

Bayesian testing using mixtures

Judith Rousseau

Joint work with K. Kanary, K. Mengersen and C. Robert

University of Oxford & Université Paris-Dauphine

Leiden

Outline

- 1 Introduction
- 2 Testing problems as estimating mixture models
- 3 Asymptotic properties
- 4 Some practical aspects and examples

Introduction

Hypothesis testing

- central problem of statistical inference

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

Hypothesis testing

- central problem of statistical inference
- dramatically differentiating feature between classical and Bayesian paradigms

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

Hypothesis testing

- central problem of statistical inference
- dramatically differentiating feature between classical and Bayesian paradigms
- wide open to controversy and divergent opinions, even within the Bayesian community

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

Testing hypotheses

- Bayesian model selection as comparison of k potential statistical models towards the selection of model that fits the data “best”

Testing hypotheses

- Bayesian model selection as comparison of k potential statistical models towards the selection of model that fits the data “best”
- tools like Bayes factor naturally include a penalisation addressing model complexity, mimicked by Bayes Information (BIC) and Deviance Information (DIC) criteria

Testing hypotheses

- Bayesian model selection as comparison of k potential statistical models towards the selection of model that fits the data “best”
- 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 imply multiple uses of data

Bayesian testing via BF

Standard Bayesian approach to testing :

► **models**

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

► **priors**

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

► **marginal likelihoods**

$$m_1(x) = \int_{\Theta_1} f_1(x|\theta_1) \pi_1(\theta_1) d\theta_1 \quad \text{and} \quad m_2(x) = \int_{\Theta_2} f_2(x|\theta_2) \pi_2(\theta_2) d\theta_2$$

► **Bayes factor** ; posterior probability :

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

the latter depends on the prior weights ω_i

Bayesian decision

Bayesian decision step

- comparing Bayes factor B_{12} with threshold (1)

Bayesian decision

Bayesian decision step

- comparing Bayes factor B_{12} with threshold (1)
- comparing posterior probability $\mathbb{P}(\mathcal{M}_1|x)$ with bound (1/2)

Bayesian decision

Bayesian decision step

- comparing Bayes factor B_{12} with threshold (1)
- comparing posterior probability $\mathbb{P}(\mathcal{M}_1|x)$ with bound (1/2)

Bayesian decision

Bayesian decision step

- comparing Bayes factor B_{12} with threshold (1)
- comparing posterior probability $\mathbb{P}(\mathcal{M}_1|x)$ with bound (1/2)

If k models, $k \geq 2$: **highest posterior probability model** is the one selected, but highly dependent on the prior modelling.

Some difficulties

- Bayesian answer to 0-1 loss function.

Some difficulties

- Bayesian answer to 0-1 loss function.
 - duality between prior/threshold

Some difficulties

- Bayesian answer to 0-1 loss function.
 - duality between prior/threshold
 - impossibility of improper (non informative) priors

Some difficulties

- Bayesian answer to 0-1 loss function.
 - duality between prior/threshold
 - impossibility of improper (non informative) priors
 - Just one number : lack of uncertainty quantification associated to it

Some difficulties

- Bayesian answer to 0-1 loss function.
 - duality between prior/threshold
 - impossibility of improper (non informative) priors
 - Just one number : lack of uncertainty quantification associated to it
- $\log n$ penalty induced is quite strong – lost of small signals

Some difficulties

- Bayesian answer to 0-1 loss function.
 - duality between prior/threshold
 - impossibility of improper (non informative) priors
 - Just one number : lack of uncertainty quantification associated to it
- $\log n$ penalty induced is quite strong – lost of small signals
- Computationnally involved– common use of BIC instead

Encompassing mixture model

► **Idea** : Given two statistical models,

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

Encompassing mixture model

► **Idea** : Given two statistical models,

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

embed both within an encompassing mixture

$$\mathcal{M}_\alpha : x_i \overset{iid}{\sim} \alpha f_1(x|\theta_1) + (1 - \alpha) f_2(x|\theta_2), \quad 0 \leq \alpha \leq 1 \quad (1)$$

Encompassing mixture model

► **Idea** : Given two statistical models,

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

embed both within an encompassing mixture

$$\mathcal{M}_\alpha : x_i \overset{iid}{\sim} \alpha f_1(x|\theta_1) + (1 - \alpha) f_2(x|\theta_2), \quad 0 \leq \alpha \leq 1 \quad (1)$$

