Partially Bayes

Xiao-Li Meng

Partially Bayes

Bivariate Normal Example

Sequential vs Simultaneous

Partially Bayes: An Inter-Perspective Method for Handling Nuisance Parameters

Xiao-Li Meng

Department of Statistics, Harvard University

November 3, 2021

Partially based on Meng (2009). **Automated Bias-variance Trade-off: Intuitive Inadmissibility or Inadmissible Intuition?** In *Frontiers of Statistical Decision Making and Bayesian Analysis* (Eds: Chen et. al.), Springer; pp 95-112.

Partially Bayes – unwilling to use full Bayes

Partially Bayes
Xiao-Li Meng
Partially Bayes
Bivariate Normal Example
Sequential vs Simultaneous

- **Put prior only on the nuisance parameter.**
- Cox, D. R. (1975) "A note on partially Bayes inference and the linear model" *Biometrika*, **92**, 399-418.
- McCullagh, P. (1990) "A note on partially Bayes inference for generalized linear models." *Technical Report 284*, Department of Statistics, The University of Chicago.
- Mukherjee, B. and Chatterjee, N. (2008) "Exploiting gene-environment independence for analysis of case control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency." *Biometrics*, **64(3)**, 685 - 694.

Partially Bayes – unable to use full Bayes...

- **Too time/energy consuming to be a full Bayesian**
- **Only partial information is available for prior specification**
- Or "incompleteness specification" — "Partially Bayes Estimates" Daniel Solomon, 1969, FSU
- Sometimes it is impossible in principle ...

Likelihood inference for Monte Carlo Integration

$$C = E[g(X)] = \int g(x)p(x)\mu(dx).$$

- Estimand is C, but the likelihood parameter is the baseline measure μ (Kong *et al.* 2003, *JRSSB*).
- How do we put a prior on μ such that (say) $C \in [C_1, C_2]$?

Partial Shrinkage

- A great benefit of using Bayesian methods is *shrinkage*.
- What happens when we are partially Bayes? Do we get a *partial shrinkage*? How much is lost compared to full shrinkage?
- How strong does my prior need to be for it to be really beneficial (e.g., for bias-variance trade-off)?
- .....

Let us take a SNoTE ...

Full Likelihood/Sampling Model

$$\{Y_1, \ldots, Y_n\} \overset{\text{iid}}{\sim} Y = \begin{pmatrix} X \\ Z \end{pmatrix} \sim N\left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right]$$

Full Prior

$$\theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Omega = \begin{pmatrix} \tau^2 & 0 \\ 0 & \xi^2 \end{pmatrix}\right]$$

Full Posterior

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \Bigg| \begin{pmatrix} \bar{X}_n \\ \bar{Z}_n \end{pmatrix} \sim N\left[W_{\Omega,\rho} \begin{pmatrix} \bar{X}_n \\ \bar{Z}_n \end{pmatrix}, W_{\Omega,\rho} \left(\frac{\Sigma}{n}\right)\right]$$

where $W_{\Omega,\rho} = (\Omega^{-1} + n\Sigma^{-1})^{-1}(n\Sigma^{-1})$ is the *shrinkage factor.*

Infer the nuisance parameter α first, Bayesianly

Partial Likelihood for α

$$\{X_1, \ldots, X_n\} \overset{\text{iid}}{\sim} N(\alpha, 1)$$

Partial Prior for α

$$\alpha \sim N(0, \tau^2)$$

Partial Posterior for α

$$\alpha \big| \bar{X}_n \sim N\left(w_\tau \bar{X}_n, w_\tau n^{-1}\right),$$

where

$$w_\tau = \frac{n}{n + \tau^{-2}}$$

Given α, then infer β, non-Bayesianly

Regression Estimator (β on α)

$$\hat{\beta}(\alpha) = \bar{Z}_n + \rho(\alpha - \bar{X}_n) \equiv \hat{\beta}(0) + \rho\alpha$$

$\hat{\beta}(\alpha)$ is a known function of α; its posterior mean under $p(\alpha|\bar{X}_n)$ is

$$\begin{aligned} \hat{\beta}_\tau^{\text{part}} \equiv E[\beta(\alpha)|\bar{X}_n] &= \hat{\beta}(0) + \rho w_\tau \bar{X}_n \\ &= w_\tau \hat{\beta}^{\text{MLE}} + (1 - w_\tau)\hat{\beta}(0) \end{aligned}$$

- $w_\tau = \frac{n}{n+\tau^{-2}}$, shrinkage factor for α based on $\bar{X}_n$ alone
- $\hat{\beta}^{MLE} = \bar{Z}_n$, the MLE of β without prior information
- $\hat{\beta}(0) = \bar{Z}_n - \rho\bar{X}_n$, the MLE of β when $\alpha = 0$

Partially Bayes for Bias-Variance Trade-off

An Appealing Compromise

$$\hat{\beta}_\tau^{\text{part}} = w_\tau \hat{\beta}^{\text{MLE}} + (1 - w_\tau)\hat{\beta}(0)$$

where $w_\tau = \dfrac{n}{n + \tau^{-2}}$: **proportion of data information**

Beneficial trade-off: $MSE(\hat{\beta}_\tau^{\text{part}}) \le MSE(\hat{\beta}^{\text{MLE}})$ if and only if

$$\frac{1}{n}\left[1 - (1 - w_\tau^2)\rho^2\right] + (1 - w_\tau)^2\rho^2\alpha^2 \le \frac{1}{n}$$

$$(\textit{assume } \rho^2 > 0) \iff \left(\frac{\alpha}{\tau}\right)^2 \le 1 + w_\tau^{-1} \qquad (A)$$

If our prior $\alpha \sim N(0, \tau^2)$ is "correct", then:

- (A) holds with at least 84% prior probability;
- (A) holds with at least 95% prior probability if $w_\tau \le 1/3$.

