**Notes on Statistical and Machine Learning**

# Model Inference and Averaging

**Chapter:** *11* **Prepared by:** *Chenxi Zhou*

This note is prepared based on

- *Chapter 8, Model Inference and Averaging* in Hastie, Tibshirani, and Friedman (2009), and
- *Chapter 14, Committee Machines* in Izenman (2009).

## I. The Bootstrap and Maximum Likelihood Methods

1. **A Review of Bootstrap Methods:** The *bootstrap* is a tool for assessing uncertainty. It can estimate the prediction error well.

   (a) *Setup:* Let $\mathbf{Z} := (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n)$ denote the training dataset, where $\mathbf{z}_i := (\mathbf{x}_i, y_i)$ for all $i = 1, 2, \cdots, n$.

   (b) *Procedures:*

   *Step 1:* Randomly draw datasets *with replacement* from the training data $\mathbf{Z}$, where each sample dataset has the same size as the original training set;

   *Step 2:* Do *Step 1* $B$ times and produce $B$ bootstrap datasets;

   *Step 3:* Fit the model to each of the bootstrap datasets and examine the behavior of the fits over the $B$ replications.

   *Remark.* This procedure is called the *nonparametric bootstrap* since the method is model-free and there is no specific parametric model to generate new datasets.

2. **A Smoothing Example:** Let the training dataset be $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n)$, with $\mathbf{z}_i = (x_i, y_i)$ for $i = 1, \cdots, n$. We assume that $x_i \in \mathbb{R}$ is a one-dimensional input and let $y_i \in \mathbb{R}$ be the outcome.

   Suppose we use a cubic spline to fit the data with three knots placed at the quartiles of the input values. This is a 7-dimensional linear space of functions, where "seven" is obtained by

   $$4 \text{ regions} \times 4 \text{ parameters in each region} - 3 \text{ knots} \times 3 \text{ constraints for each knot} = 7.$$

   Then, we represent the solution in the form

   $$\mu(x) = \sum_{j=1}^{7} \beta_j h_j(x), \tag{1}$$

where $h_j$'s are the basis functions. We can think of the function $\mu$ as representing the conditional mean $\mathbb{E}[Y \mid X = x]$. Let $\mathbf{H}$ be the $n \times 7$ matrix with $(i, j)$-th element $h_j(x_i)$.

- The least squares estimate of $\boldsymbol{\beta} := (\beta_1, \cdots, \beta_7)^\top \in \mathbb{R}^7$ is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{Y}, \tag{2}$$

  where $\mathbf{Y} := (y_1, \cdots, y_n)^\top \in \mathbb{R}^n$ is the response vector.

- The fitted value at $x$ is

$$\hat{\mu}(x) = \sum_{j=1}^{7} \hat{\beta}_j h_j(x), \tag{3}$$

  where $\hat{\beta}_j$ is the $j$-th component of $\widehat{\boldsymbol{\beta}}$.

- The estimated covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\widehat{\mathrm{Var}}[\widehat{\boldsymbol{\beta}}] = \hat{\sigma}^2 (\mathbf{H}^\top \mathbf{H})^{-1}, \tag{4}$$

  where the noise variance $\sigma^2$ is estimated by

$$\hat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^{n} \big(y_i - \hat{\mu}(x_i)\big)^2. \tag{5}$$

- Let $\mathbf{h}(x) := (h_1(x), h_2(x), \cdots, h_7(x))^\top$. The standard error of a prediction $\hat{\mu}(x) = \mathbf{h}(x)^\top \widehat{\boldsymbol{\beta}}$ is

$$\widehat{\mathrm{se}}[\widehat{\mu}(x)] = \big[\mathbf{h}(x)^\top (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{h}(x)\big]^{1/2} \hat{\sigma}. \tag{6}$$

3. **Pointwise Confidence Interval Using the Nonparametric Bootstrap in the Smoothing Example:** We apply the bootstrap in the smoothing example above to obtain the pointwise confidence interval. The procedure is outlined as follows:

   (1) Draw $B$ datasets each of size $n$ with replacement from the training set;

   (2) For each bootstrap data set $\mathbf{Z}^*$, we fit a cubic spline model $\hat{\mu}^*$;

   (3) With the $B$ bootstrap samples, form a $(1 - \alpha) \cdot 100\%$ pointwise confidence band from the percentiles at each $x$ by finding the $(\alpha/2) \times B$ largest and smallest values at each $x$.

4. **Parametric Bootstrap:** Suppose further we assume that the errors in the smoothing example follow a Gaussian distribution with mean 0 and variance $\sigma^2$, that is,

$$Y = \mu(x) + \varepsilon, \qquad \mu(x) = \sum_{j=1}^{7} \beta_j h_j(x), \qquad \varepsilon \sim \mathrm{Normal}(0, \sigma^2). \tag{7}$$

   The *parametric bootstrap* procedure is outlined as follows:

(a) Simulate new responses by adding Gaussian noise to the predicted values, that is, for $i = 1, \cdots, n$,

$$y_i^* = \hat{\mu}(x_i) + \varepsilon_i^*, \tag{8}$$

where $\varepsilon_i^* \sim \text{Normal}(0, \hat{\sigma}^2)$;

(b) Repeat the preceding process $B$ times. Then, the resulting bootstrap dataset is of the form $\{(x_i, y_i^*)\}_{i=1}^n$. We compute the $B$-spline model based on them;

(c) The *predicted values* estimated from a bootstrap sample $\mathbf{Y}^* := (y_1^*, \cdots, y_n^*)^\top$ is

$$\hat{\mu}^*(x) = \mathbf{h}(x)^\top (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{Y}^*,$$

and has the distribution

$$\hat{\mu}^*(x) \sim \text{Normal}\big(\hat{\mu}(x), \mathbf{h}(x)^\top (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{h}(x) \hat{\sigma}^2\big). \tag{9}$$

In particular, the mean of this distribution is the same as the least squares estimate, and the standard deviation is the same as that of the least squares, that is, (6).

5. **Maximum Likelihood Inference:** Consider a probability density or probability mass function $g_{\boldsymbol{\theta}}$ whose functional form is known but the parameter $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ is unknown. Suppose we have $n$ random samples from it

$$Z_1, \cdots, Z_n \overset{\text{i.i.d}}{\sim} g_{\boldsymbol{\theta}}, \tag{10}$$

Maximum likelihood is based on the likelihood function

$$L(\boldsymbol{\theta}; \mathbf{Z}) := \prod_{i=1}^n g_{\boldsymbol{\theta}}(Z_i), \tag{11}$$

where $\mathbf{Z} := (Z_1, \cdots, Z_n)$. We think $L(\boldsymbol{\theta}; \mathbf{Z})$ as a function of the parameter $\boldsymbol{\theta}$, with the data $\mathbf{Z}$ being fixed.

Typically, we work with the *log-likelihood function*, denoted by

$$\ell(\boldsymbol{\theta}; \mathbf{Z}) := \log L(\boldsymbol{\theta}; \mathbf{Z}) = \sum_{i=1}^n \log g_{\boldsymbol{\theta}}(Z_i). \tag{12}$$

Each component $\log g_{\boldsymbol{\theta}}(Z_i)$ is called a *log-likelihood component*. The maximum likelihood method chooses the value of $\widehat{\boldsymbol{\theta}}$ that maximizes $\ell(\boldsymbol{\theta}; \mathbf{Z})$.

6. **Score Function:** The first-order derivative of the log-likelihood function $\ell(\boldsymbol{\theta}; \mathbf{Z})$ with respect to $\boldsymbol{\theta}$ is called the *score function*, that is,

$$\nabla \ell(\boldsymbol{\theta}; \mathbf{Z}) = \sum_{i=1}^n \nabla \ell(\boldsymbol{\theta}; Z_i) = \sum_{i=1}^n \frac{\partial \ell(\boldsymbol{\theta}; Z_i)}{\partial \boldsymbol{\theta}}. \tag{13}$$

Assuming that the likelihood achieves its maximum in the *interior* of the parameter space, we then have $\nabla \ell(\widehat{\boldsymbol{\theta}}; \mathbf{Z}) = \mathbf{0}_p$.

