Linear regression

1 Mean squared error estimation

1.1 Minimum mean squared error estimation

We consider the problem of estimating a certain quantity of interest that we model as a random variable $\tilde{y}$. We evaluate our estimate using its average squared deviation from $\tilde{y}$, which we call the mean squared error (MSE). Imagine that the only information we have is the probability distribution of $\tilde{y}$, i.e. its probability mass function (pmf) or probability density function (pdf). For example, we want to estimate the temperature in New York tomorrow, but without up-to-date meteorological observations. We only have access to a probability distribution of temperatures obtained from historical data. Since we have no measurements related to $\tilde{y}$, we can only generate a constant estimate using the pdf or pmf of $\tilde{y}$. In that case, the best possible estimate in terms of MSE is the mean of $\tilde{y}$.

Theorem 1.1 (Minimum MSE constant estimate)**.** *For any random variable $\tilde{y}$ with finite mean, $\mathrm{E}(\tilde{y})$ is the best constant estimate of $\tilde{y}$ in terms of MSE,*

$$\mathrm{E}(\tilde{y}) = \arg\min_{c \in \mathbb{R}} \mathrm{E}\left((c - \tilde{y})^2\right). \tag{1}$$

Proof. Let $g(c) := \mathrm{E}\left((c - \tilde{y})^2\right) = c^2 - 2c\mathrm{E}\left(\tilde{y}\right) + \mathrm{E}\left(\tilde{y}^2\right)$, we have

$$g'(c) = 2(c - \mathrm{E}(\tilde{y})), \tag{2}$$
$$g''(c) = 2. \tag{3}$$

The function is strictly convex and has a minimum where the derivative equals zero, i.e. when c is equal to the mean. □

Now let us study a more interesting situation where we do have some data related to our quantity of interest, which we call the *response* (or dependent variable). We model these quantities, which we call the *features* (also known as covariates or independent variables) as a random vector $\tilde{x}$ belonging to the same probability space as $\tilde{y}$. In our temperature example, these features could be the humidity, wind speed, temperature at other locations, etc. Estimating the response from the features is a fundamental problem in statistics known as *regression*.

If we observe that $\tilde{x}$ equals a fixed value x, the uncertainty about $\tilde{y}$ is captured by the distribution of $\tilde{y}$ given $\tilde{x} = x$. Let $\tilde{w}$ be a random variable that follows that distribution. Minimizing the MSE for the fixed observation $\tilde{x} = x$ is exactly equivalent to finding a constant vector c that minimizes $\mathrm{E}[(\tilde{w} - c)^2]$. By Theorem 1.1 the optimal estimator is the mean of the distribution, i.e. the conditional mean $\mathrm{E}(\tilde{y} \,|\, \tilde{x} = x)$. Recall that in statistics an *estimator* is a function of the data that provides an estimate of our quantity of interest. The following theorem shows that this is indeed the optimal estimator in terms of MSE. The proof is identical to that of Theorem 1.1.

Carlos Fernandez-Granda, Courant Institute of Mathematical Sciences and Center for Data Science, NYU

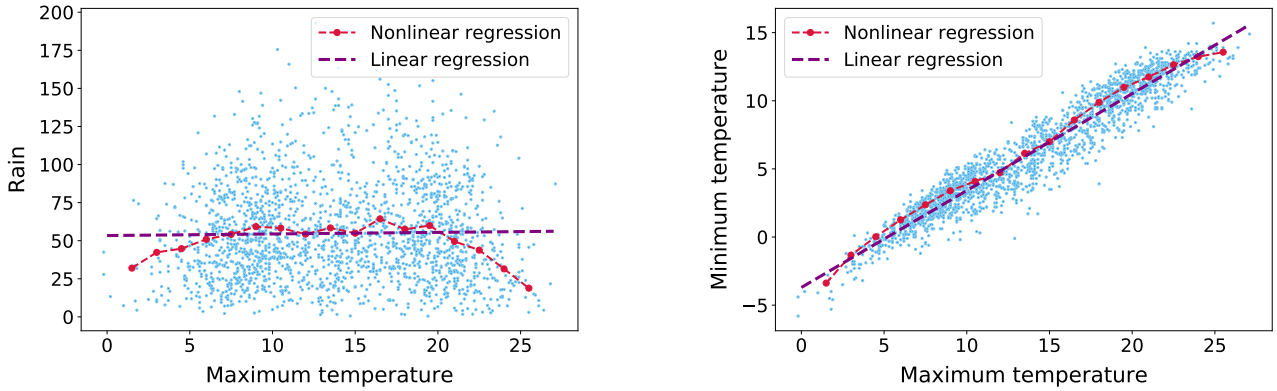


Figure 1: Regression models for weather measurements gathered at a weather station in Oxford over 150 years. On the left, the response is the monthly amount of rain, and the feature the maximum temperature during the same month. On the right, the response and the feature are the minimum and maximum monthly temperature respectively. The linear regression model is compute by minimizing the least-squares fit. The nonlinear regression model is computed by averaging the response over all values of the feature in bins of width equal to 1.5 degrees. In the case of the rain, the linear model cannot capture the fact that at high temperatures, rain and temperature are negatively correlated.

Theorem 1.2 (Minimum MSE estimator)**.** *Let $\tilde{x}$ and $\tilde{y}$ be real-valued random variables or random vectors defined in the same probability space. If $\tilde{x} = x$ then the minimum MSE estimator of $\tilde{y}$ given $\tilde{x}$ is the conditional expectation of $\tilde{y}$ given $\tilde{x} = x$, i.e.*

$$\mathrm{E}(\tilde{y} \,|\, \tilde{x} = x) = \arg\min_{w} \mathrm{E}\left[(\tilde{y} - w)^2 \,|\, \tilde{x} = x\right]. \tag{4}$$

According to the theorem, to solve the regression problem all we need to do is compute the average value of the response corresponding to every possible value of the features. The catch is that when there is more than one or two features this requires too many data. As a simple example, consider a problem with p features each taking d different values. In order to be able to perform estimation, we need to compute the expected value of the response conditioned on every possible value of the feature vector. However there are $N = d^p$ possible values! For even moderate values of p and d the number is huge: if $p = 5$, and $d = 100$ then $N = 10^{10}$! This is known as the curse of dimensionality (where dimensionality refers to the dimension of the feature vector).

1.2 Linear minimum-mean-squared-error estimation

Due to the curse of dimensionality, tackling the regression problem requires making assumptions about the relationship between the response and the features. A simple, yet often surprisingly effective, assumption is that the relationship is linear (or rather affine), i.e. there exists a constant vector $\beta \in \mathbb{R}^p$ and a constant $\beta_0 \in \mathbb{R}$ such that

$$\tilde{y} \approx \beta^T \tilde{x} + \beta_0. \tag{5}$$

Mathematically, the gradient of the regression function is constant, which means that the rate of change in the response with respect to the features does not depend on the feature values. This is illustrated in Figure 1, which compares a linear model with a nonlinear model for two simple examples where there is only one feature. The slope of the nonlinear estimate varies depending on the feature, but the slope of the linear model is constrained to be constant.

The following lemma establishes that when fitting an affine model by minimizing MSE, we can just center the response and the features, and fit a linear model without additive constants.

Lemma 1.3 (Centering works)**.** *For any $\beta \in \mathbb{R}^p$ and any random variable $\tilde{y}$ and p-dimensional random vector $\tilde{x}$,*

$$\min_{\beta_0} \mathrm{E}\left[(\tilde{y} - \tilde{x}^T\beta - \beta_0)^2\right] = \mathrm{E}\left[(c(\tilde{y}) - c(\tilde{x})^T\beta)^2\right], \tag{6}$$

where $c(\tilde{y}) := \tilde{y} - \mathrm{E}(\tilde{y})$ and $c(\tilde{x}) := \tilde{x} - \mathrm{E}(\tilde{x})$.

Proof. By Theorem 1.1, if we just optimize over β_0 the minimum is $\mathrm{E}(\tilde{y} - \tilde{x}^T\beta)$, so

$$\min_{\beta_0} \mathrm{E}\left[(\tilde{y} - \tilde{x}^T\beta - \beta_0)^2\right] = \mathrm{E}\left[(\tilde{y} - \tilde{x}^T\beta - \mathrm{E}(\tilde{y}) + \mathrm{E}(\tilde{x})^T\beta)^2\right] \tag{7}$$

$$= \mathrm{E}\left[(c(\tilde{y}) - \beta^T c(\tilde{x}))^2\right]. \tag{8}$$

□

From now on, we will assume that the response and the features are centered. The following theorem derives the linear estimator that minimizes MSE. Perhaps surprisingly, it only depends on the covariance matrix of the features and the cross-covariance between the response and the features.

