Advanced Bayesian Statistics

Yuan Liao
Department of Economics
Rutgers University

May 1, 2020

Contents

1 Bayesian variable selection

1.1 Zellner's g-priors

$$Y = X\beta + e, \quad var(e) = \sigma^2$$
$$p(\beta|\sigma) \sim N(b, g\sigma^2(X'X)^{-1})$$

$$P(\beta|\sigma^2, D) \sim N(\frac{g}{g+1}(\frac{b}{g} + \hat{\beta}), \frac{\sigma^2 g}{g+1}(X^T X)^{-1})$$

posterior mean

$$\frac{g}{g+1}(\frac{b}{g} + \hat{\beta})$$

as $g \to \infty$, prior becomes flat, posterior mean becomes OLS. Var becomes same as OLS.

1.2 setting g using empirical bayes

- Let $p(D|g)$ be the marginal likelihood
- no need prior for g.

$$\hat{g}_{EB} = \arg\max_g p(D|g)$$

why this is called EB ?

$$p(D|g) = \int \underbrace{p(D|\beta, \sigma^2, g)}_{likelihood} p(\beta, \sigma^2|g) d\beta d\sigma^2$$

So marginal likelihood integrate out other parameters.

- $$p(D|g) \propto \frac{(1+g)^d}{(1+g(1-R^2))^m}$$

where $d = \frac{n-1-p}{2}$ and $m = \frac{n-1}{2}$

$$\hat{g} = \max\{\frac{d - m(1-R^2)}{(m-d)(1-R^2)}, 0\}$$

- need $R^2 > \frac{p}{n-1}$. problematic if $R^2 \approx 0$.
- As $n \to \infty$, $\hat{g} \to \infty$

1.3 Spike and Slab priors

Let γ be a vector of indicators, where $\gamma_i = 0$ iff $\beta_i = 0$.
The "working model" is

$$Y = X_\gamma \beta_\gamma + e$$

We can consider priors

$$p(\beta_\gamma, \sigma^2, \gamma) = p(\beta_\gamma|\sigma^2, \gamma) p(\sigma^2|\gamma) p(\gamma)$$

$$p(\beta_\gamma|\sigma, \gamma) \sim N(0, g\sigma^2 (X_\gamma' X_\gamma)^{-1}), \quad p(\sigma^2|\gamma) \propto 1/\sigma^2$$

This prior makes $\log \sigma^2$ uniform.

$$p(\gamma_i = 1) = \pi_i$$

1.4 marginal likelihood

The likelihood is

$$p(Y|\beta_\gamma, \sigma^2, \gamma) \propto \exp(-\frac{\|Y - X\beta_\gamma\|^2}{2\sigma^2})$$

So

$$p(Y|\gamma) \propto p(Y, \gamma) = \int p(Y, \beta_\gamma, \sigma^2, \gamma) d\beta_\gamma d\sigma^2$$

$$= \int p(Y|\beta_\gamma, \sigma^2, \gamma) p(\beta_\gamma, \sigma^2, \gamma) d\beta_\gamma d\sigma^2$$

$$\propto (1+g)^{-q/2} S(\gamma)^{-n/2}$$

where $q = \sum_i \gamma_i$ and

$$S(\gamma) = Y'Y - \frac{g}{1+g} Y' P_{X_\gamma} Y$$

The posterior is

$$p(\gamma|Y) \propto (1+g)^{-q/2} S(\gamma)^{-n/2} \prod_i \pi_i^{\gamma_i} (1-\pi_i)^{1-\gamma_i}$$

- $g \to \infty$

  $$S(\gamma) \approx \|Y - X_\gamma \hat{\beta}_{\gamma,ols}\|^2$$

  $S(\gamma)$ encourages more complicated model. Yet, $(1+g)^{-q/2}$ encourages simpler model.

- $g \to 0$

  $$p(\gamma|Y) \propto prior(\gamma)$$

1.5 Gibbs sampler

- initial value for γ
- successively generate from $p(\gamma_i|Y, \gamma_{j \neq i})$

  $$p(\gamma_i|Y, \gamma_{j \neq i}) \propto p(Y|\gamma) p(\gamma_i)$$

  where $p(\gamma_i) = \pi_i^{\gamma_i} (1-\pi_i)^{1-\gamma_i}$

1.6 Laplace Approx

1.6.1 BIC

If equal priors are used, then max posterior is the same as max $p(D|\gamma)$, the model likelihood. Max it can be approximately solved using Laplace approximation.

In general,

$$p(D|\gamma) = \int p(D, \theta_\gamma|\gamma) d\theta_\gamma$$

We now expand $p(D, \theta_\gamma|\gamma)$. Write

$$p(D, \theta_\gamma|\gamma) = \exp(l_n(\gamma, \theta_\gamma))$$

We have

$$l_n(\gamma, \theta_\gamma) \approx l_n(\gamma, \hat{\theta}_\gamma) - \frac{1}{2}(\theta_\gamma - \hat{\theta}_\gamma)^T H (\theta_\gamma - \hat{\theta}_\gamma)$$

where

$$H_\gamma = -\nabla^2_\theta l_n(\gamma, \hat{\theta}_\gamma).$$

So

$$p(D|\gamma) \approx \exp(l_n(\gamma, \hat{\theta}_\gamma)) \int \exp(-\frac{1}{2}(\theta_\gamma - \hat{\theta}_\gamma)^T H (\theta_\gamma - \hat{\theta}_\gamma)) d\theta_\gamma$$

$$= \exp(l_n(\gamma, \hat{\theta}_\gamma))(2\pi)^{q/2}|H_\gamma|^{-1/2} = p(D, \hat{\theta}_\gamma|\gamma)(2\pi)^{q/2}|H_\gamma|^{-1/2}$$

$$= p(D|\hat{\theta}_\gamma, \gamma) p(\hat{\theta}_\gamma|\gamma)(2\pi)^{q/2}|H_\gamma|^{-1/2}$$

where $q = \dim(\theta_\gamma)$

So MLE for model selection is $\max_\gamma \log p(D|\gamma)$, which is approximately max

$$\log p(D|\hat{\theta}_\gamma, \gamma) + \log p(\hat{\theta}_\gamma|\gamma)(2\pi)^{q/2} - \frac{1}{2}\log|H_\gamma|$$

Note $p(\hat{\theta}_\gamma|\gamma)$ does not grow with n, $H_\gamma = n * h$ for some h so $|H_\gamma| = n^q|h|$. But $\log p(D|\hat{\theta}_\gamma, \gamma)$ grows with n, so dropping those do not grow with n,

$$\max \log p(D|\hat{\theta}_\gamma, \gamma) - \frac{q}{2}\log n$$

So this is BIC, which is approx posterior model. It requires $p(\theta_\gamma|\gamma)$ does not grow with rate n.

Alternatively, use bayesian,

$$p(\gamma|D) \propto p(\gamma)(\log p(D|\hat{\theta}_\gamma, \gamma) - \frac{q}{2}\log n)$$

1.7 Bayesian predictions

- We are interested in predicting y in regressions, given training data $D = (X, Y)$ and x_{new}

  Posterior

$$P(y|D, x_{new}) = \int p(y|D, \beta, x_{new})p(\beta|D, x_{new})d\beta$$

$$= \int p(y|\beta, x_{new})p(\beta|D)d\beta$$

  Suppose $p(y|\beta, x_{new}) = N(x_{new}\beta, \sigma^2)$, and $p(\beta|D) = N(\mu_n, \Sigma_n)$

  Then

$$P(y|D, x_{new}) = N(x_{new}\mu_n, \sigma_y)$$

  where

$$\sigma_y = \sigma^2 + x_{new}^T \Sigma_n x_{new}$$

  Assuming σ^2 known.

  this is homework

- With model averaging

$$P(y|D, x_{new}) = \int P(y|M, D, x_{new})P(M|D)dM$$

$$\approx \frac{1}{B}\sum_i P(y|M_i, D, x_{new})$$

  where M_i follows from posterior

2 EM

Dempster, laird and Rubin (1977)
dealt with latent variables, or missing data

2.1 The idea

- Suppose we have observed data X. We also have latent variable $Latent$, or missing data.