7. **Information Matrix:** The *information matrix* is defined to be

$$\mathbf{I}(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{\theta}; Z_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}. \tag{14}$$

   (a) *Observed Information:* When $\mathbf{I}(\boldsymbol{\theta})$ is evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, it is often called the *observed information*, that is,

$$\mathbf{I}(\widehat{\boldsymbol{\theta}}) = -\sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{\theta}; Z_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}; \tag{15}$$

   (b) *Expected Information:* The expectation of the information matrix is called the *Fisher information* (or *expected information*),

$$\mathbf{i}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{I}(\boldsymbol{\theta})]. \tag{16}$$

8. **Classical Asymptotic Results:** Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$ be the maximum likelihood estimator of $\boldsymbol{\theta}$ from the data $Z_1, \cdots, Z_n$. The sampling distribution of the maximum likelihood estimator has a limiting normal distribution, that is,

$$\widehat{\boldsymbol{\theta}} \to \text{Normal}\big(\boldsymbol{\theta}_0, \mathbf{i}(\boldsymbol{\theta}_0)^{-1}\big), \qquad \text{as } n \to \infty. \tag{17}$$

   Since $\boldsymbol{\theta}_0$ is unknown, the sampling distribution of $\widehat{\boldsymbol{\theta}}$ can be *approximated* by

$$\text{Normal}(\widehat{\boldsymbol{\theta}}, \mathbf{i}(\widehat{\boldsymbol{\theta}})^{-1}) \qquad \text{or} \qquad \text{Normal}(\widehat{\boldsymbol{\theta}}, \mathbf{I}(\widehat{\boldsymbol{\theta}})^{-1}). \tag{18}$$

   For all $j = 1, 2, \cdots, p$, the standard error of $\hat{\theta}_j$ can be estimated using the diagonal elements of the information matrix, either the expected one or the observed one; that is,

$$\sqrt{[\mathbf{i}(\widehat{\boldsymbol{\theta}})^{-1}]_{j,j}} \qquad \text{and} \qquad \sqrt{[\mathbf{I}(\widehat{\boldsymbol{\theta}})^{-1}]_{j,j}}. \tag{19}$$

   The $(1-\alpha)\cdot 100\%$ confidence interval for each $\theta_j$ can be obtained by

$$\hat{\theta}_j \pm z_{1-\alpha/2}\sqrt{[\mathbf{i}(\widehat{\boldsymbol{\theta}})^{-1}]_{j,j}} \qquad \text{or} \qquad \hat{\theta}_j \pm z_{1-\alpha/2}\sqrt{[\mathbf{I}(\widehat{\boldsymbol{\theta}})^{-1}]_{j,j}}, \tag{20}$$

   respectively, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)\cdot 100$-th percentile of the standard normal distribution.

9. **Asymptotic Distribution of Log-Likelihood Function:** Asymptotically, the log-likelihood function evaluated at the maximum likelihood estimator has a chi-squared distribution

$$2\big[\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)\big] \sim \chi^2_p, \qquad \text{as } n \to \infty, \tag{21}$$

   where $p$ is the number of components in $\boldsymbol{\theta}$ and $\chi^2_p$ is the chi-squared distribution with $p$ degrees of freedom.

With this result, a more accurate $(1-\alpha)\cdot 100\%$ confidence interval is the set of all $\boldsymbol{\theta}$ such that

$$2\big[\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)\big] \leq \chi^2_{p,1-\alpha}, \tag{22}$$

where $\chi^2_{p,1-\alpha}$ is the $(1-\alpha)\cdot 100$-th percentile of the chi-squared distribution with $p$ degrees of freedom.

10. **Example:** Consider the smoothing example where we assumed that

$$Y = \mu(x) + \varepsilon, \qquad \mu(x) = \sum_{j=1}^{7} \beta_j h_j(x), \qquad \varepsilon \sim \text{Normal}(0, \sigma^2). \tag{23}$$

With the parameter $\boldsymbol{\theta} := (\boldsymbol{\beta}^\top, \sigma^2)^\top$ and data $\mathbf{Z}$, the log-likelihood function is

$$\ell(\boldsymbol{\theta}; \mathbf{Z}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{h}(x_i)^\top\boldsymbol{\beta})^2. \tag{24}$$

Then, the maximum likelihood estimator of $\boldsymbol{\theta}$ is obtained by setting

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{Z})}{\partial \boldsymbol{\beta}} = \mathbf{0}_7, \qquad \text{and} \qquad \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{Z})}{\partial \sigma^2} = 0. \tag{25}$$

The resulting estimators of $\boldsymbol{\beta}$ and $\sigma^2$ are

$$\widehat{\boldsymbol{\beta}} := (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{Y}, \qquad \text{and} \qquad \hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}(x_i))^2, \tag{26}$$

respectively, which are the same as presented before.

11. **Connection Between Bootstrap and Maximum Likelihood:** Bootstrap is a *computer implementation* of nonparametric or parametric maximum likelihood. Bootstrap allows one to compute maximum likelihood estimates of standard errors and other statistical quantities in settings where no formulas are available.

# II. Bayesian Methods

1. **General Idea of Bayesian Inference:** In Bayesian approach, we specify

   (a) a sampling model for the data $\mathbf{z}$ given the parameters, $f(\cdot \mid \boldsymbol{\theta})$, and

   (b) a prior density function on the parameters $\boldsymbol{\theta}$, $p$.

   Then, the *posterior distribution* is computed by using the Bayes theorem

$$p(\boldsymbol{\theta} \mid \mathbf{z}) = \frac{f(\mathbf{z} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{f(\mathbf{z})}, \tag{27}$$

where

$$f(\mathbf{z}) = \int_{\Theta} f(\mathbf{z} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

Note that the posterior distribution $p(\boldsymbol{\theta} \,|\, \mathbf{z})$ reflects our updated knowledge about $\boldsymbol{\theta}$ after we see the data.

To understand $p(\cdot \,|\, \mathbf{z})$, one can draw samples from it or summarize it by the mean or the mode.

2. **Characteristics of Bayesian Approach:** The Bayesian approach
   (a) specifies the prior distribution to express the *uncertainty* present before seeing the data, and
   (b) allow the uncertainty to remain after seeing the data to be expressed in the form of the posterior distribution.

3. **Prediction Distribution:** Suppose we are given a new data point $\mathbf{z}_{\text{new}}$, the prediction distribution is calculated by

$$f(\mathbf{z}_{\text{new}} \,|\, \mathbf{z}) = \int_{\Theta} f(\mathbf{z}_{\text{new}} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, \mathbf{z}) \mathrm{d}\boldsymbol{\theta}. \tag{28}$$

   *Remark.* In the maximum likelihood approach, we just use $f(\mathbf{z}_{\text{new}} \,|\, \widehat{\boldsymbol{\theta}})$ to represent the predicted distribution of future observations and to make predictions. This does *not* account for the uncertainty in estimating $\boldsymbol{\theta}$.

4. **Bayesian Approach to the Smoothing Example:** We make the following assumptions:
   - the data $Y$ has the following parametric model

$$Y = \mu(x) + \varepsilon, \qquad \mu(x) = \sum_{j=1}^{7} \beta_j h_j(x), \qquad \varepsilon \sim \text{Normal}(0, \sigma^2);$$

   - the variance of the error $\varepsilon$, $\sigma^2$, is known;
   - the observed feature values $x_1, x_2, \cdots, x_n \in \mathbb{R}$ are *fixed* so that the randomness in the data comes *solely* from $Y$ varying around its mean $\mu$;
   - assume the following prior distribution on $\boldsymbol{\beta}$

$$\boldsymbol{\beta} \sim \text{Normal}(0, \tau \boldsymbol{\Sigma}), \tag{29}$$

     where $\tau > 0$.

Under the assumptions above, the implicit process for $\mu(x)$ is a Gaussian process with the covariance kernel

$$\begin{aligned} K(x, x') &= \operatorname{Cov}(\mu(x), \mu(x')) \\ &= \operatorname{Cov}(\mathbf{h}(x)^\top \boldsymbol{\beta}, \mathbf{h}(x')^\top \boldsymbol{\beta}) \\ &= \mathbf{h}(x)^\top \operatorname{Var}[\boldsymbol{\beta}] \mathbf{h}(x') \\ &= \tau \mathbf{h}(x)^\top \boldsymbol{\Sigma} \mathbf{h}(x'). \end{aligned}$$

Then, the posterior distribution of $\boldsymbol{\beta}$ is also Gaussian. We find the parameters associated with this posterior distribution. Notice that

$$\mathbf{Z} \,|\, \boldsymbol{\beta} \sim \text{Normal}(\mathbf{H}, \sigma^2 \mathbf{I}), \qquad \text{and} \qquad \boldsymbol{\beta} \sim \text{Normal}(\mathbf{0}, \tau \boldsymbol{\Sigma}), \tag{30}$$

where $\mathbf{H}$ is the $n \times 7$ matrix with $(i, j)$-th element $h_j(x_i)$.

