## Part I: Gaussian Distribution.

### 1° Gaussian Distribution in Vector form.

$$N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

$x \in \mathbb{R}^D$ ; $\mu \in \mathbb{R}^D$ ; $\Sigma \in \mathbb{R}^{D\times D}$

### 2° Conditional Gaussian

$x \in \mathbb{R}^D$ $\quad x \triangleq \begin{bmatrix} x_a \\ x_b \end{bmatrix}$ $\quad x_a \in \mathbb{R}^M$ $\quad x_b \in \mathbb{R}^{D-M}$

Equiv.

$\mu \in \mathbb{R}^D$ $\quad \mu \triangleq \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}$ $\quad \mu_a \in \mathbb{R}^M$ $\quad \mu_b \in \mathbb{R}^{D-M}$

$\Rightarrow \Sigma \in \mathbb{R}^{D\times D}$ $\quad \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$

$\Sigma_{aa} \in \mathbb{R}^{M\times M}$ $\quad \Sigma_{ab} \in \mathbb{R}^{M\times(D-M)}$ $\quad \Sigma_{ba} \in \mathbb{R}^{(D-M)\times M}$ $\quad \Sigma_{bb} \in \mathbb{R}^{(D-M)\times(D-M)}$

$\Sigma_{ab}^T = \Sigma_{ba}$

$\Lambda := \Sigma^{-1}$ ⇐ Precision Matrix

For the quadratic term.

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = -\frac{1}{2}\left[\begin{pmatrix} x_a \\ x_b \end{pmatrix} - \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}\right]^T \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}^{-1} \left[\begin{pmatrix} x_a \\ x_b \end{pmatrix} - \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}\right]$$

$$= -\frac{1}{2} \begin{bmatrix} x_a-\mu_a \\ x_b-\mu_b \end{bmatrix}^T \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}^{-1} \begin{bmatrix} x_a-\mu_a \\ x_b-\mu_b \end{bmatrix}$$

$$= -\frac{1}{2}(x_a-\mu_a)^T\Sigma_{aa}^{-1}(x_a-\mu_a) - \frac{1}{2}(x_a-\mu_a)^T\Sigma_{ab}^{-1}(x_b-\mu_b)$$

$$-\frac{1}{2}(x_b-\mu_b)^T\Sigma_{ba}^{-1}(x_a-\mu_a) - \frac{1}{2}(x_b-\mu_b)^T\Sigma_{bb}^{-1}(x_b-\mu_b)$$

Let the above be seen as a function of $x_a$:

$$P(x_a|x_b) = -\frac{1}{2}(x_a-\mu_a)^T\Sigma_{aa}^{-1}(x_a-\mu_a) - \frac{1}{2}(x_a-\mu_a)^T\Sigma_{ab}^{-1}(x_b-\mu_b)$$

$$-\frac{1}{2}(x_b-\mu_b)^T\Sigma_{ba}^{-1}(x_a-\mu_a) - \frac{1}{2}(x_b-\mu_b)^T\Sigma_{bb}^{-1}(x_b-\mu_b)$$

$$= -\frac{1}{2}x_a^T\Lambda_{aa}x_a + x_a^T\{\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)\} + \text{const}$$

$$= -\frac{1}{2}x_a^T\Lambda_{aa}x_a + x_a^T\cdot\Lambda_{aa}^{-1}\cdot\Lambda_{aa}\{\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)\} + \text{const}$$

For the quadratic term of the Gaussian:

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu + \text{const}$$

$$\therefore \Sigma_{a|b} = \Lambda_{aa}^{-1} = \Sigma_{aa}$$

$$\mu_{a|b} = \Lambda_{aa}^{-1}\cdot\{\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)\}$$

$$= \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b)$$

$$\Rightarrow P(x_a|x_b) \sim N(\mu_{a|b}, \Sigma_{a|b})$$

Lemma:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1}+D^{-1}CMBD^{-1} \end{bmatrix}$$

$$M = (A - BD^{-1}C)^{-1} \quad \text{"schur component"}$$

$$\Rightarrow \because \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}$$

$$\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}$$

$$\Lambda_{ab} = -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}$$

$$\Rightarrow \Sigma_{a|b} = \Lambda_{aa}^{-1} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} = Var(a) - \frac{Cov(a,b)}{Var(b)}$$

$$\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b)$$

$$= \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - x_a)$$

⇓

Lemma: If a joint distribution $p(x_a, x_b)$ is Gaussian, then the conditional distribution is Gaussian as well.

3° Marginal Gaussian

$$p(x_a) = \int p(x_a, x_b)\, dx_b$$

For the quadratic term in $p(x_a, x_b)$:

$$-\frac{1}{2}(x_a - \mu_a)^T\Sigma_{aa}^{-1}(x_a - \mu_a) - \frac{1}{2}(x_a - \mu_a)^T\Sigma_{ab}^{-1}(x_b - \mu_b)$$

$$-\frac{1}{2}(x_b - \mu_b)^T\Sigma_{ba}^{-1}(x_a - \mu_a) - \frac{1}{2}(x_b - \mu_b)^T\Sigma_{bb}^{-1}(x_b - \mu_b)$$

For terms involving $x_b$:

$$-\frac{1}{2}x_b^T\Sigma_{bb}^{-1}x_b + x_b^T\left[\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a)\right]$$

let $m := \Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a)$

$$\Rightarrow -\frac{1}{2}x_b^T\Sigma_{bb}^{-1}x_b + x_b^T m = -\frac{1}{2}(x_b - \Lambda_{bb}^{-1}m)^T\Lambda_{bb}(x_b - \Lambda_{bb}^{-1}m) + \frac{1}{2}m^T\Lambda_{bb}^{-1}m$$

$$\int \exp\left\{-\frac{1}{2}(x_b - \Lambda_{bb}^{-1}m)^T \Lambda_{bb} (x_b - \Lambda_{bb}^{-1}m)\right\} dx_b$$

= the reciprocal of the normalization coefficient of the Gaussian distribution.

