Probabilistic Approaches for Computational Biology and Medicine

Neil D. Lawrence

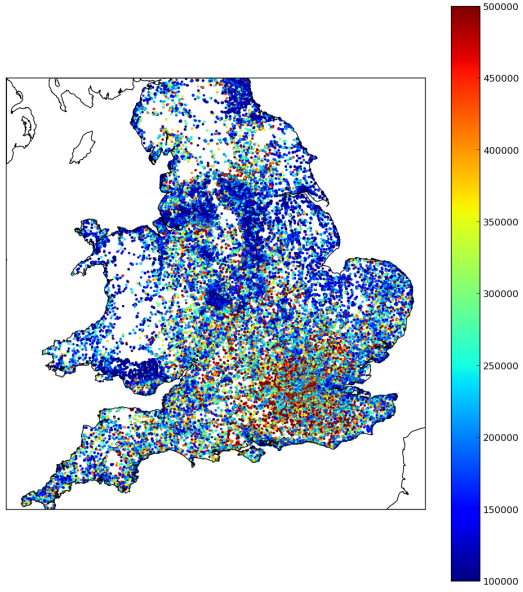
MLPM Summer School

25th September 2013

Outline

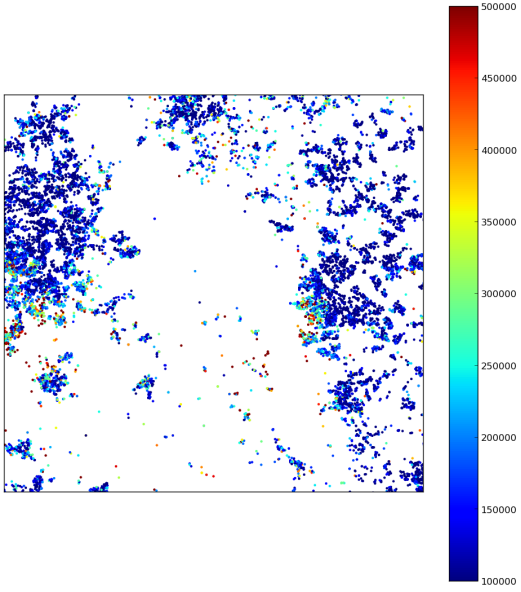
What's Changed (Changing) for Medicine?

- Modern data availability.



What's Changed (Changing) for Medicine?

- Modern data availability.



Gaussian Processes for Big Data

James Hensman*
Dept. Computer Science
The University of Sheffield
Sheffield, UK

Nicolò Fusi*
Dept. Computer Science
The University of Sheffield
Sheffield, UK

Neil D. Lawrence*
Dept. Computer Science
The University of Sheffield
Sheffield, UK

Abstract

We introduce stochastic variational inference for Gaussian process models. This enables the application of Gaussian process (GP) models to data sets containing millions of data points. We show how GPs can be variationally decomposed to depend on a set

Even to accommodate these data sets, various approximate techniques are required. One approach is to partition the data set into separate groups [e.g. Snelson and Ghahramani, 2007, Urtasun and Darrell, 2008]. An alternative is to build a low rank approximation to the covariance matrix based around 'inducing variables' [see e.g. Csató and Opper, 2002, Seeger et al., 2003, Quiñonero Candela and Rasmussen, 2005, Tit-

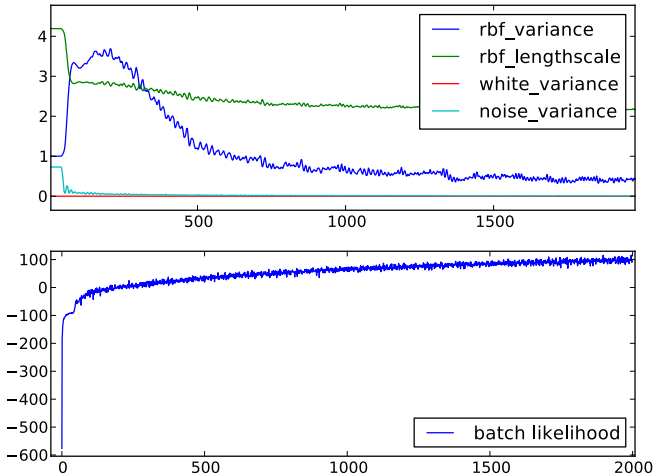


Figure 4: Convergence of the SVIGP algorithm on the two dimensional toy data

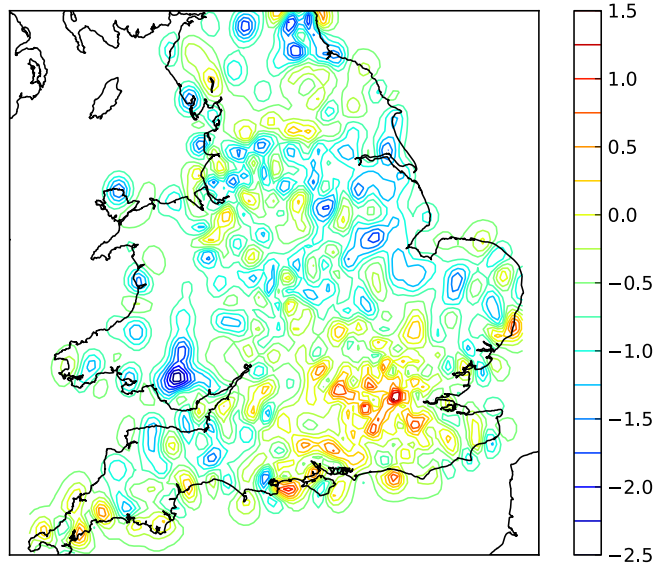


Figure 5: Variability of apartment price (logarithmically!) throughout England and Wales.

`land-registry-monthly-price-paid-data/`, which covers England and Wales, and filtered for apartments. This resulted in a data set with 75,000 entries,

What's Changed (Changing) for Medicine?

- Try Googling for: "patient data "...

Image from Wikimedia Commons

Image from Wikimedia Commons

Driver and Vehicle Licensing Agency

INF57

A brief history *of Registration*

For more information go to: **www.direct.gov.uk/motoring**

A brief history of registration

The early days

Prior to the appearance of the first railways in Britain, there was a brief development and interest in steam powered road going vehicles. In 1834, a Mr Hancock started a steam coach called the "Era", carrying up to 14 passengers from Paddington to Regents Park and the City at 6d a head. And in the following year, a Mr Church built an omnibus capable of carrying 40 passengers for the London and Birmingham Steam Carriage Company.

However, the success of the railway movement drove all such traffic off the roads.
A Parliamentary Commission of Enquiry in 1836 reported "strongly in favour of steam carriages on roads", but subsequent Acts of Parliament tended to have a discouraging and restrictive effect. **The Locomotive Act 1861** limited the weight of steam engines to 12 tons and imposed a speed limit of 10 mph.

The Locomotive Act 1865 set a speed limit of 4 mph in the country and 2 mph in towns. The 1865 Act also provided for the famous "man with a red flag". Walking 60 yards ahead of each vehicle, a man with a red flag or lantern enforced a walking pace, and warned horse riders and horse drawn traffic of the approach of a self propelled machine.

The Locomotive Amendment Act 1878 made the red flag optional under local regulations, and

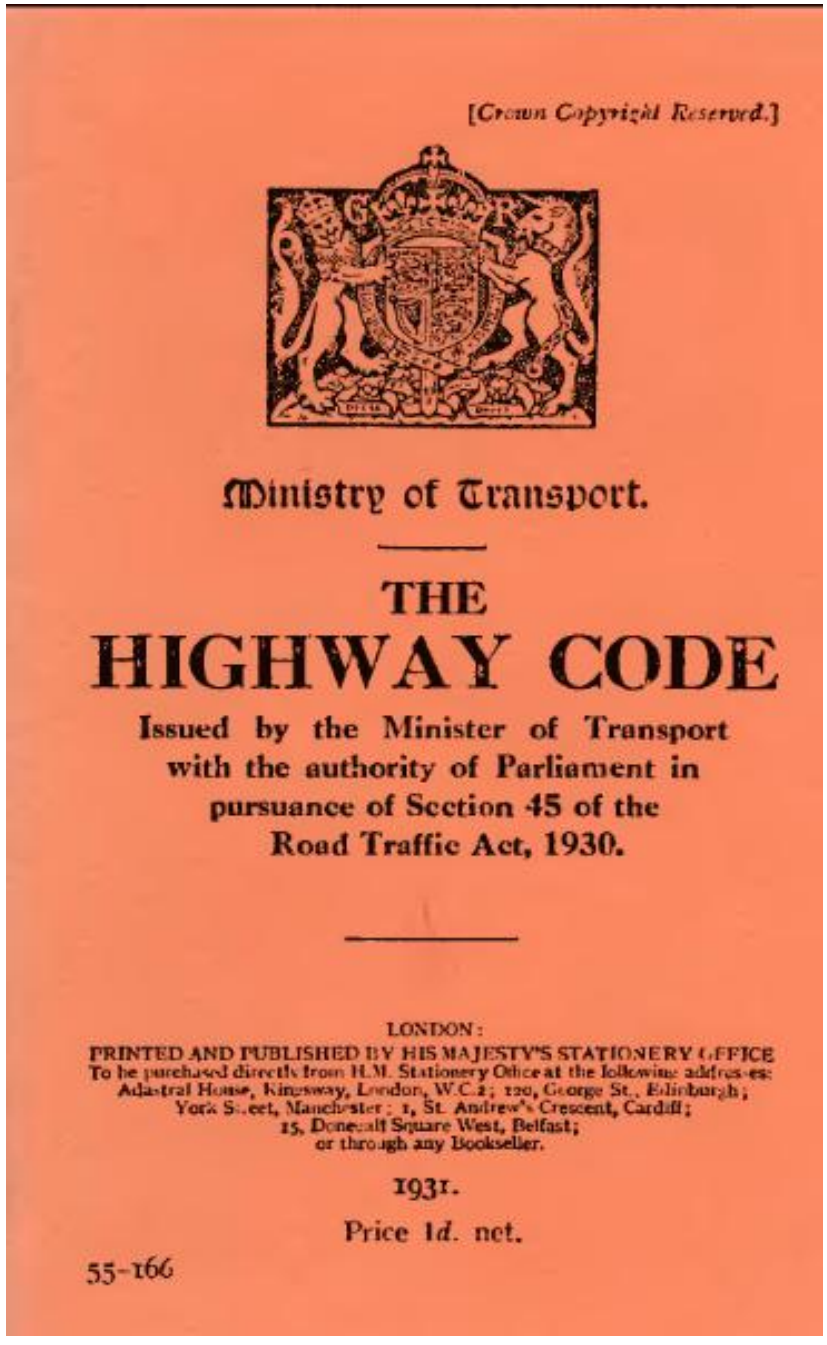


Image from Wikimedia Commons

Image from Wikimedia Commons

What's Changed (Changing) for Medicine?

- Genotyping.
- Epigenotyping.
- Transcriptome: detailed characterization of phenotype.
  - Stratification of data.

Open Data

- Automatic data curation: from curated data to curation of publicly available data.
- Open Data: http://www.openstreetmap.org/?lat=53.38086&lon=-1.48545&zoom=17&layers=M.

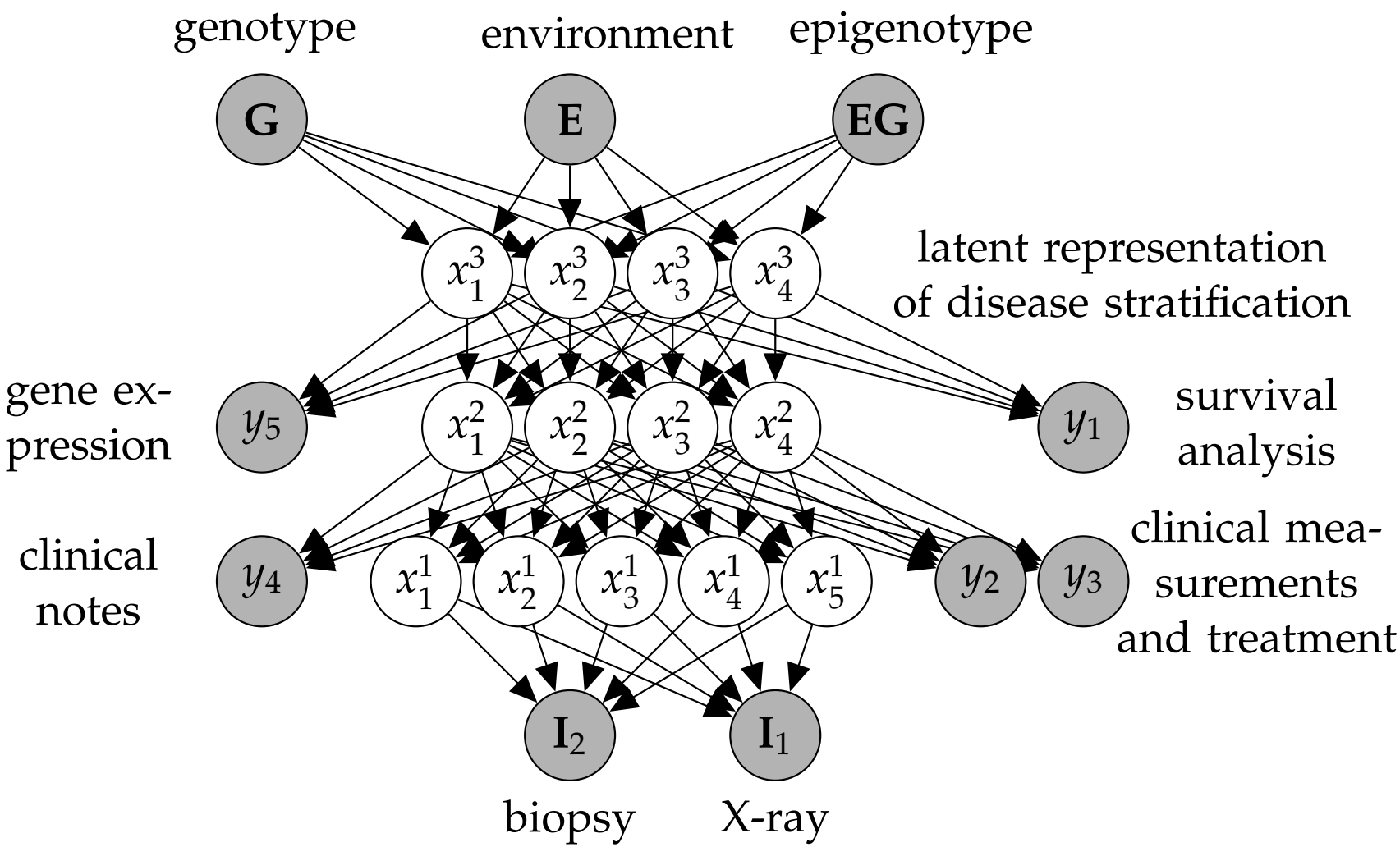
Open Data

- Automatic data curation: from curated data to curation of publicly available data.
- Open Data: http://www.openstreetmap.org/?lat=53.38086&lon=-1.48545&zoom=17&layers=M.

Data Sources

- UK Goverment Stipulation on Data Availability Telegraph Article
- Patient Access: http://www.patient.co.uk/patient-access.asp
- The midata project: Tescos, T-mobile ...
- A social network for personal health?? e.g. EMIS myHealth

Deep Health


genotype
environment
epigenotype
G
E
EG
x_1^3 x_2^3 x_3^3 x_4^3
latent representation of disease stratification
gene expression
y_5
x_1^2 x_2^2 x_3^2 x_4^2
y_1
survival analysis
clinical notes
y_4
x_1^1 x_2^1 x_3^1 x_4^1 x_5^1
y_2
y_3
clinical measurements and treatment
I_2
I_1
biopsy
X-ray

Missing Data

- If missing at random it can be marginalized.
- As data sets become very large (39 million in EMIS) data becomes extremely sparse.
- Imputation becomes impractical.

Imputation

- Expectation Maximization (EM) is gold standard imputation algorithm.
- Exact EM optimizes the log likelihood.
- Approximate EM optimizes a lower bound on log likelihood.
  - e.g. variational approximations (VIBES, Infer.net).
- Convergence is *guaranteed* to a local maxima in log likelihood.

Expectation Maximization

Require: An initial guess for missing data

Expectation Maximization

Require: An initial guess for missing data
repeat

Expectation Maximization

Require: An initial guess for missing data
  repeat
    Update model parameters (M-step)

Expectation Maximization

Require: An initial guess for missing data
  repeat
    Update model parameters (M-step)
    Update guess of missing data (E-step)

Expectation Maximization

Require: An initial guess for missing data
  repeat
    Update model parameters (M-step)
    Update guess of missing data (E-step)
  until convergence

Imputation is Impractical

- In very sparse data imputation is impractical.
- EMIS: 39 million patients, thousands of tests.
- For most people, most tests are missing.
- M-step becomes confused by poor imputation.

Direct Marginalization is the Answer

- Perhaps we need joint distribution of two test outcomes,

$$p(y_1, y_2)$$

- Obtained through marginalizing over all missing data,

$$p(y_1, y_2) = \int p(y_1, y_2, y_3, \ldots, y_p) \mathrm{d}y_3, \ldots \mathrm{d}y_p$$

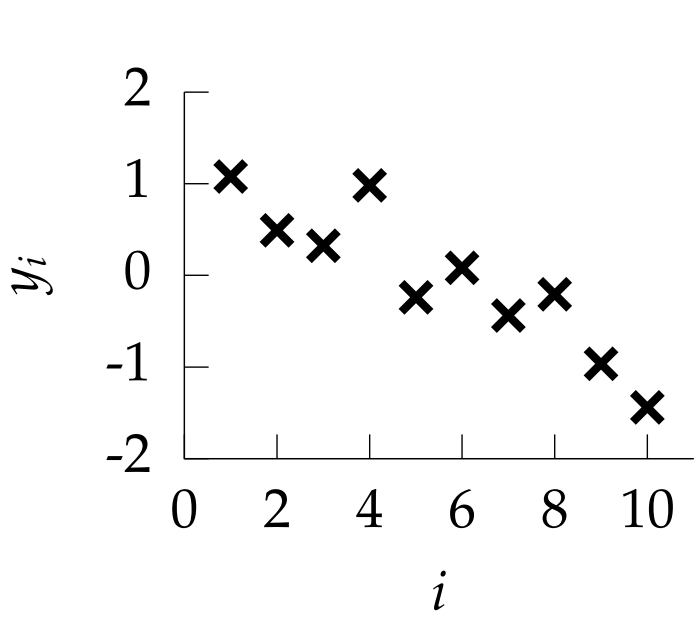
- Where $y_3, \ldots, y_p$ contains:
  1. all tests not applied to this patient
  2. all tests not yet invented!!

