# p-values and the double use of the data

J. Rousseau

ENSAE - CEREMADE, Université Paris-Dauphine

Applibugs

# Outline

1. $p$-values : some generalities
2. Nuisance parameters : troubles
3. Other candidate $p$-values
4. properties of conditional predicitive $p$-values
5. Efficient computation of $p_{cpred}$

# $p$-values : why should we use them....sometimes ?

- Calibration : ($u$-values)
  Statistic (test - model check) $T(X)$ :
  qu : What does the value $T(X) = T(x^o)$ tells us ?

# $p$-values : why should we use them....sometimes ?

- Calibration : ($u$-values)
  Statistic (test - model check) $T(X)$ :
  qu : What does the value $T(X) = T(x^o)$ tells us ?
- examples : Linear model

$$Y = X\beta + \epsilon, \quad \beta = (\beta_0, ..., \beta_p)^T, \quad x_i = (1, x_{i1}, ..., x_{ip})$$

• Test if covariate $x_{.j}$ is meaningful $\rightarrow \beta_j$ large .
Loss function

$$L(\beta, \delta) = \begin{cases} \left(\epsilon - \beta_j^2\right) \mathbf{1}_{\beta_j^2 \leq \epsilon} & \text{if} \quad \delta = 1 \\ \left(\beta_j^2 - \epsilon\right) \mathbf{1}_{\beta_j^2 > \epsilon} & \text{if} \quad \delta = 0 \end{cases}$$

Bayesian test $\delta^\pi = 1$ iff $E^\pi\left[\delta_j^2 | Y\right] > \epsilon$
choice of $\epsilon$ ? : calibration based on a $p$-value

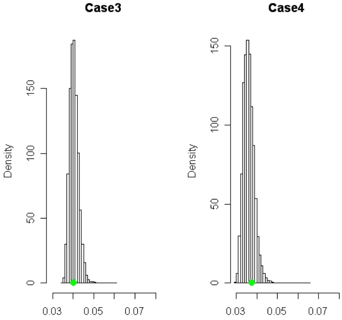
# Other examples : link with graphical model checks

Model checks $C$(data) = measure of discrepance between data and model : i.e. Elicitation : model for elicitation of quantiles

- $q_t^o$ : quantile elicited by expert, $t = 1, ..., T$
- $q_t(\theta)$ corresponding *theoretical* quantile

$$C(\text{data}) = \frac{1}{T}\sum_t (E^\pi[q_t(\theta)|\text{data}] - q_t^o)^2$$

*p*-value : summary of a graphic



# $p$-values : How not to use them

- (Sellke, berger, Bayarri) dangers for precise hypothesis testing : $D_i$ doses for different illnesses, $i = 1, ..., I$

$$H_{0i} : D_i \quad \text{inefficient} \quad H_{1i} : D_i \quad \text{efficient}$$

Model : $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $j = 1, ..., n_i$, $T_i(X) = \sqrt{n_i}|\bar{X}_i|/\sigma_i$
Simus : $\pi_0 I$ (50%) true null hypos, others $\mu_j \neq 0 \sim \mathcal{N}(0, a)$ or $\mathcal{U}(-a, a)$

- **p-value : not to be understood as proportion of false positive**
- **p-value : tail proba = rescaling in $\mathcal{U}(0, 1)$**

# $p$-values : How not to use them

- (Sellke, berger, Bayarri) dangers for precise hypothesis testing : $D_i$ doses for different illnesses, $i = 1, ..., I$

$$H_{0i} : D_i \quad \text{inefficient} \quad H_{1i} : D_i \quad \text{efficient}$$

Model : $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $j = 1, ..., n_i$, $T_i(X) = \sqrt{n_i}|\bar{X}_i|/\sigma_i$
Simus : $\pi_0 I$ (50%) true null hypos, others $\mu_j \neq 0 \sim \mathcal{N}(0, a)$ or $\mathcal{U}(-a, a)$

  - Of the $D_i$'s for which $p_i \approx 0.05$, $\geq 23\%$ are true $H_0$

- **p-value : not to be understood as proportion of false positive**
- **p-value : tail proba = rescaling in $\mathcal{U}(0, 1)$**

