Bayesian Statistics
Bayesian Estimation

Shaobo Jin

Department of Mathematics

Maximum a Posteriori Estimator

In a Bayes model, the parameter θ is a random variable with known distribution π.

- Finding the true parameter makes no sense in a Bayes model.

Data that we observe are generated in a hierarchical manner:

$$\theta \sim \pi(\theta), \qquad X \mid \theta \sim f(x \mid \theta).$$

Given the data x, we can make inference for the "current" data generating process.

Definition

The maximum a posteriori (MAP) estimator is the mode of the posterior $\pi(\theta \mid x)$ as

$$\hat{\theta}_{\mathrm{MAP}}(x) = \arg\max_{\theta} \pi(\theta \mid x).$$

MAP: Example

Note that $\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$ and $m(x)$ does not include any θ.

- The MAP estimator only requires the kernel of $f(x \mid \theta)\pi(\theta)$.
- We can skip the integration step to obtain $m(x)$.

Example

Let X_1, ..., X_n be iid $N\left(0, \sigma^2\right)$. The parameter of interest is $\theta = \sigma^{-2}$. We assume that the prior of θ is $\text{Gamma}(a, b)$. Find the MAP estimator.

MAP and 0-1 Loss

For an estimator d, consider the loss function

$$L(\theta, d) = 1(\|\theta - d\| > \epsilon),$$

where $\|\theta - d\| = \sqrt{(\theta - d)^T (\theta - d)}$. Consider the expected value of $L(\theta, d)$ under the posterior distribution:

$$\begin{aligned}
\mathrm{E}[L(\theta, d) \mid x] &= \int 1(\|\theta - d\| > \epsilon)\, \pi(\theta \mid x)\, d\theta \\
&= 1 - \mathrm{P}(\|\theta - d\| \leq \epsilon \mid x).
\end{aligned}$$

To minimize the expected loss, we want $\mathrm{P}(\|\theta - d\| \leq \epsilon \mid x)$ to be as large as possible, that is, the distribution of $\theta \mid x$ is concentrated around d.

Nuisance Parameter

Suppose that data are generated from $f(x \mid \theta, \tau)$, where θ is the parameter of interest and τ is the nuisance parameter.

- The frequentist approach will find a sufficient statistic $T(x)$ for τ and make inference for θ using the conditional distribution of $X \mid T(X)$.
- Alternatively, inference is based on the profile likelihood

  $$L(\theta) = \max_{\tau} f(x \mid \theta, \tau) = f(x \mid \theta, \hat{\tau}(\theta)),$$

  where $\hat{\tau}(\theta)$ maximizes $f(x \mid \theta, \tau)$ for fixed θ.

An advantage of the Bayes approach is that we can simply integrate out the nuisance parameter τ and make inference from the marginal posterior $\pi(\theta \mid x)$, instead of the joint posterior $\pi(\theta, \tau \mid x)$.

Neyman-Scott Problem: Example

Consider the Neyman-Scott problem, where $X_{ij} \mid \theta \sim N\left(\mu_{ij}, \sigma^2\right)$, $i = 1, ..., n$ and $j = 1, 2$. We are interested in σ^2, and μ_{ij}'s are nuisance parameters.

- The MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(x_{i1} - x_{i2}\right)^2}{4n} \xrightarrow{P} \frac{\sigma^2}{2} \neq \sigma^2.$$

- Consider the reference prior $\pi\left(\theta\right) \propto \sigma^{-1}$. The MAP of $\pi\left(\sigma^2 \mid x\right)$ is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(x_{i1} - x_{i2}\right)^2}{2n + 4} \xrightarrow{P} \sigma^2.$$

A Cautious Note

Suppose that the posterior is $\pi(\theta, \tau \mid x)$, where θ is the parameter of interest. The mode of $\pi(\theta, \tau \mid x)$ may not equal the marginal posterior mode, the mode of $\pi(\theta \mid x)$.

Example

Consider the normal-inverse-gamma model, where the posterior is

$$\mu \mid \sigma^2, x \sim N\left(\mu_n, \quad \frac{\sigma^2}{\lambda_0 + n}\right) \qquad \sigma^2 \mid x \sim \text{InvGamma}\left(a_n,\ b_n\right),$$

where μ_n, a_n, and b_n are known constants. Find the joint and marginal MAPs.

One More Issue: Existence

Example

Suppose that we observe iid $X_i \mid \theta \sim N(\theta, 1)$. The prior of θ is a mixture normal

$$\pi(\theta) = pN\left(\mu, \sigma^2\right) + (1-p)\, N\left(-\mu, \sigma^2\right).$$

Find the mode of $\pi(\theta \mid x)$.

Regularized Estimator

The MAP estimator essentially maximizes $f(x \mid \theta)\pi(\theta)$ or

$$\log f(x \mid \theta) + \log \pi(\theta),$$

provided that the logarithms are well defined. Intuitively speaking, we maximize the log-likelihood $\log f(x \mid \theta)$, but the penalty term $\log \pi(\theta)$ cannot be too big.

Suppose that data are generated from $X_i \mid \theta = \theta_0 \sim f(x \mid \theta_0)$, $i = 1, ..., n$.

- If $n^{-1} \log \pi(\theta) \to 0$ as $n \to \infty$, we should expect the MAP and the MLE to share similar properties.

One Difference Between MLE and MAP

Theorem (Invariance of MLE)

Let $\hat{\theta}_{ML}$ be the MLE of θ. Then, $g\left(\hat{\theta}_{ML}\right)$ is the MLE of $g(\theta)$ for any $g(\cdot)$.

However, MAP is not invariant with respect to reparametrization.

Example

Suppose that we observe one observation from $X \mid \theta \sim \text{Binomial}(n, \theta)$. Let the prior be $\theta \sim \text{Beta}(a_0, b_0)$, where $a_0 > 1$ and $b_0 > 1$.

1. Find the MAP estimator of θ.
2. Find the MAP estimator of $\eta = \theta / (1 - \theta)$.

Posterior Mean

An alternative to MAP is the posterior mean

$$\hat{\theta}_{\text{Mean}} = \mathrm{E}\left[\theta \mid x\right].$$

Example

Suppose that we observe one observation from $X \mid \theta \sim \text{Binomial}\,(n, \theta)$. Let the prior be $\theta \sim \text{Beta}\,(a_0, b_0)$, where $a_0 > 1$ and $b_0 > 1$.

