Quasi-Bayesian inference - pitfalls of incoherence

Jacek Osiewalski (Cracow University of Economics)

Bayesian analysis for a given statistical model:

- **probabilistic representation of initial uncertainty about all "unknowns" – not only about observations (available, missing, future) and latent variables, but also classical parameters (unknown constants)**
- **Bayesian model – joint probability (density) function** $\boldsymbol{p(y, \omega) = p(y \mid \omega)\, p(\omega)}$
- $\boldsymbol{p(y \mid \omega)}$ **– distribution of available observations given the remaining quantities**
- $\boldsymbol{p(\omega)}$ **– marginal (multivariate) distribution of all quantities that remain unknown after seeing the data (i.e., after seeing the realization of the vector** $\boldsymbol{y}$ **of available observations)**
- **Bayesian inference is based on simple, general rules of probability calculus**

1º conditioning – Bayes formula: $\boldsymbol{p(\omega \mid y) = \frac{p(y \mid \omega)\, p(\omega)}{p(y)} = \frac{p(y \mid \omega)\, p(\omega)}{\int_{\Omega} p(y \mid \omega)\, p(\omega)} \propto p(y \mid \omega)\, p(\omega)}$,

2º marginalization – deriving univariate distributions from $\boldsymbol{p(\omega \mid y)}$

“Coherent inference” – the one following strict rules of probability calculus

Quasi-Bayesian inference:

- **Bayes formula used mechanically, outside the full probabilistic context – incoherence !**
- $\boldsymbol{p(y \mid \omega) = g(y;\ \omega)}$ **corresponds to some traditional statistical model**
- $\boldsymbol{p(\omega) = f(\omega; y)}$ **is specified using given $\boldsymbol{y}$, so it cannot be the marginal distribution !!!**
- **thus $\boldsymbol{p(\omega \mid y) \propto\ g(y;\ \omega)\, f(\omega; y)}$ IS NOT the posterior in a Bayesian model with initially assumed $\boldsymbol{p(y \mid \omega)}$, but it can be the posterior in a completely different Bayesian model**
- **question: what are the *true* building blocks (statistical model and prior) corresponding to such $\boldsymbol{p(\omega \mid y)}$? it would be useful to know *true* assumptions, not only the *declared* ones**
- **fundamental pitfall of incoherence – $\boldsymbol{p(\omega \mid y)}$ corresponds to some statistical model and prior assumptions <u>to be discovered</u> !**

So-called “Empirical Bayes” (EB) is the most popular quasi-Bayesian approach, advocated in non-Bayesian, sampling-theory texts on inference in hierarchical multi-level statistical models

→ Here we show hidden assumptions behind the EB inference in hierarchical models

SOME SIMPLE EXAMPLE FIRST (Example 1)

Bayesian model: $p(y,\mu) = p(y \mid \mu)\, p(\mu) = f_N^n(y \mid \mu\, e_n, cI_n) f_N^1(\mu \mid a, v)$

Decomposition: $p(y,\mu) = p(y)\, p(\mu \mid y) = f_N^n(y \mid a\, e_n, cI_n + v\, e_n e_n') f_N^1(\mu \mid a_y, v_y)$

where $v_y = \left(\frac{n}{c} + \frac{1}{v}\right)^{-1}$, $a_y = \left(\frac{n}{c} + \frac{1}{v}\right)^{-1}\left(\frac{n}{c}\bar{y} + \frac{1}{v}a\right)$, $\bar{y} = \frac{1}{n}e_n' y$, $e_n = (1\ 1 \ldots 1)'$

Quasi-Bayesian inference: imagine a non-Bayesian statistician who agrees to use Bayes formula

$$p(\mu \mid y) \propto p(y \mid \mu)\, p(\mu)$$

but disagrees to subjectively specify a (prior mean); instead he/she puts $\bar{y}$ (sample average)
and (informally) uses $p^*(\mu) = f_N^1(\mu \mid \bar{y}, v)$ **and** $p^*(\mu \mid y) = f_N^1\left(\mu \mid \bar{y}, \left(\frac{n}{c} + \frac{1}{v}\right)^{-1}\right)$

Is there any hidden Bayesian model (sampling + prior) formally justifying such "posterior"?