Theorem 1.4 (Linear minimum MSE estimator)**.** *Let $\tilde{y}$ be a zero-mean random variable and $\tilde{x}$ a zero mean random vector with a full-rank covariance matrix equal to $\Sigma_{\tilde{x}}$, then*

$$\Sigma_{\tilde{x}}^{-1}\Sigma_{\tilde{x}\tilde{y}} = \arg\min_{\beta} \mathrm{E}\left[(\tilde{y} - \tilde{x}^T\beta)^2\right], \tag{9}$$

where $\Sigma_{\tilde{x}\tilde{y}}$ is the cross-covariance between $\tilde{x}$ and $\tilde{y}$:

$$\Sigma_{\tilde{x}\tilde{y}}[i] := \mathrm{E}\left(\tilde{x}[i]\tilde{y}\right), \quad 1 \leq i \leq p. \tag{10}$$

The MSE of this estimator equals $\mathrm{Var}(\tilde{y}) - \Sigma_{\tilde{x}\tilde{y}}^T\Sigma_{\tilde{x}}^{-1}\Sigma_{\tilde{x}\tilde{y}}$.

Proof. We have

$$\mathrm{E}((\tilde{y} - \tilde{x}^T\beta)^2) = \mathrm{E}\left(\tilde{y}^2\right) - 2\mathrm{E}\left(\tilde{y}\tilde{x}\right)^T\beta + \beta^T\mathrm{E}(\tilde{x}\tilde{x}^T)\beta \tag{11}$$

$$= \beta^T\Sigma_{\tilde{x}}\beta - 2\Sigma_{\tilde{x}\tilde{y}}^T\beta + \mathrm{Var}\left(\tilde{y}\right) := f(\beta). \tag{12}$$

The function f is a quadratic form. Its gradient and Hessian equal

$$\nabla f(\beta) = 2\Sigma_{\tilde{x}}\beta - 2\Sigma_{\tilde{x}\tilde{y}}, \tag{13}$$

$$\nabla^2 f(\beta) = 2\Sigma_{\tilde{x}}. \tag{14}$$

The data are available here.

Covariance matrices are positive semidefinite. For any vector $v \in \mathbb{R}^p$

$$v^T \Sigma_{\tilde{x}} v = \text{Var}\left(v^T \tilde{x}\right) \geq 0. \tag{15}$$

Since $\Sigma_{\tilde{x}}$ is full rank, it is actually positive definite, i.e. the inequality is strict as long as $v \neq 0$. This means that the quadratic function is strictly convex and we can set its gradient to zero to find its unique minimum. For the sake of completeness, we provide a simple proof of this. The quadratic form is exactly equal to its second-order Taylor expansion around any point $\beta_1 \in \mathbb{R}^p$. For all $\beta_2 \in \mathbb{R}^p$

$$f(\beta_2) = \frac{1}{2}(\beta_2 - \beta_1)^T \nabla^2 f(\beta_1)(\beta_2 - \beta_1) + \nabla f(\beta_1)^T (\beta_2 - \beta_1) + f(\beta_1). \tag{16}$$

The equality can be verified by expanding the expression. This means that if $\nabla f(\beta^*) = 0$ then for any $\beta \neq \beta^*$

$$f(\beta) = \frac{1}{2}(\beta - \beta^*)^T \nabla^2 f(\beta^*)(\beta - \beta^*) + f(\beta^*) > f(\beta^*) \tag{17}$$

because $\nabla^2 f(\beta^*) = \Sigma_{\tilde{x}}$ is positive definite. The unique minimum can therefore be found by setting the gradient to zero. Finally, the corresponding MSE equals

$$\begin{aligned} \text{E}\left[(\tilde{y} - \tilde{x}^T \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}\tilde{y}})^2\right] &= \text{E}(\tilde{y}^2) + \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} \text{E}(\tilde{x}\tilde{x}^T) \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}\tilde{y}} - 2\text{E}(\tilde{y}\tilde{x}^T) \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}\tilde{y}} \\ &= \text{Var}(\tilde{y}) - \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}\tilde{y}}. \end{aligned} \tag{18}$$

□

Example 1.5 (Noise cancellation)**.** We are interested in recording the voice of a pilot in a helicopter. To this end we place a microphone inside his helmet and another microphone outside. We model the measurements as

$$\tilde{x}[1] = \tilde{y} + \alpha \tilde{z} \tag{19}$$
$$\tilde{x}[2] = \alpha \tilde{y} + \tilde{z}, \tag{20}$$

where $\tilde{y}$ is a random variable modeling the voice of the pilot, $\tilde{z}$ is a random variable modeling the noise in the helicopter, and $0 < \alpha < 1$ is a constant that models the effect of the helmet. From past data, we determine that $\tilde{y}$, and $\tilde{z}$ are zero mean and uncorrelated with each other. The variances of $\tilde{y}$ and $\tilde{z}$ are equal to 1 and 100 respectively.

By independence

$$\text{Var}(\tilde{x}[1]) = 1 + 100\alpha^2, \tag{21}$$
$$\begin{aligned} \text{Var}(\tilde{x}[2]) &= \alpha^2 \text{Var}(\tilde{y}) + \text{Var}(\tilde{z}) && (22) \\ &= \alpha^2 + 100, && (23) \end{aligned}$$
$$\begin{aligned} \text{Cov}(\tilde{x}[1]\tilde{x}[2]) &= \alpha \text{E}(\tilde{y}^2) + \alpha \text{E}(\tilde{z}^2) && (24) \\ &= 101\alpha, && (25) \end{aligned}$$
$$\text{Cov}(\tilde{y}\tilde{x}[1]) = 1, \tag{26}$$
$$\text{Cov}(\tilde{y}\tilde{x}[2]) = \alpha, \tag{27}$$
$$\tag{28}$$

so

$$\Sigma_{\tilde{x}} = \begin{bmatrix} 1 + 100\alpha^2 & 101\alpha \\ 101\alpha & \alpha^2 + 100 \end{bmatrix}, \tag{29}$$

$$\Sigma_{\tilde{b}\tilde{x}} = \begin{bmatrix} 1 \\ \alpha \end{bmatrix}, \tag{30}$$

and by Theorem 1.4 the estimator equals

$$\hat{y}(\tilde{x}) = \tilde{x}^T \begin{bmatrix} 1 + 100\alpha^2 & 101\alpha \\ 101\alpha & \alpha^2 + 100 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \alpha \end{bmatrix} \tag{31}$$

$$= \tilde{x}^T \frac{1}{(1 + 100\alpha^2)(\alpha^2 + 100) - 101^2\alpha^2} \begin{bmatrix} \alpha^2 + 100 & -101\alpha \\ -101\alpha & 1 + 100\alpha^2 \end{bmatrix} \begin{bmatrix} 1 \\ \alpha \end{bmatrix} \tag{32}$$

$$= \tilde{x}^T \frac{1}{100(1 - \alpha^2)^2} \begin{bmatrix} 100(1 - \alpha^2) \\ -100\alpha(1 - \alpha^2) \end{bmatrix} \tag{33}$$

$$= \frac{\tilde{x}[1] - \alpha\tilde{x}[2]}{1 - \alpha^2}. \tag{34}$$

Notice that

$$\hat{y}(\tilde{x}) = \frac{\tilde{x}[1] - \alpha\tilde{x}[2]}{1 - \alpha^2} \tag{35}$$

$$= \frac{\tilde{y} + \alpha\tilde{z} - \alpha(\alpha\tilde{y} + \tilde{z})}{1 - \alpha^2} \tag{36}$$

$$= \tilde{y} \tag{37}$$

so the estimate is perfect! The linear estimator cancels out the noise completely by scaling the second measurement and subtracting it from the first one. △

1.3 Additive data model

In this section we analyze the linear minimum MSE estimator under the assumption that the data are indeed generated by a linear model. To account for model inaccuracy and noise we incorporate an additive scalar $\tilde{z}$. More precisely, we assume that the response equals

$$\tilde{y} = \tilde{x}^T \beta_{\text{true}} + \tilde{z}., \tag{38}$$

where $\beta_{\text{true}} \in \mathbb{R}^p$ is a vector of *true* linear coefficients $\beta_{\text{true}} \in \mathbb{R}^p$. A common assumption is that the noise $\tilde{z}$ and the features are independent. In that case, the MSE achieved by the linear minimum MSE estimator is equal to the variance of the noise. This make sense, because the noise is *unpredictable* as it is independent from the observed features.

