Multivariate Normal Distribution

Nipun Batra

August 24, 2023

IIT Gandhinagar

Multivariate Normal Distribution

$$\text{PDF}(\boldsymbol{\theta}, \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

Multivariate Normal Distribution

$$\mathrm{PDF}(\boldsymbol{\theta}, \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

- $\boldsymbol{\theta}$ is the vector of random variables (observation) for which you want to calculate the PDF.

Multivariate Normal Distribution

$$\text{PDF}(\boldsymbol{\theta}, \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

- $\boldsymbol{\theta}$ is the vector of random variables (observation) for which you want to calculate the PDF.
- k is the dimensionality of the random vector $\boldsymbol{\theta}$ (number of variables).

Multivariate Normal Distribution

$$\text{PDF}(\boldsymbol{\theta}, \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

- $\boldsymbol{\theta}$ is the vector of random variables (observation) for which you want to calculate the PDF.
- k is the dimensionality of the random vector $\boldsymbol{\theta}$ (number of variables).
- Σ is the covariance matrix

Multivariate Normal Distribution

$$\text{PDF}(\boldsymbol{\theta}, \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

- $\boldsymbol{\theta}$ is the vector of random variables (observation) for which you want to calculate the PDF.
- k is the dimensionality of the random vector $\boldsymbol{\theta}$ (number of variables).
- Σ is the covariance matrix
- $\boldsymbol{\mu}$ is the mean vector.

Bivariate Normal Distribution

$$\text{PDF}(\boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

Bivariate Normal Distribution

$$\text{PDF}(\boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

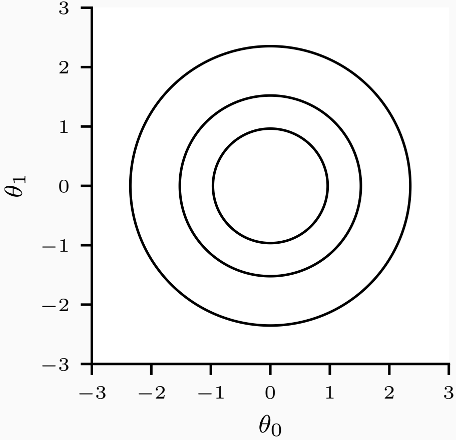
Slides heavily inspired from Richard Turner's slides

Bivariate Normal Distribution

Notebook (visualise-normal.ipynb)

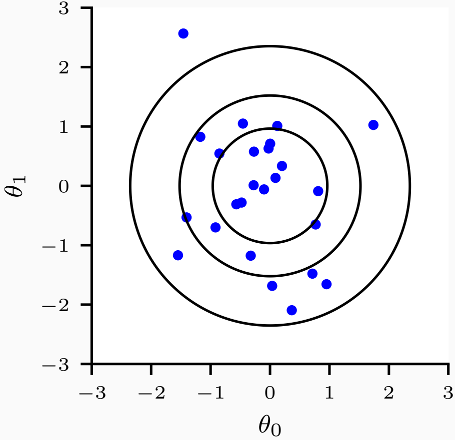
Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta-\mu)^\top \Sigma^{-1}(\theta-\mu)\right) \qquad \Sigma = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$$



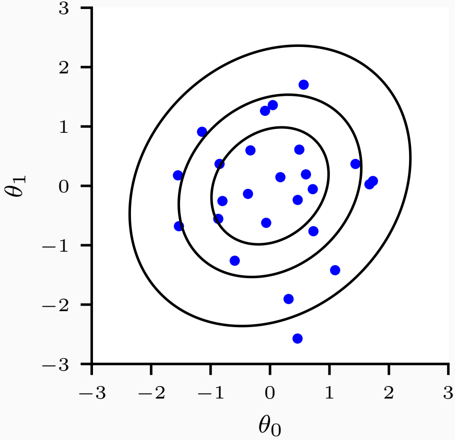
Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta-\mu)^\top \Sigma^{-1}(\theta-\mu)\right) \qquad \Sigma = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$$



