16 - Regression and learning

- Linear regression
- The linear estimator with smallest MSE
- Bias of the LMMSE
- Performance of the LMMSE
- The correlation coefficient
- Example: LMMSE when the pdf is non-Gaussian
- Polynomial regression
- Example: polynomial regression
- Learning an estimator based on data
- SLMMSE and least squares
- Example: SLMMSE
- Example: sample polynomial estimators
- Validation

Linear regression

Find the *linear* estimator $\phi(y) = Ly + c$ that minimizes

$$\boxed{\mathbf{E}\big(\|x - \phi(y)\|^2\big)}$$

- $x : \Omega \to \mathbb{R}^n$ and $y : \Omega \to \mathbb{R}^m$ are random variables.
- we can choose $L \in \mathbb{R}^{n \times m}$ and $c \in \mathbb{R}^n$
- ϕ is called the *regression* function
- called the LMMSE problem

Linear regression

Find the *linear* estimator $\phi(y) = Ly + c$ that minimizes

$$\boxed{\mathbf{E}\big(\|x - \phi(y)\|^2\big)}$$

We restrict our search to linear function $\phi : \mathbb{R}^m \to \mathbb{R}^n$

- we'll see that we don't need to know the pdf to find the optimal linear estimator
- the best linear estimator is often easier to compute than the MMSE
- if x, y are jointly Gaussian, then the best linear estimator is the MMSE estimator, which happens to be linear in this case

The linear estimator with smallest MSE

Define the error

$$\begin{aligned} z &= \phi(y) - x \\ &= \begin{bmatrix} -I & L \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + c \end{aligned}$$

The mean and covariance of z are

$$\mathbf{E}\, z = \begin{bmatrix} -I & L \end{bmatrix} \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} + c$$

$$\mathbf{cov}\, z = \begin{bmatrix} -I & L \end{bmatrix} \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} \begin{bmatrix} -I \\ L^T \end{bmatrix}$$

The linear estimator with smallest MSE

Using the mean-variance decomposition, the MSE is

$$\mathbf{E}\left(\|z\|^2\right) = \|\mathbf{E}\, z\|^2 + \mathbf{trace}(\mathbf{cov}(z))$$

$$= \left\| \begin{bmatrix} -I & L \end{bmatrix} \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} + c \right\|^2 + \mathbf{trace}\left(\begin{bmatrix} -I & L \end{bmatrix} \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} \begin{bmatrix} -I \\ L^T \end{bmatrix} \right)$$

Hence the optimal c is

$$\boxed{c_{\mathsf{opt}} = \mu_x - L\mu_y}$$

The linear estimator with smallest MSE

By completion of squares, we have

$$\begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} = \begin{bmatrix} I & \Sigma_{xy}\Sigma_y^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx} & 0 \\ 0 & \Sigma_y \end{bmatrix} \begin{bmatrix} I & 0 \\ \Sigma_y^{-1}\Sigma_{yx} & I \end{bmatrix}$$

so the MSE is

$$\begin{aligned} \mathbf{E}\big(\|z\|^2\big) &= \mathbf{trace}(\mathbf{cov}(z)) \\ &= \mathbf{trace}(\Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}) + \mathbf{trace}\big((L - \Sigma_{xy}\Sigma_y^{-1})\Sigma_y(L - \Sigma_{xy}\Sigma_y^{-1})^T\big) \end{aligned}$$

and hence the optimal L is

$$\boxed{L_{\mathsf{opt}} = \Sigma_{xy}\Sigma_y^{-1}}$$

Summary: the linear estimator with minimum MSE

The linear estimator with minimum mean square error is

$$\boxed{\phi_{\mathsf{lmmse}}(y) = \mu_x + \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y)}$$

and it achieves a mean square error of

$$\boxed{\mathbf{E}\left(\|x - \phi_{\mathsf{lmmse}}(y)\|^2\right) = \mathbf{trace}(\Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx})}$$

- Called the *linear minimum mean-square-error estimator* (LMMSE)
- Has minimum MSE among all linear estimators, even when x, y not Gaussian
- Only depends on the mean and covariance of x, y; we don't need the pdf.
- The LMMSE is the same as the MMSE when x, y are jointly Gaussian

Bias of the LMMSE

We have

$$\mathbf{E}(x - \phi_{\mathsf{lmmse}}(y)) = \mathbf{E}\big(x - \mu_x - \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y)\big)$$

$$= 0$$

and so the LMMSE is *unbiased*

Performance of the LMMSE

The optimal LMMSE satisfies

$$\mathbf{E}\big(\|x - \phi(y)\|^2\big) \leq \mathbf{trace}\,\mathbf{cov}(x)$$

- Because the trivial estimator $\phi(y) = \mu_x$ achieves an MSE of $\mathbf{cov}(x)$.
- $\mathbf{trace}\,\mathbf{cov}(x)$ is a measure of how much variation there is in x
- $\mathbf{E}\big(\|x - \phi(y)\|^2\big)$ is a measure of how much x varies about our estimate $\phi(y)$