Magical Marginalization in Gaussians

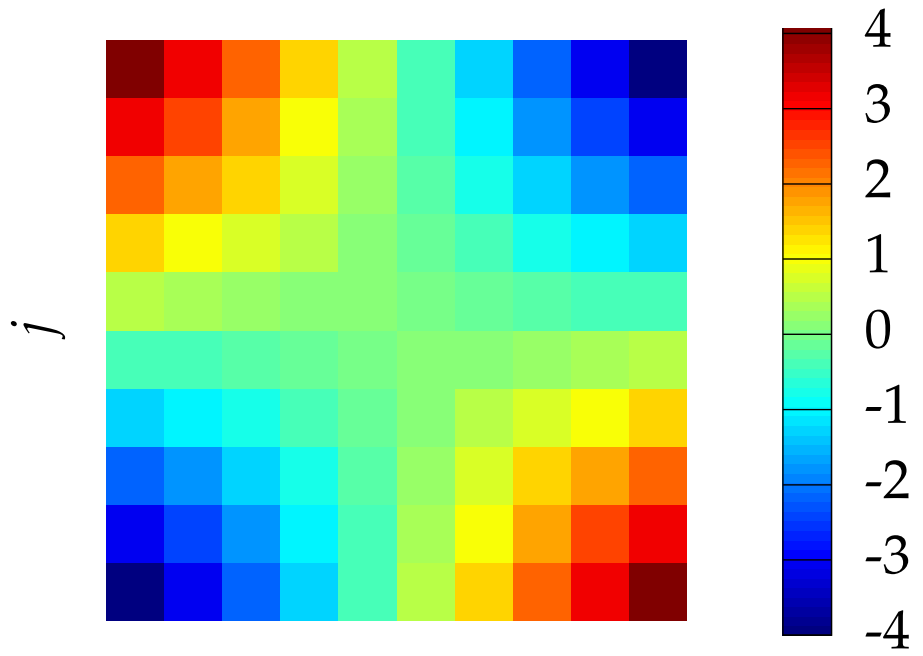
Multi-variate Gaussians

- Given 10 dimensional multivariate Gaussian, $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$.
- Generate a single correlated sample $\mathbf{y} = [y_1, y_2 \ldots y_{10}]$.
- How do we find the marginal distribution of y_1, y_2?

Gaussian Marginalization Property



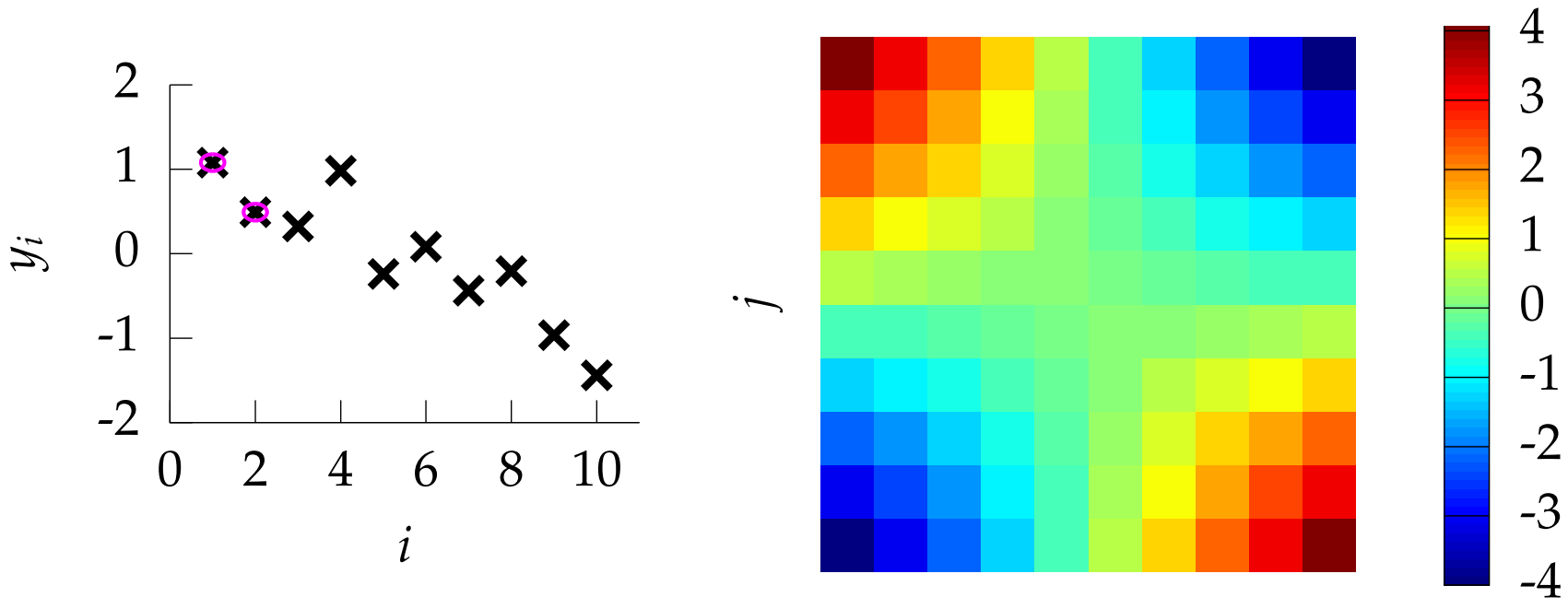
(a) A 10 dimensional sample



(b) colormap showing covariance between dimensions.

Figure: A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property

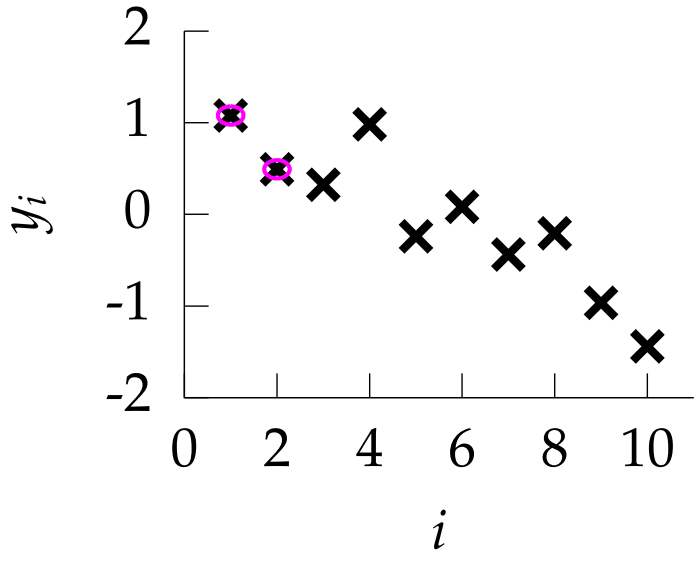


(a) A 10 dimensional sample

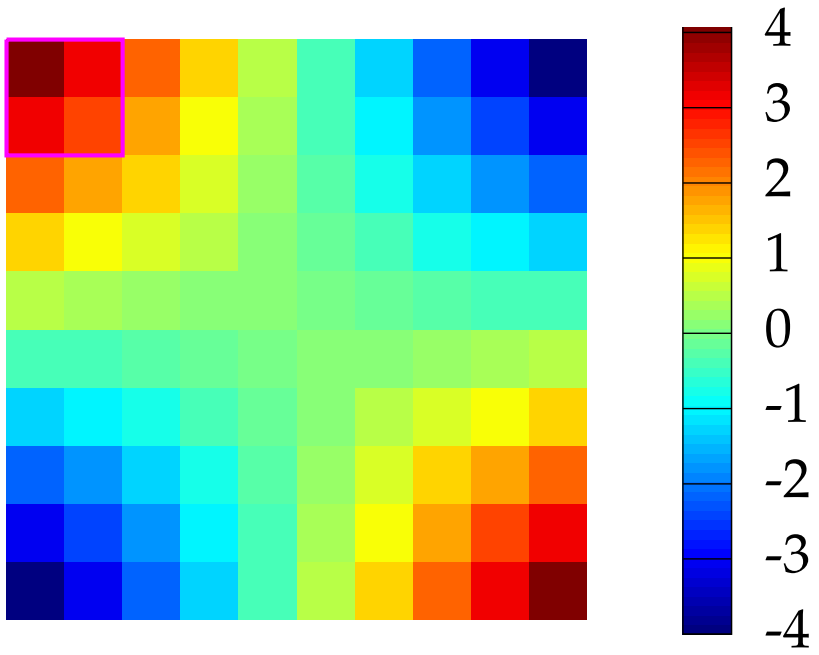
(b) colormap showing covariance between dimensions.

Figure: A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



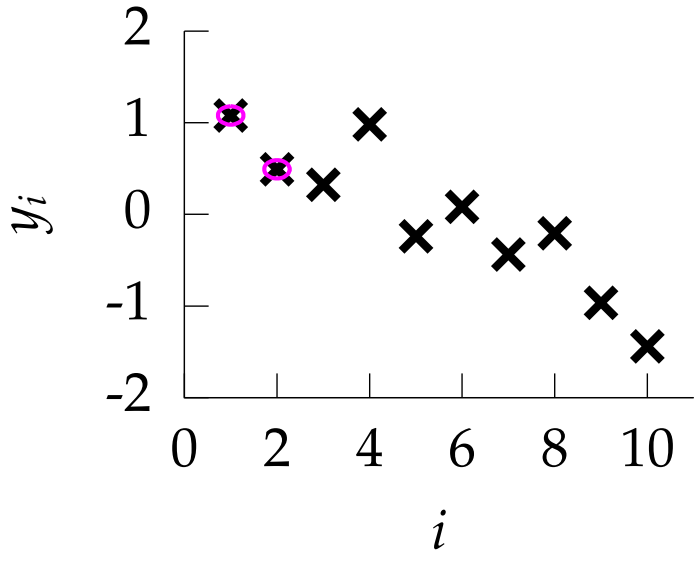
(a) A 10 dimensional sample



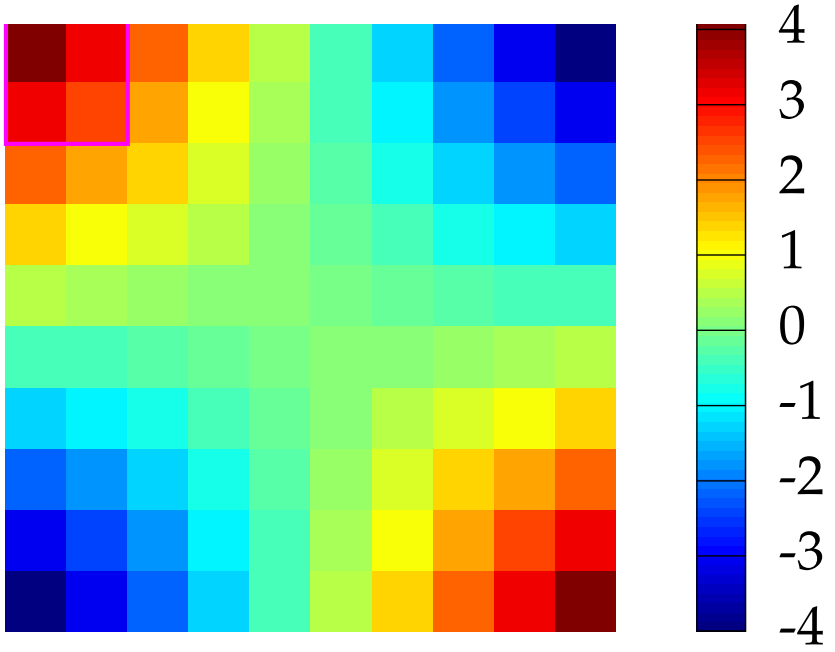
(b) colormap showing covariance between dimensions.

Figure: A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



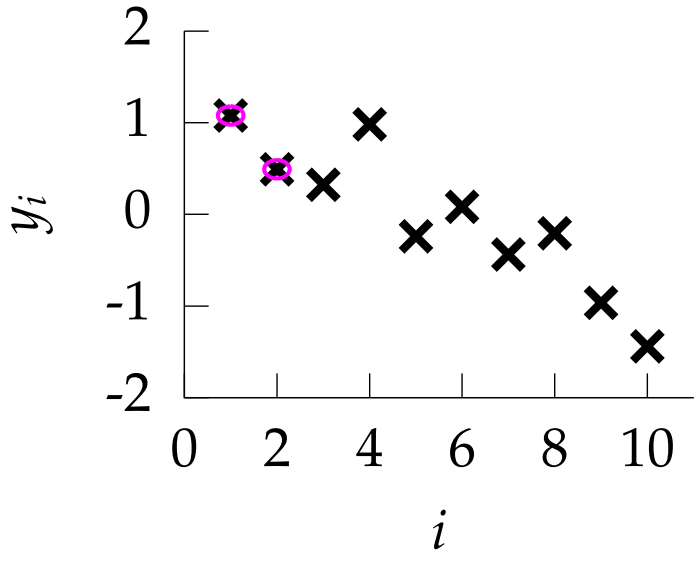
(a) A 10 dimensional sample



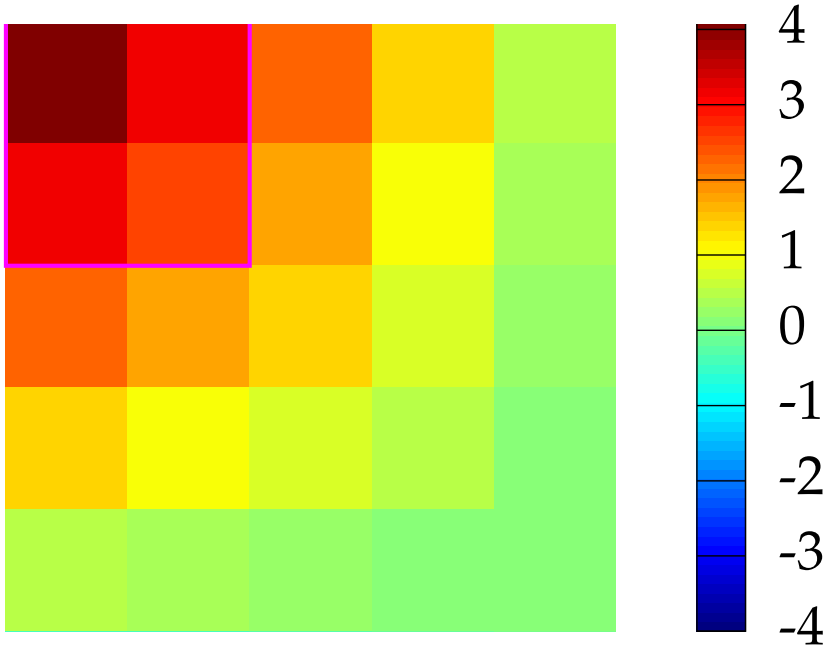
(b) colormap showing covariance between dimensions.

Figure: A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



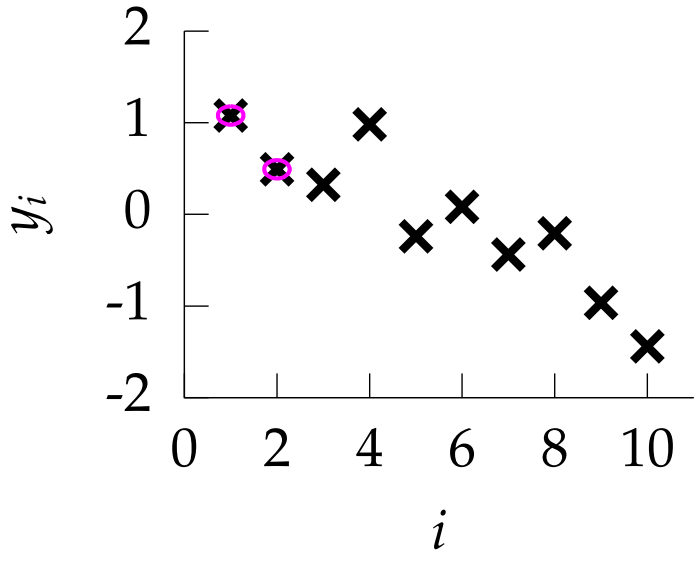
(a) A 10 dimensional sample



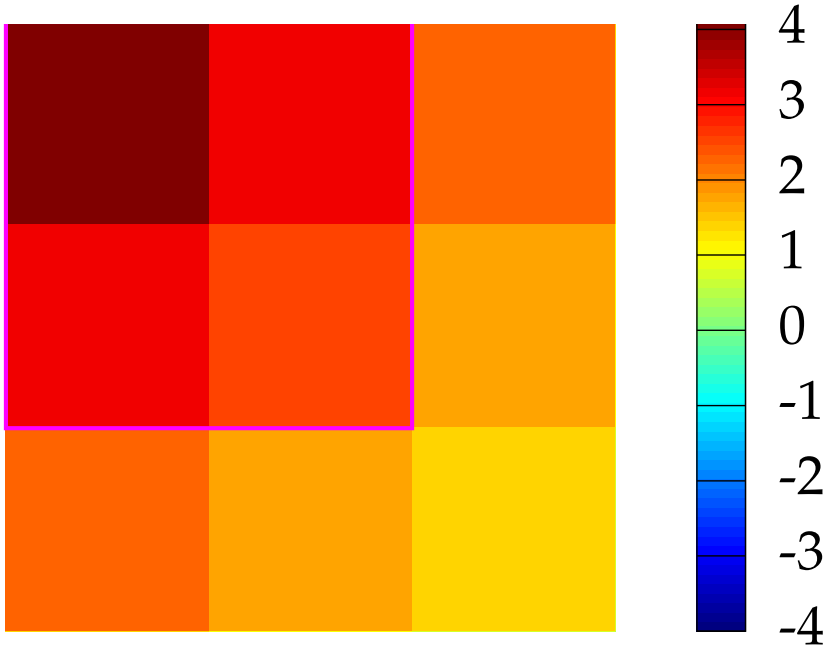
(b) colormap showing covariance between dimensions.

Figure: A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



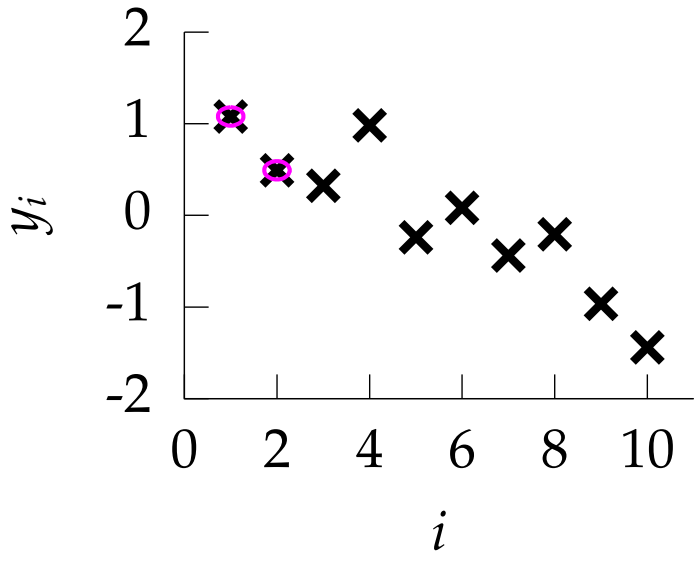
(a) A 10 dimensional sample



(b) colormap showing covariance between dimensions.

Figure: A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



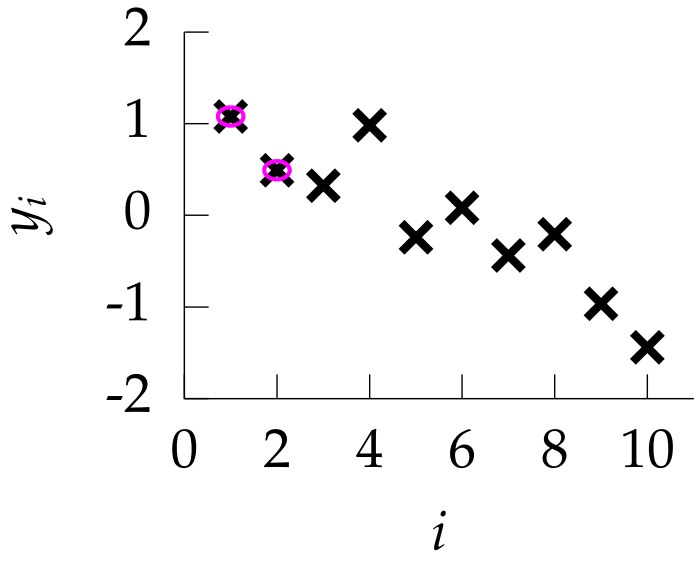
(a) A 10 dimensional sample

$$\begin{bmatrix} 1 & 0.96793 \\ 0.96793 & 1 \end{bmatrix}$$

(b) colormap showing covariance between dimensions.