Then, the density function of the posterior distribution of $\boldsymbol{\beta}$ is proportional to the product of the likelihood function and the prior density:

$$\begin{aligned} f(\boldsymbol{\beta} \,|\, \mathbf{Z}) &\propto f(\mathbf{Z} \,|\, \boldsymbol{\beta}) f(\boldsymbol{\beta}) \\ &\propto \exp\left\{ -\frac{1}{2} (\mathbf{Y} - \mathbf{H}\boldsymbol{\beta})^\top (\sigma^2 \mathbf{I})^{-1} (\mathbf{Y} - \mathbf{H}\boldsymbol{\beta}) \right\} \cdot \exp\left\{ -\frac{1}{2} \boldsymbol{\beta}^\top (\tau \boldsymbol{\Sigma})^{-1} \boldsymbol{\beta} \right\} \\ &\propto \exp\left\{ -\frac{1}{2\sigma^2} \big( \mathbf{Y}^\top \mathbf{Y} - 2\mathbf{Y}^\top \mathbf{H}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{H}^\top \mathbf{H} \boldsymbol{\beta} \big) - \frac{1}{2\tau} \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta} \right\} \\ &\propto \exp\left\{ -\frac{1}{2\sigma^2} \left[ \boldsymbol{\beta}^\top \left( \mathbf{H}^\top \mathbf{H} + \frac{\sigma^2}{\tau} \boldsymbol{\Sigma}^{-1} \right) \boldsymbol{\beta} - 2\mathbf{Y}^\top \mathbf{H} \boldsymbol{\beta} \right] \right\}. \end{aligned}$$

Supposing that the posterior mean of $\boldsymbol{\beta}$ is $\widetilde{\boldsymbol{\mu}}$ and the posterior covariance matrix is $\widetilde{\boldsymbol{\Sigma}}$, we must have

$$\begin{aligned} -\frac{1}{2} (\boldsymbol{\beta} - \widetilde{\boldsymbol{\mu}})^\top \widetilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\beta} - \widetilde{\boldsymbol{\mu}})^\top &= -\frac{1}{2} (\boldsymbol{\beta}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\beta} - 2\widetilde{\boldsymbol{\mu}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\beta} + \widetilde{\boldsymbol{\mu}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{\mu}}) \\ &= -\frac{1}{2\sigma^2} \left[ \boldsymbol{\beta}^\top \left( \mathbf{H}^\top \mathbf{H} + \frac{\sigma^2}{\tau} \boldsymbol{\Sigma}^{-1} \right) \boldsymbol{\beta} - 2\mathbf{Y}^\top \mathbf{H} \boldsymbol{\beta} + C \right], \end{aligned}$$

where $C$ is a constant independent of $\boldsymbol{\beta}$. From the equation above, we must have

$$\widetilde{\boldsymbol{\Sigma}}^{-1} = \frac{1}{\sigma^2} \left( \mathbf{H}^\top \mathbf{H} + \frac{\sigma^2}{\tau} \boldsymbol{\Sigma}^{-1} \right), \qquad \text{and} \qquad \widetilde{\boldsymbol{\mu}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\beta} = \frac{1}{\sigma^2} \mathbf{Y}^\top \mathbf{H} \boldsymbol{\beta}. \tag{31}$$

It follows that

$$\widetilde{\boldsymbol{\Sigma}} = \left( \mathbf{H}^\top \mathbf{H} + \frac{\sigma^2}{\tau} \boldsymbol{\Sigma}^{-1} \right)^{-1} \sigma^2, \tag{32}$$

and

$$\widetilde{\boldsymbol{\mu}} = \left(\mathbf{H}^\top \mathbf{H} + \frac{\sigma^2}{\tau}\boldsymbol{\Sigma}^{-1}\right)^{-1} \mathbf{H}^\top \mathbf{Y}. \tag{33}$$

In summary, the posterior distribution of $\boldsymbol{\beta}$ is Gaussian with the following parameters

$$\begin{aligned} \mathbb{E}[\boldsymbol{\beta} \,|\, \mathbf{Z}] &= \left(\mathbf{H}^\top \mathbf{H} + \frac{\sigma^2}{\tau}\boldsymbol{\Sigma}^{-1}\right)^{-1} \mathbf{H}^\top \mathbf{Y}, \\ \mathrm{Var}[\boldsymbol{\beta} \,|\, \mathbf{Z}] &= \left(\mathbf{H}^\top \mathbf{H} + \frac{\sigma^2}{\tau}\boldsymbol{\Sigma}^{-1}\right)^{-1} \sigma^2. \end{aligned} \tag{34}$$

It follows that the corresponding posterior values for $\mu(x)$ are

$$\begin{aligned} \mathbb{E}\big[\mu(x) \,|\, \mathbf{Z}\big] &= \mathbb{E}[\mathbf{h}(x)^\top \boldsymbol{\beta} \,|\, \mathbf{Z}] = \mathbf{h}(x)^\top \left(\mathbf{H}^\top \mathbf{H} + \frac{\sigma^2}{\tau}\boldsymbol{\Sigma}^{-1}\right)^{-1} \mathbf{H}^\top \mathbf{Y}, \\ \mathrm{Cov}\big[\mu(x), \mu(x') \,|\, \mathbf{Z}\big] &= \mathrm{Cov}\big[\mathbf{h}(x)^\top \boldsymbol{\beta}, \mathbf{h}(x')^\top \boldsymbol{\beta} \,|\, \mathbf{Z}\big] = \mathbf{h}(x)^\top \left(\mathbf{H}^\top \mathbf{H} + \frac{\sigma^2}{\tau}\boldsymbol{\Sigma}^{-1}\right)^{-1} \mathbf{h}(x')\sigma^2. \end{aligned}$$

*Remarks.*

(a) The prior distribution (29) of $\boldsymbol{\beta}$ when $\tau \to \infty$ is called the *non-informative* prior.

(b) As $\tau \to \infty$, the posterior distribution (34) and the bootstrap distribution (9) coincide. In Gaussian models, maximum likelihood and parametric bootstrap analyses tend to agree with Bayesian analysis that uses a non-informative prior for the free parameters.

(c) Notice that it was assumed that $\sigma^2$ is known. From a Bayesian perspective, this is *not* proper. A prior distribution on $\sigma$ should be imposed. A typical choice is $g(\sigma) \propto 1/\sigma$. Then, one calculates the posterior distribution of $\mu(x)$ and $\sigma$ and integrates out $\sigma$.

# III. Relationship Between the Bootstrap and Bayesian Inference

1. **A Simple Example:** Suppose we have $Z \sim \mathrm{Normal}(\theta, 1)$, and we specify the prior on $\theta$ as $\theta \sim \mathrm{Normal}(0, \tau)$, where $\tau > 0$. Then, the posterior distribution of $\theta$ is

$$\theta \,|\, Z = z \sim \mathrm{Normal}\left(\frac{z}{1 + 1/\tau}, \frac{1}{1 + 1/\tau}\right). \tag{35}$$

As $\tau \to \infty$, we have a *non-informative prior* and the posterior distribution becomes

$$\theta \,|\, Z = z \sim \mathrm{Normal}(z, 1), \tag{36}$$

which is the same as a parametric bootstrap distribution where we generate the bootstrap values $Z^*$ from the maximum likelihood estimate of the sampling density $\mathrm{Normal}(Z, 1)$.

*Remark.* There are *three* ingredients that make this correspondence work:

(i) We chose a non-informative prior for $\theta$ with infinite variance;

(ii) The dependence of the log-likelihood $\ell(\theta; Z)$ on the data $Z$ only through the maximum likelihood estimator $\hat{\theta}$, and therefore, we can write the log-likelihood as $\ell(\theta; Z) = \ell(\theta; \hat{\theta})$;

(iii) The symmetry of the log-likelihood in $\theta$ and $\hat{\theta}$, that is, $\ell(\theta; \hat{\theta}) = \ell(\hat{\theta}; \theta) + C$, where $C$ is a constant independent of $\theta$ and $\hat{\theta}$.

Note that (ii) and (iii) hold only for the Gaussian distribution.

**2. Another Example:** Consider a discrete sample space with $L$ categories. Let $w_j$ be the probability that a sample point falls in Category $j$, and $\hat{w}_j$ be the observed proportion in category $j$, for all $j = 1, \cdots, L$. Collectively, let

$$\mathbf{w} := (w_1, w_2, \cdots, w_L), \qquad \text{and} \qquad \widehat{\mathbf{w}} := (\hat{w}_1, \hat{w}_2, \cdots, \hat{w}_L). \tag{37}$$

We specify a prior distribution of $\mathbf{w}$ as the Dirichlet distribution with parameter $a$, that is,

$$\mathbf{w} \sim \mathrm{Dir}_L(a, \cdots, a), \tag{38}$$

and the probability density function is of the form

$$p(\mathbf{w}) \propto \prod_{\ell=1}^{L} w_\ell^{a-1}. \tag{39}$$

Furthermore, assume the data $\mathbf{x}$ are drawn from a multinomial distribution with parameters $n$ and $\mathbf{w}$, where $n$ is the total sample size and $\mathbf{w}$ is the category probability. The likelihood function is of the form

$$p(\mathbf{x} \,|\, \mathbf{w}) \propto \prod_{l=1}^{L} w_l^{x_l}. \tag{40}$$

Now, we derive the posterior distribution of $\mathbf{w}$:

$$\begin{aligned} p(\mathbf{w} \,|\, \mathbf{x}) &\propto p(\mathbf{x} \,|\, \mathbf{w}) \cdot p(\mathbf{w}) \\ &\propto \prod_{\ell=1}^{L} w_\ell^{x_\ell} \cdot \prod_{\ell=1}^{L} w_\ell^{a-1} \\ &= \prod_{\ell=1}^{L} w_\ell^{x_i+a-1}, \end{aligned}$$

and, therefore, the posterior distribution of $\mathbf{w}$ is

$$\mathrm{Dir}_L(a + x_1, \cdots, a + x_L) = \mathrm{Dir}_L(a + n\hat{w}_1, \cdots, a + n\hat{w}_L). \tag{41}$$

Letting $a \to 0$, we obtain a non-informative prior and the corresponding posterior distribution is $\mathrm{Dir}_L(n\widehat{w}_1, \cdots, n\widehat{w}_L)$.