The correlation coefficient

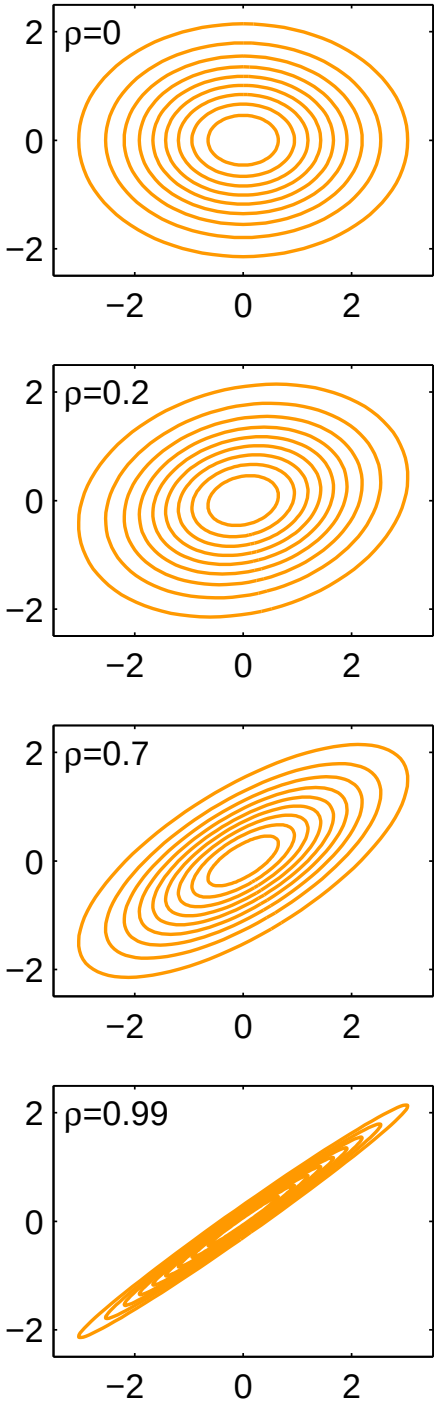
For scalar x, y, the LMMSE achieves

$$\mathbf{E}\left(\|x - \phi_{\text{lmmse}}(y)\|^2\right) = (1 - \rho^2)\Sigma_x$$

- ρ is the *correlation coefficient*

$$\rho = \frac{\Sigma_{xy}}{\sqrt{\Sigma_y \Sigma_x}}$$

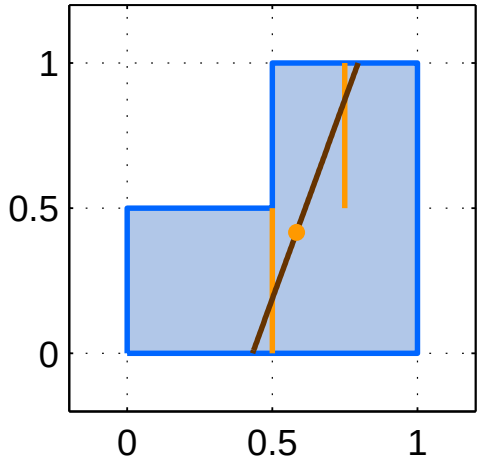
- We interpret ρ as a measure of how close x is to a linear function of y.
- When $\rho = 1$, the LMMSE achieves an MSE of zero
- When $\rho = 0$, there is no linear estimator that does better than simply $\phi(y) = \mu_x$



Example: LMMSE for a non-Gaussian PDF

(x, y) are uniformly distributed on the L-shaped region A, i.e., the pdf is

$$f(x,y) = \begin{cases} \frac{4}{3} & \text{if } (x,y) \in A \\ 0 & \text{otherwise} \end{cases}$$



- By integration we have

$$\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} = \begin{bmatrix} \frac{7}{12} \\ \frac{5}{12} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \frac{11}{144} & \frac{1}{36} \\ \frac{1}{36} & \frac{11}{144} \end{bmatrix}$$

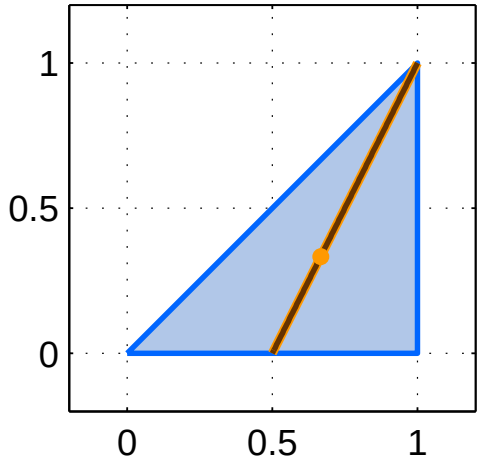
- The LMMSE is

$$\hat{x}_{\text{lmmse}} = \frac{7}{12} + \frac{4}{11}\left(y - \frac{5}{12}\right)$$

- It achieves an MSE of $\frac{35}{528} \approx 0.0663$, slightly larger than the 0.0625 achieved by the nonlinear MMSE estimator.