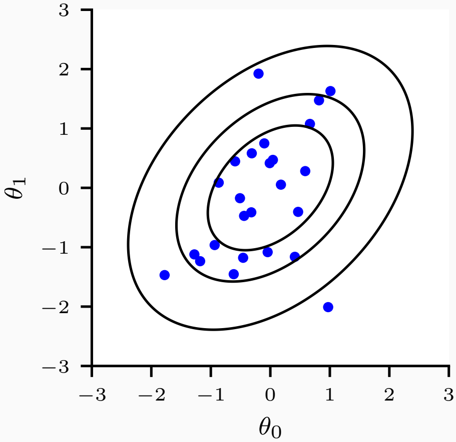
Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right) \qquad \Sigma = \begin{bmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{bmatrix}$$



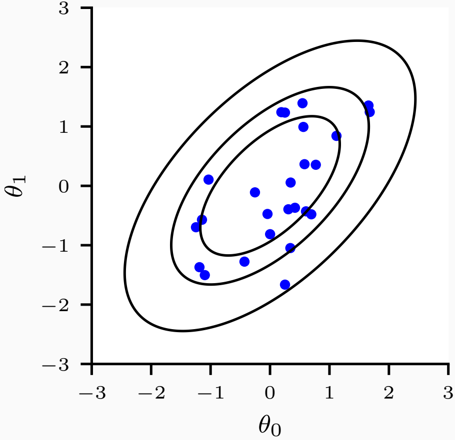
Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta-\mu)^\top \Sigma^{-1}(\theta-\mu)\right) \qquad \Sigma = \begin{bmatrix} 1.0 & 0.4 \\ 0.4 & 1.0 \end{bmatrix}$$



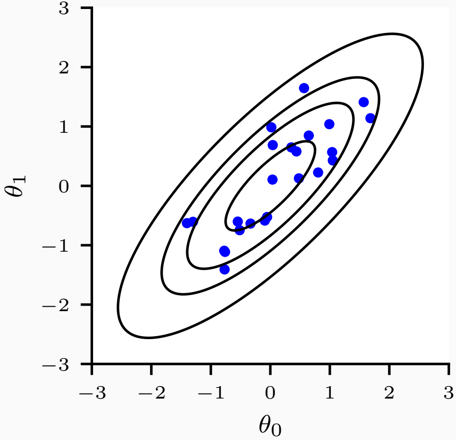
Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right) \qquad \Sigma = \begin{bmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{bmatrix}$$



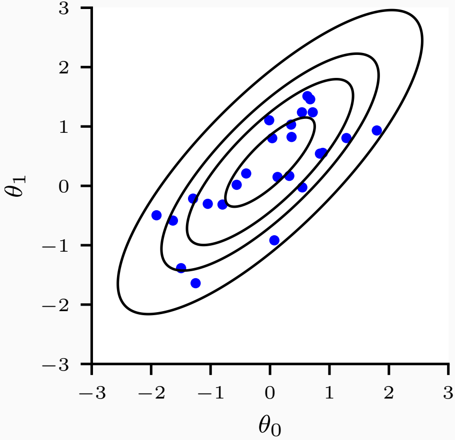
Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right) \qquad \Sigma = \begin{bmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{bmatrix}$$



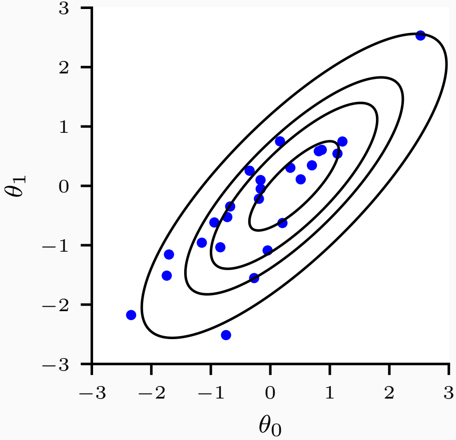
Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right) \qquad \mu = \begin{bmatrix} 0.0 \\ 0.4 \end{bmatrix}$$



Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right) \qquad \mu = \begin{bmatrix} 0.4 \\ 0.0 \end{bmatrix}$$



Bayesian Linear Regression

Nipun Batra
August 24, 2023
IIT Gandhinagar

Bayesian Linear Regression

Linear Regression

$$\boldsymbol{\theta}_{\mathrm{MLE}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Linear Regression

$$\boldsymbol{\theta}_{\mathrm{MLE}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

For θ_{MAP} estimation, we assume a Gaussian prior $p(\boldsymbol{\theta}) = \mathcal{N}\left(0, b^2 \boldsymbol{I}\right)$

Linear Regression

$$\boldsymbol{\theta}_{\mathrm{MLE}} = \left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}$$

For θ_{MAP} estimation, we assume a Gaussian prior $p(\boldsymbol{\theta}) = \mathcal{N}\left(0, b^2\boldsymbol{I}\right)$

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \frac{\sigma^2}{b^2}\boldsymbol{I}\right)^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}$$

Linear Regression

$$\boldsymbol{\theta}_{\mathrm{MLE}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

For θ_{MAP} estimation, we assume a Gaussian prior $p(\boldsymbol{\theta}) = \mathcal{N}\left(0, b^2 \boldsymbol{I}\right)$

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \left(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\sigma^2}{b^2} \boldsymbol{I}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

where $\boldsymbol{X}$ is the feature matrix, $\boldsymbol{y}$ is the corresponding ground truth values and σ is the standard deviation of Gaussian distribution in the MLE estimation.

Linear Regression using Basis Functions

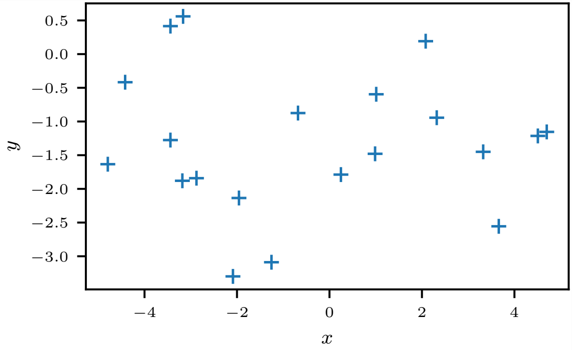


Figure 1: Data

Linear Regression using Basis Functions

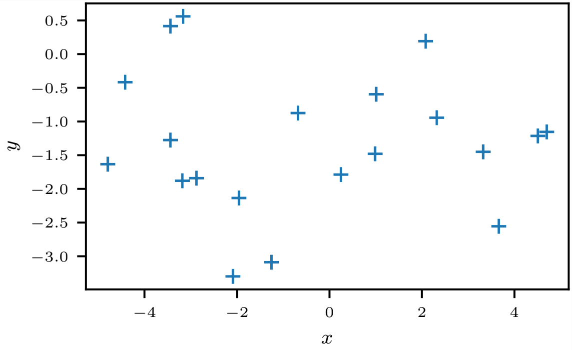


Figure 1: Data

We can use basis functions to fit a non-linear function to the data.

Linear Regression using Basis Functions

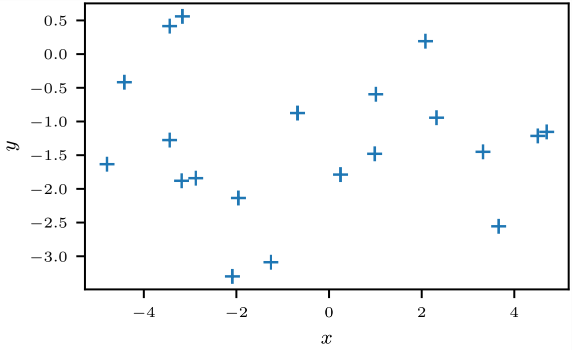


Figure 1: Data

We can use basis functions to fit a non-linear function to the data. For example we can use a polynomial basis function to fit a polynomial to the data, where $\phi_j(x) = x^j$.