Now, the bootstrap distribution can be expressed as sampling the category proportions from a multinomial distribution, that is,

$$n\widehat{\mathbf{w}}^* \sim \text{Mult}(n, \widehat{\mathbf{w}}), \tag{42}$$

where $\text{Mult}(n, \widehat{\mathbf{w}})$ denotes a multinomial distribution. This bootstrap distribution has the same support, the same mean, and nearly the same covariance matrix as the posterior distribution.

*Remark.* The bootstrap distribution represents an (approximate) nonparametric, non-informative posterior distribution for the parameter.

# IV. The Expectation-Maximization (EM) Algorithm

## IV.1 EM Algorithm for Two-component Gaussian Mixture Model

1. **Two-Component Gaussian Mixture Model:** We model the data as a mixture of two normal distributions as

$$\begin{aligned} Z_1 &\sim \text{Normal}(\mu_1, \sigma_1^2), \\ Z_2 &\sim \text{Normal}(\mu_2, \sigma_2^2), \\ Y &= (1-\Delta)\cdot Z_1 + \Delta \cdot Z_2, \end{aligned} \tag{43}$$

where $\Delta \in \{0, 1\}$ with $\mathbb{P}(\Delta = 1) = \pi$.

   (a) *Generating Process:* The generating process is the following:
      i. First generate a random variable $\Delta$ from $\{0, 1\}$ with probability $\pi$;
      ii. Depending on the outcome of $\Delta$, we sample either $Z_1$ or $Z_2$ from the corresponding distribution.

   (b) *Density Function of $Y$:* The probability density function of $Y$ is

$$g_Y(y) = (1-\pi)\cdot\phi(y; \mu_1, \sigma_1^2) + \pi\phi(y; \mu_2, \sigma_2^2), \qquad \text{for all } y \in \mathbb{R}, \tag{44}$$

   where $\phi(\,\cdot\,; \mu, \sigma^2)$ is the density function of a normal distribution with mean $\mu$ and variance $\sigma^2$.

2. **Goal:** Let $Y_1, Y_2, \cdots, Y_n$ be i.i.d samples from a two-component Gaussian mixture model. We wish to estimate the parameters $\boldsymbol{\theta} := (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)^\top$ by the method of maximum likelihood. The log-likelihood function based on $n$ samples is

$$\ell(\boldsymbol{\theta}; Y_1, Y_2, \cdots, Y_n) = \sum_{i=1}^{n} \log\big[(1-\pi)\cdot\phi(Y_i; \mu_1, \sigma_1^2) + \pi\cdot\phi(Y_i; \mu_2, \sigma_2^2)\big], \tag{45}$$

   which is numerically hard to optimize directly.

3. **Motivations of the EM Algorithm:**

(a) Observe that the unobserved variables $\{\Delta_i\}_{i=1}^n$ can only take values 0 or 1. Suppose we know the values of $\{\Delta_i\}_{i=1}^n$. The log-likelihood function becomes

$$\begin{aligned}
&\ell(\boldsymbol{\theta}; Y_1, \cdots, Y_n) \\
&= \log\left(\prod_{i=1}^n \big((1-\pi)\phi(Y_i;\mu_1,\sigma_1^2)\big)^{1-\Delta_i}\big(\pi\phi(Y_i;\mu_2,\sigma_2^2)\big)^{\Delta_i}\right) \\
&= \sum_{i=1}^n \big[(1-\Delta_i)\cdot\log\phi(Y_i;\mu_1,\sigma_1^2) + \Delta_i\cdot\log\phi(Y_i;\mu_2,\sigma_2^2)\big] \\
&\qquad + \sum_{i=1}^n \big[(1-\Delta_i)\log(1-\pi) + \Delta_i\log\pi\big].
\end{aligned}$$

Then,

i. the maximum likelihood estimators of $\mu_1$ and $\sigma_1^2$ are the sample mean and the variance for those data with $\Delta_i = 0$, respective,

ii. the maximum likelihood estimators of $\mu_2$ and $\sigma_2^2$ are the sample mean and the variance for those data with $\Delta_i = 1$, respectively, and

iii. the maximum likelihood estimator of $\pi$ is the sample proportion of $\Delta_i = 1$.

(b) However, we do *not* know $\Delta_i$'s. We proceed in an iterative fashion and substitute for each $\Delta_i$ with its expected value

$$\gamma_i(\boldsymbol{\theta}) := \mathbb{E}[\Delta_i \,|\, \boldsymbol{\theta}, Y_1, \cdots, Y_n] = \mathbb{P}(\Delta_i = 1 \,|\, \boldsymbol{\theta}, Y_1, \cdots, Y_n). \tag{46}$$

The $\gamma_i(\boldsymbol{\theta})$ is called the *responsibility* of Model 2 for the $i$-th observation.

**4. Main Idea of the EM Algorithm:**

(a) *Expectation(E)-step:* do a soft assignment of each observation to each model. That is, the current estimates of the parameters are used to assign probabilities according to the relative density of the training points under each model;

(b) *Maximization(M)-step:* the responsibilities obtained from the E-step are used in weighted maximum-likelihood fit to update the estimates of the parameters.

**5. Complete EM Algorithm:**

**Algorithm 1** EM Algorithm for Two-Component Gaussian Mixture

1: Take initial guesses for the parameters $\hat{\mu}_1$, $\hat{\sigma}_1^2$, $\hat{\mu}_2$, $\hat{\sigma}_2^2$, and $\hat{\pi}$.

2: *Expectation Step:* Compute the responsibilities

$$\hat{\gamma}_i = \frac{\hat{\pi}\phi(Y_i; \hat{\mu}_2, \hat{\sigma}_2^2)}{(1-\hat{\pi})\phi(Y_i; \hat{\mu}_1, \hat{\sigma}_1^2) + \hat{\pi}\phi(Y_i; \hat{\mu}_2, \hat{\sigma}_2^2)}, \qquad \text{for all } i = 1, \cdots, n.$$

3: *Maximization Step:* Compute the weighted means and variances

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n (1-\hat{\gamma}_i) Y_i}{\sum_{i=1}^n (1-\hat{\gamma}_i)}, \qquad \hat{\mu}_2 = \frac{\sum_{i=1}^n \hat{\gamma}_i Y_i}{\sum_{i=1}^b \hat{\gamma}_i},$$
$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^n (1-\hat{\gamma}_i)(Y_i - \hat{\mu}_1)^2}{\sum_{i=1}^n (1-\hat{\gamma}_i)}, \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^n \hat{\gamma}_i (Y_i - \hat{\mu}_2)^2}{\sum_{i=1}^n \hat{\gamma}_i},$$
$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^n \hat{\gamma}_i.$$

4: Iterate Steps 2 and 3 until convergence.

6. **EM Algorithm Computes a Local Maximum:**

   (a) Note that the maximizer of $\ell$ is achieved when a spike of infinite height is put any one data point, e.g., $\hat{\mu}_1 = Y_i$ for some $i$ and $\hat{\sigma}_1^2 = 0$ — such a solution is *not* useful;

   (b) We actually search for a good local maximum of the log-likelihood function for which $\hat{\sigma}_1^2 > 0$ and $\hat{\sigma}_2^2 > 0$. Note that there may be multiple local maxima.

## IV.2 EM Algorithm in General

1. **Motivation:** In some problems, maximizing the (log-)likelihood function directly is hard; but it is easier to enlarge the sample with latent/unobserved/missimg data and maximize a version of (log-)likelihood function with the augmented data.

