Bayesian Statistics
Hierarchical Model and Empirical Bayes

Shaobo Jin

Department of Mathematics

Hierarchical Model

It is often the case that we can easily define a probabilistic model through several levels of conditional distribution, instead of a joint distribution.

- One trivial example is our Bayesian analysis: we specify $f(x \mid \theta)$ and $\pi(\theta)$, instead of $f(x, \theta)$.

We can further introduce hierarchical modeling to

- $f(x \mid \theta)$ to model complex structure in x.
- $\pi(\theta)$ as hierarchical prior.

Incomplete Prior

It is often the case that we have some prior information, but it is not enough for us to fully determine the prior.

Example

For example, we want to specify a beta prior for success probability. Our prior information suggests that such probability is around 0.3.

We need to take the uncertainty in specifying the prior into account in our Bayesian modeling such as

$$\begin{aligned} x \mid \theta &\sim f(x \mid \theta), \\ \theta \mid \lambda &\sim \pi(\theta \mid \lambda), \\ \lambda &\sim \pi(\lambda). \end{aligned}$$

More levels in the prior can be introduced similarly.

Hierarchical Prior: Uncertainty in Prior

Example

Suppose that $X \mid \theta \sim \text{Binomial}(n, \theta)$. We want to use a beta prior $\theta \sim \text{Beta}(a_0, b_0)$. Instead of specifying the values of a_0 and b_0 such that $a_0/(a_0 + b_0) = 0.3$ directly, we reparametrize the beta density into

$$\pi(\theta) = \frac{1}{B(\mu\kappa, (1-\mu)\kappa)} \theta^{\mu\kappa - 1} (1-\theta)^{(1-\mu)\kappa - 1},$$

such that $a_0 = \mu\kappa$ and $b_0 = (1-\mu)\kappa$.

- Specifying the values of a_0 and b_0 is equivalent to specifying the values of μ_0 and κ_0 directly.
- Instead, we introduce a prior $\pi(\mu, \kappa)$ to μ and κ.

Prior Distribution

Suppose that

$$\begin{aligned} x \mid \theta &\sim f(x \mid \theta), \\ \theta \mid \lambda &\sim \pi(\theta \mid \lambda), \\ \lambda &\sim \pi(\lambda). \end{aligned}$$

- If λ is known (e.g., we specify its value direct), then $\pi(\theta \mid \lambda)$ is the usual prior. If λ is unknown, then $\pi(\lambda)$ is a prior on a prior (i.e., hyperprior).
- The prior $\pi(\theta)$ is obtained by

$$\pi(\theta) = \int \pi(\theta \mid \lambda)\, \pi(\lambda)\, d\lambda.$$

Hierarchical Prior: Another Example

A general characteristic of the hierarchical prior is that it introduces robustness.

Example

Suppose that $Y \mid \theta \sim N\left(\theta, \sigma^2\right)$, where σ^2 is known. The conjugate prior is $\theta \sim N\left(\mu_0, \lambda_0^{-1}\right)$. Instead of specifying the prior for θ directly, we consider the hierarchical prior

$$\begin{aligned} \theta \mid \tau^2 &\sim N\left(\mu_0, \tau^2\right), \\ \tau^2 &\sim \text{InvGamma}\left(a_0, b_0\right). \end{aligned}$$

This hierarchical prior turns out to be equivalent to specifying a t distribution prior on θ, which has a heavier tail than a normal prior.

Hierarchical Prior: One More Example

Example

Suppose that we want to model data from different countries

$$Y_{ij} \mid \theta_j \sim N\left(\theta_j, \sigma^2\right),$$

where Y_{ij} is the ith observation from the jth country. It is often natural to assume that data from each country is draw from a bigger population, and we assume

$$\theta_j \mid \mu \sim N\left(\mu, \omega^2\right).$$

But we don't know μ, so we specify a prior for it $\mu \sim \pi(\mu)$. This idea is Bayesian meta analysis that allows us to combine similar studies conducted by different entities.

