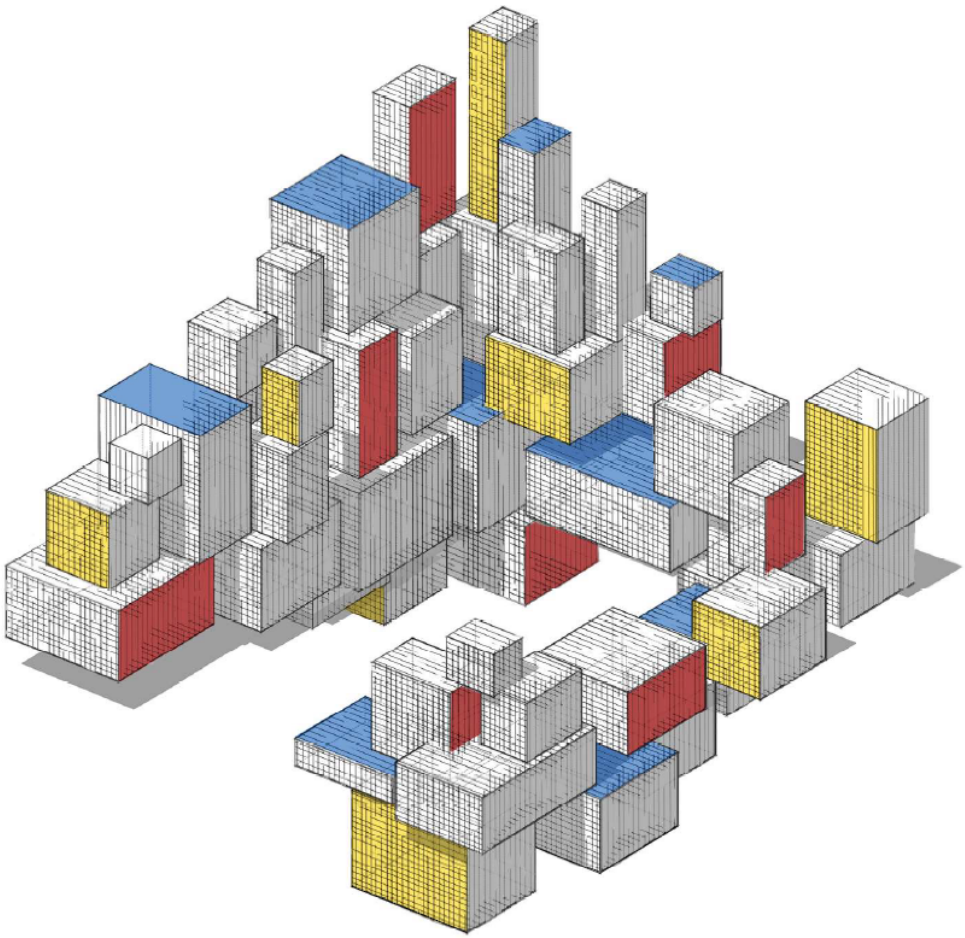
Estimating Risk



Purpose

In this lecture we discuss method to estimate the generalization risk:

- Test loss
- In-Sample Risk
- Cross-Validation

Estimating Risk

For a training set τ, prediction function class $\mathcal{G}$, and learner $g_\tau^{\mathcal{G}}$, the generalization risk is the expected loss $\ell(g_\tau^{\mathcal{G}}) = \mathbb{E}\text{Loss}(Y, g_\tau^{\mathcal{G}}(\boldsymbol{X}))$.

The most straightforward way to quantify the generalization risk is to estimate it via the test loss

$$\ell_{\mathcal{T}'}(g_\tau^{\mathcal{G}}) := \frac{1}{n'} \sum_{i=1}^{n'} \text{Loss}(Y_i', g_\tau^{\mathcal{G}}(\boldsymbol{X}_i')),$$

where $\{(\boldsymbol{X}_1', Y_1'), \ldots, (\boldsymbol{X}_{n'}', Y_{n'}')\} =: \mathcal{T}'$ is a test sample.

The generalization risk depends inherently on the training set, and so different training sets may yield significantly different estimates.

When there is a limited amount of data available, reserving a substantial proportion of the data for testing rather than training may be uneconomical.

In-Sample Risk

Due to the phenomenon of overfitting, the training loss of the learner, $\ell_\tau(g_\tau)$ is not a good estimate of the generalization risk $\ell(g_\tau)$ of the learner.