# $p$-values : How not to use them

- (Sellke, berger, Bayarri) dangers for precise hypothesis testing : $D_i$ doses for different illnesses, $i = 1, ..., I$

$$H_{0i} : D_i \quad \text{inefficient} \quad H_{1i} : D_i \quad \text{efficient}$$

  Model : $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $j = 1, ..., n_i$, $T_i(X) = \sqrt{n_i}|\bar{X}_i|/\sigma_i$
  Simus : $\pi_0 I$ (50%) true null hypos, others $\mu_j \neq 0 \sim \mathcal{N}(0, a)$ or $\mathcal{U}(-a, a)$
  - Of the $D_i$'s for which $p_i \approx 0.05$, $\geq 23\%$ are true $H_0$
  - Of the $D_i$'s for which $p_i \approx 0.01$, $\geq 7\%$ are true $H_0$

  ▶ **p-value : not to be understood as proportion of false positive**

  ▶ **p-value : tail proba = rescaling in $\mathcal{U}(0, 1)$**

# $p$-values : How not to use them

- (Sellke, berger, Bayarri) dangers for precise hypothesis testing : $D_i$ doses for different illnesses, $i = 1, ..., I$

$$H_{0i} : D_i \quad \text{inefficient} \quad H_{1i} : D_i \quad \text{efficient}$$

Model : $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $j = 1, ..., n_i$, $T_i(X) = \sqrt{n_i}|\bar{X}_i|/\sigma_i$
Simus : $\pi_0 I$ (50%) true null hypos, others $\mu_j \neq 0 \sim \mathcal{N}(0, a)$ or $\mathcal{U}(-a, a)$

  - Of the $D_i$'s for which $p_i \approx 0.05$, $\geq 23\%$ are true $H_0$
  - Of the $D_i$'s for which $p_i \approx 0.01$, $\geq 7\%$ are true $H_0$

► **p-value : not to be understood as proportion of false positive**

► **p-value : tail proba = rescaling in $\mathcal{U}(0, 1)$**

- Interesting interpretation

$$-ep \log p = \inf B_{0/1}$$

Needs uniform under the null to be valid

# Conclusion 1

- **Desirable properties of *p*-value :** *p*-value should be uniform (or close to uniform) under the null

• If a *p*-value is always conservative / anticonservative for freq. THEN also for Bayesian

$$P_\theta(p(X) \le p(x^o)|x^o) < p(x^o), \quad \forall\theta$$

Then

$$P^m(p(X) \le p(x^o)|x^o) < p(x^o)$$

# Nuisance parameters

• Model $P_\theta$, obs statistic $T(x^o) \rightarrow$ calibration

$$p = Pr\,[T(X) > T(x^o)]$$

under which proba $Pr$ ? common p-values

- $Pr = P_\theta$ impossible $\theta$ unknown

# Nuisance parameters

- Model $P_\theta$, obs statistic $T(x^o) \rightarrow$ calibration

$$p = Pr\,[T(X) > T(x^o)]$$

under which proba $Pr$ ? common p-values

- $Pr = P_\theta$ impossible $\theta$ unknown
- similar : $Pr = P[.|U]$ with $U$ sufficient stat. : not always possible

# Nuisance parameters

- Model $P_\theta$, obs statistic $T(x^o) \rightarrow$ calibration

$$p = Pr\,[T(X) > T(x^o)]$$

under which proba $Pr$ ? common p-values

- $Pr = P_\theta$ impossible $\theta$ unknown
- similar : $Pr = P[.|U]$ with $U$ sufficient stat. : not always possible
- prior : $Pr = \int_\Theta P_\theta[..]d\pi(\theta)$ impossible if $\pi$ improper

# Nuisance parameters

- Model $P_\theta$, obs statistic $T(x^o) \rightarrow$ calibration

$$p = Pr\left[T(X) > T(x^o)\right]$$

under which proba $Pr$ ? common p-values

- $Pr = P_\theta$ impossible $\theta$ unknown
- similar : $Pr = P[.|U]$ with $U$ sufficient stat. : not always possible
- prior : $Pr = \int_\Theta P_\theta[..]d\pi(\theta)$ impossible if $\pi$ improper
- plug : $Pr = P_{\hat{\theta}}[.]$ double use of dat a