Theorem 1.6 (Linear minimum MSE for additive model)**.** *Let $\tilde{x}$ and $\tilde{z}$ in Eq. (38) be zero mean and independent. Then the MSE achieved by the linear minimum MSE estimator of $\tilde{y}$ given $\tilde{x}$ is equal to the variance of $\tilde{z}$.*

Proof. By independence of $\tilde{x}$ and $\tilde{z}$, and linearity of expectation we have

$$\begin{aligned}
\text{Var}(\tilde{y}) &= \text{Var}(\tilde{x}^T \beta_{\text{true}} + \tilde{z}) && (39)\\
&= \beta_{\text{true}}^T \text{E}\left(\tilde{x}\tilde{x}^T\right) \beta_{\text{true}} + \text{Var}\left(\tilde{z}\right) && (40)\\
&= \beta_{\text{true}}^T \Sigma_{\tilde{x}} \beta_{\text{true}} + \text{Var}\left(\tilde{z}\right), && (41)\\
\Sigma_{\tilde{x}\tilde{y}} &= \text{E}\left(\tilde{x}(\tilde{x}^T \beta_{\text{true}} + \tilde{z})\right) && (42)\\
&= \Sigma_{\tilde{x}} \beta_{\text{true}}. && (43)
\end{aligned}$$

By Theorem 1.2,

$$\begin{aligned}
\text{MSE} &= \text{Var}(\tilde{y}) - \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}\tilde{y}} && (44)\\
&= \beta_{\text{true}}^T \Sigma_{\tilde{x}} \beta_{\text{true}} + \text{Var}(\tilde{z}) - \beta_{\text{true}}^T \Sigma_{\tilde{x}} \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}} \beta_{\text{true}} && (45)\\
&= \text{Var}(\tilde{z}). && (46)
\end{aligned}$$

□

2 Ordinary least squares

In Section 1 we have studied the regression problem under the idealized assumption that we have access to the true joint statistics (covariance and cross-covariance) of the response and the features. In practice, we need to perform estimation based on a finite set of data. Assume that we have available n examples consisting of feature vectors coupled with their respective response: (y_1, x_1), (y_2, x_2), $\ldots$, (y_n, x_n), where $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$ for $1 \leq i \leq n$. We define a response vector $y \in \mathbb{R}^n$, such that $y[i] := y_i$, and a feature matrix $X \in \mathbb{R}^{p \times n}$ with columns equal to the feature vectors,

$$X := \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}. \tag{47}$$

If we interpret the feature data as samples of $\tilde{x}$ and the corresponding response values as samples of $\tilde{y}$, a reasonable estimate for the covariance matrix is the sample covariance matrix,

$$\frac{1}{n} X X^T = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T. \tag{48}$$

Similarly, the cross-covariance can be approximated by the sample cross-covariance, which contains the sample covariance between each feature and the response,

$$\frac{1}{n} X y = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n x_i[1] y_i \\ \frac{1}{n} \sum_{i=1}^n x_i[2] y_i \\ \cdots \\ \frac{1}{n} \sum_{i=1}^n x_i[p] y_i \end{bmatrix}. \tag{49}$$

We obtain the following approximation to the linear minimum MSE estimator derived in Theorem 1.4,

$$\Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}\tilde{y}} \approx (X X^T)^{-1} X y. \tag{50}$$

This estimator has an alternative interpretation, which does not require probabilistic assumptions: it minimizes the least-squares fit between the observed values of the response and the linear model. In the statistics literature, this method is known as ordinary least squares (OLS).

Theorem 2.1 (Ordinary least squares)**.** *If* $X := \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \in \mathbb{R}^{p \times n}$ *is full rank and* $n \geq p$*, for any* $y \in \mathbb{R}^n$ *we have*

$$\beta_{\text{OLS}} := \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i^T \beta \right)^2 \tag{51}$$

$$= (XX^T)^{-1} Xy. \tag{52}$$

Proof.

$$\sum_{i=1}^{n} \left(y_i - x_i^T \beta \right)^2 = \|y - X^T \beta\|_2^2 \tag{53}$$

$$= \beta^T XX^T \beta - 2y^T X^T \beta + y^T y := f(\beta). \tag{54}$$

The function f is a quadratic form. Its gradient and Hessian equal

$$\nabla f(\beta) = 2XX^T \beta - 2Xy, \tag{55}$$

$$\nabla^2 f(\beta) = 2XX^T. \tag{56}$$

Since X is full rank, XX^T is positive definite because for any nonzero vector v

$$v^T XX^T v = \left|\left| X^T v \right|\right|_2^2 > 0. \tag{57}$$

By the same argument in Theorem 1.4, the unique minimum can be found by setting the gradient to zero. □

In practice, large-scale least-squares problems are not solved by using the closed-form solution, due to the computational cost of inverting the sample covariance matrix of the features, but rather by applying iterative optimization methods such as conjugate gradients.

Example 2.2 (Temperature prediction via linear regression)**.** We consider a dataset of hourly temperatures measured at weather stations all over the United States. Our goal is to design a model that can be used to estimate the temperature in Yosemite Valley from the temperatures of 133 other stations, in case the sensor in Yosemite fails. We perform estimation by fitting a linear model where the response is the temperature in Yosemite and the features are the rest of the temperatures ($p = 133$). We use 10^3 measurements from 2015 as a test set, and train a linear model using a variable number of training data also from 2015 but disjoint from the test data. In addition, we test the linear model on data from 2016. Figure 2 shows the results. With enough data, the linear model achieves an error of roughly 2.5°C on the test data, and 2.8°C on the 2016 data. The linear model outperforms a naive single-station estimate, which uses the station that best predicts the temperature in Yosemite for the training data. △

The data are available at http://www1.ncdc.noaa.gov/pub/data/uscrn/products

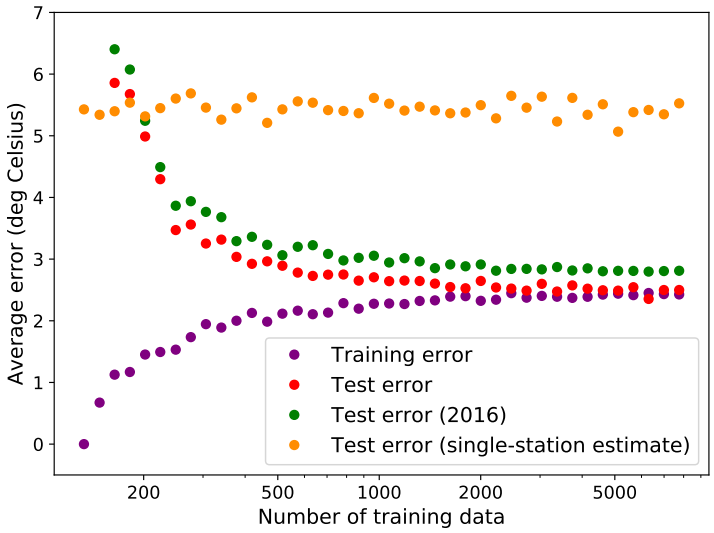


Figure 2: Performance of the OLS estimator on the temperature data described in Example 2.2. The graph shows the square root of the MSE (RMSE) achieved by the model on the training and test sets, and on the 2016 data, for different number of training data and compares it to the RMSE of the best single-station estimate.

3 The singular-value decomposition

In order to gain further insight into linear models we introduce a fundamental tool in linear algebra: the singular-value decomposition. We omit the proof of its existence, which follows from the spectral theorem for symmetric matrices.

Theorem 3.1 (Singular-value decomposition)**.** *Every real matrix $A \in R^{m \times k}$, $m \geq k$, has a singular-value decomposition (SVD) of the form*

$$A = \begin{bmatrix} u_1 & u_2 & \cdots & u_k \end{bmatrix} \begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & s_k \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \cdots & v_k \end{bmatrix}^T \tag{58}$$

$$= USV^T, \tag{59}$$

where the singular values $s_1 \geq s_2 \geq \cdots \geq s_k$ are nonnegative real numbers, the left singular vectors $u_1, u_2, \ldots u_k \in \mathbb{R}^m$ form an orthonormal set, and the right singular vectors $v_1, v_2, \ldots v_k \in \mathbb{R}^k$ also form an orthonormal set.

If $m < k$ then the SVD is of the form

$$A = \begin{bmatrix} u_1 & u_2 & \cdots & u_m \end{bmatrix} \begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & s_m \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \cdots & v_m \end{bmatrix}^T \tag{60}$$

$$= USV^T, \tag{61}$$

where $s_1 \geq s_2 \geq \cdots \geq s_m$ are nonnegative real numbers, and the singular vectors u_1, u_2, $\ldots u_m \in \mathbb{R}^m$, and v_1, v_2, $\ldots v_m \in \mathbb{R}^k$ form orthonormal sets.

