Density Estimation in Regression: Parametric Models

Maximum Likelihood Method
Efficient Estimators
Bayesian Learning (batch/online)

October 2, 2019

Agenda

- Motivational problems.
- Bayesianism.
- Frequentism.
  - Maximum-likelihood estimators (MLE).
  - Properties of MLEs.
- A comparison of Bayesianism and frequentism.

Motivational problem I

- Alan wants to estimate the distribution of shoe sizes on his town.
- He assumes they are distributed as $\mathcal{N}(\theta_0, 1)$, for some $\theta_0 \in \mathbb{R}$.
- He asks 10 neighbors for their shoe sizes: $y_1, \ldots, y_{10}$.
- How can Alan estimate θ_0?

Motivational problem I (Mean estimation for a Gaussian)

Given a sample $y_1, \ldots, y_{10}$ from $\mathcal{N}(\theta_0, 1)$, compute an estimate $\hat{\theta}$ for θ_0.

Motivational problem II

- Bob is taking AML.
- Bob wants to understand how many hours per day he may spend studying, doing sports, on social media, and sleeping in order to pass.
- He asks among colleagues who took AML last year. For each colleague $i \leq 10$, he asks them for their grade ($y_i \in \mathbb{R}$) and for these features ($x_i \in \mathbb{R}^d$).
- He wants to estimate $\mathbf{w} \in \mathbb{R}^d$ such that $\mathbf{w}^\top x_i \approx y_i$, for $i \leq 10$.

Motivational problem II (Linear regression)

Assume given $\mathbf{x}, \mathbf{y}, \xi$ with

- $\mathbf{x} \sim ?$,
- $\xi \sim \mathcal{N}(0, \epsilon I)$, and
- $\mathbf{y} = \mathbf{w}^\top \mathbf{x} + \xi$.

Given a sample $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^{d+1}$, how do we estimate $\mathbf{w}$?

Notice that $\mathbf{y} \sim \mathcal{N}\left(\mathbf{w}^\top \mathbf{x}, \epsilon I\right)$.

$\mathbf{x}$ does not depend on $\mathbf{w}$!

Why such elementary problems? This is advanced machine learning!

- Our goal today is not to derive exotic models, but to understand Bayesianism, frequentism, and their differences.
- The ideas here still also apply to current and more sophisticated problems.

Recap: Modelling assumptions for regression

Object Space: $\mathcal{O}$, measurement/feature space: $\mathcal{F} = \mathbb{R}^d \times \mathbb{R}$

Data: $\mathcal{Z} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : 1 \leq i \leq n\}$

Model: Y output, $X = (X_0, X_1, ..., X_d)$ features with $X_0 = 1$, ϵ noise, $\mathbb{E}(\epsilon) = 0$,

$$\mathbf{Y} = f(\mathbf{X}, \boldsymbol{\theta}) + \epsilon$$

Bayesian view: Both the observations (feature vector $\mathbf{X}$ and output variable $\mathbf{Y}$) and the parameters $\boldsymbol{\theta}$ of the regression model are random variables!

Parametric Statistics: the functional form of the likelihood $\mathbf{P}(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})$ is given; we want to estimate the parameters $\boldsymbol{\theta}$ of the likelihood $\mathbf{P}(\text{data}|\text{model})$.

Non-Parametric Statistics: we sample $\mathbf{X}, \mathbf{Y}$ to estimate the likelihood.

Statistical Learning Theory: we minimize the empirical risk $\min_{f \in \mathcal{C}} \hat{R}(f, \mathcal{Z}^{\text{train}})$ directly without estimating the likelihood.

Thomas Bayes and his Terminology

The model is also a random variable!

prior: $\mathbf{P}(\text{model})$

likelihood: $\mathbf{P}(\text{data}|\text{model})$

posterior: $\mathbf{P}(\text{model}|\text{data})$

evidence: $\mathbf{P}(\text{data})$

Bayes Rule $$\mathbf{P}(\text{model}|\text{data}) = \frac{\mathbf{P}(\text{data}|\text{model})\ \mathbf{P}(\text{model})}{\mathbf{P}(\text{data})}$$

Bayesianism and Alan's shoes

- Alan defines a prior $\mathbf{P}(\theta)$ on the mean shoe size θ.

$$\mathbf{P}(\theta) = \mathcal{N}(42, 1)\,.$$

- Alan defines the likelihood $\mathbf{P}(y_1, \ldots, y_n \mid \theta)$.

$$\mathbf{P}(y_1, \ldots, y_n \mid \theta) = \prod_{i \leq n} \mathbf{P}(y_i \mid \theta) = \prod_{i \leq n} \mathcal{N}(y_i \mid \theta, 1)\,.$$

- Alan then computes the posterior $\mathbf{P}(\theta \mid y_1, \ldots, y_n)$.

$$\mathbf{P}(\theta \mid y_1, \ldots, y_n) = \mathcal{N}\left(\frac{n}{n+1}\bar{Y} + \frac{1}{n+1}42, \frac{1}{n+1}\right),$$

where $\bar{Y} = \frac{1}{n}\sum_{i \leq n} y_i$.

The posterior's mean is a weighted sum of the sample mean and the prior's mean!

Bayesianism and Alan's shoes

Is it OK to allow Alan to define a prior that interferes with his estimations?