  The complete data is $(X, Latent)$, whose joint distr depends on unknown parameter θ.

- If we knew $Latent$, then the full likelihood is

$$L_{complete}(Latent, \theta) = p(X, Latent|\theta)$$

  then estimate θ is just MLE.

- But we do not know W. The idea is to replace $L_{complete}$ by its expectation with respect to Latent, which is conditional posterior

$$p(Latent|X, \theta)$$

  Treating Latent is "parameter", this is Bayesian.

  Assume the conditional distribution known:

$$p(Latent|X, \theta)$$

- The EM:

  E-step: at $j+1$ step, compute

$$E^j \log L_{complete}(Latent, \theta)$$

  as a function of θ. The expectation is wrt $p(Latent|X, \theta^j)$

  M-step: estimate θ:

$$\theta^{j+1} = \arg\max E^j \log L_{complete}(Latent, \theta)$$

- The EM algorithm can be sensitive to the choice of the initial point

2.2 The math behind EM

- Ideally we want to max the marginal likelihood

$$p(X|\theta)$$

We now look at the E step in detail.

$$E^j \log L_{complete}(Latent, \theta) = \log p(X|\theta) + B(\theta, \theta^j)$$

- Proof:

$$E^j \log L_{complete}(Latent, \theta) = \int p(latent|X, \theta^j) \log p(X, Latent|\theta) dLatent$$

$$= \int p(latent|X, \theta^j) \log(p(Latent|X, \theta) p(X|\theta)) dLatent$$

$$= \underbrace{\int p(latent|X, \theta^j) \log p(Latent|X, \theta) dLatent}_{B} + \log p(X|\theta) \int p(latent|X, \theta^j) dLatent$$

$$:= B(\theta, \theta^j) + \log p(X|\theta)$$

- The M step: use $B(\theta^j)$, so

$$\theta^{j+1} = \max_{\theta} E^j \log L_{complete}(Latent, \theta) = \max B(\theta, \theta^j) + \log p(X|\theta)$$

$$\theta^{j+1} \approx \max \log p(X|\theta) + B(\theta^j, \theta^j).$$

- $\log p(X|\theta^j)$ is monotonically increasing:

$$\log p(X|\theta^{j+1}) = E^j(\theta^{j+1}) - B(\theta^{j+1}; \theta^j) \geq E^j(\theta^j) - B(\theta^{j+1}; \theta^j)$$

$$= \log p(X|\theta^j) + \underbrace{B(\theta^j, \theta^j) - B(\theta^{j+1}; \theta^j)}_{\geq 0}$$

in fact

$$B(\theta^j, \theta^j) - B(\theta^{j+1}; \theta^j) = KL(a||b) \geq 0$$

where

$$a = p(latent|X, \theta^j), \quad b = p(latent|X, \theta^{j+1})$$

2.3 Example in linear regression

2.3.1 Missing one parameter

$$Y = X_1\beta + X_2 b + e, \quad e \sim N(0,1)$$

where we treat b as "latent variable", or "latent data". We only interested in β

- full likelihood

  assume prior independence :

  $$p(D, b|\beta) = p(D|b,\beta)p(b|\beta) = p(D|b,\beta)p(b)$$

  as a function of β.

- posterior $p(b|D,\beta^j)$: this is the conditional posterior of b

  The expected log likeihood

  $$El(\beta) = E^*(\log p(D|b,\beta) + \log p(b)) = E^* \log p(D|b,\beta) + c$$

  where E^* is wrt $p(b|D,\beta^j)$.

- For Gaussian model

  suppose $\sigma = 1$ and g-prior with prior zero mean,

  $$P(\beta, b|D) \sim N(cOLS, c(X^TX)^{-1})$$

  where $c = \frac{g}{g+1}$.

  then

  $$p(b|D,\beta) \sim N(\mu, \Sigma)$$

  which is conditional normal from multivariate normal.

- The algorithm:

  E step:

  $$El(\beta) \approx \frac{1}{B}\sum_b \log p(D|, b_b, \beta)$$

  where $b_b \sim p(b|D,\beta^j)$, which is marginal posterior

M step:

$$\beta^{j+1} = \arg\min_{\beta} \sum_{b} \|Y - X_1\beta - X_2 b_b\|^2$$

$$(X_1^T X_1)^{-1} X_1^T (Y - X_2\bar{b}) = \beta$$

where $\bar{b}$ is posterior mean from $p(b|D, \beta^j)$.

2.3.2 Missing data

$$Y = X\beta + e$$

Suppose some Y are missing, so

$$Y = (Y_o, Y_m)$$

- full likelihood

$$p(Y|X, \beta) \sim normal$$

- posterior $p(Y_m|Y_o, X, \beta^j) = p(Y_m|X, \beta^j)$ because data are iid conditionally on X.

  To see this:

$$p(Y_m|Y_o, X, \beta^j) = \frac{p(Y_m, Y_o|X, \beta)}{p(Y_o|X, \beta)} = \frac{\frac{1}{(2\pi\sigma^2)^{m+o}} \exp(-\frac{1}{2\sigma^2} \sum_i (y_i - X_i\beta)^2)}{\frac{1}{(2\pi\sigma^2)^{o}} \exp(-\frac{1}{2\sigma^2} \sum_{i\in o} (y_i - X_i\beta)^2)}$$

$$= \frac{1}{(2\pi\sigma^2)^m} \exp(-\frac{1}{2\sigma^2} \sum_{i\in m} (y_i - X_i\beta)^2) = p(Y_m|X, \beta^j) = N(X\beta^j, \sigma^2)$$

- E step:

$$E \log p(Y_o, Y_m|X, \beta) = \int p(Y_m|X, \beta^j) \log p(Y_o, Y_m|X, \beta) dY_m$$

$$= -\int p(Y_m|X, \beta^j) \sum_i (y_i - x_i\beta)^2 dY_m$$

$$= -\int p(Y_m|X, \beta^j) \sum_{i\in m} (y_i - x_i\beta)^2 dY_m - \sum_{i\in o} (y_i - x_i\beta)^2 \int p(Y_m|X, \beta^j) dY_m$$

$$= -\sum_{i\in m}\int p(y_i|x_i,\beta^j)(y_i - x_i\beta)^2 dy_i - \sum_{i\in o}(y_i - x_i\beta)^2$$

$$= -\sum_{i\in m} E[(Y_i - x_i\beta)^2|X_i,\beta^j] - \sum_{i\in o}(y_i - x_i\beta)^2$$

where $Y_i \sim p(y_i|x_i,\beta^j) = N(x_i\beta^j,\sigma^2)$

Note that the first term is

$$\sum_{i\in m} E[(Y_i - x_i\beta)^2|X_i,\beta^j] = \sum_{i\in m} E[(Y_i - x_i\beta^j + x_i\beta^j - x_i\beta)^2|X_i,\beta^j]$$

$$= m\sigma^2 + \sum_{i\in m}(x_i\beta^j - x_i\beta)^2$$

So

$$E\log p(Y_o,Y_m|X,\beta) \propto -\sum_{i\in o}(y_i - x_i\beta)^2 - \sum_{i\in m}(x_i\beta^j - x_i\beta)^2$$

- The algorithm:

  M step:

$$\beta^{j+1} = \arg\min_\beta \sum_{i\in o}(y_i - x_i\beta)^2 - \sum_{i\in m}(x_i\beta^j - x_i\beta)^2$$

  solution: let $W = (X_o^TX_o + X_m^TX_m)^{-1}X_m^TX_m$ and $\hat\beta_o$ = OLS using (X_o,Y_o) only

$$\beta^{j+1} = (I - W)\hat\beta_o + W\beta^j$$

$$= (I - W)\sum_{k=0}^{j} W^k\hat\beta_o + W^{j+1}\beta^0$$

  When $X_m^TX_m = O(X_o^TX_o)$, i.e., $\|W\| < 1$, not overly many missing, then $\beta^{j+1} \to \hat\beta_o$, using missing data only

3 Gaussian mixture models

3.1 Mixing densities

- To model an unknown pdf $p(x)$, use

$$\sum_k p(x|k)w_k$$

  where w_k is weighting, and $p(x|k)$ is a collection of known pdf. To ensure this is pdf, weighting are negative and sum to one.