For the left term $\frac{1}{2}m^T\Lambda_{bb}m$ and the remaining terms in the quadratic form:

$$\frac{1}{2}\left[\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a)\right]^T \Lambda_{bb}^{-1}\left[\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a)\right]$$

$$-\frac{1}{2}x_a^T\Lambda_{aa}x_a + x_a^T(\Lambda_{aa}\mu_a + \Lambda_{ab}\mu_b) + \text{const.}$$

$$= -\frac{1}{2}x_a^T(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})x_a + x_a^T(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1}\mu_a + \text{const}$$

⇒ The mean and variance of $p(x_a)$:

$$\Sigma_a = (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1}$$

$$\mu_a = \Sigma_a(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})\mu_a$$

$$\because \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}^{-1} = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$$

$$\therefore \Sigma_{aa} = (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1}$$

$$\therefore E[x_a] = \mu_a \qquad cov(x_a) = \Sigma_{aa}.$$

Summary:

$x \sim \mathcal{N}(x \mid \mu, \Sigma) \quad ; \Lambda := \Sigma^{-1}$

$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$

$\Lambda = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}$

$\Rightarrow p(x_a \mid x_b) = \mathcal{N}(x \mid \mu_{a|b}, \Lambda_{aa}^{-1})$

$\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab} (x_b - \mu_b)$

$p(x_a) = \int p(x_a, x_b)\, dx_b = \mathcal{N}(x_a \mid \mu_a, \Sigma_{aa})$

4° Bayesian's theorem for Gaussian Random variables.

Suppose: we are given a Gaussian marginal distribution $p(x)$ and a Gaussian conditional distribution $p(y|x)$; $p(y|x)$ has a mean that is linear function of $x$ and a covariance that is independent of $x$ "Linear Gaussian Model"

we want to find $p(y)$ and $p(x|y)$.

$p(x) = \mathcal{N}(x \mid \mu, \Lambda^{-1})$

$p(y|x) = \mathcal{N}(y \mid Ax+b, L^{-1})$

$x \in \mathbb{R}^D, \quad y \in \mathbb{R}^M \qquad A \in \mathbb{R}^{M \times D}$

let $z := \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^{D+M}$

$\Rightarrow p(z) = p(x)\, p(y|x)$

$\ln p(z) = \ln p(x) + \ln p(y|x)$

$= -\frac{1}{2}(x-\mu)^T \Lambda (x-\mu) - \frac{1}{2}(y-Ax-b)^T L (y-Ax-b) + \text{const}$

For the second-order terms:

$-\frac{1}{2}x^T(\Lambda + A^T L A)x - \frac{1}{2}y^T L y + \frac{1}{2}y^T L A x + \frac{1}{2}x^T A^T L y$

$= -\frac{1}{2}\begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} \Lambda + A^T L A & -A^T L \\ -LA & L \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$

let $R = \begin{bmatrix} \Lambda + A^T L A & -A^T L \\ -LA & L \end{bmatrix}$

$\Rightarrow$ The second-order term: $-\frac{1}{2}z^T R z$

By the Lemma: $\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix}$

$M = (A - BD^{-1}C)^{-1}$

$\text{cov}(z) = R^{-1} = \begin{bmatrix} \Lambda + A^T L A & -A^T L \\ -LA & L \end{bmatrix}^{-1}$

$= \begin{bmatrix} \Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{bmatrix}$

$\Rightarrow$ For the linear terms:

$x^T \Lambda \mu - x^T A^T L b + y^T L b = \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} \Lambda\mu - A^T L b \\ Lb \end{bmatrix}$

$\Rightarrow E[z] = R^{-1} \begin{bmatrix} \Lambda\mu - A^T L b \\ Lb \end{bmatrix}$

$$= \begin{bmatrix} \Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1}+A\Lambda^{-1}A^T \end{bmatrix} \begin{bmatrix} \Lambda\mu - A^TLb \\ Lb \end{bmatrix}$$

$$= \begin{bmatrix} \mu \\ A\mu+b \end{bmatrix}$$

$$p(y) = \int p(z)\, dx$$

$$E[y] = A\mu+b$$

$$\text{cov}(y) = \Sigma_{yy} = L^{-1} + A\Lambda^{-1}A^T$$

For $x|y$:

$$E[x|y] = \mu_x - \Lambda_{xx}^{-1}\Lambda_{xy}(y-\mu_y) = E[x] + \frac{\text{cov}(x,y)}{\text{var}(y)}(y-\mu_y)$$

$$= (\Lambda + A^TLA)^{-1}\{A^TL(y-b)+\Lambda\mu\}$$

$$\text{cov}(x|y) = (\Lambda + A^TLA)^{-1}$$

## Part II. Bayesian Curve-fitting.

Settings: given $N$ input values $(x_1,\ldots,x_N)^T$, and the respective target values $(t_1,\ldots,t_N)^T$

