COMP 7070 Advanced Topics in AI and ML January 15, 2025

Lecture 1: Logistics and Preliminaries

Instructor: Yifan Chen *Scribes: Yujia Yin* *Proof reader: Yifan Chen, Xiong Peng*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

1.1 Gaussian Distribution

The Gaussian distribution, also known as the normal distribution, is a continuous probability distribution that is symmetric around its mean. It is characterized by its mean ($\boldsymbol{\mu}$) and covariance matrix ($\boldsymbol{\Sigma}$). In this section, we study the matrix form of the distribution for the random variable $\boldsymbol{X} \in \mathbb{R}^n$ as follows:

$$p(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^n} \cdot \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu})\right),$$

here, $\boldsymbol{\mu}$ is the mean vector of size $n \times 1$, and $\boldsymbol{\Sigma}$ is the covariance matrix, which describes the shape and orientation of the distribution. The term $|\boldsymbol{\Sigma}|$ represents the determinant of the covariance matrix.

1.1.1 T-test

In a t-test, we consider two independent groups of observations. Suppose each group consists of n observations, and let $\boldsymbol{X}_n^y$ and $\boldsymbol{X}_n^b$ be two independent random vectors, following a multivariate normal distribution:

$$\boldsymbol{X}_n^y \sim N(\boldsymbol{\mu}_y, \sigma_y^2 \boldsymbol{I}), \quad \boldsymbol{X}_n^b \sim N(\boldsymbol{\mu}_b, \sigma_b^2 \boldsymbol{I}),$$

where $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_b$ are both mean vectors, $\boldsymbol{I}$ is the identity matrix.

Under the null hypothesis $H_0 : \boldsymbol{\mu}_y = \boldsymbol{\mu}_b$ and alternative hypothesis $H_1 : \boldsymbol{\mu}_y > \boldsymbol{\mu}_b$, we define a new random variable $\boldsymbol{X}_n^y - \boldsymbol{X}_n^b$. Since under H_0 both $\boldsymbol{X}_n^y$ and $\boldsymbol{X}_n^b$ are normally distributed and independent, their difference also follows a normal distribution:

$$\boldsymbol{X}_n^y - \boldsymbol{X}_n^b \sim N(\boldsymbol{0}, (\sigma_y^2 + \sigma_b^2)\boldsymbol{I}).$$

1.1.2 t distribution

The basic form of the t distribution is:

$$T = \frac{z}{\sqrt{s/d}},$$

where the random variables z and s satisfy the following conditions:

1. $z \sim N(0, 1)$,
2. $s \sim \chi^2(d)$, where d is the degrees of freedom,
3. z and s are independent.

In the case of estimating a population mean, the t distribution arises in the following way:

$$T = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim T(n-1), \quad \text{where } \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2. \tag{1.1}$$

Proof. We can rewrite eq. (1.1) to match the basic form of the t distribution:

$$T = \frac{(\bar{X}-\mu)/\sqrt{\sigma^2/n}}{\hat{\sigma}/\sqrt{n}/\sqrt{\sigma^2/n}} \triangleq \frac{z}{\sqrt{s/(n-1)}}. \tag{1.2}$$

1. We can easily verify that condition item 1 holds because the numerator in eq. (1.2) follows a standard normal distribution.

2. Next, we prove condition item 2 holds in our case. we need to show that: $s = \frac{n-1}{\sigma^2}\hat{\sigma}^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \bar{X})^2$ follows a chi-square distribution with $n-1$ degrees of freedom.

   Let $\boldsymbol{X} = \{x_1, x_2, \cdots, x_n\}^\top$ be the vector of observed data points. Using vector notation, we have:

$$\begin{aligned} s &= \frac{n-1}{\sigma^2}\hat{\sigma}^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\bar{X})^2 \\ &= \frac{1}{\sigma^2}(\boldsymbol{X} - \frac{1}{n}\cdot\mathbf{1}\cdot\mathbf{1}^\top\boldsymbol{X})^\top(\boldsymbol{X} - \frac{1}{n}\cdot\mathbf{1}\cdot\mathbf{1}^\top\boldsymbol{X}) \\ &= \frac{1}{\sigma^2}\left[(\boldsymbol{I} - \frac{1}{n}\cdot\mathbf{1}\cdot\mathbf{1}^\top)\boldsymbol{X}\right]^\top\left[(\boldsymbol{I} - \frac{1}{n}\cdot\mathbf{1}\cdot\mathbf{1}^\top)\boldsymbol{X}\right]. \end{aligned} \tag{1.3}$$