- each $p(x|k)$ is known up to a parameter, so write

$$\sum_k p(x|k,\theta_k)w_k$$

  We assume

$$p(x|k,\theta_k) \sim N(\mu_k, \sigma_k^2)$$

  The goal is to estimate θ_k, w_k

- "missing data problem":

  Each observation, X1, ... Xn, is drawn from one of these K distributions, but we are not told from which one.

  w_k provides the probability that a sample has been drawn from $p(x|k)$

- We assume

$$z_{ik} = 1\{x_i \text{ is processed by model k}\}$$

  For each i, $z_{i,k} = 1$ only for a single value of k and zero for the rest.

  Now we still observe x_i for i=1...n.

  For each X, we do not know which model it is from , nor do we know these model θ_ks.

  Here

$$w_k = P(z_{ik} = 1)$$

- E step

  (i) Full likelihood

$$p(X, Z|\beta, W) = \prod_i \prod_k [w_k N(x_i; \mu_k, \sigma_k^2)]^{z_{ik}}$$

$$\log p(X, Z|\theta, W) = \sum_{ik} z_{ik} [\log w_k + \log N(x_i; \mu_k, \sigma_k^2)]$$

  (ii) Now $p(z_{ik}|X, \beta^j, W^j)$ is , assuming independence over i,

$$p(z_{ik}|X, \beta, W) = p(z_{ik}|x_i, \beta, W) = \frac{p(z_{ik}, x_i|\theta, W)}{p(x_i|\theta, W)} = \frac{p(x_i|z_{ik}, \theta, W)p(z_{ik}|\theta, W)}{p(x_i|\theta, W)}$$

$$= \frac{N(x_i; \mu_k, \sigma_k^2)p(z_{ik}|\theta, W)}{\sum_d N(x_i; \mu_d, \sigma_d^2)w_d}$$

  So

$$E^j(z_{ik}) = P(z_{ik} = 1|X, \beta^j, W^j) = \frac{N(x_i; \mu_k^j, \sigma_k^j)w_k^j}{\sum_{d \le K} N(y_i; x_i^T \theta_d^j, \sigma^j)w_d^j}$$

  (iii) E-step

$$E^j(\theta, W) := \sum_{ik} E^j(z_{ik}) [\log w_k + \log N(x_i; \mu_k, \sigma_k^2)]$$

- M step:

  max $E^j(\theta, W)$ wrt θ, σ, W

  (i) First solve W^{j+1} :

$$\max \sum_{ik} E^j(z_{ik}) \log w_k : \quad w_k \ge 0, \sum w_k = 1$$

$$w_k^{j+1} = \frac{1}{n} \sum_i E^j(z_{ik})$$

  (ii) Solve for μ, σ

For each k,

$$\theta_k^{j+1} = \arg\max \sum_i E^j(z_{ik})[\log N(x_i; \mu_k, \sigma_k^2)]$$

which is weighted MLE, each observation is weighted by $E^j(z_{ik})$

Taking derivative, get

$$\mu_k^{j+1} = \frac{\sum_i E^j(z_{ik}) x_i}{\sum_i E^j(z_{ik})}$$

(iii) solve for σ_k^{j+1}

Still weighted MLE

$$\sigma^{j+1} = \frac{\sum_i E^j(z_{ik})(x_i - \mu_k^{j+1})^2}{\sum_i E^j(z_{ik})}$$

3.2 Mixing regressions

- Consider for k=1... K

$$y_k = \theta_k^T x + e$$

with common x. also $var(e_k) = \sigma^2$ is common

Then under normal,

$$y_k \sim N(x^T \theta_k, \sigma^2)$$

- We assume

$$z_{ik} = 1\{x_i \text{ is processed by model k}\}$$

This means: $z_{i,k} = 1$ if

$$y_i = \theta_k x_i + e_i$$

For each i, $z_{i,k} = 1$ only for a single value of k and zero for the rest.

Now we still observe (x_i, y_i) for i=1...n.

For output Y, we do not know which model it is from , nor do we know these model θ_ks.

- Consider a mixture model

$$p(y|\theta, \sigma, W) = \sum_k w_k N(y; x^T \theta_k, \sigma^2)$$

We treat Z as latent variable, and apply EM.

Here

$$w_k = P(z_{ik} = 1|x_i)$$

- E step

  (i) Full likelihood

$$p(Y, Z|\beta, W) = \prod_i \prod_k [w_k N(y_i; x_i^T \theta_k, \sigma^2)]^{z_{ik}}$$

$$\log p(Y, Z|\beta, W) = \sum_{ik} z_{ik} [\log w_k + \log N(y_i; x_i^T \theta_k, \sigma^2)]$$

  (ii) Now $p(z_{ik}|Y, X, \beta^j, W^j)$ is

$$p(z_{ik}|Y, X, \beta, W) = p(z_{ik}|y_i, x_i, \beta, W) = \frac{p(z_{ik}, y_i|x_i, \theta, W)}{p(y_i|x_i, \theta, W)} = \frac{p(y_i|x_i, z_{ik}, \theta, W) p(z_{ik}|x_i, \theta, W)}{p(y_i|x_i, \theta, W)}$$

$$= \frac{N(y_i; x_i^T \theta_k, \sigma^2) p(z_{ik}|x_i, \theta, W)}{\sum_d N(y_i; x_i^T \theta_d, \sigma^2) w_d}$$

  So

$$E^j(z_{ik}) = P(z_{ik} = 1|Y, X, \beta^j, W^j) = \frac{N(y_i; x_i^T \theta_k^j, \sigma^j) w_k^j}{\sum_{d \le K} N(y_i; x_i^T \theta_d^j, \sigma^j) w_d^j}$$

  So

  (iii) E-step

$$E^j(\theta, \sigma, W) = \sum_{ik} E^j(z_{ik}) [\log w_k + \log N(y_i; x_i^T \theta_k, \sigma^2)]$$

- M step:

  max $E^j(\theta, \sigma, W)$ wrt θ, σ, W

  (i) First solve W^{j+1} :

$$\max \sum_{ik} E^j(z_{ik}) \log w_k : \quad w_k \ge 0, \sum w_k = 1$$

$$w_k^{j+1} = \frac{1}{n} \sum_i E^j(z_{ik})$$

(ii) Solve for θ, σ

For each k,

$$\theta_k^{j+1} = \arg\max \sum_i E^j(z_{ik})[\log N(y_i; x_i^T \theta_k, \sigma^2)]$$

which is weighted MLE, each observation is weighted by $E^j(z_{ik})$

Taking derivative, get

$$\theta_k^{j+1} = (X^T \Gamma_k X)^{-1} X^T \Gamma_k Y$$

HW: find Γ_k

(iii) solve for σ^{j+1}

Still weighted MLE

$$\sigma^{j+1} = \frac{1}{n} \sum_{ik} E^j(z_{ik})(y_i - \theta_k^{j+1} x_i)^2$$

3.3 Mixture of experts

- We have K learners, $f_1, ..., f_K$, each called "expert".
  For example, $f_k = \theta_k^T x$. so each expert is a linear regression model
- Each expert is associated with a gating parameter, or weights:

  $$w_1 ... w_K$$

  The "mixing of learners" are

  $$\sum_k w_k f_k$$

  The mixture regression is an example

  $$f_k(x) = N(x^T \theta_k, \sigma^2)$$

- In the general case, the gatings are also functions of the input variables. Then the "mixture of experts " is

$$\sum_k g_k(x) f_k(x)$$

where we replace the new notation g for w, as "gates"