MLE and MAP

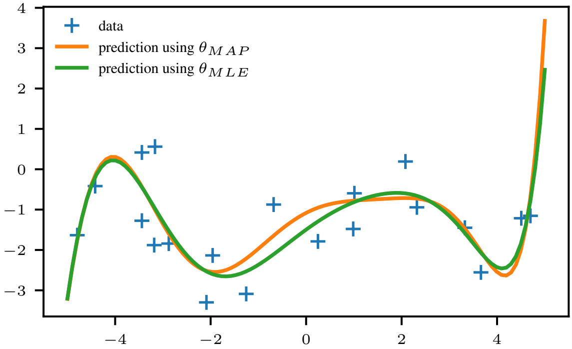


Figure 2: MLE and MAP

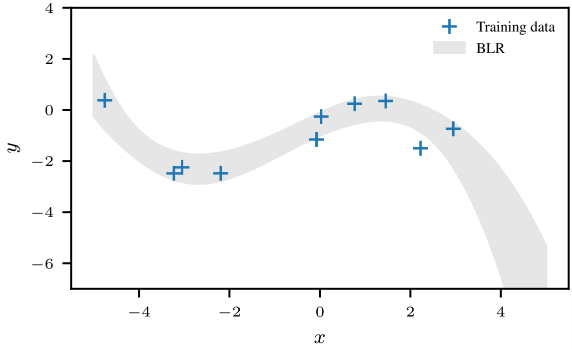


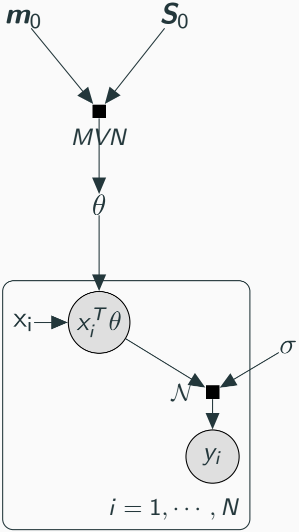
Figure 3: Bayesian linear regression

Bayes Rule

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$$

- $P(\theta|D)$ is called the posterior
- $P(D|\theta)$ is called the likelihood
- $P(\theta)$ is called the prior
- $P(D)$ is called the evidence

Bayesian Linear Regression



Bayesian Linear Regression

In Bayesian linear regression, we consider the model:

$$\text{prior}: \quad p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{m}_0, \boldsymbol{S}_0\right)$$

with $\boldsymbol{m}_0$ and $\boldsymbol{S}_0$ as the mean and covariance matrix and

$$\text{likelihood}: p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}\left(y \mid \boldsymbol{x}^\top \boldsymbol{\theta}, \sigma^2\right)$$

Bayes Rule

Given a training set of inputs $\boldsymbol{x}_n \in \mathbb{R}^D$ and corresponding observations $y_n \in \mathbb{R}$, $n = 1, \ldots, N$, we compute the posterior over the parameters using Bayes' theorem as

$$p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathcal{Y} \mid \mathcal{X})}$$

where $\mathcal{X}$ is the set of training inputs and $\mathcal{Y}$ the collection of corresponding training targets.

Posterior

We find the closed form solution of posterior $p(\boldsymbol{\theta} \mid \mathcal{X}$ to be a normal distribution with mean $\boldsymbol{m}_N$ and covariance matrix $\boldsymbol{S}_N$

$$p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) = \mathcal{N}\left(\boldsymbol{\theta} \mid \boldsymbol{m}_N, \boldsymbol{S}_N\right)$$

$$\boldsymbol{S}_N = \left(\boldsymbol{S}_0^{-1} + \sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}$$

$$\boldsymbol{m}_N = \boldsymbol{S}_N\left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sigma^{-2}\boldsymbol{X}^\top\boldsymbol{y}\right)$$

where the subscript N indicates the size of the training set.