The uncertainty of the target values are expressed as probability distribution.

$$p(t|x,w,\beta) = N(t|y(x,w),\beta^{-1}), \quad w \text{ is the parameter}$$

we have a belief of the parameter $w$, we express this belief as the form of a prior distribution:

$$p(w|\alpha) = N(w|0, \alpha^{-1}I) = \left(\frac{\alpha}{2\pi}\right)^{\frac{M+1}{2}} \exp\left\{-\frac{\alpha}{2}w^T w\right\}$$

Now, we are given a known data set $X_{data}$ and their target value $T_{data}$ along with a new point $x$, we want to predict the distribution of $y$

$\Rightarrow$ which is

$$p\{t|x, X_{data}, T_{data}\} = \int p\{t|x,w\}\, p\{w|X_{data}, T_{data}\}\, dw$$

$$p\{t|x,w\} \sim N(t|y(x,w), \beta^{-1})$$

$p\{w|X_{data}, T_{data}\}$ is the posterior distribution over $w$

## Part III: Bayesian Regression.

∴ We assume the distribution of our target value is:

$$p\{t|X,w,\beta\} = \prod_{n=1}^{N} N(t_n|w^T\phi(x_n), \beta^{-1})$$

which is the exponential of a quadratic function of $w$.

∴ The corresponding of a conjugate prior is:

$$p(w) = N(w|m_0, S_0)$$

By the conjugate property of exponentials, the posterior is also a Gaussian.

$$p(w|t) = N(w|m_N, S_N)$$

∵ By the conditional Gaussian:

$$E[x|y] = \Sigma\{A^T L(y-b) + \Lambda\mu\}$$

$$\text{cov}[x|y] = \Sigma = (\Lambda + A^T L A)^{-1}$$

∴ $$S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi$$

$$m_N = S_N(S_0^{-1} m_0 + \beta \Phi^T t)$$

∵ The mode of a Gaussian is its mean

∴ The maximum posterior estimation of the weight vector is:

$$W_{MAP} = m_N = S_N(S_0^{-1} m_0 + \beta \Phi^T t)$$

If the data points arrive sequentially, then the posterior distribution at any stage serve as the prior distribution of the next point.

We can view this mode as a point estimate of $w$.

Formally, a point estimate in the Bayesian model is defined as the estimator that minimizes the Bayesian risk. (Bayes Estimator)

Bayesian risk: $R(\theta, \delta) = E_\theta[L(\theta, \delta(x))] = \int_X L(\theta, \delta(x)) f(x|\theta)\, dx$

$$B_\pi(\hat{\theta}) = \int R(\theta, \hat{\theta})\, \pi(\theta)\, d\theta = E[E[L(\theta, L(x))|\theta]]$$

$$\hat{\theta}^* = \arg\min_{\hat{\theta}} B_\pi(\hat{\theta})$$

By the tower property:

$$B_\pi(\hat{\theta}) = E[L(\theta, \delta(x))] = E[E[L(\theta, \delta(x))|\theta]]$$

Theorem:

If: 1° there exists $\delta_0$ an estimator of $\theta$ with finite risk for all $\theta$

2° there exists a $\delta'(x)$ that minimizes:

$E[L(\theta, \delta(x)) \mid X=x]$ for almost every $x$

then $\delta'(x)$ is a Bayes estimator for $\theta$

proof:

$$E[L(\theta, \delta'(x)) \mid X=x] \le E[L(\theta, \delta(x)) \mid X=x]$$ for almost every $x$

By taking expectation over $X$:

$$E[L(\theta, \delta'(x))] \le E[L(\theta, \delta(x))]$$

$\therefore \hat{\theta}^* = \arg\min_{\hat{\theta}} B_\pi(\hat{\theta}) = \arg\min_{\hat{\theta}} E[L(\theta, \hat{\theta}) \mid X=x]$

If we take the loss function as MSE:

we want to minimize the posterior risk:

$$\min E[\pi(\theta)(\hat{\theta} - \theta)^2 \mid X=x]$$

∵ The expectation has nothing to do with $\hat{\theta}$

∵ Take $\hat{\theta}$ out of the expectation:

$$\hat{\theta}^2 E[\pi(\theta) | x=x] - 2\hat{\theta} E[\pi(\theta)\theta | x=x] + E[\pi(\theta)\theta^2 | x=x]$$

↑
A convex quadratic function of $\hat{\theta}$

$$\therefore \hat{\theta}^* = \frac{E[\pi(\theta)\theta | x=x]}{E[\pi(\theta) | x=x]}$$

$$\Rightarrow \hat{\theta}^* = E_\theta(\theta | data) = \frac{\int_\theta \theta P(x=x|\theta)\pi(\theta) d\theta}{\int_\theta P(x=x|\theta)\pi(\theta) d\theta}$$

which is just the posterior mean.