   Let $\boldsymbol{P} := \boldsymbol{I} - \frac{1}{n}\cdot\mathbf{1}\cdot\mathbf{1}^\top$, which is a projection matrix satisfying $\boldsymbol{P}^2 = \boldsymbol{P}$ and $\boldsymbol{P}\cdot\mathbf{1} = 0$:

$$\boldsymbol{P} = \bar{\boldsymbol{U}}\begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}\bar{\boldsymbol{U}}^\top = \boldsymbol{U}\boldsymbol{U}^\top, \tag{1.4}$$

   where $\boldsymbol{U} \in \mathbb{R}^{n\times(n-1)}$ is orthogonal. From eqs. (1.3) and (1.4), we obtain:

$$s = \frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{X} = \frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{X}).$$

   Let $\mathbf{Y}_{n-1} := \boldsymbol{U}^\top\boldsymbol{X} \sim N(\mu\cdot\boldsymbol{U}^\top\cdot\mathbf{1}, \sigma^2\cdot\boldsymbol{U}^\top\boldsymbol{U})$, where $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_{n-1}$. We can also infer that $\boldsymbol{P}\cdot\mathbf{1} = 0$ based on the properties of the projection matrix, which implies that $\boldsymbol{U}^\top\cdot\mathbf{1} = 0$. Finally, we have:

$$s = \frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{X}) = \frac{1}{\sigma^2}\mathbf{Y}_{n-1}^\top\mathbf{Y}_{n-1} = \frac{1}{\sigma^2}\sum_{i=1}^{n-1}Y_i^2 \sim \chi^2(n-1).$$

   Thus, condition item 2 holds.

3. To prove the independence between z and s, we calculate the covariance between $\bar{X}$ and $\boldsymbol{Y}$.

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{Y}, \bar{X}) &= E[\boldsymbol{Y}\cdot\bar{X}] - E[\boldsymbol{Y}]\cdot E[\bar{X}] \\ &= E[\boldsymbol{U}^\top\boldsymbol{X}\cdot\frac{1}{n}\cdot\boldsymbol{X}^\top\cdot\mathbf{1}] && (E[\boldsymbol{Y}] = 0) \\ &= \frac{1}{n}\boldsymbol{U}^\top(E[\boldsymbol{X}\boldsymbol{X}^\top])\cdot\mathbf{1} && (\text{linearity of expection}) \\ &= \frac{1}{n}\boldsymbol{U}^\top\left[(\mu\cdot\mathbf{1})(\mu\cdot\mathbf{1})^\top + \sigma^2\boldsymbol{I}\right]\cdot\mathbf{1} && (\text{covariance structure of } \boldsymbol{X}) \\ &= \frac{1}{n}\boldsymbol{U}^\top\mu^2\cdot\mathbf{1}\cdot\mathbf{1}^\top\cdot\mathbf{1} + \frac{1}{n}\boldsymbol{U}^\top\sigma^2\cdot\boldsymbol{I}\cdot\mathbf{1} = \mathbf{0}_{n-1}. && (\boldsymbol{U}^\top\mathbf{1} = 0) \end{aligned} \tag{1.5}$$

   The eq. (1.5) shows that condition item 3 holds.

$\square$

The t distribution is commonly used in hypothesis testing and in the construction of t-tests. It allows for inference about population means when the population standard deviation is unknown.

1.1.3 Maximum Likelihood Estimator

The Maximum Likelihood Estimator (MLE) of μ is the value that maximizes the likelihood function, i.e.,

$$\hat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} \prod_{i=1}^{n} p(\mathbf{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Next, we simplify the likelihood function by applying a logarithm, which converts the product into a sum. This transformation facilitates computation and differentiation. For example, taking the logarithm of an exponential function often leads to a more tractable additive form. In the case of MLE, this allows us to rewrite the optimization problem as follows:

$$\sum \ln p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}).$$

Next, we use matrix trace to simplify the problem:

$$\begin{aligned}
&\min_{\boldsymbol{\mu}} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \\
\iff &\min \operatorname{Tr}\left[(\boldsymbol{X} - \mathbf{1} \cdot \boldsymbol{\mu}^\top)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \mathbf{1} \cdot \boldsymbol{\mu}^\top) \right] \\
:= &f(\boldsymbol{\mu}).
\end{aligned}$$