- If he only asks a few neighbors, the prior gives robustness to the estimation in case he gets an unusual sample $y_1, \ldots, y_n$. The prior acts then as a "regularization term" that avoids overfitting the sample.
- He may also use "flat" priors, if he wants to avoid interfering with the sample's observations.

But we are in the era of big data! Why don't we just get more data?

- You don't need priors if you have enough data and are estimating simple models, but what if you are training more sophisticated models?
- Even big data may be biased... (predicting recidivism).

Bayesianism and Bob's class

- Bob defines a prior $\mathbf{P}(\mathbf{w})$ on the weights of the linear regression model. $\mathbf{P}(\mathbf{w}) = \mathcal{N}\left(0, \alpha^{-1} I\right)$.
- Bob defines the likelihood:

$$\mathbf{P}((x_1, y_1), \ldots, (x_n, y_n) \mid \mathbf{w}) = \prod_{i \leq n} \mathbf{P}((x_i, y_i) \mid \mathbf{w}) = \prod_{i \leq n} \mathbf{P}(y_i \mid x_i, \mathbf{w}) \mathbf{P}(x_i \mid \mathbf{w})$$
$$= \prod_{i \leq n} \mathcal{N}\left(y_i \mid \mathbf{w}^\top x_i, \epsilon I\right) \mathbf{P}(x_i).$$

- Bob then computes the posterior: $\mathbf{P}(\mathbf{w} \mid (x_1, y_1), \ldots, (x_n, y_n)) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}, \mathbf{S})$, where $\mathbf{m} = \ldots$ and $\mathbf{S} = \ldots$
- One can show that

$$\log \mathbf{P}(\mathbf{w} \mid (x_1, y_1), \ldots, (x_n, y_n)) = -\frac{1}{2\epsilon} \sum_{i \leq n} \left(y_i - \mathbf{w}^\top x_i\right)^2 - \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} + \text{const.}$$

In Bayesian linear regression, you are doing least-squares regression with a regularization term!

Bayesianism (Encyclopedia Britannica)

Bayesian methods

The methods of statistical inference previously described are often referred to as classical methods. Bayesian methods (so called after the English mathematician Thomas Bayes) provide alternatives that allow one to combine prior information about a population parameter with information contained in a sample to guide the statistical inference process. A prior probability distribution for a parameter of interest is specified first. Sample information is then obtained and combined through an application of Bayes's theorem to provide a posterior probability distribution for the parameter. The posterior distribution provides the basis for statistical inferences concerning the parameter.

A key, and somewhat controversial, feature of Bayesian methods is the notion of a probability distribution for a population parameter. According to classical statistics, parameters are constants and cannot be represented as random variables. Bayesian proponents argue that, if a parameter value is unknown, then it makes sense to specify a probability distribution that describes the possible values for the parameter as well as their likelihood. The Bayesian approach permits the use of objective data or subjective opinion in specifying a prior distribution. With the Bayesian approach, different individuals might specify different prior distributions. Classical statisticians argue that for this reason Bayesian methods suffer from a lack of objectivity. Bayesian proponents argue that the classical methods of statistical inference have built-in subjectivity (through the choice of a sampling plan) and that the advantage of the Bayesian approach is that the subjectivity is made explicit.

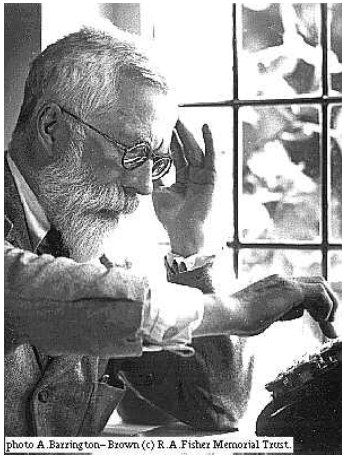
Bayesian methods have been used extensively in statistical decision theory (see below Decision analysis). In this context, Bayes's theorem provides a mechanism for combining a prior probability distribution for the states of nature with sample information to provide a revised (posterior) probability distribution about the states of nature. These posterior probabilities are then used to make better decisions.

Bayes's theorem

Bayes's theorem is in probability theory, a means for revising predictions in light of relevant evidence, also known as conditional probability or inverse probability. The theorem was discovered among the papers of the English Presbyterian minister and mathematician Thomas Bayes and published posthumously in 1763. Related to the theorem is Bayesian inference, or Bayesianism, based on the assignment of some a priori distribution of a parameter under investigation. In 1854 the English logician George Boole criticized the subjective character of such assignments, and Bayesianism declined in favour of confidence intervals and hypothesis tests, now basic research methods.

Ronald A. Fisher and Frequentism

Fisher, Ronald Aylmer (1890-1962): founder of frequentist statistics together with Jerzey Neyman & Karl Pearson.

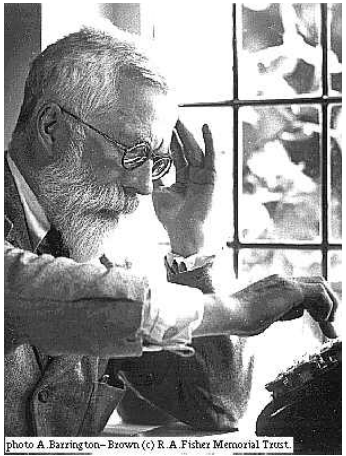


British mathematician and biologist who invented revolutionary techniques for applying statistics to natural sciences.

