Fitting a Model:
Maximum Likelihood Estimation (MLE)

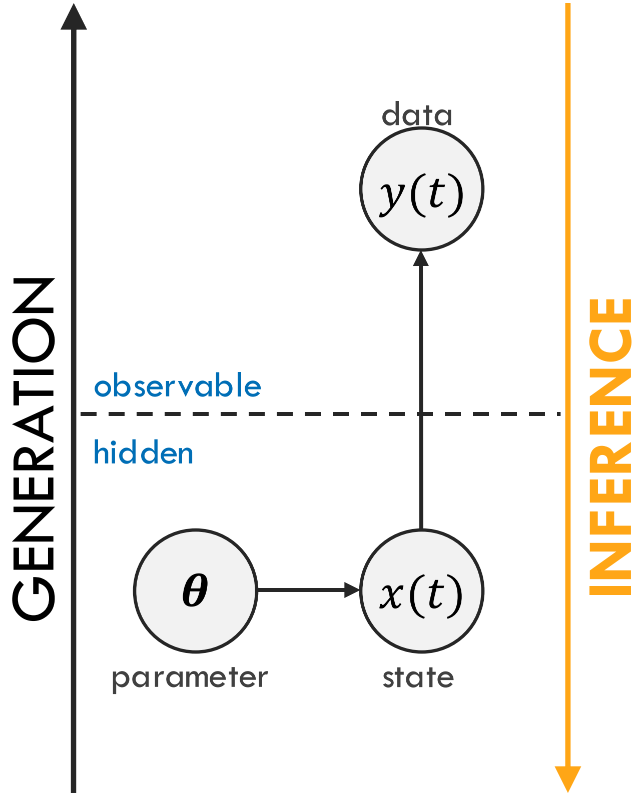
Florian M. Schönleitner

Based on slides by Stefan Frässle and Katharina Wellstein

Recap: generative modeling

Last talk:

- ✓ Building a model
- ✓ Simulating data



This talk:

- ? Fitting the model to observed data

Recap: Bayesian workflow

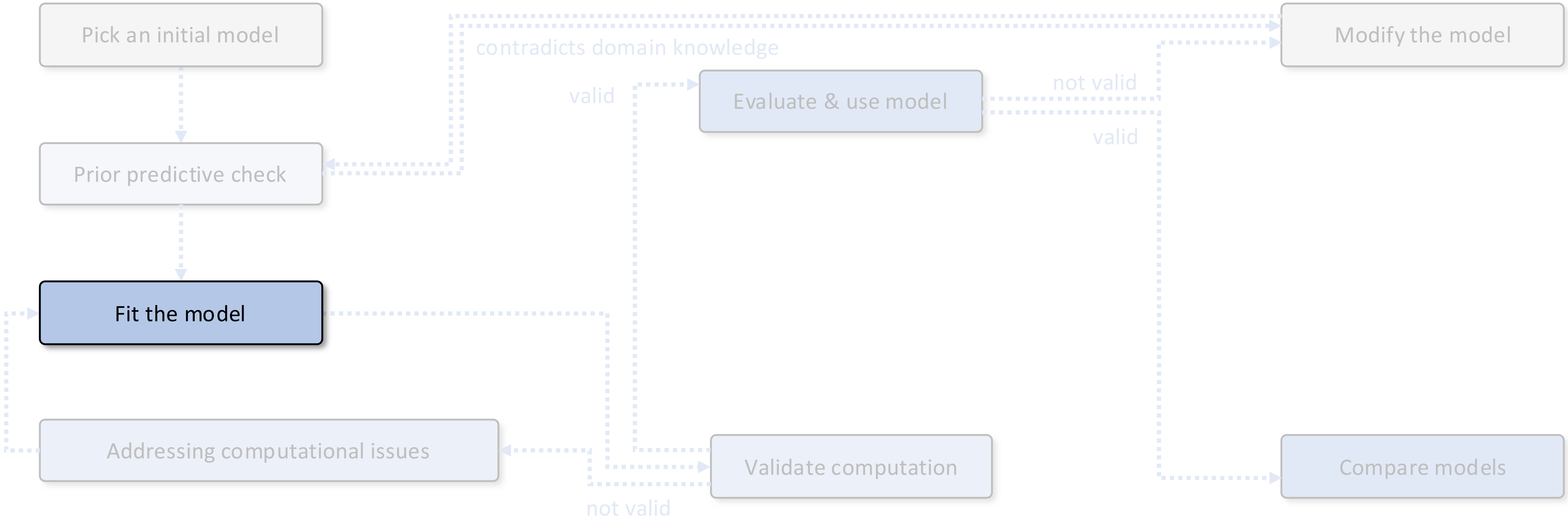


Figure from Gelman et al., 2020, *arXiv*

MLE: maximum likelihood estimator

Principle:

Find the parameters θ for which the **acquired data Y is most likely** under the model m.

$$\theta_{MLE} = \arg\max_{\theta \in \Theta} \underbrace{p(Y \mid \theta, m)}_{\text{Likelihood}}$$

where

$$p(Y \mid \theta, m) = p(y_{1 \ldots T} \mid \theta, m)$$

m	model
Θ	parameter space
θ	model parameters
θ_{MLE}	mle estimate of θ
Y	observed dataset
y_t	single observation
T	number of trials

Example: slot machines

Understand how people learn to maximize their rewards in a case where the most rewarding choice is initially unknown.

vs.