Consider $\tilde{p}(y,\mu) = p(y \mid \mu)\, p^*(\mu) = f_N^n(y - \mu\, e_n \mid 0, cI_n)\, f_N^1(\mu - \bar{y} \mid 0, v)$

it decomposes into $\tilde{p}(\mu \mid y) = p^*(\mu \mid y)$ **and** $\tilde{p}(y) \propto \exp\left(-\frac{1}{2c} y' M y\right)$, $M = I_n - \frac{1}{n} e_n e_n'$

or $\tilde{p}(y \mid \mu) = f_N^n\left(y \,\middle|\, \mu\, e_n, c\left(I_n - \frac{c}{n(c+nv)} e_n e_n'\right)\right)$ **and** $\tilde{p}(\mu)$ **constant (!!!)**

true sampling model assumes dependence (equi-correlation); true prior is flat, improper

MAIN PART: Statistical models with hierarchical structure

conditional distribution of observations: $\boldsymbol{p(y|\theta) = g(y;\theta),\ \ y\epsilon Y, \theta\epsilon\Theta}$;

distribution of random parameters (latent variables): $\boldsymbol{f_0(\theta;\alpha),\ \ \alpha\epsilon A} \subseteq \mathbb{R}^{\boldsymbol{s}}$;

joint distribution (α fixed):

$$\boldsymbol{p(y|\theta)\, f_0(\theta;\alpha) = g(y;\theta)\, f_0(\theta;\alpha) = f_1(\theta|y;\alpha)\, h(y;\alpha)} \quad \leftarrow \textbf{decomposition}$$

$\boldsymbol{h(y;\alpha)}$ ← **marginal distribution of** $\boldsymbol{y}$

$\boldsymbol{f_1(\theta|y;\alpha) = \frac{g(y;\theta)\, f_0(\theta;\alpha)}{h(y;\alpha)} \ \propto\ g(y;\theta)\, f_0(\theta;\alpha)}$ ← **conditional distribution of** $\boldsymbol{\theta}$ **(Bayes formula)**

SIMPLE EXAMPLE OF A HIERARCHICAL MODEL (Example 2)

$\boldsymbol{\theta_i}$ – **unobservable characteristic, randomly distributed over** $\boldsymbol{n}$ **observed units** $(\boldsymbol{i = 1, \dots, n})$**,**

$\boldsymbol{\theta = (\theta_1 \dots \theta_n)', \ \ \theta_i \sim iiN(\alpha, d), \ \ d > 0}$ **known;**

$\boldsymbol{x_i = (x_{i1} \dots x_{im})', \ \ x_{ij} \sim iiN(\theta_i, c_0)}$ $(\boldsymbol{j = 1, \dots, m})$ **– independent measurements of** $\boldsymbol{\theta_i}$ **(**$\boldsymbol{c_0}$ **known)**

$\boldsymbol{y_i = \frac{1}{m} e'_m x_i = \bar{x}_{i.}}$ **– sufficient statistic (for fixed** $\boldsymbol{\theta_i}$**);** $\boldsymbol{y_i \sim iiN(\theta_i, c), \quad c = \frac{c_0}{m}, \ \ y = (y_1 \dots y_n)'}$

$$\boldsymbol{p(y|\theta) = f_N^n(y|\theta, cI_n), \ \ f_0(\theta; \alpha) = f_N^n(\theta|\alpha e_n, dI_n)}$$

Decomposition of the product $\boldsymbol{p(y|\theta)\, f_0(\theta; \alpha)}$ **into** $\boldsymbol{f_1(\theta|y; \alpha)\, h(y; \alpha)}$**, where**

$\boldsymbol{h(y; \alpha) = \int_{\mathbb{R}^n} p(y|\theta)\ f_0(\theta; \alpha)\, d\theta = f_N^n(y|\alpha e_n, (c + d) I_n)}$**,**