2. **Notation:** Let

   - $\mathbf{Z}_{\text{obs}}$ be the *observed* data,
   - $\ell(\boldsymbol{\theta}; \mathbf{Z}_{\text{obs}}) := \log f(\mathbf{Z}_{\text{obs}} \,|\, \boldsymbol{\theta})$ be the log-likelihood function of $\boldsymbol{\theta}$ that depends on the observed data $\mathbf{Z}_{\text{obs}}$ and the parameter $\boldsymbol{\theta}$,
   - $\mathbf{Z}_{\text{mis}}$ be the *unobserved/missing* data,
   - $\mathbf{Z} := (\mathbf{Z}_{\text{obs}}, \mathbf{Z}_{\text{mis}})$ be the *complete* data, and
   - $\ell_0(\boldsymbol{\theta}; \mathbf{Z}) := \log f_0(\mathbf{Z}; \boldsymbol{\theta})$ be the log-likelihood function the $\boldsymbol{\theta}$ based on the complete data $\mathbf{Z}$.

**3. EM Algorithm:**

**Algorithm 2** EM Algorithm for General Case

1: Take initial guesses for the parameters $\widehat{\boldsymbol{\theta}}^{(0)}$;
2: *Expectation(E-) Step:* At the $j$-th step, compute

$$Q(\boldsymbol{\theta}', \widehat{\boldsymbol{\theta}}^{(j)}) := \mathbb{E}[\ell_0(\boldsymbol{\theta}'; \mathbf{Z}) \,|\, \mathbf{Z}_{\text{obs}}, \widehat{\boldsymbol{\theta}}^{(j)}]$$

as a function of the dummy argument $\boldsymbol{\theta}'$;
3: *Maximization(M-) Step:* Determine the new estimate $\widehat{\boldsymbol{\theta}}^{(j+1)}$ as the maximizer of $Q(\boldsymbol{\theta}', \widehat{\boldsymbol{\theta}}^{(j)})$ over $\boldsymbol{\theta}'$;
4: Iterate Steps 2 and 3 until convergence.

**4. Rationale of EM Algorithm:** First note that

$$f_1(\mathbf{Z}_{\text{mis}} \,|\, \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}') = \frac{f_0(\mathbf{Z}_{\text{mis}}, \mathbf{Z}_{\text{obs}} \,|\, \boldsymbol{\theta}')}{f(\mathbf{Z}_{\text{obs}} \,|\, \boldsymbol{\theta}')}. \tag{47}$$

Hence,

$$f(\mathbf{Z}_{\text{obs}} \,|\, \boldsymbol{\theta}') = \frac{f_0(\mathbf{Z}_{\text{mis}}, \mathbf{Z}_{\text{obs}} \,|\, \boldsymbol{\theta}')}{f_1(\mathbf{Z}_{\text{mis}} \,|\, \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}')} = \frac{f_0(\mathbf{Z} \,|\, \boldsymbol{\theta}')}{f_1(\mathbf{Z}_{\text{mis}} \,|\, \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}')}. \tag{48}$$

In terms of the log-likelihood function, we have

$$\ell(\boldsymbol{\theta}'; \mathbf{Z}_{\text{obs}}) = \ell_0(\boldsymbol{\theta}'; \mathbf{Z}) - \ell_1(\boldsymbol{\theta}'; \mathbf{Z}_{\text{mis}} \,|\, \mathbf{Z}_{\text{obs}}), \tag{49}$$

where $\ell_1$ is the log-likelihood function based on the conditional density $f_1(\cdot \,|\, \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}')$. Now, taking the expectations of both sides of (49) with respect to the distribution of $\mathbf{Z} \,|\, \mathbf{Z}_{\text{obs}}$ that depends on $\boldsymbol{\theta}$ gives

$$\begin{aligned} \ell(\boldsymbol{\theta}'; \mathbf{Z}_{\text{obs}}) &= \mathbb{E}[\ell_0(\boldsymbol{\theta}'; \mathbf{Z}) \,|\, \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}] - \mathbb{E}[\ell_1(\boldsymbol{\theta}'; \mathbf{Z} \,|\, \mathbf{Z}_{\text{obs}}) \,|\, \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}] \\ &= Q(\boldsymbol{\theta}', \boldsymbol{\theta}) - R(\boldsymbol{\theta}', \boldsymbol{\theta}), \end{aligned}$$

where

$$Q(\boldsymbol{\theta}', \boldsymbol{\theta}) := \mathbb{E}[\ell_0(\boldsymbol{\theta}'; \mathbf{Z}) \,|\, \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}],$$

and

$$R(\boldsymbol{\theta}', \boldsymbol{\theta}) := \mathbb{E}[\ell_1(\boldsymbol{\theta}'; \mathbf{Z} \,|\, \mathbf{Z}_{\text{obs}}) \,|\, \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}].$$

Analysis of $Q(\boldsymbol{\theta}', \boldsymbol{\theta})$ and $R(\boldsymbol{\theta}', \boldsymbol{\theta})$:

(a) Analysis of $R(\boldsymbol{\theta}', \boldsymbol{\theta})$: Note that $R(\boldsymbol{\theta}', \boldsymbol{\theta})$ is the expectation of a log-likelihood function of a density function (indexed by $\boldsymbol{\theta}'$) with respect to the same density indexed by $\boldsymbol{\theta}$. It is maximized as a function of $\boldsymbol{\theta}'$ when $\boldsymbol{\theta}' = \boldsymbol{\theta}$, by Jensen's inequality;

(b) Analysis of $Q(\boldsymbol{\theta}', \boldsymbol{\theta})$: In the M-step, we maximize $Q(\boldsymbol{\theta}', \boldsymbol{\theta})$ over $\boldsymbol{\theta}'$. Hence, if $\boldsymbol{\theta}'$ maximizes $Q(\boldsymbol{\theta}', \boldsymbol{\theta})$, we have

$$\ell(\boldsymbol{\theta}'; \mathbf{Z}) - \ell(\boldsymbol{\theta}; \mathbf{Z}) = [Q(\boldsymbol{\theta}', \boldsymbol{\theta}) - Q(\boldsymbol{\theta}, \boldsymbol{\theta})] - [R(\boldsymbol{\theta}', \boldsymbol{\theta}) - R(\boldsymbol{\theta}, \boldsymbol{\theta})] \geq 0,$$

where the first term is nonnegative and the second term is nonpositive.

Thus, the EM iteration never decreases the value of the log-likelihood function.

*Remark.* In the M-step, a full maximization is *not* necessary. We only need to find a value $\widehat{\boldsymbol{\theta}}^{(j+1)}$ so that $Q(\boldsymbol{\theta}', \widehat{\boldsymbol{\theta}}^{(j)})$ increases as a function of the first argument, i.e.,

$$Q(\widehat{\boldsymbol{\theta}}^{(j+1)}, \widehat{\boldsymbol{\theta}}^{(j)}) > Q(\widehat{\boldsymbol{\theta}}^{(j)}, \widehat{\boldsymbol{\theta}}^{(j)}).$$

The resulting algorithm is called the *generalized EM algorithm*.

**5. EM Algorithm as a Minorization Procedure:**

(a) *Basic Definition and Motivation:* A function $g(x, y)$ is said to *minorize* a function $f(x)$ if

$$g(x, y) \leq f(x), \qquad g(x, x) = f(x), \qquad \text{for all } x, y \text{ in the domain.}$$

This is useful for maximizing $f$ since $f$ is non-decreasing under the update

$$x^{(j+1)} = \arg\max_x g(x, x^{(j)}).$$

To see this, we note that

$$f(x^{(j)}) \overset{\text{(i)}}{=} g(x^{(j)}, x^{(j)}) \overset{\text{(ii)}}{\leq} g(x^{(j+1)}, x^{(j)}) \overset{\text{(iii)}}{\leq} f(x^{(j+1)}),$$

where (i) and (iii) follow from the fact that $g$ minorizes $f$, and (ii) is because $x^{(j+1)}$ maximizes the function $x \mapsto g(x, x^{(j)})$.

(b) *MM Algorithm:* The algorithm of maximizing $f$ via the minorization function $g$ is known as the *MM algorithms*, for "Minorize-Maximize". Details are given in Algorithm 3.