# Nuisance parameters

- Model $P_\theta$, obs statistic $T(x^o) \rightarrow$ calibration

$$p = Pr\,[T(X) > T(x^o)]$$

under which proba $Pr$ ? common p-values

- $Pr = P_\theta$ impossible $\theta$ unknown
- similar : $Pr = P[.|U]$ with $U$ sufficient stat. : not always possible
- prior : $Pr = \int_\Theta P_\theta[..]d\pi(\theta)$ impossible if $\pi$ improper
- plug : $Pr = P_{\hat\theta}[.]$ double use of dat a
- post : $Pr = \int_\Theta P_\theta[.]d\pi(\theta|X)$ idem plug

# Double use of the data...

$p_{plug}$ and $p_{post}$ double use of the data :

- Finite sample example : Bayarri and Berger, 2000

$$X_i \sim \mathcal{N}(0, \sigma^2), \quad \theta = \sigma^2, \quad T(x) = |\bar{x}_n|, \quad \hat{\sigma}^2 = s^2 + \bar{x}_n$$

$$\pi(\sigma) = 1/\sigma$$

$$p_{plug}(x^o) = 2\left[1 - \Phi\left(\frac{\sqrt{n}|\bar{x}|^o}{\hat{\sigma}^o}\right)\right], \quad p_{post} = 2\left(1 - T_n\left(\frac{\sqrt{n}|\bar{x}|^o}{\hat{\sigma}^o}\right)\right)$$

If $|\bar{x}|^o / s^o \to +\infty$ : both are bounded from below

# Double use of the data...

$p_{plug}$ and $p_{post}$ double use of the data :

- Finite sample example : Bayarri and Berger, 2000

$$X_i \sim \mathcal{N}(0, \sigma^2), \quad \theta = \sigma^2, \quad T(x) = |\bar{x}_n|, \quad \hat{\sigma}^2 = s^2 + \bar{x}_n$$

$$\pi(\sigma) = 1/\sigma$$

$$p_{plug}(x^o) = 2\left[1 - \Phi\left(\frac{\sqrt{n}|\bar{x}|^o}{\hat{\sigma}^o}\right)\right], \quad p_{post} = 2\left(1 - T_n\left(\frac{\sqrt{n}|\bar{x}|^o}{\hat{\sigma}^o}\right)\right)$$

If $|\bar{x}|^o/s^o \to +\infty$ : both are bounded from below

- Asymptotic : Robins et al. $p_{plug}$ conservative , $p_{post}$ worse :

$$P_\sigma[p_{plug} \leq \alpha] < \alpha + o(1)$$

Not a good measure of model fit : too much in favour of model

# Other candidate $p$-values

- Conditional predictive

$$p_{cpred}(x^o) = \int_\Theta P_\theta\left[T(X) > T(x^o) | U = U(x^o), x^o\right] d\pi(\theta|U)$$

- No double use of data. True Bayesian $p$-value.
- If *good* $U$ $\pi(\theta|U)$. Importance of the choice of $U$

asymptotically equivalent....

# Other candidate $p$-values

- Conditional predictive

$$p_{cpred}(x^o) = \int_\Theta P_\theta\left[T(X) > T(x^o) | U = U(x^o), x^o\right] d\pi(\theta|U)$$

  - No double use of data. True Bayesian $p$-value.
  - If *good* $U$ $\pi(\theta|U)$. Importance of the choice of $U$

- partial predictive

$$P_{part}(x^o) = \int \mathbb{I}_{t \leq t(x^o)} f(t|\theta)\pi(\theta|x^o \setminus t^o) d\theta$$

asymptotically equivalent....

# Study of $p_{cpred}$

▶ **Choice of $U$ : $U = \hat{\theta}$MLE**

• Why ? : MLE asympt. suff statist.

$$P_\theta \left[ T(X) > T(x^o) | \hat{\theta} = \hat{\theta}(x^o), x^o \right] \approx \text{independent}(\theta)$$

• Result : Whatever $T(X)$

$$p_{cpred}(x) \approx \mathcal{U}(0,1) \quad \text{under } P_\theta$$

• Whatever $T$, influence of prior to order $n^{-3/2}$ only

• Good higher order properties also. if $T(X)$ asymptotically stable ($\mathcal{N}$)

$$p_{cpred} = \mathcal{U}(0,1) + O(n^{-3/2})$$

• $p_{cpred}(T^o)$ : implicit studentization of $T$ (pivotal)