Proof

$$\text{Posterior :} \quad p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathcal{Y} \mid \mathcal{X})}$$

$$\text{Likelihood : } p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}\right)$$

$$\text{Prior : } p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta} \mid \boldsymbol{m}_0, \boldsymbol{S}_0\right)$$

Proof

The sum of the log-prior and the log-likelihood is

$$\log \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}\right) + \log \mathcal{N}\left(\boldsymbol{\theta} \mid \boldsymbol{m}_0, \boldsymbol{S}_0\right)$$

$$= -\frac{1}{2}\Big(\sigma^{-2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{m}_0)^\top \boldsymbol{S}_0^{-1} (\boldsymbol{\theta} - \boldsymbol{m}_0)\Big) + \text{const}$$

We ignore the constant term independent of $\boldsymbol{\theta}$. We now factorize, which yields

We ignore the constant term independent of $\boldsymbol{\theta}$. We now factorize, which yields

$$= -\frac{1}{2}\Big(\sigma^{-2}\boldsymbol{y}^\top\boldsymbol{y} - 2\sigma^{-2}\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\theta}^\top\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\theta}^\top\boldsymbol{S}_0^{-1}\boldsymbol{\theta}$$
$$-2\boldsymbol{m}_0^\top\boldsymbol{S}_0^{-1}\boldsymbol{\theta} + \boldsymbol{m}_0^\top\boldsymbol{S}_0^{-1}\boldsymbol{m}_0\Big)$$

We ignore the constant term independent of $\boldsymbol{\theta}$. We now factorize, which yields

$$= -\frac{1}{2}\Big(\sigma^{-2}\boldsymbol{y}^\top\boldsymbol{y} - 2\sigma^{-2}\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\theta}^\top\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\theta}^\top\boldsymbol{S}_0^{-1}\boldsymbol{\theta}$$
$$-2\boldsymbol{m}_0^\top\boldsymbol{S}_0^{-1}\boldsymbol{\theta} + \boldsymbol{m}_0^\top\boldsymbol{S}_0^{-1}\boldsymbol{m}_0\Big)$$

$$= -\frac{1}{2}\left(\boldsymbol{\theta}^\top\left(\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{S}_0^{-1}\right)\boldsymbol{\theta} - 2\left(\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{S}_0^{-1}\boldsymbol{m}_0\right)^\top\boldsymbol{\theta}\right)$$
$$+ \text{ const}$$

Posterior

Now, we evaluate the posterior distribution,

$$p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) = \exp(\log p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y})) \propto \exp(\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}))$$

$$\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\theta}^\top \left(\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{S}_0^{-1}\right)\boldsymbol{\theta} - 2\left(\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{S}_0^{-1}\boldsymbol{m}_0\right)^\top\boldsymbol{\theta}\right)\right)$$

Normalizing the posterior distribution

We now normalize this Gaussian distribution into the form that is proportional to $\mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{m}_N, \boldsymbol{S}_N)$, i.e., we need to identify the mean $\boldsymbol{m}_N$ and the covariance matrix $\boldsymbol{S}_N$.

Normalizing the posterior distribution

We now normalize this Gaussian distribution into the form that is proportional to $\mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{m}_N, \boldsymbol{S}_N)$, i.e., we need to identify the mean $\boldsymbol{m}_N$ and the covariance matrix $\boldsymbol{S}_N$.

To do this, we use the concept of completing the squares. The desired log posterior is

Normalizing the posterior distribution

We now normalize this Gaussian distribution into the form that is proportional to $\mathcal{N}\left(\boldsymbol{\theta} \mid \boldsymbol{m}_N, \boldsymbol{S}_N\right)$, i.e., we need to identify the mean $\boldsymbol{m}_N$ and the covariance matrix $\boldsymbol{S}_N$.