**Algorithm 3** MM Algorithm

**Require:** $x^{(0)}$ in the domain.

1: **for** $j = 0, 1, 2, \cdots$ **do**

2: *Minorization Step:* Compute $g(x, x^{(j)})$;

3: *Maximization Step:* Maximization $g$ with respect to the first argument

$$x^{(j+1)} := \arg\max_x g(x, x^{(j)});$$

4: **end for**

(c) *Connection to EM Algorithm:* We show that the EM algorithm is an example of the MM algorithms. To start with, we show

$$S(\boldsymbol{\theta}', \boldsymbol{\theta}) := Q(\boldsymbol{\theta}', \boldsymbol{\theta}) + \log f(\mathbf{Z}_{\text{obs}} \,|\, \boldsymbol{\theta}) - Q(\boldsymbol{\theta}, \boldsymbol{\theta})$$

minorizes the log-likelihood function of the observed data $\ell(\boldsymbol{\theta}'; \mathbf{Z}_{\text{obs}})$.
Note that $S(\boldsymbol{\theta}', \boldsymbol{\theta}') = \ell(\boldsymbol{\theta}'; \mathbf{Z}_{\text{obs}})$, since

$$S(\boldsymbol{\theta}', \boldsymbol{\theta}') = Q(\boldsymbol{\theta}', \boldsymbol{\theta}') + \log f(\mathbf{Z}_{\text{obs}} \,|\, \boldsymbol{\theta}') - Q(\boldsymbol{\theta}', \boldsymbol{\theta}') = \log f(\mathbf{Z}_{\text{obs}} \,|\, \boldsymbol{\theta}') = \ell(\boldsymbol{\theta}'; \mathbf{Z}_{\text{obs}}).$$

Also, we have $S(\boldsymbol{\theta}', \boldsymbol{\theta}) \le \ell(\boldsymbol{\theta}'; \mathbf{Z}_{\text{obs}})$ for all $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$, since

$$\begin{aligned} S(\boldsymbol{\theta}', \boldsymbol{\theta}) - \ell(\boldsymbol{\theta}'; \mathbf{Z}_{\text{obs}}) &= Q(\boldsymbol{\theta}', \boldsymbol{\theta}) + \ell(\boldsymbol{\theta}; \mathbf{Z}_{\text{obs}}) - Q(\boldsymbol{\theta}, \boldsymbol{\theta}) - \ell(\boldsymbol{\theta}'; \mathbf{Z}_{\text{obs}}) \\ &= [Q(\boldsymbol{\theta}', \boldsymbol{\theta}) - \ell(\boldsymbol{\theta}'; \mathbf{Z}_{\text{obs}})] - [Q(\boldsymbol{\theta}, \boldsymbol{\theta}) - \ell(\boldsymbol{\theta}; \mathbf{Z}_{\text{obs}})] \\ &= R(\boldsymbol{\theta}', \boldsymbol{\theta}) - R(\boldsymbol{\theta}, \boldsymbol{\theta}) \\ &\le 0, \end{aligned}$$

by Jensen's inequality, using the notation defined earlier.
Hence, the function $S(\boldsymbol{\theta}', \boldsymbol{\theta})$ minorizes $\ell(\boldsymbol{\theta}'; \mathbf{Z}_{\text{obs}})$. By maximizing $S(\boldsymbol{\theta}', \boldsymbol{\theta})$ in the first argument sequentially, we can increase the value of the log-likelihood function at the observed data steadily, which achieves the goal of maximizing $\ell(\,\cdot\,; \mathbf{Z}_{\text{ob}})$.

**6. EM Algorithm as a Maximization-Maximization Procedure:**

(a) *Derivation:* Consider the function

$$G(\boldsymbol{\theta}', \tilde{f}) := \mathbb{E}_{\tilde{f}}[\ell_0(\boldsymbol{\theta}'; \mathbf{Z})] - \mathbb{E}_{\tilde{f}}[\log \tilde{f}(\mathbf{Z}_{\text{mis}})], \tag{50}$$

where $\tilde{f}$ is any density function over the latent/missing data $\mathbf{Z}_{\text{mis}}$.
Then, we have

$$\begin{aligned} G(\boldsymbol{\theta}', \tilde{f}) &= \int \tilde{f}(\mathbf{z}_{\text{mis}}) \log f_0(\mathbf{Z}_{\text{obs}}, \mathbf{z}_{\text{mis}} \,|\, \boldsymbol{\theta}') \mathrm{d}\mathbf{z}_{\text{mis}} - \int \tilde{f}(\mathbf{z}_{\text{mis}}) \log \tilde{f}(\mathbf{z}_{\text{mis}}) \mathrm{d}\mathbf{z}_{\text{mis}} \\ &= -\int \tilde{f}(\mathbf{z}_{\text{mis}}) \log \frac{\tilde{f}(\mathbf{z}_{\text{mis}})}{f_0(\mathbf{Z}_{\text{obs}}, \mathbf{z}_{\text{mis}} \,|\, \boldsymbol{\theta}')} \mathrm{d}\mathbf{z}_{\text{mis}} \\ &= -\int \tilde{f}(\mathbf{z}_{\text{mis}}) \log \frac{\tilde{f}(\mathbf{z}_{\text{mis}})}{f_0(\mathbf{Z}_{\text{obs}}, \mathbf{z}_{\text{mis}} \,|\, \boldsymbol{\theta}') \cdot \frac{f(\mathbf{Z}_{\text{obs}} \,|\, \boldsymbol{\theta}')}{f(\mathbf{Z}_{\text{obs}} \,|\, \boldsymbol{\theta}')}} \mathrm{d}\mathbf{z}_{\text{mis}} \\ &= -\int \tilde{f}(\mathbf{z}_{\text{mis}}) \log \frac{\tilde{f}(\mathbf{z}_{\text{mis}})}{f_1(\mathbf{z}_{\text{mis}} \,|\, \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}') f(\mathbf{Z}_{\text{obs}} \,|\, \boldsymbol{\theta}')} \mathrm{d}\mathbf{z}_{\text{mis}} \\ &= -\mathrm{KL}(\tilde{f} \,\|\, f_1(\,\cdot \,|\, \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}')) + \log f(\mathbf{Z}_{\text{obs}} \,|\, \boldsymbol{\theta}'), \end{aligned}$$

where $\mathrm{KL}(p \,\|\, q) = \int p(x) \log(p(x)/q(x)) \mathrm{d}x$ is the Kullback-Leibler divergence between $p$ and $q$, is always nonnegative, and is minimized when $p = q$.
Therefore, with a fixed value $\boldsymbol{\theta}'$, if we let $\tilde{f} = f_1(\,\cdot \,|\, \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}')$, $G(\boldsymbol{\theta}, \,\cdot\,)$ is maximized in the second argument, and

$$G(\boldsymbol{\theta}, f_1(\,\cdot \,|\, \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}')) = \log f(\mathbf{Z}_{\text{obs}} \,|\, \boldsymbol{\theta}'),$$

which is exactly the observed data log-likelihood. Maximizing this second argument of $G$ is equivalent to the E-step in Algorithm 2.

(b) *EM Algorithm as a Maximization-Maximization Procedure:* The EM algorithm can be viewed as a joint maximization method for $G$ over $\boldsymbol{\theta}'$ and $\tilde{f}$, by fixing one argument and maximizing over the other. More precisely,

i. the maximizer over $\tilde{f}$ for a fixed $\boldsymbol{\theta}'$ is $f_1(\cdot \mid \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}')$, which is the distribution computed by the E-step;

ii. in the M-step, we maximize $G(\boldsymbol{\theta}', \tilde{f})$ over $\boldsymbol{\theta}'$ with $\tilde{f} = f_1(\cdot \mid \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta})$ fixed: this is the same as maximizing the first term

$$\mathbb{E}_{f_1(\cdot \mid \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta})}[\ell_0(\boldsymbol{\theta}'; \mathbf{Z})] = \mathbb{E}[\ell_0(\boldsymbol{\theta}'; \mathbf{Z}) \mid \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta}],$$

since the second term in $G$ does *not* involve $\boldsymbol{\theta}'$.

# V. MCMC for Sampling from the Posterior

1. **Goal:** With a Bayesian model, one would like to draw samples from the posterior distribution to make inferences about the parameters.

2. **Gibbs Sampling:** Suppose we have random variables $U_1, U_2, \cdots, U_K$ and we want to draw a sample from their joint distribution. Suppose drawing from the joint distribution is difficult but drawing from the conditional distribution

$$\mathbb{P}(U_j \mid U_1, \cdots, U_{j-1}, U_{j+1}, \cdots, U_K), \qquad \text{for all } j = 1, \cdots, K,$$

is easy. The *Gibbs sampling procedure* alternatively samples from each of these conditional distributions and, when the process stabilizes, provides a sample from the desired joint distribution.

**Algorithm 4** Gibbs Sampling

1: Take initial values for $U_k^{(0)}$, for all $k = 1, \cdots, K$;
2: Repeat for $t = 1, 2, \cdots$:
For $k = 1, \cdots, K$, generate $U_k^{(t)}$ from

$$\mathbb{P}\big(U_k^{(t)} \mid U_1^{(t-1)}, U_2^{(t-1)}, \cdots, U_{k-1}^{(t-1)}, U_{k+1}^{(t-1)}, \cdots, U_K^{(t-1)}\big);$$

3: Continue Step 2 until the joint distribution of $(U_1^{(t)}, U_2^{(t)}, \cdots, U_K^{(t)})$ does *not* change.

*Remarks.*

(a) Under regularity conditions, it can be shown that this procedure eventually stabilizes, and the resulting random variables are indeed a sample from the joint distribution of $(U_1, U_2, \cdots, U_K)$. This occurs despite the fact that the samples $(U_1^{(t)}, U_2^{(t)}, \cdots, U_K^{(t)})$ are *not* independent for different values of $t$.

(b) Gibbs sampling produces a *Markov chain* whose stationary distribution is the true joint distribution, and hence the term "Markov chain Monte Carlo".

