Lesson 6: validation

the BDC group

September 24, 2019

Outline

- scalable learning
- connecting the pieces together
- peer instructions
- a hands-on example

Scalable learning

Connecting the pieces together

The pieces of the puzzle, up to now:

- estimating the model parameters
- validating the model structure
- estimating the uncertainties

- validating the data

Connecting the pieces together - piece 1: estimating the model parameters

Most important strategies, up to now:

- Ordinary Least Squares
- Principal Component Regression
- Partial Least Squares

Ordinary Least Squares

Assumptions:
$$\begin{cases} \text{data generation model:} \quad y_t = f\left(u_t;\theta\right) + v_t \\ \text{dataset:} \quad \mathcal{D} = \left\{\left(u_t, y_t\right)\right\}_{t=1,\dots,N} \\ \text{hypothesis space:} \quad \theta \in \Theta \end{cases}$$

Formulation:
$$\widehat{\theta}_{\mathrm{LS}} = \arg\min_{\theta\in\Theta} \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} f\left(u_1;\theta\right) \\ \vdots \\ f\left(u_N;\theta\right) \end{bmatrix} \right\|^2 = \arg\min_{\theta\in\Theta} \sum_{t=1}^{N} \Big(y_t - f\left(u_t;\theta\right) \Big)^2$$

Principal Components Regression

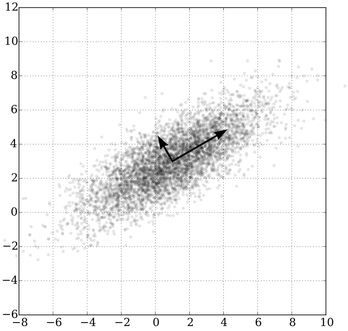
step 1: $X = U\Sigma V^T \quad \rightarrow \quad$ select U^{PC} (i.e., the first n components)

step 2: $\widehat{\theta}_{\mathrm{PCR}} = \arg\min_{\theta\in\Theta} \sum_{t=1}^{N} \left(y_t - U^{\mathsf{PC}}\theta\right)^2$

Principal Components Regression

step 1: $X = U\Sigma V^T \quad \rightarrow \quad$ select U^{PC} (i.e., the first n components)

step 2: $\widehat{\theta}_{\mathrm{PCR}} = \arg\min_{\theta\in\Theta} \sum_{t=1}^{N} \left(y_t - U^{\mathsf{PC}}\theta \right)^2$

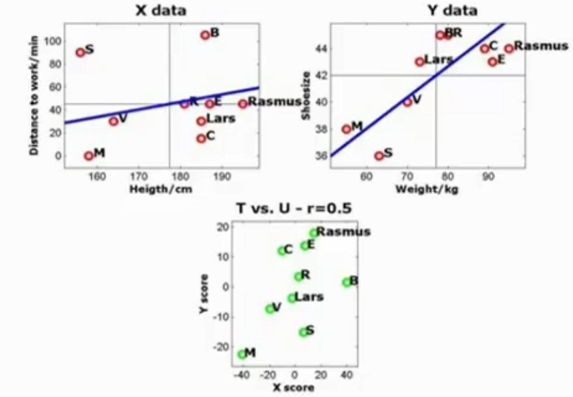


Partial Least Squares

$$\begin{cases} X = TP^T + E \\ Y = UQ^T + F \end{cases} \tag{1}$$

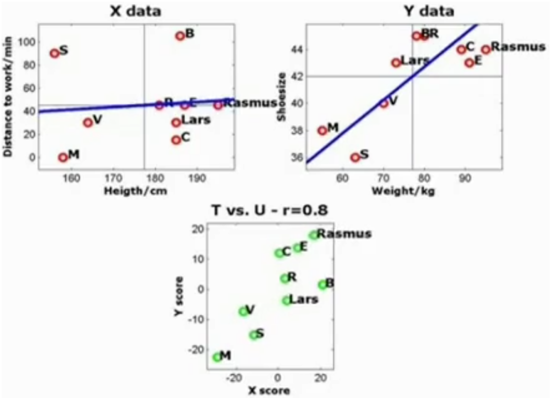
Partial Least Squares

$$\begin{cases} X = TP^T + E \\ Y = UQ^T + F \end{cases} \quad (1)$$



Partial Least Squares

$$\begin{cases} X = TP^T + E \\ Y = UQ^T + F \end{cases} \tag{1}$$



?

Connecting the pieces together

The pieces of the puzzle, up to now:

- estimating the model parameters
- validating the model structure
- estimating the uncertainties

- validating the data

Connecting the pieces together - piece 2: validating the model structure

First question, even before seeing the data: are the *internal* and *external* validity satisfied?

Connecting the pieces together - piece 2: validating the model structure

First question, even before seeing the data: are the *internal* and *external* validity satisfied? For us, in a rough way:

- drawing conclusions (i.e., make models) ignoring some unknown input $\implies$ losing internal validity

Connecting the pieces together - piece 2: validating the model structure

First question, even before seeing the data: are the *internal* and *external* validity satisfied? For us, in a rough way:

- drawing conclusions (i.e., make models) ignoring some unknown input $\Longrightarrow$ losing internal validity
- drawing conclusions (i.e., make models) ignoring some part of the input space (AND being in a situation where this leads to poor generalization capabilities) $\Longrightarrow$ losing external validity

Connecting the pieces together - piece 2: validating the model structure

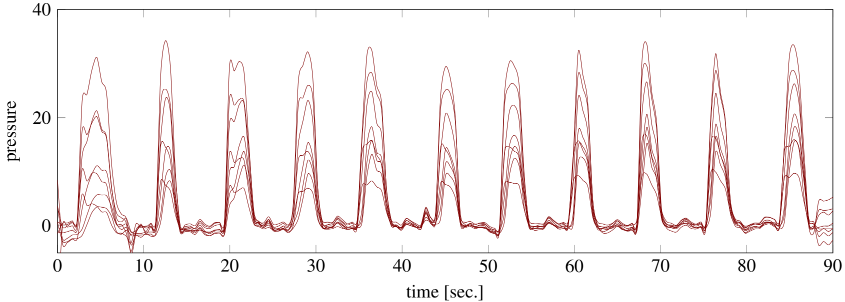
First question, even before seeing the data: are the *internal* and *external* validity satisfied? For us, in a rough way:

- drawing conclusions (i.e., make models) ignoring some unknown input $\Longrightarrow$ losing internal validity
- drawing conclusions (i.e., make models) ignoring some part of the input space (AND being in a situation where this leads to poor generalization capabilities) $\Longrightarrow$ losing external validity

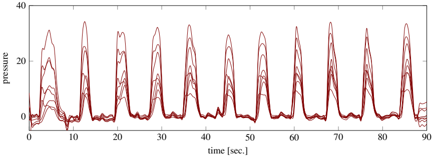
Question 1

why "... space AND being ... "?