Full Bayesian Treatment

Consider again

$$\begin{aligned} x \mid \theta &\sim f(x \mid \theta), \\ \theta \mid \lambda &\sim \pi(\theta \mid \lambda), \\ \lambda &\sim \pi(\lambda). \end{aligned}$$

Neither θ nor λ is known. A full Bayesian treatment of such hierarchical model is based on the joint prior

$$\pi(\theta, \lambda) = \pi(\theta \mid \lambda)\pi(\lambda)$$

and the joint posterior

$$\pi(\theta, \lambda \mid x) \propto f(y \mid \theta, \lambda)\pi(\theta, \lambda) = f(y \mid \theta)\pi(\theta \mid \lambda)\pi(\lambda).$$

For example, we can use MCMC to draw posterior samples from $\pi(\theta, \lambda \mid x)$.

Hierarchical Prior: Example

Example

Suppose that $X \mid \theta \sim \text{Binomial}(n, \theta)$. Consider the hierarchical prior

$$\begin{aligned} \theta \mid \mu, \kappa &\sim \text{Beta}\left(\mu\kappa,\ (1-\mu)\,\kappa\right), \\ \mu &\sim \text{Uniform}\left[0, 1\right], \\ \kappa &\sim \text{Exp}\left(1\right), \end{aligned}$$

where the priors of μ and κ are independent. Find the posterior distributions.

Empirical Bayes

Instead of directly introducing a hyperprior on θ and apply the full Bayesian treatment, the empirical Bayes approach estimates the unknown hyperparameters from the marginal distribution and use the Bayes formula treating $\hat{\pi}$ as a prior.

- Suppose that $x \mid \theta \sim f(x \mid \theta)$ and $\theta \mid \lambda \sim \pi(\theta \mid \lambda)$, where λ is unknown.
- If λ were known, the posterior is the usual

$$\pi(\theta \mid x) \quad \propto \quad f(x \mid \theta)\, \pi(\theta \mid \lambda).$$

- The empirical Bayes approach estimates λ from $f(x \mid \lambda)$, and apply the "plug-in principle" by treating $\pi\left(\theta \mid \hat{\lambda}\right)$ as the prior of θ.

Empirical Bayes: Example

Example

Find a point estimator of θ using empirical Bayes.

1. Suppose that we observe independent $X_i \mid \theta_i \sim \text{Binomial}(m, \theta_i)$, $i = 1, ..., n$. Consider the independent prior $\theta_i \mid a, b \sim \text{Beta}(a,\ b)$. The posterior is $\theta_i \mid x_i \sim \text{Beta}(a + x_i, b + m - x_i)$. If a and b are known, the posterior mean is

$$\text{E}[\theta \mid x, a, b] \quad = \quad \frac{a + x}{a + b + m}.$$

2. Suppose that $x \mid \theta \sim N_p(\theta, I_p)$ and the prior distributions ofθ_i are independent $N\left(0, \tau^2\right)$. If τ^2 is known, the posterior is

$$\pi(\theta \mid x) \quad \sim \quad N\left(\frac{\tau^2}{\tau^2 + 1} x,\ \frac{\tau^2}{\tau^2 + 1} I_p\right).$$

Hierarchical Data

The hierarchical structure is not only limited to prior. We can also use the conditional distributions to help us specify the likelihood part.