- The posterior is $\text{Beta}\,(a_0 + x, b_0 + n - x)$.
- Hence, $\hat{\theta}_{\text{Mean}} = \frac{a_0+x}{a_0+b_0+n}$.

Posterior Mean and L_2 Loss

Consider the weighted L_2 loss

$$L_W(\theta, d) = (\theta - d)^T W (\theta - d),$$

where W is a $p \times p$ positive definite matrix and d is an estimator of θ using x.

Theorem

Suppose that there exists an estimator d such that

$$E[L_W(\theta, d) \mid x] = \int L_W(\theta, d) \pi(\theta \mid x) \, d\theta < \infty,$$

Then, the posterior mean minimizes $E[L_W(\theta, d) \mid x]$, where W does not depend on θ.

Posterior Mean versus MAP

Suppose that the posterior is $\pi(\theta, \tau \mid x)$, where θ is the parameter of interest.

- The mode of $\pi(\theta, \tau \mid x)$ may not equal the marginal posterior mode, the mode of $\pi(\theta \mid x)$.
- But the marginal posterior mean is the same as the joint posterior mean.

Example

Consider the normal-inverse-gamma model, where the posterior is

$$\mu \mid \sigma^2, x \sim N\left(\mu_n, \quad \frac{\sigma^2}{\lambda_0 + n}\right) \qquad \sigma^2 \mid x \sim \mathrm{InvGamma}\left(a_n,\ b_n\right),$$

where μ_n, a_n, and b_n are known constants. The posterior mean is the mean of $\mathrm{InvGamma}(a_n,\ b_n)$.

Posterior Mean versus MAP

To obtain the closed form expression of $\mathrm{E}[\theta \mid x]$, we need the normalizing constant of $\pi(\theta \mid x)$.

- The MAP estimator only requires the kernel $f(x \mid \theta)\pi(\theta)$. We can skip the integration step to obtain $m(x)$.
- Even we know $m(x)$, the integral to get $\mathrm{E}[\theta \mid x]$ may not be tractable.

But if we can sample from $\pi(\theta \mid x)$, we don't need to compute $m(x)$ nor evaluate the integral for $\mathrm{E}[\theta \mid x]$.

- Suppose that we have a sample θ_1, ..., θ_m from $\pi(\theta \mid x)$, then we can approximate the posterior mean by

$$\frac{1}{m}\sum_{j=1}^{m}\theta_j.$$

Posterior Mean versus MAP

It can even happen that the posterior mean does not exist, even though the posterior is proper.

Example

Let X_1, ..., X_n be iid from a two parameter Weibull distribution

$$f(x \mid \theta, \beta) = \frac{\beta x^{\beta-1}}{\theta^{\beta}} \exp\left\{-\left(\frac{x}{\theta}\right)^{\beta}\right\}, \quad x > 0,\ \theta > 0,\ \beta > 0.$$

Consider the proper priors

$$\begin{aligned} \pi(\theta \mid \beta) &= \frac{\beta b_0^{a_0}}{\Gamma(a_0)} \frac{1}{\theta^{a_0\beta+1}} \exp\left(-\frac{b_0}{\theta^{\beta}}\right), \text{ "InvGamma" prior} \\ \pi(\beta) &= \frac{d_0^{c_0}}{\Gamma(c_0)} \beta^{c_0-1} \exp(-d_0\beta). \text{ Gamma prior} \end{aligned}$$

With probability 1, the posterior mean of θ^k does not exist for any $k > 0$.

Posterior Mean versus MAP

It can happen that the likelihood involves intractable integrals. Hence, the MAP is not easy to obtain but we can sample easily from the posterior.

Example

Suppose that $Y_{ij} \mid Z_i, \beta, \lambda \sim \text{Bernoulli}\left(p_{ij}\right)$, $i = 1, ..., n$, $j = 1, ..., k$, where

$$p_{ij} = \frac{\exp\left(\beta_j + \lambda z_i\right)}{1 + \exp\left(\beta_j + \lambda z_i\right)}.$$

But we only observe $\{Y_{ij}\}$.

Invariance of Posterior Mean

The posterior mean is not invariant with respect to reparametrization either.

Example

Suppose that we observe one observation from $X \mid \theta \sim \text{Binomial}\left(n, \theta\right)$. Let the prior be $\theta \sim \text{Beta}\left(a_0, b_0\right)$, where $a_0 > 1$ and $b_0 > 1$.

1. Find the posterior mean of θ.
2. Find the posterior mean of $\eta = \theta / \left(1 - \theta\right)$.

Predict New Value

In frequentist statistics, the prediction of a new observation z after observing x is

$$\hat{z}(x) = \int z f\left(z \mid x, \hat{\theta}\right) dz.$$

In Bayesian statistics, the predictive distribution of a new observation z after observing x is

$$f(z \mid x) = \int f(z \mid x, \theta) \pi(\theta \mid x) d\theta.$$

A predictor can be the predictive mean

$$\hat{z}(x) = \int z f(z \mid x) dz,$$

or the predictive mode $\max_z f(z \mid x)$.

Derive Predictive Distribution

Example

Consider an iid sample $(X_1, ..., X_n)$ from Poisson (θ). The prior of θ is Gamma (a_0, b_0) with density

$$\pi(\theta) = \frac{b_0^{a_0}}{\Gamma(a_0)} \theta^{a_0 - 1} \exp(-b_0 \theta).$$

1. Find the posterior $\pi(\theta \mid x)$.
2. Let z be a future value. Find the predictive distribution $f(z \mid x)$.
3. Propose a predictor of z.

Multiple Linear Regression

A multiple linear regression is

$$Y_i = x_i^T \beta + \epsilon_i, \quad i = 1, ..., n,$$

where Y_i is the response, x_i is the vector of covariates (or regressors, or features), and β is the vector of unknown regression parameter.

In matrix notation, the model is

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}.$$

Some examples are:

1. Y is apartment price, $\boldsymbol{Z}$ includes crime rate, number of rooms, size of the apartment, year of construction, etc.
2. Y is waste water flow rate, $\boldsymbol{Z}$ includes temperature, precipitation, date of the year, time, etc.

Normal Linear Model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

The usual assumptions are

1. $\mathrm{E}\left[\epsilon \mid X\right] = 0$,
2. $\mathrm{Var}\left(\epsilon \mid X\right) = \Sigma$, where $\Sigma > 0$.

A typical assumption is $\Sigma = \sigma^2 I_n$, where I_n is an $n \times n$ identity matrix.