Example - introduction

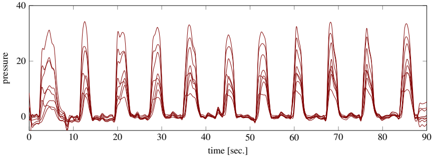


Example - population validity



Definition: population validity = how well the sample used can be extrapolated to a population as a whole *(type of external validity)*

Example - ecological validity



Definition: ecological validity = how well the findings can be extrapolated to real life settings *(type of external validity)*

?

Bias vs. variance

our approach: $\theta =$ unknown and deterministic; $\widehat{\theta} =$ estimator of θ

Bias vs. variance

our approach: θ = unknown and deterministic; $\widehat{\theta}$ = estimator of θ

Question 2

is $\widehat{\theta}$ always a random variable?

- *yes*
- *no*

Bias vs. variance

our approach: θ = unknown and deterministic; $\widehat{\theta}$ = estimator of θ

Question 2

is $\widehat{\theta}$ always a random variable?

- *yes*
- *no*

usual case: $\widehat{\theta}$ = random variable

Mean Squared Error

how do we weight the difference between θ *and* $\widehat{\theta}$*?*

Mean Squared Error

how do we weight the difference between θ and $\widehat{\theta}$?

squared error committed by a specific realization of $\widehat{\theta}$: $\left\|\theta - \widehat{\theta}\right\|^2$

Mean Squared Error

how do we weight the difference between θ and $\widehat{\theta}$?

squared error committed by a specific realization of $\widehat{\theta}$: $\left\|\theta - \widehat{\theta}\right\|^2$

mean squared error committed by $\widehat{\theta}$: $\mathbb{E}\left[\left\|\theta - \widehat{\theta}\right\|^2\right]$

Remark 1: the MSE is a function of θ!

$$\mathrm{MSE}(\theta) = \mathbb{E}\left[\left\|\theta - \widehat{\theta}\right\|^2\right]$$

Remark 1: the MSE is a function of θ!

$$\mathrm{MSE}(\theta) = \mathbb{E}\left[\left\|\theta - \widehat{\theta}\right\|^2\right]$$

Example: $\quad y_t \sim \mathcal{N}(\mu, 1) \qquad \widehat{\mu} = 3$

Remark 1: the MSE is a function of θ!

$$\mathrm{MSE}(\theta) = \mathbb{E}\left[\left\|\theta - \widehat{\theta}\right\|^2\right]$$

Example: $y_t \sim \mathcal{N}(\mu, 1) \qquad \widehat{\mu} = 3$

Question 3

$\mathrm{MSE}(3) = ?$

- *0*
- *1*
- *10*

Remark 1: the MSE is a function of θ!

$$\mathrm{MSE}(\theta) = \mathbb{E}\left[\left\|\theta - \widehat{\theta}\right\|^2\right]$$

Example: $\quad y_t \sim \mathcal{N}(\mu, 1) \qquad \widehat{\mu} = 3$

fundamental message: given $\widehat{\theta}$, that estimator may have excellent performance for certain specific θs and awful performance for other ones!

Remark 2: the MSE is defined over the measure of $y_1, \ldots, y_N$!

$$\mathbb{E}\left[\left\|\theta - \widehat{\theta}\right\|^2\right] = \int_{\mathcal{Y}^N} \left\|\theta - \widehat{\theta}\left(y_1, \ldots, y_N\right)\right\|^2 dp\left(y_1, \ldots, y_N; \theta\right)$$

Remark 2: the MSE is defined over the measure of $y_1, \dots, y_N$!

$$\mathbb{E}\left[\left\|\theta - \widehat{\theta}\right\|^2\right] = \int_{\mathcal{Y}^N} \left\|\theta - \widehat{\theta}\left(y_1, \dots, y_N\right)\right\|^2 dp\left(y_1, \dots, y_N; \theta\right)$$

important implication: the MSE cannot be computed!

Remark 2: the MSE is defined over the measure of $y_1, \ldots, y_N$!

$$\mathbb{E}\left[\left\|\theta - \widehat{\theta}\right\|^2\right] = \int_{\mathcal{Y}^N} \left\|\theta - \widehat{\theta}\left(y_1, \ldots, y_N\right)\right\|^2 dp\left(y_1, \ldots, y_N; \theta\right)$$

important implication: the MSE cannot be computed!

Strategy: estimate some alternative quantity from the data:

$$\begin{cases} y_t = f\left(u_t; \theta\right) + v_t \\ \widehat{y}_t = \widehat{f}\left(u_t; \widehat{\theta}\right) \end{cases} \quad \mapsto \quad \frac{1}{N} \sum_{t=1}^{N} \left(y_t - \widehat{y}_t\right)^2$$

Remark 2: the MSE is defined over the measure of $y_1, \ldots, y_N$!

$$\mathbb{E}\left[\left\|\theta - \widehat{\theta}\right\|^2\right] = \int_{\mathcal{Y}^N} \left\|\theta - \widehat{\theta}\left(y_1, \ldots, y_N\right)\right\|^2 dp\left(y_1, \ldots, y_N; \theta\right)$$

important implication: the MSE cannot be computed!

Strategy: estimate some alternative quantity from the data:

$$\begin{cases} y_t = f\left(u_t; \theta\right) + v_t \\ \widehat{y}_t = \widehat{f}\left(u_t; \widehat{\theta}\right) \end{cases} \quad \mapsto \quad \frac{1}{N} \sum_{t=1}^{N} \left(y_t - \widehat{y}_t\right)^2$$

'training-vs-test' and cross-validation are examples

the bias - variance tradeoff

Decomposing the MSE in two interesting terms

$$\mathbb{E}\left[\left\|\widehat{\theta} - \theta\right\|^2\right]$$

Decomposing the MSE in two interesting terms

$$\mathbb{E}\left[\left\|\widehat{\theta}-\theta\right\|^2\right] \quad = \mathbb{E}\left[\left\|\widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right]+\mathbb{E}\left[\widehat{\theta}\right]-\theta\right\|^2\right]$$

Decomposing the MSE in two interesting terms

$$
\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\theta}-\theta\right\|^2\right] &= \mathbb{E}\left[\left\|\widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right]+\mathbb{E}\left[\widehat{\theta}\right]-\theta\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\mathcal{V}+\mathcal{B}\right\|^2\right]
\end{aligned}
\qquad
\begin{cases}
\mathcal{V} := \widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right] \\
\mathcal{B} := \mathbb{E}\left[\widehat{\theta}\right]-\theta
\end{cases}
$$