(c) We do *not* need to know the explicit form of the conditional densities, but just need to be able to sample from them.

(d) If the explicit form of the conditional density $\mathbb{P}(U_k \,|\, U_l, l \neq k)$ is available, we can estimate the marginal density of $U_k$ by

$$\widehat{\mathbb{P}}_{U_k}(u) = \frac{1}{M-m+1} \sum_{t=m}^{M} \mathbb{P}(u \,|\, U_l^{(t)}, l \neq k), \tag{51}$$

where we average over the last $M - m + 1$ members of the sequence to allow for an initial "burn-in" period before stationarity is reached.

**3. Connection between Gibbs Sampling and EM Algorithm:** We focus on the EM algorithm for the two-component Gaussian mixture model.

(a) *Setup:*

i. Treat the latent data $\mathbf{W}$ from the EM algorithm to be another parameter for the Gibbs sampler;

ii. We fix the variances $\sigma_1^2$ and $\sigma_2^2$ and the mixing proportion $\pi$ at their maximum likelihood values, denoted by $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, and $\hat{\pi}$, respectively, so that the only unknown parameters in the model are the mean parameters $\mu_1$ and $\mu_2$.

(b) *Gibbs sampling procedure:*

**Algorithm 5** Gibbs Sampling for Two-component Gaussian Mixture Model

1: Take initial values for $\mu_1^{(0)}$ and $\mu_2^{(0)}$;

2: Repeat for $t = 1, 2, \cdots$:

i. For $i = 1, \cdots, n$, generate $\Delta_i^{(t)} \in \{0, 1\}$ according to the following responsibilities

$$\mathbb{P}(\Delta_i^{(t)} = 1) = \hat{\gamma}_i(\mu_1^{(t-1)}, \mu_2^{(t-1)}), \qquad \text{for all } i = 1, 2, \cdots, n;$$

ii. Set

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n (1 - \Delta_i^{(t)}) Y_i}{\sum_{i=1}^n (1 - \Delta_i^{(t)})}, \quad \text{and} \quad \hat{\mu}_2 = \frac{\sum_{i=1}^n \Delta_i^{(t)} Y_i}{\sum_{i=1}^n \Delta_i^{(t)}},$$

and generate $\mu_1^{(t)} \sim \text{Normal}(\hat{\mu}_1, \hat{\sigma}_1^2)$ and $\mu_2^{(t)} \sim \text{Normal}(\hat{\mu}_2, \hat{\sigma}_2^2)$;

3: Continue Step 2 until the joint distribution of $(\Delta_1^{(t)}, \cdots, \Delta_n^{(t)}, \mu_1^{(t)}, \mu_2^{(t)})$ does *not* change.

(c) *Connection:* Note Steps 2(i) and 2(ii) in Algorithm 5 are the same as the E and M steps of the EM algorithm, except that we sample rather than maximize:

i. In Step 2(i), rather than compute the maximum likelihood responsibilities $\gamma_i := \mathbb{E}[\Delta_i \,|\, \boldsymbol{\theta}, \mathbf{Z}]$, the Gibbs sampling procedure simulates the latent data $\Delta_i$ from the distributions $\mathbb{P}(\Delta_i \,|\, \boldsymbol{\theta}, \mathbf{Z})$;

ii. In step 2(ii), rather than compute the maximizers of the posterior distribution $\mathbb{P}(\mu_1, \mu_2, \Delta_1, \cdots, \Delta_n \,|\, \mathbf{Z})$, we simulate from the conditional distribution $\mathbb{P}(\mu_1, \mu_2 \,|\, \Delta_1, \cdots, \Delta_n, \mathbf{Z})$.

(d) *Remarks:* The above mixture model was *simplified*. More realistically, one would

i. put a prior distribution on the variances $\sigma_1^2$ and $\sigma_2^2$ and the mixing proportion $\pi$, and

ii. include separate Gibbs sampling steps in which we sample from their posterior distributions, conditional on the other parameters.

# VI. Bagging

1. **Setup:** Let the training data be $\mathbf{Z} := \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where we assume $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, for all $i = 1, 2, \cdots, n$, and consider a regression problem. Suppose $y_i = f(\mathbf{x}_i) + \varepsilon_i$ for all $i = 1, \cdots, n$, where $\varepsilon_i \overset{\text{i.i.d}}{\sim} \text{Normal}(0, \sigma^2)$.

2. **Goal:** The goal is to predict $f(\mathbf{x}_0)$ at some $\mathbf{x}_0 \in \mathbb{R}^p$.

3. **Bagging for Regression:** *Bootstrap aggregation*, or *bagging*, averages the predictions over a collection of bootstrap samples, and can reduce the *variance* of the predictions.

(a) *Procedure:*

i. Draw bootstrap samples from $\mathbf{Z}$, denoted by $\mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(B)}$;

ii. Fit a model on each $\mathbf{Z}^{(b)}$, denoted by $\hat{f}^{(b)}$, and obtain $\hat{f}^{(b)}(\mathbf{x}_0)$, for all $b = 1, \cdots, B$;

iii. The *bagging estimate* at $\mathbf{x}_0$ is

$$\hat{f}_{\text{bag}}(\mathbf{x}_0) = \frac{1}{B} \sum_{n=1}^{B} \hat{f}^{(b)}(\mathbf{x}_0). \tag{52}$$

(b) *A different view:* Let $\widehat{F}$ be the empirical distribution putting equal probability $1/n$ on each data point $(\mathbf{x}_i, y_i)$. The "true" bagging estimate is defined by

$$\mathbb{E}_{\widehat{F}}[\hat{f}^*(\mathbf{x}_0)], \tag{53}$$

where $\mathbf{Z}^* := \{(\mathbf{x}_i^*, y_i^*)\}_{i=1}^n$, each $(\mathbf{x}_i^*, y_i^*) \overset{\text{i.i.d}}{\sim} \widehat{F}$, and $\hat{f}^*$ is the model fit using $\mathbf{Z}^*$. Then, (52) is a Monte Carlo estimate of the true bagging estimate, approaching it as $B \to \infty$.

4. **Bagging for $W$-class Classification:**

(a) *Goal:* To predict the class label of an observation $\mathbf{x}_0 \in \mathbb{R}^p$.

(b) *Procedure:* Suppose for each model fit on each bootstrap sample $\mathbf{Z}^{(b)}$, denoted by $\hat{f}^{(b)}: \mathbb{R}^p \to \mathbb{R}^W$, we obtain a classifier, i.e.,

$$\widehat{G}^{(b)}(\mathbf{x}_0) := \underset{w=1,2,\cdots,W}{\arg\max}\ \hat{f}^{(b)}(\mathbf{x}_0).$$

The bagged estimate $\hat{f}_{\text{bag}}(\mathbf{x}_0)$ is a $W$-dimensional vector, $(p_1(\mathbf{x}_0), p_2(\mathbf{x}_0), \cdots, p_W(\mathbf{x}_0))^\top \in \mathbb{R}^W$, with $p_w(\mathbf{x}_0)$ being the proportion of the number of models predicting Class $w$ at $\mathbf{x}_0$. The *bagged classifier* selects the class with the most "votes" from the $B$ models, i.e.,

$$\begin{aligned}\widehat{G}_{\text{bag}}(\mathbf{x}_0) &= \underset{w=1,2,\cdots,W}{\arg\max}\ \hat{f}_{\text{bag}}(\mathbf{x}_0) \\ &= \underset{w=1,2,\cdots,W}{\arg\max}\ (p_1(\mathbf{x}), p_2(\mathbf{x}), \cdots, p_W(\mathbf{x}))^\top.\end{aligned}$$

(c) *An Alternative Procedure:* In the cases where we require the class-probability estimates at $\mathbf{x}_0$, we assume the model $\hat{f}^{(b)}(\mathbf{x}_0)$ can produce the $W$-dimensional vector with each component being the predicted class-probability of the corresponding class at $\mathbf{x}_0$. We can then average these $B$ $W$-dimensional vectors, yielding the bagged estimate of the class-probability.

5. **Benefits of Bagging:** Bagging can dramatically reduce the *variance* of unstable procedures, leading to improved predictions — this is because averaging can reduce the variance and leave bias unchanged.