After defining $f(\boldsymbol{\mu})$, we calculate its derivative:

$$df = \operatorname{Tr}\left[\left(\frac{\partial f}{\partial \boldsymbol{\mu}} \right)^\top d\boldsymbol{\mu} \right].$$

To compute the derivative of $f(\boldsymbol{\mu})$, let's recall how to differentiate the trace of a matrix. For matrices $\boldsymbol{ABC}$, the trace of the product is denoted as $\operatorname{Tr}(\boldsymbol{ABC})$, and its derivative is:

$$df = \operatorname{Tr}\left[\boldsymbol{dA} \cdot \boldsymbol{BC} + \boldsymbol{A} \cdot \boldsymbol{dB} \cdot \boldsymbol{C} + \boldsymbol{AB} \cdot \boldsymbol{dC} \right]. \tag{1.6}$$

Additionally, based on properties of matrix traces, we know that:

$$\operatorname{Tr}(\boldsymbol{AB}) = \operatorname{Tr}(\boldsymbol{BA}). \tag{1.7}$$

Using eqs. (1.6) and (1.7), the derivative of the MLE is:

$$df = \operatorname{Tr}\left[d(\boldsymbol{X} - \mathbf{1} \cdot \boldsymbol{\mu}^\top)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \mathbf{1} \cdot \boldsymbol{\mu}^\top) + \mathbf{0} + (\boldsymbol{X} - \mathbf{1} \cdot \boldsymbol{\mu}^\top)^\top \boldsymbol{\Sigma}^{-1} d(\boldsymbol{X} - \mathbf{1} \cdot \boldsymbol{\mu}^\top) \right],$$

let $\boldsymbol{A}^\top = \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \mathbf{1} \cdot \boldsymbol{\mu}^\top)$, so we have:

$$\begin{aligned}
df &= \operatorname{Tr}\left[d(\boldsymbol{X} - \mathbf{1} \cdot \boldsymbol{\mu}^\top)^\top \cdot \boldsymbol{A}^\top + \mathbf{0} + \boldsymbol{A} \cdot d(\boldsymbol{X} - \mathbf{1} \cdot \boldsymbol{\mu}^\top) \right] \\
&= \operatorname{Tr}\left[-\mathbf{1} \cdot d\boldsymbol{\mu}^\top \cdot \boldsymbol{A}^\top + \boldsymbol{A} \cdot (-d\boldsymbol{\mu}) \cdot \mathbf{1}^\top \right] \\
&= \operatorname{Tr}\left[-2 \cdot -\mathbf{1}^\top \cdot \boldsymbol{A} \cdot d\boldsymbol{\mu} \right].
\end{aligned}$$

Thus, we derive the first derivative of $f(\boldsymbol{\mu})$. Since $f(\boldsymbol{\mu})$ is a convex function, the point at which the first derivative is zero corresponds to the global minimum. Therefore, we deduce:

$$\begin{aligned}\frac{\partial f}{\partial \mu} &= -2 \cdot \boldsymbol{A} \cdot \boldsymbol{1} = 0 \\ &\Rightarrow \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{1} \cdot \boldsymbol{\mu}^\top) \cdot \boldsymbol{1} = 0 \\ &\Rightarrow \boldsymbol{\mu} = \frac{1}{n} \cdot \boldsymbol{X}^\top \cdot \boldsymbol{1}.\end{aligned}$$

Thus, we obtain the value of $\boldsymbol{\mu}$.

1.2 Linear Regression

1.2.1 Linear Model

Assume we have a linear model described by the equation:

$$\boldsymbol{Y} = \begin{pmatrix}\boldsymbol{1}_n & \boldsymbol{X}\end{pmatrix} \cdot \begin{pmatrix}\boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1\end{pmatrix} + \boldsymbol{e}, \quad \boldsymbol{X} \in \mathbb{R}^{n \times d}, \quad \boldsymbol{\beta}_1 \in \mathbb{R}^{d \times 1}. \tag{1.8}$$

In the eq. (1.8), $\boldsymbol{X}$ is the design matrix of independent variables, and $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1$ are the regression coefficients. The error term $\boldsymbol{e}$ typically assumed to satisfy the Gaussian-Markov conditions:

1. Zero mean: $E[\boldsymbol{e}] = \boldsymbol{0}$,
2. Homoscedasticity and no autocorrelation: $\mathrm{Var}[\boldsymbol{e}] = \sigma^2 \boldsymbol{I}_{n \times n}$.