Maximum likelihood method

Ronald A. Fisher and Frequentism

Fisher, Ronald Aylmer (1890-1962): founder of frequentist statistics together with Jerzey Neyman & Karl Pearson.



Maximum likelihood method

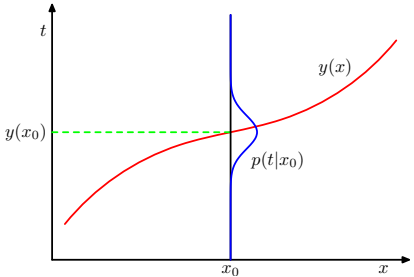
1. Define a parametric model.
2. Define the likelihood as a function of the parametric model.
3. Compute an estimator by maximizing the likelihood.

Regression problem - revisited

Regression problem

- Estimate the conditional probability $\mathbf{P}(Y|X)$
- Calculate the regression function $y(x)$ at $X = x_0$

$$y(x_0) = \int y\, \mathbf{P}(y|X = x_0) dy$$



Bishop: *Patt. Rec. & Mach. Learning*, (2006), fig 1.28

What is the challenge?

- We usually observe only one sample for a given x_0!
- Estimation problem $\mathbf{P}(Y|X)$ is solvable for "simple" regression function, e.g., smooth regressors $y(x)$.
- Neighboring samples provide information to estimate x_0.

Frequentism and Alan's shoes

1. Alan defines a parametric model θ for a random shoe size: $\mathcal{N}(\theta, 1)$.
2. Alan defines the likelihood function:

$$\mathbf{P}(y_1, \ldots, y_n \mid \theta) = \prod_{i \leq n} \mathbf{P}(y_i \mid \theta) = \prod_{i \leq n} \mathcal{N}(y_i \mid \theta, 1).$$

3. Alan computes the estimator $\hat{\theta}_{ML} = \arg\max_\theta \mathbf{P}(y_1, \ldots, y_n \mid \theta)$.

$$\hat{\theta}_{ML} = \frac{1}{n} \sum_{i \leq n} y_i.$$

Frequentism and Bob's class

1. Bob defines a parametric model for the grades of the students in the classroom $\mathbf{y} = \mathbf{w}^\top \mathbf{x} + \xi$, where $\xi \sim \mathcal{N}(0, \epsilon I)$.
2. Bob defines the likelihood function:

$$\begin{aligned}
\mathbf{P}\left((x_1, y_1), \ldots, (x_n, y_n) \mid \mathbf{w}\right) = \prod_{i \leq n} \mathbf{P}\left(x_i, y_i \mid \mathbf{w}\right) &= \prod_{i \leq n} \mathbf{P}\left(y_i \mid x_i, \mathbf{w}\right) \mathbf{P}\left(x_i \mid \mathbf{w}\right) \\
&= \prod_{i \leq n} \mathbf{P}\left(y_i \mid x_i, \mathbf{w}\right) \mathbf{P}\left(x_i\right) \\
&= \prod_{i \leq n} \mathbf{P}\left(y_i \mid x_i, \mathbf{w}\right) \prod_{i \leq n} \mathbf{P}\left(x_i\right) \\
&= \text{const} \prod_{i \leq n} \mathbf{P}\left(y_i \mid x_i, \mathbf{w}\right).
\end{aligned}$$

Frequentism and Bob's class

1. Bob defines a parametric model for the grades of the students in the classroom $\mathbf{y} = \mathbf{w}^\top \mathbf{x} + \xi$, where $\xi \sim \mathcal{N}(0, \epsilon I)$.
2. Bob defines the likelihood function:
   $\mathbf{P}((x_1, y_1), \ldots, (x_n, y_n) \mid \mathbf{w}) = \text{const} \prod_{i \leq n} \mathbf{P}(y_i \mid x_i, \mathbf{w})$.
3. Estimate $\hat{\mathbf{w}}_{ML}$ by maximizing the log likelihood:

$$\begin{aligned}
\log \left(\text{const} \prod_{i \leq n} \mathbf{P}(y_i \mid x_i, \mathbf{w}) \right) &= \log \text{const} + \log \left(\prod_{i \leq n} \mathbf{P}(y_i \mid x_i, \mathbf{w}) \right) \\
&= \text{const} + \sum_{i \leq n} \log \mathbf{P}(y_i \mid x_i, \mathbf{w}) \\
&= \text{const} - \frac{1}{2} \sum_{i \leq n} \left(y_i - \mathbf{w}^\top x_i \right)^2 .
\end{aligned}$$

That is, $\hat{\mathbf{w}}_{ML} = \arg\max_{\mathbf{w}} -\frac{1}{2} \sum_{i \leq n} \left(y_i - \mathbf{w}^\top x_i \right)^2 = \arg\min_{\mathbf{w}} \sum_{i \leq n} \left(y_i - \mathbf{w}^\top x_i \right)^2$.

How should we estimate a density?

Data are i.i.d. drawn $\forall i,\ (X_i, Y_i) \sim \mathbf{P}(X, Y)$

Estimation: We introduce a parametric density $\mathbf{P}(X, Y|\theta)$ that measures the likelihood of (X, Y) given parameter θ.

Example: Consider linear regression with $Y = \beta^\mathsf{T} X +$ noise. The difference $Y - \beta^\mathsf{T} X$ is assumed to be normally distributed.