Decomposing the MSE in two interesting terms

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\theta}-\theta\right\|^2\right] &= \mathbb{E}\left[\left\|\widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right]+\mathbb{E}\left[\widehat{\theta}\right]-\theta\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\mathcal{V}+\mathcal{B}\right\|^2\right] \\
&= \mathbb{E}\left[\left(\mathcal{V}+\mathcal{B}\right)^T\left(\mathcal{V}+\mathcal{B}\right)\right]
\end{aligned}$$

$$\begin{cases} \mathcal{V} := \widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right] \\ \mathcal{B} := \mathbb{E}\left[\widehat{\theta}\right]-\theta \end{cases}$$

Decomposing the MSE in two interesting terms

$$
\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\theta} - \theta\right\|^2\right] &= \mathbb{E}\left[\left\|\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right] + \mathbb{E}\left[\widehat{\theta}\right] - \theta\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\mathcal{V} + \mathcal{B}\right\|^2\right] \\
&= \mathbb{E}\left[\left(\mathcal{V} + \mathcal{B}\right)^T\left(\mathcal{V} + \mathcal{B}\right)\right] \\
&= \mathbb{E}\left[\left\|\mathcal{V}\right\|^2 + \left\|\mathcal{B}\right\|^2 + 2\mathcal{V}^T\mathcal{B}\right]
\end{aligned}
$$

$$
\begin{cases}
\mathcal{V} := \widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right] \\
\mathcal{B} := \mathbb{E}\left[\widehat{\theta}\right] - \theta
\end{cases}
$$

Decomposing the MSE in two interesting terms

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\theta}-\theta\right\|^2\right] &= \mathbb{E}\left[\left\|\widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right]+\mathbb{E}\left[\widehat{\theta}\right]-\theta\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\mathcal{V}+\mathcal{B}\right\|^2\right] \\
&= \mathbb{E}\left[\left(\mathcal{V}+\mathcal{B}\right)^T\left(\mathcal{V}+\mathcal{B}\right)\right] \\
&= \mathbb{E}\left[\left\|\mathcal{V}\right\|^2+\left\|\mathcal{B}\right\|^2+2\mathcal{V}^T\mathcal{B}\right]
\end{aligned}$$

$$\begin{cases} \mathcal{V} := \widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right] \\ \mathcal{B} := \mathbb{E}\left[\widehat{\theta}\right]-\theta \end{cases}$$

$$\mathbb{E}\left[\mathcal{V}^T\mathcal{B}\right] = \mathbf{0}$$

Decomposing the MSE in two interesting terms

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\theta}-\theta\right\|^2\right] &= \mathbb{E}\left[\left\|\widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right]+\mathbb{E}\left[\widehat{\theta}\right]-\theta\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\mathcal{V}+\mathcal{B}\right\|^2\right] \\
&= \mathbb{E}\left[(\mathcal{V}+\mathcal{B})^T(\mathcal{V}+\mathcal{B})\right] \\
&= \mathbb{E}\left[\left\|\mathcal{V}\right\|^2+\left\|\mathcal{B}\right\|^2+2\mathcal{V}^T\mathcal{B}\right] \\
&= \mathbb{E}\left[\left\|\mathcal{V}\right\|^2\right]+\left\|\mathcal{B}\right\|^2
\end{aligned}$$

$$\begin{cases} \mathcal{V} := \widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right] \\ \mathcal{B} := \mathbb{E}\left[\widehat{\theta}\right]-\theta \end{cases}$$

$$\mathbb{E}\left[\mathcal{V}^T\mathcal{B}\right] = \mathbf{0}$$

Decomposing the MSE in two interesting terms

$$
\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\theta}-\theta\right\|^2\right] &= \mathbb{E}\left[\left\|\widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right]+\mathbb{E}\left[\widehat{\theta}\right]-\theta\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\mathcal{V}+\mathcal{B}\right\|^2\right] \\
&= \mathbb{E}\left[(\mathcal{V}+\mathcal{B})^T(\mathcal{V}+\mathcal{B})\right] \\
&= \mathbb{E}\left[\left\|\mathcal{V}\right\|^2+\left\|\mathcal{B}\right\|^2+2\mathcal{V}^T\mathcal{B}\right] \\
&= \mathbb{E}\left[\left\|\mathcal{V}\right\|^2\right]+\left\|\mathcal{B}\right\|^2 \\
&= \text{"variance"} + \text{"bias}^2\text{"}
\end{aligned}
$$

$$
\begin{cases}
\mathcal{V} := \widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right] \\
\mathcal{B} := \mathbb{E}\left[\widehat{\theta}\right]-\theta
\end{cases}
$$

$$
\mathbb{E}\left[\mathcal{V}^T\mathcal{B}\right] = \mathbf{0}
$$

Definitions

ideal model: $y_t = f(u_t) + v_t$ our model: $y_t = \widehat{f}\left(u_t, \widehat{\theta}\right)$

underfitting $=$ a $\widehat{f}\left(\cdot, \widehat{\theta}\right)$ that misses the fundamental features of f_0

overfitting $=$ a $\widehat{f}\left(\cdot, \widehat{\theta}\right)$ that follows v_t instead of $f(u_t)$.

Remarking fact: the model complexity affects the bias - variance tradeoff

$$y_t = \prod_{k=1}^{n} \theta_k u_t^k + v_t$$

Remarking fact: the model complexity affects the bias - variance tradeoff

$$y_t = \prod_{k=1}^{n} \theta_k u_t^k + v_t$$

we will see how this strongly connects to the Ockham's razor

Quiz time!

Question 3

Underfitting is associated to

- *high bias and low variance*
- *low bias and high variance*

Quiz time!

Question 4

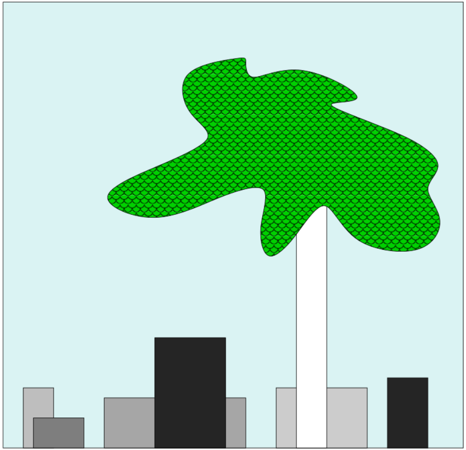
Overfitting is associated to:

- *high bias and low variance*
- *low bias and high variance*

?