Figure: A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



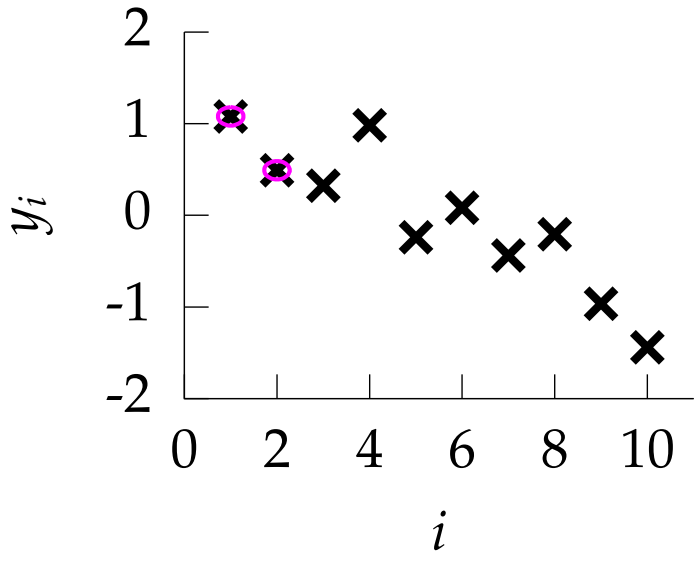
(a) A 10 dimensional sample

$$\begin{bmatrix} 4.1 & 3.1111 \\ 3.1111 & 2.5198 \end{bmatrix}$$

(b) covariance between y_1 and y_2.

Figure: A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



(a) A 10 dimensional sample

$$\begin{bmatrix} 1 & 0.96793 \\ 0.96793 & 1 \end{bmatrix}$$

(b) correlation between y_1 and y_2.

Figure: A sample from a 10 dimensional correlated Gaussian distribution.

Bishop


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP

Outline

Regression Examples

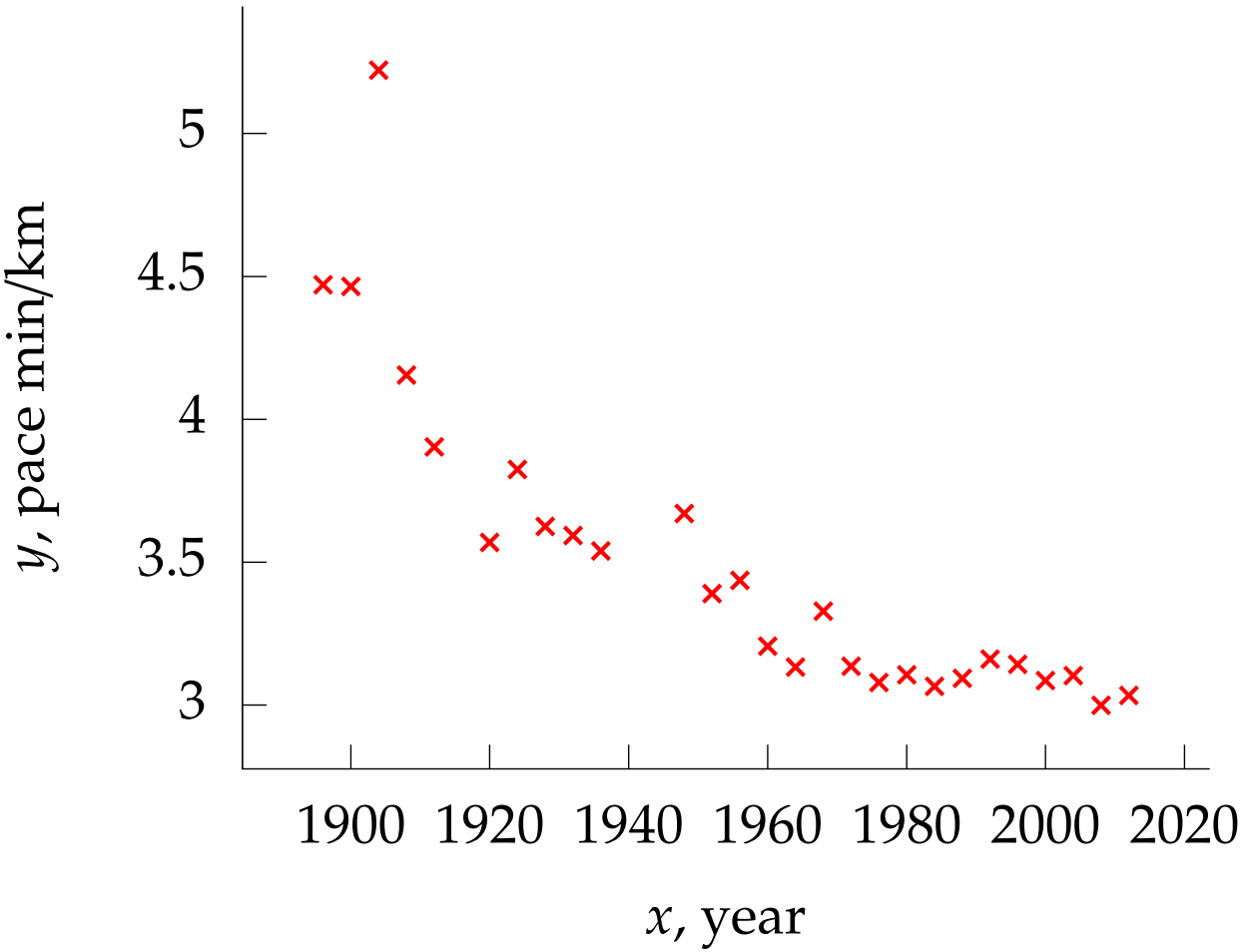
- Predict a real value, y_i given some inputs $\mathbf{x}_i$.
- Predict quality of meat given spectral measurements (Tecator data).
- Radiocarbon dating, the C14 calibration curve: predict age given quantity of C14 isotope.
- Predict quality of different Go or Backgammon moves given expert rated training data.

Olympic Marathon Data

- Gold medal times for Olympic Marathon since 1896.
- Marathons before 1924 didn't have a standardised distance.
- Present results using pace per km.
- In 1904 Marathon was badly organised leading to very slow times.

Image from Wikimedia Commons
http://bit.ly/16kMKHQ

Olympic Marathon Data



Olympic Marathon Data.

What is Machine Learning?

data

- data: observations, could be actively or passively acquired (meta-data).

What is Machine Learning?

data +

- data: observations, could be actively or passively acquired (meta-data).

What is Machine Learning?

data + model

- data: observations, could be actively or passively acquired (meta-data).
- model: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.

What is Machine Learning?

data + model =

- data: observations, could be actively or passively acquired (meta-data).
- model: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.

What is Machine Learning?

data + model = prediction

- data: observations, could be actively or passively acquired (meta-data).
- model: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.
- prediction: an action to be taken or a categorization or a quality score.

Regression: Linear Releationship

$$y = mx + c$$

- y: winning time/pace.

Regression: Linear Releationship

$$y = mx + c$$

- y: winning time/pace.
- x: year of Olympics.

Regression: Linear Releationship

$$y = mx + c$$

- y: winning time/pace.
- x: year of Olympics.
- m: rate of improvement over time.

Regression: Linear Releationship

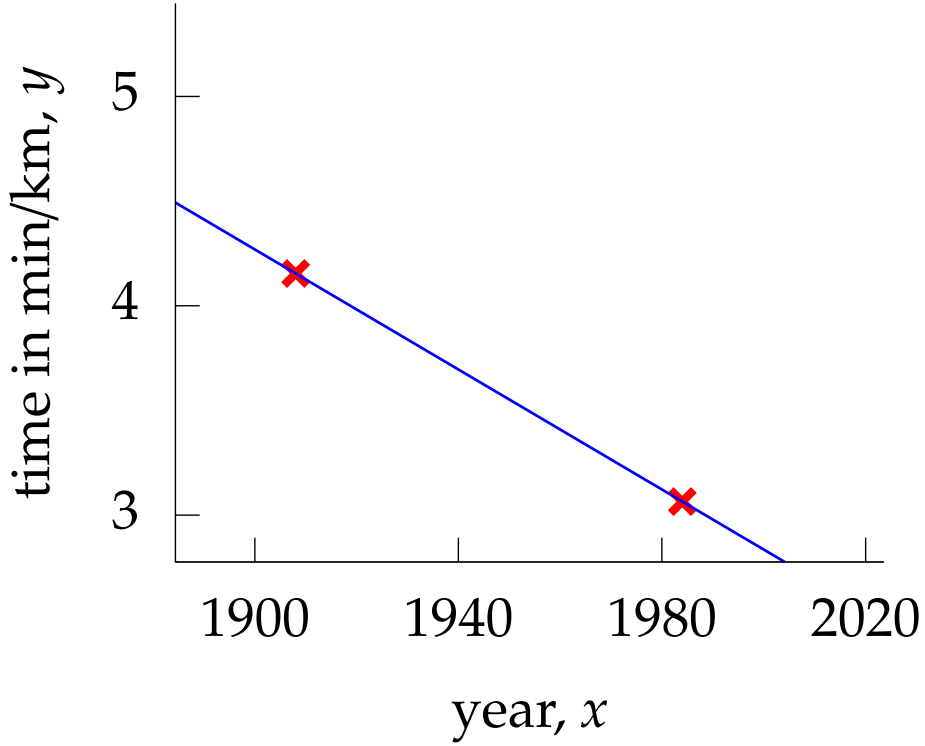
$$y = mx + c$$

- y: winning time/pace.
- x: year of Olympics.
- m: rate of improvement over time.
- c: winning time at year 0.

Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

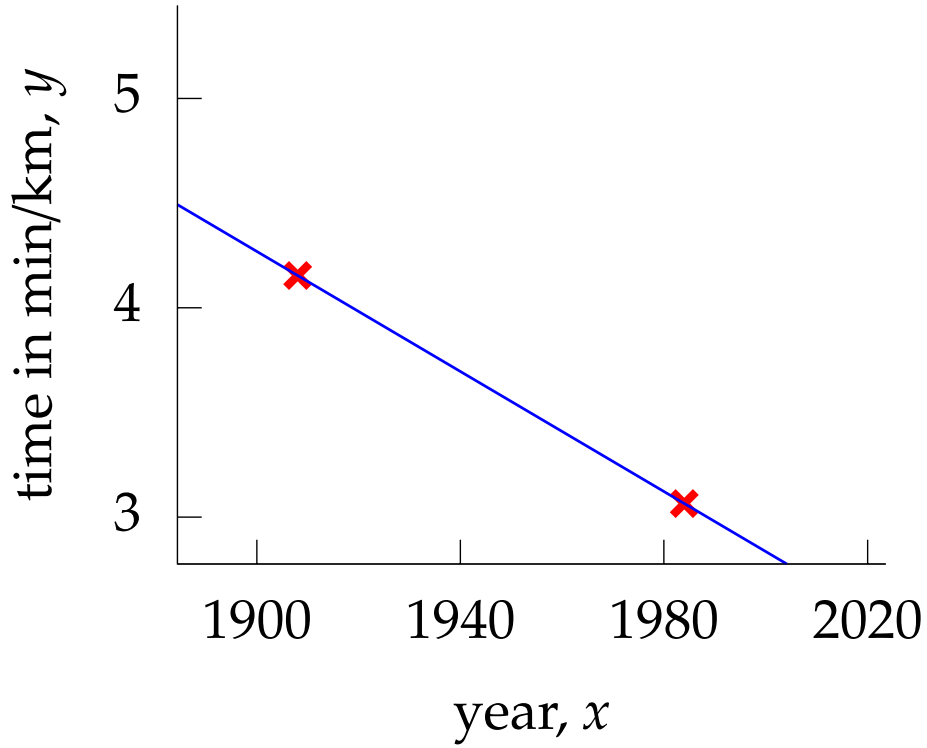
$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$



Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

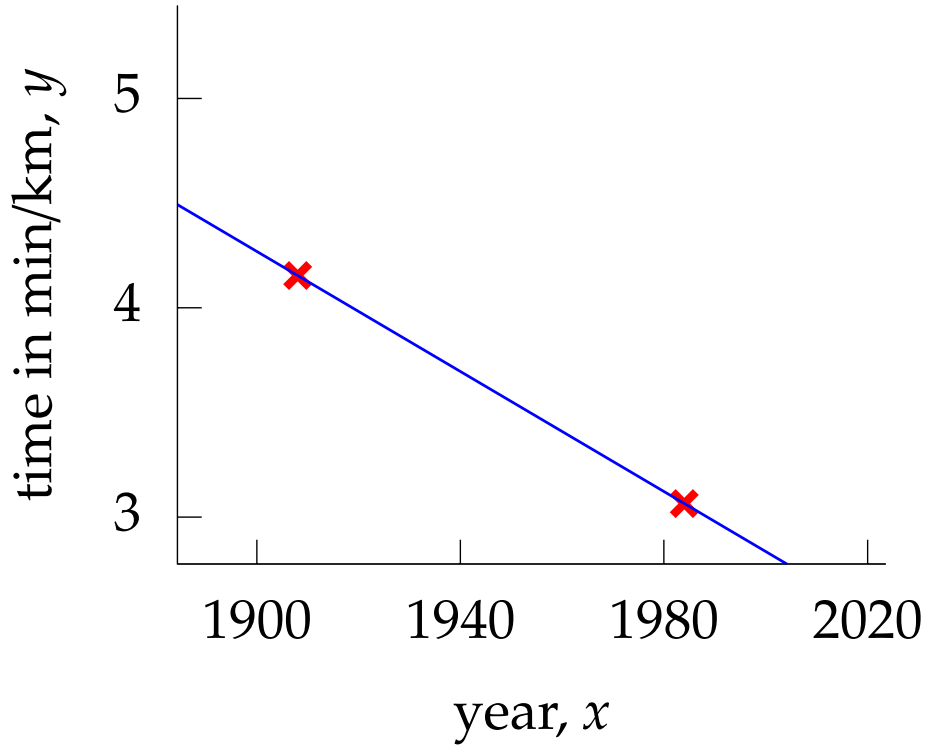
$$y_1 - y_2 = m(x_1 - x_2)$$



Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

$$\frac{y_1 - y_2}{x_1 - x_2} = m$$

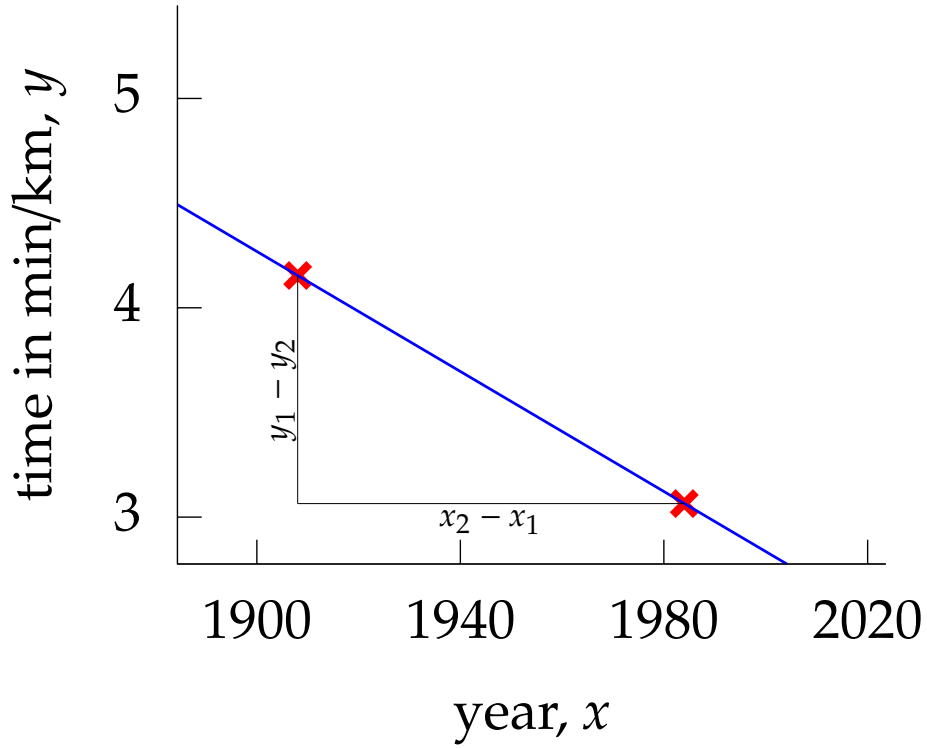


Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

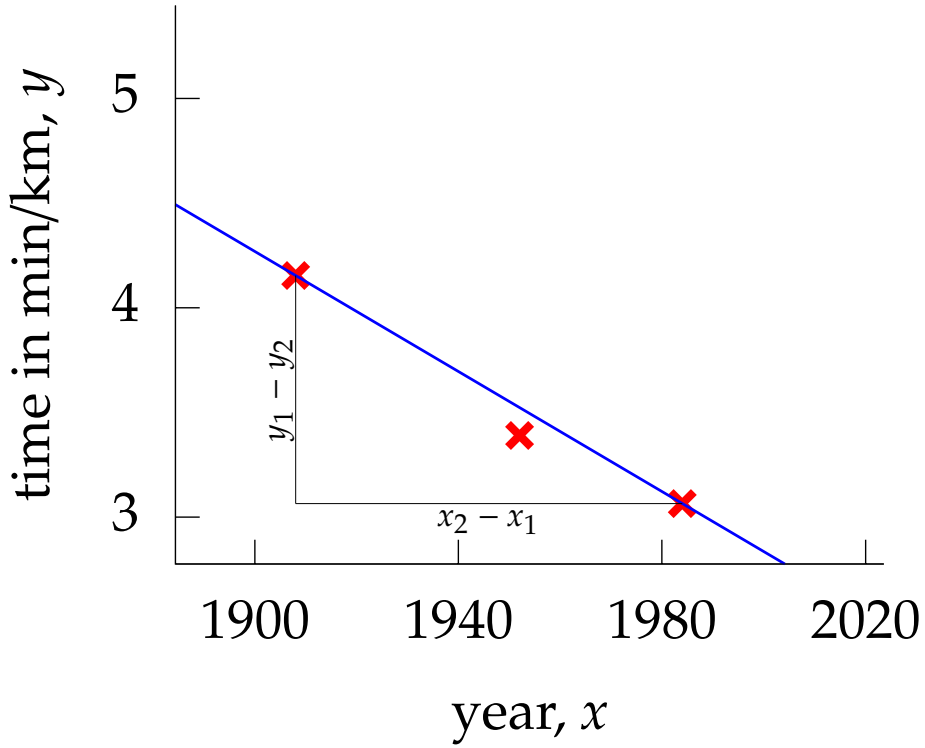
$$c = y_1 - mx_1$$



Two Simultaneous Equations

How do we deal with three simultaneous equations with only two unknowns?