- A natural example is the Gaussian mixture model, where

$$f\left(x \mid \theta\right) = pN\left(\mu_1, \sigma_1^2\right) + \left(1-p\right) N\left(\mu_2, \sigma_2^2\right).$$

- This likelihood is generally difficult to handle. But we can introduce a latent variable $Z \sim \text{Bernoulli}\left(p\right)$ and specify the likelihood as

$$\begin{aligned} f\left(x \mid z, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) &= \left[N\left(\mu_1, \sigma_1^2\right)\right]^z \left[N\left(\mu_2, \sigma_2^2\right)\right]^{1-z}, \\ f\left(z \mid p\right) &= p^z \left(1-p\right)^{1-z}. \end{aligned}$$

Gaussian Mixture Model

Consider the Gaussian mixture model

$$f(x \mid \theta) = pN\left(\mu_1, \sigma_1^2\right) + (1-p)\, N\left(\mu_2, \sigma_1^2\right).$$

In a full Bayesian treatment, we can develop a Gibbs sampler to sample from the posterior

$$f\left(z, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p \mid x\right).$$

However, MCMC will encounter various problems.

- Label switching: for simplicity let $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the likelihood of $\left(p, \mu_1, \mu_2, \sigma^2\right) = \left(0.2, 0, -1, \sigma^2\right)$ will produce the same likelihood as $\left(p, \mu_1, \mu_2, \sigma^2\right) = \left(0.8, -1, 0, \sigma^2\right)$.
- Component collapsing: a mixture of K normals can yield $\mu_i = \mu_j$ and $\sigma_i = \sigma_j$ even $i \neq j$.

Latent Variable Model

The finite mixture model is a special case of a latent variable model, where the mixture indicators $\{z_i\}$ are latent. Another example of a latent variable model is the measurement error model.

- Suppose that we want to observe $z \sim f(z \mid \lambda)$, but we only observe $x = z + e$, where e is the measurement error.
- If z observed, then the usual Bayes analysis yields

$$\pi(\lambda \mid z) \quad \propto \quad f(z \mid \lambda)\, \pi(\lambda).$$

- Since only x is observed, we need to use a hierarchical model

$$\pi(z, \lambda \mid x) \quad \propto \quad f(x \mid z)\, f(z \mid \lambda)\, \pi(\lambda),$$

where we have treated z as nuisance parameters.

Combining Different Models

A different view relative to model selection is to combine the contributions of several models as in ensemble learning. Let Δ be the quantity of interest such as

- average treatment effect of a drug,
- a future value.

Bayesian model selection chooses a model k^* using $\mathrm{P}(\mathcal{M}_k \mid x)$ and estimates Δ by

$$\hat{\Delta} = \mathrm{E}[\Delta \mid x, \mathcal{M}_{k^*}].$$

Bayesian model averaging (BMA) takes a weighted average instead as

$$\hat{\Delta} = \sum_k \mathrm{E}[\Delta \mid x, \mathcal{M}_k] \mathrm{P}(\mathcal{M}_k \mid x).$$

Posterior of Δ

The posterior of Δ is given by

$$f(\Delta \mid x) = \sum_k f(\Delta \mid \mathcal{M}_k, x) \, \mathrm{P}(\mathcal{M}_k \mid x),$$

where $f(\Delta \mid \mathcal{M}_k, x)$ is the posterior of Δ under model k. The posterior mean of Δ is

$$\begin{aligned} \mathrm{E}[\Delta \mid x] &= \int \Delta \sum_k f(\Delta \mid \mathcal{M}_k, x) \, \mathrm{P}(\mathcal{M}_k \mid x) \, d\Delta \\ &= \sum_k \mathrm{P}(\mathcal{M}_k \mid x) \underbrace{\int \Delta f(\Delta \mid \mathcal{M}_k, x) \, d\Delta}_{=\mathrm{E}[\Delta \mid x, \mathcal{M}_k]}, \end{aligned}$$

which is the BMA estimator of Δ.

Three Scenarios

The posterior probability

$$\mathrm{P}\left(\mathcal{M}_k \mid x\right) \quad = \quad \mathrm{P}\left(\mathcal{M}_k \text{ is the true model} \mid x\right)$$

can be strange if all candidate models are wrong, especially when we need to specify the prior probability that $\mathcal{M}_k$ is the true model.