The ordinary least squares (OLS) estimator of β minimizes $\left(y - X\beta\right)^T \left(y - X\beta\right)$, and the minimizer is

$$\hat{\beta}_{\mathrm{OLS}} = \left(X^T X\right)^{-1} X^T y.$$

Normal Linear Model

In the normal linear model, we further assume that ϵ is normal: $\epsilon \mid X \sim N_n(0, \Sigma)$. Hence,

$$Y \mid X, \beta, \Sigma \quad \sim \quad N(X\beta,\ \Sigma).$$

The likelihood function is

$$f(y \mid X, \beta, \Sigma) \quad = \quad \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left\{-\frac{1}{2}(y - X\beta)^T \Sigma^{-1} (y - X\beta)\right\}$$

If $\Sigma = \sigma^2 I$, the the MLE of β coincides with the OLS estimator:

$$\hat{\beta}_{\text{ML}} \quad = \quad \left(X^T X\right)^{-1} X^T y.$$

For notation simplicity, we will treat X as fixed and drop it from conditioning.

Bayesian Linear Model: Known Σ

Suppose that Σ is completely known, i.e., β is the only unknown parameter.

Result

The conjugate prior for β is $N_p\left(\mu_0,\, \Lambda_0^{-1}\right)$. The posterior is $\beta \mid y \sim N\left(\mu_n,\, \Lambda_n^{-1}\right)$, where

$$\begin{aligned} \Lambda_n &= \Lambda_0 + X^T \Sigma^{-1} X, \\ \mu_n &= \Lambda_n^{-1}\left(\Lambda_0 \mu_0 + X^T \Sigma^{-1} y\right). \end{aligned}$$

Suppose that we observe a new x_0 and want to predict the new y_0. If $y_0 \mid \beta \sim N\left(x_0^T \beta,\, \sigma^2\right)$, where σ^2 is known, then the predictive distribution is

$$y_0 \mid y \sim N\left(x_0^T \mu_n,\, \sigma^2 + x_0^T \Lambda_n^{-1} x_0\right).$$

Ridge Regression

Suppose that $Y \mid \beta \sim N_n\left(X\beta, \sigma^2 I_n\right)$, where σ^2 is known. Let $\mu_0 = 0$ and $\Lambda_0 = \frac{\lambda}{\sigma^2} I_n$, that is

$$\beta \;\sim\; N_p\left(0,\, \frac{\sigma^2}{\lambda} I_n\right).$$

The posterior is

$$\beta \mid y \;\sim\; N_p\left(\left(X^T X + \lambda I_n\right)^{-1} X^T y,\, \frac{X^T X + \sigma^2 I_n}{\sigma^2}\right).$$

The posterior mean and MAP give the ridge regression estimator that minimizes

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta} \left(y - X\beta\right)^T \left(y - X\beta\right) + \lambda \beta^T \beta \\ &= \arg\max_{\beta} -\frac{1}{2\sigma^2}\left(y - X\beta\right)^T \left(y - X\beta\right) - \frac{1}{2\sigma^2/\lambda} \beta^T \beta. \end{aligned}$$

Laplace Prior

Consider the independent Laplace prior

$$\beta_i \overset{iid}{\sim} \text{Laplace}\left(0, \frac{\sigma^2}{\lambda}\right).$$

The posterior satisfies

$$\pi\left(\beta \mid y\right) \propto \exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}\left(y - X\beta\right)^T\left(y - X\beta\right) + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right]\right\}.$$

The MAP gives the lasso regression estimator that minimizes

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta}\frac{1}{2}\left(y - X\beta\right)^T\left(y - X\beta\right) + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \\ &= \arg\max_{\beta} -\frac{1}{\sigma^2}\left[\frac{1}{2}\left(y - X\beta\right)^T\left(y - X\beta\right) + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right].\end{aligned}$$

Bayesian Linear Model: Unknown σ^2

Suppose that $\Sigma = \sigma^2 I_n$, but σ^2 is unknown. The parameter is $\theta = \left(\beta, \sigma^2\right)$.

- The likelihood is

$$\begin{aligned} f\left(y \mid \beta, \sigma^2\right) &= \frac{\exp\left\{-\frac{1}{2}\left(y - X\beta\right)^T\left(\sigma^2 I_n\right)^{-1}\left(y - X\beta\right)\right\}}{\left(2\pi\right)^{n/2}\sqrt{\det\left(\sigma^2 I_n\right)}} \\ &\propto \frac{1}{\left(\sigma^2\right)^{n/2}}\exp\left\{-\frac{\beta^T X^T X\beta - 2y^T X\beta}{2\sigma^2}\right\}. \end{aligned}$$

- The conjugate prior is

$$\begin{aligned} \beta \mid \sigma^2 &\sim N_p\left(\mu_0, \sigma^2\Lambda_0^{-1}\right), \\ \sigma^2 &\sim \text{InvGamma}\left(a_0, b_0\right), \end{aligned}$$

  a normal-inverse-gamma distribution.

Normal-Inverse-Gamma Distribution

Definition

A random vector $X \in \mathbb{R}^p$ and a positive random scalar $\lambda > 0$ follow a normal-inverse-gamma (NIG) distribution if

$$X \mid \lambda \sim N_p(\mu,\ \lambda\Sigma),\ \text{and } \lambda \sim \text{InvGamma}(a, b).$$

It is denoted by $(X, \lambda) \sim \text{NIG}(a, b, \mu, \Sigma)$. The joint density is

$$f(x, \lambda) = c \exp\left\{-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2\lambda} - \frac{b}{\lambda}\right\} \frac{1}{\lambda^{a+p/2+1}},$$

where the constant c is given by

$$c = \frac{b^a}{(2\pi)^{p/2} \Gamma(a) \sqrt{\det(\Sigma)}}.$$

Marginal Distribution

A random vector $X \in \mathbb{R}^p$ follows a multivariate t-distribution $t_v(\mu, \Sigma)$, if its density is

$$f(x) = \frac{\Gamma\left(\frac{v+p}{2}\right)}{\Gamma\left(\frac{v}{2}\right) v^{p/2} \pi^{p/2} \sqrt{\det(\Sigma)}} \left[1 + \frac{1}{v}(x-\mu)^T \Sigma^{-1} (x-\mu)\right]^{-(v+p)/2},$$

where v is the degrees of freedom, $\mu = \mathrm{E}[X]$ for $v > 1$, and $\mathrm{var}(X) = \frac{v}{v-2}\Sigma$ for $v > 2$.