$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$
$$y_3 = mx_3 + c$$



Overdetermined System

- With two unknowns and two observations:

$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$

Overdetermined System

- With two unknowns and two observations:

$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$

- Additional observation leads to *overdetermined* system.

$$y_3 = mx_3 + c$$

Overdetermined System

- With two unknowns and two observations:

$$\begin{aligned} y_1 =& mx_1 + c \\ y_2 =& mx_2 + c \end{aligned}$$

- Additional observation leads to *overdetermined* system.

$$y_3 = mx_3 + c$$

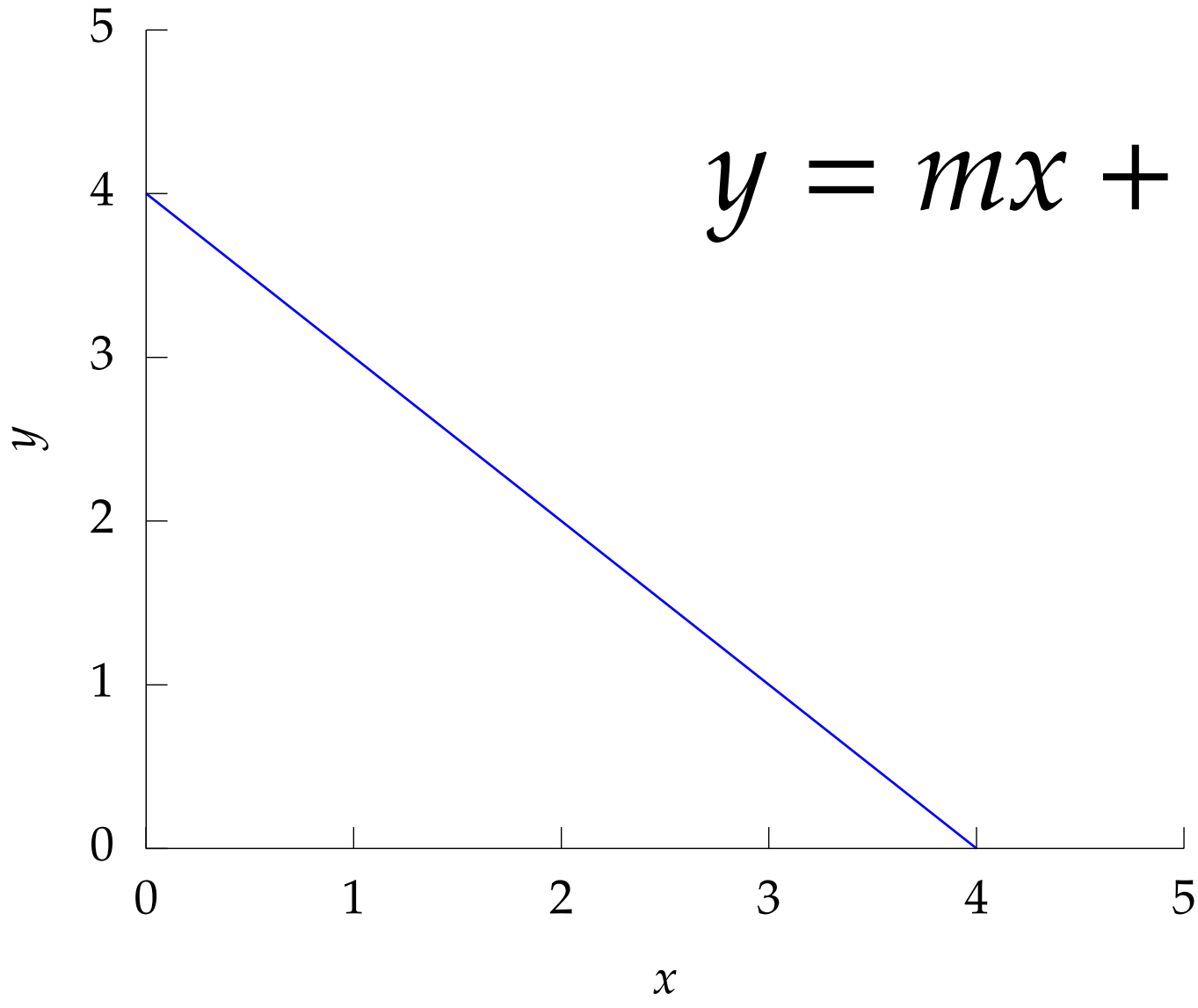
- This problem is solved through a noise model $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

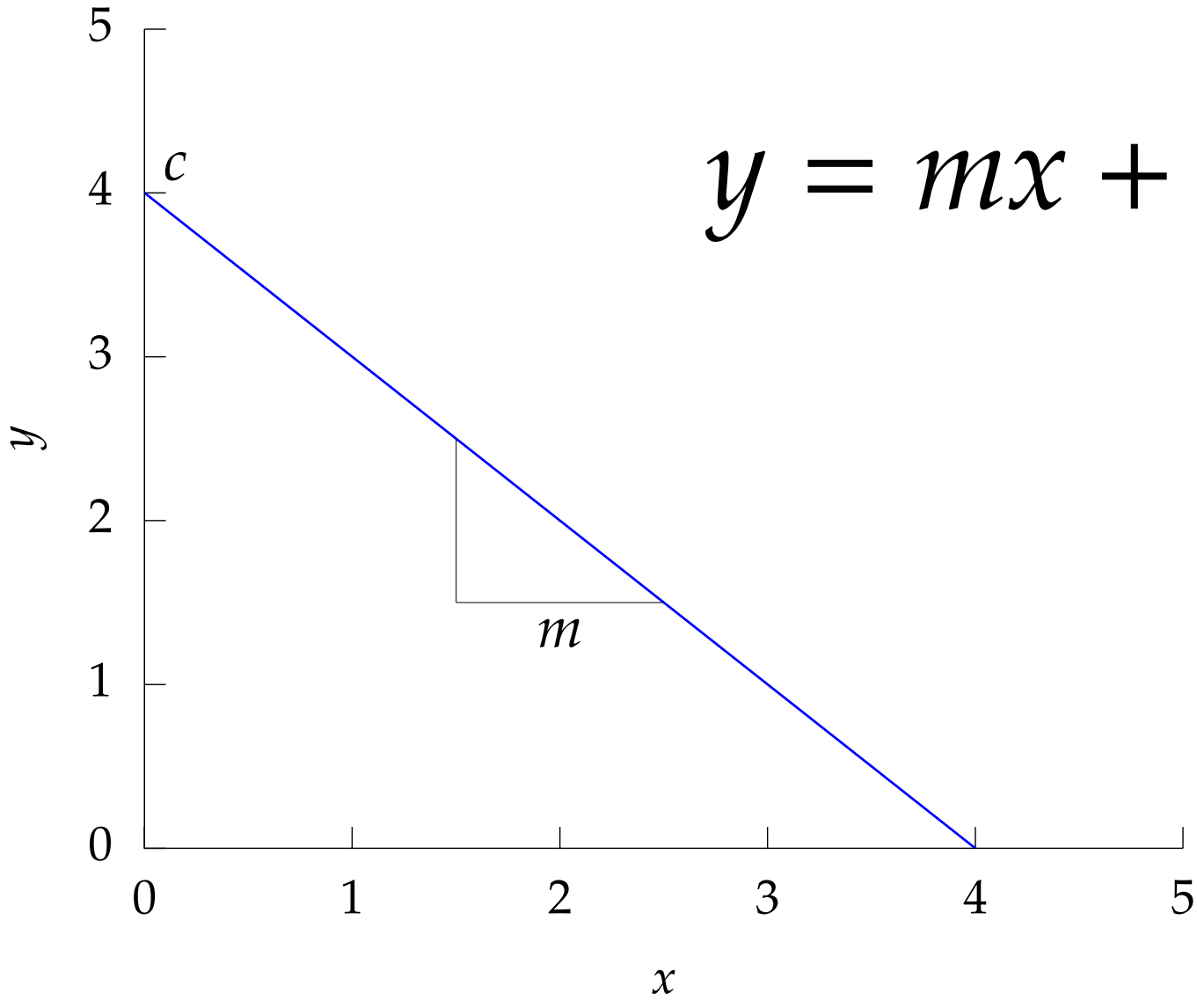
$$\begin{aligned} y_1 &= mx_1 + c + \epsilon_1 \\ y_2 &= mx_2 + c + \epsilon_2 \\ y_3 &= mx_3 + c + \epsilon_3 \end{aligned}$$

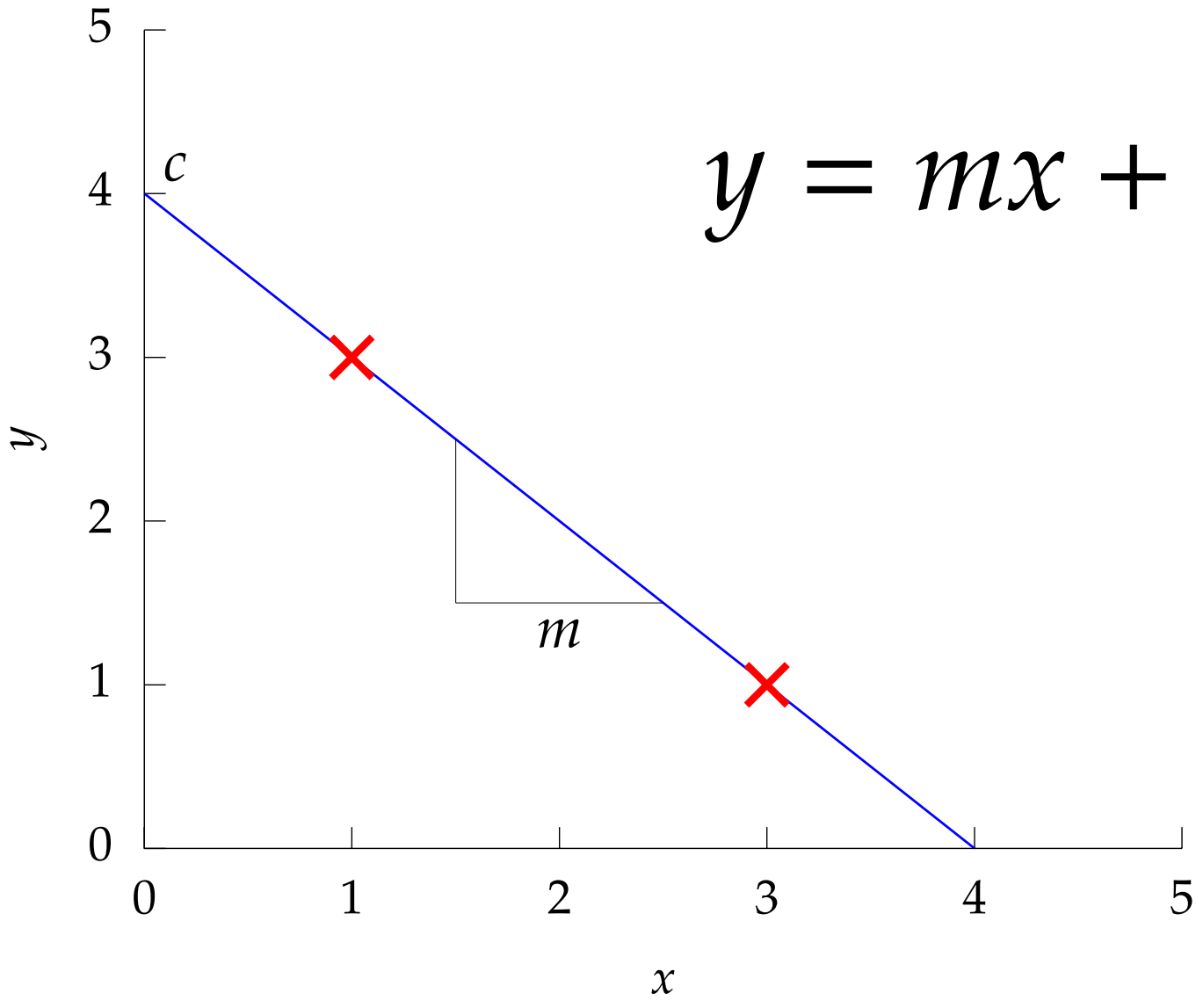
Noise Models

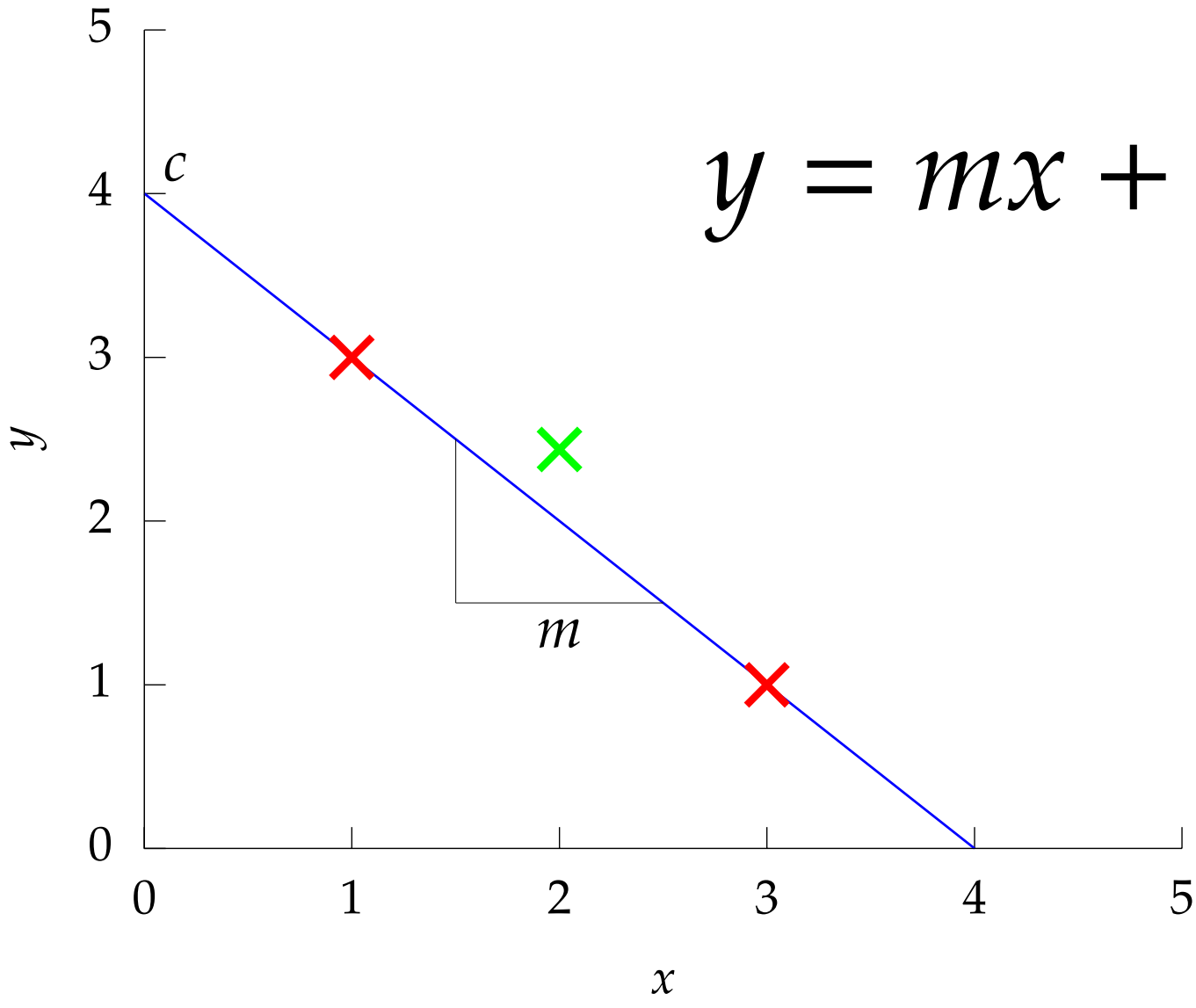
- We aren't modeling entire system.
- Noise model gives mismatch between model and data.
- Gaussian model justified by appeal to central limit theorem.
- Other models also possible (Student-t for heavy tails).
- Maximum likelihood with Gaussian noise leads to *least squares*.