Inference: Choose the parameter value θ such that the observations $\mathcal{Z} = \{(X_i, Y_i), 1 \leq i \leq n\}$ are most probably.

$\Rightarrow$ maximum likelihood method

We will adopt the strategy of parametric density estimation. This approach requires to define a parametric form of the density (e.g., Gaussian, γ-distributed, etc.) and to estimate the respective location and width parameters. For Gaussians these parameters are mean and variance.

Recall: Maximum Likelihood Estimation

Data: Given is a sample set $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$.

Likelihood of the data set: $\mathbf{P}(\mathcal{X}|\theta) = \prod_{i \leq n} p(x_i|\theta)$

Estimation principle: Select the parameter $\hat{\theta}$ which maximizes the likelihood, i.e., the probability of the data given the parameter

$$\hat{\theta} \in \arg\max_\theta \mathbf{P}(\mathcal{X}|\theta)$$

The random variable $\hat{\theta}(\mathcal{X})$ is called an estimator for the parameter θ. The theory of estimators covers a large body of knowledge in statistics, where notions like consistency, unbiasedness, sufficiency and efficiency are introduced.

Procedure: Find the extremum of the log-likelihood function

Score $\Lambda(\theta) \stackrel{\triangle}{=} \nabla_\theta \log \mathbf{P}(\mathcal{X}|\theta) = \frac{\partial}{\partial\theta} \sum_{i \leq n} \log p(x_i|\theta) = 0$

Example: Multivariate Normal Distribution

Expectation values of a normal distribution and its estimation

Determine the mean of Gaussian distributed samples x_i:

$$\begin{aligned}
\log p(x_i|\theta) &= -\frac{1}{2}(x_i-\mu)^\mathsf{T}\Sigma^{-1}(x_i-\mu) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| \\
\frac{\partial}{\partial\mu}\sum_{i\le n}\log p(x_i|\theta) &= \frac{1}{2}\sum_{i\le n}\Sigma^{-1}(x_i-\mu) + \frac{1}{2}\sum_{i\le n}\left((x_i-\mu)^\mathsf{T}\Sigma^{-1}\right)^\mathsf{T} = 0 \\
\Sigma^{-1}\sum_{i\le n}(x_i-\mu) &= 0 \quad \Rightarrow \quad \hat{\mu}_n = \frac{1}{n}\sum_{i\le n}x_i \qquad \text{estimator for } \mu
\end{aligned}$$

Average value formula results from the quadratic form.

Unbiasedness: $\mathbb{E}[\hat{\mu}_n] = \frac{1}{n}\sum_{i\le n}\mathbb{E}x_i = \mathbb{E}[x] = \mu$

ML estimation of the variance (1d case)

$$\begin{aligned}
\frac{\partial}{\partial \sigma^2} \sum_{i \le n} \log p(x_i|\theta) &= -\frac{\partial}{\partial \sigma^2} \left(\sum_{i \le n} \frac{1}{\sigma^2} \|x_i - \mu\|^2 + \frac{n}{2} \log(2\pi\sigma^2) \right) \\
&= \frac{1}{2} \sum_{i \le n} \sigma^{-4} \|x_i - \mu\|^2 - \frac{n}{2} \sigma^{-2} \; = \; 0 \\
\Rightarrow \quad \hat{\sigma}_n^2 &= \frac{1}{n} \sum_{i \le n} \|x_i - \mu\|^2
\end{aligned}$$

$$\text{Multivariate case} \quad \hat{\Sigma}_n = \frac{1}{n} \sum_{i \le n} (x_i - \mu)(x_i - \mu)^{\mathsf{T}}$$

$\hat{\Sigma}_n$ is biased, e.g., $\mathbb{E}\hat{\Sigma}_n \neq \Sigma$, if μ is unknown.

Why ML-estimators? (Back to Alan's shoe shop)

- Eve proposes $\theta_{\text{median}} := \text{median}(y_1, \ldots, y_n)$.
- Melchor proposes $\theta_{\text{first}} := y_1$.

Eve and Melchor argue that their estimators are also effective, as $\mathbb{E}[\theta_{\text{median}}] = \mathbb{E}[\theta_{\text{first}}] = \theta_0$, the real mean shoe size. Furthermore, they require less calculations than in maximum likelihood.

Which is better in estimating the real mean shoe size θ_0? θ_{ML}, θ_{median}, or θ_{first}? Why?

Three reasons:

- θ_{ML} is consistent. More precisely, $\theta_{ML} \to \theta_0$ in probability as $n \to \infty$. In contrast, θ_{first} remains with a normal distribution, no matter n.
- θ_{ML} is asymptotically normal. More precisely, $1/\sqrt{n}\,(\theta_{ML} - \theta_0)$ converges in distribution to a random variable with distribution $\mathcal{N}\left(0, J^{-1}(\theta) I(\theta) J^{-1}(\theta)\right)$.
- θ_{ML} is asymptotically efficient. More precisely, θ_{ML} minimizes $\mathbb{E}\left[(\theta_{ML} - \theta_0)^2\right]$, as $n \to \infty$.