1. The $\mathcal{M}$-closed setting means that one of the candidate models is the true data generating process.
2. The $\mathcal{M}$-complete setting means that but the true data generating process can be conceptualized, but it is not one of the candidate models due to, for example, model complexity or lack of information.
3. The $\mathcal{M}$-open setting means that the data generating process cannot be conceptualized and all candidate models are wrong.

Bayesian Stacking

Let $S(P, Q)$ be a scoring rule to measure the similarity between two probability measure P and Q. Let p and q be the corresponding densities. Then,

$$S(P, Q) = \int s(P, w)\, q(w)\, dw,$$

for some function $s(\cdot, \cdot)$. Bayesian stacking maximizes such similarity

$$S\left(\sum_k w_k f(\tilde{x} \mid x, \mathcal{M}_k),\ f_{\text{true}}(\tilde{x} \mid x)\right)$$

with respect to weights $\{w_k\}$ under the restriction that

$$\sum_k w_k = 1,\ 0 \leq 1 w_k \leq 1,\ \forall k.$$

Scoring Rule

Two commonly used scoring rules are

1. log score: $s(P, x) = \log p(x)$ such that $S(Q, Q) - S(P, Q)$ is the KL divergence.
   - Taking $p(x) = \sum_k w_k f(\tilde{x} \mid x, \mathcal{M}_k)$ is the same as maximizing the similarity between the stacked predictive distribution and the true predictive distribution.
2. energy score: $s(P, x) = \frac{1}{2}\mathrm{E}_P\left[\left\|X - \tilde{X}\right\|^\beta\right] - \mathrm{E}_P\left[\|X - x\|^\beta\right]$, where X and $\tilde{X}$ are two iid random variables, and the expectations are taken with respect to P.
   - If $\beta = 2$, it reduces to $s(P, x) = -\|\mathrm{E}_P[X] - x\|^2$.
   - Maximizing the scoring rule is equivalent to minimizing the squared prediction error.

Leave-One-Out Cross Validation

However, we don't know $f_{\text{true}}(\tilde{x} \mid x)$ that is needed to evaluate

$$S\left(\sum_k w_k f(\tilde{x} \mid x, \mathcal{M}_k),\ f_{\text{true}}(\tilde{x} \mid x)\right).$$

One alternative is to use leave-one-out cross validation as

$$\min_w \frac{1}{n}\sum_{i=1}^{n} s\left(\sum_k w_k f(x_i \mid x_{-i}, \mathcal{M}_k),\ x_i\right),$$

where

$$f(x_i \mid x_{-i}, \mathcal{M}_k) = \int f(x_i \mid \theta_k, \mathcal{M}_k)\, \pi(\theta_k \mid x_{-i}, \mathcal{M}_k)\, d\theta_k.$$

The stacked estimate of the predictive density is

$$\hat{f}(\tilde{x} \mid x) = \sum_k \hat{w}_k f(\tilde{x} \mid x, \mathcal{M}_k).$$

BMA and Bayesian Stacking

For BMA, it is alleged that, as $n \to \infty$,

- if one of the candidate models is the true model, say $\mathcal{M}_{k^*}$ is the true model,
- or, if all candidate models are misspecified and $\mathcal{M}_{k^*}$ has the smallest Kullback-Leibler divergence to the true model,

then $\mathrm{P}(\mathcal{M}_{k^*} \mid x) \to 1$.

In contrast to Bayesian model averaging,

- no prior $\mathrm{P}(\mathcal{M}_k)$ is needed in Bayesian stacking.
- Bayesian stacking is intended for the case where all candidate models are misspecified.

Importance Sampling

It is computationally intensive to compute the leave-one-out (LOO) predictive density

$$f\left(x_i \mid x_{-i}, \mathcal{M}_k\right) = \int p\left(x_i \mid \theta_k, \mathcal{M}_k\right) \pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right) d\theta_k$$

for each i, because we have to refit $\mathcal{M}_k$ n times to obtain all $\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)$.