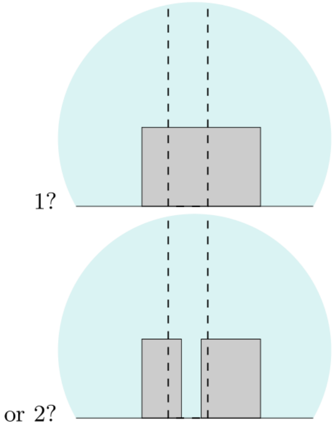
Ockham's razor

from David J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*



Ockham's razor

from David J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*



Connecting the pieces together - piece 2: validating the model structure

Strategy: estimate the MSE (or other performance indexes) from the data:

$$\begin{cases} y_t = f\left(u_t; \theta\right) + v_t \\ \widehat{y_t} = \widehat{f}\left(u_t; \widehat{\theta}\right) \end{cases} \qquad \mapsto \qquad \frac{1}{N} \sum_{t=1}^{N} \left(y_t - \widehat{y_t}\right)^2$$

Connecting the pieces together - piece 2: validating the model structure

Strategy: estimate the MSE (or other performance indexes) from the data:

$$\begin{cases} y_t = f\left(u_t;\theta\right) + v_t \\ \widehat{y_t} = \widehat{f}\left(u_t;\widehat{\theta}\right) \end{cases} \qquad \mapsto \qquad \frac{1}{N}\sum_{t=1}^{N}\left(y_t - \widehat{y_t}\right)^2$$

Most important strategies, up to now:

- dividing the dataset into training / test / validation
- cross-validation

Why is estimating the MSE on the training data a very bad idea?

Why is estimating the MSE on the training data a very bad idea?

$$y \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad \widehat{\mu} : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \widehat{\mu}\left(\cdot\right) = \text{estimator of } \mu \tag{2}$$

Why is estimating the MSE on the training data a very bad idea?

$$y \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad \widehat{\mu} : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \widehat{\mu}\left(\cdot\right) = \text{estimator of } \mu \tag{2}$$

$$R = \mathbb{E}\left[\|\mu - \widehat{\mu}\|_2^2\right]$$

Why is estimating the MSE on the training data a very bad idea?

$$y \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad \widehat{\mu} : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \widehat{\mu}\left(\cdot\right) = \text{estimator of } \mu \tag{2}$$

$$\begin{aligned} R &= \mathbb{E}\left[\|\mu - \widehat{\mu}\|_2^2\right] \\ &= \mathbb{E}\left[\|\mu - y + y - \widehat{\mu}\|_2^2\right] \end{aligned}$$

Why is estimating the MSE on the training data a very bad idea?

$$y \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad \widehat{\mu} : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \widehat{\mu}\left(\cdot\right) = \text{estimator of } \mu \tag{2}$$

$$\begin{aligned} R &= \mathbb{E}\left[\|\mu - \widehat{\mu}\|_2^2\right] \\ &= \mathbb{E}\left[\|\mu - y + y - \widehat{\mu}\|_2^2\right] \\ &= \mathbb{E}\left[\|\mu - y\|_2^2\right] + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\mathbb{E}\left[\left(\mu - y\right)^T \left(y - \widehat{\mu}\right)\right] \end{aligned}$$

Why is estimating the MSE on the training data a very bad idea?

$$y \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad \widehat{\mu} : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \widehat{\mu}\left(\cdot\right) = \text{estimator of } \mu \tag{2}$$

$$\begin{aligned}
R &= \mathbb{E}\left[\|\mu - \widehat{\mu}\|_2^2\right] \\
&= \mathbb{E}\left[\|\mu - y + y - \widehat{\mu}\|_2^2\right] \\
&= \mathbb{E}\left[\|\mu - y\|_2^2\right] + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\mathbb{E}\left[\left(\mu - y\right)^T\left(y - \widehat{\mu}\right)\right] \\
&= n\sigma^2 + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\mathbb{E}\left[\left(y - \mu\right)^T\left(\widehat{\mu} - y\right)\right]
\end{aligned}$$

Why is estimating the MSE on the training data a very bad idea?

$$y \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad \widehat{\mu} : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \widehat{\mu}\left(\cdot\right) = \text{estimator of } \mu \tag{2}$$

$$\begin{aligned}
R &= \mathbb{E}\left[\|\mu - \widehat{\mu}\|_2^2\right] \\
&= \mathbb{E}\left[\|\mu - y + y - \widehat{\mu}\|_2^2\right] \\
&= \mathbb{E}\left[\|\mu - y\|_2^2\right] + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\mathbb{E}\left[\left(\mu - y\right)^T\left(y - \widehat{\mu}\right)\right] \\
&= n\sigma^2 + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\mathbb{E}\left[\left(y - \mu\right)^T\left(\widehat{\mu} - y\right)\right] \\
&= -n\sigma^2 + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\sum_{i=1}^{n} \operatorname{cov}\left(y_i, \widehat{\mu}_i\right)
\end{aligned}$$

Why is estimating the MSE on the training data a very bad idea?

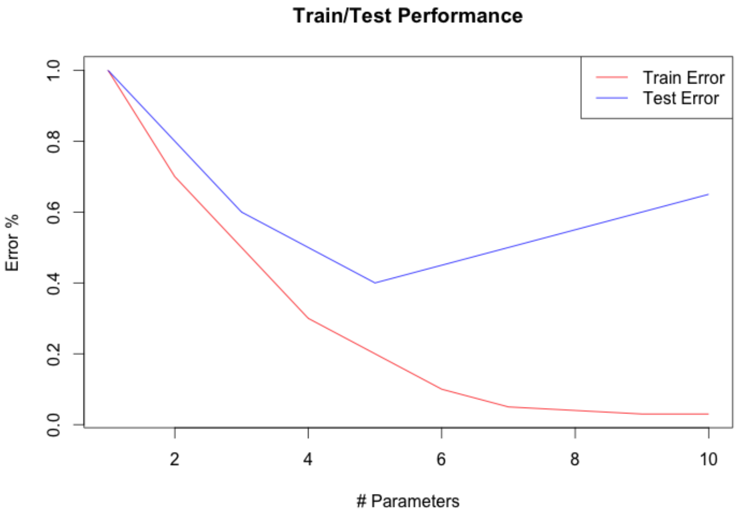
$$y \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad \widehat{\mu} : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \widehat{\mu}\left(\cdot\right) = \text{estimator of } \mu \tag{2}$$

$$\begin{aligned}
R &= \mathbb{E}\left[\|\mu - \widehat{\mu}\|_2^2\right] \\
&= \mathbb{E}\left[\|\mu - y + y - \widehat{\mu}\|_2^2\right] \\
&= \mathbb{E}\left[\|\mu - y\|_2^2\right] + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\mathbb{E}\left[(\mu - y)^T\left(y - \widehat{\mu}\right)\right] \\
&= n\sigma^2 + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\mathbb{E}\left[(y - \mu)^T\left(\widehat{\mu} - y\right)\right] \\
&= -n\sigma^2 + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\sum_{i=1}^{n} \operatorname{cov}\left(y_i, \widehat{\mu}_i\right) \\
\implies \quad & \widehat{R} = -n\sigma^2 + \|y - \widehat{\mu}\|_2^2 + 2\sum_{i=1}^{n} \operatorname{cov}\left(y_i, \widehat{\mu}_i\right)
\end{aligned} \tag{3}$$