The predictive distribution:

$\tilde{X}$, $\tilde{t}$ come from the training set, we want to predict the label of the testing set $x$

$$\Rightarrow p(t | x, \tilde{X}, \tilde{t}) = \int p(t | w, \beta) p(w | \alpha, \beta) dw$$

$$p(t | w, \beta) = N(t | y(x, w), \beta^{-1})$$

$$p(w | \alpha, \beta) = N(w | m_N, S_N)$$

∵ $p(t | x, \tilde{X}, \tilde{t})$ equals a convolution of two Gaussian distribution.

∴ By making the result of the conditional Gaussian.

$$p(y) = \int p(x,y)\,dx = \int p(y|x)\,p(x)\,dx$$

$$= N(y \mid A\mu + b,\ L^{-1} + A\Lambda^{-1}A^T)$$

$$\Rightarrow p(t|x, \hat{x}, \hat{t}) = N\left\{t \mid m_N^T \phi(x),\ \frac{1}{\beta} + \phi(x)^T S_N \phi(x)\right\}$$

## Part IV: Ridge Regression and Bayesian Regression.

The Bayesian regression assumes the parameter $w$ and $\sigma^2$ to be the random variables, while take $X$ and $Y$ to be fixed.

let $w \in \mathbb{R}^D$:

$$w \mid \sigma^2 \sim N(0_D,\ \sigma^2 \lambda^{-1} I_{DD})$$

$$\sigma^2 \sim IG(\alpha_0, \beta_0)$$

IG is the inverse Gamma distribution. why choose IG?

Due to the computation advantage of the conjugate priors.

$$f(x, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, x^{-\alpha-1}\, e^{-\frac{\beta}{x}}\, I(x>0)$$

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\,dx$$

⇒ The joint probability of $w$ and $\sigma^2$ are:

$$P\{w, \sigma^2 \mid \tilde{t}, \tilde{X}\} \propto P\{\tilde{t} \mid \tilde{X}, w, \sigma^2\} \cdot P\{w \mid \sigma^2\} \cdot P\{\sigma^2\}$$

Due to the Gaussian assumption of the data distribution.

$$\propto \sigma^{-n} \exp\left[-\frac{1}{2}\sigma^{-2}(\tilde{t} - \tilde{X}w)^T(\tilde{t} - \tilde{X}w)\right] \cdot \sigma^{-D} \exp\left[-\frac{1}{2}\sigma^{-2}\lambda w^T w\right] \cdot (\sigma^2)^{-\alpha_0 - 1} \exp\left[-\frac{1}{2}\sigma^{-2}\beta_0\right]$$

$$= \sigma^{-n-D-2\alpha_0-2} \exp\left\{-\frac{1}{2}\sigma^{-2}\left[(\tilde{t} - \tilde{X}w)^T(\tilde{t} - \tilde{X}w) + \lambda w^T w + \beta_0\right]\right\}.$$

↑

We can consider the above as a multivariate Gaussian distribution. with respective to $w$, with terms involving

$$\Rightarrow (\tilde{t} - \tilde{X}w)^T(\tilde{t} - \tilde{X}w) + \lambda w^T w$$

$$= \tilde{t}^T\tilde{t} - w^T\tilde{X}^T\tilde{t} - \tilde{t}^T\tilde{X}w + w^T\tilde{X}^T\lambda w + \lambda w^T w$$

$$= \tilde{t}^T\tilde{t} - w^T(\tilde{X}^T\tilde{X} + \lambda I_{DD})(\tilde{X}^T\tilde{X} + \lambda I_{DD})^{-1} \cdot \tilde{X}^T \cdot \tilde{t}$$

$$- \tilde{t}^T\tilde{X}(\tilde{X}^T\tilde{X} + \lambda I_{DD})^{-1}(\tilde{X}^T\tilde{X} + \lambda I_{DD}) \cdot w + w^T(\tilde{X}^T\tilde{X} + \lambda I_{DD})w.$$

Recall that in Ridge regression, the analytical solution for $w$ is:

$$w(\lambda) = (\tilde{X}^T\tilde{X} + \lambda I_{DD})^{-1} \cdot \tilde{X} \cdot \tilde{t}$$

∴ The above is equivalent to:

$$\tilde{t}^T\tilde{t} - w^T(\tilde{X}^T\tilde{X} + \lambda I_{DD}) \cdot w(\lambda) - w(\lambda)^T(\tilde{X}^T\tilde{X} + \lambda I_{DD})w + w^T(\tilde{X}^T\tilde{X} + \lambda I_{DD})w$$

$$= \tilde{t}^T\tilde{t} - \tilde{t}^T\tilde{X}(\tilde{X}^T\tilde{X} + \lambda I_{DD})^{-1}\tilde{X}^T\tilde{t} + [w - w(\lambda)]^T(\tilde{X}^T\tilde{X} + \lambda I_{DD})[w - w(\lambda)]$$

$$\Rightarrow \quad p\{w, \sigma^2 | \bar{t}, \bar{x}\} \propto f_w(w | \bar{t}, \bar{x}) \cdot f(\sigma^2 | \bar{t}, \bar{x})$$

$$\text{with} \quad f_w(w | \bar{t}, \bar{x}) \propto \exp\left\{-\frac{1}{2}\sigma^2 [w - w(\lambda)]^T [\bar{x}^T \bar{x} + \lambda I_{DD}] [w - w(\lambda)]\right\}$$

$$f\{\sigma^2 | \bar{t}, \bar{x}\} \propto (\sigma^2)^{-\kappa_0 - 1} \exp\left\{-\frac{1}{2}\sigma^2 \beta_0\right\}$$