Result

For the NIG distribution $(X, \lambda) \sim \mathrm{NIG}(a, b, \mu, \Sigma)$, the marginal distributions are

$$\begin{aligned} \lambda &\sim \mathrm{InvGamma}(a,\ b), \\ X &\sim t_{2a}\left(\mu,\ \frac{b}{a}\Sigma\right). \end{aligned}$$

Posterior Distribution

Result

Under the conjugate prior, the posterior distribution is

$$\begin{aligned} \beta \mid y, \sigma^2 &\sim N\left(\mu_n,\ \sigma^2 \Lambda_n^{-1}\right), \\ \sigma^2 \mid y &\sim \text{InvGamma}\left(a_n,\ b_n\right), \end{aligned}$$

where

$$\begin{aligned} \Lambda_n &= X^T X + \Lambda_0, \\ \mu_n &= \Lambda_n^{-1}\left(\Lambda_0 \mu_0 + X^T y\right), \\ a_n &= \frac{n}{2} + a_0, \\ b_n &= b_0 + \frac{1}{2}\left(y^T y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n\right). \end{aligned}$$

That is, $\left(\beta, \sigma^2\right) \mid y \sim \text{NIG}\left(a_n, b_n, \mu_n, \Lambda_n^{-1}\right)$.

Marginal Posterior of Normal Linear Model

Under the conjugate prior, the posterior distribution is

$$\begin{aligned} \beta \mid y, \sigma^2 &\sim N\left(\mu_n,\ \sigma^2 \Lambda_n^{-1}\right), \\ \sigma^2 \mid y &\sim \text{InvGamma}\left(a_n,\ b_n\right), \end{aligned}$$

that is $\left(\beta, \sigma^2\right) \mid y \sim \text{NIG}\left(a_n, b_n, \mu_n, \Lambda_n^{-1}\right)$. Then,

$$\begin{aligned} \beta \mid y &\sim t_{2a_n}\left(\mu_n,\ \frac{b_n}{a_n}\Lambda_n^{-1}\right), \\ \sigma^2 \mid y &\sim \text{InvGamma}\left(a_n, b_n\right). \end{aligned}$$

Predictive Distribution

Result

Suppose that we observe a new x_0 and want to predict the new y_0. Assume that $y_0 \perp y \mid \beta, \sigma^2$. Under the conjugate prior, the predictive distribution is

$$y_0 \mid y \sim t_{2a_n}\left(x_0^T \mu_n, \frac{b_n}{a_n}\left(1 + x_0^T \Lambda_n^{-1} x_0\right)\right),$$

same expectation as σ^2 were known.

Ridge Regression Again

Suppose that $Y \mid \beta \sim N_n\left(X\beta, \sigma^2 I_n\right)$, where σ^2 is unknown. Let $\mu_0 = 0$ and $\Lambda_0 = \lambda I_n$, that is

$$\beta \mid \sigma^2 \sim N_p\left(\mu_0, \frac{\sigma^2}{\lambda} I_p\right), \qquad \sigma^2 \sim \text{InvGamma}\left(a_0, b_0\right).$$

The posterior satisfies

$$\beta \mid y, \sigma^2 \sim N\left(\mu_n,\ \sigma^2 \Lambda_n^{-1}\right), \qquad \beta \mid y \sim t_{2a_n}\left(\mu_n,\ \frac{b_n}{a_n}\Lambda_n^{-1}\right),$$

where

$$\mu_n = \left(X^T X + \lambda I_n\right)^{-1} X y$$

coincides with the ridge regression estimator.

Laplace Prior Again

Consider the independent Laplace prior

$$\beta_j \mid \sigma \overset{iid}{\sim} \text{Laplace}\left(0, \frac{\sigma^2}{\lambda}\right).$$

The posterior satisfies

$$\pi\left(\beta, \sigma^2 \mid y\right) \propto \frac{\pi\left(\sigma^2\right)}{\left(\sigma^2\right)^{p+n/2}} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}\left(y - X\beta\right)^T\left(y - X\beta\right) + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right]\right\}.$$

The MAP gives the lasso regression estimator.

Tuning Parameter

The tuning parameter λ is often selected using cross validation in ridge/lasso regression.

In Bayesian linear model, we can also treat λ as an unknown variable and use a prior for λ. A hierarchical setup can be

$$
\begin{aligned}
y \mid \beta, \sigma^2 &\sim N\left(X\beta, \sigma^2 I_n\right) \\
\beta \mid \sigma^2, \lambda &\sim N_p\left(0, \frac{\sigma^2}{\lambda} I_p\right) \\
\sigma^2 &\sim \mathrm{InvGamma}\left(a_0, b_0\right), \\
\lambda &\sim \mathrm{InvGamma}\left(c_0, d_0\right).
\end{aligned}
$$

That is, the prior is $\pi\left(\beta, \sigma^2, \lambda\right) = \pi\left(\beta \mid \sigma^2, \lambda\right) \pi\left(\sigma^2\right) \pi\left(\lambda\right)$.

Maximum Likelihood Estimator

The maximum likelihood estimator (MLE) of the model

$$Y = X\beta + e, \quad e \mid X \sim N_n\left(0, \sigma^2 I_n\right)$$

is given by

$$\hat{\beta}_{\mathrm{ML}} = \left(X^T X\right)^{-1} X^T y.$$

Its sampling distribution is

$$\hat{\beta}_{\mathrm{ML}} \mid \sigma^2 \sim N_p\left(\beta,\ \sigma^2 \left(X^T X\right)^{-1}\right).$$

The Zellner's g-prior is given by $\beta \mid \sigma^2 \sim N_p\left(\mu_0,\ g\sigma^2 \left(X^T X\right)^{-1}\right)$, where the constant $g > 0$.