Understanding asymptotic efficiency

A measure for quantifying how good an estimator $\hat{\theta}$ is the expected mean squared error:

$$\mathbb{E}\left[\left(\hat{\theta} - \theta_0\right)^2\right].$$

Is there an estimator $\hat{\theta}^*$ that reaches $\mathbb{E}\left[\left(\hat{\theta}^* - \theta_0\right)^2\right] = 0$? The Cramer-Rao bound shows that there is not necessarily such an estimator:

$$\mathbb{E}\left[\left(\hat{\theta} - \theta_0\right)^2\right] \geq \frac{1}{I_n(\theta_0)}, \text{ for any estimator } \hat{\theta}.$$

Here, $I_n(\theta_0)$ is the Fisher information:
$I_n(\theta_0) = \mathbb{V}_{\theta_0}\left[\left[\frac{\partial}{\partial\theta}\log \mathbf{P}\left(Y_1, \ldots, Y_n \mid \theta\right)\right]_{\theta_0}\right] = n\mathbb{V}_{\theta_0}\left[\left[\frac{\partial}{\partial\theta}\log \mathbf{P}\left(Y \mid \theta\right)\right]_{\theta_0}\right].$

Understanding asymptotic efficiency

Is there at least an estimator $\hat{\theta}_{\text{eff}}$ that is efficient? That is,

$$\mathbb{E}\left[\left(\hat{\theta}_{\text{eff}} - \theta_0\right)^2\right] = \frac{1}{I_n(\theta_0)}.$$

Yes! The ML estimator $\hat{\theta}_{ML}$ is asymptotically efficient:

$$\lim_{n\to\infty} \mathbb{E}\left[\left(\hat{\theta}_{ML} - \theta_0\right)^2\right] = \frac{1}{I_n(\theta_0)}$$

Warning! Stein estimator

For finite samples, the ML estimator is not necessarily efficient!

Consider a multivariate random variable with distribution $\mathcal{N}\left(\theta, \sigma^2 I\right)$ with range $\mathbb{R}^d$ and $d \geq 4$.

If we sample a single point $\mathbf{y}$ from this distribution then the Stein estimator

$$\hat{\theta}_{JS} := \left(1 - \frac{(m-2)\sigma^2}{\|\mathbf{y}\|^2}\right)\mathbf{y}$$

is better than the ML-estimator $\hat{\theta}_{ML} = \mathbf{y}$.

That is,

$$\mathbb{E}\left[\left(\hat{\theta}_{JS} - \theta_0\right)^2\right] \leq \mathbb{E}\left[\left(\hat{\theta}_{ML} - \theta_0\right)^2\right], \qquad \text{for any } \theta_0$$

and the inequality is strict for some θ_0.

Conclusion

Bayesian method:

- Allows priors.
- Provides a distribution when estimating parameters.
- Requires efficient integration methods, when computing posteriors.
- The prior often induces a regularization term.

Frequentist method:

- Does not allow priors.
- Provides a single-point when estimating parameters.
- Requires only differentiation methods, when estimating parameters.
- MLE estimators are consistent, equivariant, asymptotically normal, and asymptotically efficient.

Remarks on ML-estimation

Bias of an estimator: $\text{bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta$.

The bias measures how much the expected value of the estimator deviates from the true parameter value. The design of unbiased estimators ($\mathbb{E}[\hat{\theta}_n] = \theta$) and bias reduction has been considered as a goal of statistics, although the work of Stein has shown that there exists biased estimators which are better in the least squares sense than any unbiased estimator.

Consistent estimator:

A point estimator $\hat{\theta}_n$ of a parameter θ is consistent if $\hat{\theta}_n \xrightarrow{\mathbf{P}} \theta$, i.e.,
$\forall \epsilon \; \mathbf{P}\{|\hat{\theta}_n - \theta| > \epsilon\} \overset{n\to\infty}{\longrightarrow} 0$,

Efficiency

ML estimators are **asymptotically efficient** estimators, i.e.,

$$\lim_{n\to\infty} \left(\mathbb{V}[\hat{\theta}^{\text{ML}}(x_1, \ldots, x_n)] \; I(\theta) \right)^{-1} = 1, \quad I = \mathbb{V}\left[\frac{\partial \log \mathbf{P}(x|\theta)}{\partial \theta} \right]$$

Rao Cramer inequality

Let $\hat{\theta}(X_1, \ldots, X_n)$ be an unbiased estimator of θ. Then the Rao Cramer inequality holds for the class of unbiased estimators

$$\begin{aligned} \int \left(\theta - \hat{\theta}\right)^2 \mathbf{P}(x_1, \ldots, x_n; \theta) dx_1 \ldots dx_n &\geq \frac{1}{I^{(n)}(\theta)} \quad \text{with} \\ \int \left[\frac{\partial}{\partial \theta} \ln \mathbf{P}(x_1, \ldots, x_n; \theta)\right]^2 \mathbf{P}(x_1, \ldots, x_n; \theta) d\mathcal{X} &= I^{(n)}(\theta) \end{aligned}$$

$I^{(n)}(\theta)$ is the Fisher information. The variance of an estimator is bounded from below by the inverse Fisher information which has been proven by Rao & Cramer.