► **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

New interpretation : estimation problem instead of a testing one

- Bayes estimate of the weight α replaces posterior probability of model $\mathcal{M}_1$,

New interpretation : estimation problem instead of a testing one

- Bayes estimate of the weight α replaces posterior probability of model $\mathcal{M}_1$,
- α : proportion of the population which is well represented by $\mathcal{M}_1$

New interpretation : estimation problem instead of a testing one

- Bayes estimate of the weight α replaces posterior probability of model $\mathcal{M}_1$,
- α : proportion of the population which is well represented by $\mathcal{M}_1$
- Allows for measures of uncertainty on the model adequation

L_1 posterior consistency

Posterior consistency holds for embedded mixture model
[under mild conditions]

$$\mathcal{M}_\alpha : x_i \stackrel{iid}{\sim} \alpha f_1(x|\theta_1) + (1 - \alpha)f_2(x|\theta_2), \quad 0 \leq \alpha \leq 1$$

Theorem

*Under typical regularity conditions on $\theta_j \rightarrow f_j(\cdot|\theta_j)$, $\Theta_j \subset \mathbb{R}^{d_j}$,
and prior π on $\alpha, \theta_1, \theta_2$*

$$\mathbb{P} \left(\|f_{\alpha, \theta} - f^*\|_1 \leq \frac{M\sqrt{\log n}}{\sqrt{n}} \mid \mathbf{x}^n \right) = 1 + o_p(1)$$

► **Non parametric** Obvious extension to non parametric cases
, with nonparametric rates

Parameter recovery 1 : well separated models

► identifiability condition

$$P_{\alpha, \underline{\theta}} = P_{\alpha', \underline{\theta}'} \Rightarrow \alpha = \alpha', \quad \underline{\theta} = \underline{\theta}'$$

► separability

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

► Regularity $f_{j, \theta_j} : C^2$.

Then $\forall \alpha^* \in [0, 1], \theta_j^*$

$$\mathbb{P} \left(|\alpha - \alpha^*| \leq \frac{M\sqrt{\log n}}{\sqrt{n}} \mid \mathbf{x}^n \right) = 1 + o_p(1)$$

Parameter recovery 1 : embedded models

$$\mathcal{M}_\alpha : x_i /, \overset{iid}{\sim} \alpha f_1(x|\theta) + (1 - \alpha)f_2(x|\theta, \psi), \quad 0 \leq \alpha \leq 1$$

- ▶ **prior** $\pi(\alpha)\pi(\theta)\pi(\psi)$, $[f_2(\cdot|\theta, 0) = f_1(\cdot|\theta)]$
- ▶ **(non) identifiability** Under $\mathcal{M}_1 : P_{\theta, \alpha} = P^*$

$$\Rightarrow \alpha = 1, \quad \theta_1 = \theta_1^*, \quad \theta_2 = (\theta_1^*, \psi)$$

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

- ▶ **Regularity conditions** + Linear indepdce of score in (α, θ)

Then with $\pi(\alpha) = \text{Beta}(a_1, a_2)$.

- If $a_2 < d_\psi$ **concentrates around 1**

$$\Pi \left(|\alpha - 1| \leq \frac{M_n}{\sqrt{n}} \middle| \mathbf{x}^n \right) = 1 + o_p(1), \quad \forall M_n \uparrow +\infty$$

Parameter recovery 1 : embedded models

$$\mathcal{M}_\alpha : x_i /, \stackrel{iid}{\sim} \alpha f_1(x|\theta) + (1 - \alpha) f_2(x|\theta, \psi), \quad 0 \leq \alpha \leq 1$$

- ▶ **prior** $\pi(\alpha)\pi(\theta)\pi(\psi)$, $[f_2(\cdot|\theta, 0) = f_1(\cdot|\theta)]$
- ▶ **(non) identifiability** Under $\mathcal{M}_1 : P_{\theta, \alpha} = P^*$

$$\Rightarrow \alpha = 1, \quad \theta_1 = \theta_1^*, \quad \theta_2 = (\theta_1^*, \psi)$$

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

- ▶ **Regularity conditions** + Linear indepdce of score in (α, θ)

Then with $\pi(\alpha) = \text{Beta}(a_1, a_2)$.