$\Rightarrow$ The conditional prior mean of $w$ is $E[w | \sigma^2, \bar{x}, \bar{t}] = w(\lambda)$

$$\Rightarrow \quad w_{MAP} = w(\lambda) = (\bar{x}^T \bar{x} + \lambda I_{DD})^{-1} \bar{x}^T \bar{t}$$

The penalty parameter $\lambda$ can be seen as the belief of the precision of the prior distribution. ( the scaled parameter of the distribution)

If $\lambda$ is large, we are confident about the prior

Actually, any penalized estimator has a Bayesian equivalent

The Bayesian estimator is:

$$\hat{w}_{map} = \arg\max_w \pi\{w | x = x\} = \arg\max_w \frac{p\{x = x | w\}\, \pi(w)}{\int p\{x = x | w\}\, \pi(w)\, dw}$$

$$\Rightarrow \hat{w}_{MAP} = \arg\max_w \log p\{x = x | w\} + \log[\pi(w)]$$

↑ log likelihood

$\because \log(\pi(w)) \propto \sigma^{-2}\lambda\|w\|_2^2$ If we use normal prior

$\therefore$ If we use other prior distribution, we can model other penalty terms.

## Part V: Conjugate priors.

Recall the general process of the Bayesian curve-fitting:

prior distribution: $p(\theta)$

$$p(\theta|x) = \frac{p(x|\theta)\cdot p(\theta)}{p(x)}$$

$$p(x) = \int p(x|\theta)\, p(\theta)\, d\theta$$

$$p(x_{new}|x) = \int p(x_{new}|\theta)\, p(\theta|x)\, d\theta$$

$\therefore$ Compute the integral every time is computational complex

$\therefore$ We want to

1° $p(x) = \int p(x|\theta)\, p(\theta)\, d\theta$ can be obtained tractably.

2° $p(x_{new}|x) = \int p(x_{new}|\theta)\, p(\theta|x)\, d\theta$ can be obtained tractably.

Take univariate Gaussian as an example:

$$p\{x|u,\sigma^2\} \propto (\sigma^2)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}\sigma^{-2}(x-u)^2\}.$$

1° Conjugacy for the mean.

Assume the conjugate prior of the mean is Gaussian:

$$p\{\mu|\mu_0,\sigma_0^2\} = (\sigma_0^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\sigma_0^{-2}(\mu-\mu_0)^2\right\}$$

To obtain the posterior $p\{\mu|x\}$

If only one data point is observed:

$$E[\mu|x] = E[\mu] + \frac{Cov(\mu,x)}{Var(x)}(x - E[x])$$

$$Var[\mu|x] = Var(\mu) - \frac{Cov^2(x,\mu)}{Var(x)}$$

Write: $x = \mu + \sigma\varepsilon \qquad \varepsilon \sim N(0,1)$

$$\mu = \mu_0 + \sigma_0 \cdot \delta \qquad \delta \sim N(0,1)$$

$$E[x] = E[\mu] + \sigma E[\varepsilon] = \mu_0$$

$$Var(x) = E[(x-\mu_0)^2] = E[\mu - \mu_0 + \sigma\varepsilon]^2 = \sigma^2 + \sigma_0^2$$

$$Cov(x,\mu) = E[(x-\mu_0)(\mu-\mu_0)] = E[(\mu-\mu_0+\sigma\varepsilon)(\mu-\mu_0)] = \sigma_0^2$$

$$\therefore \mu_{post} := E[\mu|x=x] = \mu_0 + \frac{\sigma_0^2}{\sigma^2+\sigma_0^2}(x-\mu_0) = \frac{\sigma_0^2}{\sigma+\sigma_0^2}x + \frac{\sigma^2}{\sigma^2+\sigma_0^2}\mu_0$$

$$\sigma^2_{post} := Var(\mu|x=x) = \sigma_0^2 - \frac{\sigma_0^4}{\sigma^2+\sigma_0^2} = \frac{\sigma\sigma_0^2}{\sigma^2+\sigma_0^2}$$

Expressed in precision:

$$\tau = \frac{1}{\sigma^2} \quad ; \quad \tau_0 = \frac{1}{\sigma_0^2}$$

$$E[\mu|x=x] = \frac{\tau}{\tau+\tau_0}x + \frac{\tau_0}{\tau+\tau_0}\mu_0$$

$$\tau_{post} = \bar{\tau} + \tau_0.$$

For the multiple data input:

$$\chi = (x_1, \dots, x_n), \quad x_i \sim N(\mu, \sigma^2)$$

Expressed in precision:

$$P(x|\mu,\tau) \propto \tau^{\frac{1}{2}} e^{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2$$

$$= \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2$$

$\sum_{i=1}^{n}(x_i-\bar{x})^2$ is constant

$\bar{x}$ is a sufficient statistics with respect to $\mu$.

∴ The problem is reduced to univariate case.

$$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$$

$$\therefore \quad \mu_{post} = \frac{n\tau}{n\tau+\tau_0}\bar{x} + \frac{\tau_0}{n\tau+\tau_0}\mu_0$$

$$\tau_{post} = n\tau + \tau_0.$$

$$P\{x_{new}|x,\tau\} = \int P\{x_{new}|x,\mu,\tau\}\, P\{\mu|x,\tau\}\, d\mu$$

$$= \int P\{x_{new}|\mu,\tau\}\, P\{\mu|x,\tau\}\, d\mu$$

This is again the convolution of two Gaussian

∴ $x_{new}|x,\tau$ is Gaussian distributed with mean $\mu_{post}$, variance $\sigma^2_{post} + \sigma^2$.