$p($$|$$) = 0.8$

$p($$|$$) = 0.2$

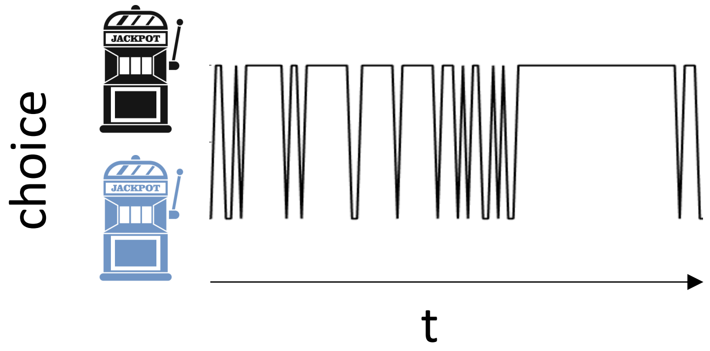
Observations

Dataset:

Choice y_t in each trial t

$$Y = (y_1, \dots, y_T)$$

Experiment



Based on slides by Alex Hess

Wilson & Collins, 2019, *eLife*

Specifying the likelihood function

Model 1
Random choice

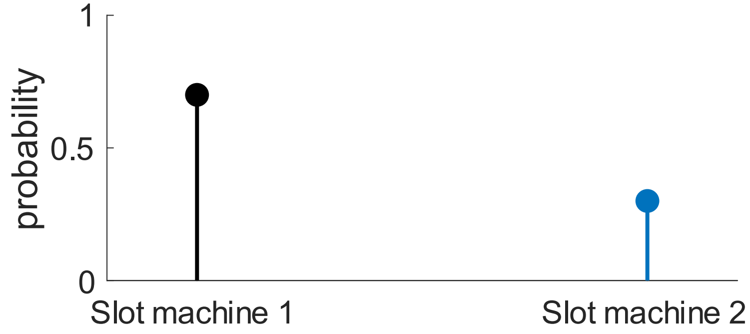
$$p_t^1 = b$$
$$p_t^0 = 1 - b$$

$$0 \le b \le 1$$

$$\boldsymbol{\theta} = \{\boldsymbol{b}\}$$

For a single trial t:

$$p(\, y_t \mid \theta, m\,) = \theta^{y_t}(1-\theta)^{(1-y_t)}$$
$$= Bernoulli$$



$$Y \overset{iid}{=} \{y_1, \dots, y_T\}$$

For all trials 1 … T:

$$p(\, Y \mid \theta, m\,) = p(\, y_{1\dots T} \mid \theta, m\,) = \prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}$$

Maximising the likelihood function

$$\underset{\theta \in \Theta}{\arg\max}$$

Likelihood function

$$p(Y \mid \theta, m) = \prod_{t=1}^{T} p(y_t \mid \theta, m)$$

Analytical solution

Is the likelihood tractable?
Is the likelihood differentiable?

→ Solve $\frac{d}{d\theta} p(Y \mid \theta, m) \overset{!}{=} 0$ and find maximum

Numerical solution

Use numerical optimization routines available in different software (MATLAB, Python, Julia, etc.)

→ Implement $p(Y \mid \theta, m)$ and find the maximum

Maximising the likelihood function

Analytic solution

$$p(Y \mid \theta, m) = \prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}$$

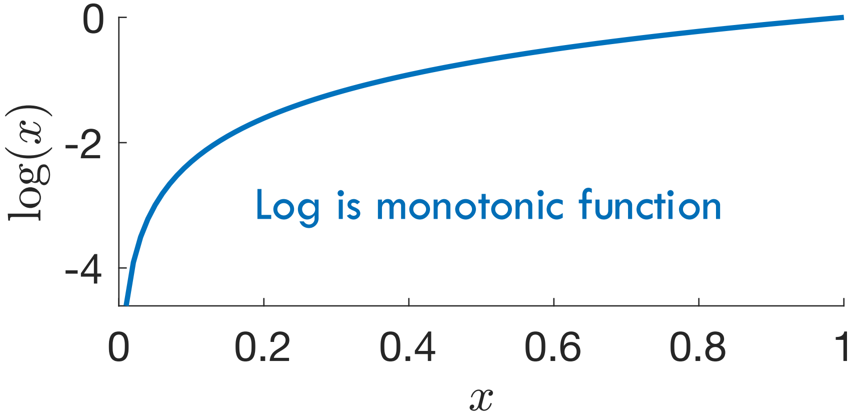
$\downarrow$ 1.

$$\log p(Y \mid \theta, m) = \log \prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}$$

$$= \sum_{t=1}^{T} \log \theta^{y_t}(1-\theta)^{(1-y_t)}$$

1. Change product to sum by log-transformation:

$$\log\left(\prod_t x_t\right) = \sum_t \log x_t$$



Maximising the likelihood function
Analytic solution

$$p(Y \mid \theta, m) = \prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}$$

$$\Big\downarrow 1.$$

$$\log p(Y \mid \theta, m) = \log \prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}$$

$$= \sum_{t=1}^{T} \log \theta^{y_t}(1-\theta)^{(1-y_t)}$$

$$\Big\downarrow 2.$$

$$= \sum_{t=1}^{T} y_t \log(\theta) + (1-y_t)\log(1-\theta)$$

1. Change product to sum by log-transformation:
$$\log\left(\prod_t x_t\right) = \sum_t \log x_t$$

2. Logarithm of a power:
$$\log(x^a) = a \log x$$

Maximising the likelihood function
Analytic solution

$$\frac{d}{d\theta}\sum_{t=1}^{T} y_t \log(\theta) + (1 - y_t)\log(1-\theta) \overset{!}{=} 0$$

$$\left(\frac{d}{d\theta}\log(\theta)\right)\left(\sum_{t=1}^{T} y_t\right) + \left(\frac{d}{d\theta}\log(1-\theta)\right)\left(\sum_{t=1}^{T} 1 - y_t\right) \overset{!}{=} 0$$

$$\frac{1}{\theta(1-\theta)}\left(\sum_{t=1}^{T} y_t - \theta\sum_{t=1}^{T} y_t - \theta T + \theta\sum_{t=1}^{T} y_t\right) \overset{!}{=} 0$$

$$\sum_{t=1}^{T} y_t - \theta T \overset{!}{=} 0$$

MLE estimate

$$\boxed{\theta_{MLE} = \frac{1}{T}\sum_{t=1}^{T} y_t}$$

$\downarrow$

MLE estimate is arithmetic mean of data!