Suppose that, for each k, we fit $\mathcal{M}_k$ using all the data and obtain L draws from the posterior $\pi\left(\theta_k \mid x, \mathcal{M}_k\right)$. Then,

$$f\left(x_i \mid x_{-i}, \mathcal{M}_k\right) = \int p\left(x_i \mid \theta_k, \mathcal{M}_k\right) \frac{\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)}{\pi\left(\theta_k \mid x, \mathcal{M}_k\right)} \pi\left(\theta_k \mid x, \mathcal{M}_k\right) d\theta_k,$$

where the importance weight is

$$w_i\left(\theta_k, \mathcal{M}_k\right) = \frac{\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)}{\pi\left(\theta_k \mid x, \mathcal{M}_k\right)}.$$

Normalized Importance Sampling

We can rewrite the importance weight as

$$w_i\left(\theta_k, \mathcal{M}_k\right) = \frac{\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)}{\pi\left(\theta_k \mid x, \mathcal{M}_k\right)} \propto \frac{f\left(x_{-i} \mid \theta_k, \mathcal{M}_k\right)\pi\left(\theta_k \mid \mathcal{M}_k\right)}{f\left(x \mid \theta_k, \mathcal{M}_k\right)\pi\left(\theta_k \mid \mathcal{M}_k\right)} \propto \frac{1}{f\left(x_i \mid \theta_k, \mathcal{M}_k\right)}.$$

The normalized importance sampling estimator is

$$\hat{f}^{\mathrm{NIS}}\left(x_i \mid x_{-i}, \mathcal{M}_k\right) = \frac{\sum_{l=1}^{L} w_i\left(\theta_k^{(l)}, \mathcal{M}_k\right) p\left(x_i \mid \theta_k, \mathcal{M}_k\right)}{\sum_{l=1}^{L} w_i\left(\theta_k^{(l)}, \mathcal{M}_k\right)},$$

where we sample $\theta_k^{(l)}$, $l = 1, ..., L$, from $\pi\left(\theta_k \mid x, \mathcal{M}_k\right)$.

However, the importance weights can be unstable if the distribution has a long tail that makes the importance weight very large.

Generalized Pareto Distribution

Theorem

Under suitable conditions on the random variable X, if the threshold u_0 is high enough, the conditional distribution of $X \mid X > u$ converges to a three-parameter generalized Pareto distribution (GPD), as $u \to \infty$. Its density is given by

$$f(x; u, \sigma, k) = \begin{cases} \frac{1}{\sigma}\left[1 + \frac{k(x-u)}{\sigma}\right]^{-1-1/k}, & \text{if } c \neq 0, \\ \frac{1}{\sigma}\exp\left(\frac{x-u}{\sigma}\right), & \text{if } c = 0, \end{cases}$$

for $y > u$ and $\sigma > 0$.

Pareto Smoothed Importance Sampling

Pareto Smoothed Importance Sampling stabilizes the large importance weights. Without loss of generality, suppose that $\left\{w_i\left(\theta_k^{(l)}, \mathcal{M}_k\right)\right\}$ has been ordered in increasing order.

- Consider the largest $N = \left\lfloor \min\left(0.2L,\, 3\sqrt{L}\right)\right\rfloor$ importance weights.
- We fit a GPD to $\left(w_i\left(\theta_k^{(L-N+1)}, \mathcal{M}_k\right), ..., w_i\left(\theta_k^{(L)}, \mathcal{M}_k\right)\right)$ with $u = w_i\left(\theta_k^{(L-N)}, \mathcal{M}_k\right)$.
- These N tail importance weights are replaced by

$$\min\left\{F^{-1}\left(\frac{z-1/2}{M}\right),\ \max_i w_i\left(\theta_k^{(l)}, \mathcal{M}_k\right)\right\}, \qquad z = 1, ..., M,$$

where F^{-1} is the inverse distribution function of the fitted GPD. The other importance weights are unchanged.