1.2.2 Square Loss

To estimate $\bar{\boldsymbol{\beta}}$, we minimize the squared loss function, which is given by:

$$L(\bar{\boldsymbol{\beta}}) = \frac{1}{2n}(Y - \bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}})^\top (Y - \bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}}), \quad \text{where} \quad \bar{\boldsymbol{X}} = (\boldsymbol{1}_n, \boldsymbol{X}), \quad \bar{\boldsymbol{\beta}} = \begin{pmatrix}\boldsymbol{\beta_0} \\ \boldsymbol{\beta_1}\end{pmatrix}.$$

Since the squared loss function is convex, we find its global minimum by setting its gradient to zero. Taking the derivative with respect to $\bar{\boldsymbol{\beta}}$ and setting it to zero, we obtain:

$$\nabla_{\bar{\boldsymbol{\beta}}} L(\bar{\boldsymbol{\beta}}) = -\frac{1}{n}\bar{\boldsymbol{X}}^\top (\boldsymbol{Y} - \bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}}) = \boldsymbol{0}.$$

Solving for $\bar{\boldsymbol{\beta}}$, we derive the closed-form solution:

$$\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}} \bar{\boldsymbol{\beta}} = \bar{\boldsymbol{X}}^\top \boldsymbol{Y}. \tag{1.9}$$

Since eq. (1.9) uniquely determines the optimal estimate of $\bar{\boldsymbol{\beta}}$, we denote the solution as $\hat{\bar{\boldsymbol{\beta}}}$:

$$\hat{\bar{\boldsymbol{\beta}}} = (\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}})^{-1} \bar{\boldsymbol{X}}^\top \boldsymbol{Y}.$$

Next, we analyze the properties of $\hat{\bar{\boldsymbol{\beta}}}$, focusing on its unbiasedness and variance.

Lemma 1.1. *The estimator $\hat{\bar{\boldsymbol{\beta}}}$ is unbiased.*

Proof. For $\boldsymbol{Y} = \bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}} + \boldsymbol{e}$, we can deduce:

$$\begin{aligned}E[\hat{\bar{\boldsymbol{\beta}}}] &= E\left[(\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}})^{-1} \bar{\boldsymbol{X}}^\top (\bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}} + \boldsymbol{e})\right] \\ &= (\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}})^{-1}(\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}})\bar{\boldsymbol{\beta}} + (\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}})^{-1}(\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}})E[\boldsymbol{e}].\end{aligned}$$

Since $E[\mathbf{e}] = \boldsymbol{0}$, we have:

$$E[\hat{\bar{\boldsymbol{\beta}}}] = \bar{\boldsymbol{\beta}}.$$

□

Lemma 1.2. *The variance of* $\hat{\bar{\boldsymbol{\beta}}}$ *is given by* $\mathrm{Var}(\hat{\bar{\boldsymbol{\beta}}}) = \sigma^2(\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}})^{-1}$.

Proof. Using the formula for the variance of a linear transformation, we deduce:

$$\begin{aligned}\mathrm{Var}(\hat{\bar{\boldsymbol{\beta}}}) &= (\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}^\top \cdot \mathrm{Var}(\boldsymbol{Y}) \cdot \bar{\boldsymbol{X}}(\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}})^{-1} \\ &= \sigma^2(\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}}(\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}})^{-1} \\ &= \sigma^2(\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}})^{-1}.\end{aligned}$$

□

While we obtained $\hat{\bar{\boldsymbol{\beta}}}$ by minimizing the squared loss, we now show that this solution also corresponds to the Maximum Likelihood Estimator (MLE) under the assumption of Gaussian errors.

Lemma 1.3. $\hat{\bar{\boldsymbol{\beta}}}$ *is the MLE of* $\bar{\boldsymbol{\beta}}$ *under the Gaussian error term* $\boldsymbol{e}$.