Advantages of maximum likelihood estimation

1. **Easy to compute:** simple optimisation problem

Advantages of maximum likelihood estimation

1. **Easy to compute:** simple optimisation problem

2. **Interpretable:** often θ_{MLE} is intuatively interpretable wrt. model parameters (see MLE of random choice model)

Advantages of maximum likelihood estimation

1. **Easy to compute:** simple optimisation problem

2. **Interpretable:** often θ_{MLE} is intuatively interpretable wrt. model parameters (see MLE of random choice model)

3. **Asymptotic properties:** consistency (true parameter value recovered) and efficiency (lowest possible parameter variance)

Advantages of maximum likelihood estimation

1. **Easy to compute:** simple optimisation problem

2. **Interpretable:** often θ_{MLE} is intuatively interpretable wrt. model parameters (see MLE of random choice model)

3. **Asymptotic properties:** consistency (true parameter value recovered) and efficiency (lowest possible parameter variance)

4. **Invariant to reparameterization:** if $\theta_{MLE} = \mathrm{MLE}(\theta)$ for $\theta \in \Theta$, then $g(\theta_{MLE}) = \mathrm{MLE}(\mathrm{g}(\theta))$ for $g: \mathbb{R} \to \mathbb{R}$

Limitations of maximum likelihood estimation

1. **Point estimate:** θ_{MLE} is a point estimate → no representation of uncertainty

Limitations of maximum likelihood estimation

1. **Point estimate:** θ_{MLE} is a point estimate → no representation of uncertainty

2. **Existence & uniqueness:** The MLE might not be unique or even non existent, depending on properties of the likelihood function and parameter space

Limitations of maximum likelihood estimation

1. **Point estimate:** θ_{MLE} is a point estimate → no representation of uncertainty

2. **Existence & uniqueness:** The MLE might not be unique or even non existent, depending on properties of the likelihood function and parameter space

3. **Overfitting:** MLE is limited to a finite set of observed datapoints

Limitations of maximum likelihood estimation

3. **Overfitting:** MLE is limited to observed data but does not (explicitly) take into account prior information