- If $a_2 < d_\psi$ **concentrates around 1**

$$\mathbb{P} \left(|\alpha - 1| \leq \frac{M_n}{\sqrt{n}} \mid \mathbf{x}^n \right) = 1 + o_p(1), \quad \forall M_n \uparrow +\infty$$

- If $a_2 > d_\psi$ **concentrates away from 1**

$$\mathbb{P} (|\alpha - 1| \leq e_n \mid \mathbf{x}^n) = o_p(1), \quad \forall e_n = o(1)$$

- If $a_2 < d_\psi$: posterior on α concentrates around 1 at the rate $1/\sqrt{n}$ (no spurious $\log n$) under $\mathcal{M}_1$.

- If $a_2 < d_\psi$: posterior on α concentrates around 1 at the rate $1/\sqrt{n}$ (no spurious $\log n$) under $\mathcal{M}_1$.
- Under $\mathcal{M}_2$: concentration of α to 0 at the rate $1/\sqrt{n}$.

- If $a_2 < d_\psi$: posterior on α concentrates around 1 at the rate $1/\sqrt{n}$ (no spurious $\log n$) under $\mathcal{M}_1$.
- Under $\mathcal{M}_2$: concentration of α to 0 at the rate $1/\sqrt{n}$.
- BF under $\mathcal{M}_2$ exponentially decreasing & under $\mathcal{M}_1$ goes to ∞ at the rate $n^{d_\psi/2}$.

- If $a_2 < d_\psi$: posterior on α concentrates around 1 at the rate $1/\sqrt{n}$ (no spurious $\log n$) under $\mathcal{M}_1$.
- Under $\mathcal{M}_2$: concentration of α to 0 at the rate $1/\sqrt{n}$.
- BF under $\mathcal{M}_2$ exponentially decreasing & under $\mathcal{M}_1$ goes to ∞ at the rate $n^{d_\psi/2}$.
- What about separation rate ?

Separation rate : behaviour under *contiguous* alternatives

$$f_n^* = f_{2,\theta_{1,n},\psi_n},$$

Then $\forall M_n \uparrow +\infty, \exists M' > 0$

$$\sup_{\theta_{1,n} \in \Theta_1, \|\psi_n\| \geq M_n/\sqrt{n}} E_{\theta_{1,n},\psi_n} \pi \left[|\alpha - 1| \leq M' M_n^2 / \sqrt{n} | \mathbf{x}^n \right] = o(1)$$

► **Testing decision rule** : if $\mathcal{M}_1$ chosen as soon as

$$\Pi(\alpha > 1 - M_0/\sqrt{n} | \mathbf{x}^n) > 1/2, \quad M_0 \text{ large}$$

then separation rate : detect $\mathcal{M}_2$ as soon as

$$\|\psi\| \gtrsim \sqrt{M_0/n}$$

Obviously more info in $\Pi(\alpha | \mathbf{x}^n)$

Towards possibly improper priors

- **idea : use common parameterization**

$$\mathcal{M}_1 : f_{1,\theta_1}, \quad \mathcal{M}_2 : f_{2,\theta_2}, \quad \alpha f_{1,\theta_1} + (1 - \alpha) f_{2,\theta_2}$$

Common parameterization e.g.

$$\eta = E_{1,\theta_1}(X) = E_{2,\theta_2}(X), \quad \Rightarrow \quad \theta_1(\eta, u_1), \quad \theta_2(\eta, u_2)$$

allows for improper prior on η

- **Used in the embedded model**

$$\alpha f_{1,\theta} + (1 - \alpha) f_{2,\theta,\psi}$$

About implementation

- Simpler computationally (typically) : standard algorithms are available for Bayesian mixture estimation

About implementation

- Simpler computationally (typically) : 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

About implementation

- Simpler computationally (typically) : 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

About implementation

- Simpler computationally (typically) : 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
- variability of posterior distribution on α allows for a more thorough assessment of the strength of this support

About implementation

- Simpler computationally (typically) : 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
- variability of posterior distribution on α allows for a more thorough assessment of the strength of this support
- Calibration using prior predictive if necessary