Posterior Distribution

Result

Under the g-prior $\beta \mid \sigma^2 \sim N_p\left(\mu_0,\ g\sigma^2\left(X^TX\right)^{-1}\right)$ and $\sigma^2 \sim \mathrm{InvGamma}\left(a_0, b_0\right)$, the posterior distribution is

$$\begin{aligned}
\beta \mid y, \sigma^2 &\sim N\left(\mu_n,\ \frac{g}{g+1}\sigma^2\left(X^TX\right)^{-1}\right), \\
\sigma^2 \mid y &\sim \mathrm{InvGamma}\left(\frac{n}{2}+a_0,\ b_n\right),
\end{aligned}$$

where

$$\begin{aligned}
\mu_n &= \frac{1}{g+1}\mu_0 + \frac{g}{g+1}\left(X^TX\right)^{-1}X^Ty, \\
b_n &= b_0 + \frac{1}{2}\left(y^Ty - \frac{g}{g+1}y^TX\left(X^TX\right)^{-1}X^Ty\right) \\
&\quad + \frac{1}{2}\left(\frac{1}{g+1}\mu_0^TX^TX\mu_0 - \frac{2}{g+1}y^TX\mu_0\right).
\end{aligned}$$

Detour: Gradient and Hessian of Linear Form

Consider the function

$$f(x) = a_1 x_1 + a_2 x_2,$$

where $x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$ is a column vector. The gradient is

$$\frac{\partial f(x)}{\partial x} = \begin{bmatrix} \partial f / \partial x_1 \\ \partial f / \partial x_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}.$$

The Hessian matrix is

$$\frac{\partial^2 f(x)}{\partial x \partial x^T} = \begin{bmatrix} \partial^2 f / \partial x_1^2 & \partial^2 f / \partial x_1 \partial x_2 \\ \partial^2 f / \partial x_2 \partial x_1 & \partial^2 f / \partial x_2^2 \end{bmatrix} = 0_{2 \times 2}.$$

Detour: Gradient and Hessian of Quadratic Form

Consider

$$\begin{aligned} f(x) &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= a_{11}x_1^2 + a_{12}x_1x_2 + a_{21}x_2x_1 + a_{22}x_2^2. \end{aligned}$$

The gradient is

$$\frac{\partial f(x)}{\partial x} = \begin{bmatrix} \partial f / \partial x_1 \\ \partial f / \partial x_2 \end{bmatrix} = \begin{bmatrix} 2a_{11}x_1 + a_{12}x_2 + a_{21}x_2 \\ a_{12}x_1 + a_{21}x_1 + 2a_{22}x_2 \end{bmatrix}.$$

The Hessian matrix is

$$\frac{\partial^2 f(x)}{\partial x \partial x^T} = \begin{bmatrix} 2a_{11} & a_{12} + a_{21} \\ a_{12} + a_{21} & 2a_{22} \end{bmatrix}.$$

General Results for Linear and Quadratic Form

If $f(x) = a^T x$ with a and x being $p \times 1$ column vectors, then

$$\begin{aligned} \frac{\partial f(x)}{\partial x} &= a, \\ \frac{\partial^2 f(x)}{\partial x \partial x^T} &= 0_{p \times p}. \end{aligned}$$

If $f(x) = x^T A x$ with x being a $p \times 1$ column vector, then

$$\begin{aligned} \frac{\partial f(x)}{\partial x} &= \left(A + A^T\right) x, \\ \frac{\partial^2 f(x)}{\partial x \partial x^T} &= A + A^T. \end{aligned}$$

Jacobian Matrix

Suppose that the output of $f(x)$ is a $m \times 1$ vector, where the input x is a $p \times 1$ vector. The Jacobian matrix of f is defined to be

$$\frac{\partial f(x)}{\partial x^T} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x^T} \\ \frac{\partial f_2(x)}{\partial x^T} \\ \vdots \\ \frac{\partial f_m(x)}{\partial x^T} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} & \cdots & \frac{\partial f_1(x)}{\partial x_p} \\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} & \cdots & \frac{\partial f_2(x)}{\partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \frac{\partial f_m(x)}{\partial x_2} & \cdots & \frac{\partial f_m(x)}{\partial x_p} \end{bmatrix}_{m \times p}.$$

Example: Compute Jacobian Matrix

Example

Find the Jacobian matrix of $f(x) = \begin{bmatrix} a_1 & a_2 \\ b_1 & -b_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

- Note that

$$\frac{\partial f_1(x)}{\partial x} = \frac{\partial a_1 x_1 + a_2 x_2}{\partial x} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \qquad \frac{\partial f_2(x)}{\partial x} = \frac{\partial b_1 x_1 - b_2 x_2}{\partial x} = \begin{bmatrix} b_1 \\ -b_2 \end{bmatrix}.$$

- Hence, the Jacobian matrix is

$$\frac{\partial f(x)}{\partial x^T} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x^T} \\ \frac{\partial f_2(x)}{\partial x^T} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ b_1 & -b_2 \end{bmatrix}.$$

In general, we have

$$\frac{\partial Ax}{\partial x^T} = A.$$

Jeffreys Prior

Result

Consider the linear model $Y \mid \beta, \sigma^2 \sim N_n\left(X\beta, \sigma^2 I_n\right)$. The Fisher information of the above model is

$$\mathcal{I}\left(\beta, \sigma^2\right) = \begin{bmatrix} \frac{1}{\sigma^2} X^T X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}.$$

Hence, the Jeffreys prior is

$$\pi\left(\beta, \sigma^2\right) \propto \frac{1}{(\sigma^2)^{p/2+1}},$$

and the independent Jeffreys prior is

$$\pi\left(\beta, \sigma^2\right) \propto \frac{1}{\sigma^2}.$$

Independent Jeffreys Prior

The independent Jeffreys prior for the linear model $Y \mid \beta, \sigma^2 \sim N_n\left(X\beta, \sigma^2 I_n\right)$ is

$$\pi\left(\beta, \sigma^2\right) \quad \propto \quad \frac{1}{\sigma^2}.$$

Consider the change of variables $\beta = \beta$ and $\tau = \log \sigma^2$. Then,

$$\pi\left(\beta, \tau\right) \quad \propto \quad \frac{1}{\sigma^2}\left|\det\left(\frac{\partial \begin{bmatrix} \beta \\ \sigma^2 \end{bmatrix}}{\partial \begin{bmatrix} \beta & \tau \end{bmatrix}}\right)\right| = 1.$$

Hence, the independent Jeffreys prior means that the prior is uniform on $\left(\beta, \log \sigma^2\right)$.