One reason for this is that we use the same data for both training the model and assessing its risk.

An alternative is to estimate the in-sample risk of the learner g_τ:

$$\ell_{\text{in}}(g_\tau) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\,\text{Loss}(Y_i', g_\tau(\boldsymbol{x}_i)), \tag{1}$$

where each response Y_i' is drawn from $f(y \mid \boldsymbol{x}_i)$, independently. This can be estimated by the sample average

$$\frac{1}{n}\sum_{i=1}^{n} \text{Loss}(Y_i', g_\tau(\boldsymbol{x}_i)).$$

Optimism

For a fixed training set τ, we can compare the training loss of the learner with the in-sample risk. Their difference,

$$\mathrm{op}_\tau = \ell_{\mathrm{in}}(g_\tau) - \ell_\tau(g_\tau),$$

is called the optimism (of the training loss), because it measures how much the training loss underestimates (is optimistic about) the unknown in-sample risk. Mathematically, it is simpler to work with the expected optimism:

$$\mathbb{E}[\mathrm{op}_{\mathcal{T}} \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n] =: \mathbb{E}_{\mathbf{X}}\, \mathrm{op}_{\mathcal{T}},$$

where the expectation is taken over a random training set $\mathcal{T}$, conditional on $\boldsymbol{X}_i = \boldsymbol{x}_i$, $i = 1, \ldots, n$.

In the simpified notation, $\mathbb{E}_{\mathbf{X}}$ denotes the expectation operator conditional on $\boldsymbol{X}_i = \boldsymbol{x}_i, i = 1, \ldots, n$.

The expected optimism can be expressed in terms of the (conditional) covariance between the observed and predicted response.

Theorem: Expected Optimism

For the squared-error loss and 0–1 loss with 0–1 response, the expected optimism is

$$\mathbb{E}_{\mathbf{X}}\,\mathrm{op}_{\mathcal{T}} = \frac{2}{n}\sum_{i=1}^{n}\mathbb{C}\mathrm{ov}_{\mathbf{X}}(g_{\mathcal{T}}(\boldsymbol{x}_i), Y_i). \tag{2}$$

Proof

All expectations are taken conditional on $\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n$. Let Y_i be the response for $\boldsymbol{x}_i$ and let $\widehat{Y}_i = g_{\mathcal{T}}(\boldsymbol{x}_i)$ be the predicted value. Note that the latter depends on $Y_1, \ldots, Y_n$. Also, let Y_i' be an independent copy of Y_i for the *same* $\boldsymbol{x}_i$, as in (1). In particular, Y_i' has the same distribution as Y_i and is statistically independent of all $\{Y_j\}$, including Y_i, and therefore is also independent of $\widehat{Y}_i$. We have

$$\begin{aligned}\mathbb{E}_{\mathbf{X}} \operatorname{op}_{\mathcal{T}} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathbf{X}}\left[(Y_i' - \widehat{Y}_i)^2 - (Y_i - \widehat{Y}_i)^2\right] = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathbf{X}}\left[(Y_i - Y_i')\widehat{Y}_i\right] \\ &= \frac{2}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{\mathbf{X}}[Y_i\widehat{Y}_i] - \mathbb{E}_{\mathbf{X}}Y_i\,\mathbb{E}_{\mathbf{X}}\widehat{Y}_i\right) = \frac{2}{n}\sum_{i=1}^{n} \mathbb{C}\mathrm{ov}_{\mathbf{X}}(\widehat{Y}_i, Y_i).\end{aligned}$$

The expected optimism indicates how much the training loss deviates from the expected in-sample risk. Since the covariance of independent random variables is zero, the expected optimism is zero if the learner $g_{\mathcal{T}}$ is statistically independent from the responses $Y_1, \ldots, Y_n$.

Example: Polynomial Regression (cont.)