$\boldsymbol{f_1(\theta|y; \alpha) = f_N^n\left(\theta \middle| \frac{d^{-1}}{c^{-1}+d^{-1}} \alpha e_n + \frac{c^{-1}}{c^{-1}+d^{-1}} y, \ \frac{1}{c^{-1}+d^{-1}} I_n\right)}$ **(final precision = sample + prior)**

$\boldsymbol{E(\theta|y; \alpha) = w \cdot \alpha e_n + (1 - w) \cdot y, \quad w = \frac{d^{-1}}{c^{-1}+d^{-1}} \epsilon (0, 1)}$ **(**$\boldsymbol{w}$ **= prior precision / final precision)**

$\boldsymbol{E(\theta|y; \alpha)}$ **– point in** $\boldsymbol{\Theta = \mathbb{R}^n}$ **lying on the line segment between** $\boldsymbol{(\alpha\ \alpha \dots \alpha)'}$ **and** $\boldsymbol{(y_1\ y_2 \dots y_n)'}$

$\boldsymbol{f_1(\theta|y; \alpha)}$ **follows Bayes Theorem for any fixed** $\boldsymbol{\alpha}$**, so then we have coherence; but how to get** $\boldsymbol{\alpha}$**?**

Empirical Bayes (EB)

inference on $\boldsymbol{\theta}$ based on the conditional distribution $\boldsymbol{f_1(\theta|y;\alpha)}$ obtained using Bayes Theorem, BUT for some point estimate of unknown $\boldsymbol{\alpha\epsilon A}$, e.g., using so-called type II maximum likelihood:

$$\boldsymbol{\widehat{\alpha} = \widehat{\alpha}_{ML} = \arg\ \max\ L(\alpha;y) = \arg\ \max\ h(y;\alpha),\quad \alpha\epsilon A}$$

So EB uses $\quad \boldsymbol{\widehat{p}(\theta|y) = f_1(\theta|y,\widehat{\alpha})\ \propto\ p(y|\theta)f_0(\theta;\widehat{\alpha})}$,

i.e. the "posterior" corresponding to the "prior" with hyper-parameter based on $\boldsymbol{y}$!!!

EXAMPLE 2 (continued)

$$\boldsymbol{L(\alpha;y) = h(y;\alpha) = f_N^n(y|\alpha e_n,(c+d)I_n) = (2\pi\cdot\frac{c+d}{n})^{\frac{1}{2}}\ f_N^1\left(\alpha\middle|\overline{y},\frac{c+d}{n}\right)\ f_N^n(My|0,(c+d)I_n)},$$

$$\boldsymbol{\widehat{\alpha} = \widehat{\alpha}_{ML} = \overline{y} = \frac{1}{n}e_n'y\ ,\qquad M = I_n - \frac{1}{n}e_ne_n'},$$

$$\boldsymbol{\widehat{p}(\theta|y) = f_1(\theta|y,\widehat{\alpha}) = f_N^n(\theta|\widehat{\theta}_{EB}\ ,\ \frac{1}{c^{-1}+d^{-1}}I_n),\quad \widehat{\theta}_{EB} = \frac{d^{-1}}{c^{-1}+d^{-1}}\overline{y}e_n + \frac{c^{-1}}{c^{-1}+d^{-1}}y}$$

- **uncertainty about $\boldsymbol{\alpha}$ not taken into account**
- **obvious incoherence of inferences on $\boldsymbol{\theta}$**

Bayesian hierarchical model (BHM)

$p(y,\omega) = p(y,\theta,\alpha) = p(y|\theta)\, p(\theta|\alpha)\, p(\alpha),\;\; p(\alpha)$ **– the prior for** $\alpha \epsilon A$

