Chapter 3: Maximum Likelihood & Bayesian Parameter Estimation

3.1 introduction

For Bayes Formula

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

To compute **posterior probability** $P(w_j|x)$, we need to know **Prior probability** $P(w_j)$ and **Likelihood** $p(x|w_j)$.

For prior probability, there is no difficulty in computing: $P(w_j) = \frac{|\mathcal{D}_j|}{\sum_{i=1}^{c}}|\mathcal{D}_i|$

However, class-conditional probability density function $p(x|w_j)$ is hard to assess. The main problems are listed below:

- Training samples always seems to be numerically limited.
- When the dimensionality of the feature space is high, here comes the problem of computaion complexity.

To solve the problem, we assume $p(x|w_j)$ has a certain **parametric form**.

- Assumption 1: **Maximum-Likelihood(ML) Estimation**
  - View parameters as quantities whose value are **fixed but unknown**

- Estimate parameter values by **maximizing the likelihood** (probability) of observing the actual training examples.
- Assumption 2: **Bayesian Estimation**
  - View parameters as **random variables** having some known prior distribution.
  - Observation of the actual training examples transforms parameters' **prior distribution into a posterior distribution** (via Bayes theorem).

3.2 Maximum-Likelihood Estimation

3.2.1 Basic principle

Settings:

- Likelihood function for each category is governed by some **fixed but unknown** parameters, i.e. $p(x|w_j, \theta_j)(1 \leq j \leq c)$

Task:

- Estimate the parameters $\{\theta_j\}_{j=1}^{c}$ form the training data $\{\mathcal{D}_j\}_{j=1}^{c}$

Assuming that Examples in $\mathcal{D}_j$ give no information about the parameters θ_j if $i \neq j$, each category can work separately hence we can simplify the notations $\mathcal{D}_j \Rightarrow \mathcal{D}; \theta_j \Rightarrow \theta$ without loss of generality.

Given that:

$$x_k \sim p(x|\theta)$$

- θ: Parameters to be estimated
- $\mathcal{D}$: A set of i.i.d. (independent and identically distributed) examples $\{x_1, x_2, ..., x_n\}$

Here we have the **objective function**, i.e. **the likelihood of θ w.r.t. (with respect to) the set of observed examples**:

$$p(\mathcal{D}|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$$

The maximum-likelohood estimation:

$$\hat{\theta} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

- Intuitively, $\hat{\theta}$ best agrees with the actually observed examples.

Assuming the number of unknown parameters is p, let $\theta = (\theta_1, \dots, \theta_p) \in \mathbf{R}^d$ be a d-dimensional vector.

Note ∇_θ as the gradient operator:

$$\nabla_\theta = \left(\frac{\partial}{\partial\theta_1}, \dots, \frac{\partial}{\partial\theta_p}\right)^T$$

Define the **log-likelihood function** $l(\theta) = \ln p(\mathcal{D}|\theta)$:

$$\hat{\theta} = \arg\max_{\theta} p(\mathcal{D}|\theta) \Leftrightarrow \hat{\theta} = \arg\max_{\theta} l(\theta)$$

Hence we have:

$$\nabla_\theta l = \nabla_\theta \Big(\sum_{k=1}^{n} \ln p(x_k|\theta) \Big) = \sum_{k=1}^{n} \nabla_\theta \ln p(x_k|\theta)$$

Necessary conditions for ML estimate $\hat{\theta}$:

$$\nabla_\theta l|_{\theta=\hat{\theta}} = 0 \qquad \text{(a set of } p \text{ eqautions)}$$

3.2.2 The Gaussian Case: unknown μ

$$x_k \sim N(\mu, \Sigma) \quad (k = 1, \cdots, n)$$

- suppose Σ is known
- $\theta = \{\mu\}$

Here we have:

$$\begin{aligned}
p(x_k|\mu) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x_k-\mu)^t\Sigma^{-1}(x_k-\mu)} \\
&\Downarrow \\
\ln p(x_k|\mu) &= -\frac{1}{2}(x_k-\mu)^t\Sigma^{-1}(x_k-\mu) - \frac{1}{2}ln[(2\pi)^d|\Sigma|] \\
&= -\frac{1}{2}ln\big[(2\pi)^d|\Sigma|\big] - \frac{1}{2}x_k^t\Sigma^{-1}x_k + \mu^t\Sigma^{-1}x_k - \frac{1}{2}\mu^t\Sigma^{-1}\mu \\
&\Downarrow \\
\nabla_\mu \ln p(x_k|\mu) &= \Sigma^{-1}(x_k-\mu)
\end{aligned}$$