Potential practical difficulties when doing model validation

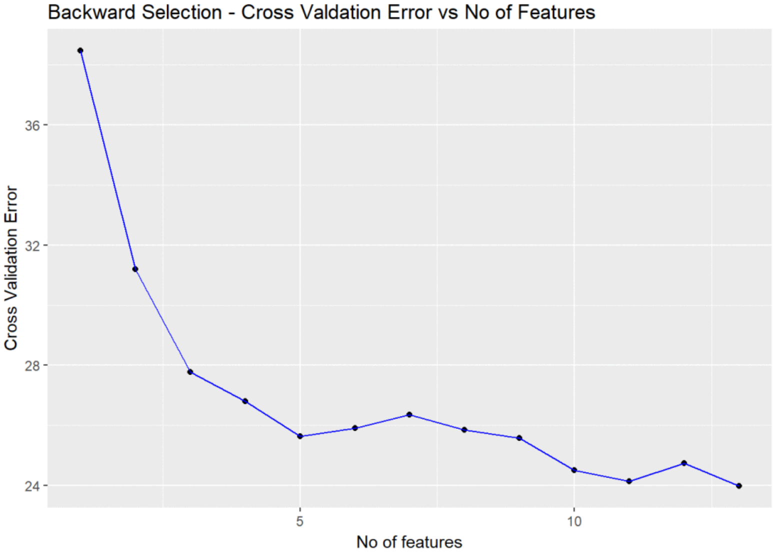
- lack of data
- lack of control of the input variables
- uncertainty about the underlying probability distributions and correlations

Practical examples: which model structure would you select?

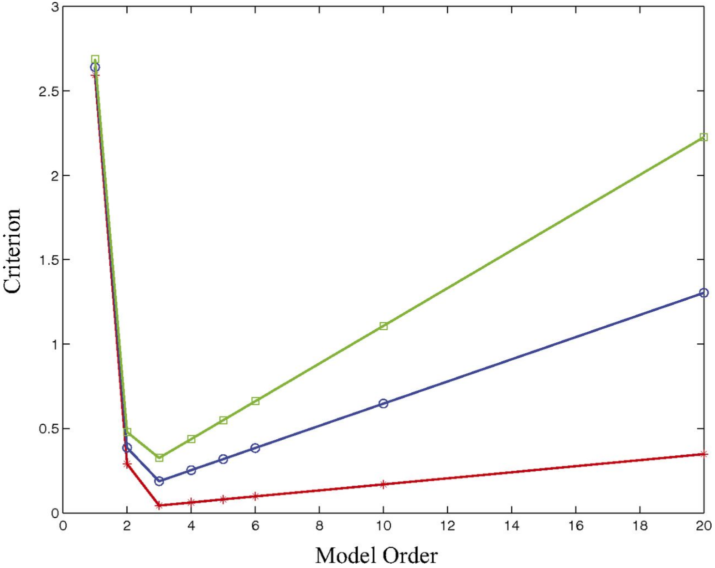


Practical examples: which model structure would you select?

Practical examples: which model structure would you select?



Practical examples: which model structure would you select?



?

Connecting the pieces together

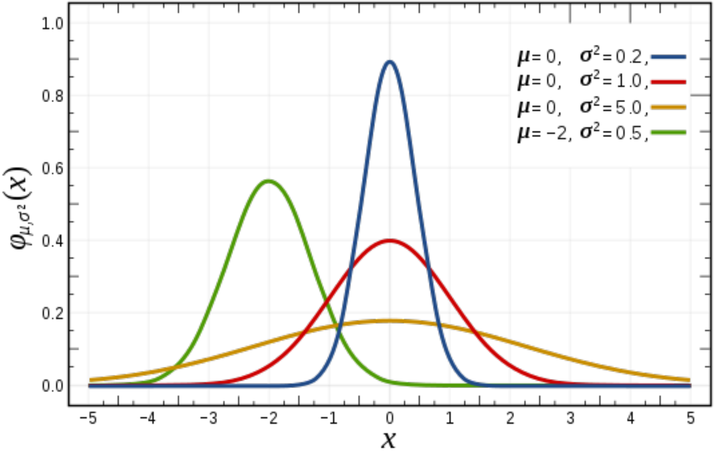
The pieces of the puzzle, up to now:

- estimating the model parameters
- validating the model structure
- estimating the uncertainties

- validating the data

Connecting the pieces together - piece 3: estimating the uncertainty on the model parameters

Typical aim: estimating mean and variance, assuming a Gaussian distribution:



Simplest strategy: jackknifing

Purpose: estimate *bias* and *variance* of the estimator;

Simplest strategy: jackknifing

Purpose: estimate *bias* and *variance* of the estimator; steps:

1. define the reduced datasets $X_{[i]} := \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$

Simplest strategy: jackknifing

Purpose: estimate *bias* and *variance* of the estimator; steps:

1. define the reduced datasets $X_{[i]} \coloneqq \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$
2. compute the corresponding estimates $\widehat{\theta}_{[i]} = \widehat{\theta}\left(X_{[i]}\right)$

Simplest strategy: jackknifing

Purpose: estimate *bias* and *variance* of the estimator; steps:

1. define the reduced datasets $X_{[i]} := \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$
2. compute the corresponding estimates $\widehat{\theta}_{[i]} = \widehat{\theta}\left(X_{[i]}\right)$
3. compute the average "reduced" estimate $\widehat{\theta}_{\text{ave}} := \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_{[i]}$

Simplest strategy: jackknifing

Purpose: estimate *bias* and *variance* of the estimator; steps:

1. define the reduced datasets $X_{[i]} := \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$
2. compute the corresponding estimates $\widehat{\theta}_{[i]} = \widehat{\theta}\left(X_{[i]}\right)$
3. compute the average "reduced" estimate $\widehat{\theta}_{\text{ave}} := \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \widehat{\theta}_{[i]}$
4. compute the estimated bias and variance

$$\widehat{\text{bias}}\left(\widehat{\theta}\right)_{\text{jk}} := \frac{n-1}{n} \sum_{i=1}^{n} \left(\widehat{\theta}_{[i]} - \widehat{\theta}\right) \qquad \widehat{\text{var}}\left(\widehat{\theta}\right)_{\text{jk}} := \frac{n-1}{n} \sum_{i=1}^{n} \left(\widehat{\theta}_{[i]} - \widehat{\theta}_{\text{ave}}\right)^2$$