We continue the polynomial regression example, where the components of the response vector $\boldsymbol{Y} = [Y_1, \ldots, Y_n]^\top$ are independent and normally distributed with variance $\ell^* = 25$ (the irreducible error) and expectations $\mathbb{E}Y_i = g^*(\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}^*$, $i = 1, \ldots, n$. Using the least-squares estimator $\widehat{\boldsymbol{\beta}} = \mathbf{X}^+ \boldsymbol{y}$, the expected optimism (2) is

$$\frac{2}{n}\sum_{i=1}^{n} \mathbb{C}\mathrm{ov}_{\mathbf{X}}\left(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}, Y_i\right) = \frac{2}{n}\mathrm{tr}\left(\mathbb{C}\mathrm{ov}_{\mathbf{X}}\left(\mathbf{X}\widehat{\boldsymbol{\beta}}, \boldsymbol{Y}\right)\right) = \frac{2}{n}\mathrm{tr}\left(\mathbb{C}\mathrm{ov}_{\mathbf{X}}\left(\mathbf{X}\mathbf{X}^+\boldsymbol{Y}, \boldsymbol{Y}\right)\right)$$

$$= \frac{2\mathrm{tr}\left(\mathbf{X}\mathbf{X}^+\mathbb{C}\mathrm{ov}_{\mathbf{X}}\left(\boldsymbol{Y}, \boldsymbol{Y}\right)\right)}{n} = \frac{2\ell^*\mathrm{tr}\left(\mathbf{X}\mathbf{X}^+\right)}{n} = \frac{2\ell^* p}{n}.$$

Here, we used the cyclic property of the trace:
$\mathrm{tr}(\mathbf{X}\mathbf{X}^+) = \mathrm{tr}(\mathbf{X}^+\mathbf{X}) = \mathrm{tr}(\mathbf{I}_p)$, assuming that $\mathrm{rank}(\mathbf{X}) = p$.

Therefore, an estimate for the in-sample risk (1) is:

$$\widehat{\ell}_{\text{in}}(g_\tau) = \ell_\tau(g_\tau) + 2\ell^* p/n. \tag{3}$$

Instead of computing the test loss to assess the best model complexity p, we could simply have minimized the training loss plus the correction term $2\ell^* p/n$. In practice, ℓ^* also has to be estimated.