To get the maximum-likelihood estimate $\hat{\mu}$, let

$$\nabla_u l(\mu) = \sum_{k=1}^{n} \nabla_\mu \ln p(x_k|\mu) = \sum_{k=1}^{n} \Sigma^{-1}(x_k - \hat{\mu}) = 0$$

hence we have:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

- **Intuitive result**: ML estimate for the unknown μ is just the artithmetic average of training samples - **sample mean**

3.2.3 The Gaussian Case: Unknown μ and Σ

$$x_k \sim N(\mu, \Sigma) \quad (k = 1, \cdots, n)$$

- μ and Σ are unknown $\Rightarrow \theta = \{\mu, \Sigma\}$

3.2.3.1 Univariate case

$$\begin{aligned}
p(x_k|\theta) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \qquad \Big(\theta = [\theta_1, \theta_2]^T = [\mu, \sigma^2]^T\Big) \\
&\Downarrow \\
\ln p(x_k|\theta) &= -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k-\theta_1)^2 \\
&\Downarrow \\
\nabla_\theta \ln p(x_k|\theta) &= \Big[\frac{x_k-\theta_1}{\theta_2}, -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2}\Big]^T
\end{aligned}$$

To get the maximum-likelihood estimate $\hat{\theta}$, let:

$$\nabla_\theta l(\theta) = \sum_{k=1}^{n} \nabla_\theta \ln p(x_k|\theta) = \Big[\sum_{k=1}^{n} \frac{x_k - \theta_1}{\theta_2}, \sum_{k=1}^{n}(-\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2})\Big]^T = 0$$

hence we have:

$$\hat{\theta}_1 = \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$
$$\hat{\theta}_2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$

3.2.3.2 Multivariate case

$$p(x_k|\theta) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x_k-\mu)^t\Sigma^{-1}(x_k-\mu)}$$

Here we have:

$$\begin{aligned}\ln p(x_k|\theta) &= -\frac{1}{2}(x_k-\mu)^t\Sigma^{-1}(x_k-\mu) - \frac{1}{2}ln[(2\pi)^d|\Sigma|] \\ &= -\frac{1}{2}ln\big[(2\pi)^d|\Sigma|\big] - \frac{1}{2}x_k^t\Sigma^{-1}x_k + \mu^t\Sigma^{-1}x_k - \frac{1}{2}\mu^t\Sigma^{-1}\mu\end{aligned}$$

- **Properties of Matrix Differentiation**
  - Addition and subtraction: $d(X \pm Y) = dX \pm dY$
  - Matrix multiplication: $d(XY) = YdX + XdY$
  - Transpose: $d(X^T) = (dX)^T$
  - Trace: $dtr(X) = tr(dX)$
  - Inverse: $dX^{-1} = -X^{-1}dXX^{-1}$
    - Use $XX^{-1} = I$ to prove and $dI = 0$
  - Determinant: $d|X| = tr(X^*dX)$, when Matrix X is invertible, $d|X| = |X|tr(X^{-1}dX)$
- **Rules of matrix calculus**
  - $\frac{\partial X^TAX}{\partial X} = (A + A^T)X$, when A is a real symmetric matrix matrix, $\frac{\partial X^TAX}{\partial X} = 2AX$
  - When A is a real symmetric matrix, $\frac{\partial X^TAX}{\partial A} = XX^T$, $\frac{\partial |A|}{\partial A} = A^{-1}|A|$, $\frac{\partial \ln|A|}{\partial A} = A^{-1}$
  - $\frac{\partial (X^{-1})}{\partial t} = -X^{-1}\frac{\partial X}{\partial t}X^{-1}$

According to the matrix differentiation rules, we have:

$$\begin{aligned}\nabla_\theta \ln p(x_k|\theta) &= \Big[\nabla_\mu \ln p(x_k|\mu), \nabla_\Sigma \ln p(x_k|\mu)\Big]^T \\ &= \Big[\Sigma^{-1}x_k - \Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}x_kx_k^t\Sigma^{-1} - \Sigma^{-1}x_k\mu^t\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\mu\mu^t\Sigma^{-1}\Big]^T \\ &= \Big[\Sigma^{-1}x_k - \Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(x_k-\mu)(x_k-\mu)^t\Sigma^{-1}\Big]^T\end{aligned}$$

