EbayesThresh of Laplace Prior with Heterogeneous Variance

Kan Xu

April 3, 2017

Contents

1 Model

The Bayesian model extended the assumption of homogeneous variance of observations:

$$x_j|\mu_j, s_j \sim N(\mu_j, s_j^2) \tag{1.1}$$
$$\mu_j \sim (1-w)\delta_0(\mu_j) + w\gamma_a(\mu_j) \tag{1.2}$$

in which

$$\gamma_a(x) = \frac{1}{2}ae^{-a|x|}$$

is the Laplace prior with zero mean. $1-w$ is the weight of probability mass at zero.

The objective is to estimate μ_j with its posterior distribution given information observed (x_j and s_j). Posterior median and mean are used as estimators of μ_j.

1.1 Derivation

The posterior distribution of μ_j is

$$\mu_j|x_j, s_j \sim (1-w_p)\delta_0(\mu_j) + w_p f_p(\mu_j|x_j, s_j)$$

in which w_p is the posterior weight of non-zero means.

In the following analysis, ϕ is used to represent the density of a standard normal distribution, Φ the cdf of a standard normal, and $\tilde{\Phi} = 1 - \Phi$.

Define $g(x, s)$ to be the convolution of a normal distribution with standard deviation s and a Laplace distribution with parameter a.

$$\begin{aligned} g(x,s) &= \int_D \frac{1}{s}\phi(\frac{x-\mu}{s})\gamma_a(\mu)d_\mu \\ &= \frac{1}{2}ae^{\frac{a^2s^2}{2}}(e^{-ax}\Phi(\frac{x-s^2a}{s}) + e^{ax}\tilde{\Phi}(\frac{x+s^2a}{s})) \end{aligned}$$

in which D is the domain of μ.

$$\begin{aligned} w_p &= P(\mu_j \neq 0 | x_j, s_j) \\ &= \frac{w \cdot g(x,s)}{(1-w)\frac{1}{s}\phi(\frac{x}{s}) + w \cdot g(x,s)} \\ &= \frac{w(\beta(x,s)+1)}{1+w\beta(x,s)} \end{aligned}$$

in which

$$\begin{aligned} \beta(x,s) &= \frac{g(x,s)}{\frac{1}{s}\phi(\frac{x}{s})} - 1 \\ &= \frac{1}{2}as\left(\frac{\Phi(\frac{x-s^2a}{s})}{\phi(\frac{x-s^2a}{s})} + \frac{\tilde{\Phi}(\frac{x+s^2a}{s})}{\phi(\frac{x+s^2a}{s})}\right) - 1 \end{aligned}$$

The non-zero part of μ_j's posterior distribution is

$$f_p(\mu|x,s) = \begin{cases} \frac{e^{-ax}\frac{1}{s}\phi(\frac{\mu-(x-s^2a)}{s})}{e^{-ax}\Phi(\frac{x-s^2a}{s}) + e^{ax}\tilde{\Phi}(\frac{x+s^2a}{s})}, & \mu > 0 \\ \frac{e^{ax}\frac{1}{s}\phi(\frac{\mu-(x+s^2a)}{s})}{e^{-ax}\Phi(\frac{x-s^2a}{s}) + e^{ax}\tilde{\Phi}(\frac{x+s^2a}{s})}, & \mu < 0 \end{cases}$$

Mean of $f_p(\mu|x,s)$ is

$$\mu_m(x,s) = x - \frac{as^2(e^{-ax}\Phi(\frac{x-s^2a}{s}) - e^{ax}\tilde{\Phi}(\frac{x+s^2a}{s}))}{e^{-ax}\Phi(\frac{x-s^2a}{s}) + e^{ax}\tilde{\Phi}(\frac{x+s^2a}{s})}$$

when $x > 0$.

Thus, the posterior mean is $w_p \cdot \mu_m(x,s)$.

Posterior median will either have the same sign as the original value or be zero. Define

$$\begin{aligned} \tilde{F}_p(\mu|x,s) &= \int_\mu^\infty f_p(\mu|x,s)d\mu \\ &= \frac{e^{-ax}\tilde{\Phi}(\frac{\mu-(x-s^2a)}{s})}{e^{-ax}\Phi(\frac{x-s^2a}{s}) + e^{ax}\tilde{\Phi}(\frac{x+s^2a}{s})} \end{aligned}$$

Suppose $x > 0$. The posterior median $\mu_d(x,s)$ is the solution of

$$w_p\tilde{F}_p(\mu|x,s) = \frac{1}{2}, \text{ if } w_p\tilde{F}_p(0|x,s) > \frac{1}{2}$$

Otherwise, the posterior median is zero.

For the case of negative observations, we can transform them into their additive inverse, follow the same analysis as above and assign the correct sign to the posterior mean or median. Thus, assume x to be positive in the following analysis.