The SVD provides a very intuitive geometric interpretation of the action of a matrix $A \in \mathbb{R}^{m \times k}$ on a vector $w \in \mathbb{R}^k$, as illustrated in Figure 3:

1. Rotation of w to align the component of w in the direction of the ith right singular vector v_i with the ith axis:

$$V^T w = \sum_{i=1}^{k} \langle v_i, w \rangle e_i, \tag{62}$$

   where e_i is the ith standard basis vector.

2. Scaling of each axis by the corresponding singular value

$$SV^T w = \sum_{i=1}^{k} s_i \langle v_i, w \rangle e_i. \tag{63}$$

3. Rotation to align the ith axis with the ith left singular vector

$$USV^T w = \sum_{i=1}^{k} s_i \langle v_i, w \rangle u_i. \tag{64}$$

Another consequence of the spectral theorem for symmetric matrices is that the maximum scaling produced by a matrix is equal to the maximum singular value. The maximum is achieved when the matrix is applied to any vector in the direction of the right singular vector v_1. If we restrict our attention to the orthogonal complement of v_1, then the maximum scaling is the second singular value, due to the orthogonality of the singular vectors. In general, the direction of maximum scaling orthogonal to the first $i-1$ left singular vectors is equal to the ith singular value and occurs in the direction of the ith singular vector.

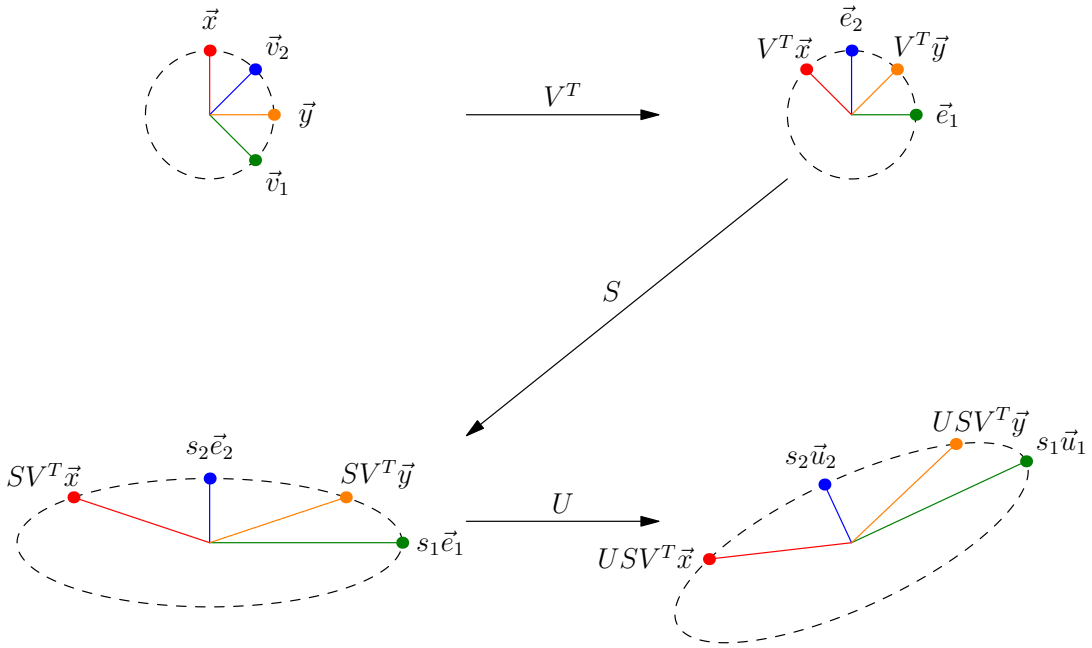
Theorem 3.2. *For any matrix $A \in \mathbb{R}^{m \times k}$, the singular values satisfy*

$$s_1 = \max_{\{\|w\|_2 = 1 \mid w \in \mathbb{R}^k\}} \|Aw\|_2, \tag{65}$$

$$s_i = \max_{\{\|w\|_2 = 1 \mid w \in \mathbb{R}^k, w \perp v_1, \ldots, v_{i-1}\}} \|Aw\|_2, \tag{66}$$

$$\tag{67}$$

(a) $s_1 = 3$, $s_2 = 1$.



(b) $s_1 = 3$, $s_2 = 0$.

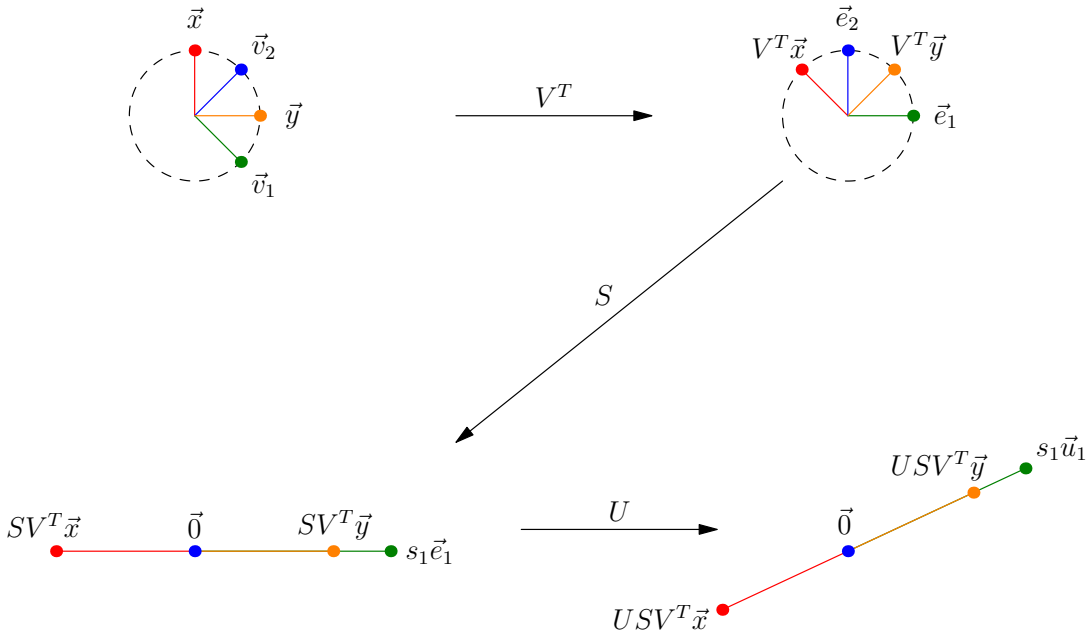


Figure 3: The action of any matrix can be decomposed into three steps: rotation to align the right singular vectors to the axes, scaling by the singular values and a final rotation to align the axes with the left singular vectors. In image (b) the second singular value is zero, so the matrix projects two-dimensional vectors onto a one-dimensional subspace.

and the right singular vectors satisfy

$$v_1 = \underset{\{\|w\|_2=1 \mid w\in\mathbb{R}^k\}}{\arg\max} \|Aw\|_2, \tag{68}$$

$$v_i = \underset{\{\|w\|_2=1 \mid w\in\mathbb{R}^k, w\perp v_1,\ldots,v_{i-1}\}}{\arg\max} \|Aw\|_2, \qquad 2 \leq i \leq k. \tag{69}$$

The SVD provides a geometric interpretation of the OLS estimator derived in Theorem 2.1. Let $X = USV^T$ be the SVD of the feature matrix, then

$$\beta_{\text{OLS}} = \left(XX^T\right)^{-1} Xy \tag{70}$$

$$= (US^2U^T)^{-1}USV^Ty \tag{71}$$

$$= US^{-2}U^TUSV^Ty \tag{72}$$

$$= US^{-1}V^Ty. \tag{73}$$

The OLS estimator is obtained by inverting the action of the feature matrix. This is achieved by computing the components of the response vector in the direction of the right singular vectors, scaling by the inverse of the corresponding singular values, and then rotating so that each component is aligned with the corresponding left singular vector. The matrix $\left(XX^T\right)^{-1} X$ is called a left inverse or pseudoinverse of X^T because $\left(XX^T\right)^{-1} XX^T = I$.

4 Analysis of ordinary-least-squares estimation

In order to study the properties of OLS, we study an additive model. The training data are equal to the n-dimensional vector

$$\tilde{y}_{\text{train}} := X^T\beta_{\text{true}} + \tilde{z}_{\text{train}}, \tag{74}$$

where $X \in \mathbb{R}^{p\times n}$ contains n p-dimensional feature vectors. The noise $\tilde{z}_{\text{train}}$ is modeled as an n-dimensional iid Gaussian vector with zero mean and variance σ^2. The feature matrix X is fixed and deterministic. OLS is equivalent to maximum-likelihood estimation under this model.