Posterior with Jeffreys Prior

Theorem

Consider the linear regression model $Y \mid \beta, \sigma^2 \sim N_n\left(X\beta, \sigma^2 I_n\right)$. *Let the prior be*

$$\pi\left(\beta, \sigma^2\right) \quad \propto \quad \left(\sigma^2\right)^{-m}.$$

The posterior is

$$\begin{aligned} \beta \mid \sigma^2, y &\sim N_p\left(\mu_n,\, \sigma^2\left(X^T X\right)^{-1}\right), \\ \sigma^2 \mid y &\sim InvGamma\left(\frac{n-p}{2} + m - 1,\, \frac{1}{2} y^T\left(I_n - H\right) y\right), \end{aligned}$$

where $\mu_n = \left(X^T X\right)^{-1} X^T y$ *and* $H = X\left(X^T X\right)^{-1} X^T$ *is the hat matrix.*

MLE versus Posterior

The previous theorem shows that

$$\beta - \mu_n \mid \sigma^2, y \quad \sim \quad N_p\left(0,\ \sigma^2 \left(X^T X\right)^{-1}\right).$$

If maximum likelihood is used to estimate, then the MLE is $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$ and

$$\hat{\beta} - \beta \mid \beta, \sigma^2 \quad \sim \quad N_p\left(0,\ \sigma^2 \left(X^T X\right)^{-1}\right).$$

Posterior Predictive Checks

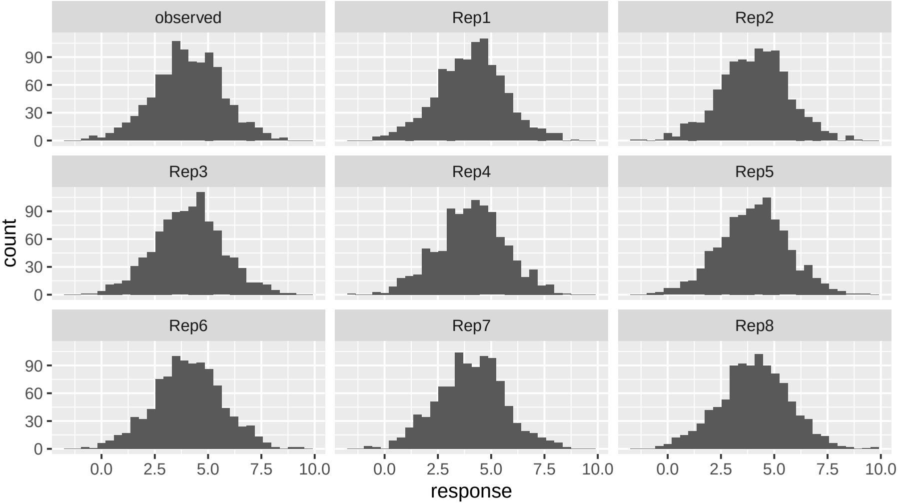
Posterior predictive check is a way to investigate whether our model can capture some relevant aspects of the data.

- We simulate data x_{sim} from the posterior predictive distribution

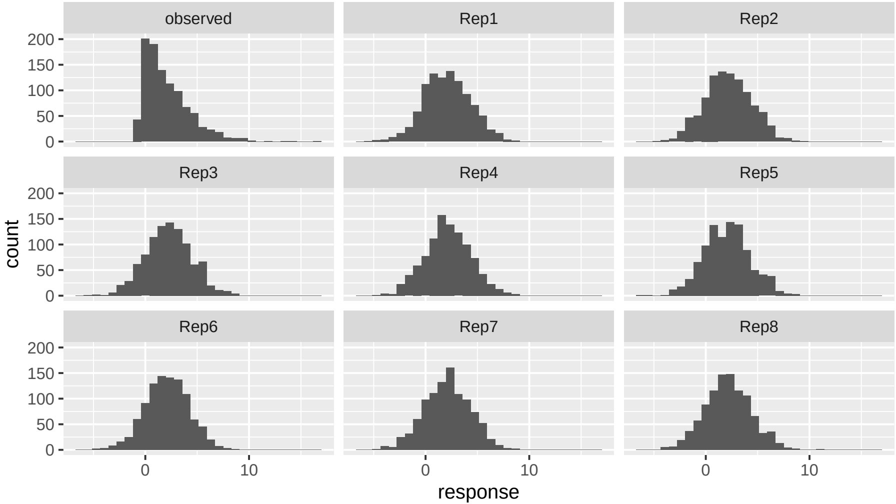
$$f\left(x_{\text{new}} \mid x\right) = \int f\left(x_{\text{new}} \mid x, \theta\right) \pi\left(\theta \mid x\right) d\theta.$$

- We can compare what our model predicts with the observed data, or compare statistics applied to the simulated data with the same statistics applied to the observed data.

Model 1


observed
Rep1
Rep2
Rep3
Rep4
Rep5
Rep6
Rep7
Rep8
count
90
60
30
0
0.0
2.5
5.0
7.5
10.0
response

Model 2


observed
Rep1
Rep2
Rep3
Rep4
Rep5
Rep6
Rep7
Rep8
count
200
150
100
50
0
0
10
response

Gaussian Process

Definition

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Let f be a scalar-valued function. We denote a Gaussian process by

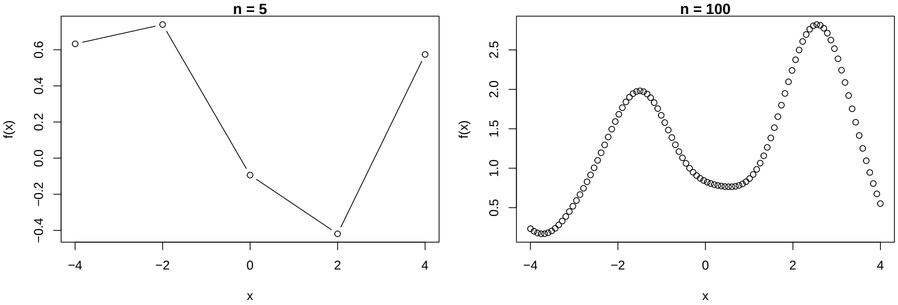
$$f(x) \sim \mathrm{GP}\left(m(x), k\left(x, x'\right)\right),$$

where $x \in \mathbb{R}^p$,

$$\begin{aligned} m(x) &= \mathrm{E}\left[f(x)\right], \text{ mean function} \\ k\left(x, x'\right) &= \operatorname{cov}\left(f(x), f\left(x'\right)\right). \text{ covariance function} \end{aligned}$$

Gaussian Process As Smooth Function

By the definition, the joint distribution of any finite $\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$ is multivariate normal. For a large enough n, the multivariate normal vector seems to produce a smooth function in x.