2 Algorithm

beta.laplace $\beta(x,s)$ is calculated. In practice, $\frac{\tilde{\Phi}(x)}{\phi(x)}$ is approximated by $\frac{1}{x}$ when $x > 35$, while $\frac{\Phi(x)}{\phi(x)}$ takes value $\frac{\Phi(35)}{\phi(35)}$ when $x > 35$ due to a limit of numerical accuracy R language can reach (or

it will take infinite value). The same approximation as $\frac{\tilde{\Phi}(x)}{\phi(x)}$ for $\frac{\Phi(x)}{\phi(x)}, x < -35$ is followed due to symmetry of these two functions.

postmean.laplace Posterior mean is calculated. $\Phi(x)$ and $\tilde{\Phi}(x)$ will take value at 35 when $x > 35$ in case that both functions take value 0 when standard deviation is large.

postmed.laplace Posterior median is calculated. If posterior median is positive,

$$\begin{aligned}
&w_p\tilde{F}_p(\mu|x,s) = \frac{1}{2}\\
\Leftrightarrow &\frac{e^{-ax}\tilde{\Phi}(\frac{\mu-(x-s^2a)}{s})}{e^{-ax}\Phi(\frac{x-s^2a}{s}) + e^{ax}\tilde{\Phi}(\frac{x+s^2a}{s})} = \frac{(1-w)\frac{1}{s}\phi(\frac{x}{s}) + w\cdot g(x,s)}{2wg(x,s)}\\
\Leftrightarrow &\tilde{\Phi}(\frac{\mu-(x-s^2a)}{s}) = (aw)^{-1}(1+w\beta(x,s))\frac{1}{s}\phi(\frac{x-s^2a}{s})\\
\Leftrightarrow &\mu_d(x,s) = x - s^2a - s\Phi^{-1}(z(x,s))
\end{aligned}$$

in which

$$z(x,s) = a^{-1}(w^{-1}+\beta(x,s))\frac{1}{s}\phi(\frac{x-s^2a}{s})$$

As $\frac{x-s^2a}{s} \to \infty$, $z(x,s)$ converges to $\frac{1}{2}$. This approximate value of $z(x,s)$ will be used when $\frac{x-s^2a}{s}$ is large ($\frac{x-s^2a}{s} > 25$ in practice) lest $\beta(x,s)$ become infinity and $\phi(\frac{x-s^2a}{s})$ go to zero. Notice that when x is small and s is small (or in some other cases, e.g. x is large and s is large), $z(x,s)$ can be larger than 1. Therefore,

$$\mu_d(x,s) = \max\{0, x - s^2a - s\Phi^{-1}(\min\{1, z(x,s)\})$$

tfromw(prior='laplace') When $x - s^2a - s\Phi^{-1}(z(x,s)) < 0$, the posterior median will be set to zero. Thus, there is a posterior median threshold $t(s,w,a)$ such that the estimate of μ is zero whenever $|x| < t(s,w,a)$. This threshold, t, satisfies

$$\begin{aligned}
0 &= t - s^2a - s\Phi^{-1}(z(t,s))\\
\Leftrightarrow &\Phi(\frac{t-s^2a}{s}) = z(t,s)
\end{aligned}$$

If there is no solution to the above equation or there is negative root to the above equation, t will be set to infinity. The root can be found by binary search in the interval $[0, 25s + s^2a]$ in practice since the left hand side goes to 1 while the right hand side goes to $\frac{1}{2}$ when x is large. Different from the original problem with homogeneous variance, threshold might vary by standard deviation, which means t_j might not be equal to t_i if $s_i \neq s_j$.

wfromt(prior='laplace') w can be written in terms of t, s and a given the above formula:

$$w(t,s,a) = (as\frac{\Phi(\frac{t-s^2a}{s})}{\phi(\frac{t-s^2a}{s})} - \beta(t,s))^{-1} \tag{2.1}$$

w is monotonically declining with t, given s and a (which I observed from plots with different s and a, but still needs rigorous proof). Note that the value of w might be different for different threshold and standard deviation.

wandafromx w and a will be estimated using marginal log likelihood maximization in the empirical bayes sense if not provided. The marginal log likelihood is

$$\begin{aligned}
l(w,a|x_j,s_j,j=\{1,2,\cdots,n\}) &= log(\Pi_{i=1}^n f(x_j|w,a,s_j))\\
&= \sum_{i=1}^n log((1-w)\frac{1}{s_j}\phi(\frac{x_j}{s_j}) + w\cdot g(x_j,s_j))
\end{aligned}$$

To maximize the above log likelihood function is the same as to maximize

$$S(w,a|x_j,s_j,j=\{1,2,\cdots,n\}) = \sum_{i=1}^n log(1+w\beta(x_i,s_i))$$

Each threshold t_j is required to be ex-ante upper bounded by universal threshold, $0 < t_j < s_j\sqrt{2log(n)}$, which does not allow a very large probability mass at zero. Thus, w is constrained by the intersection of the range of $w(t_j, s_j, a)$ in Equation 2.1, given the domain of t_j:

$$w \in W(a, s_j, j = \{1, 2, \cdots, n\}) = \bigcap_{j=1}^{n} \{w(t_j, s_j, a) : 0 < t_j < s_j\sqrt{2log(n)}\}$$

I argue that the optimization problem with constraints can be solved by solving the following problem instead:

$$\max_{t_m, a} \; S(t_m, a | x_j, s_j, j = \{1, 2, \cdots, n\}) = \sum_{i=1}^{n} log(1 + w(t_m, s_m, a)\beta(x_i, s_i))$$

where $m = argmin_{k=1,2,\cdots,n} \{s_k\}$.

Thus, $\hat{w} = w(\hat{t}_m, s_m, \hat{a})$. (Still thinking about this part. My original argument has some problems.)

3 Question

1. It seems that the conception of 'thresholding' only exists when posterior median is treated as estimator. However, the w is estimated based on the constraint that the threshold is controlled by universal upper bound $s_j\sqrt{2log(n)}$ for all methods (posterior mean, median, hard and soft thresholding).

2. The universal threshold constraint $0 < t_j < s_j\sqrt{2log(n)}$ seems to be only used to search for the optimal w. When taking posterior median as the estimator, the actual threshold used might not be bounded by this universal threshold. For example, if some x is large and its posterior median is zero, the threshold should be at least as large as x. However, the article doesn't discuss whether the threshold used here is controlled by universal threshold or not.