Lemma 4.1. *If the training data are interpreted as a realization of the random vector in Eq. (74) the maximum-likelihood estimate of the coefficients is equal to the OLS estimator.*

Proof. The likelihood is the probability density function of $\tilde{y}_{\text{train}}$ evaluated at the observed data y_{train} and interpreted as a function of the coefficient vector β,

$$\mathcal{L}_{y_{\text{train}}}(\beta) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}\left|\left|y_{\text{train}} - X^T\beta\right|\right|_2^2\right). \tag{75}$$

The maximum-likelihood estimate equals

$$\beta_{\text{ML}} = \arg\max_\beta \mathcal{L}_{y_{\text{train}}}(\beta) \tag{76}$$

$$= \arg\max_\beta \log \mathcal{L}_{y_{\text{train}}}(\beta) \tag{77}$$

$$= \arg\min_\beta \left|\left|y_{\text{train}} - X^T\beta\right|\right|_2^2. \tag{78}$$

□

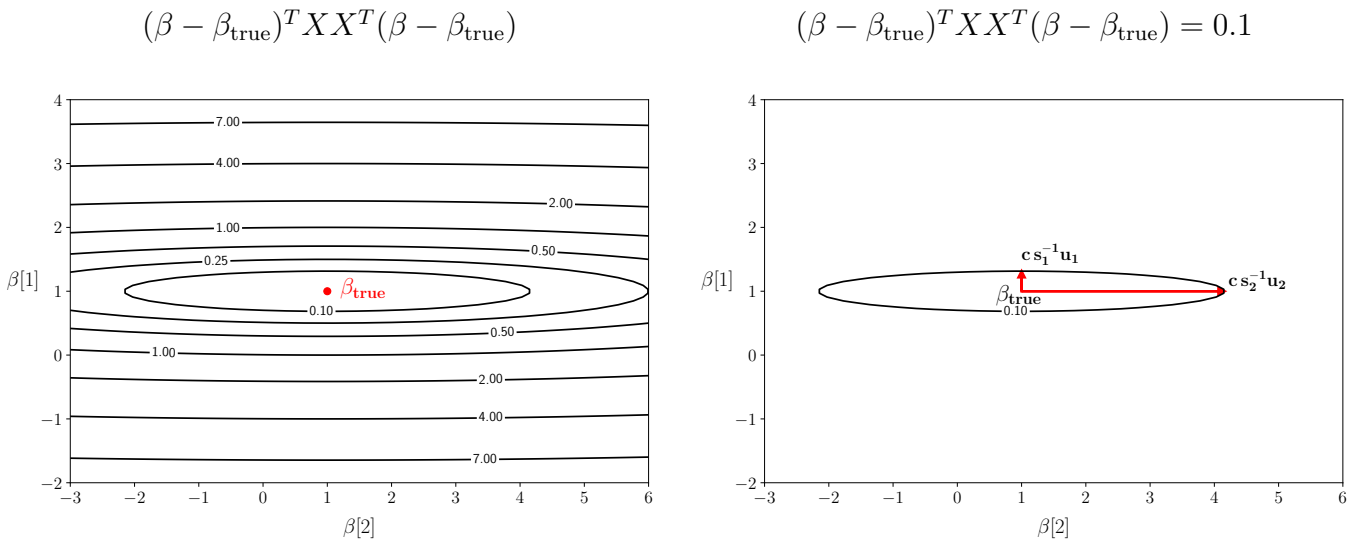


Figure 4: Deterministic quadratic component of the least-squares cost function (see Eq. (80)) for an example with two features where the left singular vectors of X align with the horizontal and vertical axes, and the singular values equal 1 and 0.1. The quadratic form is an ellipsoid centered at β_{true} with axes aligned with the left singular vectors. The curvature of the quadratic is proportional to the square of the singular values.

In the following sections we analyze the OLS coefficient estimate, as well as its corresponding training and test errors when the training data follow the additive model.

4.1 Analysis of OLS coefficients

If the data follow the additive model in (74), the OLS cost function can be decomposed into a deterministic quadratic form centered at β_{true} and a random linear function that depends on the noise,

$$\begin{aligned}
\arg\min_{\beta} \|\tilde{y}_{\text{train}} - X^T\beta\|_2^2 &= \arg\min_{\beta} \|\tilde{z}_{\text{train}} - X^T(\beta - \beta_{\text{true}})\|_2^2 \qquad (79)\\
&= \arg\min_{\beta} (\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) - 2\tilde{z}_{\text{train}}^T X^T(\beta - \beta_{\text{true}}) + \tilde{z}_{\text{train}}^T\tilde{z}_{\text{train}}\\
&= \arg\min_{\beta} (\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) - 2\tilde{z}_{\text{train}}^T X^T\beta. \qquad (80)
\end{aligned}$$

Figure 4 shows the quadratic component for a simple example with two features. Let $X = USV^T$ be the SVD of the feature matrix. The contour lines of the quadratic form are ellipsoids, defined by the equation

$$\begin{aligned}
(\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) &= (\beta - \beta_{\text{true}})^T US^2U^T(\beta - \beta_{\text{true}}) \qquad (81)\\
&= \sum_{i=1}^{p} s_i^2(u_i^T(\beta - \beta_{\text{true}}))^2 = c^2 \qquad (82)
\end{aligned}$$

for a constant c. The axes of the ellipsoid are the left singular vectors of X. The curvature in those directions is proportional to the square of the singular values, as shown in Figure 4. Due

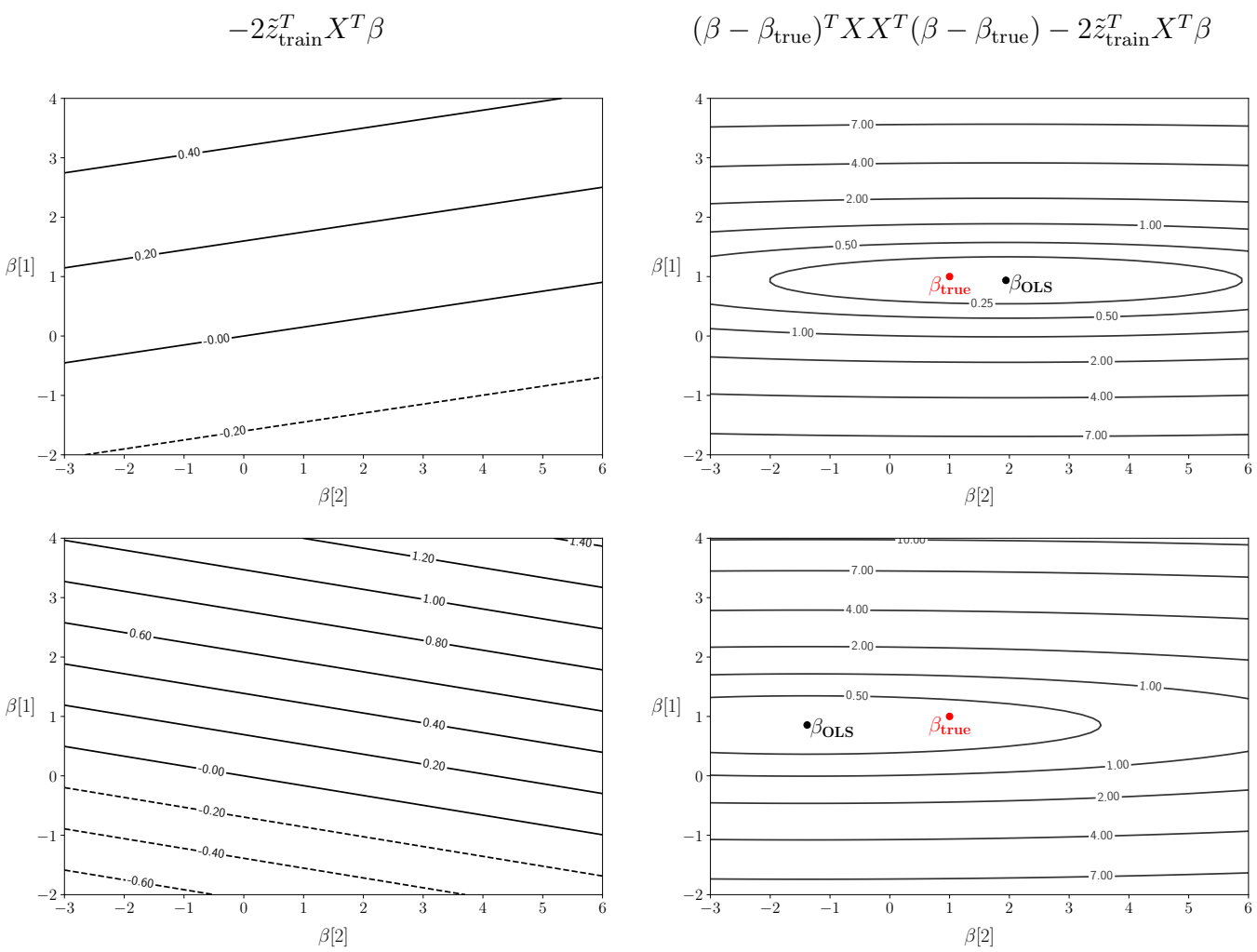


Figure 5: The left column show two realizations of the random linear component of the least-squares cost function (see Eq. (80)) for the example in Figure 4. The right column shows the corresponding cost function, which is a quadratic centered at a point that does not coincide with β_{true} due to the linear term. The minimum of the quadratic is denoted by β_{OLS}.

to the random linear component, the minimum of the least-squares cost function is not at β_{true}. Figure 4 shows this for a simple example. The following theorem shows that the minimum of the cost function is a Gaussian random vector centered at β_{true}.

