B}(a_1, a_2)$, with $a_2 < d_2$, and if the prior $\pi_{\theta_1,\psi}$ is absolutely continuous with positive and continuous density at $(\theta_1^*, 0)$, then for $M_n \longrightarrow \infty$

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

Conclusion

- many applications of the Bayesian paradigm concentrate on the comparison of scientific theories and on testing of null hypotheses
- natural tendency to default to Bayes factors
- poorly understood sensitivity to prior modeling and posterior calibration

Time is ripe for a paradigm shift

Conclusion

Time is ripe for a paradigm shift

- original testing problem replaced with a better controlled estimation target
- allow for posterior variability over the component frequency as opposed to deterministic Bayes factors
- range of acceptance, rejection and indecision conclusions easily calibrated by simulation
- posterior medians quickly settling near the boundary values of 0 and 1
- potential derivation of a Bayesian *b*-value by looking at the posterior area under the tail of the distribution of the weight

Prior modelling

Time is ripe for a paradigm shift

- Partly common parameterisation always feasible and hence allows for reference priors
- removal of the absolute prohibition of improper priors in hypothesis testing
- prior on the weight α shows sensitivity that naturally vanishes as the sample size increases
- default value of $a_0 = 0.5$ in the Beta prior

Computing aspects

Time is ripe for a paradigm shift

- proposal that does not induce additional computational strain
- when algorithmic solutions exist for both models, they can be recycled towards estimating the encompassing mixture
- easier than in standard mixture problems due to common parameters that allow for original MCMC samplers to be turned into proposals
- Gibbs sampling completions useful for assessing potential outliers but not essential to achieve a conclusion about the overall problem