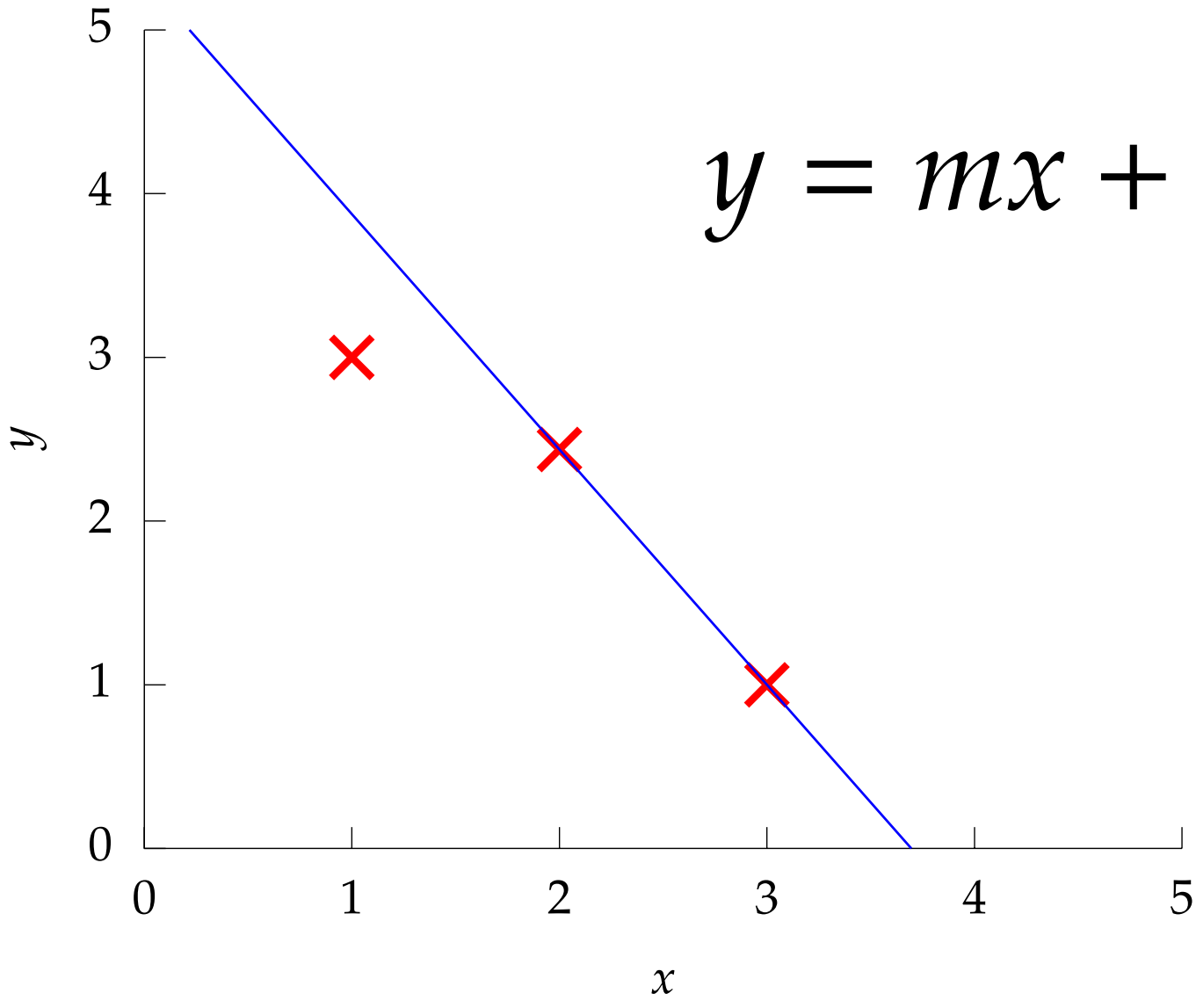
$$y = mx + c$$

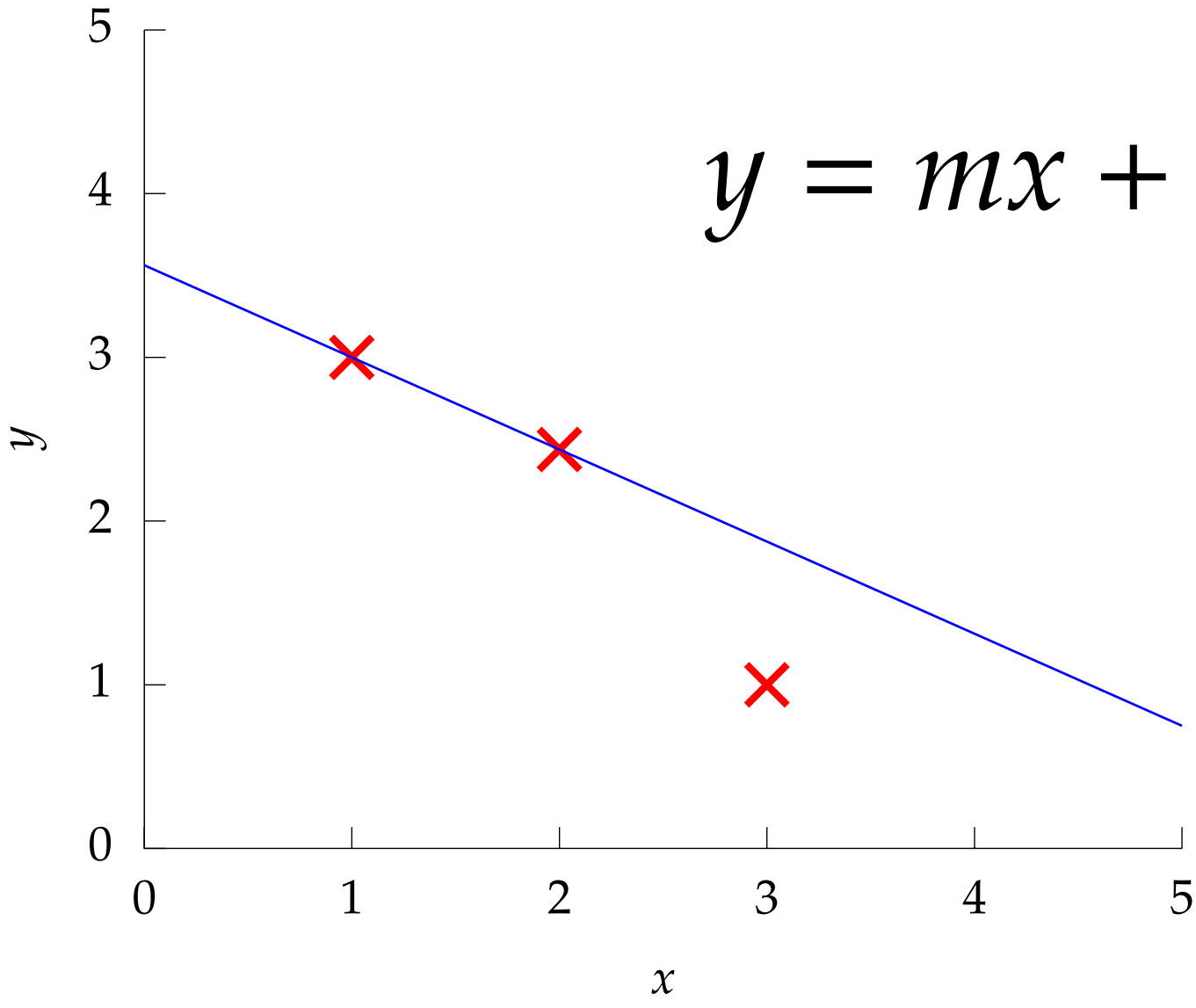

$y = mx + c$
5
4
3
2
1
0
y
0
1
2
3
4
5
x

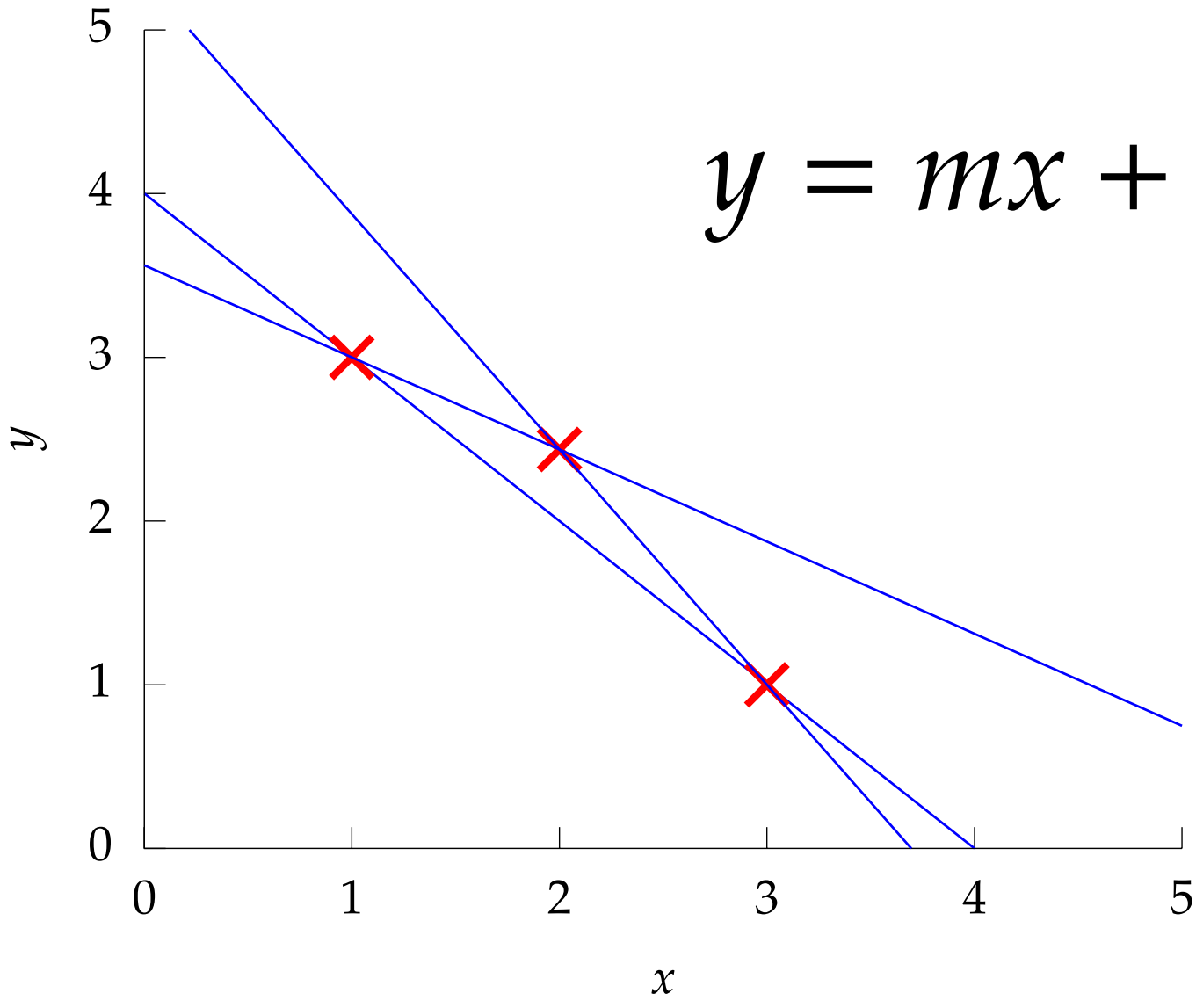

$y = mx + c$
c
m
5
4
3
2
1
0
0
1
2
3
4
5
y
x


$y = mx + c$
c
m
y
x
0
1
2
3
4
5


$y = mx + c$
c
m
y
x
0
1
2
3
4
5


$y = mx + c$
5
4
3
2
1
0
0
1
2
3
4
5
y
x


$y = mx + c$
5
4
3
2
1
0
0
1
2
3
4
5
y
x


$y = mx + c$
5
4
3
2
1
0
0
1
2
3
4
5
y
x

point 1: $x = 1, y = 3$

$$3 = m + c$$

point 2: $x = 3, y = 1$

$$1 = 3m + c$$

point 3: $x = 2, y = 2.5$

$$2.5 = 2m + c$$

point 1: $x = 1, y = 3$

$$3 = m + c + \epsilon_1$$

point 2: $x = 3, y = 1$

$$1 = 3m + c + \epsilon_2$$

point 3: $x = 2, y = 2.5$

$$2.5 = 2m + c + \epsilon_3$$

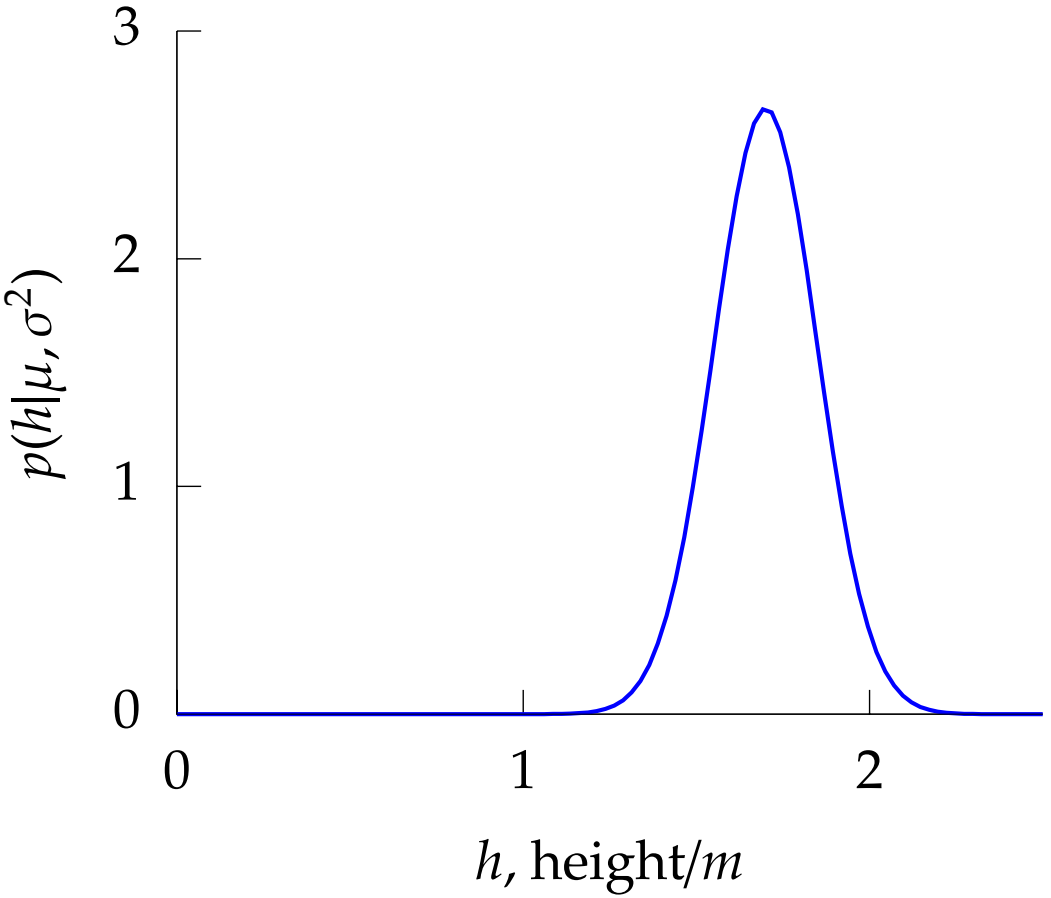
The Gaussian Density

- Perhaps the most common probability density.

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$
$$\triangleq \mathcal{N}\left(y|\mu, \sigma^2\right)$$

- The Gaussian density.

Gaussian Density



The Gaussian PDF with $\mu = 1.7$ and variance $\sigma^2 = 0.0225$. Mean shown as red line. It could represent the heights of a population of students.

Gaussian Density

$$\mathcal{N}\left(y|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

σ^2 is the variance of the density and μ is the mean.

Two Important Gaussian Properties

Sum of Gaussians

- Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

Two Important Gaussian Properties

Sum of Gaussians

- Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

And the sum is distributed as

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Two Important Gaussian Properties

Sum of Gaussians

- Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

And the sum is distributed as

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

(*Aside*: As sum increases, sum of non-Gaussian, finite variance variables is also Gaussian [central limit theorem].)

Two Important Gaussian Properties

Sum of Gaussians

- Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

And the sum is distributed as

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

(*Aside*: As sum increases, sum of non-Gaussian, finite variance variables is also Gaussian [central limit theorem].)

Two Important Gaussian Properties

Scaling a Gaussian

- Scaling a Gaussian leads to a Gaussian.

Two Important Gaussian Properties

Scaling a Gaussian

- Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

Two Important Gaussian Properties

Scaling a Gaussian

- Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

And the scaled density is distributed as

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

A Probabilistic Process

- Set the mean of Gaussian to be a function.

$$p\left(y_i|x_i\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - f\left(x_i\right)\right)^2}{2\sigma^2}\right).$$

- This gives us a 'noisy function'.
- This is known as a process.

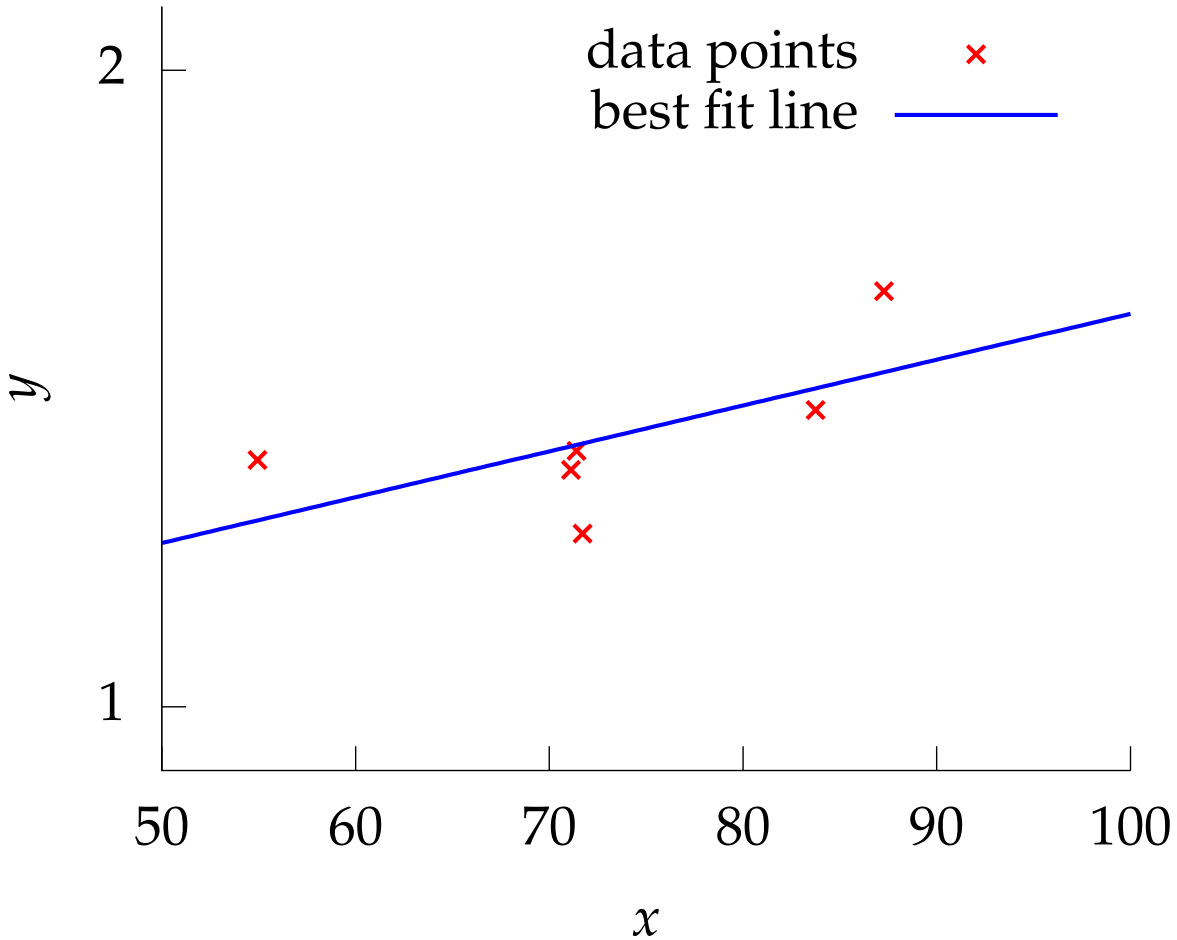
Height as a Function of Weight

- In the standard Gaussian, parametized by mean and variance.
- Make the mean a linear function of an *input*.
- This leads to a regression model.

$$y_i = f(x_i) + \epsilon_i,$$
$$\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right).$$

- Assume y_i is height and x_i is weight.

Linear Function



A linear regression between x and y.

Data Point Likelihood

- Likelihood of an individual data point

$$p\left(y_i|x_i, m, c\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - mx_i - c\right)^2}{2\sigma^2}\right).$$

- Parameters are gradient, m, offset, c of the function and noise variance σ^2.

Data Set Likelihood

- If the noise, ϵ_i is sampled independently for each data point.
- Each data point is independent (given m and c).
- For independent variables:

$$p(\mathbf{y}) = \prod_{i=1}^{n} p(y_i)$$

Data Set Likelihood

- If the noise, ϵ_i is sampled independently for each data point.
- Each data point is independent (given m and c).
- For independent variables:

$$p(\mathbf{y}|\mathbf{x}, m, c) = \prod_{i=1}^{n} p(y_i|x_i, m, c)$$

Data Set Likelihood

- If the noise, ϵ_i is sampled independently for each data point.
- Each data point is independent (given m and c).
- For independent variables:

$$p(\mathbf{y}|\mathbf{x}, m, c) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - mx_i - c)^2}{2\sigma^2}\right).$$

Data Set Likelihood

- If the noise, ϵ_i is sampled independently for each data point.
- Each data point is independent (given m and c).
- For independent variables:

$$p(\mathbf{y}|\mathbf{x}, m, c) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - mx_i - c)^2}{2\sigma^2}\right).$$

Log Likelihood Function

- Normally work with the log likelihood:

$$L(m, c, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \sum_{i=1}^{n} \frac{(y_i - mx_i - c)^2}{2\sigma^2}.$$

Consistency of Maximum Likelihood

- If data was really generated according to probability we specified.
- Correct parameters will be recovered in limit as $n \to \infty$.
- This can be proven through sample based approximations (law of large numbers) of "KL divergences".
- Mainstay of classical statistics.

Probabilistic Interpretation of the Error Function

- Probabilistic Interpretation for Error Function is Negative Log Likelihood.
- *Minimizing* error function is equivalent to *maximizing* log likelihood.
- Maximizing *log likelihood* is equivalent to maximizing the *likelihood* because log is monotonic.
- Probabilistic interpretation: Minimizing error function is equivalent to maximum likelihood with respect to parameters.