To get the maximum-likelihood estimate $\hat{\theta}$, let:

$$\nabla_\theta l(\theta) = \sum_{k=1}^{n} \nabla_\theta \ln p(x_k|\theta) = \Big[\sum_{k=1}^{n} \Sigma^{-1}x_k - \Sigma^{-1}\hat{\mu}, \sum_{k=1}^{n}\big(-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(x_k-\hat{\mu})(x_k-\hat{\mu})^t\Sigma^{-1}\big)\Big]^T = 0$$

hence we have:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$
$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})(x_k - \hat{\mu})^t$$

Intuitive result:

- $\hat{\mu}$ is the artithmetic average of **n vector** x_k
- $\hat{\Sigma}$ is the artithmetic average of **n matrix** $(x_k - \hat{\mu})(x_k - \hat{\mu})^t$

3.2.4 Biased/Unbiased Estimator

Biased estimator of Σ, whose artithmetic expectation does not equal to the real variance:

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})(x_k - \hat{\mu})^t$$
$$\mathcal{E}\left[\frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})(x_k - \hat{\mu})^t\right] = \frac{n-1}{n}\Sigma \neq \Sigma$$

Unbiased estimator of Σ:

$$\mathcal{C} = \frac{n-1}{n}\hat{\Sigma}$$

When $n \to \infty$, the estimator converges to the real Unbiased estimator. We call it **Asyptotically unbiased estimator**:

$$\lim_{n\to\infty} \mathcal{E}(\hat{\Sigma}) = \Sigma$$

3.3 Bayesian Estimation

Settings:

- The **parametric form** of the likelihood function for each category is known $p(x|w_j, \theta_j) \quad (1 \leq j \leq c)$
- θ_j is considered to be random variables instead of being fixed (but unknown) values

In this case, we can no longer make a single ML estimate $\hat{\theta}_j$. Instead, we fully exploit training samples to make the estimation.

$$P(w_j|x) \Rightarrow P(w_j|x, \mathcal{D}^*) \qquad (\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_c)$$

3.3.1 class-conditional probability density function

Two assumptions:

1. $p(w_j|\mathcal{D}^*) = p(w_j)$
2. $p(x|w_j, \mathcal{D}^*) = p(x|w_j, \mathcal{D}_j)$: training samples in $\mathcal{D}_i$ have no impact on calculating $p(x|w_j, \mathcal{D}_j)$ if $i \neq j$

$$
\begin{aligned}
p(w_j|x,\mathcal{D}^*) &= \frac{p(w_j,x,\mathcal{D}^*)}{p(x,\mathcal{D}^*)} = \frac{p(w_j,x,\mathcal{D}^*)}{\sum_{i=1}^{c} p(w_i,x,\mathcal{D}^*)} \\
&= \frac{p(\mathcal{D}^*)\cdot p(w_j|\mathcal{D}^*)\cdot p(x|w_j,\mathcal{D}^*)}{p(\mathcal{D}^*)\cdot \sum_{i=1}^{c} p(w_i|\mathcal{D}^*)\cdot p(x|w_i,\mathcal{D}^*)} \\
&= \frac{p(w_j|\mathcal{D}^*)\cdot p(x|w_j,\mathcal{D}^*)}{\sum_{i=1}^{c} p(w_i|\mathcal{D}^*)\cdot p(x|w_i,\mathcal{D}^*)} \\
&= \frac{p(w_j)\cdot p(x|w_j,\mathcal{D}_j)}{\sum_{i=1}^{c} p(w_i)\cdot p(x|w_i,\mathcal{D}_i)}
\end{aligned}
$$

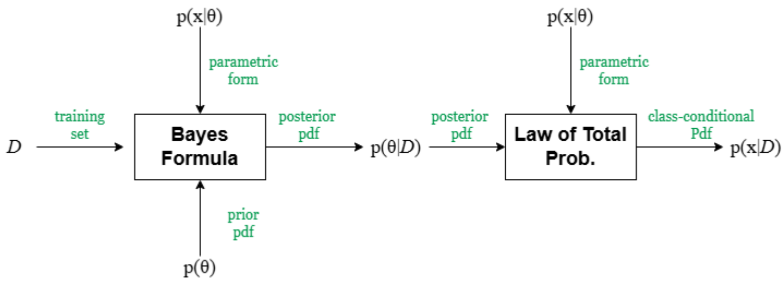
According to the above formula, we find that the **key problem** is to determine the **prior probability** $p(x|w_j,\mathcal{D}_j)$