   We focus on the mean square error (MSE). Assume the training observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are independently drawn from a distribution $\mathbb{P}$, and consider the ideal aggregate estimator $f_{\text{ag}}(\mathbf{x}) := \mathbb{E}_{\mathbb{P}}[\hat{f}^*(\mathbf{x})]$. Here, $\mathbf{x}$ is fixed and the bootstrap dataset $\mathbf{Z}^*$ consists of observations $(\mathbf{x}_i^*, y_i^*)$, for all $i = 1, 2, \cdots, n$ sampled from $\mathbb{P}$ [1]. We have

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}[(Y - \hat{f}^*(\mathbf{x}))^2] &= \mathbb{E}_{\mathbb{P}}[(Y - f_{\text{ag}}(\mathbf{x}) + f_{\text{ag}}(\mathbf{x}) - \hat{f}^*(\mathbf{x}))^2] \\ &= \mathbb{E}_{\mathbb{P}}[(Y - f_{\text{ag}}(\mathbf{x}))^2] + \mathbb{E}_{\mathbb{P}}[(f_{\text{ag}}(\mathbf{x}) - \hat{f}^*(\mathbf{x}))^2] \\ &\geq \mathbb{E}_{\mathbb{P}}[(Y - f_{\text{ag}}(\mathbf{x}))^2]. \qquad (54)\end{aligned}$$

   Note that the term $\mathbb{E}_{\mathbb{P}}[(f_{\text{ag}}(\mathbf{x}) - \hat{f}^*(\mathbf{x}))^2]$ comes from the *variance* of $\hat{f}^*(\mathbf{x})$ around its mean $f_{\text{ag}}(\mathbf{x})$. Therefore, true population aggregation never increases MSE and bagging (which draws samples from data) will often decrease mean-squared error.

   *Remark.* The argument above does *not* hold for classification under 0-1 loss, due to the non-additivity of bias and variance. In this setting, bagging a good classifier can make it better, but bagging a bad classifier can make it worse.

6. **Drawback of Bagging:** Note that when we bag a model, any simple structure in the model is *lost*. Thus, the results from bagging are *not* easy to interpret.

[1] Note that $f_{\text{ag}}(\mathbf{x})$ is a bagging estimate, drawing bootstrap samples from the *actual population* $\mathbb{P}$ rather than the data. It is *not* an estimate that we can use in practice, but is convenient for analysis.

# VII. Model Averaging and Stacking

1. **Setup:** Suppose we have a set of candidate models, $\{\mathcal{M}_1, \cdots, \mathcal{M}_M\}$, for the training data $\mathbf{Z}$.

2. **Bayesian Model Averaging:** Suppose the quantity $\zeta$ is of interest. The posterior distribution of $\zeta$ is

$$\mathbb{P}(\zeta \mid \mathbf{Z}) = \sum_{m=1}^{M} \mathbb{P}(\zeta \mid \mathcal{M}_m, \mathbf{Z}) \mathbb{P}(\mathcal{M}_m \mid \mathbf{Z}), \tag{55}$$

   and the posterior mean is

$$\mathbb{E}[\zeta \mid \mathbf{Z}] = \sum_{m=1}^{M} \mathbb{E}[\zeta \mid \mathcal{M}_m, \mathbf{Z}] \mathbb{P}(\mathcal{M}_m \mid \mathbf{Z}). \tag{56}$$

   Thus, the Bayesian estimate of $\zeta$ is a weighted average of the individual predictions, with the weights proportional to the posterior probability of each model.

3. **Committee Method:** *Committee methods* take a simple *unweighted* average of the predictions from each model, essentially giving *equal* probability to each model.

4. **Frequentist Model Averaging:** Suppose we have predictions $\hat{f}_1(\mathbf{x}), \hat{f}_2(\mathbf{x}), \cdots, \hat{f}_M(\mathbf{x})$. To average them, we can seek the weights $\mathbf{w} := (w_1, w_2, \cdots, w_M)^\top$ to minimize the squared-error loss, i.e.,

$$\widehat{\mathbf{w}} := \arg\min_{\mathbf{w}} \left\{ \mathbb{E}_{\mathbb{P}} \left[ \left( Y - \sum_{m=1}^{M} w_m \hat{f}_m(\mathbf{x}) \right)^2 \right] \right\}. \tag{57}$$

   Here, $\mathbf{x}$ is fixed and the $n$ observations in $\mathbf{Z}$ are distributed according to $\mathbb{P}$.

   The solution to (57) is

$$\widehat{\mathbf{w}} = \mathbb{E}_{\mathbb{P}}[\hat{\mathbf{f}}(\mathbf{x}) \hat{\mathbf{f}}(\mathbf{x})^\top]^{-1} \mathbb{E}_{\mathbb{P}}[\hat{\mathbf{f}}(\mathbf{x}) Y], \tag{58}$$

   where $\hat{\mathbf{f}}(\mathbf{x}) := (\hat{f}_1(\mathbf{x}), \hat{f}_2(\mathbf{x}), \cdots, \hat{f}_M(\mathbf{x}))^\top$.

   It follows that the full regression has smaller error than any single model

$$\mathbb{E}_{\mathbb{P}} \left[ \left( Y - \sum_{m=1}^{M} \hat{w}_m \hat{f}_m(\mathbf{x}) \right)^2 \right] \leq \mathbb{E}_{\mathbb{P}} \left[ \left( Y - \hat{f}_m(\mathbf{x}) \right)^2 \right], \qquad \text{for all } m = 1, \cdots, M,$$

   so combining models *never* makes things worse, at the population level.

   *Remarks.*

   - This approach can not be used in practice, as $\widehat{\mathbf{w}}$ depends on the distribution $\mathbb{P}$, which is typically unknown.

- This approach does *not* take the *model complexity* into consideration.

5. **Stacked Generalization (Stacking):** Let $\hat{f}_m^{(-i)}(\mathbf{x})$ be the prediction at $\mathbf{x}$, using Model $m$, applied to the dataset with the $i$-th training observation removed. The *stacking estimate* of the weights is obtained from the least squares linear regression of $y_i$ on $\hat{f}_m^{(-i)}(\mathbf{x})$, for all $m = 1, 2, \cdots, M$:

$$\widehat{\mathbf{w}}^{\text{stacking}} := \arg\min_{\mathbf{w}} \sum_{i=1}^{n} \left[ y_i - \sum_{m=1}^{M} w_m \hat{f}_m^{(-i)}(\mathbf{x}_i) \right]^2, \tag{59}$$

The final prediction is $\sum_{m=1}^{M} \hat{w}_m^{\text{stacking}} \hat{f}_m(\mathbf{x})$. Note that this $\mathbf{x}$ may be a new data point not present in the dataset to fit the $M$ models.

*Remarks.*

(a) Benefit: By using the cross-validated predictions $\hat{f}_m^{(-i)}(\mathbf{x})$, stacking avoids giving unfairly high weight to models with higher complexity.

(b) Better results can be obtained by restricting the weights to be nonnegative, and to sum to 1.

6. **Connection between Stacking and Leave-one-out Cross-validation:** If we restrict the minimization in (59) to weight vectors $\mathbf{w}$ that have *one* unit weight and the rest zero, this leads to a model choice with the smallest leave-one-out cross-validation error.

7. **Final Remarks on Stacking:**

(a) Stacking combines multiple models with the estimated optimal weights. This will often lead to better prediction, but less interpretability than the choice of only one of the $M$ models.

(b) The stacking idea is actually more general than described above:

i. One can use *any* learning method, not just linear regression, to combine the models;

ii. the weights could depend on the input location $\mathbf{x}$.

In this way, learning methods are "stacked" on top of one another, to improve prediction performance.

# VIII. Stochastic Search: Bumping

1. **Bumping:**

(a) *Main Idea: Bumping* uses bootstrap sampling to move randomly through the model space.

(b) *Benefit:* Bumping is helpful in problems where fitting methods finds many local minima, and can help avoid getting stuck in poor solutions.

(c) *Procedure:*

i. Draw bootstrap samples $\mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(B)}$, and fit a model to each, giving predictions $\hat{f}^{(b)}(\mathbf{x})$, for all $b = 1, \cdots, B$, at the input point $\mathbf{x}$;

ii. Choose the model that produces the smallest prediction error, averaged over the original *training set*.

(d) *Example:* Suppose we use the squared-error loss function. Choose the model obtained from bootstrap sample $\hat{b}$, where

$$\hat{b} := \underset{b=1,\cdots,B}{\arg\min} \left\{ \sum_{i=1}^{n} [y_i - \hat{f}^{(b)}(\mathbf{x}_i)]^2 \right\}. \tag{60}$$

The corresponding model predictions are $\hat{f}^{(\hat{b})}(\mathbf{x})$.

(e) *A Remark.* We include the original training sample in the set of bootstrap samples, so that the method is free to pick the original model if it has the lowest training error.

2. **Caveat:** Since bumping compares different models on the training data, one must ensure that the models have roughly the *same complexity*.

3. **Other Applications:** Bumping can help in problems where it is *difficult* to optimize the fitting criterion (perhaps because of a lack of smoothness). The trick is to optimize a different, more convenient criterion over the bootstrap samples, and then choose the model producing the best results for the *desired criterion* on the training sample.

4. **Comparison with Bagging:** Bagging *averages* the outputs from different models, but bumping only picks the *best* one.

# References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Vol. 1. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Izenman, Alan J (Mar. 2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. en. Springer Science & Business Media.