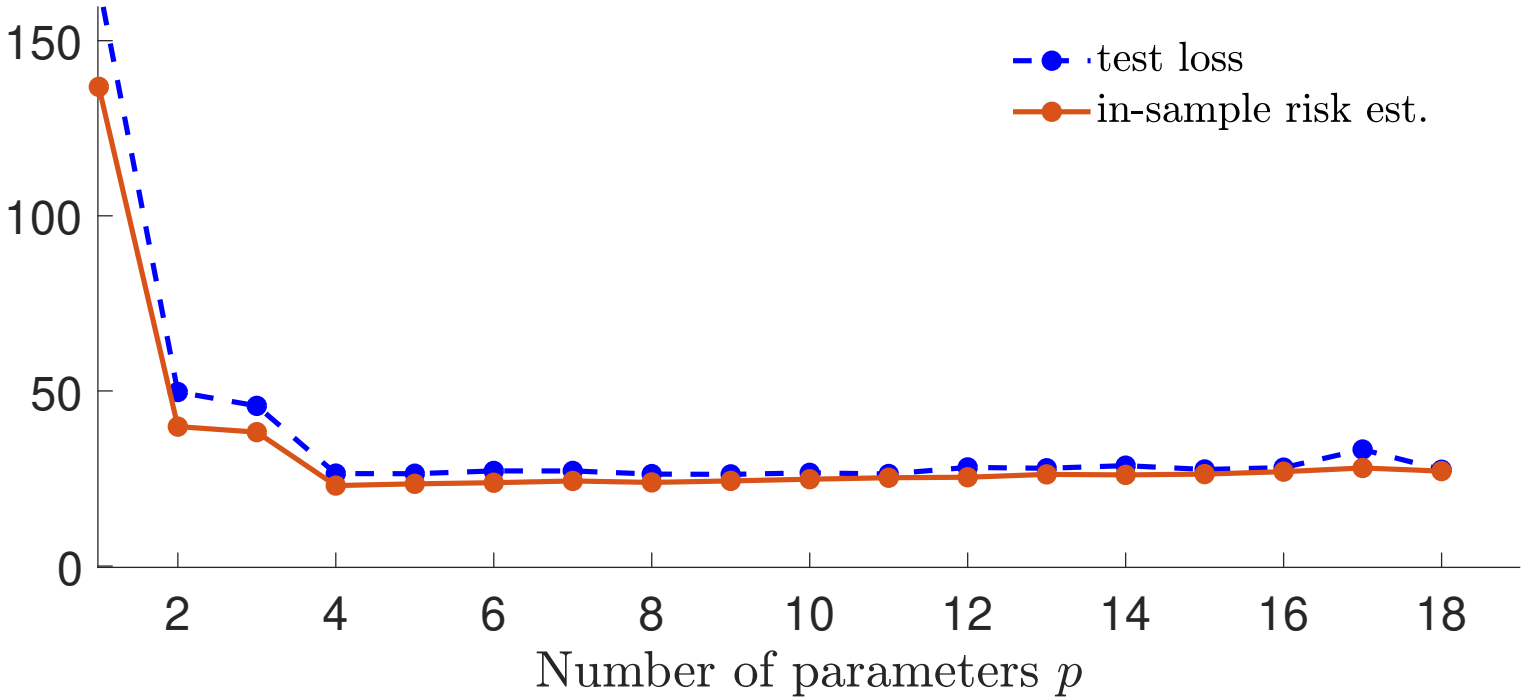


Figure: In-sample risk estimate $\widehat{\ell}_{\text{in}}(g_\tau)$ (red) and test loss (blue) as a function of the number of parameters p of the model.

Cross-Validation

In general, for complex function classes $\mathcal{G}$, it is very difficult to derive simple formulas of the approximation and statistical errors, let alone for the generalization risk or expected generalization risk.

When there is an abundance of data, the easiest way to assess the generalization risk for a given training set τ is to obtain a test set τ' and evaluate the test loss:

$$\ell_{\tau'}(g_\tau^{\mathcal{G}}) := \frac{1}{n'} \sum_{i=1}^{n'} \text{Loss}(y_i', g_\tau^{\mathcal{G}}(\boldsymbol{x}_i')).$$

When a sufficiently large test set is not available but computational resources are cheap, one can instead gain direct knowledge of the expected generalization risk via a computationally intensive method called cross-validation.

Cross-Validation

- Make multiple identical copies of the data set, and partition each copy into different training (blue) and test sets (pink).
- For each of these sets, estimate the model parameters using only training data and then predict the responses for the test set.
- The average loss between the predicted and observed responses is then a measure for the predictive power of the model.

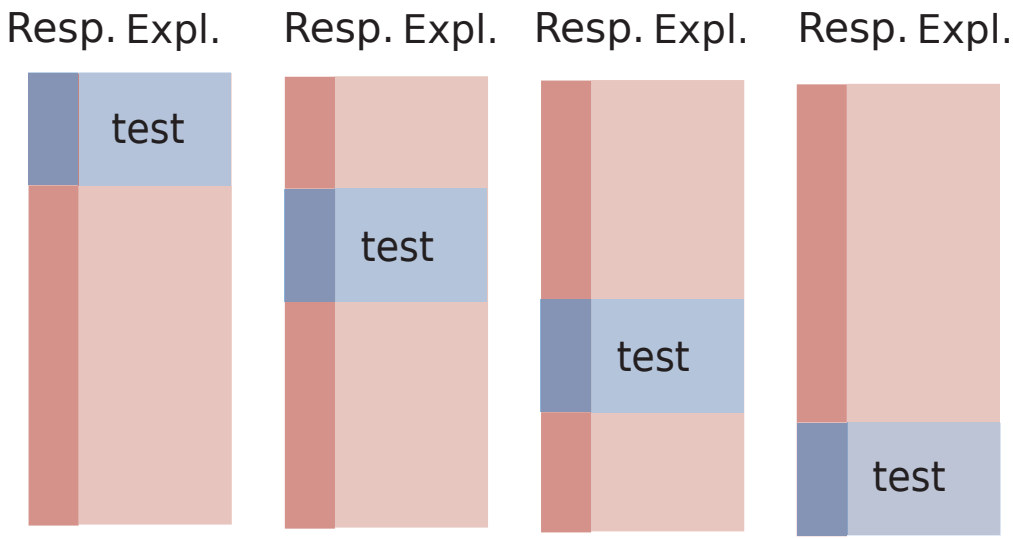


Figure: Four-fold cross-validation.

Partition a data set $\mathcal{T}$ of size n into K folds $C_1, \ldots, C_K$ of sizes $n_1, \ldots, n_K$. Typically $n_k \approx n/K$, $k = 1, \ldots, K$.

Let ℓ_{C_k} be the test loss when using C_k as test data and all remaining data, denoted $\mathcal{T}_{-k}$, as training data.

Each ℓ_{C_k} is an unbiased estimator of the generalization risk for training set $\mathcal{T}_{-k}$; that is, for $\ell(g_{\mathcal{T}_{-k}})$.