Relative reduction on MSE achieved ...

Relative Reduction: $RR \equiv \frac{MSE(\hat{\beta}^{MLE}) - MSE(\hat{\beta}^{part}_{\tau})}{MSE(\hat{\beta}^{MLE})}$

$$RR = \rho^2(1 - w_\tau)\left[1 + w_\tau\left(1 - \left(\frac{\alpha}{\tau}\right)^2\right)\right]$$

- Maximal gain: $\max_\alpha RR = \rho^2(1 - w_\tau^2)$ (when $\alpha = 0$)
- Average gain : $E(RR) = \rho^2(1 - w_\tau)$ (under prior)

No Free Lunch ....

- Setting τ high increases the "correctness" of the prior but decreases the benefit of using the partial knowledge;
- Always a bias-variance/robustness-efficiency trade-off!

Automated Bias-variance Trade-off?

Deterministically, $MSE(\hat{\beta}^{\text{part}}_{\tau}) \leq MSE(\hat{\beta}^{\text{MLE}})$ if and only if

$$|\alpha| \leq \sqrt{2\tau^2 + n^{-1}}$$

- Knowledge about τ^2 is critical for achieving favorable bias-variance trade-off via $\hat{\beta}^{\text{part}}_{\tau}$;
- Small τ is risky, but large τ is not beneficial;
- How about letting the data to choose τ, such as empirical Bayes?
- Does not seem possible to estimate τ^2 by $\hat{\tau}^2$ from the *same* data and to guarantee (even for a given n)

$$R(\hat{\beta}^{\text{part}}_{\hat{\tau}}; (\alpha, \beta)) \leq R(\hat{\beta}^{\text{MLE}}; (\alpha, \beta))$$

Partially Bayes Risk

$$R(\hat{\beta};(\alpha,\beta)) = \int_y (\hat{\beta}(y)-\beta)^2 p(y|\alpha,\beta)\mu(dy).$$

Given a prior on α only, define Partially Bayes Risk

$$r_\pi(\hat{\beta};\beta) = \int R(\hat{\beta};(\alpha,\beta))\pi(d\alpha).$$

The key question then is what class of priors $\{\pi(\alpha)\}$ will render

$$r_\pi(\hat{\beta}^{\text{part}}_{\hat{\tau}};\beta) \le r_\pi(\hat{\beta}^{\text{MLE}};\beta), \quad \forall\beta?$$

- Such theoretical results help to provide practitioners with **principled guidelines** on when being partially Bayesian is beneficial.
- A few Ph.D. theses ...

Sequential Partially Bayes

- Imagine we have $\alpha \sim N(0, \tau^2)$ and *independently* $\beta \sim N(0, \zeta^2)$
- Sequential: Partially Bayes for α and then Bayes inference for β.

Given α, $\hat{\beta}(\alpha) = \bar{Z}_n + \rho(\alpha - \bar{X}_n)$ is sufficient for β.

$$\hat{\beta}(\alpha)|\beta \sim N\left(\beta,\ n_\rho^{-1}\right), \quad \beta \sim N(0, \zeta^2)$$

where $n_\rho = n/(1-\rho^2)$. Letting $w_{\zeta,\rho} = n_\rho/(n_\rho + \zeta^{-2})$, we have

$$\beta|\hat{\beta}(\alpha) \sim N\left(w_{\zeta,\rho}\hat{\beta}(\alpha), w_{\zeta,\rho}n_\rho^{-1}\right),$$

The Sequential Partially Bayes estimator for β

$$\beta^{\text{seqe}}_{\zeta\tau,\rho} = w_{\zeta,\rho}\left(\hat{\beta}(0) + w_\tau(\rho\bar{X}_n)\right) = w_{\zeta,\rho}\beta^{\text{part}}_\tau.$$

- **Is this same as the full Bayes estimator for β?**

Simultaneous Partially Bayes

Carrying out the following two procedures simultaneously:

- Partially Bayes for α and then Bayes inference for β.
- Partially Bayes for β and then Bayes inference for α.

Because $\mathrm{E}[\beta|\bar{X}_n, \bar{Z}_n] = w_{\zeta,\rho}\mathrm{E}\left[\hat{\beta}(\alpha)|\bar{X}_n, \bar{Z}_n\right]$

$$\mathrm{E}[\beta|\bar{X}_n, \bar{Z}_n] = w_{\zeta,\rho}\left(\hat{\beta}(0) + \rho\mathrm{E}[\alpha|\bar{X}_n, \bar{Z}_n]\right);$$

$$\mathrm{E}[\alpha|\bar{X}_n, \bar{Z}_n] = w_{\tau,\rho}\left(\hat{\alpha}(0) + \rho\mathrm{E}[\beta|\bar{X}_n, \bar{Z}_n]\right).$$

Solving them simultaneously leads to

$$\beta^{\text{full}}_{\tau,\zeta} = w_{\zeta\tau,\rho}\hat{\beta}^{\text{part}}_{\tau},$$

$$\text{with} \quad w_{\zeta\tau,\rho} = \frac{n_{\rho,\tau}}{n_{\rho,\tau} + \zeta^{-2}}, \quad \text{and} \quad n_{\rho,\tau} = \frac{n}{1 - \rho^2(1 - w_\tau)}.$$

- **Same as the full Bayes estimator for** β