Theorem 4.2. *If the training data follow the additive model in Eq. (74) and X is full rank, the OLS coefficient*

$$\tilde{\beta}_{\text{OLS}} := \arg\min_{\beta} \left|\left|\tilde{y}_{\text{train}} - X^T\beta\right|\right|_2, \tag{83}$$

is a Gaussian random vector with mean β_{true} and covariance matrix $\sigma^2 U S^{-2} U^T$, where $X = USV^T$ is the SVD of the feature matrix.

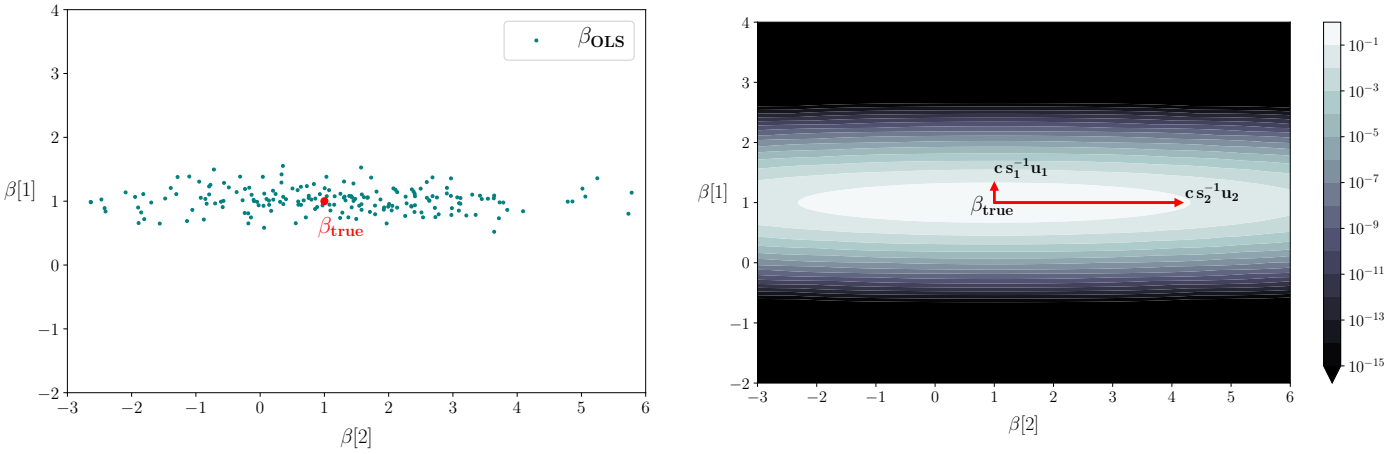


Figure 6: The left image is a scatterplot of OLS estimates corresponding to different noise realizations for the example in Figure 5. The right image is a heatmap of the distribution of the OLS estimate, which is centered at β_{true} and has covariance matrix $\sigma^2 U S^{-2} U^T$, as established in Theorem 4.2.

Proof. We have

$$\begin{aligned}
\beta_{\text{OLS}} &= (XX^T)^{-1} X \tilde{y}_{\text{train}} && (84)\\
&= (XX^T)^{-1} XX^T \beta_{\text{true}} + (XX^T)^{-1} X \tilde{z}_{\text{train}} && (85)\\
&= \beta_{\text{true}} + (XX^T)^{-1} X \tilde{z}_{\text{train}} && (86)\\
&= \beta_{\text{true}} + U S^{-1} V^T \tilde{z}_{\text{train}}. && (87)
\end{aligned}$$

The result then follows from Theorem 3.4 in the lecture notes on the covariance matrix. □

Figure 6 shows a scatterplot of the OLS estimates corresponding to different noise realizations, as well as the distribution of the OLS estimate. The contour lines of the distribution are ellipsoidal with axes aligned with the left singular vectors of the feature matrix. The variance along those axes is proportional to the inverse of the squared singular values. If there are singular values that are very small, the variance in the direction of the corresponding singular vector can be very large, as is the case along the horizontal axis of Figure 6.

4.2 Training error

In order to analyze the training error of the OLS estimator, we leverage a geometric perspective. The OLS estimator approximates the response vector y using a linear combination of the corresponding features. Each feature corresponds to a row of X. The linear coefficients weight these rows. This means that the estimator is equal to *the vector in the row space of the feature matrix X that is closest to y*. By definition, that vector is the orthogonal projection of y onto $\text{row}(X)$. Figure 7 illustrates this with a simple example with two features and three examples.

Lemma 4.3. *Let $X \in \mathbb{R}^{p \times n}$ be full-rank feature matrix, where $n \geq p$, and let $y \in \mathbb{R}^n$ be a response*

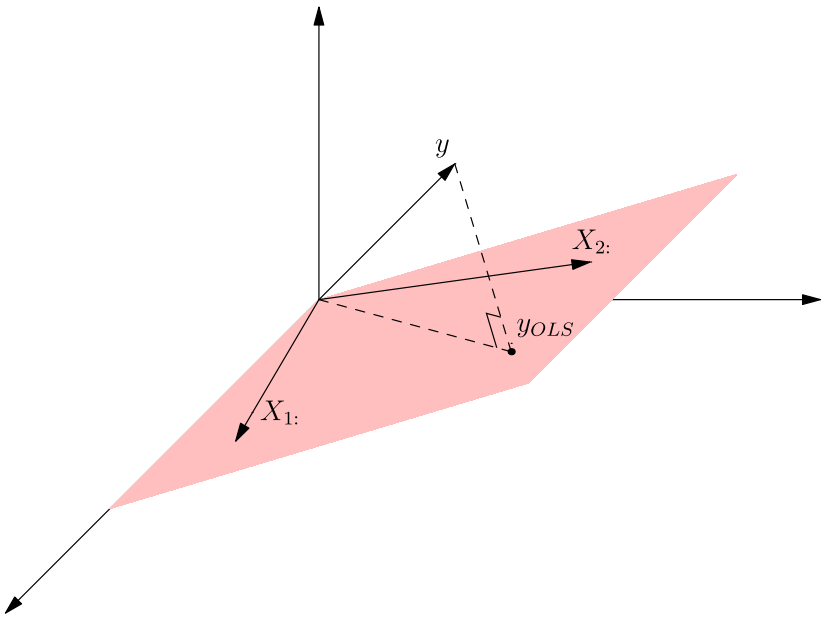


Figure 7: Illustration of Lemma 4.3 for a problem with two features corresponding to the two rows of the feature matrix $X_{1:}$ and $X_{2:}$. The least-squares solution is the orthogonal projection of the data onto the subspace spanned by these vectors.

vector. The OLS estimator $X^T\beta_{\text{OLS}}$ *of* y *given* X*, where*

$$\beta_{\text{OLS}} := \arg\min_{\beta\in\mathbb{R}^p} \left|\left|y - X^T\beta\right|\right|_2, \tag{88}$$

is equal to the orthogonal projection of y *onto the row space of* X*.*

Proof. Let USV^T be the SVD of X. By Eq. (73)

$$X^T\beta_{\text{OLS}} = X^T US^{-1}V^T y \tag{89}$$
$$= VSU^T US^{-1}V^T y \tag{90}$$
$$= VV^T y. \tag{91}$$

Since the rows of V form an orthonormal basis for the row space of X the proof is complete. □

The result provides an intuitive interpretation for the training error achieved by OLS for the additive model in Eq. (74): it is the projection of the noise vector onto the subspace spanned by the feature vectors.