Example: LMMSE for non-Gaussian PDF

Let's return to the uniform pdf on the triangle.



- By integration we have

$$\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ \frac{1}{3} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \frac{1}{18} & \frac{1}{36} \\ \frac{1}{36} & \frac{1}{18} \end{bmatrix}$$

- The LMMSE is

$$\hat{x}_{\mathsf{lmmse}} = \frac{1}{3} + \frac{1}{2}\left(y - \frac{2}{3}\right)$$

  Since the MMSE is linear, it is the same as the LMMSE (but computed differently)

- The MSE is $\frac{1}{24}$

More general regression

Find the estimator of the form

$$\phi(y) = a_1 f_1 + a_2 f_2 + \cdots + a_d f_d$$

which minimizes the *mean square error*.

- The functions $f_j : \mathbb{R}^m \to \mathbb{R}^n$ are called *regressors*
- Often we don't need the pdf; just the expected value of particular functions

Polynomial regression

For scalar x and y, find the degree d polynomial estimator with minimum MSE.

$$\phi(y) = a_0 + a_1 y + a_2 y^2 + \cdots + a_d y^d$$

The MSE is

$$\begin{aligned}
\mathbf{E}(a_0 + a_1 y + \cdots + a_d y^d - x)^2 &= \mathbf{E}(a^T q - x)^2 \\
&= a^T \mathbf{E}(qq^T) a - 2a^T \mathbf{E}(qx) + \mathbf{E}\, x^2 \\
&= \begin{bmatrix} a \\ 1 \end{bmatrix}^T \begin{bmatrix} \mathbf{E}(qq^T) & \mathbf{E}(qx) \\ \mathbf{E}(qx)^T & \mathbf{E}(x^2) \end{bmatrix} \begin{bmatrix} a \\ 1 \end{bmatrix}
\end{aligned}$$

where

$$a = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{bmatrix} \qquad q = \begin{bmatrix} 1 \\ y \\ \vdots \\ y^d \end{bmatrix} \qquad qq^T = \begin{bmatrix} 1 & y & y^2 & \dots \\ y & y^2 & & \\ y^2 & & & \\ \vdots & & & \end{bmatrix}$$

Polynomial regression

The optimal a is

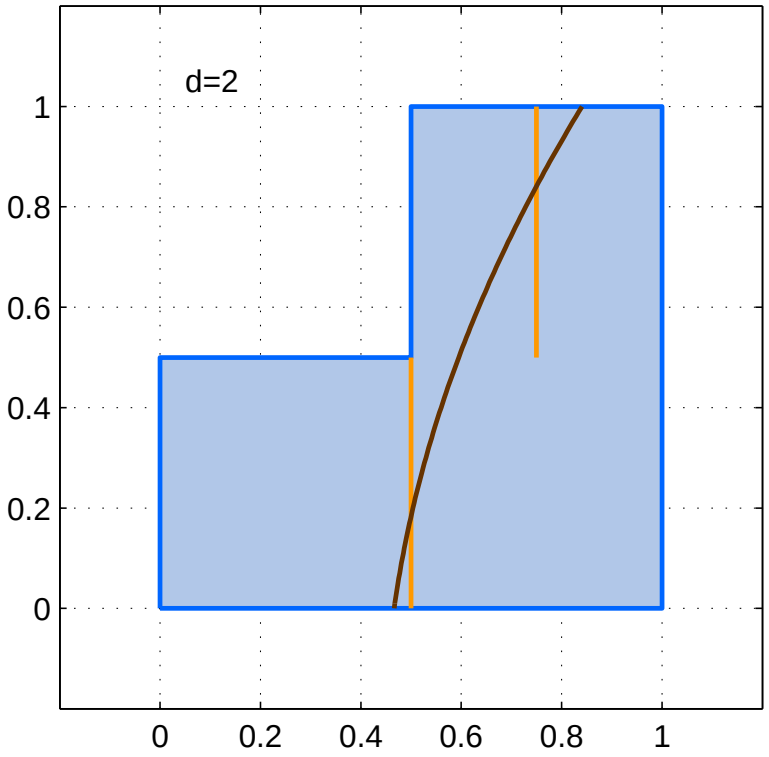
$$\boxed{a_{\text{opt}} = -\left(\mathbf{E}(qq^T)\right)^{-1}\mathbf{E}(qx)}$$