$\omega = (\theta,\alpha)$**, conditional independence:** $y \perp \alpha \mid \theta$ **– leads to** $p(y|\omega) = p(y|\theta)$

$$p(y|\theta) = g(y;\theta),\;\; p(\theta|\alpha) = f_0(\theta;\alpha) \quad \textbf{– the same as in EB}$$

final decomposition of Bayesian model: $p(y,\theta,\alpha) = p(y)\, p(\theta,\alpha|y) = p(y)\, p(\alpha|y)\, p(\theta|y,\alpha)$

$$p(\theta|y,\alpha) = \frac{p(y|\theta)\, p(\theta|\alpha)}{p(y|\alpha)} = \frac{g(y;\theta)\, f_0(\theta;\alpha)}{h(y;\alpha)} = f_1(\theta|y;\alpha)$$

$$p(\alpha|y) = \frac{p(y|\alpha)\, p(\alpha)}{p(y)} = \frac{h(y;\alpha)\, p(\alpha)}{p(y)}$$

$$p(y) = \int_A p(y|\alpha)\; p(\alpha)\, d\alpha$$

Remarks:

- $p(\theta|y) = \int_A f_1(\theta|y;\alpha)\, p(\alpha|y)\, d\alpha$ **– uncertainty about α is formally taken into account**
- **Bayes Theorem is used twice: for latent variables (given parameters) and for parameters**

EXAMPLE 2 (continued) – Bayesian hierarchical model with:

$$p(y|\theta) = f_N^n(y|\theta, cI_n),\;\; p(\theta|\alpha) = f_N^n(\theta|\alpha e_n, dI_n),\;\; p(\alpha) = f_N^1(\alpha|a, v)$$

* $p(\theta) = \int_{-\infty}^{+\infty} p(\theta|\alpha)\, p(\alpha)\, d\alpha = f_N^n(\theta|ae_n, dI_n + ve_ne_n')$
* $p(\theta|y) \;\propto\; p(y|\theta)\, p(\theta) = f_N^n(y|\theta, cI_n)\, p(\theta)$

or, equivalently, $p(\theta|y) = \int_{-\infty}^{+\infty} p(\theta|y,\alpha)\, p(\alpha|y)\, d\alpha = \int_{-\infty}^{+\infty} f_1(\theta|y;\alpha)\, p(\alpha|y)\, d\alpha$

where $\quad p(\alpha|y) = f_N^1(\alpha|\left(\frac{n}{c+d}+\frac{1}{v}\right)^{-1}\left(\frac{n}{c+d}\bar{y}+\frac{a}{v}\right), \left(\frac{n}{c+d}+\frac{1}{v}\right)^{-1})$

Finally:

$$p(\theta|y) = f_N^n(\theta|\frac{c^{-1}}{c^{-1}+d^{-1}}y + \frac{d^{-1}}{c^{-1}+d^{-1}}\left(\frac{n}{c+d}+\frac{1}{v}\right)^{-1}\left(\frac{n}{c+d}\bar{y}+\frac{a}{v}\right)\cdot e_n,$$
$$\frac{1}{c^{-1}+d^{-1}}I_n + \left(\frac{n}{c+d}+\frac{1}{v}\right)^{-1}\left(\frac{d^{-1}}{c^{-1}+d^{-1}}\right)^2 e_ne_n').$$

If $v^{-1} \approx 0$**, then** $p(\alpha) \approx const$, $p(\theta) \propto$ **exp**$\left(-\frac{1}{2d}\;\theta' M\theta\right)$, $M = I_n - \frac{1}{n}e_ne_n'$, **and**

$p(\theta|y) \approx f_N^n(\theta|\hat{\theta}_{EB}, \frac{1}{c^{-1}+d^{-1}}I_n + \frac{c^2}{n(c+d)}e_ne_n');\quad \frac{c^2}{n(c+d)}e_ne_n'$ **- reflects uncertainty about α!**

If also n is large enough, then $p(\theta|y) \approx \hat{p}(\theta|y)$; asymptotically, incoherence does not matter

Small-sample interpretation of *Empirical Bayes*

For a given EB form of $\widehat{p}(\theta|y)$, we seek for $\widetilde{p}(y|\theta)$ and $\widetilde{p}(\theta)$ that lead to the Bayesian model $\widetilde{p}(y,\theta) = \widetilde{p}(y|\theta)\ \widetilde{p}(\theta)$ of the form

$$\widetilde{p}(y,\theta) = k(y)\, p(y|\theta)\, p(\theta|\alpha = \widehat{\alpha}) = k(y)\ g(y;\theta)\ f_0(\theta;\widehat{\alpha}),$$

resulting in $\widehat{p}(\theta|y)$ as the true posterior, i.e.