- Here we parametrize $g_k(x)$, by for example,

4 Gaussian process: modeling the regression function

$$Y = f(X) + e, \quad e \sim N(0, \sigma^2)$$

4.1 Reproducing Kernel Hilbert Space

- kernel function

$$K(x, x;') = cov(f(x), f(x'))$$

(1) Gaussian kernel

$$\exp(-\frac{\|x - x'\|^2}{2c})$$

(2) quadratic kernel

$$(1 + \|x - x'\|^2)^{-a}, a \geq 0$$

(3) Ornstein - Uhlenbeck Kernel

(4) Linear kernel

$$x * x'$$

which is not stationary (as function of $x - x'$)

- PDS K: for any $x_1 ... x_N$

$$(K(x_i, x_j))_{N \times N}$$

is symmetric and semi-positive definite

- The kernel helps to shape a Hilbert space

  Proof

  Step 1:

  Define a space

$$H = \{\sum_{i=1}^{M} a_i K(x_i, .) : a \in \mathbb{R}^M, M \in \mathbb{R}, x \in \mathcal{X}^M\}$$

  Step 2: Inner product on H_0:

  For any two functions in H_0:

$$f(.) = \sum_{i=1}^{M} a_i K(x_i, .), \quad g(.) = \sum_{j=1}^{J} b_j K(y_i, .)$$

  f is determined by the choice (a, x); g is determined by the choice (b, y).

  Define

$$< f, g >= \sum_{i=1}^{M} \sum_{j=1}^{J} a_i b_j K(x_i, y_j)$$

  Then the norm

$$\|f\|_K^2 =< f, f >\geq 0$$

  because K is PDS.

  It is also easy to check the inner product is linear

  Step 3: completeness: H_0 is not complete yet, but we can build a space $H_0 \subset H$, so that H_0 is dense in it. So H is complete and is Hilbert.

  any Cauchy sequence converges to an element in H

  Step 4: reproducing property:

  for any

$$f(.) = \sum_{i=1}^{M} a_i K(x_i, .)$$

any $y \in \mathcal{X}$

$$f(y) = \sum_{i=1}^{M} a_i K(x_i, y).$$

Now consider $g(.) = K(y, .)$ by definition,

$$< f, g >= \sum_{i=1}^{M} a_i K(x_i, y)$$

So we have proved

$$f(y) =< f, K(y, .) >$$

- Mercer's theorem: if $K(.)$ is semipositive definite, as a kernel of a linear operator T

$$[Tg](w) := \int K(w, s)g(s)ds$$

Then T has eigen-decomposition λ_j, ϕ_j, so that

$$K(w, s) = \sum_{j=1}^{\infty} \lambda_j \phi_j(w)\phi_j(s)$$

Then truncate at J

$$K(w, s) = \phi(w)^T diag(\lambda)\phi(s) + r(w, s)$$

- Thus the covariance kernel matrix

$$K(x_i, x_j) = \phi_i^T \Lambda_J \phi_j + r_{ij}$$

the matrix

$$K = \Phi \Lambda_J \Phi^T + R \approx \bar{\Phi}\bar{\Phi}^T$$

where $R \approx 0$.

So the covariance kernel matrix is approximately low rank

4.2 Gaussian process

- A random process,
$$\{g(Z,t) : t \in T\}$$
indexed by t, is called a Gaussian process (GP) if for any finite number of points, t1... tM, the joint distribution
$$g(Z,t_1)...g(Z,t_M)$$
is Gaussian

- covariance function
$$K(x,x;') = cov(g(Z,x), g(Z,x'))$$

- It is **stationary** if $Eg(Z,x)$ does not depend on x, and covariance function is a function of $x - x'$.

4.3 Predictive density

- Suppose in the regression,
$$y = f(X) + e, \quad e \sim N(0,\sigma^2)$$
f is a zero mean gaussian process conditionally on $X1..Xn$, in the following sense:
$$\{f(X) : X \in \mathcal{X}\}$$
is GP, indexed by X.
$$g(Z,t) := f(X)$$
where $Z = f$, which is random; $t = X$, which is fixed index. So the randomness comes from f, while X is held as fixed index. So this is a Bayesian definition.

- Then conditionally on X
$$F = f(X_1)...f(X_n)$$
is zero mean, covariance function $K(X_i, X_j)$.

- For predictions, suppose we have a new x_{new}

  The goal is to find

$$p(y_{new}|Y, X, x_{new})$$

- Method (1)

  Let $\bar{F} = (f(xnew), F)'$

  Let $\bar{Y} = (y_{new}, Y)$. Supress the conditions on X, x_{new}, the goal is to find

$$p_|(\bar{Y}) := p(\bar{Y}|x, x_{new})$$

  where $p_|$ means conditional on X, Xnew.

  We know

$$p_|(\bar{Y}|\bar{F}) \sim N(\bar{F}, \sigma^2 I)$$
$$p_|(\bar{F}) \sim N(0, K(n+1))$$

  So $p_l(\bar{Y})$ is also normal.

$$E_|(\bar{Y}) = E_|E_|(\bar{Y}|\bar{F}) = E_|\bar{F} = 0$$
$$Var_|(\bar{Y}) = E_|Var_|(\bar{Y}|\bar{F}) + Var_|E_|(\bar{Y}|\bar{F}) = \sigma^2 I + Var_|(\bar{F}) = \sigma^2 I + K(n+1)$$

  Hence $y_{new}|Y, X, Xnew$ is also Gaussian, because the joint is Gaussian. Let

$$K(n+1) = \begin{pmatrix} k(new) & \Sigma_{12} \\ \Sigma_{21} & K(n) \end{pmatrix}$$

  The mean is

$$\Sigma_{12}(\sigma^2 + K(n))^{-1} Y$$

  The variance is

$$\sigma^2 + M$$
$$M := k(new) - \Sigma_{12}(\sigma^2 + K(n))^{-1}\Sigma_{21}$$

- Method (2)

$$p(y_{new}|Y, X, x_{new}) = \int p(y_{new}|\bar{F}, Y, X, x_{new}) p(\bar{F}|Y, X, x_{new}) d\bar{F}$$

(a) $p(\bar{F}|Y,X,x_{new}) \propto p(\bar{F}|X,X_{new})p(Y|\bar{F},X,X_{new}) = N(\bar{F};0,K(n+1))N(Y;F,\sigma^2 I)$

$$\propto \exp(-\frac{1}{2}\bar{F}^T K(n+1)^{-1}\bar{F})\exp(-\frac{1}{2\sigma^2}\|Y-F\|^2)$$

(b)

$$p(y_{new}|\bar{F},Y,X,x_{new}) = N(y_{new};f(xnew),\sigma^2) \propto \exp(-\frac{1}{2\sigma^2}(y_{new}-f(xnew))^2)$$

(c) Together, note that $f(xnew) = e_1^T\bar{F}$ and $F = A\bar{F}$, where $A = [0,I]$

$$p(y_{new}|Y,X,x_{new}) \propto \int \exp(-\frac{1}{2\sigma^2}(y_{new}-f)^2 - \frac{1}{2}\bar{F}^T K(n+1)^{-1}\bar{F} - \frac{1}{2\sigma^2}\|Y-F\|^2)d\bar{F}$$

$$\propto \int \exp(-\frac{1}{2\sigma^2}(y_{new}-e_1^T\bar{F})^2 - \frac{1}{2}\bar{F}^T K(n+1)^{-1}\bar{F} - \frac{1}{2\sigma^2}\|Y-A\bar{F}\|^2)d\bar{F}$$

Calculations yield

$$p(y_{new}|Y,X,x_{new}) \sim N(mean,var)$$

for $e^T = (1,0,...0)$

$$G = I + \sigma^2 K_{n+1}^{-1}, \quad A = (0,I)$$

$$mean = \frac{e^T G^{-1} A^T Y}{1 - e^T G^{-1} e}$$

$$var = \frac{\sigma^2}{1 - e^T G^{-1} e}$$

- Matlab

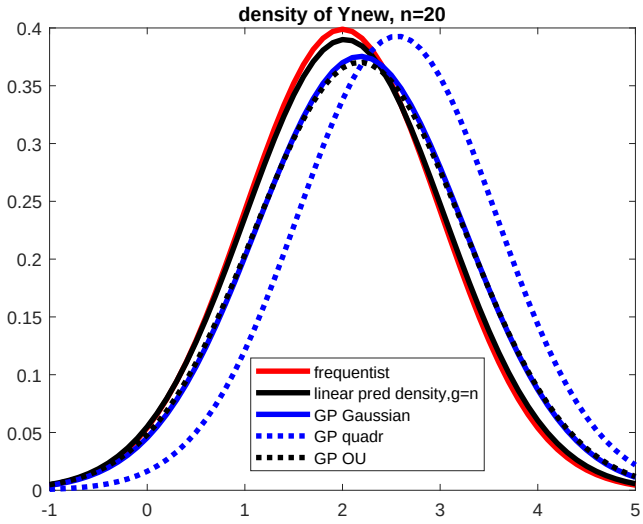
  Method (1) GP, with gaussian kernel $K(n) = \exp(-0.5|x-x'|^2)$.