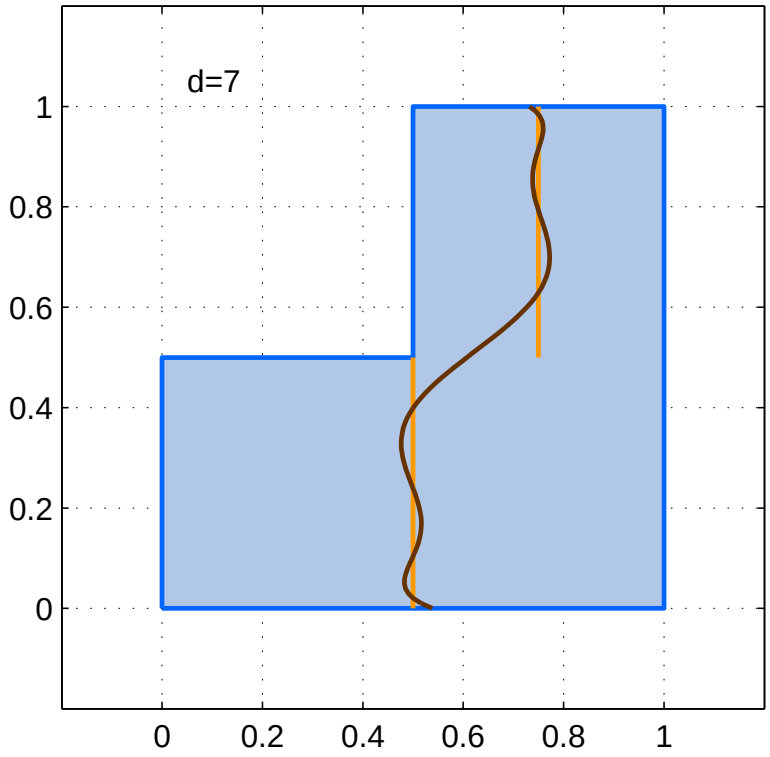
- We need to know the *moments*

$$\mathbf{E}\,qx = \begin{bmatrix} \mathbf{E}\,x \\ \mathbf{E}\,xy \\ \vdots \\ \mathbf{E}\,xy^d \end{bmatrix} \qquad \mathbf{E}\,qq^T = \begin{bmatrix} 1 & \mathbf{E}\,y & \mathbf{E}\,y^2 & \dots \\ \mathbf{E}\,y & \mathbf{E}\,y^2 & & \\ \mathbf{E}\,y^2 & & & \\ \vdots & & & \end{bmatrix}$$

Example: polynomial regression

For the same L-shaped region as before, the polynomial MMSE estimators of degree 2 and 7 are below.





Learning an estimator from data

Find the linear estimator $\phi(y) = Ly + c$ that minimizes the *sample MSE*

$$\frac{1}{N}\sum_{i=1}^{N}\|x_i - \phi(y_i)\|^2$$

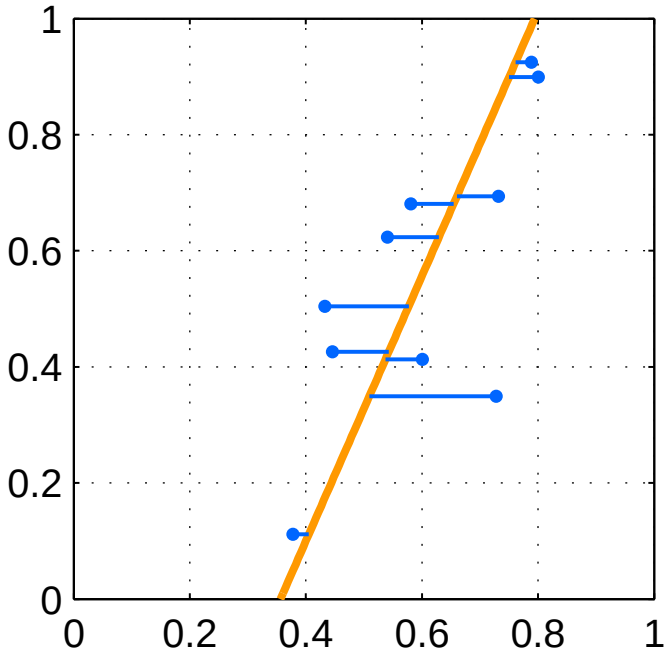
- We are given N data pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
- We do not know the pdf of (x, y)
- We are *learning* an estimator based on the data
- The sample MSE is approximately equal to the true MSE

Least squares

The estimator that minimizes the sample MSE is the *least-squares fit* of the data

Because with linear estimator $\phi(y) = Ly + c$, the sample MSE is

$$\frac{1}{N}\sum_{i=1}^{N}\|x_i - Ly_i - c\|^2$$



Called *learning* the estimator, or the *method of moments*, or the *sample LMMSE*, or *least-squares*

The SLMMSE (i.e., least squares)