Gaussian Process Regression

Consider the Gaussian process regression model

$$y = f(x) + \epsilon,$$

where $\epsilon \mid \sigma^2 \sim N\left(0, \sigma^2\right)$ and $f(x) \sim \mathrm{GP}\left(0, k\left(x, x'\right)\right)$. If we have observed n observations from this model, then

$$Y \mid \sigma^2 \sim N\left(0,\ K(X, X) + \sigma^2 I_n\right),$$

where

$$K(X, X) = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{bmatrix}.$$

Recap: Conditional Distribution

Result: Conditional Distribution of Multivariate Gaussian Distribution

Suppose that $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim N_p\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$ such that $\Sigma_{22} > 0$. Then,

$$Y_1 \mid Y_2 = y_2 \quad \sim \quad N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(y_2 - \mu_2\right),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

Predicted Value

Suppose that we want to predict the response value based on a new X_*. Then,

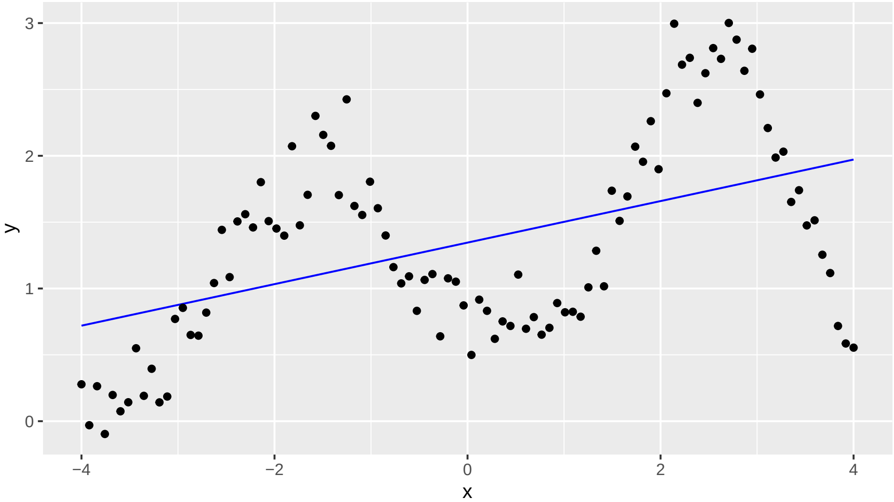
$$\begin{bmatrix} f(X_*) \\ Y \end{bmatrix} \mid \sigma^2 \quad \sim \quad N\left(0, \begin{bmatrix} K(X_*, X_*) & K(X_*, X) \\ K(X, X^*) & K(X, X) + \sigma^2 I_n \end{bmatrix}\right).$$

Hence, $f(X_*) \mid Y, \sigma^2$ is also Gaussian with

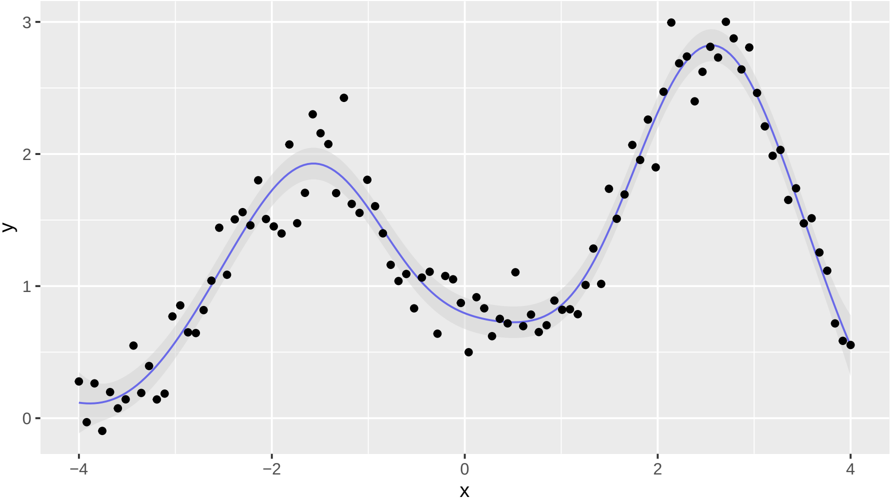
$$\begin{aligned} \text{mean} \quad & K(X_*, X)\left[K(X, X) + \sigma^2 I_n\right]^{-1} y, \\ \text{covariance} \quad & K(X_*, X_*) - K(X_*, X)\left[K(X, X) + \sigma^2 I_n\right]^{-1} K(X, X^*). \end{aligned}$$

The fitted function is then $K(X_*, X)\left[K(X, X) + \sigma^2 I_n\right]^{-1} y$.

Fitted Function Curve: $k(x, z) = x^T \Lambda_0^{-1} z$



Fitted Function Curve: $k(x,z) = \exp\left\{-\|x-z\|_2^2/2\right\}$



Bayesian Linear Model

Consider the linear regression model

$$y = f(x) + \epsilon, \qquad f(x) = x^T \beta,$$

where $\epsilon \mid \sigma^2 \sim N\left(0, \sigma^2\right)$.

- Under the conjugate prior $\beta \sim N_p\left(0,\ \Lambda_0^{-1}\right)$, the posterior is $\beta \mid y \sim N\left(\mu_n,\ \Lambda_n^{-1}\right)$, where

$$\mu_n = \left(\Lambda_0 + X^T X\right)^{-1} X^T y.$$

- Suppose that we observe a new x_0. and want to predict the new y_0. The predictive distribution is

$$y_0 \mid y \sim N\left(x_0^T \mu_n,\ \sigma^2 + x_0^T \Lambda_n^{-1} x_0\right).$$

Transform x

If we transform $x \in \mathbb{R}^p$ and obtain $\phi(x) \in \mathbb{R}^d$, then we can consider the linear regression model

$$y = f(x) + \epsilon, \qquad f(x) = \phi^T(x)\gamma,$$

where $\epsilon \mid \sigma^2 \sim N\left(0, \sigma^2\right)$.

- Under the conjugate prior $\gamma \sim N_d\left(0,\ \Omega_0^{-1}\right)$, the predictive distribution is

$$y_0 \mid y, \sigma^2 \quad \sim \quad N\left(\phi^T(x_0)\mu_n,\ \sigma^2 + \phi^T(x_0)\Omega_n^{-1}\phi(x_0)\right),$$

  where $\mu_n = \left(\sigma^2\Omega_0 + \phi^T(X)\phi(X)\right)^{-1}\phi^T(X)y$.
- The predictor is not linear in x_0 but linear in $\phi(x_0)$.

Kernel Function

A function $\kappa(x, z)$ is a kernel function if

1. it is symmetric, $\kappa(x, z) = \kappa(z, x)$,
2. the kernel matrix K with (i, j)th entry $\kappa(x_i, x_j)$ is positive semi-definite for all x_1, ..., x_n.

Example

Show that $\kappa(x, z) = x^T \Lambda_0^{-1} z$ is a kernel function for a symmetric Λ_0.