Error Function

- Negative log likelihood is the error function leading to an error function

$$E(m, c, \sigma^2) = \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - mx_i - c)^2 .$$

- Learning proceeds by minimizing this error function for the data set provided.

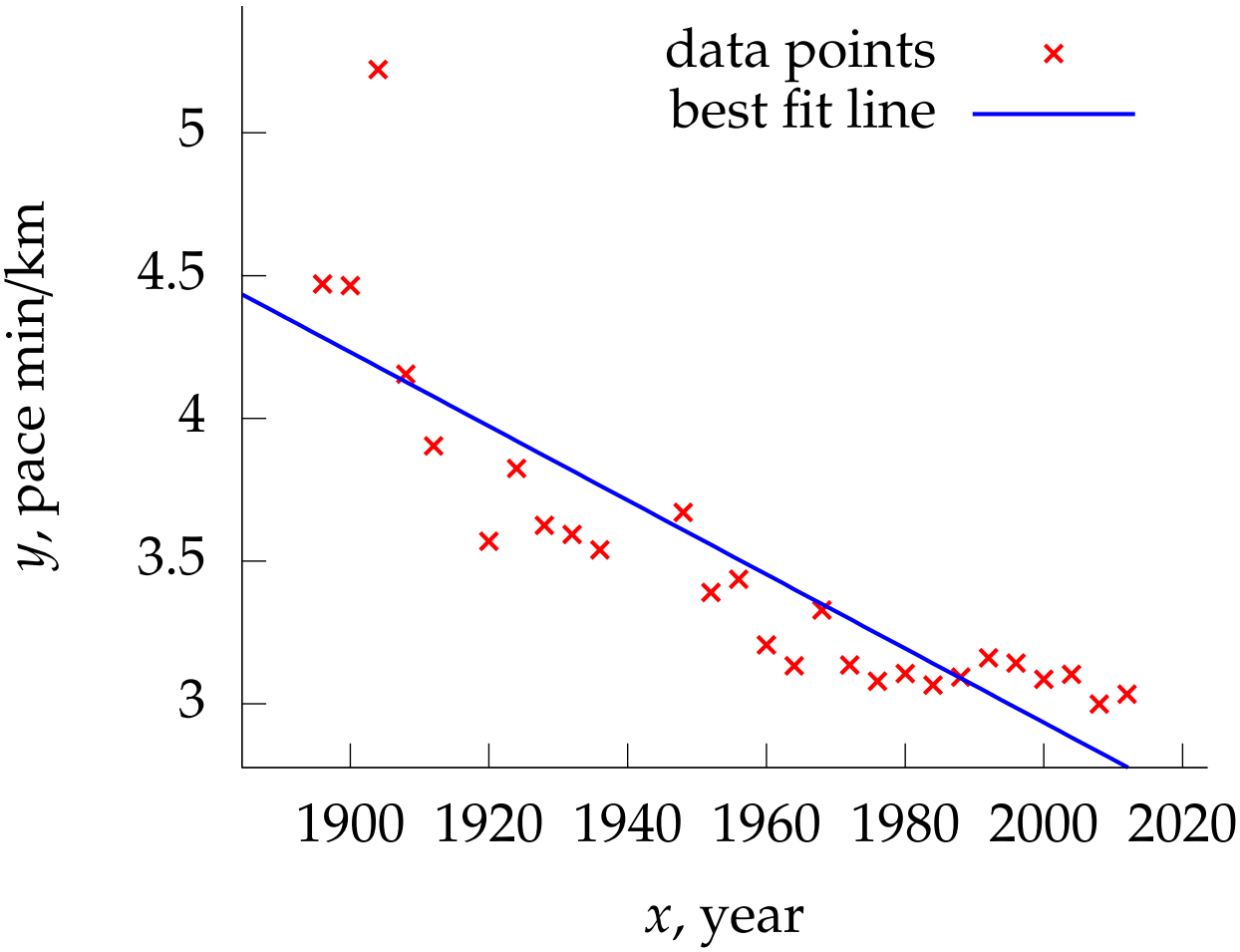
Connection: Sum of Squares Error

- Ignoring terms which don't depend on m and c gives

$$E(m, c) \propto \sum_{i=1}^{n} (y_i - f(x_i))^2$$

  where $f(x_i) = mx_i + c$.
- This is known as the *sum of squares* error function.
- Commonly used and is closely associated with the Gaussian likelihood.

Linear Function



Linear regression for Male Olympics Marathon Gold Medal times.

Reading

- Section 1.2.5 of Bishop up to equation 1.65.
- Section 1.1-1.2 of Rogers and Girolami for fitting linear models.

Multi-dimensional Inputs

- Multivariate functions involve more than one input.
- Height might be a function of weight and gender.
- There could be other contributory factors.
- Place these factors in a feature vector $\mathbf{x}_i$.
- Linear function is now defined as

$$f(\mathbf{x}_i) = \sum_{j=1}^{q} w_j x_{i,j} + c$$

Vector Notation

mo

- Write in vector notation,

$$f(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i + c$$

- Can absorb c into $\mathbf{w}$ by assuming extra input x_0 which is always 1.

$$f(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i$$

Log Likelihood for Multivariate Regression

- The likelihood of a single data point is

$$p\left(y_i|x_i\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \mathbf{w}^\top \mathbf{x}_i\right)^2}{2\sigma^2}\right).$$

- Leading to a log likelihood for the data set of

$$L(\mathbf{w}, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{\sum_{i=1}^{n}\left(y_i - \mathbf{w}^\top \mathbf{x}_i\right)^2}{2\sigma^2}.$$

- And a corresponding error function of

$$E(\mathbf{w}, \sigma^2) = \frac{n}{2}\log\sigma^2 + \frac{\sum_{i=1}^{n}\left(y_i - \mathbf{w}^\top \mathbf{x}_i\right)^2}{2\sigma^2}.$$

Expand the Brackets

$$
\begin{aligned}
E(\mathbf{w}, \sigma^2) =& \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2 - \frac{1}{\sigma^2} \sum_{i=1}^{n} y_i \mathbf{w}^\top \mathbf{x}_i \\
&+ \frac{1}{2\sigma^2} \sum_{i=1}^{n} \mathbf{w}^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{w} + \text{const.} \\
=& \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2 - \frac{1}{\sigma^2} \mathbf{w}^\top \sum_{i=1}^{n} \mathbf{x}_i y_i \\
&+ \frac{1}{2\sigma^2} \mathbf{w}^\top \left[\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right] \mathbf{w} + \text{const.}
\end{aligned}
$$

Multivariate Derivatives

- We will need some multivariate calculus.
- For now some simple multivariate differentiation:

$$\frac{\mathrm{d}\mathbf{a}^\top\mathbf{w}}{\mathrm{d}\mathbf{w}} = \mathbf{a}$$

and

$$\frac{\mathrm{d}\mathbf{w}^\top\mathbf{A}\mathbf{w}}{\mathrm{d}\mathbf{w}} = \left(\mathbf{A} + \mathbf{A}^\top\right)\mathbf{w}$$

or if $\mathbf{A}$ is symmetric (*i.e.* $\mathbf{A} = \mathbf{A}^\top$)

$$\frac{\mathrm{d}\mathbf{w}^\top\mathbf{A}\mathbf{w}}{\mathrm{d}\mathbf{w}} = 2\mathbf{A}\mathbf{w}.$$

Differentiate

Differentiating with respect to the vector $\mathbf{w}$ we obtain

$$\frac{\partial L(\mathbf{w}, \beta)}{\partial \mathbf{w}} = \beta \sum_{i=1}^{n} \mathbf{x}_i y_i - \beta \left[\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right] \mathbf{w}$$

Leading to

$$\mathbf{w}^* = \left[\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right]^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i,$$

Rewrite in matrix notation:

$$\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top = \mathbf{X}^\top \mathbf{X}$$

$$\sum_{i=1}^{n} \mathbf{x}_i y_i = \mathbf{X}^\top \mathbf{y}$$

Update Equations

- Update for $\mathbf{w}^*$.

$$\mathbf{w}^* = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

- The equation for ${\sigma^2}^*$ may also be found

$${\sigma^2}^* = \frac{\sum_{i=1}^{n} \left(y_i - {\mathbf{w}^*}^\top \mathbf{x}_i\right)^2}{n}.$$

Reading

- Section 1.3 of Rogers and Girolami for Matrix & Vector Review.

Basis Functions

Nonlinear Regression

- Problem with Linear Regression—$\mathbf{x}$ may not be linearly related to $\mathbf{y}$.
- Potential solution: create a feature space: define $\phi(\mathbf{x})$ where $\phi(\cdot)$ is a nonlinear function of $\mathbf{x}$.
- Model for target is a linear combination of these nonlinear functions

$$f(\mathbf{x}) = \sum_{j=1}^{K} w_j \phi_j(\mathbf{x}) \tag{1}$$

Quadratic Basis

- Basis functions can be global. E.g. quadratic basis:

$$[1, x, x^2]$$

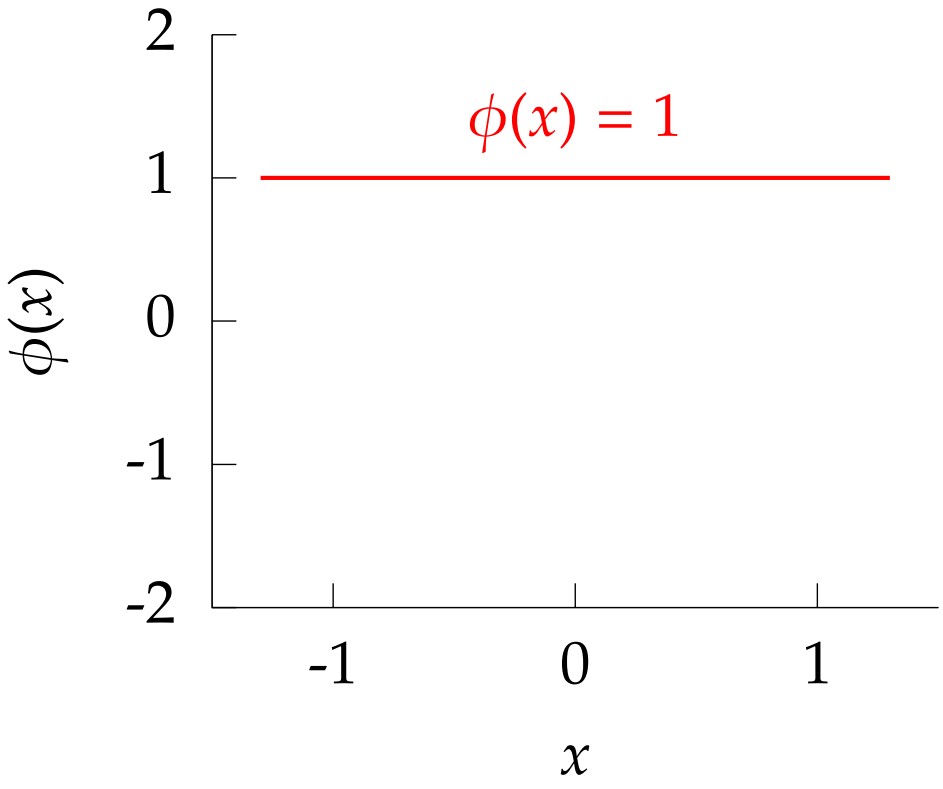


Figure: A quadratic basis.

Quadratic Basis

- Basis functions can be global. E.g. quadratic basis:

$$[1, x, x^2]$$

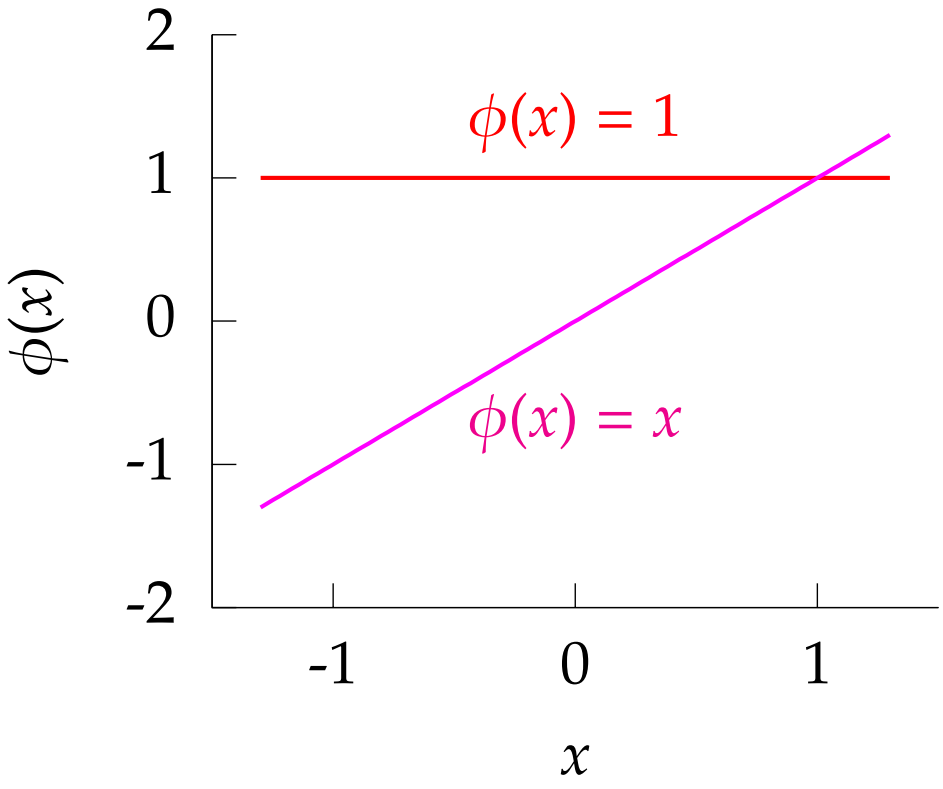


Figure: A quadratic basis.

Quadratic Basis

- Basis functions can be global. E.g. quadratic basis:

$$[1, x, x^2]$$

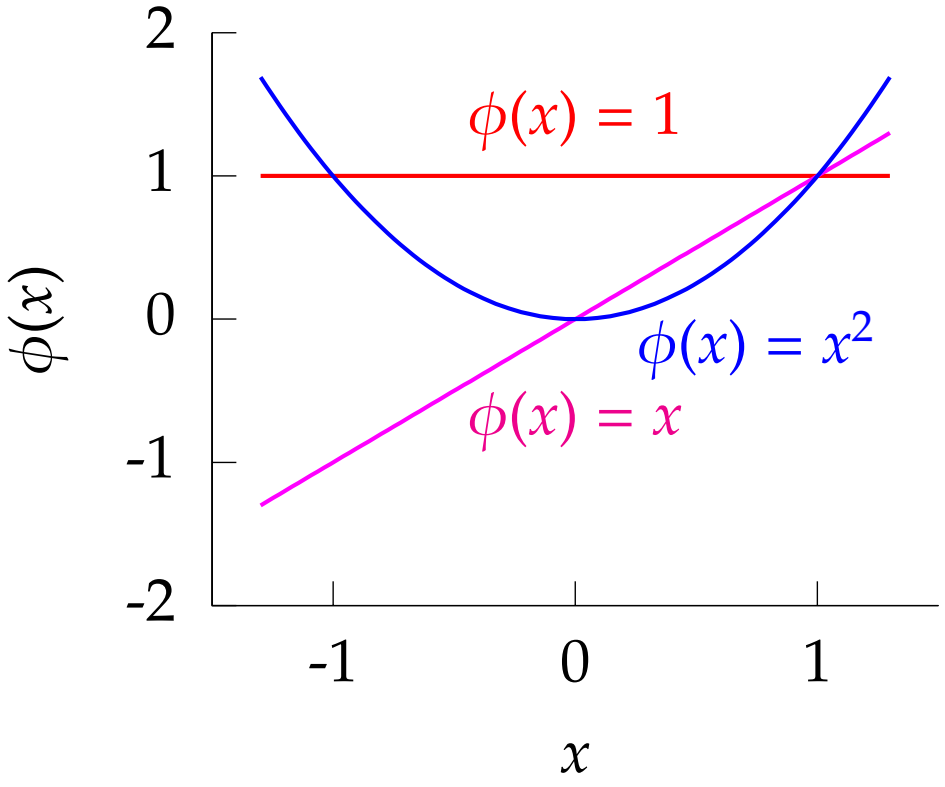


Figure: A quadratic basis.

Functions Derived from Quadratic Basis

$$f(x) = w_1 + w_2 x + w_3 x^2$$

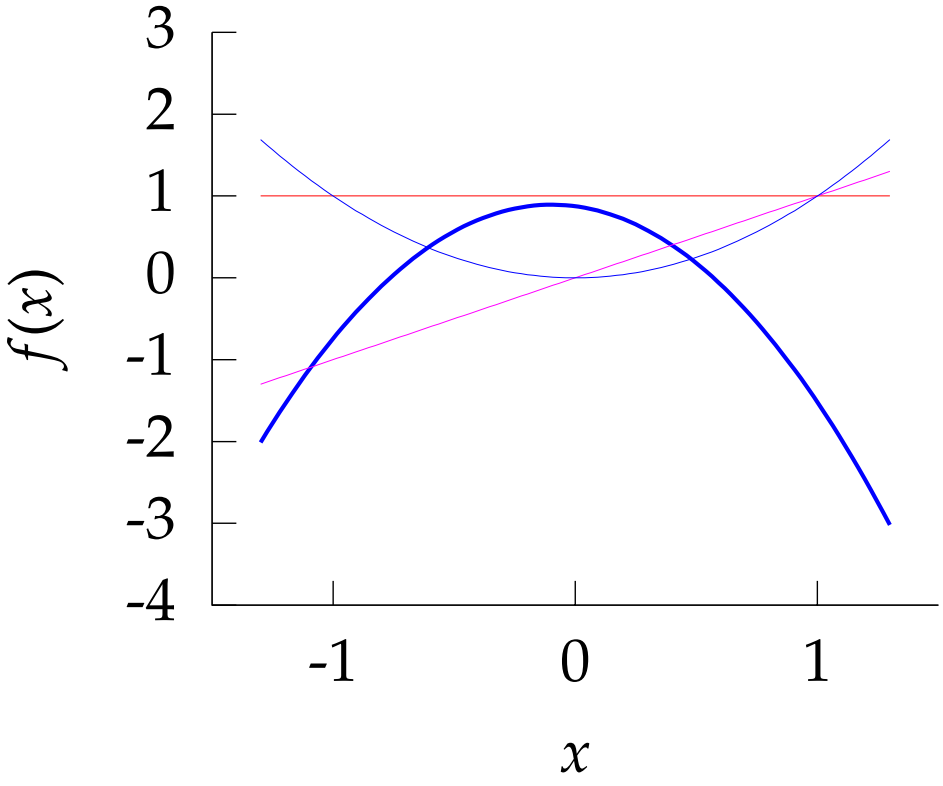


Figure: Function from quadratic basis with weights $w_1 = 0.87466$, $w_2 = -0.38835$, $w_3 = -2.0058$.