We treat each class independently to simplify the class-conditional pdf notation $p(x|w_j,\mathcal{D}_j)$ as $p(x|\mathcal{D})$

Given the assumption that $p(x|\mathcal{D})$ is a distribution that relies on the parameters θ and θ itself is a random variable with prior probability $p(\theta)$, we can take the uncertainty of parameters into consideration through integration:

$$
\begin{aligned}
p(x|\mathcal{D}) &= \int p(x,\theta|\mathcal{D})d\theta \qquad (\theta : \text{random variable w.r.t. parametric form}) \\
&= \int p(x|\theta,D)p(\theta|\mathcal{D})d\theta \\
&= \int p(x|\theta)p(\theta|\mathcal{D})d\theta \qquad (x \text{ is independent of } \mathcal{D} \text{ given } \theta)
\end{aligned}
$$

3.3.2 General Procedure of Bayesian Estimation



- **Phase 1**: prior pdf $\rightarrow$ posterior pdf (for θ)

$$
\begin{aligned}
p(\theta|\mathcal{D}^*) &= \frac{p(\theta,\mathcal{D})}{p(\mathcal{D})} \\
&= \frac{p(\theta)\cdot p(\mathcal{D}|\theta)}{\int p(\theta,\mathcal{D})d\theta} \\
&= \frac{p(\theta)\cdot p(\mathcal{D}|\theta)}{\int p(\theta)\cdot p(\mathcal{D}|\theta)d\theta} \\
p(D|\theta) &= \prod_{k=1}^{n} p(x_k|\theta)
\end{aligned}
$$

- **Phase 2**: posterior pdf (for θ) $\rightarrow$ class-conditional pdf (for x)

$$
p(x|\mathcal{D}) = \int p(\theta|\mathcal{D})\cdot p(x|\theta)d\theta
$$

- **Phase 3**: $P(w_j|x,\mathcal{D}^*) = \frac{P(w_j)\cdot p(x|w_j,\mathcal{D}_j)}{\sum_{i=1}^{c} P(w_i)\cdot p(x|w_i,\mathcal{D}_i)}$

3.3.3 The Gaussian Case

Use the **Bayesian estimation** to calculate the posterior pdf $P(\theta|\mathcal{D})$ of θ and then design $p(x|\mathcal{D})$ for classication design. In this case, we assume: $p(x|\mu) \sim N(\mu, \Sigma)$

3.3.3.1 Univariate case: Unknown μ

In this case: $p(x|\mu) \sim N(\mu, \sigma^2)$ and $\theta = \mu$ (σ^2 is known)

- **Phase 1**: prior pdf $\rightarrow$ posterior pdf (for θ)
  - We assume the prior pdf $p(\mu) \sim N(\mu_0, \sigma_0^2)$ (μ_0, σ_0^2 are known)
    - **Notice**: the key assumption is to assume the unknown parameters **follow one specific distribution** instead of the specific form of normal distribution
  - According to the fact that $p(\mu|\mathcal{D})$ is **an exponential function** of a quadratic function of μ, $p(\mu|\mathcal{D})$ is **a normal pdf** as well: $p(\mu|\mathcal{D}) \sim N(\mu_n, \sigma_n^2)$. We can equate the coefficients in both form ($\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^n x_k$):
    - $\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \Rightarrow \sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}$
      - σ_n^2 represents the **uncertainty** of the estimation. When $n \rightarrow \infty$, $\sigma_n \rightarrow \frac{\sigma^2}{n} \rightarrow 0$, which means the decrease of the uncertainty.
    - $\frac{\mu_n}{\sigma_n^2} = \frac{1}{\sigma^2}\sum_{k=1}^n x_k + \frac{\mu_0}{\sigma_0^2} \Rightarrow \mu_n = \frac{\sigma_n^2}{\sigma^2}\sum_{k=1}^n x_k + \frac{\sigma_n^2}{\sigma_0^2}\mu_0 = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$
      - μ_n represents the **best estimation** of the unknown parameter μ after observing n training samples.
      - according to the formula, we can see that μ_n is **the linear combination** of the sample mean value $\hat{\mu}_n$ and the prior estimation μ_0.
      - when $n \rightarrow \infty$, $\mu_n \rightarrow \hat{\mu}_n$, which means the estimation approaches the sample mean value