Illustration 1 : simple separated models

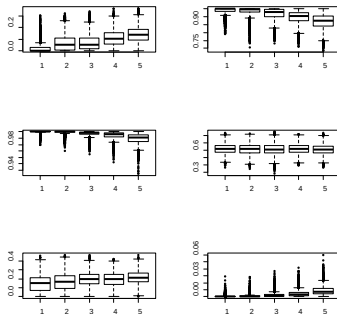
$$\mathcal{M}_1 : \mathcal{N}(\mu, 1), \quad \mathcal{M}_2 : \mathcal{N}(\mu, 2), \quad \alpha \sim \text{Beta}(a_1, a_1)$$

Testing hypothesis as a mixture estimation problem

1. Test $N(0, \text{var}=1)$ vs $N(0, \text{var}=2)$

Hyperparameters $(0.1, 0.1), (0.2, 0.2), (0.5, 0.5), (1, 1), (2, 2)$

Sample from $N(0, 2)$, $N(0, 1)$, $N(0.5)$, $N(0, 1.3)$, $N(0, 1.7)$, $N(0, 5)$



Nonparametric example : goodness of fit

$$\mathcal{M}_1 : \mathcal{N}(\mu, \sigma^2), \quad \text{vs} \quad \mathcal{M}_2 : \text{not } \mathcal{N}(\mu, \sigma^2)$$

► prior embedding model

$$f_\alpha(x) = \alpha \mathcal{N}(\mu, \sigma^2) + (1-\alpha) \int \mathcal{N}(\mu', \sigma^2) dP(\mu'), \quad \alpha \sim \text{Beta}(a_1, a_1)$$

$$\mu | \sigma^2 \sim \mathcal{N}(0, \sigma^2 \tau^2), \quad \sigma^2 \sim IG, \quad P \sim DP(M, \mathcal{N}(0, \sigma^2 \tau^2))$$

► How choose M, a_1 : theoretical insights

- If true $f^* \neq \mathcal{N}$, then L_1 posterior contraction rate implies

$$\Pi(\alpha < 1 - Cd(f^*, \mathcal{N}) | \mathbf{x}^n) = 1 + o_p(1)$$

NP GoF : impact of M, a_1

- If $f^* = \mathcal{N}(\mu_*, \sigma_*^2)$, then Nguyen

$$\Pi(\textcolor{red}{W}(\textcolor{red}{Q}, \delta_{(\mu_*, \sigma_*)}) \leq \epsilon | \mathbf{x}^n) = 1 + o_p(1), \quad Q = (\alpha \delta_{(\mu)} + (1 - \alpha) P) \otimes \delta_{(\sigma)}$$

Which implies that under the posterior dist

$$|E_Q(\mu) - \mu^*| + |E_Q(\mu - \mu_*)^2 + 2\sigma_*(\sigma - \sigma_*)| + (\sigma - \sigma_*)^2 \lesssim \frac{\sqrt{\log n}}{\sqrt{n}}$$

Prior mass comparison implies that

$$\Pi(\alpha < 1 - \epsilon | \mathbf{x}^n) = o_p(1)$$

as soon as $\textcolor{red}{M} + \textcolor{red}{a}_1 \wedge 1/2 > 2\textcolor{red}{a}_1$

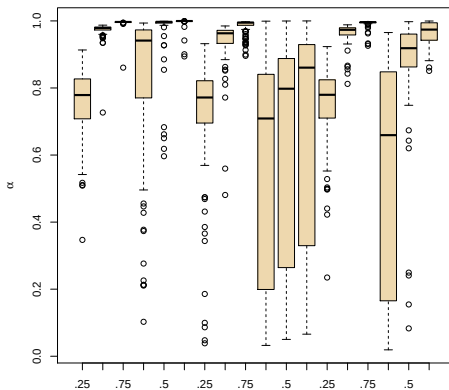


FIGURE – Boxplot of the posterior 25%, 50%, and 75% quantiles of α ; 100 replications of N simulations from (a) $\mathcal{N}(0, 1)$, ($N = 100$); (b) $\mathcal{N}(0, 1)$, ($N = 1000$); (c) a t_6 , ($N = 100$); (d) a t_6 ($N = 1000$); (e) a t_9 distribution ($N = 100$); (g) a t_9 distribution ($N = 1000$)

- Different interpretation of our tests : proportion of the data fitted by the model.
- closer to estimation
- valid theoretically
- For the moment iid or regression models. Not clear how to extend the idea to time series
- I wouldn't use it for variable selection in high dimension