The linear estimator which minimizes the *sample MSE* is

$$\phi_{\mathsf{slmmse}}(y) = \bar{x} + R_{xy}R_y^{-1}(y - \bar{y})$$

it achieves a sample MSE of

$$\frac{1}{N}\sum_{i=1}^{N}\left(\|x - \phi_{\mathsf{slmmse}}(y)\|^2\right) = \mathbf{trace}(R_x - R_{xy}R_y^{-1}R_{yx})$$

- We use the following mean and covariance; note the constant is $1/N$ not $1/(N-1)$.

$$\begin{bmatrix}\bar{x}\\ \bar{y}\end{bmatrix} = \frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix}x_i\\ y_i\end{bmatrix} \qquad \begin{bmatrix}R_x & R_{xy}\\ R_{yx} & R_y\end{bmatrix} = \frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix}x_i - \bar{x}\\ y_i - \bar{y}\end{bmatrix}\begin{bmatrix}x_i - \bar{x}\\ y_i - \bar{y}\end{bmatrix}^T$$

Proof

Let z_i be the error $z_i = x_i - Ly_i - c$. Then

$$\bar{z} = \begin{bmatrix} -I & L \end{bmatrix} \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} + c \qquad\qquad z_i - \bar{z} = \begin{bmatrix} -I & L \end{bmatrix} \begin{bmatrix} x_i - \bar{x} \\ y_i - \bar{y} \end{bmatrix}$$

Then the sample MSE is

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} \|z_i\|^2 &= \frac{1}{N}\sum_{i=1}^{N} \|z_i - \bar{z} + \bar{z}\|^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} \|z_i - \bar{z}\|^2 + \|\bar{z}\|^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} \left\| \begin{bmatrix} -I & L \end{bmatrix} \begin{bmatrix} x_i - \bar{x} \\ y_i - \bar{y} \end{bmatrix} \right\|^2 + \left\| \begin{bmatrix} -I & L \end{bmatrix} \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} + c \right\|^2 \end{aligned}$$

Hence the optimal c is $c_{\text{opt}} = \bar{x} - L\bar{y}$

Proof continued

Now define (similarly for Y and Z)

$$X = \begin{bmatrix} x_1 & x_2 & \dots & x_N \end{bmatrix} \qquad \bar{X} = \begin{bmatrix} \bar{x} & \bar{x} & \dots & \bar{x} \end{bmatrix}$$

Then with the optimal c, the sample MSE is

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} \|z_i\|^2 &= \frac{1}{N}\sum_{i=1}^{N} \left\| \begin{bmatrix} -I & L \end{bmatrix} \begin{bmatrix} x_i - \bar{x} \\ y_i - \bar{y} \end{bmatrix} \right\|^2 \\ &= \frac{1}{N} \left\| \begin{bmatrix} -I & L \end{bmatrix} \begin{bmatrix} X - \bar{X} \\ Y - \bar{Y} \end{bmatrix} \right\|_F^2 \\ &= \frac{1}{N} \mathbf{trace} \begin{bmatrix} -I & L \end{bmatrix} \begin{bmatrix} X - \bar{X} \\ Y - \bar{Y} \end{bmatrix} \begin{bmatrix} X - \bar{X} \\ Y - \bar{Y} \end{bmatrix}^T \begin{bmatrix} -I \\ L^T \end{bmatrix} \\ &= \mathbf{trace} \begin{bmatrix} -I & L \end{bmatrix} \begin{bmatrix} R_x & R_{xy} \\ R_{yx} & R_y \end{bmatrix} \begin{bmatrix} -I \\ L^T \end{bmatrix} \end{aligned}$$

and (as for the LMMSE) completion of squares gives the result.

Least squares and SLMMSE

- Both the sample means and covariances are *consistent*.

  i.e., as we collect more data, the sample means and covariances converge to the true means and covariances.

- Hence one can show that (for nice pdfs) the SLMMSE converges to the true LMMSE

- We can compute the SLMMSE without any *model*

- If in addition we have a model, such as the joint pdf or the covariances, then we can also analyze the error.

Three methods for constructing estimators

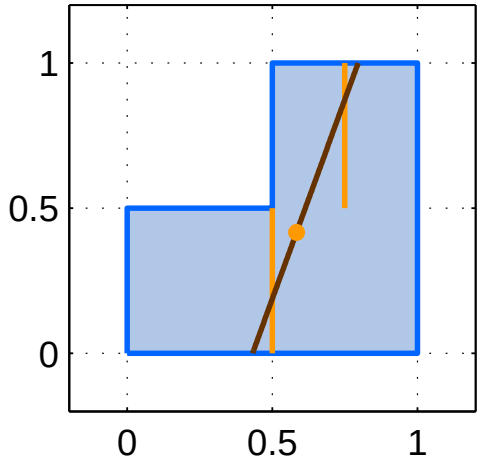
- *The MMSE*: Uses the joint pdf of x, y; optimal estimator $\phi(y) = \mathbf{E}(x \mid y = y_{\text{meas}})$
- *The LMMSE*: Uses only the mean and covariance of x, y; gives optimal estimator $\phi(y) = \mu_x + \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y)$
- The *sample LMMSE*: Uses only data; optimal estimator $\phi(y) = \bar{x} + R_{xy}R_y^{-1}(y - \bar{y})$

- The sample LMMSE converges to the LMMSE as the amount of data $N \to \infty$.
- The LMMSE is the same as the MMSE if x, y are jointly Gaussian.