Functions Derived from Quadratic Basis

$$f(x) = w_1 + w_2 x + w_3 x^2$$

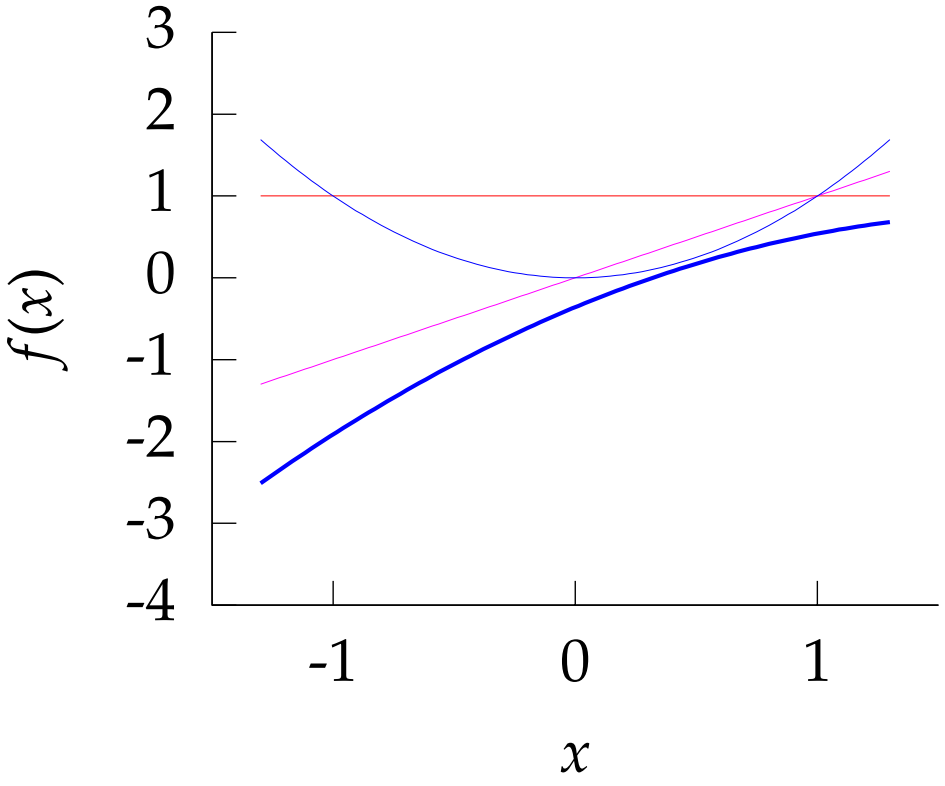


Figure: Function from quadratic basis with weights $w_1 = -0.35908$, $w_2 = 1.2274$, $w_3 = -0.32825$.

Functions Derived from Quadratic Basis

$$f(x) = w_1 + w_2 x + w_3 x^2$$

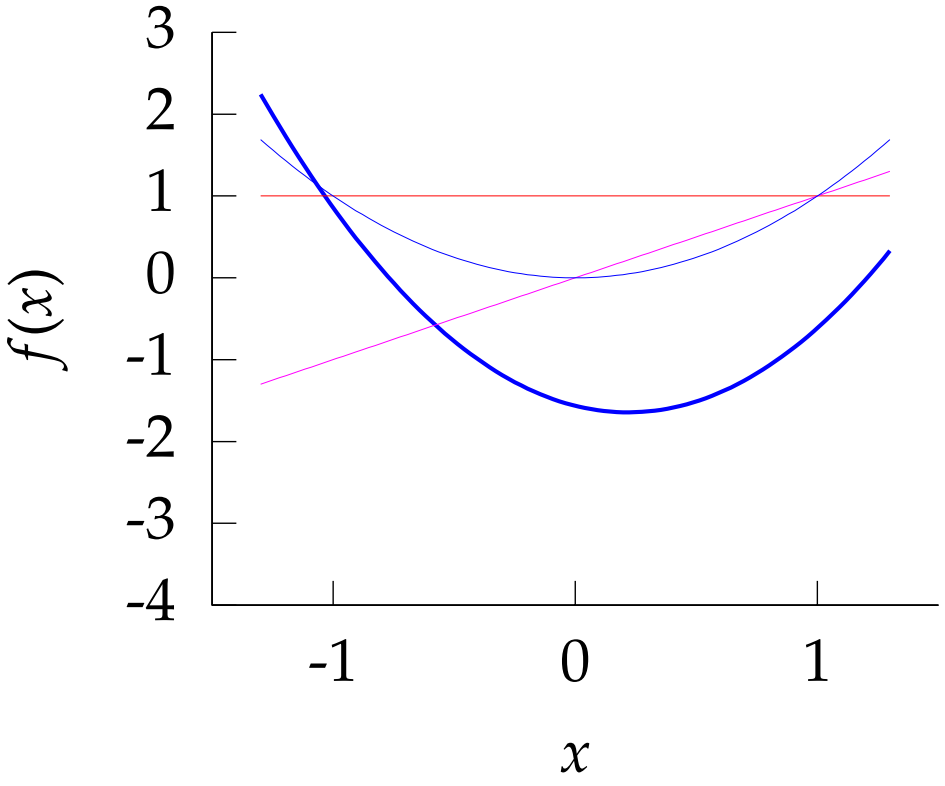


Figure: Function from quadratic basis with weights $w_1 = -1.5638$, $w_2 = -0.73577$, $w_3 = 1.6861$.

Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis
  $\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{\ell^2}\right)$

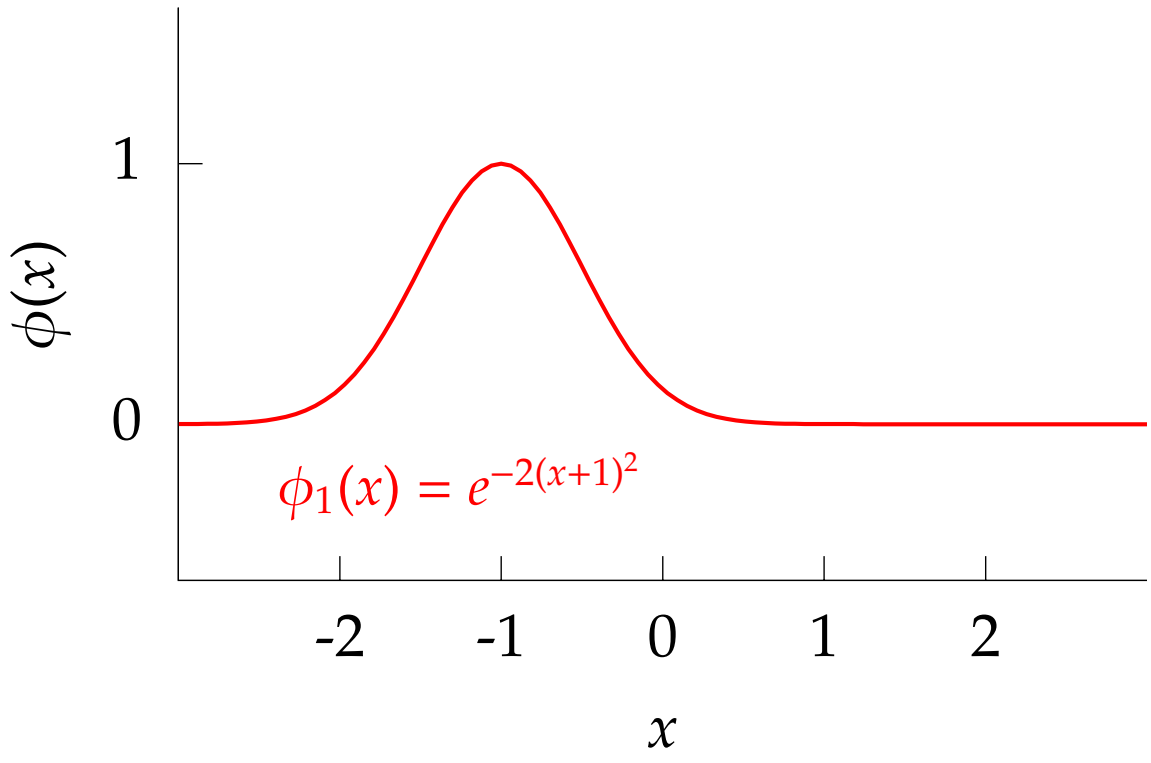


Figure: Radial basis functions.

Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis
  $\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{\ell^2}\right)$

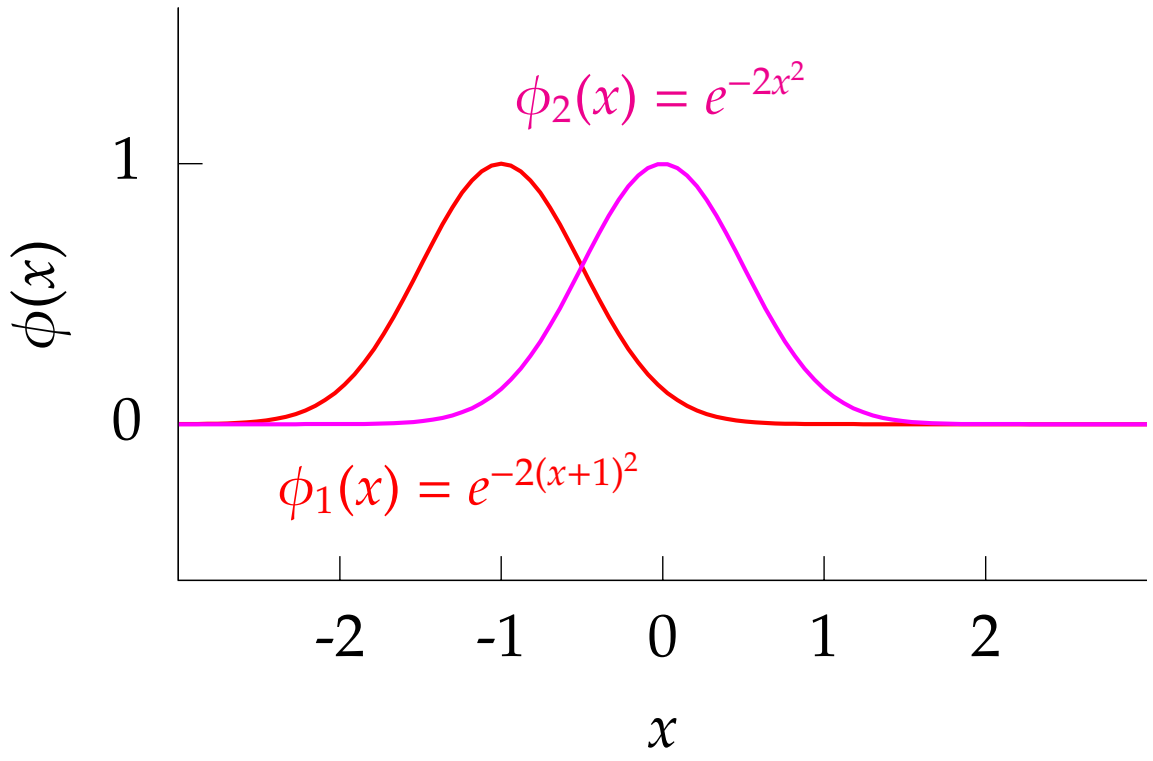


Figure: Radial basis functions.

Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis
  $\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{\ell^2}\right)$

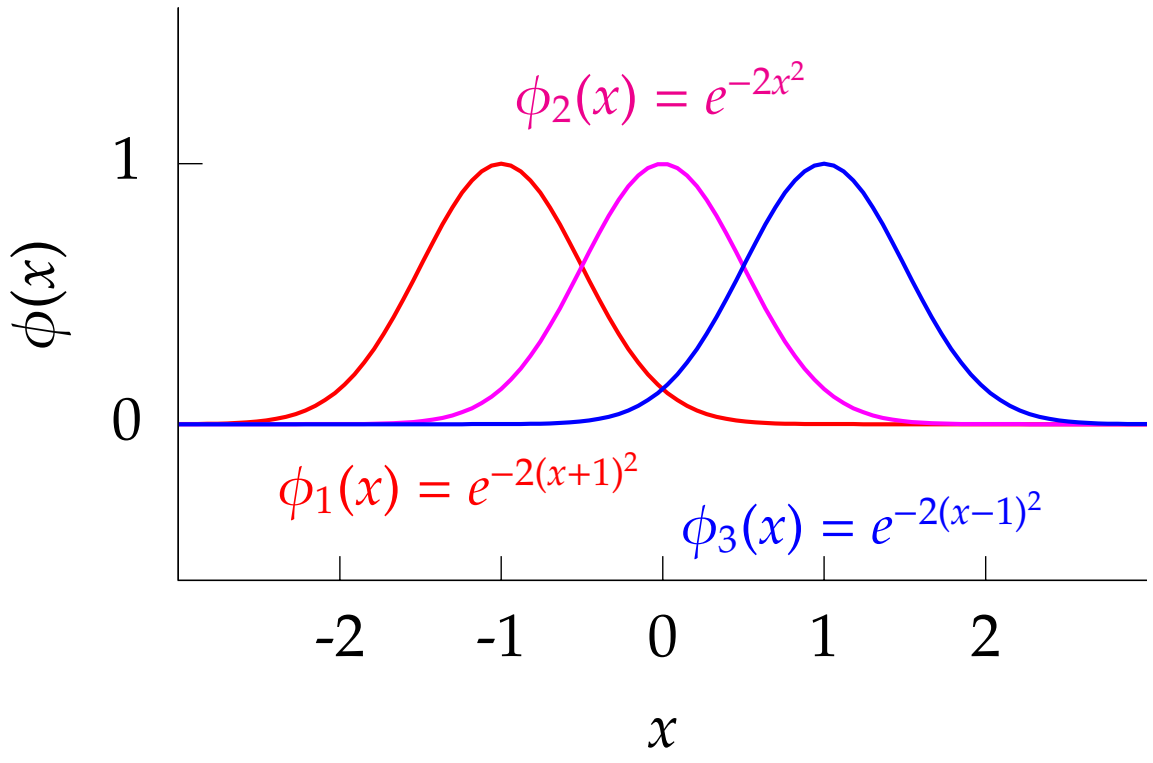


Figure: Radial basis functions.

Functions Derived from Radial Basis

$$f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

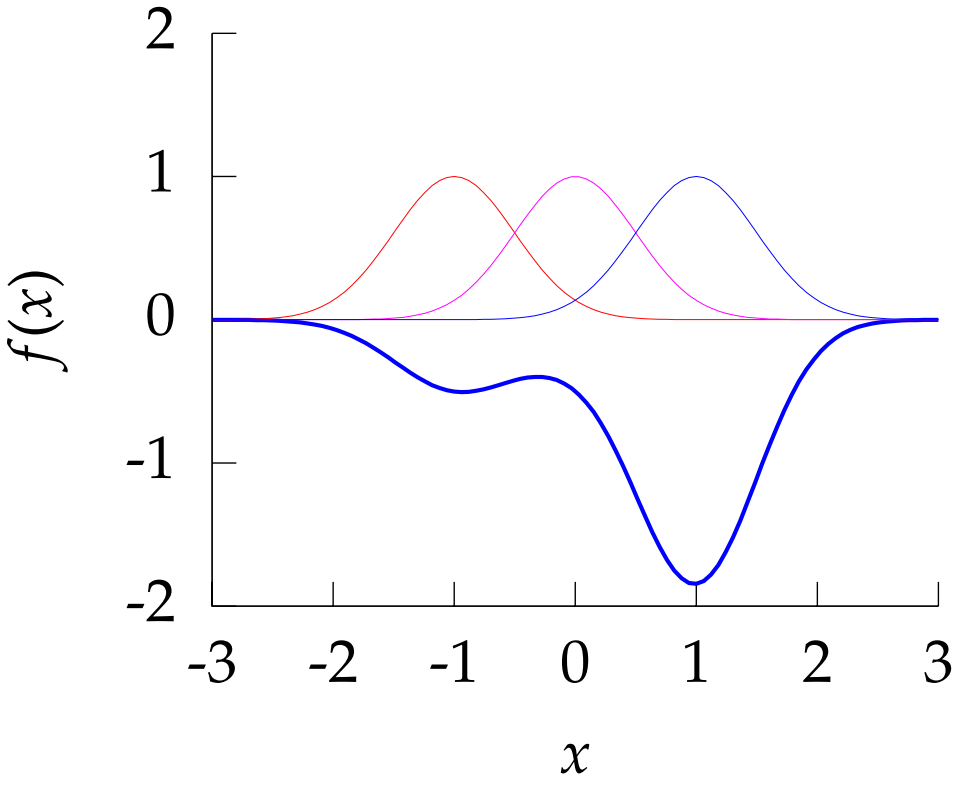


Figure: Function from radial basis with weights $w_1 = -0.47518$, $w_2 = -0.18924$, $w_3 = -1.8183$.

Functions Derived from Radial Basis

$$f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

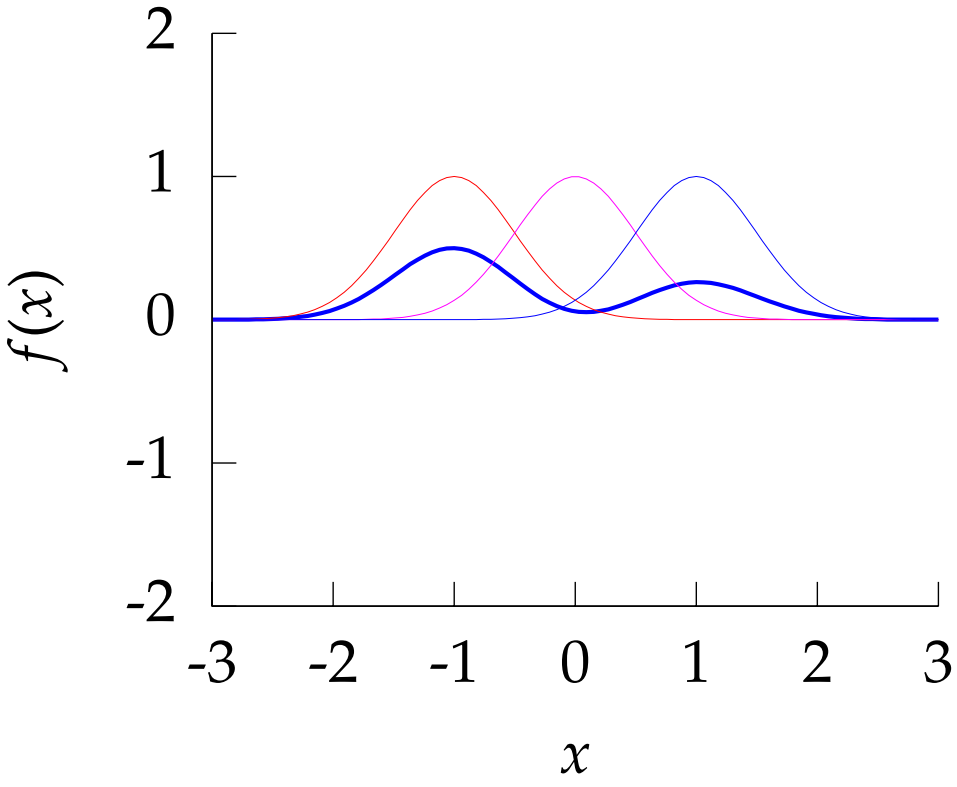


Figure: Function from radial basis with weights $w_1 = 0.50596$, $w_2 = -0.046315$, $w_3 = 0.26813$.