# some other properties : good and bad

- **Equivalence with triple bootstrap** $p_{cpred}$ gives the same result as the following :

  1. *1 rst order bootstrap*

     $$p_1(T(x^o)) = P_{\hat{\theta}}\left[T(X) > T(x^o)\right] \overset{asy}{\neq} \mathcal{U}(0,1)$$

- **2 difficulties**

# some other properties : good and bad

- **Equivalence with triple bootstrap** $p_{cpred}$ gives the same result as the following :
  1. *1 rst order bootstrap*

     $$p_1(T(x^o)) = P_{\hat{\theta}}[T(X) > T(x^o)] \overset{asy}{\neq} \mathcal{U}(0,1)$$

  2. *2 nd order bootstrap* New : $T_2(X) = p_1(T(X))$, new PV

     $$p_2(T(x^o)) = P_{\hat{\theta}}[T_2(X) > T_2(x^o)] = \mathcal{U} + O(n^{-1}))$$

- **2 difficulties**

# some other properties : good and bad

- **Equivalence with triple bootstrap** $p_{cpred}$ gives the same result as the following :

  1. *1 rst order bootstrap*

     $$p_1(T(x^o)) = P_{\hat{\theta}}[T(X) > T(x^o)] \overset{asy}{\neq} \mathcal{U}(0,1)$$

  2. *2 nd order bootstrap* New : $T_2(X) = p_1(T(X))$, new PV

     $$p_2(T(x^o)) = P_{\hat{\theta}}[T_2(X) > T_2(x^o)] = \mathcal{U} + O(n^{-1}))$$

  3. *3 rd order bootstrap* $T_3(X) = p_2(T(X))$,

     $$p_3(T(x^o)) = P_{\hat{\theta}}[T_3(X) > T_3(x^o)] = p_{cpred}(x^o) + 0(n^{-3/2})$$

- **2 difficulties**

# some other properties : good and bad

- **Equivalence with triple bootstrap** $p_{cpred}$ gives the same result as the following :

1. *1 rst order bootstrap*

   $$p_1(T(x^o)) = P_{\hat{\theta}}[T(X) > T(x^o)] \overset{asy}{\neq} \mathcal{U}(0,1)$$

2. *2 nd order bootstrap* New : $T_2(X) = p_1(T(X))$, new PV

   $$p_2(T(x^o)) = P_{\hat{\theta}}[T_2(X) > T_2(x^o)] = \mathcal{U} + O(n^{-1}))$$

3. *3 rd order bootstrap* $T_3(X) = p_2(T(X))$,

   $$p_3(T(x^o)) = P_{\hat{\theta}}[T_3(X) > T_3(x^o)] = p_{cpred}(x^o) + 0(n^{-3/2})$$

- **2 difficulties**
  - Discrete observations

# some other properties : good and bad

- **Equivalence with triple bootstrap** $p_{cpred}$ gives the same result as the following :
  1. *1 rst order bootstrap*

     $$p_1(T(x^o)) = P_{\hat{\theta}}\left[T(X) > T(x^o)\right] \overset{asy}{\neq} \mathcal{U}(0,1)$$

  2. *2 nd order bootstrap* New : $T_2(X) = p_1(T(X))$, new PV

     $$p_2(T(x^o)) = P_{\hat{\theta}}\left[T_2(X) > T_2(x^o)\right] = \mathcal{U} + O(n^{-1}))$$

  3. *3 rd order bootstrap* $T_3(X) = p_2(T(X))$,

     $$p_3(T(x^o)) = P_{\hat{\theta}}\left[T_3(X) > T_3(x^o)\right] = p_{cpred}(x^o) + 0(n^{-3/2})$$

- **2 difficulties**
  - Discrete observations
  - computation

# General algorithm :

1. $\theta^j \sim \pi(\theta|\hat{\theta}), j = 1, ..., J$

- **Time consuming** In particular : $T(x^{(j)})$ for all $j$ !!! long if
$T(x) = E^{\pi_1}[h(\psi, \theta)|x]$
ex : $\eta = (\theta, \psi)$ with $H_0 : \psi = 0$ ($h(\eta) = \|\psi\|^2$)