$$
\begin{aligned}
p(\mu|\mathcal{D}) &= \frac{p(\mu,\mathcal{D})}{p(\mathcal{D})} = \frac{p(\mu)p(\mathcal{D}|\mu)}{\int p(\mu)p(\mathcal{D}|\mu)du} \\
&= \alpha p(\mu)p(\mathcal{D}|\mu) \qquad (\int p(\mu)p(\mathcal{D}|\mu)du \text{ is a constant}) \\
&= \alpha p(\mu)\prod_{k=1}^n p(x_k|\mu) \qquad (\text{examples in } \mathcal{D} \text{ are i.i.d.}) \\
&= \alpha \cdot \left[\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{(x_k-\mu_0)^2}{2\sigma_0^2}}\right] \cdot \left[\prod_{k=1}^n \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_k-\mu)^2}{2\sigma^2}}\right] \\
&= \alpha^{'} \cdot e^{-\frac{1}{2}\left(\left(\frac{\mu-\mu_0}{\sigma_0^2}\right)^2 + \sum_{k=1}^n\left(\frac{u-x_k}{\sigma}\right)^2\right)} \\
&= \alpha^{''} \cdot e^{-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^n x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]} \\
p(\mu|\mathcal{D}) &= \frac{1}{\sqrt{2\pi}\sigma_n}e^{-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2} = \alpha^{''} \cdot e^{-\frac{1}{2}\left(\frac{1}{\sigma_n^2}\mu^2 - 2\frac{\mu_n}{\sigma_n^2}\mu\right)}
\end{aligned}
$$

- **Phase 2**: posterior pdf (for θ) $\rightarrow$ class-conditional pdf (for x)
  - According to the fact that $p(x|\mathcal{D})$ is **an exponential function** of a quadratic function of x, $p(x|\mathcal{D})$ is **a normal pdf** as well: $p(x|\mathcal{D}) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$

$$
\begin{aligned}
p(x|\mathcal{D}) &= \int p(\mu|\mathcal{D}) \cdot p(x|\mu) d\mu \\
&= \int \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} du \\
&= \int \frac{1}{2\pi\sigma\sigma_n} e^{-\frac{1}{2}\left[\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_n^2}\right)\mu^2 - 2\left(\frac{x}{\sigma^2}+\frac{\mu_n}{\sigma_n^2}\right)\mu + (\frac{x^2}{\sigma^2}+\frac{\mu_n^2}{\sigma_n^2})\right]} du \\
&= \frac{1}{2\pi\sigma\sigma_n} e^{\left[-\frac{1}{2}\cdot\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right]} \cdot \int e^{\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu - \frac{\sigma_n^2 x + \sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right]} du \\
&= \frac{1}{2\pi\sigma\sigma_n} e^{\left[-\frac{1}{2}\cdot\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right]} \cdot f(\sigma, \sigma_n)
\end{aligned}
$$

- **Phase 3**: $P(w_j|x, \mathcal{D}^*) = \frac{P(w_j)\cdot p(x|w_j,\mathcal{D}_j)}{\sum_{i=1}^c P(w_i)\cdot p(x|w_i,\mathcal{D}_i)}$

3.3.3.2 Multivariate case: Unknown μ

- **Setting**

$$\theta = \mu \quad (\Sigma \text{ is known})$$

- **Assumptions**

$$
\begin{gathered}
p(x|\mathcal{D}) \sim N(\mu, \Sigma) \\
p(\mu) \sim N(\mu_0, \Sigma_0)
\end{gathered}
$$

- **Results** ($\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^n x_k$)

$$
\begin{aligned}
&p(\mu|\mathcal{D}) \sim N(\mu_n, \Sigma_n) \qquad p(x|\mathcal{D}) \sim N(\mu_n, \Sigma + \Sigma_n) \\
&\mu_n = \Sigma_0\big(\Sigma_0 + \frac{1}{n}\Sigma\big)^{-1}\hat{\mu}_n + \frac{1}{n}\Sigma\big(\Sigma_0 + \frac{1}{n}\Sigma\big)^{-1}\mu_0 \\
&\Sigma_n = \Sigma_0\big(\Sigma_0 + \frac{1}{n}\Sigma\big)^{-1}\frac{1}{n}\Sigma
\end{aligned}
$$

3.3 Maximum-Likelihood Estimation vs Bayesian Estimation

Infinite examples	ML estimation **=** Bayesian estimation
Complexity	ML estimation **<** Bayesian estimation
Interpretability	ML estimation **>** Bayesian estimation
Prior Knowledge	ML estimation **<** Bayesian estimation

3.4 Error

Bayes error + Model error + Estimation error