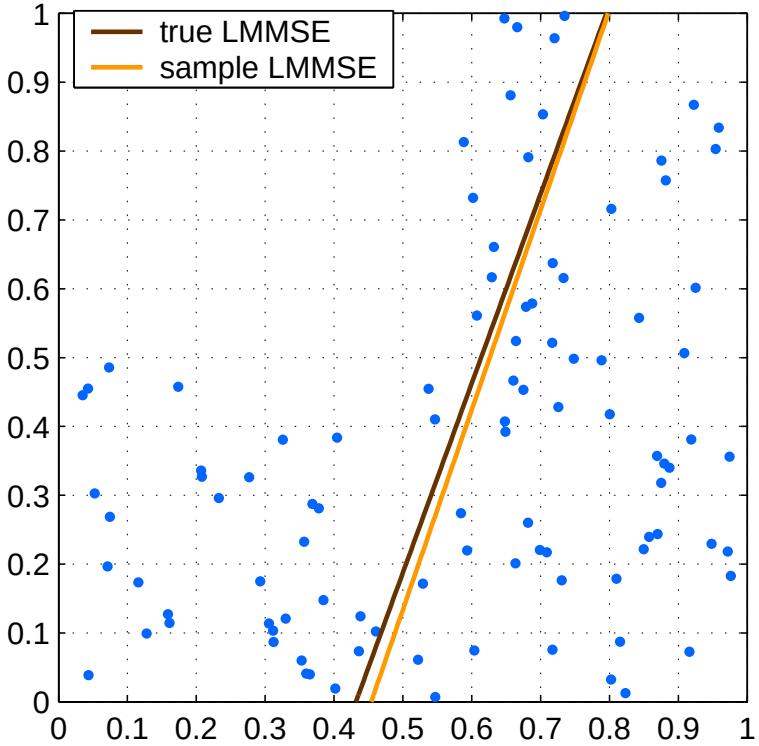
Example: SLMMSE

(x, y) are uniformly distributed on the L-shaped region A, i.e., the pdf is

$$f(x,y) = \begin{cases} \frac{4}{3} & \text{if } (x,y) \in A \\ 0 & \text{otherwise} \end{cases}$$



We construct a least-squares fit based on 100 data points



Example: sample polynomial estimators

For scalar x, y, find the estimator of the form

$$\phi(y) = a_0 + a_1 y + a_2 y^2 + \cdots + a_d y^d$$

that minimizes the sample MSE.

- Again, this is just a least-squares problem
- The solution uses the *sample moments*