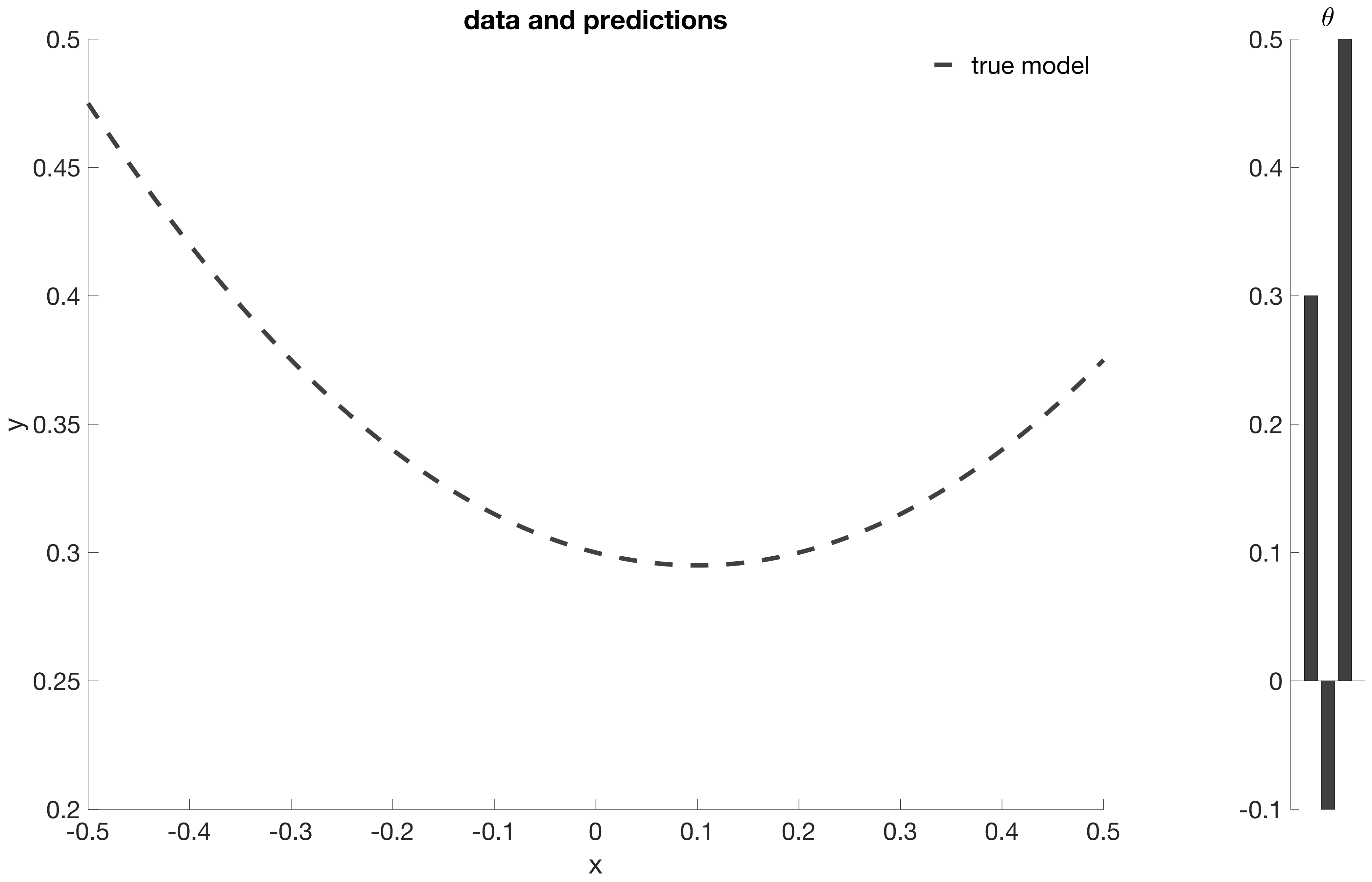
Example

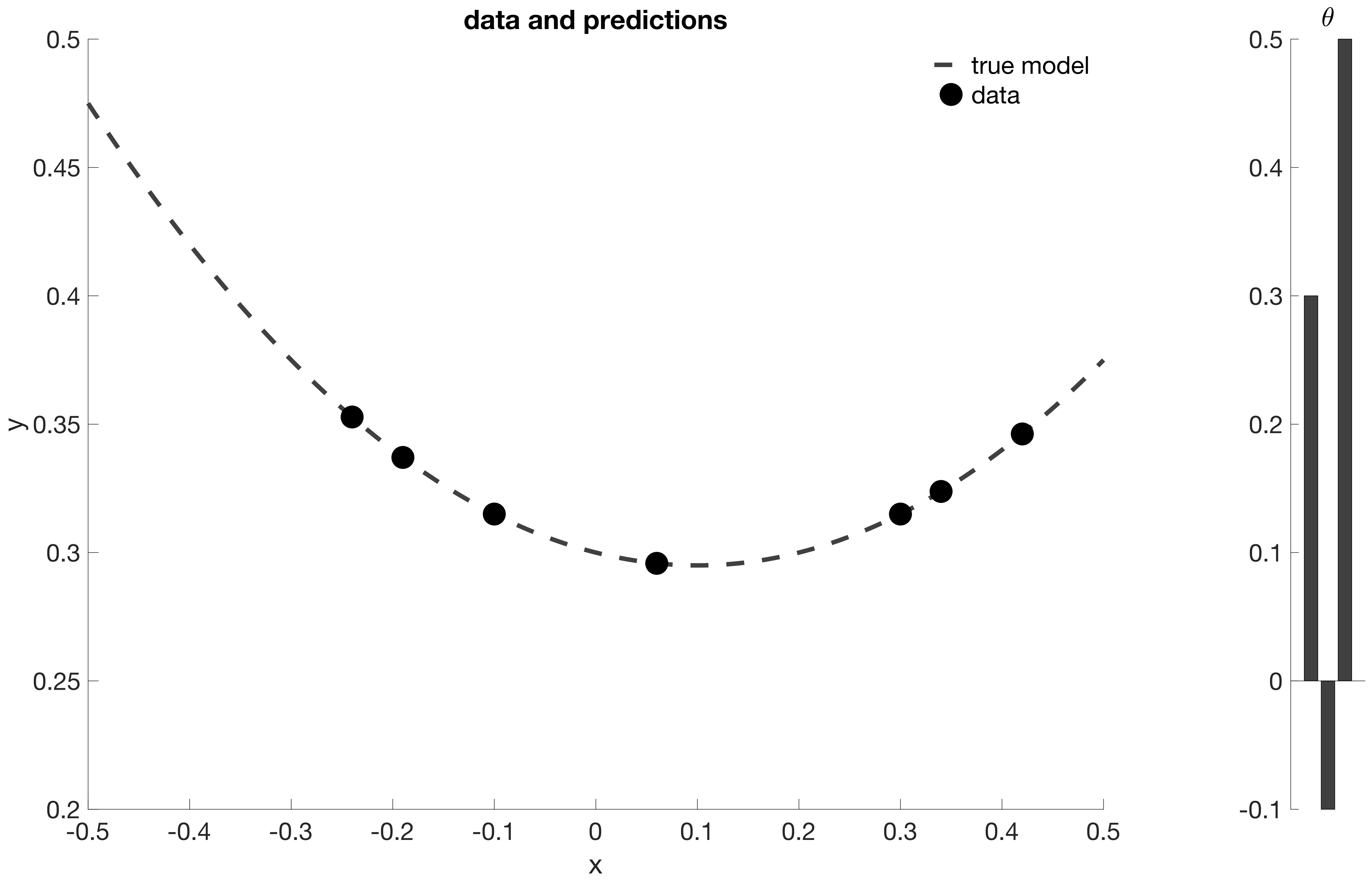
Polynomial model of order P with Gaussian noise

$$y = \theta_0 + \theta_1 x + \dots + \theta_P x^P + \epsilon = \boldsymbol{x\theta} + \epsilon$$