# General algorithm :

1. $\theta^j \sim \pi(\theta|\hat{\theta})$, $j = 1, ..., J$
2. new data $x^{(j)}|\theta^j \sim f(x^{(j)}|\hat{\theta}, \theta^j)$

- **Time consuming** In particular : $T(x^{(j)})$ for all $j$ !!! long if
$T(x) = E^{\pi_1}[h(\psi, \theta)|x]$
ex : $\eta = (\theta, \psi)$ with $H_0 : \psi = 0$ ($h(\eta) = \|\psi\|^2$)

# General algorithm :

1. $\theta^j \sim \pi(\theta|\hat{\theta}), j = 1, ..., J$
2. new data $x^{(j)}|\theta^j \sim f(x^{(j)}|\hat{\theta}, \theta^j)$
3. for each $j$ compute $T(x^{(j)})$

- **Time consuming** In particular : $T(x^{(j)})$ for all $j$ ! ! ! long if $T(x) = E^{\pi_1}[h(\psi, \theta)|x]$
ex : $\eta = (\theta, \psi)$ with $H_0 : \psi = 0$ ($h(\eta) = \|\psi\|^2$)

# General algorithm :

1. $\theta^j \sim \pi(\theta|\hat{\theta}), j = 1, ..., J$
2. new data $x^{(j)}|\theta^j \sim f(x^{(j)}|\hat{\theta}, \theta^j)$
3. for each $j$ compute $T(x^{(j)})$
4. $$\hat{p}(x^o) = \frac{1}{J}\sum_{j=1}^{J} \mathbb{I}_{T(x^{(j)}>T(x^o)}$$

- **Time consuming** In particular : $T(x^{(j)})$ for all $j$ ! ! ! long if $T(x) = E^{\pi_1}[h(\psi, \theta)|x]$
ex : $\eta = (\theta, \psi)$ with $H_0 : \psi = 0$ $(h(\eta) = \|\psi\|^2)$

$$T(x^{(j)}) = E^{\pi_1}[h(\psi,\theta)|x^{(j)}], \quad j = 1, ..., J \quad \eta = (\theta,\psi)$$

- ONE MCMC : Compute $(\eta^t)_{t=1}^T$ MCMC for $\pi_1(\eta|x^{(1)})$

$$T(x^{(j)}) = E^{\pi_1}[h(\psi,\theta)|x^{(j)}], \quad j = 1, ..., J \quad \eta = (\theta, \psi)$$

- ONE MCMC : Compute $(\eta^t)_{t=1}^T$ MCMC for $\pi_1(\eta|x^{(1)})$
- Importance sampling : $j \geq 2$

$$E^{\pi}\left[h(\eta)|x^{(j)}\right] = \frac{\int h(\eta)\frac{f(x^{(j)}|\eta)}{f(x^{(1)}|\eta)} d\pi_1(\eta|x^{(1)})}{\int \frac{f(x^{(j)}|\eta)}{f(x^{(1)}|\eta)} d\pi_1(\eta|x^{(1)})}$$

so

$$\hat{T}(x^{(j)}) = \frac{\sum_{t=T_1}^{T} h(\eta^t) w(\eta^t; x^{(j)})}{\sum_{t=T_1}^{T} w(\eta^t; x^{(j)})}, \quad w(\eta^t; x^{(j)}) = \frac{f(x^{(j)}|\eta)}{f(x^{(1)}|\eta)}$$

# Possible improvements of IS

- **IS : weights (sometimes ) unstable** (large dimensions)
- **Simple re-centering**

$$\tilde{\eta}^t = \eta_t + \hat{\eta}(x^{(j)}) - \hat{\eta}(x^{(1)})$$

where $\hat{\eta}(x) = E^\pi(\eta|x)$, MLE ...

$$\tilde{w}(\tilde{\eta}) = \frac{\pi_1(\tilde{\eta}) f(x^{(j)}|\tilde{\eta})}{\pi_1(\eta) f(x^{(1)}|\eta)}$$

• result

$$Var\left(\frac{\tilde{w}(\tilde{\eta})}{\sum_t \tilde{w}(\tilde{\eta}_t)}\right) \approx 0$$