Let $\Lambda := \frac{\partial \log \mathbf{P}(x;\theta)}{\partial \theta}$ denote the score function. Then the Rao-Cramer inequality is shown by using the Cauchy-Schwarz inequality and the fact that the expected score $\mathbb{E}[\Lambda] = \mathbb{E}\frac{\partial \log \mathbf{P}(x;\theta)}{\partial \theta} = 0$ vanishes, i.e.,

$$\begin{aligned} \left(\mathbb{E}[(\Lambda - \mathbb{E}[\Lambda])(\hat{\theta} - \mathbb{E}[\hat{\theta}])]\right)^2 &\leq \mathbb{E}[(\Lambda - \mathbb{E}[\Lambda])^2]\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] \\ &= \mathbb{E}[\Lambda^2]\mathbb{V}[\hat{\theta}]. \end{aligned}$$

The left hand side equates to $(\mathbb{E}[\Lambda\hat{\theta}])^2 = 1$. Furthermore, the unbiasedness of the estimator $\mathbb{E}[\hat{\theta}] = \theta$ is exploited.

Importance of the Maximum Likelihood Method

Def.: The sequence $\xi_1, \xi_2, \ldots$ of r.v. converges **in distribution** to the r.v. ξ (abbr. $\xi_n \xrightarrow{d} \xi$), if for every bounded continuous function $f = f(x)$ holds

$$\mathbb{E}f(\xi_n) \to \mathbb{E}f(\xi) \quad n \to \infty$$

Theorem: Convergence of ML estimators to best model θ_0

$$\sqrt{n}(\hat{\theta}_n^{\mathsf{ML}} - \theta_0) \quad \xrightarrow{d} \quad \mathcal{N}(0, J^{-1}(\theta_0) I(\theta_0) J^{-1}(\theta_0)) \quad \text{for} \quad n \to \infty$$

$$\text{with} \quad J(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathbf{P}(x|\theta)}{\partial\theta\partial\theta^{\mathsf{T}}}\right]$$

$$I(\theta) = \mathbb{V}\left[\frac{\partial \log \mathbf{P}(x|\theta)}{\partial\theta}\right]$$

$I = J$, if the true distribution $\mathbf{P}(x|\theta^\star)$ belongs to the family of parametric distributions $\{\mathbf{P}(x|\theta)\}$; (realizable case $\theta_0 = \theta^\star$).

Proof sketch $\left(\Lambda_n(\theta) := \sum_{i\leq n} \frac{\partial}{\partial\theta} \log p(x_i|\theta)\right)$

θ_0 denotes the model which is closest to the true model.

The following Taylor expansion holds for ML estimators:

$$\overset{\text{condition for ML estimator}}{0 \overset{!}{=} \Lambda_n(\hat{\theta})} = \Lambda_n(\theta_0) + \frac{\partial^2}{\partial\theta\partial\theta^\intercal} \sum_{i\leq n} \log p(x_i|\tilde{\theta})(\hat{\theta} - \theta_0)$$

where $\tilde{\theta}$ is on the line between $\hat{\theta}$ and θ_0.

$$\begin{aligned}
\Rightarrow \quad \sqrt{n}(\hat{\theta} - \theta_0) &= \frac{1}{\sqrt{n}} \Bigg(\underbrace{-\frac{1}{n} \frac{\partial^2}{\partial\theta\partial\theta^\intercal} \sum_{i\leq n} \log p(x_i|\theta)\Big|_{\tilde{\theta}}}_{\overset{n\to\infty}{\longrightarrow} -\mathbb{E}_p\left[\frac{\partial^2}{\partial\theta\partial\theta^\intercal} \log p(x|\theta)\Big|_{\tilde{\theta}}\right] = J} \Bigg)^{-1} \Lambda_n(\theta_0) \\
\frac{1}{\sqrt{n}}\Lambda_n(\theta_0) &= \frac{1}{\sqrt{n}} \sum_{i\leq n} \frac{\partial}{\partial\theta} \log p(x_i|\theta)\Big|_{\theta=\theta_0}
\end{aligned}$$

is a r.v. with covariance matrix $\mathbb{V}_p\left[\frac{\partial}{\partial\theta} \log p(x|\theta_0)\right] = \mathbb{V}_p[\Lambda] = I$.

Since $\Lambda_n(\theta_0)$ is a sum of n random variables, the distribution of $\Lambda_n(\theta_0)$ converges (under mild regularity assumptions) asymptotically toward the normal distribution with $\mu = \mathbb{E}[\Lambda_n(\theta_0)] = 0$, i.e., $\theta_0 = \theta^\star$.

$$\frac{1}{\sqrt{n}}\Lambda_n(\theta_0) \quad \xrightarrow{\mathsf{d}} \quad \mathcal{N}(0, I)$$

$$\frac{1}{\sqrt{n}}\left(-\frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta^\mathsf{T}}\sum_{i\leq n}\log p(x_i|\tilde{\theta})\right)^{-1}\Lambda_n(\theta_0) \quad \xrightarrow{\mathsf{d}} \quad J^{-1}\mathcal{N}(0, I)$$

$$\xrightarrow{\mathsf{d}} \quad \mathcal{N}(0, J^{-1}IJ^{-1})$$

The last convergence statement holds since $\mathbb{E}[J^{-1}\Lambda_n(\theta_0)/\sqrt{n}] = 0$ is true provided that $\mathbb{E}[\Lambda_n(\theta_0)/\sqrt{n}] = 0$ holds. Therefore,
$\mathbb{V}[J^{-1}\Lambda_n(\theta_0)/\sqrt{n}] = \mathbb{E}[(J^{-1}\Lambda_n(\theta_0)/\sqrt{n})(J^{-1}\Lambda_n(\theta_0)/\sqrt{n})^\mathsf{T}] = \mathbb{E}[J^{-1}IJ^{-1}]$.