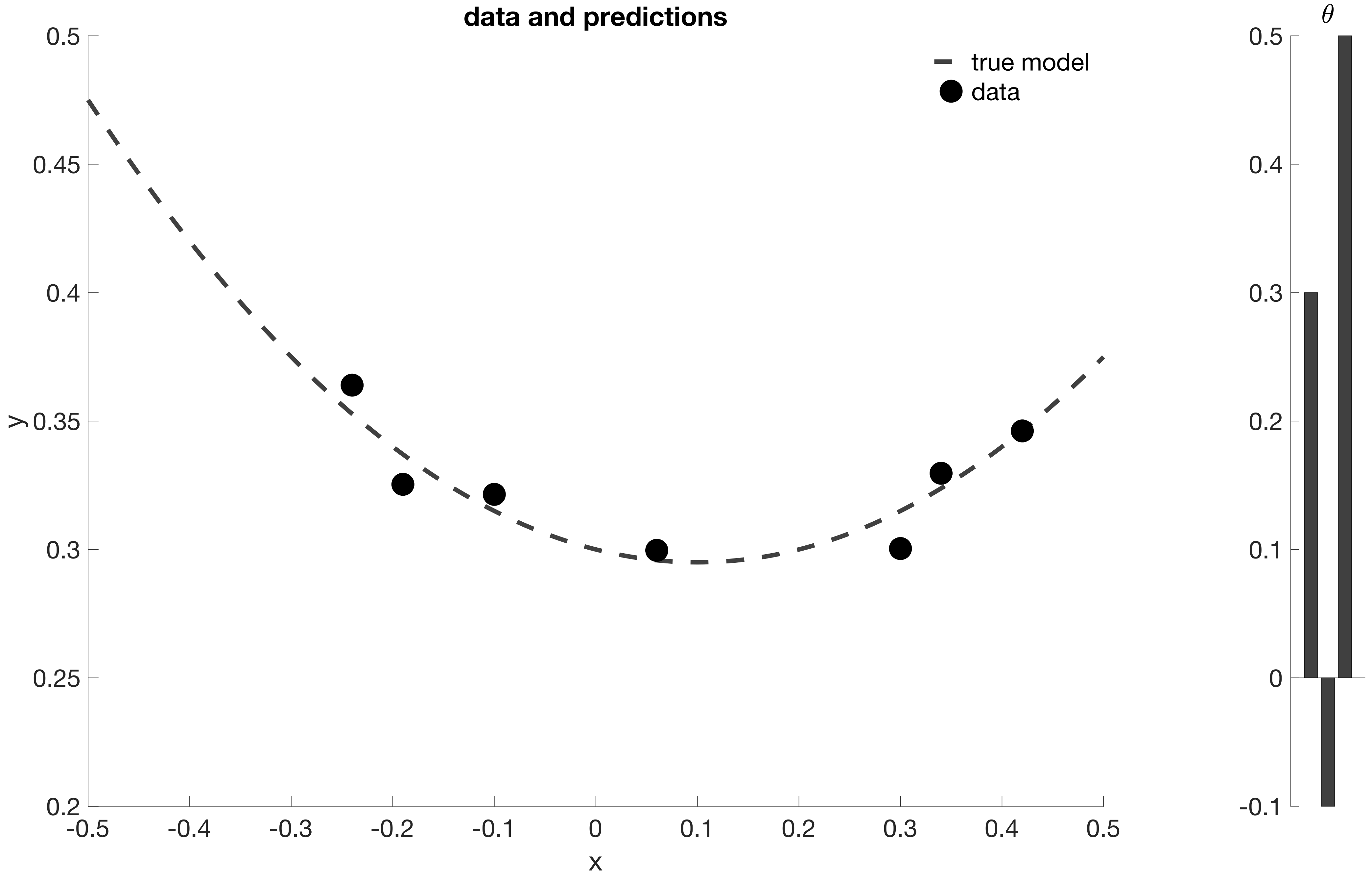
where

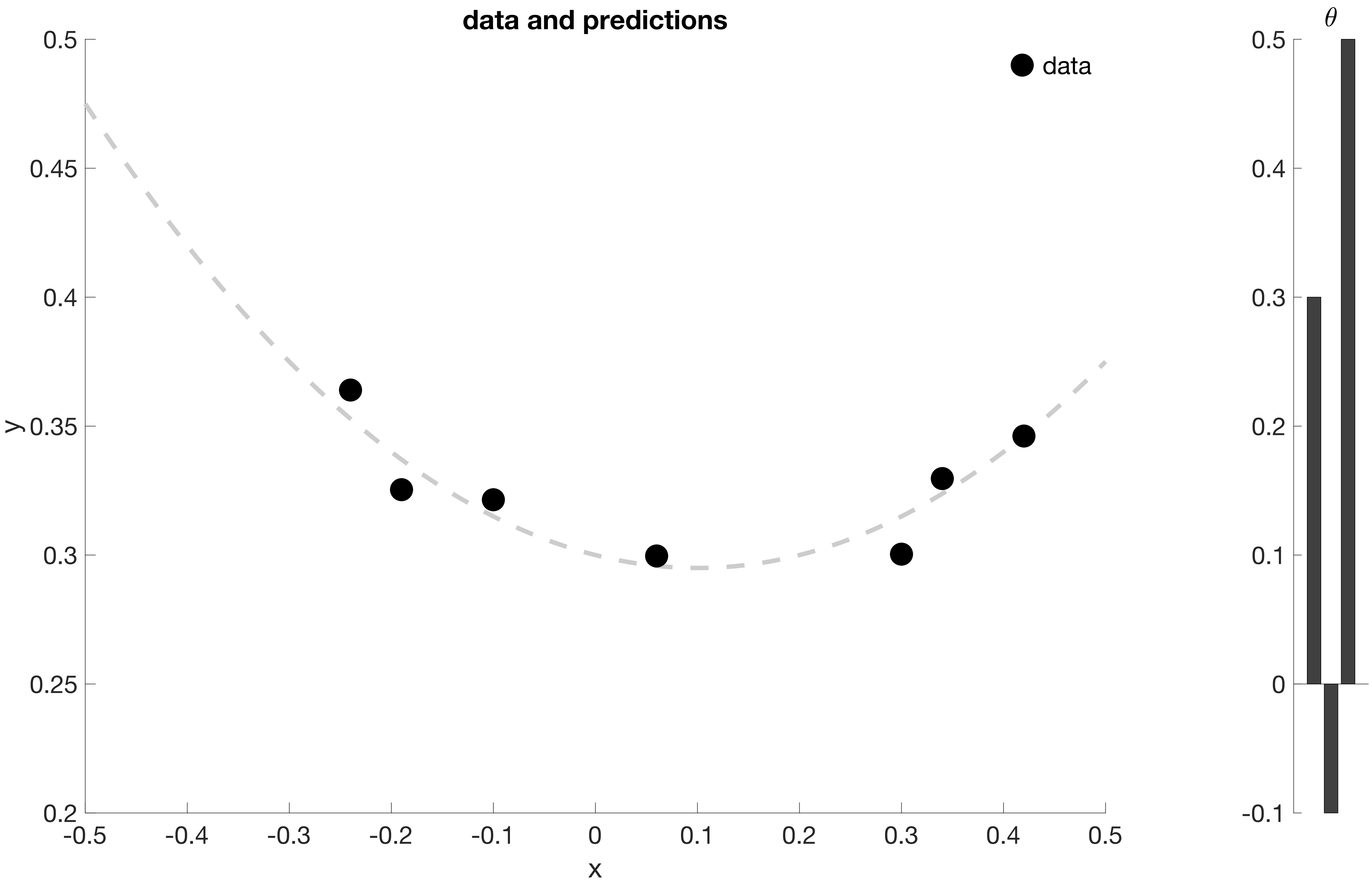
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