$$\mathbf{E}\, y^k \approx \frac{1}{N} \sum_{i=1}^{N} y_i^k$$

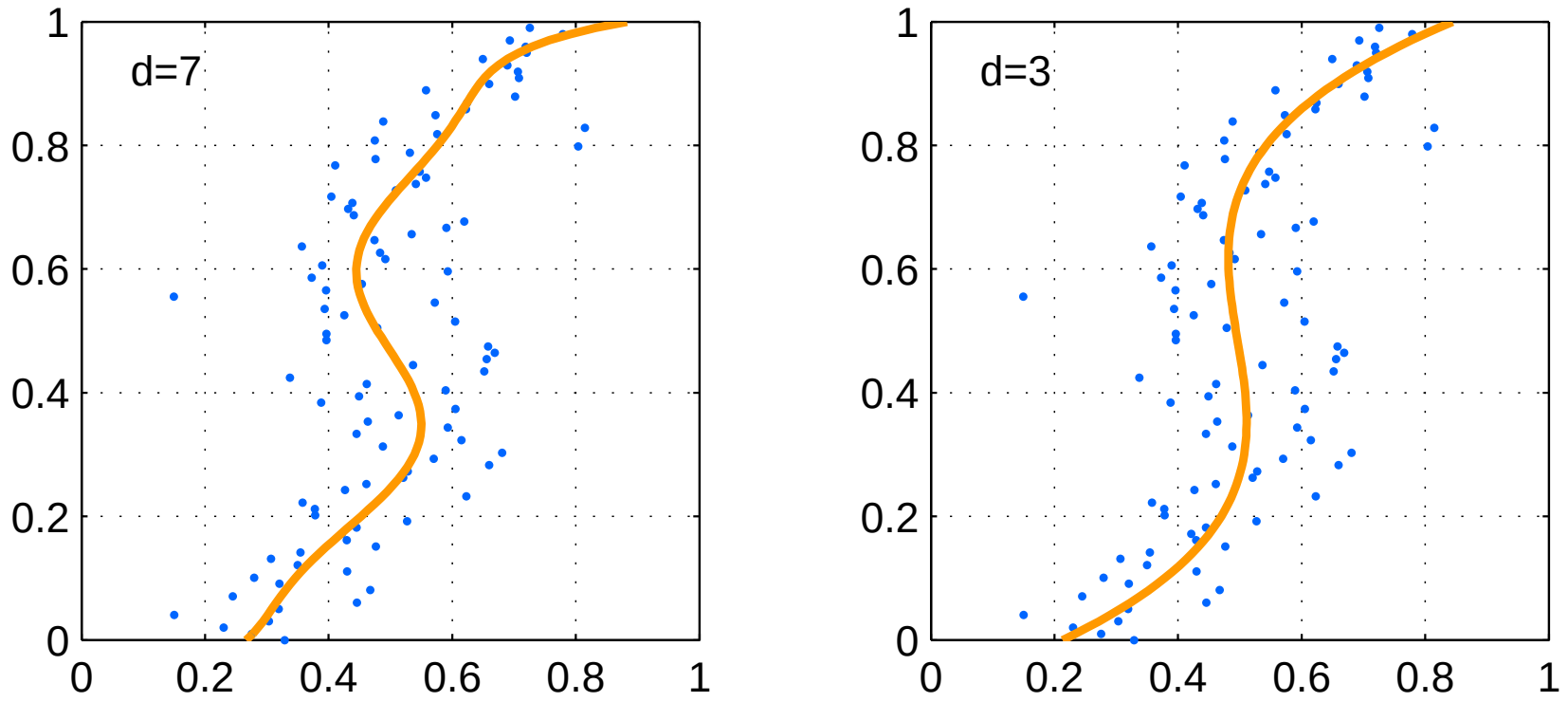
Example: learning polynomial estimators

Suppose the data comes from the model

$$x = f(y) + w$$

where $w \sim \mathcal{N}(0, 0.01)$ and f is the cubic $f = \frac{173}{400} + \frac{18\,y}{25} - \frac{51\,y^2}{20} + 3\,y^3$

The data, and fits of degree 3 and 7 are below.