4.4 Asymptotics of the predictive density

- To see $var \sim O(1/n)$, the Mercer's theorem shows

  $$K_{n+1} \approx \bar{\Phi}\bar{\Phi}^T$$

  which is low rank



Then first apply SVD to show

$$(I + \sigma^2 K^{-1})^{-1} = I - (K\sigma^{-2} + I)^{-1}$$

Hence

$$e^T G^{-1} e = 1 - e^T (K\sigma^{-2} + I)^{-1} e$$

Then apply Woodbury's identity to show (suppose $\sigma^2 = 1$)

$$e^T (K\sigma^{-2} + I)^{-1} e \approx e^T (\bar{\Phi}\bar{\Phi}^T + I)^{-1} e \approx 1 - \bar{\phi}_1^T (1 + \bar{\Phi}^T \bar{\Phi})^{-1} \bar{\phi}_1 = 1 + O(1/n)$$

given the assumption (x is fixed, not random)

$$\sum_i \|\phi(x_i)\|^2 = O(n)$$

So

$$e^T G^{-1} e = O(1/n)$$

Hence

$$var = \frac{\sigma^2}{1 - e^T G^{-1} e} \approx \sigma^2$$

Formal proof of this is HW

5 Objective priors

5.1 Jeffreys' prior

- Fisher's criticism of "uninformative prior".

  If no information on θ, then no information as well on $h(\theta)$.

  Example: $Exp(\lambda)$ or $Exp(1/\theta)$? where $\lambda = 1/\theta$.

  Say flat prior is on θ: $\theta \in Unif(0, A)$. Then

$$P(\lambda < t) = P(1/\theta < t) = P(\theta > t^{-1}) = \int_{1/t}^{A} \frac{1}{A} dx = \frac{1}{A}(A - 1/t)$$

  Thus

$$p_\lambda(t) = \frac{dP(\lambda < t)}{dt} = \frac{1}{A} t^{-2}, \quad t > 1/A$$

  This is an informative prior.

- More generally, Suppose we put flat prior on h. Then

$$p(\theta) = p(h(\theta)) |\frac{dh}{d\theta}| = |\frac{dh}{d\theta}|$$

  Suppose $h(\theta) = \theta^3$. Then

$$p(\theta) = 3\theta^2$$

  which is INFORMATIVE.

- Fisher's criticism leads to thoughts about what "non-informative prior" means

  So what does "informative" means? flat prior is not necessarily "informative" or "uninformative".

  We need to re-define "informative" or "uninformative"

  " finding prior distributions that have a minimal impact as possible on the data "

- "Uninformative" here means:

  invariance to reparametrization

If there is a method for us to find $p(\theta)$, then for any change of variable $h(\theta)$, and if we start from the beginning using the same method on h directly, we would get the same thing.

Say, this method gives me priors

$$p_\theta, \quad p_h$$

, say $\theta = \theta(h)$, then

$$p_h(h) = p_\theta(\theta(h))|\frac{d\theta}{dh}|$$

- Jeffrey prior

$$p(\theta) \propto \sqrt{|detI(\theta)|}$$

$$I(\theta) = -E(\nabla^2 \log L(\theta)) = E(\nabla \log L(\theta))^2$$

It can be proved that for any transformation $h(\theta)$

$$p(h) \propto \sqrt{detI(h)}$$

Proof. say $\theta = \theta(h)$

then

$$p_h(h) = p_\theta(\theta(h))|\frac{d\theta}{dh}|$$

now use Wikipedia, and other examples

Example: $N(\theta, \sigma^2)$. What is the Jeffreys' prior on θ ?

Simple calculation yields

$$I(\theta, \sigma) = \sigma^{-2} I_2$$

so the joint J prior is $1/\sigma^2$. This does not depend on μ.

5.2 Shannon's information theory

- Entropy X

$$H(X) = -E_X \log p(X) = -\int p(x) \log p(x) dx$$

which represents the chaos of the distr of X.

- Conditional entropy $X|Y$:

$$H(X|Y) = -\int p(x|y)\log p(x|y)dx$$

- Joint entropy decomposition:

$$H(X,Y) = EH(X|Y) + H(Y)$$

$EH(X|Y)$: after knowing Y, the remaining chaos of X.

So the above says:

$$\text{total chaos} = \text{chaos of Y} + \text{remaining chaos of X after knowing Y}$$

$$X \cup Y = Y \cup [X \backslash (X \cap Y)]$$

Proof

$$H(X,Y) = -\iint p(x,y)\log p(x,y)dxdy = -\int p(x,y)\log p(x|y)dxdy$$

$$-\int \log p(y)\int p(x,y)dxdy$$

$$= -\int p(y)\int p(x|y)\log p(x|y)dxdy - \int p(y)\log p(y)dy$$

$$= \int p(y)H(X|y)dy + H(Y) = EH(X|Y) + H(Y)$$

- The above equality implies:

$$EH(X|Y) = \text{remaining chaos of X after knowing Y} = X \backslash (X \cap Y)$$

So

$$H(X) - EH(X|Y) = \text{reduced chaos of X, after knowing Y} = X \cap Y$$

$$= \text{the "Information" about X, carried by Y}$$

Note that $X \cap Y$ is symmetric, so this is also the same as "the information" about Y, carried by X. In other words,

$$H(X) - EH(X|Y) = \text{mutual information}$$

Proof

Note

$$EH(X|Y) = -\int p(x,y) \log p(x|y) dxdy$$

$$H(X) = -\int p(x,y) \log p(x) dxdy$$

$$X \cap Y = I(X,Y) := H(X) - EH(X|Y) = -\int p(x,y) \log p(x) dxdy + \int p(x,y) \log p(x|y) dxdy$$

$$= (1) \quad \int p(x,y) \log \frac{p(x|y)}{p(x)} dxdy$$

$$= (2) \quad \int p(x,y) \log \frac{p(x,y)}{p(y)p(x)} dxdy$$

(2) shows symmetry.

- look at (1) again.

$$I(X,Y) = \int p(y)p(x|y) \log \frac{p(x|y)}{p(x)} dxdy = E_Y K(p(x|y)||p(x))$$

recall

$$K(P||Q) = \int \log \frac{dP}{dQ} dP$$

Note KP divergence is not symmetric, but mutual information is.

So mutual infor measures how close $p(x|y)$ is to $p(x)$. The closer, the more independent (X,Y), the less information.

- If $X \perp Y$, then $I = 0$, $X \cap Y = 0$.

5.3 Reference prior

- Now let

$$Y = D, \quad X = \theta$$

$$I(\theta, D) = E_D K(posterior||prior)) = \text{information about } \theta \text{ carried by data}$$

If data has lots of information about θ, posterior and prior are then very far from each other.