Simplest strategy: jackknifing

Purpose: estimate *bias* and *variance* of the estimator; steps:

1. define the reduced datasets $X_{[i]} := \{X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n\}$
2. compute the corresponding estimates $\widehat{\theta}_{[i]} = \widehat{\theta}\left(X_{[i]}\right)$
3. compute the average "reduced" estimate $\widehat{\theta}_{\text{ave}} := \frac{1}{n}\sum_{i=1}^{n} \widehat{\theta}_{[i]}$
4. compute the estimated bias and variance

$$\widehat{\text{bias}}\left(\widehat{\theta}\right)_{\text{jk}} := \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_{[i]} - \widehat{\theta}\right) \qquad \widehat{\text{var}}\left(\widehat{\theta}\right)_{\text{jk}} := \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_{[i]} - \widehat{\theta}_{\text{ave}}\right)^2$$

$\rightarrow$ can be generalized to k deletions

More sophisticated strategy: bootstrapping

More sophisticated strategy: bootstrapping

1. generate B new datasets with the same dimension as the original one sampling with replacement
2. compute the B estimates
3. compute some statistics on the B estimates

More sophisticated strategy: bootstrapping

1. generate B new datasets with the same dimension as the original one sampling with replacement
2. compute the B estimates
3. compute some statistics on the B estimates
4. if one wants to have a direct estimate of the predictive performance, average the performance of the various B estimates on the corresponding “out-of-bag” samples

More sophisticated strategy: bootstrapping

1. generate B new datasets with the same dimension as the original one sampling with replacement
2. compute the B estimates
3. compute some statistics on the B estimates
4. if one wants to have a direct estimate of the predictive performance, average the performance of the various B estimates on the corresponding "out-of-bag" samples

? how big shall B be ?

Flashback: bias and variance tradeoff

Alternative strategies:

- jackknifing
- bootstrapping
- CV *(i.e., go directly towards estimating the MSE / whatever performance index is desired)*

Comparisons

- the bootstrap handles skewed distributions better
- the jackknife is suitable for smaller original data samples

how to use the expressions for the estimated variance

Using $\widehat{P}$ for finding confidence intervals on $\widehat{\theta}$

$$\left(\widehat{\theta} - \theta\right) \sim \mathcal{N}\left(0, \widehat{P}\right)$$

Using $\widehat{P}$ for finding confidence intervals on $\widehat{\theta}$

$$\left(\widehat{\theta} - \theta\right) \sim \mathcal{N}\left(0, \widehat{P}\right) \qquad \Longrightarrow \qquad \left(\widehat{\theta}^{(k)} - \theta^{(k)}\right) \sim \mathcal{N}\left(0, \widehat{P}_{(kk)}\right)$$

Using $\widehat{P}$ for finding confidence intervals on $\widehat{\theta}$

$$\left(\widehat{\theta} - \theta\right) \sim \mathcal{N}\left(0, \widehat{P}\right) \qquad \Longrightarrow \qquad \left(\widehat{\theta}^{(k)} - \theta^{(k)}\right) \sim \mathcal{N}\left(0, \widehat{P}_{(kk)}\right)$$

$$\Longrightarrow \qquad \mathbb{P}\left[\left|\widehat{\theta}^{(k)} - \theta^{(k)}\right| \geq \alpha\right] \approx \sqrt{\frac{1}{2\pi \widehat{P}_{(kk)}}} \int_{|x| \geq \alpha} \exp\left(-\frac{1}{2} \frac{x^2}{\widehat{P}_{(kk)}}\right) dx$$

Using $\widehat{P}$ for finding confidence intervals on $\widehat{\theta}$

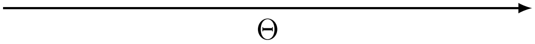
$$\left(\widehat{\theta} - \theta\right) \sim \mathcal{N}\left(0, \widehat{P}\right) \qquad \Longrightarrow \qquad \left(\widehat{\theta}^{(k)} - \theta^{(k)}\right) \sim \mathcal{N}\left(0, \widehat{P}_{(kk)}\right)$$

$$\Longrightarrow \qquad \mathbb{P}\left[\left|\widehat{\theta}^{(k)} - \theta^{(k)}\right| \geq \alpha\right] \approx \sqrt{\frac{1}{2\pi \widehat{P}_{(kk)}}} \int_{|x| \geq \alpha} \exp\left(-\frac{1}{2} \frac{x^2}{\widehat{P}_{(kk)}}\right) dx$$

$$\Longrightarrow \qquad \left(\widehat{\theta} - \theta\right)^T \widehat{P}^{-1} \left(\widehat{\theta} - \theta\right) \sim \chi^2(K)$$

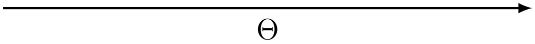
Confidence intervals: these mysterious objects. . .

$$\begin{cases} \widehat{\theta} \in \Theta \\ C \subseteq \Theta \end{cases}$$



Confidence intervals: these mysterious objects...

$$\begin{cases} \widehat{\theta} \in \Theta \\ C \subseteq \Theta \end{cases}$$



Definition 1 (Confidence interval)

$C : \mathcal{D} \mapsto 2^{\Theta}$ *is a C.I. with confidence level* α *if*

$$\inf_{\theta \in \Theta} \mathbb{P}\left[\mathcal{D} \,:\, C(\mathcal{D}) \ni \theta\right] \geq \alpha$$

Confidence intervals: these mysterious objects. . .

$$\begin{cases} \widehat{\theta} \in \Theta \\ C \subseteq \Theta \end{cases}$$

Θ

Definition 1 (Confidence interval)

$C : \mathcal{D} \mapsto 2^{\Theta}$ *is a C.I. with confidence level* α *if*

$$\inf_{\theta \in \Theta} \mathbb{P}\left[\mathcal{D} \, : \, C(\mathcal{D}) \ni \theta\right] \geq \alpha$$

a C.I. is not a statement about θ

?

Connecting the pieces together

The pieces of the puzzle, up to now:

- estimating the model parameters
- validating the model structure
- estimating the uncertainties

- validating the data

Connecting the pieces together - piece 4: validating the data

Can happen at every step

- validating the data
- estimating the model parameters
- validating the data
- validating the model structure
- validating the data
- estimating the uncertainties
- validating the data

Connecting the pieces together - piece 4: validating the data

Most important strategies, up to now:

- Hotelling's T^2
- F- and Q-residuals *(to be seen now)*

Hotelling's T^2

Fundamental question: are these

$$x_1, \ldots, x_{n_x} \qquad\qquad y_1, \ldots, y_{n_y} \tag{4}$$

identically distributed?