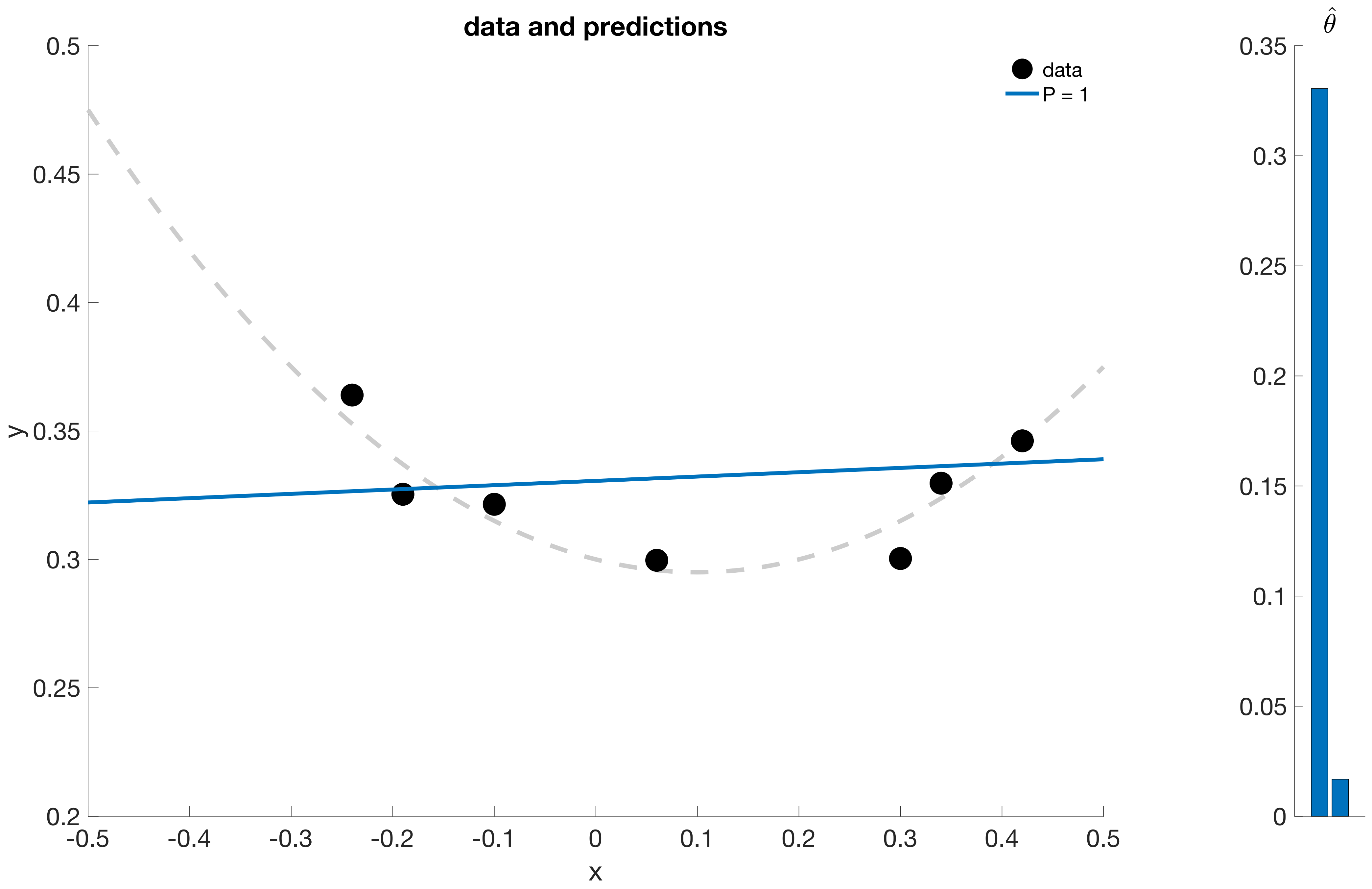

data and predictions
true model
0.5
0.45
0.4
0.35
0.3
0.25
0.2
y
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
x
θ
0.5
0.4
0.3
0.2
0.1
0
-0.1