- We want the data to play as much role as possible, and thus we need prior to be as far from the posterior as possible. This is "uninformative" means.

$$p(\theta) = \arg\max_{p(\theta)} \lim_{n} I(\theta, D)$$

This is known as Reference prior

- Now let us solve it asymptotically

$$I(\theta, Y) = H(\theta) - E_D H(\theta|D) = H(\theta) + E_D E_{\theta|D} \log p(\theta|D)$$
$$\approx H(\theta) + E_D E_{\theta|D} \log \phi(\theta; D)$$

where $\phi(\theta; D)$ is the pdf of $N(MLE, \frac{1}{n}S)$, S is inverse Fisher information.

- In the one dim case, This simplifies

$$\approx K(det(S)^{1/2}||p(\theta)) = \int p(\theta) \log \frac{\sqrt{det(S(\theta))}}{p(\theta)} d\theta$$

this is max at Jeffrey's prior.

6 Posterior large sample properties

6.1 concentration rate

Suppose

$$p(\beta|D) = \frac{L_n(\beta)\pi(\beta)}{\int L_n(\beta)\pi(\beta)d\beta}$$

We now show that

$$P(\beta \in B|D) \to^P 0.$$

where

$$B = \{\|\beta - \beta_0\| \geq r_n\}$$

Proof. Let

$$p(\beta \in B|D) = \frac{\int_B L_n(\beta)/L_n(\beta_0)\pi(\beta)d\beta}{\int L_n(\beta)/L_n(\beta_0)\pi(\beta)d\beta} = \frac{J_B}{J}$$

Let

$$J_B = \int_B L_n(\beta)/L_n(\beta_0)\pi(\beta)d\beta$$

$$J = \int L_n(\beta)/L_n(\beta_0)\pi(\beta)d\beta$$

6.1.1 lower bound of J

We now show J is not too small.

This part is hardest, we follow Shen and Wasserman (2001)

Let

$$K(\beta) = \frac{1}{n}E_{D_n}\log\frac{L_n(\beta_0)}{L_n(\beta)}$$

$$V(\beta) = \frac{1}{n}Var_{D_n}\log\frac{L_n(\beta_0)}{L_n(\beta)}$$

$$K_n(\beta) = \frac{1}{n}\log\frac{L_n(\beta_0)}{L_n(\beta)}$$

Kullback-leibler divergence between $p(Y|\beta_0)$ and $p(Y|\beta)$.

So

$$E_{D_n}K_n(\beta) = K(\beta), \quad Var_{D_n}K_n(\beta) = \frac{1}{n}V(\beta)$$

$$\frac{L_n(\beta)}{L_n(\beta_0)} = \exp(-\log\frac{L_n(\beta_0)}{L_n(\beta)}) = \exp(-nK_n(\beta))$$

$$J = \int \exp(-nK_n(\beta))\pi(\beta)d\beta$$

In the normal case,

$$\log L_n(\beta) = c - \frac{\|Y - X\beta\|^2}{2\sigma^2}$$

$$K_n(\beta) = -\frac{\|Y - X\beta_0\|^2}{2n\sigma^2} + \frac{\|Y - X\beta\|^2}{2n\sigma^2}$$

now for some $b_n \to 0$,let

$$W_n = \{\beta : \frac{K_n(\beta) - K(\beta)}{\sqrt{V(\beta)}} > \sqrt{b_n}\}$$

$$S_n = \{\beta : K(\beta) < b_n, \quad V(\beta) < b_n\}$$

set W_n^c and S_n should be "relatively large":

$$\sup_\beta P_{D_n}(\beta \in W_n) \leq \sup_\beta \frac{Var_{D_n}(K_n(\beta))}{V(\beta)b_n} = \frac{1}{nb_n}$$

$$J \geq \int_{W_n^c \cap S_n} \exp(-nK_n(\beta))\pi(\beta)d\beta$$

$$= \int_{W_n^c \cap S_n} \exp(-n(\frac{K_n(\beta) - K(\beta)}{\sqrt{V(\beta)}})\sqrt{V(\beta)}) \exp(-nK(\beta))\pi(\beta)d\beta$$

$$\geq \int_{W_n^c \cap S_n} \exp(-n\sqrt{b_n}\sqrt{V(\beta)}) \exp(-nK(\beta))\pi(\beta)d\beta$$

$$\geq \int_{W_n^c \cap S_n} \exp(-2nb_n)\pi(\beta)d\beta = \exp(-2nb_n)\pi(W_n^c \cap S_n)$$

$$= \exp(-2nb_n)[\pi(S_n) - \pi(S_n \cap W_n)]$$

To look at $E_{D_n}\pi(S_n \cap W_n)$:

$$E_{D_n}\pi(S_n \cap W_n) = \int \pi(S_n \cap W_n)L_n(\beta_0)dD_n = \int\int 1_{\beta \in S_n \cap W_n}\pi(\beta)d\beta L_n(\beta_0)dD_n$$

$$= \int 1_{\beta \in S_n} \int 1_{\beta \in W_n} L_n(\beta_0)dD_n\pi(\beta)d\beta$$

Note: S_n does NOT depend on data, only W_n does.

$$\int 1_{\beta \in W_n} L_n(\beta_0)dD_n = E_{D_n}1\{\beta \in W_n\} = P_{D_n}(\beta \in W_n)$$

So

$$E_{D_n}\pi(S_n \cap W_n) = \int 1_{\beta \in S_n} P_{D_n}(\beta \in W_n)\pi(\beta)d\beta \leq \frac{1}{nb_n}\pi(\beta \in S_n)$$

Hence for some a_n,

$$P_{D_n}(J < a_n\pi(S_n)) \le P_{D_n}(\exp(-2nb_n)[\pi(S_n) - \pi(S_n \cap W_n)] < a_n\pi(S_n))$$

$$= P_{D_n}(\pi(S_n \cap W_n) > \pi(S_n) - a_n\pi(S_n)\exp(2nb_n))$$

$$\le \frac{1}{nb_n}\frac{1}{(1 - a_n\exp(2nb_n))}$$

can choose a_n so that $1 - a_n\exp(2nb_n)$ is a constant, for example, $a_n = \frac{1}{2}\exp(-2nb_n)$ Then

$$P_{D_n}(J > \frac{1}{2}\exp(-2nb_n)\pi(S_n)) \ge 1 - \frac{2}{nb_n}$$

6.1.2 upper bound of J_B

Method 1

The key proof for posterior convergence is usually based on a "testing technique" Schwartz (1965). Let E_n be a random event, such that E_n^c holds with high prob, to be determined. The usual tradition of the notation is that E_n^c holds likely, instead of E_n. We hope

$$P_{D_n}(E_n) \le small, \quad \sup_{\beta \in B} P_{D_n,\beta}(E_n^c) \le small$$

where $P_{D_n,\beta}$ means the probability measure with respect to the data, when the true value is β.

$$p(\beta \in B|D) = \frac{J_B}{J}1_{E_n} + \frac{J_B}{J}1_{E_n^c} \le 1_{E_n} + \frac{J_B}{J}1_{E_n^c}$$

We need to bound both.

step 1

For 1_{E_n}:

$$E_{D_n}(1_{E_n}) = P_{D_n}(E_n)$$

Hence for some c_n,

$$P_{D_n}(1_{E_n} > c_n) \le P_{D_n}(E_n)\frac{1}{c_n}$$

For $\frac{J_B}{J}1_{E_n^c}$,

$$E_{D_n}(J_B 1_{E_n^c}) = E_{D_n}1_{E_n^c}\int_B \frac{L_n(\beta)}{L_n(\beta_0)}\pi(\beta)d\beta = \int 1_{E_n^c}\int_B \frac{L_n(\beta)}{L_n(\beta_0)}\pi(\beta)L_n(\beta_0)dD_n d\beta$$

$$= \int 1_{E_n^c} \int_B L_n(\beta)\pi(\beta) d\beta dD_n = \int_B \int 1_{E_n^c} L_n(\beta) dD_n \pi(\beta) d\beta = \int_B E_{D_n,\beta}(1_{E_n^c})\pi(\beta) d\beta$$
$$\leq \pi(\beta \in B) \sup_{\beta \in B} E_{D_n,\beta}(1_{E_n^c}) = \pi(\beta \in B) \sup_{\beta \in B} P_{D_n,\beta}(E_n^c)$$

where $E_{D_n,\beta}$ means the expectation with respect to the data, when the true value is β.