Rewrite Predictive Distribution

Let $\Phi = \phi(X) = \begin{bmatrix} \phi^T(x_1) \\ \vdots \\ \phi^T(x_n) \end{bmatrix}$. We can show that

$$\Omega_0^{-1}\Phi^T\left(\sigma^2 I_n + \Phi\Omega_0^{-1}\Phi^T\right)^{-1} = \left(\sigma^2\Omega_0 + \Phi^T\Phi\right)^{-1}\Phi^T.$$

Hence, the predictor from the predictive distribution is

$$\begin{aligned} \phi^T(x_0)\mu_n &= \phi^T(x_0)\left(\sigma^2\Omega_0 + \Phi^T\Phi\right)^{-1}\Phi^T y \\ &= \phi^T(x_0)\Omega_0^{-1}\Phi^T\left(\sigma^2 I_n + \Phi\Omega_0^{-1}\Phi^T\right)^{-1} y, \end{aligned}$$

where $\phi^T(x_0)\Omega_0^{-1}\Phi^T$ is a $1 \times n$ vector with elements $\left\{\phi^T(x_0)\Omega_0^{-1}\phi(x_i)\right\}$ and $\Phi\Omega_0^{-1}\Phi^T$ is a $n \times n$ matrix with elements $\left\{\phi(x_i)\Omega_0^{-1}\phi^T(x_j)\right\}$.

Example

Show that $\kappa(x, z) = \phi^T(x)\Omega_0^{-1}\phi(z)$ is a kernel function for a symmetric Ω_0.

Predictive Distribution Using Kernel Function

If $\kappa(x,z)$ is a kernel function, then we can find a function $\psi()$ such that $\kappa(x,z) = \psi^T(x)\psi(z)$.

- $\kappa(x,z) = \phi^T(x)\Omega_0^{-1}\phi(z) = \left[\Omega_0^{-1/2}\phi(x)\right]^T \Omega_0^{-1/2}\phi(z)$, where $\psi(x) = \Omega_0^{-1/2}\phi(x)$.

We can express the predictor from the predictive distribution as

$$\phi^T(x_0)\mu_n = K(x_0,X)\left[\sigma^2 I_n + K(X,X)\right]^{-1} y,$$

where

$$K(x_0,X) = \left\{\phi^T(x_0)\Omega_0^{-1}\phi(x_i)\right\} = \left\{\psi^T(x_0)\psi(x_i)\right\}$$

is a $1\times n$ vector and

$$K(X,X) = \left\{\phi^T(x_i)\Omega_0^{-1}\phi(x_j)\right\} = \left\{\psi^T(x_i)\psi(x_j)\right\}$$

is a $n\times n$ matrix.

Kernel Trick

Our predictor $\phi^T(x_0)\,\mu_n$ depends on x only through the inner products $\psi^T(x)\,\psi(z)$ such as $\left\{\psi^T(x_0)\,\psi(x_i)\right\}$ and $\left\{\psi^T(x_i)\,\psi(x_j)\right\}$.

- Kernel trick is a commonly used trick to create new features from your original observed features, if our prediction depends on x only through inner products.
- By varying the kernel function, we obtain different sets of $\phi(x)$ and $\psi(x)$ as our new features.

Create New Feature

If $\kappa(x,z)$ is a kernel function, then we will have an eigen-decomposition

$$\kappa(x,z) = \sum_{m=1}^{\infty} \rho_m e_m(x) e_m(z),$$

for some eigenvalues ρ_k and eigenfunctions $e_m(x)$.

It can possibly be viewed as infinite new features have been created as

$$\kappa(x,z) = \sum_{m=1}^{\infty} \underbrace{\sqrt{\rho_m} e_m(x)}_{\text{new feature } \psi_m(x)} \underbrace{\sqrt{\rho_m} e_m(z)}_{\text{new feature } \psi_m(z)}.$$

Bayesian Regression and Gaussian Process

The predictive distribution $y_0 \mid y, \sigma^2$ is Gaussian with

$$\begin{aligned} \text{mean} \qquad & K\left(x_0, X\right)\left[\sigma^2 I_n + K\left(X, X\right)\right]^{-1} y, \\ \text{variance} \qquad & \sigma^2 + \phi^T\left(x_0\right)\left(\Omega_0 + \sigma^{-2}\Phi^T\Phi\right)^{-1}\phi\left(x_0\right). \end{aligned}$$

We can show that the variance is equivalent to

$$K\left(x_0, x_0\right) - K\left(x_0, X\right)\left(\sigma^2 I_n + K\left(X, X\right)\right)^{-1} K\left(X, x_0\right).$$

Recall that in Gaussian process regression, $f\left(x_0\right) \mid y, \sigma^2$ is also Gaussian with

$$\begin{aligned} \text{mean} \qquad & K\left(x_0, X\right)\left[\sigma^2 I_n + K\left(X, X\right)\right]^{-1} y, \\ \text{covariance} \qquad & K\left(x_0, x_0\right) - K\left(x_0, X\right)\left[K\left(X, X\right) + \sigma^2 I_n\right]^{-1} K\left(X, x_0\right). \end{aligned}$$

They are the same thing!

Prior on Function

Consider the linear regression model

$$y = f(x) + \epsilon, \qquad f(x) = \phi^T(x)\gamma,$$

where $\epsilon \mid \sigma^2 \sim N\left(0, \sigma^2\right)$.

- The conjugate prior $\gamma \sim N_d\left(0,\ \Omega_0^{-1}\right)$ implies that

$$f(x) \quad \sim \quad N\left(0,\ \phi^T(x)\,\Omega_0^{-1}\phi(x)\right).$$

  It can be viewed as the function has a Gaussian prior.

- The prior distribution of any set of function values satisfies

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} = \phi(X)\gamma \quad \sim \quad N_n\left(0,\ \phi(X)\,\Omega_0^{-1}\phi^T(X)\right).$$

Posterior on Function

The corresponding posterior is

$$\gamma \mid y, \sigma^2 \quad \sim \quad N\left(\mu_n,\, \Omega_n^{-1}\right),$$

where

$$\mu_n \quad = \quad \left(\sigma^2 \Omega_0 + \phi^T(X)\,\phi(X)\right)^{-1} \phi^T(X)\, y.$$

It can be viewed as the function has a Gaussian posterior

$$f(x) \mid y, \sigma^2 \quad \sim \quad N\left(\mu_n,\, \phi^T(x)\,\Omega_n^{-1}\phi(x)\right).$$

The predictive distribution is also Gaussian.