2° Conjugacy for variance

$$p\{x \mid \mu, \sigma^2\} \propto (\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\sigma^{-2}(x-\mu)^2\right\}$$

$$= (\sigma^2)^{a} e^{-b/\sigma^2}$$

Assume the conjugate prior distribution of $\sigma^2$ is inverse Gamma:

$$p\{\sigma^2 \mid \alpha, \beta\} = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} e^{\frac{-\beta}{\sigma^2}}$$

The posterior:

$$p\{\sigma^2 \mid x, \mu, \alpha, \beta\} \propto p\{x \mid \mu, \sigma^2\}\, p\{\sigma^2 \mid \alpha, \beta\}$$

$$\propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right\} (\sigma^2)^{-\alpha-1} e^{-\beta/\sigma^2}$$

$$= (\sigma^2)^{-(\alpha+\frac{n}{2})-1} \exp\left\{-(\beta + \tfrac{1}{2}\textstyle\sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2)\right\}$$

$$\sim IGa\left(\alpha+\frac{n}{2},\ \beta+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

Expressed in terms of the precision:

$\because \sigma^2 \sim IGa(\alpha, \beta)$

$\therefore \tau = \frac{1}{\sigma^2} \sim Ga(\alpha, \beta)$

$$\Rightarrow p\{\tau \mid x, \mu, \alpha, \beta\} \propto p\{x \mid \mu, \tau\}\, p\{\tau \mid \alpha, \beta\}$$

$$\propto \tau^{\frac{n}{2}} \exp\left\{-\frac{\tau}{2}\textstyle\sum_{i=1}^{n}(x_i-\mu)^2\right\} \tau^{\alpha-1} \cdot \exp\{-\beta\tau\}$$

$$= \tau^{\alpha+\frac{n}{2}+1} \exp\left\{-\left(\beta+\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\right\}$$

$$\Rightarrow \quad \tau \mid x \sim Gac\left(\alpha+\frac{n}{2},\ \beta+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$p\{x_{new} \mid x, \mu, \alpha, \beta\} = \int p\{x_{new} \mid x, \mu, \tau\}\, p\{\tau \mid x, \mu, \alpha, \beta\}\, d\tau$$

$$= \int p\{x_{new} \mid \mu, \tau\}\, p\{\tau \mid x, \mu, \alpha, \beta\}\, d\tau$$

$$= \int \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha+1} e^{-\beta\tau} \left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{\tau}{2}(x-\mu)^2} d\tau$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(2\pi)^{\frac{1}{2}}} \int \tau^{\alpha-\frac{1}{2}} e^{-\beta\tau} e^{-\frac{\tau}{2}(x-\mu)^2} d\tau$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(2\pi)^{\frac{1}{2}}} \int \tau^{\alpha-\frac{1}{2}} e^{-\left(\beta+\frac{1}{2}(x-\mu)^2\right)\tau} d\tau$$

// An unnormalized Gamma integral

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(2\pi)^{\frac{1}{2}}} \frac{\Gamma\left(\alpha+\frac{1}{2}\right)}{\left(\beta+\frac{1}{2}(x-\mu)^2\right)^{\alpha+\frac{1}{2}}}$$

$$= \frac{\Gamma\left(\alpha+\frac{1}{2}\right)}{\Gamma(\alpha)} \frac{1}{(2\pi\beta)^{\frac{1}{2}}} \frac{1}{\left(1+\frac{1}{2\beta}(x-\mu)^2\right)^{\alpha+\frac{1}{2}}}$$

// Generalized student's t.

let $p := \frac{\alpha}{2}$, $\lambda := \frac{\alpha}{\beta}$.

$$p\{x \mid \mu, \alpha, \beta\} = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \left(\frac{\lambda}{\pi p}\right)^{\frac{1}{2}} \frac{1}{\left(1+\frac{\lambda}{p}(x-\mu)^2\right)^{\frac{p+1}{2}}}$$

## Part VI: Markov Chain Monte Carlo.

When we don't use conjugate prior, we can't yield a tractable closed form formula of the posterior.

△ How to use Markov Chain:

1° Construct a Markov chain whose stationary probability equals to the desired probability

2° Iterate until the chain is converged and reaches stationary.

3°. Draw sample from this stationary chain.

△ Monte-Carlo Integration.

Assume the analytical solution of $\mathbb{E}[h|y] = \int h(y)\pi(y)\,dy$ is impossible

Draw enough samples from $\pi(\cdot)$: $y^{(1)}, \dots, y^{(n)}$

⇒ The estimated expectation: $\hat{\mathbb{E}}[h|y] = n^{-1}\sum_{i=1}^{n} h(y_i)$

By the Law of large numbers, it's a consistent estimator.

For the Bayesian framework:

$$P\{w \mid \tilde{t}, \vec{x}\} = \int_0^{\infty} P\{w, \sigma^2 \mid \tilde{t}, \vec{x}\}\, d\sigma^2 = \int_0^{\infty} P\{w \mid \sigma^2, \tilde{t}, \vec{x}\}\, P\{\sigma^2 \mid \tilde{t}, \vec{x}\}\, d\sigma^2$$

∵ The prior distribution of $\sigma^2$ is known.