Proof. From the linear model in eq. (1.8), we can easily get that $\boldsymbol{Y}$ is following a normal distribution. The likelhood function of $\boldsymbol{Y}$ given $\bar{\boldsymbol{X}}$ is:

$$L(\bar{\boldsymbol{\beta}}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}})^\top(\boldsymbol{Y} - \bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}})\right).$$

Taking the logarithm, we have:

$$\log L(\bar{\boldsymbol{\beta}}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}})^\top(\boldsymbol{Y} - \bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}}).$$

To find MLE of $\bar{\boldsymbol{\beta}}$, we maximize $\log L(\bar{\boldsymbol{\beta}}, \sigma^2)$. Since the first term is independent of $\bar{\boldsymbol{\beta}}$, we only need to minimize $(\boldsymbol{Y} - \bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}})^\top(\boldsymbol{Y} - \bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}})$. We take its derivative with respect to $\bar{\boldsymbol{\beta}}$ and set it to $\mathbf{0}$:

$$\begin{aligned}&\nabla_{\bar{\boldsymbol{\beta}}}(\boldsymbol{Y} - \bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}})^\top(\boldsymbol{Y} - \bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}}) = \mathbf{0} \\ \Rightarrow \quad &-2\bar{\boldsymbol{X}}^\top(\boldsymbol{Y} - \bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{\beta}}}) = \mathbf{0}\end{aligned}$$

Thus, we obtain the solution:

$$\hat{\bar{\boldsymbol{\beta}}} = (\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}^\top\boldsymbol{Y}.$$

□

1.2.3 Population Risk

In this section, we investigate the population risk of our linear model. We remark that during the training process, we are minimizing the *empirical risk* on the fixed design matrix $\bar{\boldsymbol{X}}$.

Specifically, we assume a new random sample $\mathbf{x}$ and the corresponding label $y = \mathbf{x}^\top\boldsymbol{\beta} + \epsilon$; the risk is expressed as below:

$$E_{\mathbf{x},y,\hat{\boldsymbol{\beta}}}\left(y - \mathbf{x}^\top\hat{\boldsymbol{\beta}}\right)^2 = E_{\hat{\boldsymbol{\beta}}}\left[E_{\mathbf{x},y}\left(y - \boldsymbol{x}^\top\hat{\boldsymbol{\beta}}\right)^2 \mid \hat{\boldsymbol{\beta}}\right].$$

Considering that the noise term ϵ is a random variable independent of the feature vector $\mathbf{x}$, and that its expectation is zero, the population risk depends on: $E_{\mathbf{x},\hat{\boldsymbol{\beta}}}\left[\mathbf{x}^\top\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^\top\mathbf{x}\right]$ For simplicity, we refer to this as the prediction error in the well-specified case, where the model is correctly specified. We are equivalently minimizing:

$$\begin{aligned}\min E_{\mathbf{x},\hat{\boldsymbol{\beta}}}&\left[\mathbf{x}^\top\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^\top\mathbf{x}\right] \\ &= E_{\hat{\boldsymbol{\beta}}}\left[E_{\mathbf{x}}\ \mathrm{Tr}\left(\mathbf{x}\mathbf{x}^\top(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mid \hat{\boldsymbol{\beta}}\right)\right] \\ &= E_{\hat{\boldsymbol{\beta}}}\left[\mathrm{Tr}\left(E[\mathbf{x}\mathbf{x}^\top](\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top\right) \mid \hat{\boldsymbol{\beta}}\right] \\ &= \mathrm{Tr}\left(E[\mathbf{x}\mathbf{x}^\top] \cdot E\left[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top\right]\right) \\ &= \sigma^2\ \mathrm{Tr}\left(E[\mathbf{x}\mathbf{x}^\top]\left(\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}}\right)^{-1}\right).\end{aligned} \tag{1.10}$$

In eq. (1.10), we recall that $\bar{\boldsymbol{X}}$ is the fixed design matrix in the training process. In other words, during the training, we can manipulate $\bar{\boldsymbol{X}}$ to minimize the population risk, while other quantities remain fixed. The expectation $E[\mathbf{x}\mathbf{x}^\top]$ represents the second-moment matrix of the feature distribution and remains constant across realizations of $\mathbf{x}$, whereas individual samples $\mathbf{x}$ are random. Assuming $E[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{I}$, the preceding expression simplifies to: $\sigma^2 \operatorname{Tr}\left[(\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}})^{-1}\right] = \operatorname{Tr}\left[\operatorname{Var}(\hat{\boldsymbol{\beta}})\right]$, indicating the connection between population risk and parameter variance in the case of linear regression.