Hotelling's T^2

Fundamental question: are these

$$x_1, \ldots, x_{n_x} \qquad\qquad y_1, \ldots, y_{n_y} \tag{4}$$

identically distributed? Algorithm:

$$\overline{x} := \frac{1}{n_x} \sum_i x_i \qquad\qquad \overline{y} := \frac{1}{n_y} \sum_i y_i \tag{5}$$

$$\Sigma_x := \frac{1}{n_x - 1} \sum_i (x_i - \overline{x})(x_i - \overline{x})^T \qquad\qquad \Sigma_y := \frac{1}{n_y - 1} \sum_i (y_i - \overline{y})(y_i - \overline{y})^T \tag{6}$$

$$\Sigma := \frac{(n_x - 1)\,\Sigma_x + (n_y - 1)\,\Sigma_y}{n_x + n_y - 2} \tag{7}$$

Hotelling's T^2

Fundamental question: are these

$$x_1, \ldots, x_{n_x} \qquad\qquad y_1, \ldots, y_{n_y} \tag{4}$$

identically distributed? Algorithm:

$$\overline{x} := \frac{1}{n_x} \sum_i x_i \qquad\qquad \overline{y} := \frac{1}{n_y} \sum_i y_i \tag{5}$$

$$\Sigma_x := \frac{1}{n_x - 1} \sum_i \left(x_i - \overline{x}\right)\left(x_i - \overline{x}\right)^T \qquad\qquad \Sigma_y := \frac{1}{n_y - 1} \sum_i \left(y_i - \overline{y}\right)\left(y_i - \overline{y}\right)^T \tag{6}$$

$$\Sigma := \frac{\left(n_x - 1\right)\Sigma_x + \left(n_y - 1\right)\Sigma_y}{n_x + n_y - 2} \tag{7}$$

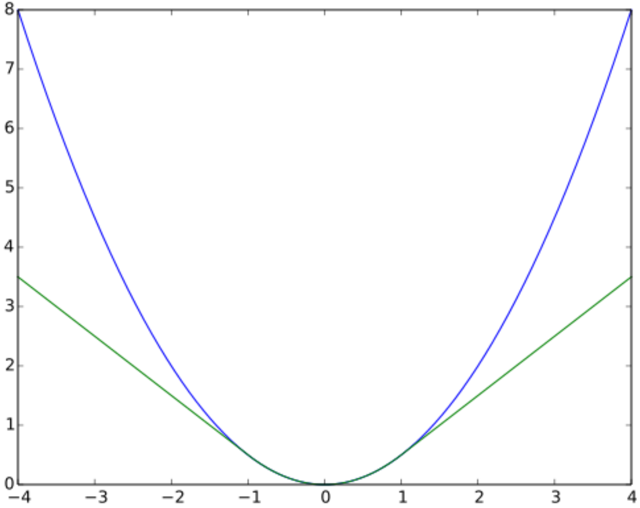
$$t^2 := \frac{n_x n_y}{n_x + n_y} \left(\overline{x} - \overline{y}\right) \Sigma \left(\overline{x} - \overline{y}\right)^T \tag{8}$$

Causes of outliers

- measurement error
- heavy tail distributions
- mixture models

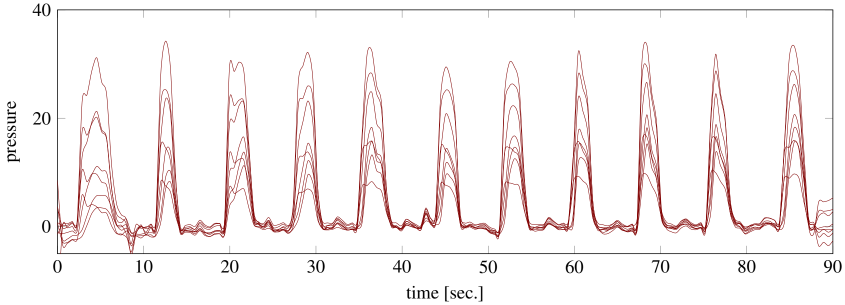
Norms to treat outliers

Norms to treat outliers



?

Hands-on example - vaginal pressure data



Vaginal pressure data: state variables

- $m_{\mathrm{a}}(t)$:= number of motor units that are in an *active state* at time t, and that are activated by a voluntary drive
- $m_{\mathrm{f}}(t)$:= number of motor units that are in a *fatigued state* at time t
- $m_{\mathrm{r}}(t)$:= number of motor units that are in a *resting state* at time t
- $u(t)$:= muscular activation signal, sometimes referred to as the “brain stimulus” or “brain force”
- M := total number of motor units present in the muscles, assumed to be constant over time, i.e., $m_{\mathrm{a}}(t) + m_{\mathrm{f}}(t) + m_{\mathrm{r}}(t) = M$ for all t

Vaginal pressure data: state dynamics

$$\dot{m}_{\mathrm{a}}(t) = -\theta_{\mathrm{a}\mapsto\mathrm{f}} m_{\mathrm{a}}(t) + \theta_{\mathrm{f}\mapsto\mathrm{a}} m_{\mathrm{f}}(t) + u(t)\theta_{\mathrm{r}\mapsto\mathrm{a}} m_{\mathrm{r}}(t) - \big(1 - u(t)\big)\theta_{\mathrm{a}\mapsto\mathrm{r}} m_{\mathrm{a}}(t) \qquad (9)$$

Vaginal pressure data: state dynamics

$$\dot{m}_{\mathrm{a}}(t) = -\theta_{\mathrm{a}\mapsto\mathrm{f}} m_{\mathrm{a}}(t) + \theta_{\mathrm{f}\mapsto\mathrm{a}} m_{\mathrm{f}}(t) + u(t)\theta_{\mathrm{r}\mapsto\mathrm{a}} m_{\mathrm{r}}(t) - \big(1 - u(t)\big)\theta_{\mathrm{a}\mapsto\mathrm{r}} m_{\mathrm{a}}(t) \tag{9}$$

$$\begin{cases} m_{\mathrm{f}}(k+1) = & \phi_{\mathrm{f}\mapsto\mathrm{a}} m_{\mathrm{f}}(k) + (1 - \phi_{\mathrm{a}\mapsto\mathrm{f}})\, m_{\mathrm{a}}(k) \\ m_{\mathrm{a}}(k+1) = & \phi_{\mathrm{a}\mapsto\mathrm{f}} m_{\mathrm{a}}(k) + (1 - \phi_{\mathrm{f}\mapsto\mathrm{a}})\, m_{\mathrm{f}}(k) \\ & + u(k)\phi_{\mathrm{r}\mapsto\mathrm{a}} m_{\mathrm{r}}(k) - \big(1 - u(k)\big)\phi_{\mathrm{a}\mapsto\mathrm{r}} m_{\mathrm{a}}(k) \\ m_{\mathrm{r}}(k) = & M - m_{\mathrm{a}}(k) - m_{\mathrm{f}}(k) \end{cases} \tag{10}$$