The K-fold cross-validation loss is:

$$\mathrm{CV}_K := \sum_{k=1}^{K} \frac{n_k}{n} \ell_{C_k}(g_{\mathcal{T}_{-k}}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Loss}(g_{\mathcal{T}_{-\kappa(i)}}(\boldsymbol{x}_i), y_i),$$

where $\kappa(i)$ indicates to which of the K folds observation i belongs.

As the average is taken over varying training sets $\{\mathcal{T}_{-k}\}$, it estimates the expected generalization risk $\mathbb{E}\,\ell(g_{\mathcal{T}})$, rather than the generalization risk $\ell(g_\tau)$ for the particular training set τ.

Polynomial Regression (cont.)

For the polynomial regression example, we can calculate a K-fold cross-validation loss with a nonrandom partitioning of the training set using the following code.

polyregCV.py

```
from polyreg3 import *

K_vals = [5, 10, 100]  # number of folds
cv = np.zeros((len(K_vals), max_p))   # cv loss
X = np.ones((n, 1))



... see GitHub code ...
```

The figure shows the cross-validation loss for $K \in \{5, 10, 100\}$. The case $K = 100$ corresponds to the leave-one-out cross-validation.

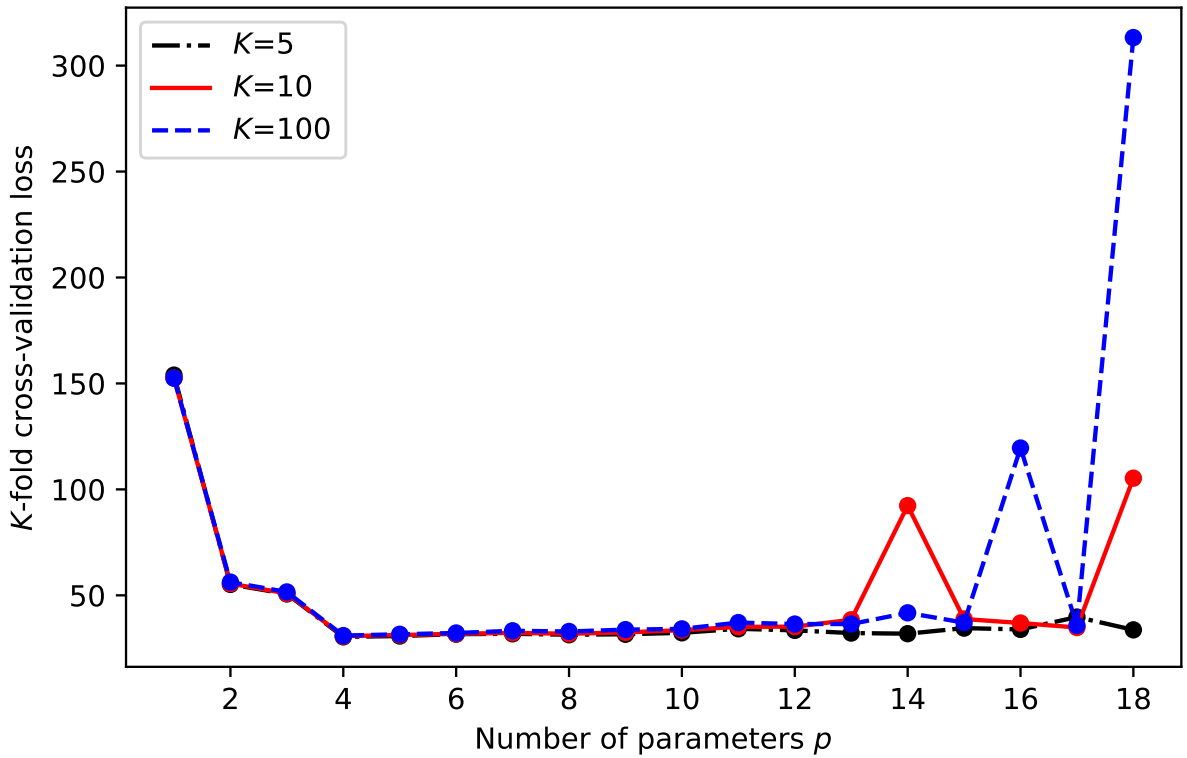


Figure: K-fold cross-validation for the polynomial regression example.