Realizable Model

Let the correct model be in the set $\{p(x|\theta)\}$, i.e., .

$$\begin{aligned}
\tilde{K} &:= \frac{\partial}{\partial\theta}\int \frac{\partial \log p(x|\theta)}{\partial\theta^\mathsf{T}}\bigg|_{\theta=\theta_0} p(x|\theta_0)dx \\
&= \frac{\partial^2}{\partial\theta\partial\theta^\mathsf{T}} \underbrace{\int p(x|\theta_0)dx}_{=1,\ \forall\theta_0} = 0
\end{aligned}$$

Also:
$$\begin{aligned}
\tilde{K} &= \int \frac{\partial^2}{\partial\theta\partial\theta^\mathsf{T}} \log p(x|\theta)\bigg|_{\theta=\theta_0} p(x|\theta_0)dx + \\
&\quad \int \frac{\partial}{\partial\theta} \log p(x|\theta)\bigg|_{\theta=\theta_0} \underbrace{\frac{\partial}{\partial\theta^\mathsf{T}} p(x|\theta)\bigg|_{\theta=\theta_0}}_{p(x|\theta_0)\,\frac{\partial}{\partial\theta^\mathsf{T}}\log p(x|\theta)\big|_{\theta=\theta_0}} dx \\
&= -J + I \\
\Longrightarrow \quad J &= I
\end{aligned}$$

□

Summary on Maximum Likelihood Estimators

Consistency: MLEs are consistent, i.e., $\hat{\theta}_n^{\text{ML}} \xrightarrow{p} \theta_0$!

Equivariance

If $\hat{\theta}_n$ is MLE of θ then $g(\hat{\theta}_n)$ is MLE of $g(\theta)$.

Let $\tau = g(\theta)$ be a function of θ and $h = g^{-1}$ denotes the inverse of g. Then $\hat{\theta}_n = h(\hat{\tau}_n)$. For any τ, $\mathcal{L}(\tau) = \prod_i f(x_i|h(\tau)) = \prod_i f(x_i|\theta) = \mathcal{L}(\theta)$. Hence, for any τ, $\mathcal{L}(\tau) = \mathcal{L}(\theta) \leq \mathcal{L}_n(\hat{\theta}) = \mathcal{L}_n(\hat{\tau})$. □

Asymptotic normality: $\sqrt{n}(\theta - \hat{\theta}_n) \rightsquigarrow \mathcal{N}(0, J^{-1}IJ^{-1})$

Asymptotic efficiency

For well-behaved estimators, the MLE has the smallest variance for large n.

Bayesian Learning

Parameter distribution: θ is considered to be a random variable with distribution $p(\theta|\mathcal{X})$.

Assumption: $p(x)$ is unknown ($X \sim p(x)$), $p(x|\theta)$ is known.

"Wanted": $p(X = x|\mathcal{X})$, i.e., the probability of x for given the sample set $\mathcal{X}$, (*class conditional density*)

$$p(X = x|\mathcal{X}) \quad = \quad \int \underbrace{p(x, \theta|\mathcal{X})}_{p(x|\theta,\mathcal{X})p(\theta|\mathcal{X})} d\theta \quad = \quad \int p(x|\theta)p(\theta|\mathcal{X})d\theta$$

Note! Class conditional density depends on the posterior by an integration. ($p(x|\theta, \mathcal{X}) = p(x|\theta)$ since $x_i \in \mathcal{X}$ and x are i.i.d).

Approximation: $p(\theta|\mathcal{X}) \sim \delta(\theta - \hat{\theta}) \;\Rightarrow\; p(x|\mathcal{X}) \backsimeq p(x|\hat{\theta})$

Bayesian learning of a normal distribution

Distribution assumption

$p(x|\mu) \sim \mathcal{N}(\mu, \sigma^2)$, $p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$, let $\mathcal{X} = \{x_1, \ldots, x_n\}$

$$
\begin{aligned}
p(\mu|\mathcal{X}) &= \alpha \prod_{i \leq n} p(x_i|\mu) p(\mu) \\
&= \alpha \prod_{i \leq n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right) \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right) \\
&= \alpha' \exp\left(-\frac{1}{2}\left(\sum_{i \leq n}\left(\frac{x_i - \mu}{\sigma}\right)^2 + \left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right)\right) \\
&= \alpha'' \exp\left(-\frac{1}{2}\left(\mu^2\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) - 2\left(\frac{1}{\sigma^2}\sum_{i \leq n} x_i + \frac{\mu_0}{\sigma_0^2}\right)\mu\right)\right)
\end{aligned}
$$

Exponent is a quadratic form $\Rightarrow$ Gaussian distribution.

$p(\mu|\mathcal{X})$ is called reproducing density with conjugate prior $p(\mu)$.
($\alpha, \alpha', \alpha''$ are normalization constants.)