∴ 1° Draw $\epsilon^{2,(1)}, \epsilon^{2,(2)}, \ldots, \epsilon^{2,(D)}$ from $p\{\epsilon^2 | \bar{t}, \bar{x}\}$

2° Draw $\{w^{(t)}\}_{t=1}^{T}$ from $p\{w | \epsilon^{2,(t)}, \bar{x}, \bar{t}\}$, $t = 1, \ldots, T$.

3°. ∵ From the above procedure, $\epsilon^2$ has been integrated out.

∴ the resulting $\beta^{(1)}, \ldots, \beta^{(T)}$ is from $p\{\beta | \bar{t}, \bar{x}\}$

How to sample from $\pi(y)$ is the key in this process.

Metropolis – Hasting samper as an example.

Task 1: Construct a Markov chain to model $\pi(\cdot)$.

For a markov chain $\{\gamma_t\}$, $\gamma_t \in \mathbb{R}^D$

The stationary distribution satisfies:

$$q(y_a) = \int_{\mathbb{R}^D} f_{\gamma_{t+1}|\gamma_t}(\gamma_{t+1} = y_a \mid \gamma_t = y_b)\, q(y_b)\, dy_b$$

The stationary distribution must satisfy the balance equation.

$$f_{\gamma_{t+1}|\gamma_t}(y_b | y_a)\, q(y_a) = f_{\gamma_{t+1}|\gamma_t}(y_a | y_b)\, q(y_b)$$

For mcmc, the stationary distribution is the target density where the sample is generated.

However, the transition kernel is unknown.

To have the train to work, we need to compute its transition probability.

For an arbitrary kernel, it's unlikely to be the valid one.

$$f_{\gamma_{t+1}|\gamma_t}(y_b | y_a)\, q(y_a) > f_{\gamma_{t+1}|\gamma_t}(y_a | y_b)\, q(y_b)$$

We need to correct this arbitrary kernel.

Given the current state $Y_t = y_t$, this kernel produce a suggestion $y_{t+1}$ for the next state of the Markov chain. Check this by the balance equation. If the value is too far from the stationary distribution, reject it and repeat until we find an acceptable one.

The probability of acceptance given the current state $Y_t = y_t$: $A(y_{t+1} | y_t)$

⇒ The constructed transition probability is equivalent to:

$$h_{Y_{t+1}|Y_t}(y_a | y_b) = g_{Y_{t+1}|Y_t}(y_a | y_b)\, A(y_a | y_b) + r(y_a)\, \delta_{y_b}(y_a)$$

$A(y_a | y_b)$ is the acceptance probability of the suggestion $y_a$:

$$A(y_a | y_b) = \min\left\{1, \frac{\pi(y_a)}{\pi(y_b)} \frac{g_{Y_{t+1}|Y_t}(y_b | y_a)}{g_{Y_{t+1}|Y_t}(y_a | y_b)}\right\}$$

$\delta_{y_b}(y_b)$ is a Dirac function.

Dirac function is a generalized function that can be used in describing random variables with mixture of a continuous part and a discrete part. It is characterized by returning the value 0 anywhere besides its mass.

In our case, the discrete part is $y_b$ since we want to model

the case that the chain remains the same state at $t+1$

$$r(y_a) = 1 - \int_{\mathbb{R}^D} q_{Y_{t+1}|Y_t}(y_a|y_b)\, A(y_a|y_b)\, dy_a$$

The probability to reject $y_a$.

We want to verify if the above defined transition kernel is valid for the Markov chain:

$$h_{Y_{t+1}|Y_t}(y_a|y_b)\, q(y_b) = h_{Y_{t+1}|Y_t}(y_b|y_a)\, q(y_a)$$

$$\Rightarrow q_{Y_{t+1}|Y_t}(y_a|y_b)\, A(y_a|y_b)\, q(y_b) + \boxed{r(y_a)\,\delta_{y_b|y_a}\, q(y_b)} \quad \text{"exists only when } y_a = y_b\text{"}$$

$$= q_{Y_{t+1}|Y_t}(y_b|y_a)\, A(y_b|y_a)\, q(y_a) + r(y_b)\,\delta_{y_a|y_b}\, q(y_a)$$

$$\Rightarrow q_{Y_{t+1}|Y_t}(y_a|y_b)\, A(y_a|y_b)\, q(y_b) = q_{Y_{t+1}|Y_t}(y_b|y_a)\, A(y_b|y_a)\, q(y_a)$$

To verify this equation:

If $A(y_a|y_b) < 1$:

$$\Rightarrow q_{Y_{t+1}|Y_t}(y_a|y_b) \frac{q(y_a)\, q_{Y_{t+1}|Y_t}(y_b|y_a)}{q(y_b)\, q_{Y_{t+1}|Y_t}(y_a|y_b)} \cdot q(y_b)$$

$$= q(y_a)\, q_{Y_{t+1}|Y_t}(y_b|y_a)$$

$\because A(y_a|y_b) < 1 \quad \therefore A(y_b|y_a) = 1$

∴ The equality holds.