$$\widetilde{p}(\theta|y) = \widehat{p}(\theta|y) \ \propto\ g(y;\theta)\ f_0(\theta;\widehat{\alpha}).$$

From the form of $\widetilde{p}(y,\theta)$ we obtain the (implicit) prior $\widetilde{p}(\theta) = \int_Y k(y)\ g(y;\theta)\ f_0(\theta;\widehat{\alpha})\, dy$

and then the (implicit) conditional distribution of observations

$$\widetilde{p}(y|\theta) = \frac{\widetilde{p}(y,\theta)}{\widetilde{p}(\theta)} = k(y)\,\frac{g(y;\theta)\ f_0(\theta;\widehat{\alpha})}{\widetilde{p}(\theta)}.$$

If both f_0 and k are not constant in y, then $\widetilde{p}(y|\theta) \neq p(y|\theta) = g(y;\theta)$ and the *true* conditional distribution of observations is different from the *initially assumed* (declared) one.

EXAMPLE 2 (continued)

$$\tilde{p}(y,\theta) = k(y)g(y;\theta)\, f_0(\theta;\hat{\alpha}) = k\ \ f_N^n(y|\theta,\ cI_n)\, f_N^n(\theta|\bar{y}e_n,\ dI_n)$$

$$= f_N^n\left(y\middle|\theta,\left(\frac{1}{c}I_n + \frac{1}{dn}e_n e_n'\right)^{-1}\right)\ k(2\pi)^{-\frac{n}{2}}\exp(-\frac{1}{2d}\theta' M\theta)$$

From $\tilde{p}(y,\theta)$ we easily derive:

$\tilde{p}(\theta) = k(2\pi)^{-\frac{n}{2}}\exp\left(-\frac{1}{2d}\theta' M\theta\right)$ – improper (only σ–finite), but informative (favors equal θ_i)

(for $v^{-1} \approx 0$ we get $p(\theta) \approx \tilde{p}(\theta)$, so the *declared* prior coincides with the *true* one)

$$\tilde{p}(y|\theta) = f_N^n\left(y\middle|\theta,\left(\frac{1}{c}I_n + \frac{1}{dn}e_n e_n'\right)^{-1}\right)$$

– conditional distribution with equally correlated observations (instead of independent ones!!!)

$$\tilde{V}(y|\theta) = c\left(I_n - \frac{c}{n(c+d)}e_n e_n'\right) \quad \Rightarrow \quad \widetilde{Corr}\left(y_i, y_j\middle|\theta\right) = \frac{c}{(n-1)c+nd} \quad (i \neq j),$$

***true* $\tilde{p}(y|\theta)$ is <u>qualitatively</u> different from *declared* $p(y|\theta)$; problem disappears when $n \to \infty$**

Concluding remarks

- **From the purely Bayesian perspective, using Bayes formula with "prior" dependent on actual data is completely incoherent.**
- **Is this, however, of any interest to a non-Bayesian statistician? Perhaps such incoherent quasi-Bayesian approach generates inference tools that are better in terms of sampling-theory properties...**
- **Remind that, under certain regularity conditions, Bayesian decision functions (estimators) are admissible (cannot be improved – in terms of risk – uniformly in the parameter space) and form complete classes of such decision functions.**
- **Here it has been shown that incoherent, quasi-Bayesian approaches can be interpreted as Bayesian for <u>other sampling models, not for the *declared* (assumed) ones</u>.**
- **When a quasi-Bayesian procedure is not Bayesian for the declared sampling model, it may produce inadmissible decision functions (within this sampling model).**
- **Being coherent (i.e., being Bayesian and obeying rules of probability) prevents from such risks – and it does so for (almost) free...**

THANK YOU FOR YOUR ATTENTION!