Vaginal pressure data: state dynamics

$$\dot{m}_{\mathrm{a}}(t) = -\theta_{\mathrm{a}\mapsto\mathrm{f}} m_{\mathrm{a}}(t) + \theta_{\mathrm{f}\mapsto\mathrm{a}} m_{\mathrm{f}}(t) + u(t)\theta_{\mathrm{r}\mapsto\mathrm{a}} m_{\mathrm{r}}(t) - \big(1 - u(t)\big)\theta_{\mathrm{a}\mapsto\mathrm{r}} m_{\mathrm{a}}(t) \tag{9}$$

$$\begin{cases} m_{\mathrm{f}}(k+1) = & \phi_{\mathrm{f}\mapsto\mathrm{a}} m_{\mathrm{f}}(k) + (1 - \phi_{\mathrm{a}\mapsto\mathrm{f}})\, m_{\mathrm{a}}(k) \\ m_{\mathrm{a}}(k+1) = & \phi_{\mathrm{a}\mapsto\mathrm{f}} m_{\mathrm{a}}(k) + (1 - \phi_{\mathrm{f}\mapsto\mathrm{a}})\, m_{\mathrm{f}}(k) \\ & + u(k)\phi_{\mathrm{r}\mapsto\mathrm{a}} m_{\mathrm{r}}(k) - \big(1 - u(k)\big)\phi_{\mathrm{a}\mapsto\mathrm{r}} m_{\mathrm{a}}(k) \\ m_{\mathrm{r}}(k) = & M - m_{\mathrm{a}}(k) - m_{\mathrm{f}}(k) \end{cases} \tag{10}$$

$$\begin{cases} \phi_{\mathrm{f}\mapsto\mathrm{a}} := 1 - \theta_{\mathrm{f}\mapsto\mathrm{a}} T \\ \phi_{\mathrm{a}\mapsto\mathrm{f}} := 1 - \theta_{\mathrm{a}\mapsto\mathrm{f}} T \\ \phi_{\mathrm{a}\mapsto\mathrm{r}} := \theta_{\mathrm{a}\mapsto\mathrm{r}} T \\ \phi_{\mathrm{r}\mapsto\mathrm{a}} := \theta_{\mathrm{r}\mapsto\mathrm{a}} T \end{cases} \tag{11}$$

Vaginal pressure data: estimation

$$
\begin{aligned}
m_{\mathrm{a}}(k+1) = \; & \Big(\phi_{\mathrm{a}\mapsto\mathrm{f}} - \phi_{\mathrm{a}\mapsto\mathrm{r}} - (\phi_{\mathrm{r}\mapsto\mathrm{a}} - \phi_{\mathrm{a}\mapsto\mathrm{r}})\, u(k)\Big) m_{\mathrm{a}}(k) \\
& + \Big(1 - \phi_{\mathrm{f}\mapsto\mathrm{a}} - \phi_{\mathrm{r}\mapsto\mathrm{a}} u(k)\Big) (1 - \phi_{\mathrm{a}\mapsto\mathrm{f}}) \left(\sum_{\tau=0}^{k-1} \phi_{\mathrm{f}\mapsto\mathrm{a}}^{k-1-\tau} m_{\mathrm{a}}(\tau) \right) + \phi_{\mathrm{r}\mapsto\mathrm{a}} M u(k)
\end{aligned} \tag{12}
$$

Vaginal pressure data: estimation

$$
\begin{aligned}
m_{\mathrm{a}}(k+1) = \; & \Big(\phi_{\mathrm{a}\mapsto\mathrm{f}} - \phi_{\mathrm{a}\mapsto\mathrm{r}} - \left(\phi_{\mathrm{r}\mapsto\mathrm{a}} - \phi_{\mathrm{a}\mapsto\mathrm{r}}\right) u(k)\Big) m_{\mathrm{a}}(k) \\
& + \Big(1 - \phi_{\mathrm{f}\mapsto\mathrm{a}} - \phi_{\mathrm{r}\mapsto\mathrm{a}} u(k)\Big) \left(1 - \phi_{\mathrm{a}\mapsto\mathrm{f}}\right) \left(\sum_{\tau=0}^{k-1} \phi_{\mathrm{f}\mapsto\mathrm{a}}^{k-1-\tau} m_{\mathrm{a}}(\tau) \right) + \phi_{\mathrm{r}\mapsto\mathrm{a}} M u(k)
\end{aligned} \tag{12}
$$

or, in a more compact way,

$$
m_a\left(k+1\right) = f\left(m_a\left(k\right), u\left(k\right)\;;\;\theta\right) \tag{13}
$$

Vaginal pressure data: estimation

$$
\begin{aligned}
m_{\mathrm{a}}(k+1) = \quad & \Big(\phi_{\mathrm{a}\mapsto\mathrm{f}} - \phi_{\mathrm{a}\mapsto\mathrm{r}} - (\phi_{\mathrm{r}\mapsto\mathrm{a}} - \phi_{\mathrm{a}\mapsto\mathrm{r}})\, u(k)\Big) m_{\mathrm{a}}(k) \\
& + \Big(1 - \phi_{\mathrm{f}\mapsto\mathrm{a}} - \phi_{\mathrm{r}\mapsto\mathrm{a}} u(k)\Big) (1 - \phi_{\mathrm{a}\mapsto\mathrm{f}}) \left(\sum_{\tau=0}^{k-1} \phi_{\mathrm{f}\mapsto\mathrm{a}}^{k-1-\tau} m_{\mathrm{a}}(\tau) \right) + \phi_{\mathrm{r}\mapsto\mathrm{a}} M u(k)
\end{aligned}
\tag{12}
$$

or, in a more compact way,

$$
m_a\left(k+1\right) = f\left(m_a\left(k\right), u\left(k\right)\ ;\ \theta\right) \tag{13}
$$

$\Longrightarrow$ *naturally leads to a nonlinear LS formulation*

What do we want to do?

- estimate the parameters of person A and the associated uncertainty on this estimate *(i.e., her physiological status)*
- check if the model structure is meaningful
- detect outliers in the measurements stream *(i.e., detect if the sensor is breaking)*

Vaginal pressure data: first step

how to compute $m_a(k)$ and $u(k)$ from the data?