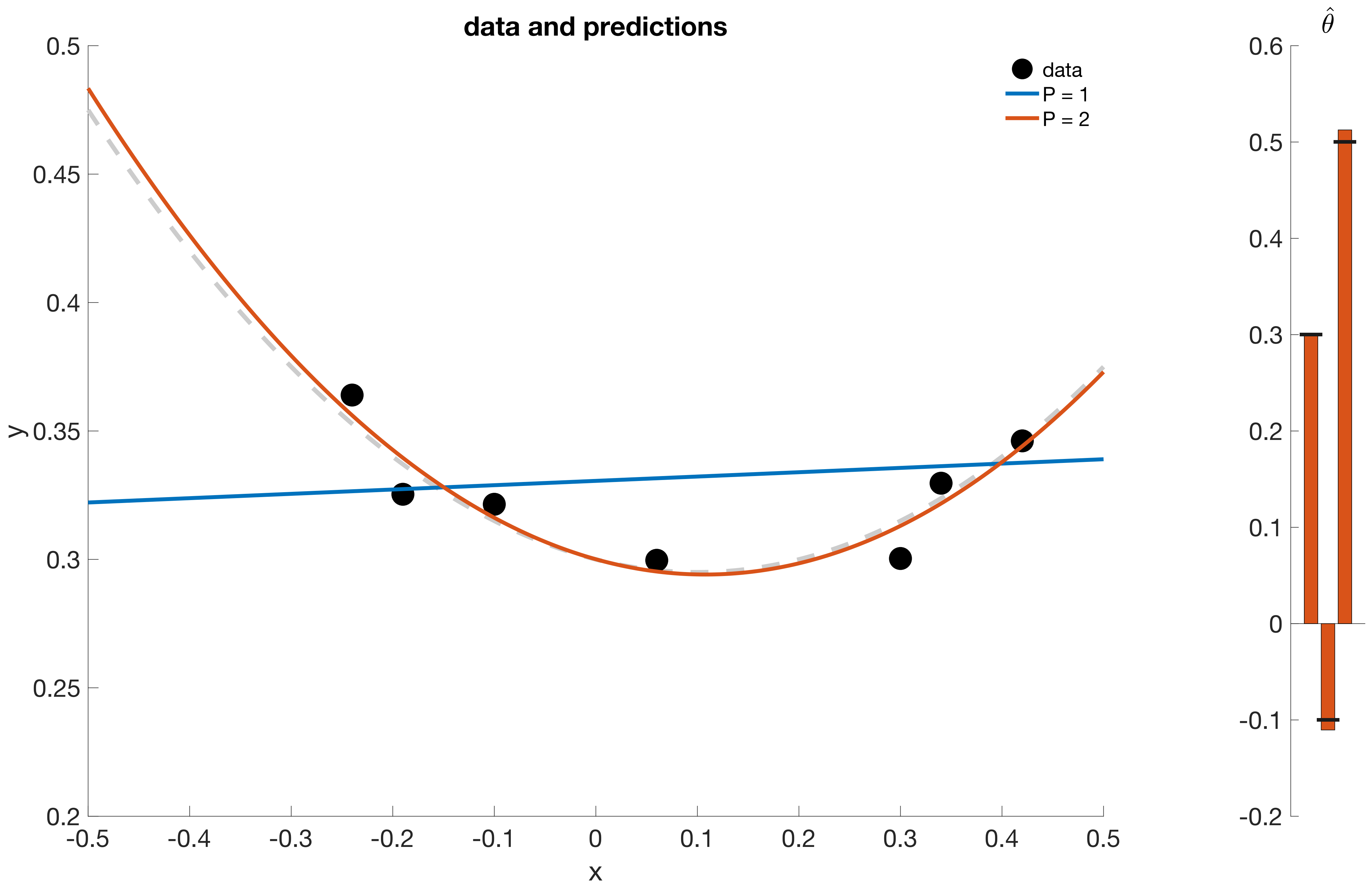


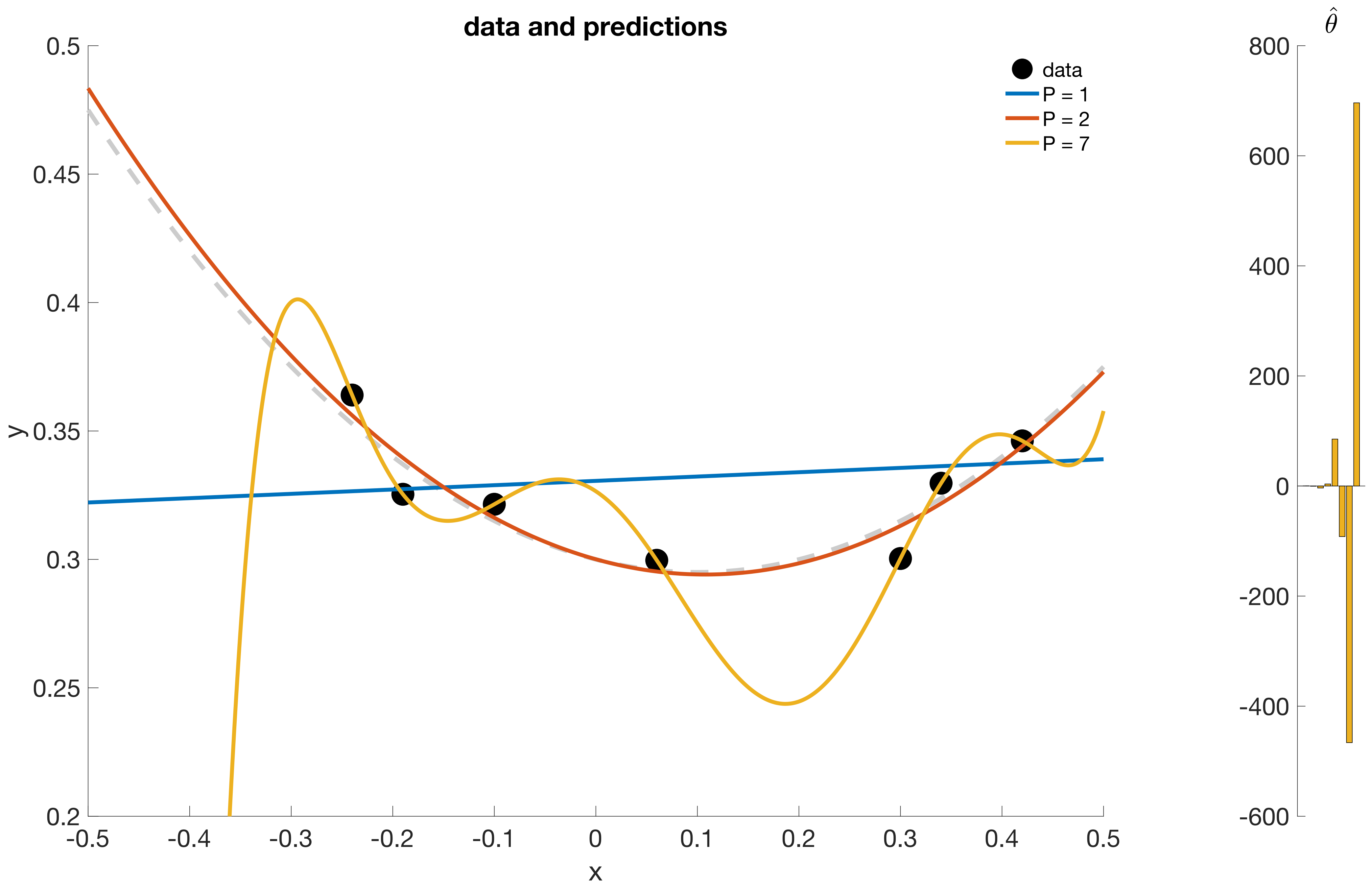
data and predictions
true model
data
y
x
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.2
0.25
0.3
0.35
0.4
0.45
0.5
θ
-0.1
0
0.1
0.2
0.3
0.4
0.5


data and predictions
θ
true model
data
y
x
0.5
0.45
0.4
0.35
0.3
0.25
0.2
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.5
0.4
0.3
0.2
0.1
0
-0.1