$q(\cdot)$ is the valid stationary distribution can be verified by:

$$\int_{\mathbb{R}^D} h_{Y_{t+1}|Y_t}(y_a|y_b)\, q(y_b)\, dy_b = \int_{\mathbb{R}^D} h_{Y_{t+1}|Y_t}(y_b|y_a)\, q(y_a)\, dy_b$$

$$= \int_{\mathbb{R}^D} g_{Y_{t+1}|Y_t}(y_b|y_a)\, A(y_b|y_a)\, q(y_a)\, dy_b + \int_{\mathbb{R}^D} r(y_b)\, \delta_{y_a}(y_b)\, q(y_a)\, dy_b$$

$$= q(y_a) \int_{\mathbb{R}^D} g_{Y_{t+1}|Y_t}(y_b|y_a)\, A(y_b|y_a)\, dy_b + q(y_a) \int_{\mathbb{R}^D} r(y_b)\, \delta_{y_a}(y_b)\, dy_b$$

$$= q(y_a)\left[1 - r(y_a)\right] + r(y_a)\, q(y_a) = q(y_a)$$

## Part VII Empirical Bayes

Does not fully specify the priors, but only the form of the priors, leaving the hyperparameters to be estimated.

Assume: $w$ and $\epsilon^2$ are endowed with the conjugate priors.

$$w|\epsilon^2 \sim N(0_D, \epsilon^2 \lambda^{-1} I_{DD})$$

$$\epsilon^2 \sim IG(\alpha_0, \beta_0)$$

Thus three parameters to be estimated.

$$\lambda, \alpha_0, \beta_0$$

The basic idea is to integrate out the model parameter and results only a marginal posterior with only data and hyperparameters.

$$\hat{\lambda}_{eb} = \arg\max_{\lambda} \int_0^\infty \int_{\mathbb{R}^D} p\{w, \sigma^2 \mid \tilde{t}, \tilde{X}\} \, dw \, d\sigma^2$$

$$= \arg\max_{\lambda} \int_0^\infty \int_{\mathbb{R}^D} \sigma^{-n} \exp\left\{-\tfrac{1}{2}\sigma^{-2}(\tilde{t}-\tilde{X}w)^T(\tilde{t}-\tilde{X}w)\right\}$$

$$\cdot \sigma^{-D} \exp\left\{-\tfrac{1}{2}\sigma^{-2}\lambda w^T w\right\} \cdot (\sigma^2)^{-\alpha_0-1} \exp\{-\beta_0\sigma^{-2}\} \, dw \, d\sigma^2$$

$$= \arg\max_{\lambda} \int_0^\infty \int_{\mathbb{R}^D} \underbrace{\sigma^{-n} \exp\left\{-\tfrac{1}{2}\sigma^{-2}\left[\tilde{t}^T\tilde{t} - \tilde{t}^T\tilde{X}(\tilde{X}^T\tilde{X}+\lambda I_{DD})^{-1}\tilde{X}^T\tilde{t}\right]\right\}}_{\text{invariant to } \beta}$$

$$\underbrace{\cdot \sigma^{-D} \exp\left\{-\tfrac{1}{2}\sigma^{-2}\left[w-w(\lambda)\right]^T(\tilde{X}^T\tilde{X}+\lambda I_{DD})\left[w-w(\lambda)\right]\right\}}_{\text{"unnormalized Gaussian"}}$$

$$\underbrace{\cdot (\sigma^2)^{-\alpha_0-1} \exp\{-\beta_0\sigma^{-2}\}}_{\text{"unnormalized IG"}} \, d\beta \, d\sigma^2$$

$$= \arg\max_{\lambda} \int_0^\infty \sigma^{-n} \exp\left\{-\tfrac{1}{2}\sigma^{-2}\left[\tilde{t}^T\tilde{t} - \tilde{t}^T\tilde{X}(\tilde{X}^T\tilde{X}+\lambda I_{DD})^{-1}\tilde{X}^T\tilde{t}\right]\right\}$$

$$\cdot |\tilde{X}^T\tilde{X}+\lambda I_{DD}|^{-\frac{1}{2}} (\sigma^2)^{-\alpha_0-1} \exp\{-\beta_0\sigma^{-2}\} \, d\sigma^2$$

$$= \arg\max_{\lambda} |\tilde{X}^T\tilde{X}+\lambda I_{DD}|^{-\frac{1}{2}} \int_0^\infty \exp\left\{-\sigma^{-2}\left[\beta_0 + \tfrac{1}{2}\left[\tilde{t}^T\tilde{t} - \tilde{t}^T\tilde{X}(\tilde{X}^T\tilde{X}+\lambda I_{DD})^{-1}\tilde{X}^T\tilde{t}\right]\right]\right\}$$

$$\cdot (\sigma^2)^{-\alpha_0-\frac{1}{2}-1} \, d\sigma^2 \qquad \text{"unnormalized IG"}$$

$$= \arg\max_{\lambda} \; |\tilde{X}^T \tilde{X} + \lambda I_{DD}|^{-\frac{1}{2}} \; b_1^{-\alpha_0 - \frac{n}{2}}$$

where $b_1 = \beta_0 + \frac{1}{2}\left[\tilde{Y}^T \tilde{Y} - \tilde{Y}^T \tilde{X} (\tilde{X}^T \tilde{X} + \lambda I_{DD})^{-1} \tilde{X}^T \tilde{Y}\right]$