So for some d_n,

$$P_{D_n}(J_B 1_{E_n^c} > d_n) \leq \pi(\beta \in B) \sup_{\beta \in B} P_{D_n,\beta}(E_n^c) \frac{1}{d_n}$$

step 2
together

$$p(\beta \in B|D) \leq 1_{E_n} + \frac{J_B}{J} 1_{E_n^c}$$

$$P_{D_n}(J_B 1_{E_n^c} < d_n) \geq 1 - \pi(\beta \in B) \sup_{\beta \in B} P_{D_n,\beta}(E_n^c) \frac{1}{d_n}$$

$$P_{D_n}(1_{E_n} < c_n) \geq 1 - P_{D_n}(E_n) \frac{1}{c_n}$$

$$P_{D_n}(J > \frac{1}{2} \exp(-2nb_n)\pi(S_n)) \geq 1 - \frac{2}{nb_n}$$

Hence with probability $1 - m$, where $m = \frac{2}{nb_n} + \pi(\beta \in B) \sup_{\beta \in B} P_{D_n,\beta}(E_n^c) \frac{1}{d_n} + P_{D_n}(E_n) \frac{1}{c_n}$

$$P(\beta \in B|D) \leq c_n + \frac{d_n}{\frac{1}{2} \exp(-2nb_n)\pi(S_n)}$$

we need $m \to 0$ and c_n and d_n as small as possible, so take

$$nb_n \to \infty$$

$$P_{D_n}(E_n) = \frac{1}{3} c_n m$$

$$\pi(B) \sup_{\beta \in B} P_{D_n,\beta}(E_n^c) = \frac{1}{3} d_n m$$

Thus

$$P(\beta \in B|D) \leq \frac{3}{m} P_{D_n}(E_n) + \frac{3\pi(B)}{\frac{1}{2} m\pi(S_n)} \exp(2nb_n) \sup_{\beta \in B} P_{D_n,\beta}(E_n^c)$$

step 3 let us now find E_n. can set

$$E_n = \{\|\widehat{\beta} - \beta_0\| \geq k_n\}.$$

and note $\|\widehat{\beta} - \beta_0\| \geq \|\beta - \beta_0\| - \|\widehat{\beta} - \beta\|$

$$\sup_{\beta \in B} P_{D_n,\beta}(E_n^c) = \sup_{\|\beta - \beta_0\| > r_n} P_{D_n,\beta}(\|\widehat{\beta} - \beta_0\| < k_n) \leq \sup_{\|\beta - \beta_0\| > r_n} P_{D_n,\beta}(\|\widehat{\beta} - \beta_0\| < k_n)$$

$$\leq \sup_{\|\beta - \beta_0\| > r_n} P_{D_n,\beta}(\|\widehat{\beta} - \beta\| > r_n - k_n) = \sup_{\|\beta - \beta_0\| > r_n} P_{D_n,\beta}(\|\widehat{\beta} - \beta\| > r_n/2)$$

set $k_n = r_n/2$,

So need exponential bound. For normal model,

$$P_{D_n}(E_n) + P_{D_n,\beta}(\|\widehat{\beta} - \beta\| > r_n/2) \leq C \exp(-Cnr_n^2)$$

then

$$P(\beta \in B|D) \leq \frac{3}{m} C \exp(-Cnr_n^2) + \frac{3\pi(B)}{\frac{1}{2} m\pi(S_n)} C \exp(-Cnr_n^2 + 2nb_n)$$

let

$$2b_n = \frac{C}{2} r_n^2$$

$$P(\beta \in B|D) \leq \frac{3}{m} C \exp(-Cnr_n^2) + \frac{3\pi(B)}{\frac{1}{2} m\pi(S_n)} C \exp(-\frac{C}{2} nr_n^2)$$

let $nr_n^2 \to \infty$, suppose

$$\pi(S_n) \exp(\frac{C}{2} nr_n^2) \to \infty$$

$m^{-1} \to \infty$ slower than $\pi(S_n) \exp(\frac{C}{2} nr_n^2)$

Method 2 recall

$$p(\beta \in B|D) = \frac{J_B}{J}$$

$$\frac{L_n(\beta)}{L_n(\beta_0)} = \exp(-\log\frac{L_n(\beta_0)}{L_n(\beta)}) = \exp(-nK_n(\beta))$$

alternative bound for

$$J_B = \int_B L_n(\beta)/L_n(\beta_0)\pi(\beta)d\beta = \int_B \exp(-nK_n(\beta))\pi(\beta)d\beta$$

$$K_n(\beta) = \frac{1}{n}\log\frac{L_n(\beta_0)}{L_n(\beta)}$$

Suppose for $a_n \to 0$,

$$P_{D_n}(\inf_{\beta\in B} K_n(\beta) > a_n) \to 1,$$

Then on the event $\inf_{\beta\in B} K_n(\beta) > a_n$,

$$J_B \le \exp(-na_n)\pi(B)$$

Also,

$$P_{D_n}(J > \frac{1}{2}\exp(-2nb_n)\pi(S_n)) \ge 1 - \frac{2}{nb_n}$$

Hence

$$p(\beta \in B|D) \le \frac{\exp(2nb_n - na_n)\pi(B)}{\frac{1}{2}\pi(S_n)} = \frac{\exp(-\frac{1}{2}na_n)\pi(B)}{\frac{1}{2}\pi(S_n)}$$

let $2b_n = \frac{1}{2}a_n$, need

$$na_n \to \infty$$

$$\pi(S_n)\exp(\frac{1}{2}na_n) \to \infty$$

key question: what is a_n ? For normal model,

$$K_n(\beta) = -\frac{\|Y - X\beta_0\|^2}{2n\sigma^2} + \frac{\|Y - X\beta\|^2}{2n\sigma^2}$$

then $a_n \sim r_n$. More general verification of $P_{D_n}(\inf_{\beta\in B} K_n(\beta) > a_n) \to 1$, is the stochastic equicontinuity of empirical process (Ossiander 1987)

6.2 Bernstein von Mises theorem

- The BvM shows the asymptotic normality

Recall that the likelihood is

$$L_n(\beta) = \exp(l_n(\beta)).$$

- Consider two expansions

  (1) Here $l_n(\beta)$ has expansion

  $$l_n(\beta) = l_n(\beta_0) + (\beta - \beta_0)'\nabla l_n(\beta_0) - \frac{n}{2}(\beta - \beta_0)' J_n(\beta_0)(\beta - \beta_0) + R_n(\beta)$$

  So this looks like a normal likelihood. We now formalize this result.

  (2)

  $$0 = \nabla l_n(MLE) = \nabla l_n(\beta_0) - nJ_n(\beta_0)(MLE - \beta_0) + r_n$$

  So

  $$MLE \approx \beta_0 + \frac{1}{n} J_n(\beta_0)^{-1} \nabla l_n(\beta_0)$$

  This motivates:

  Given β, we make a transformation

  $$h = \sqrt{n}(\beta - \beta_0) - \frac{1}{\sqrt{n}} J_n(\beta_0)^{-1} \nabla l_n(\beta_0)$$

  Equivalently

  $$\beta = \frac{1}{\sqrt{n}} h + \underbrace{\beta_0 + \frac{1}{n} J_n(\beta_0)^{-1} \nabla l_n(\beta_0)}_{T_n}$$

  We do so , because when $\beta = \hat{\beta}_{MLE}$, $h \approx 0$ (in absence of R_n). So for a general β, h measures how far it is from $\hat{\beta}_{MLE}$.