Coefficient comparison

$p(\mu|\mathcal{X}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right)$

$$\Rightarrow \quad \frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} = \frac{n+\sigma^2/\sigma_0^2}{\sigma^2}$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \quad \text{with} \quad \hat{\mu}_n = \frac{1}{n}\sum_{i\leq n} x_i$$

Note: the ratio σ^2/σ_0^2 measures the effective number of samples which correspond to the a priori information! (Idea of *conjugate priors*)

Solve for μ_n, σ_n^2

$$\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2} \qquad \mu_n = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$$

Class conditional density

$$
\begin{aligned}
p(x|\mathcal{X}) &= \int p(x|\mu)p(\mu|\mathcal{X})d\mu \\
&= \frac{1}{2\pi\sigma\sigma_n}\exp\left(-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right)f(\sigma,\sigma_n) \\
\text{with}\quad f(\sigma,\sigma_n) &= \int \exp\left(-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}(\mu-\frac{\sigma_n^2 x+\sigma^2\mu_n}{\sigma^2+\sigma_n^2})^2\right)d\mu
\end{aligned}
$$

Summary $p(x|\mathcal{X}) \sim \mathcal{N}(\mu_n, \sigma^2+\sigma_n^2)$

Remark: The maximum likelihood method only estimates the parameters $\hat{\mu}, \hat{\sigma}$, but not the distribution!

Multivariate case is analogous:

$$
\begin{aligned}
\Sigma_n^{-1} &= n\Sigma^{-1}+\Sigma_0^{-1} \\
\Sigma_n^{-1}\mu_n &= n\Sigma^{-1}\hat{\mu}_n+\Sigma_0^{-1}\mu_0
\end{aligned}
$$

Bayesian learning of the mean of normal distributions in one and two dimensions.

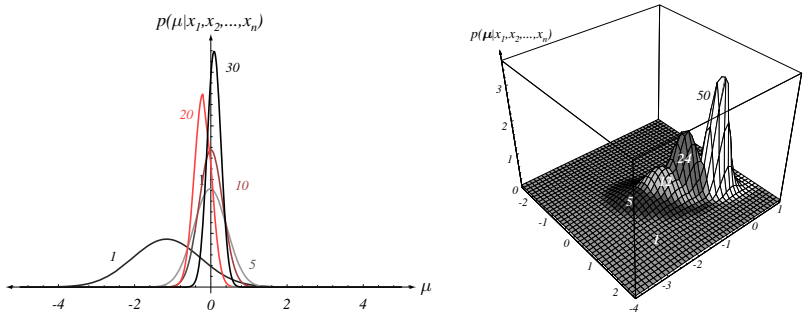


Fig. 3.2: The densities in one or two dimensions are indexed by the number of sample points used for estimation. The convergence to a more and more peaked distribution is apparent. Duda, Hart, Stork, (2000) *Pattern Classification, Wiley*

Recursive Bayesian Estimation

Assumption: Data are available in sequential order, e.g., $\mathcal{X}^n = (x_1, x_2, \ldots, x_n)$ are used one datum after another to estimate the data distribution.

$$\text{likelihood of data} \quad p(\mathcal{X}^n|\theta) = p(x_n|\theta)\underbrace{p(\mathcal{X}^{n-1}|\theta)}_{\text{likelihood of the data set without } x_n}$$

$$\text{posterior} \quad p(\theta|\mathcal{X}^n) = \frac{p(x_n|\theta)\overbrace{p(\theta|\mathcal{X}^{n-1})}^{\text{Prior for } n\text{-th estimate}}}{\int p(x_n|\theta)p(\theta|\mathcal{X}^{n-1})d\theta}$$

$$\text{prior} \quad p(\theta|\mathcal{X}^0) = p(\theta)$$

Online estimation: these equations estimate the density in an incremental fashion; close relation to *online* algorithms.

Differences: ML-/Bayes Estimation

Asymptotic equivalence: for reasonable a priori distributions (the true solution should have a non-vanishing a priori probability) the results are equal in the asymptotic limit.

Advantages of the ML method: we only need differential calculus and gradient descent techniques.

Disadvantage of the Bayes method: we have to integrate.

Non-informative Priors: a “flat” prior yields little information, since the model uncertainty is not significantly reduced.
See also effective sample sizes σ^2/σ_0^2 for normal distributions.

Schematic Behavior of Bias & Variance

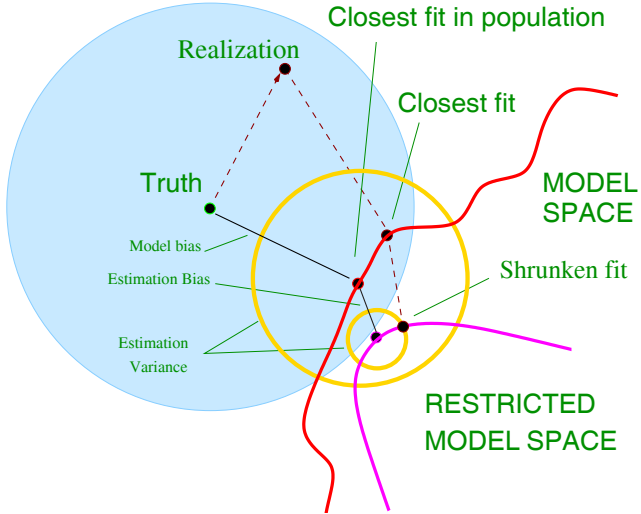


FIGURE 7.2. Schematic of the behavior of bias and variance.
(Hastie, Tibshirani, Friedman (2009) The Elements of Statistical Learning), p. 225