Lemma 4.4. *If the training data follow the additive model in Eq. (74) and* X *is full rank, the training error of the OLS estimator* $X^T\tilde{\beta}_{\text{OLS}}$ *is the projection of the noise onto the orthogonal complement of the row space of* X*.*

Proof. By Lemma 4.3

$$\begin{aligned}
\tilde{y}_{\text{train}} - X^T \tilde{\beta}_{\text{OLS}} &= \tilde{y}_{\text{train}} - \mathcal{P}_{\text{row}(X)}\, \tilde{y}_{\text{train}} && (92)\\
&= X^T \beta_{\text{true}} + \tilde{z}_{\text{train}} - \mathcal{P}_{\text{row}(X)}\, (X^T \beta_{\text{true}} + \tilde{z}_{\text{train}}) && (93)\\
&= X^T \beta_{\text{true}} + \tilde{z}_{\text{train}} - X^T \beta_{\text{true}} - \mathcal{P}_{\text{row}(X)}\, \tilde{z}_{\text{train}} && (94)\\
&= \mathcal{P}_{\text{row}(X)^\perp}\, \tilde{z}_{\text{train}}. && (95)
\end{aligned}$$

□

We define the average training square error as the average error incurred by the OLS estimator on the training data,

$$\widetilde{E}^2_{\text{train}} := \frac{1}{n} \left|\left| \tilde{y}_{\text{train}} - X^T \tilde{\beta}_{\text{OLS}} \right|\right|_2^2. \tag{96}$$

By Lemma 4.4, if the noise is Gaussian, then the training error is the projection of an n-dimensional iid Gaussian random vector onto the subspace orthogonal to the span of the feature vectors. The iid assumption means that the Gaussian distribution is isotropic. The dimension of this subspace equals $n - p$, so the fraction of the variance in the Gaussian vector that lands on it should be approximately equal to $1 - p/n$. The following theorem establishes that this is indeed the case.

Theorem 4.5. *If the training data follow the additive model in Eq. (74) and X is full rank, then the mean of the average training error defined in Eq. (96) equals*

$$\mathrm{E}\left(\widetilde{E}^2_{\text{train}}\right) = \sigma^2 \left(1 - \frac{p}{n}\right) \tag{97}$$

and its variance equals

$$\mathrm{Var}(\widetilde{E}^2_{\text{train}}) = \frac{2\sigma^4 (n-p)}{n^2}. \tag{98}$$

Proof. By Lemma 4.4

$$\begin{aligned}
n\widetilde{E}^2_{\text{train}} &= \left|\left| \mathcal{P}_{\text{row}(X)^\perp}\, \tilde{z}_{\text{train}} \right|\right|_2^2 && (99)\\
&= \tilde{z}^T_{\text{train}} V_\perp V_\perp^T V_\perp V_\perp^T \tilde{z}_{\text{train}} && (100)\\
&= \left|\left| V_\perp^T \tilde{z}_{\text{train}} \right|\right|_2^2, && (101)
\end{aligned}$$

where the columns of $V_\perp$ are an orthonormal basis for $\text{row}(X)^\perp$. $V_\perp^T \tilde{z}_{\text{train}}$ is a Gaussian vector of dimension $n - p$ with covariance matrix

$$\begin{aligned}
\Sigma_{V_\perp^T \tilde{z}_{\text{train}}} &= V_\perp^T \Sigma_{\tilde{z}_{\text{train}}} V_\perp && (102)\\
&= V_\perp^T \sigma^2 I V_\perp && (103)\\
&= \sigma^2 I. && (104)
\end{aligned}$$

The error is therefore equal to the square ℓ_2 norm of an iid Gaussian random vector. Let $\tilde{w}$ be a d-dimensional zero-mean Gaussian random vector $\tilde{w}$ with unit variance. The expected value of its

square ℓ_2 norm is

$$\mathrm{E}\left(||\tilde{w}||_2^2\right) = \mathrm{E}\left(\sum_{i=1}^{d} \tilde{w}[i]^2\right) \tag{105}$$

$$= \sum_{i=1}^{d} \mathrm{E}\left(\tilde{w}[i]^2\right) \tag{106}$$

$$= d. \tag{107}$$

The mean square equals

$$\mathrm{E}\left[\left(||\tilde{w}||_2^2\right)^2\right] = \mathrm{E}\left[\left(\sum_{i=1}^{d} \tilde{w}[i]^2\right)^2\right] \tag{108}$$

$$= \sum_{i=1}^{d}\sum_{j=1}^{d} \mathrm{E}\left(\tilde{w}[i]^2\tilde{w}[j]^2\right) \tag{109}$$

$$= \sum_{i=1}^{d} \mathrm{E}\left(\tilde{w}[i]^4\right) + 2\sum_{i=1}^{d-1}\sum_{j=i+1}^{d} \mathrm{E}\left(\tilde{w}[i]^2\right)\mathrm{E}\left(\tilde{w}[j]^2\right) \tag{110}$$

$$= 3d + d(d-1) \quad \text{(the 4th moment of a standard Gaussian equals 3)} \tag{111}$$

$$= d(d+2), \tag{112}$$

so the variance equals

$$\mathrm{Var}\left(||\tilde{w}||_2^2\right) = \mathrm{E}\left[\left(||\tilde{w}||_2^2\right)^2\right] - \mathrm{E}^2\left(||\tilde{w}||_2^2\right) \tag{113}$$

$$= 2d. \tag{114}$$

As d grows, the relative deviation of the squared norm of the Gaussian vector from its mean decreases proportionally to $\sqrt{2/d}$, as shown in Figure 8. Geometrically, the probability density concentrates close to the surface of a sphere with radius $\sqrt{d}$. By definition of the training error, we have

$$\widetilde{E}^2_{\text{train}} = \frac{1}{n}\left|\left|V_\perp^T \tilde{z}_{\text{train}}\right|\right|_2^2 \tag{115}$$

$$= \frac{\sigma^2}{n}\left|\left|\tilde{w}\right|\right|_2^2, \tag{116}$$

so the result follows from setting $d := n - p$ in Eqs. (107) and (114). □

The variance of the square error scales with $1/n$, which implies that the error concentrates around its mean with high probability as the number of examples in the training set grows.

For large n, the error equals the variance of the noisy component σ^2, which is the error achieved by the true coefficients β_{true}. When $n \approx p$, however, the error can be much smaller. This is bad news. We cannot possibly fit the noisy component of the response using the features, because they are

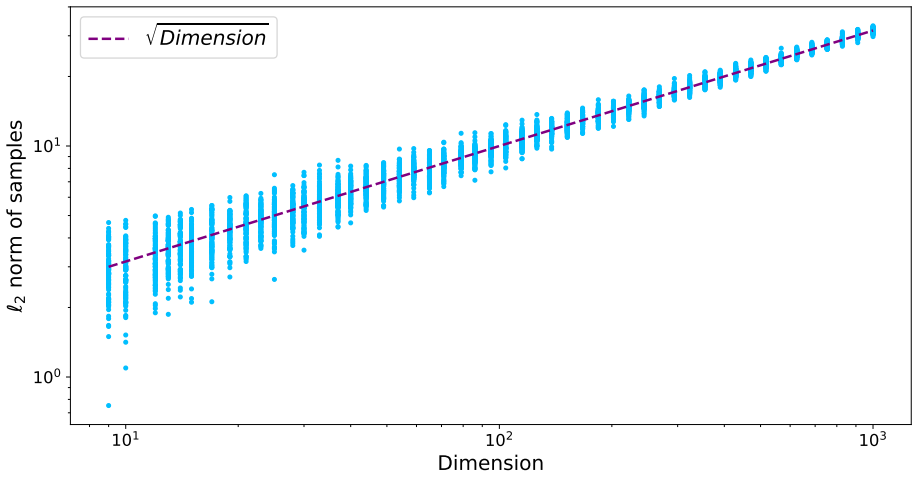


Figure 8: The graphs shows the ℓ_2 norm of 100 independent samples from standard Gaussian random vectors in different dimensions. The norms of the samples concentrate around the square root of the dimension.

independent. Therefore, if an estimator achieves an error of less than σ it must be *overfitting* the training noise, which will result in a higher generalization error on held-out data. This suggests that the OLS estimator overfits the training data when the number of examples is small with respect to the number of features. Figure 9 shows that this is indeed the case for the dataset in Example 2.2. In fact, the training error is proportional to $\left(1 - \frac{p}{n}\right)$ as predicted by our theoretical result.

4.3 Test error

The test error of an estimator quantifies its performance on held-out data, which have not been used to fit the model. We model the test data as

$$\tilde{y}_{\text{test}} := \tilde{x}_{\text{test}}^T \beta_{\text{true}} + \tilde{z}_{\text{test}}. \tag{117}$$

The linear coefficients are the same as in the training set, but the features and noise are different. The features are modeled as a p-dimensional random vector $\tilde{x}_{\text{test}}$ with zero mean (the features are assumed to be centered) and the noise $\tilde{z}_{\text{test}}$ is a zero-mean Gaussian random variable with the same variance σ^2 as the training noise. The training and test noise are assumed to be independent from each other and from the features. Our goal is to characterize the test error

$$\widetilde{E}_{\text{test}} := \tilde{y}_{\text{test}} - \tilde{x}_{\text{test}}^T \tilde{\beta}_{\text{OLS}} \tag{118}$$

$$= \tilde{z}_{\text{test}} + \tilde{x}_{\text{test}}^T \left(\beta_{\text{true}} - \tilde{\beta}_{\text{OLS}}\right), \tag{119}$$

where $\tilde{\beta}_{\text{OLS}}$ is computed from the training data.