- The B-v-M theorem: detailed proof can be found from Ghosh and Ramamoorthi (2003, Springer)

  Let $p_n(\beta) = p(\beta|D)$.

  we consider $p_n(\beta)$, but in the new transformation

  $$\frac{1}{\sqrt{n}} p_n\left(\frac{1}{\sqrt{n}} h + \beta_0 + \frac{1}{n} J_n(\beta_0)^{-1} \nabla l_n(\beta_0)\right) := p_n^*(h)$$

we aim to show

$$p_n^*(h) \approx N(0, J_n(\beta_0)^{-1}) \sim \phi(h)$$

Proof. Define

$$T_n = \beta_0 + \frac{1}{n} J_n(\beta_0)^{-1} \nabla l_n(\beta_0)$$

then

$$\beta = \frac{1}{\sqrt{n}} h + T_n$$

T_n looks like MLE.

Use these notation, replace β by $\frac{1}{\sqrt{n}} h + T_n$,

$$l_n(\beta) = l_n(\beta_0) + (\beta - \beta_0)' \nabla l_n(\beta_0) - \frac{n}{2} (\beta - \beta_0)' J_n(\beta_0)(\beta - \beta_0) + R_n(\beta)$$

$$= l_n(\beta_0) + (\frac{1}{\sqrt{n}} h + T_n - \beta_0)' \nabla l_n(\beta_0) - \frac{n}{2} (\frac{1}{\sqrt{n}} h + T_n - \beta_0)' J_n(\beta_0)(\frac{1}{\sqrt{n}} h + T_n - \beta_0) + R_n(\frac{1}{\sqrt{n}} h + T_n)$$

$$= C - \frac{1}{2} h' J_n(\beta_0) h + R_n(\frac{1}{\sqrt{n}} h + T_n)$$

Recall

$$p(\beta|D) = \frac{1}{C} \exp(l_n(\beta)) \pi(\beta)$$

then replace β with $\frac{1}{\sqrt{n}} h + T_n$

$$p_n^*(h) = \frac{\pi(\frac{h}{\sqrt{n}} + T_n) \exp(w(h))}{C_n}$$

$$w(h) = -\frac{1}{2} h' J_n(\beta_0) h + R_n(\frac{h}{\sqrt{n}} + T_n)$$

where C_n is the integration wrt h.

Now we are hoping

$$\frac{\pi(\frac{h}{\sqrt{n}} + T_n) \exp(w(h))}{C_n} \approx \phi(h) \sim N(0, J_n(\beta_0)^{-1})$$

So we aim to show

$$\exp(-\frac{1}{2} h' J_n(\beta_0) h + R_n(\frac{h}{\sqrt{n}} + T_n)) \pi(\frac{h}{\sqrt{n}} + T_n) \approx \exp(-\frac{1}{2} h' J_n(\beta_0) h) \pi(\beta_0)$$

(i) when $|h| < M$, this is true, note $T_n \approx \beta_0$

(ii) when $M < |h| < \sqrt{n}$, the LHS can be made smaller than $\exp(-\frac{1}{4}h' J_n(\beta_0)h)$, which then is small if M is sufficiently large.

(iii) when $|h| > \sqrt{n}$. Both sides go to zero.

we only give the intuition up to here.

6.3 confidence interval

6.3.1 MCMC revisits

Metroplis-Hastings

The algorithm:

1. choose a starting β^0
2. generate ξ from $q(\beta^j|\xi)$
3. update β^{j+1} using

$$\beta^{j+1} = \begin{cases} \xi & pro = \rho(\beta^j, \xi) \\ \beta^j & pro = 1 - \rho(\beta^j, \xi) \end{cases}$$

where

$$\rho(x, y) = \min(1, \frac{L_n(y)\pi(y)}{L_n(x)\pi(x)} \frac{q(x|y)}{q(y|x)})$$

we use

$$q(x|y) \sim e^{-\frac{|x-y|^2}{2}}$$

6.3.2 Large sample property of MCMC confidence interval

Let c_{a1} and c_{a2} be the upper and lower quantiles of MCMC:

$$P(c_{a2} < \beta < c_{a1}|D) = 0.95$$

We now show that

$$P(c_{a2} < \beta_0 < c_{a1}) \to 0.95$$

Proof. let

$$F(x|D) = \int^x p(\beta|D)d\beta$$

$$F(\beta_0 + \frac{s}{\sqrt{n}}|D) = \int^{\beta_0 + \frac{s}{\sqrt{n}}} p(\beta|D)d\beta$$

Let $U_n = \frac{1}{\sqrt{n}} J_n^{-1} \nabla l_n(\beta_0)$ (at trueth)
recall

$$\beta = \frac{1}{\sqrt{n}} h + \beta_0 + \frac{1}{\sqrt{n}} U_n$$

change var to h,

$$P(\beta < \beta_0 + \frac{s}{\sqrt{n}}|D) = F(\beta_0 + \frac{s}{\sqrt{n}}|D) = \int_{h+U_n<s} p_n^*(h)dh \approx \int_{h+U_n<s} \phi(h)dh$$

$$= \int_{x<s} \phi(x - U_n)d(x - U_n) = \int^s N(x; U_n, J_n^{-1})dx$$

$$= P(N(U_n, J_n^{-1}) < s|U_n)$$

let $Z = N(0, J_n^{-1})$,
now let $\beta_0 + \frac{s_1^*}{\sqrt{n}} = c_{a1}$, and $\beta_0 + \frac{s_2^*}{\sqrt{n}} = c_{a2}$,

$$0.95 = P(c_{a2} < \beta < c_{a1}|D) = P(\beta < c_{a1}|D) - P(\beta < c_{a2}|D)$$
$$\approx P(N(U_n, J_n^{-1}) < s_1^*|U_n) - P(N(U_n, J_n^{-1}) < s_2^*|U_n)$$
$$= P(Z < s_1^* - U_n|U_n) - P(Z < s_2^* - U_n|U_n)$$
$$= P(s_2^* - U_n < Z < s_1^* - U_n|U_n)$$

So let us define

$$m1 = s_1^* - U_n, \quad m2 = s_2^* - U_n$$

which satisfy

$$P(m2 < Z < m1) = 0.95$$

On the other hand,

$$P(c_{a2} < \beta_0 < c_{a1}) = P(s_1^* > 0, s_2^* < 0) = P(U_n < -m2, U_n > -m1)$$

$$= P(m2 < U_n < m1) = 0.95$$

suppose $U_n \sim N(0, J^{-1})$.
it remains to show $U_n \to^d N(0, J_n^{-1})$. Note $U_n = \frac{1}{\sqrt{n}} J_n^{-1} \nabla l_n$.

$$Var(U_n) = J_n^{-1} \frac{1}{n} Var(\nabla l_n) J_n^{-1}$$

this is true because $\frac{1}{n} Var(\nabla l_n) = J_n$

7 Term project: Estimate the density of SP500 returns

Obtain data from SP500 index, as well as its constituents. It does not matter whether the panel is balanced, because all we need for each individual stock is its mean and variance.

Estimate the density of SP500 using three methods. Plot the three in the same plot

- Use full MoN to estimate index density,

  you need to:

  (1) give the iteration scheme

  (2) plot the histogram with the density

  (3) try some number of mixtures

- Use MoN to estimate index density, but assuming σ_k, μ_k are known, only need to estimate weights

  you need to:

  (0) Estimate individual σ_k, μ_k from each return

  (1) give the iteration scheme for the weights

  (2) plot the histogram with the density

  (3) the number of mixtures is just the number of stocks used

- Estimate the SP500 predictive density, given the model

$$Y_t = f(X_t) + e_t, \quad X_t = Y_{t-1}, \quad e \sim N(0, \sigma^2)$$

  you need to:

  (1) σ^2 is assumed known, simply estimated by sample variance

  (2) Use "data" $Y_1, ..., Y_T$ in the following way:

  let

$$(Y_1, ..., Y_{T-1}) \text{be "X"}$$

$$(Y_2, ..., Y_T) \text{be "Y"}$$

$$Y_T \text{ be "Xnew"}$$

To goal is to use the above to get the posterior predict density for $Y_{new} = Y_{T+1}$

(3) Use Gaussian process