# Iterative algorithm

- MCMC : $(\eta^t)_t$ from $\pi_1(\eta|x^{(1)})$

# Iterative algorithm

- MCMC : $(\eta^t)_t$ from $\pi_1(\eta|x^{(1)})$
- simple IS (no re-centering) $w(\eta^t; x^{(j)})$ , $j = 2, ..., J$

$$\text{compute} \quad \hat{\eta}(x^{(j)}) = \frac{\sum_{t=T_1}^{T} \eta^t w(\eta^t; x^{(j)})}{\sum_{t=T_1}^{T} w(\eta^t; x^{(j)})}$$

# Iterative algorithm

- MCMC : $(\eta^t)_t$ from $\pi_1(\eta|x^{(1)})$
- simple IS (no re-centering) $w(\eta^t; x^{(j)})$ , $j = 2, ..., J$

$$\text{compute} \quad \hat{\eta}(x^{(j)}) = \frac{\sum_{t=T_1}^{T} \eta^t w(\eta^t; x^{(j)})}{\sum_{t=T_1}^{T} w(\eta^t; x^{(j)})}$$

- re-centering based on $\hat{\eta}(x^{(j)})$ :

$$\eta' = \eta + \hat{\eta}(x^{(j)}) - \hat{\eta}(x^{(1)})$$

Compute $\hat{T}(x^{(j)})$

# Iterative algorithm

- MCMC : $(\eta^t)_t$ from $\pi_1(\eta|x^{(1)})$
- simple IS (no re-centering) $w(\eta^t; x^{(j)})$ , $j = 2, ..., J$

$$\text{compute} \quad \hat{\eta}(x^{(j)}) = \frac{\sum_{t=T_1}^{T} \eta^t w(\eta^t; x^{(j)})}{\sum_{t=T_1}^{T} w(\eta^t; x^{(j)})}$$

- re-centering based on $\hat{\eta}(x^{(j)})$ :

$$\eta' = \eta + \hat{\eta}(x^{(j)}) - \hat{\eta}(x^{(1)})$$

  Compute $\hat{T}(x^{(j)})$
- reiterate if need be

# example : influence of covariates

$$X_i = \beta_0 + \beta Z_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1, ..., n$$

$$T(X) = E^{\pi}\left[\|\beta\|^2 | X\right]$$

Scatter plots of ‘Exact’ *p*-values against approximated *p*-values

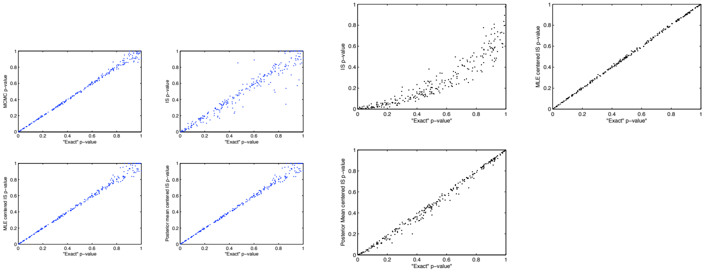



FIG.: Left : $H_1$ includes one covariate, Right : 9 covariates

# conclusion

- $p$-value = calibration = Uniform under a distribution of interest

# conclusion

- $p$-value = calibration = Uniform under a distribution of interest
- posterior predictive $p$-value : can be biased $\Rightarrow$ Does that mean that posterior predictive might tell us a different story that what we imagine

# conclusion

- $p$-value = calibration = Uniform under a distribution of interest
- posterior predictive $p$-value : can be biased $\Rightarrow$ Does that mean that posterior predictive might tell us a different story that what we imagine
- conditional predictive $p$-value : no (less) double use of the data

# conclusion

- $p$-value = calibration = Uniform under a distribution of interest
- posterior predictive $p$-value : can be biased $\Rightarrow$ Does that mean that posterior predictive might tell us a different story that what we imagine
- conditional predictive $p$-value : no (less) double use of the data
- choice of $U = MLE$

# conclusion

- $p$-value = calibration = Uniform under a distribution of interest
- posterior predictive $p$-value : can be biased $\Rightarrow$ Does that mean that posterior predictive might tell us a different story that what we imagine
- conditional predictive $p$-value : no (less) double use of the data
- choice of $U = MLE$
- computational issues : IS can be quite effective (prior sensitivity, also) improvemenst : re-cntering, multiple MCMC (Gajda et al.)