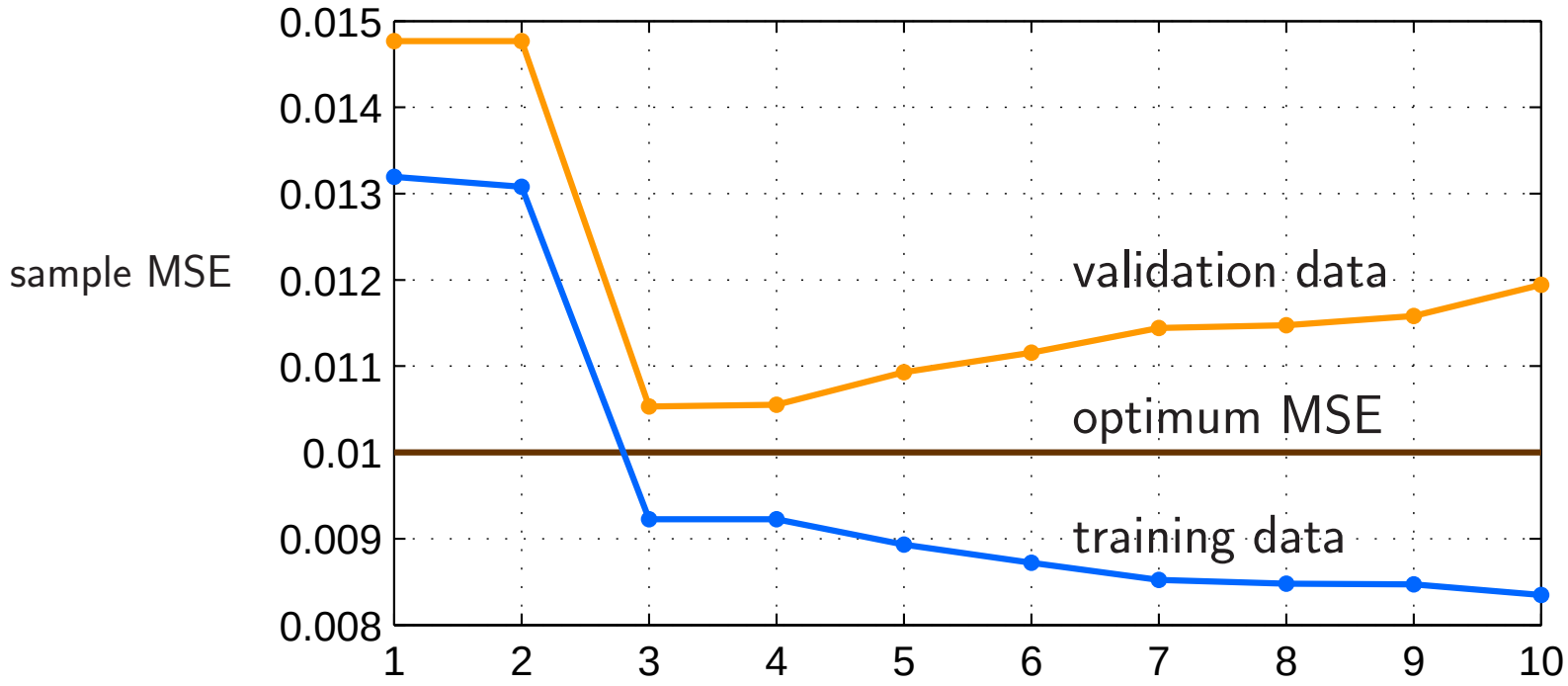


Validation

As the degree d increases:

- the optimal degree d fit achieves a smaller MSE on the *training data*
- but the *predictive ability* of the model on other data becomes worse

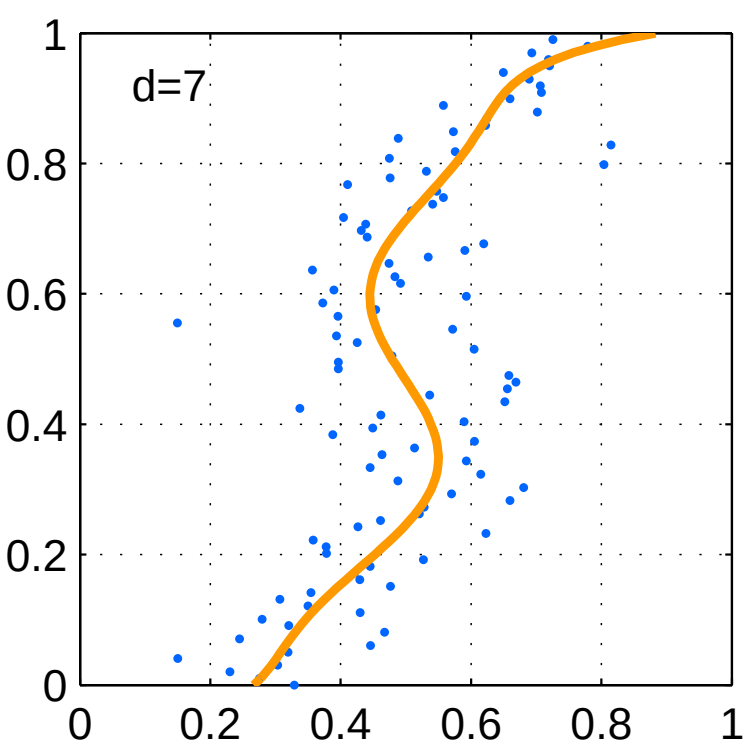
Cross-validation: measure performance on a separate set of data, called *validation data*.

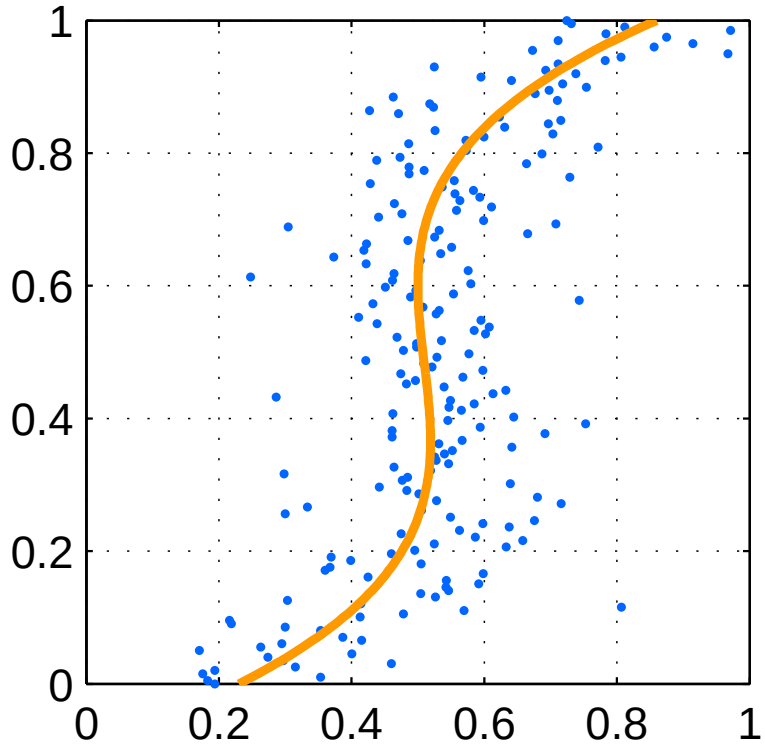


The above plot suggests that $d = 3$ is a good choice.

Example: validation

When d is large, the model is *over-fitted* to the training data.





The plot on the right shows $x = f(y)$.