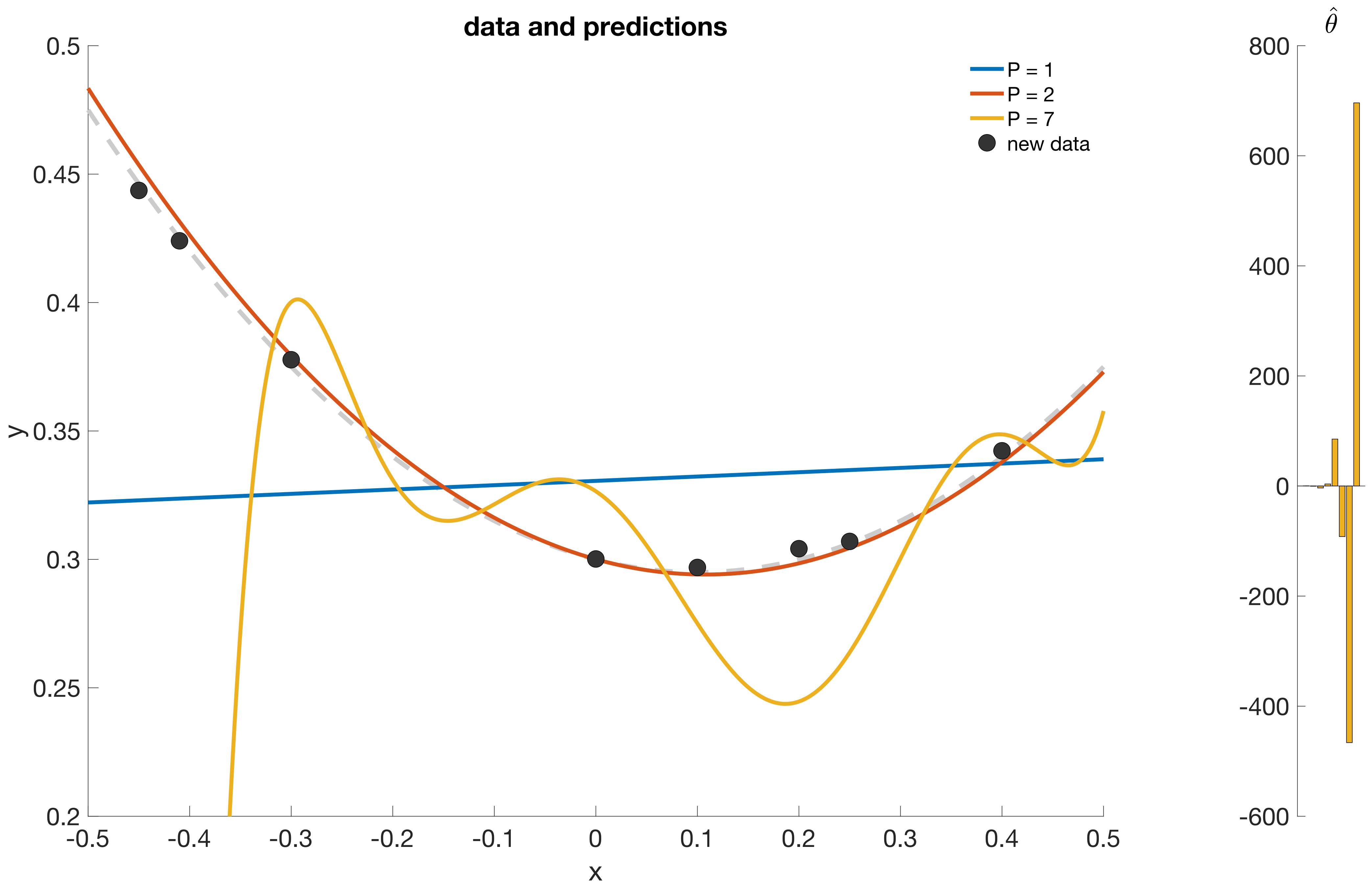

data and predictions
data
y
x
0.5
0.45
0.4
0.35
0.3
0.25
0.2
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
θ
0.5
0.4
0.3
0.2
0.1
0
-0.1


data and predictions
data
P = 1
y
x
0.5
0.45
0.4
0.35
0.3
0.25
0.2
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
$\hat{\theta}$
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0


data and predictions
data
P = 1
P = 2
y
x
0.5
0.45
0.4
0.35
0.3
0.25
0.2
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
$\hat{\theta}$
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2



data and predictions
data
P = 1
P = 2
P = 7
0.5
0.45
0.4
0.35
0.3
0.25
0.2
y
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
x
$\hat{\theta}$
800
600
400
200
0
-200
-400
-600


data and predictions
P = 1
P = 2
P = 7
new data
y
x
0.5
0.45
0.4
0.35
0.3
0.25
0.2
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
$\hat{\theta}$
800
600
400
200
0
-200
-400
-600

Maximum likelihood vs. full Bayesian inference

$$p(\theta \mid y, m) = \frac{p(y \mid \theta, m)p(\theta, m)}{p(y \mid m)}$$

Maximum-a-posteriori (**MAP**) estimation:

→ Point estimate of the posterior

→ Under a flat prior MAP=MLE

Variational Bayes (**VB**), sampling-based (**MCMC**) techniques

→ Full posterior densities

Translational Neuromodeling Unit

Acknowledgement

Special thanks to my TNU colleagues

Stefan Frässle

Katharina Wellstein

Jakob Heinzle

Klaas Enno Stephan

QUESTIONS?

Schoenleitner@biomed.ee.ethz.ch

https://github.com/computational-psychiatry-course/cpc2024

https://www.linkedin.com/in/florian-schönleitner-34201210a/