Theorem 4.6 (Test mean square error)**.** *If the training data follow the additive model in Eq. (74), X is full rank, and the test data follow the model in Eq. (117), then the mean square of the test*

error equals

$$\mathrm{E}(\widetilde{E}_{\text{test}}^2) = \sigma^2 \left(1 + \sum_{i=1}^{p} \frac{\mathrm{Var}(u_i^T \tilde{x}_{\text{test}})}{s_i^2}\right), \tag{120}$$

where $\Sigma_{\tilde{x}_{\text{test}}}$ *is the covariance matrix of the feature vector,* $s_1, \ldots, s_p$ *are the singular values of* X *and* $v_1, \ldots, v_p$ *are the right singular vectors.*

Proof. By assumption, the two components of the test error in Eq. (119) are independent, so the variance of their sum is the sum of their variances:

$$\mathrm{Var}\left(\tilde{y}_{\text{test}} - \tilde{x}_{\text{test}}^T \tilde{\beta}_{\text{OLS}}\right) = \sigma^2 + \mathrm{Var}\left(\tilde{x}_{\text{test}}^T \left(\beta_{\text{true}} - \tilde{\beta}_{\text{OLS}}\right)\right) \tag{121}$$

Since everything is zero mean, this also holds for the mean square. Let USV^T be the SVD of X. The coefficient error equals

$$\beta_{\text{OLS}} - \beta_{\text{true}} = \sum_{i=1}^{p} \frac{v_i^T \tilde{z}_{\text{train}}}{s_i} u_i, \tag{122}$$

by Theorem 4.2. This implies

$$\mathrm{E}\left[\left(\tilde{x}_{\text{test}}^T \left(\beta_{\text{true}} - \tilde{\beta}_{\text{OLS}}\right)\right)^2\right] = \mathrm{E}\left[\left(\sum_{i=1}^{p} \frac{v_i^T \tilde{z}_{\text{train}}\, u_i^T \tilde{x}_{\text{test}}}{s_i}\right)^2\right] \tag{123}$$

$$= \sum_{i=1}^{p} \frac{\mathrm{E}\left[(v_i^T \tilde{z}_{\text{train}})^2\right] \mathrm{E}\left[(u_i^T \tilde{x}_{\text{test}})^2\right]}{s_i^2}, \tag{124}$$

where the second equality holds because when we expand the square, the cross terms cancel due to the independence assumptions and linearity of expectation. For $i \neq j$

$$\mathrm{E}\left(\frac{v_i^T \tilde{z}_{\text{train}}\, u_i^T \tilde{x}_{\text{test}}}{s_i} \frac{v_j^T \tilde{z}_{\text{train}}\, u_j^T \tilde{x}_{\text{test}}}{s_j}\right) = \frac{\mathrm{E}\left(u_i^T \tilde{x}_{\text{test}} u_j^T \tilde{x}_{\text{test}}\right)}{s_i s_j} v_i^T \mathrm{E}\left(\tilde{z}_{\text{train}} \tilde{z}_{\text{train}}^T\right) v_j \tag{125}$$

$$= \frac{\mathrm{E}\left(u_i^T \tilde{x}_{\text{test}} u_j^T \tilde{x}_{\text{test}}\right)}{s_i s_j} v_i^T v_j \tag{126}$$

$$= 0. \tag{127}$$

By linearity of expectation, we conclude

$$\mathrm{E}\left[\left(\tilde{x}_{\text{test}}^T \left(\beta_{\text{true}} - \tilde{\beta}_{\text{OLS}}\right)\right)^2\right] = \sum_{i=1}^{p} \frac{v_i^T \mathrm{E}(\tilde{z}_{\text{train}} \tilde{z}_{\text{train}}^T) v_i u_i^T \mathrm{E}(\tilde{x}_{\text{test}} \tilde{x}_{\text{test}}^T) u_i}{s_i^2} \tag{128}$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{u_i^T \Sigma_{\tilde{x}_{\text{test}}} u_i}{s_i^2}, \tag{129}$$

because the covariance matrix of the training noise $\tilde{z}_{\text{train}}$ equals $\sigma^2 I$. □

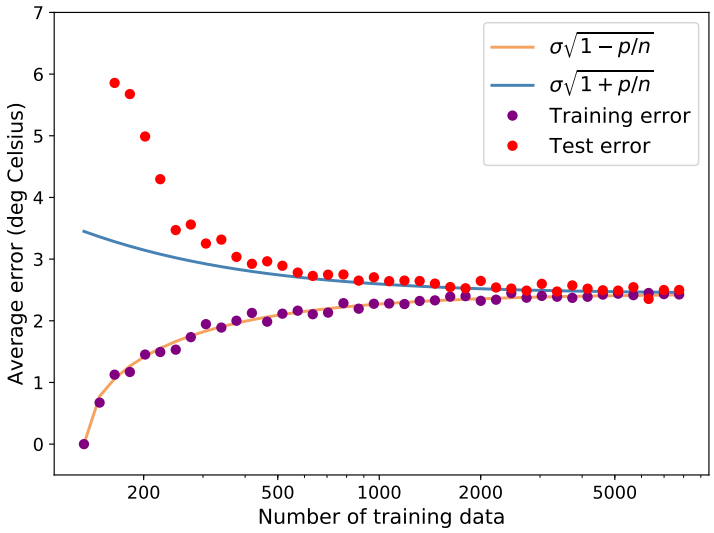


Figure 9: Comparison of the theoretical approximation for the training and test error of the OLS with the actual errors on the temperature data described in Example 2.2. The parameter σ is fixed based on the asymptotic value of the error.

The square of the ith singular value of the training covariance matrix is proportional to the sample variance of the training data in the direction of the ith singular value u_i,

$$\frac{s_i^2}{n} = \frac{u_i X X^T u_i}{n} \tag{130}$$

$$= u_i^T \Sigma_{\mathcal{X}} u_i \tag{131}$$

$$= \operatorname{var}\left(\mathcal{P}_{u_i} \mathcal{X}\right). \tag{132}$$

If this sample variance is a good approximation to the variance of the test data in that direction then

$$\mathrm{E}(\widetilde{E}_{\mathrm{test}}^2) \approx \sigma^2 \left(1 + \frac{p}{n}\right). \tag{133}$$

However, if the training data is not large enough, the sample covariance matrix of the training data may not provide a good estimate of the feature variance in every direction. In that case, there may be terms in the test error where s_i is very small, due to correlations between the features, but the true directional variance is not. Figure 10 shows that some of the singular values of the training matrix in the temperature prediction are indeed minuscule. Unless the test variance in that direction cancel them out, this results in a large test error.

Intuitively, estimating the contribution of low-variance components of the feature vector to the linear coefficients requires amplifying them. This also amplifies the training noise in those directions. When estimating the response, this amplification is neutralized by the corresponding small directional variance of the test features as long as it occurs in the right directions (which is the case if the sample covariance matrix is a good approximation to the test covariance matrix). Otherwise, it will result in a high response error. This typically occurs when the number of training

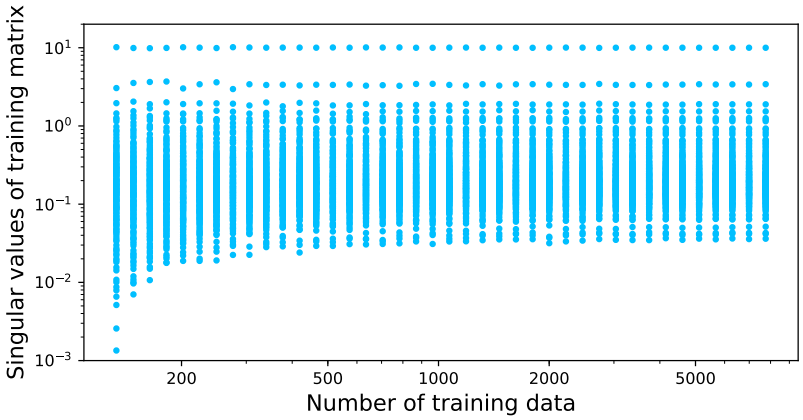


Figure 10: Singular values of the training matrix in Example 2.2 for different numbers of training data.

data is small with respect to the number of features. The effect is apparent in Figure 2 for small values of n.