To do this, we use the concept of completing the squares. The desired log posterior is

$$\begin{aligned}
\log \mathcal{N}\left(\boldsymbol{\theta} \mid \boldsymbol{m}_N, \boldsymbol{S}_N\right) &= -\frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{m}_N\right)^{\top} \boldsymbol{S}_N^{-1}\left(\boldsymbol{\theta}-\boldsymbol{m}_N\right)+\text { const } \\
&= -\frac{1}{2}\left(\boldsymbol{\theta}^{\top} \boldsymbol{S}_N^{-1} \boldsymbol{\theta}-2 \boldsymbol{m}_N^{\top} \boldsymbol{S}_N^{-1} \boldsymbol{\theta}+\boldsymbol{m}_N^{\top} \boldsymbol{S}_N^{-1} \boldsymbol{m}_N\right) .
\end{aligned}$$

Normalizing the posterior distribution

We factorize the quadratic form $(\boldsymbol{\theta} - \boldsymbol{m}_N)^\top \boldsymbol{S}_N^{-1} (\boldsymbol{\theta} - \boldsymbol{m}_N)$ into a term that is quadratic in $\boldsymbol{\theta}$ alone, a term that is linear in $\boldsymbol{\theta}$, and a constant term. This allows us now to find $\boldsymbol{S}_N$ and $\boldsymbol{m}_N$ by matching the expressions, which yields

$$\begin{aligned} \boldsymbol{S}_N^{-1} &= \boldsymbol{X}^\top \sigma^{-2} \boldsymbol{I} \boldsymbol{X} + \boldsymbol{S}_0^{-1} \\ \Longrightarrow \boldsymbol{S}_N &= \left(\sigma^{-2} \boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{S}_0^{-1}\right)^{-1} \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{m}_N^\top \boldsymbol{S}_N^{-1} &= \left(\sigma^{-2} \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{S}_0^{-1} \boldsymbol{m}_0\right)^\top \\ \Longrightarrow \boldsymbol{m}_N &= \boldsymbol{S}_N \left(\sigma^{-2} \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{S}_0^{-1} \boldsymbol{m}_0\right). \end{aligned}$$

Posterior Predictive Distribution

Goal: Find $p\,(y_* \mid \mathcal{X}, \mathcal{Y}, \boldsymbol{x}_*)$

$$
\begin{aligned}
p(y_* \mid \mathcal{X}, \mathcal{Y}, \boldsymbol{x}_*) &= \int p(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) \mathrm{d}\boldsymbol{\theta} \\
&= \int \mathcal{N}(y_* \mid \boldsymbol{x}_*^\top \boldsymbol{\theta}, \sigma^2) \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{m}_N, \boldsymbol{S}_N) \mathrm{d}\boldsymbol{\theta} \\
&= \mathcal{N}(y_* \mid \boldsymbol{x}_*^\top \boldsymbol{m}_N, \boldsymbol{x}_*^\top \boldsymbol{S}_N \boldsymbol{x}_* + \sigma^2)
\end{aligned}
$$

Posterior Predictive Distribution

Goal: Find $p\left(y_* \mid \mathcal{X}, \mathcal{Y}, \boldsymbol{x}_*\right)$

$$\begin{aligned}
p(y_* \mid \mathcal{X}, \mathcal{Y}, \boldsymbol{x}_*) &= \int p(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) \mathrm{d}\boldsymbol{\theta} \\
&= \int \mathcal{N}(y_* \mid \boldsymbol{x}_*^\top \boldsymbol{\theta}, \sigma^2) \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{m}_N, \boldsymbol{S}_N) \mathrm{d}\boldsymbol{\theta} \\
&= \mathcal{N}(y_* \mid \boldsymbol{x}_*^\top \boldsymbol{m}_N, \boldsymbol{x}_*^\top \boldsymbol{S}_N \boldsymbol{x}_* + \sigma^2)
\end{aligned}$$

Two kinds of uncertainty:

- **Aleatoric uncertainty**: Uncertainty in the data - given as σ^2
- **Epistemic uncertainty**: Uncertainty in the model - given as $\boldsymbol{x}_*^\top \boldsymbol{S}_N \boldsymbol{x}_*$

Posterior Predictive Distribution

- TFP blog: Aleatoric v/s Epistemic Uncertainty
- MML book: Figure 9.4

Bayesian Updation

Bishop book: Figure 3.7