Functions Derived from Radial Basis

$$f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

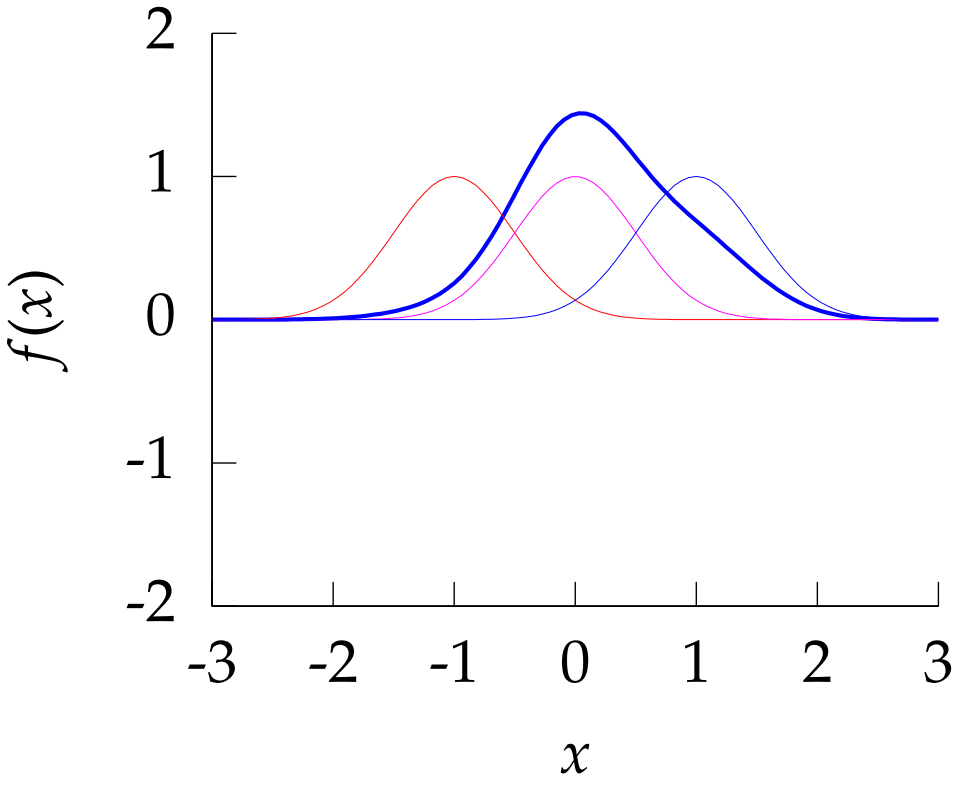
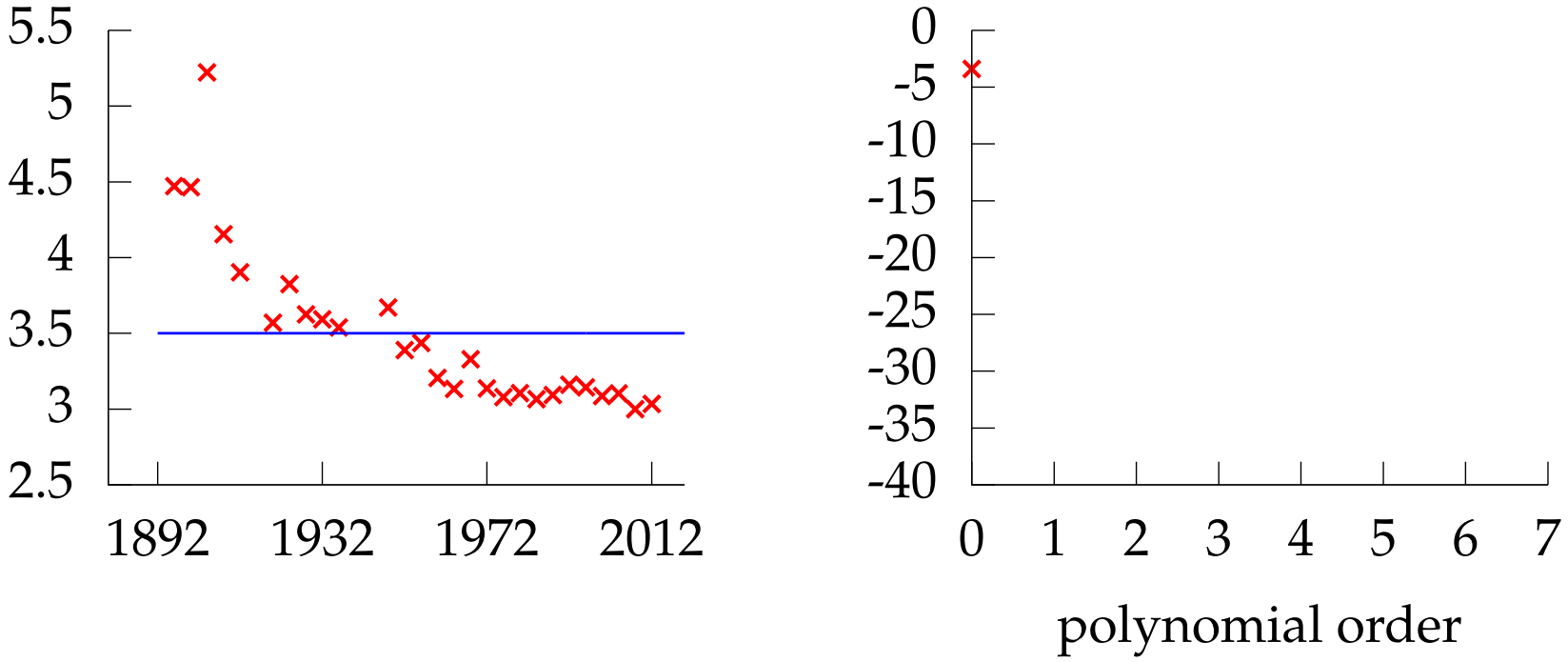


Figure: Function from radial basis with weights $w_1 = 0.07179$, $w_2 = 1.3591$, $w_3 = 0.50604$.

Reading

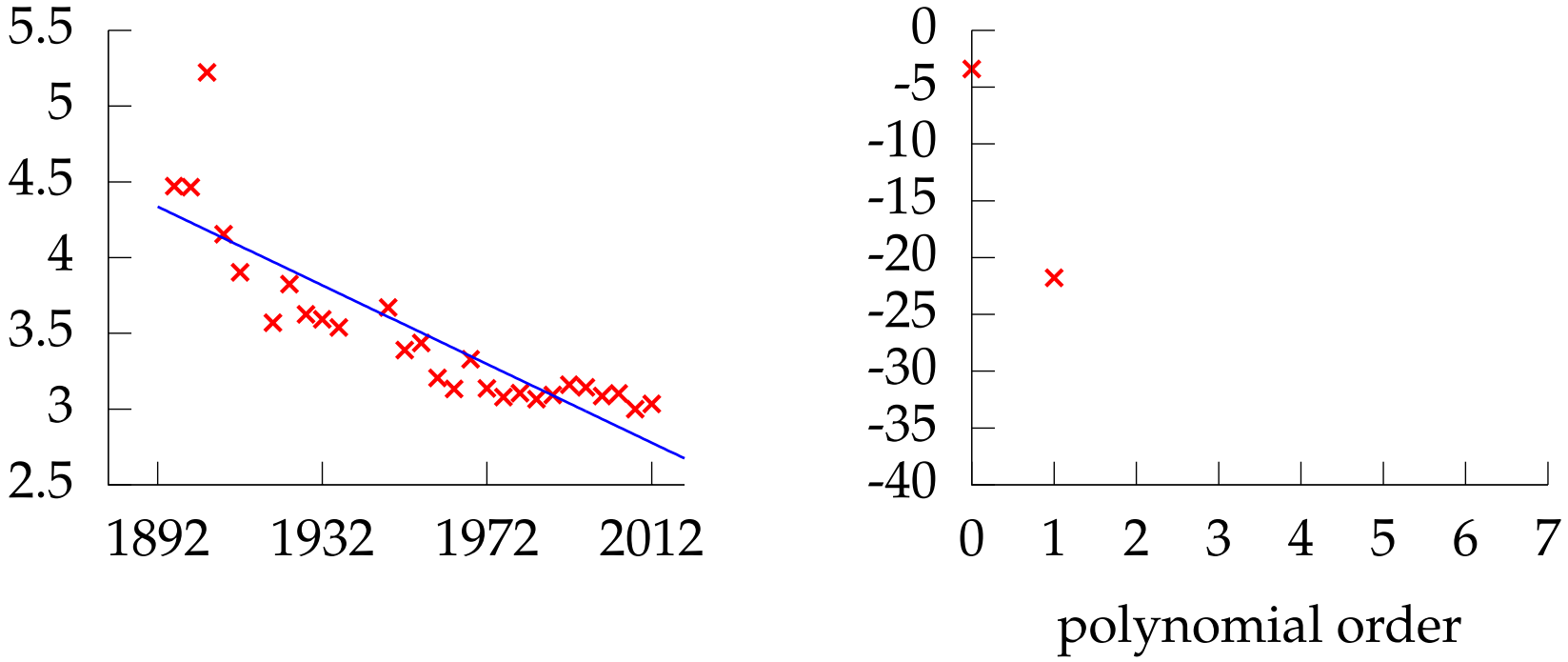
- Chapter 1, pg 1-6 of Bishop.
- Section 1.4 of Rogers and Girolami.
- Chapter 3, Section 3.1 of Bishop up to pg 143.

Polynomial Fits to Olympics Data



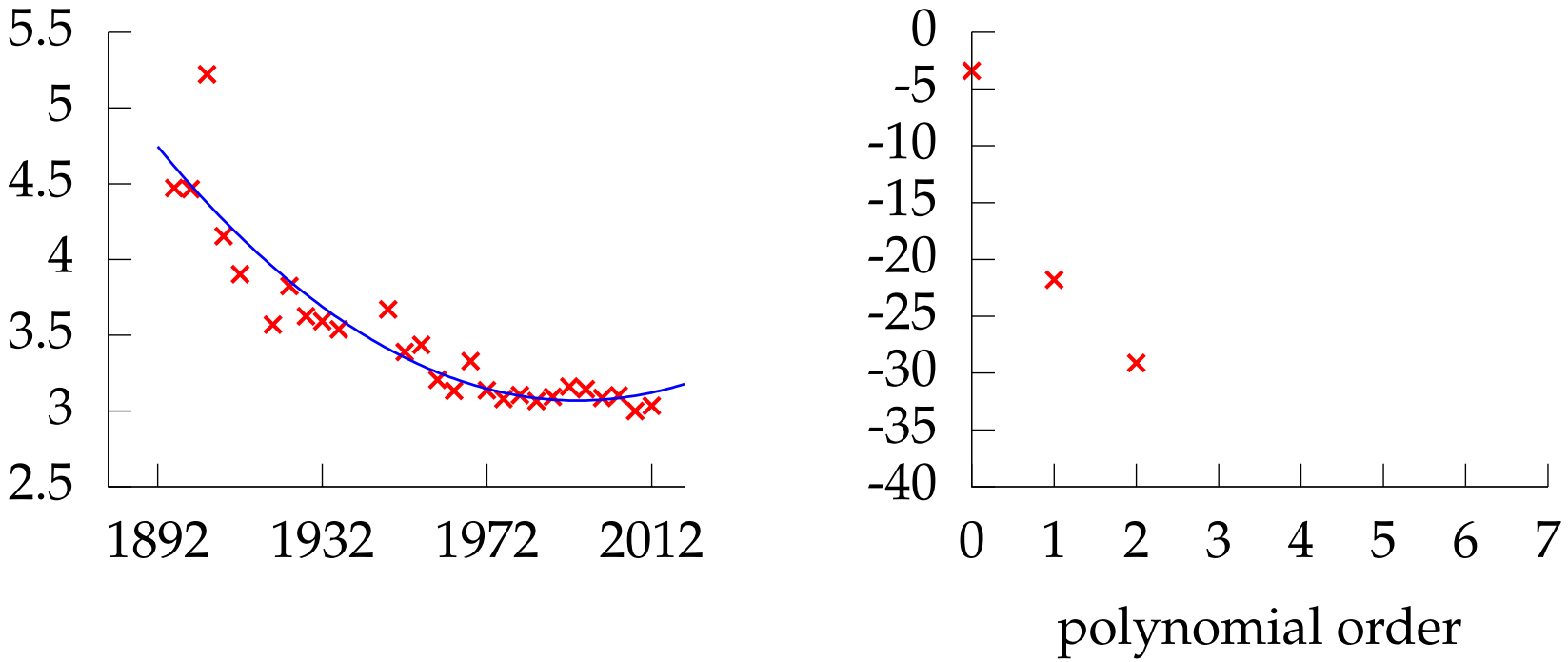
Left: fit to data, *Right*: model error. Polynomial order 0, model error -3.3989, $\sigma^2 = 0.286$, $\sigma = 0.535$.

Polynomial Fits to Olympics Data



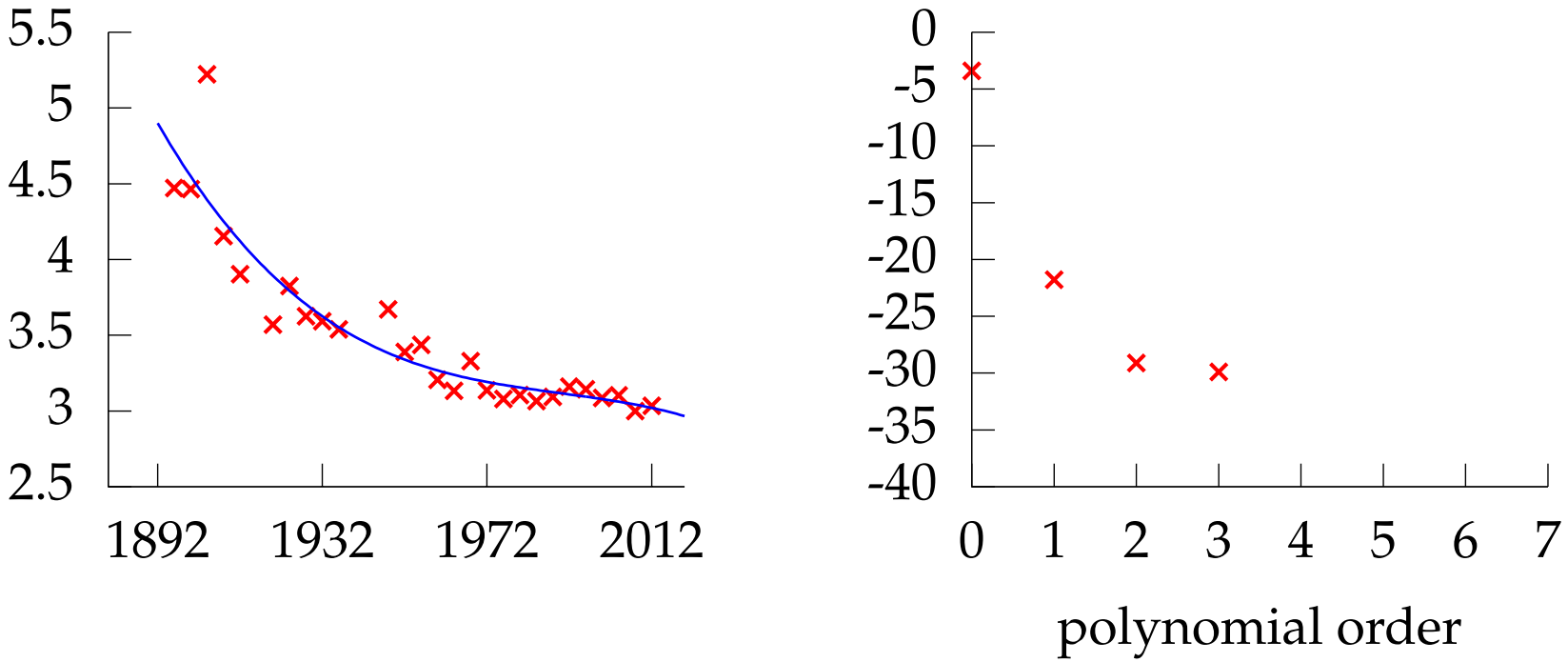
Left: fit to data, *Right*: model error. Polynomial order 1, model error -21.772, $\sigma^2 = 0.0733$, $\sigma = 0.271$.

Polynomial Fits to Olympics Data



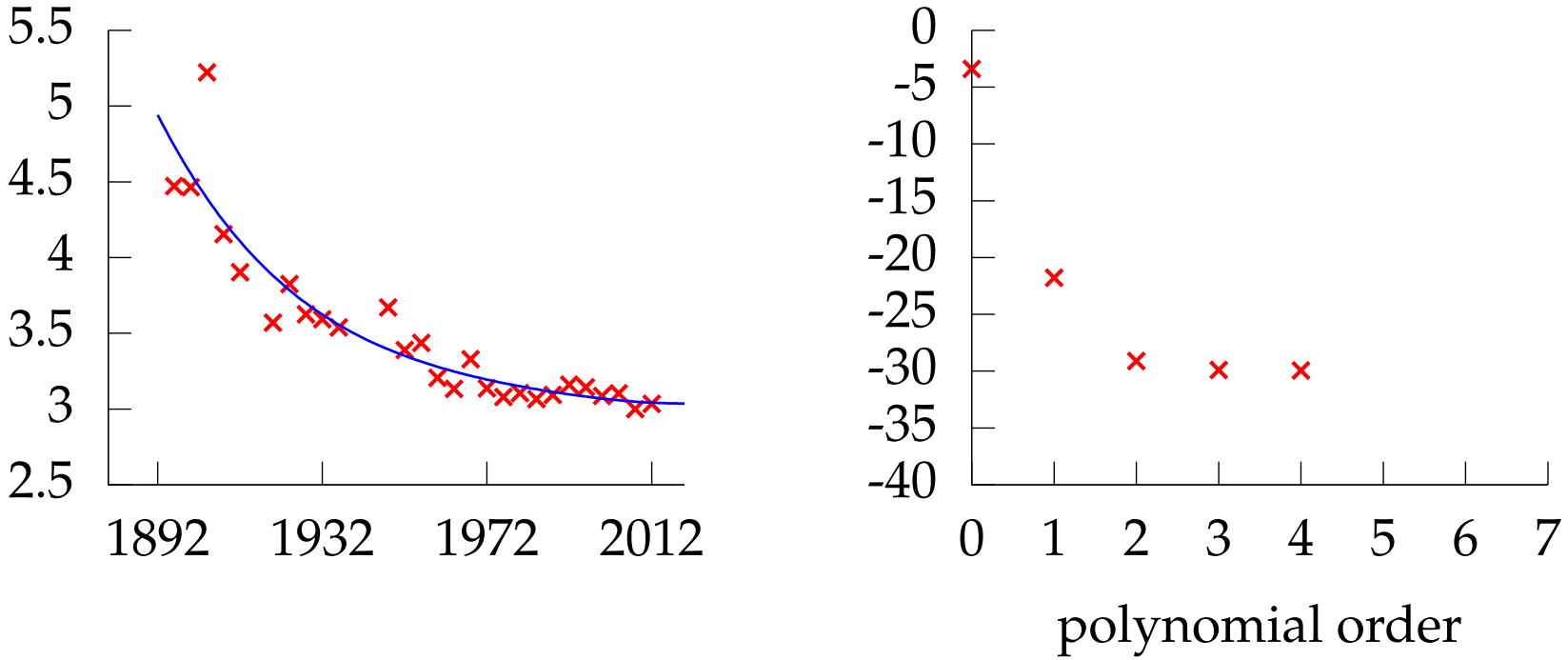
Left: fit to data, *Right*: model error. Polynomial order 2, model error -29.101, $\sigma^2 = 0.0426$, $\sigma = 0.206$.

Polynomial Fits to Olympics Data



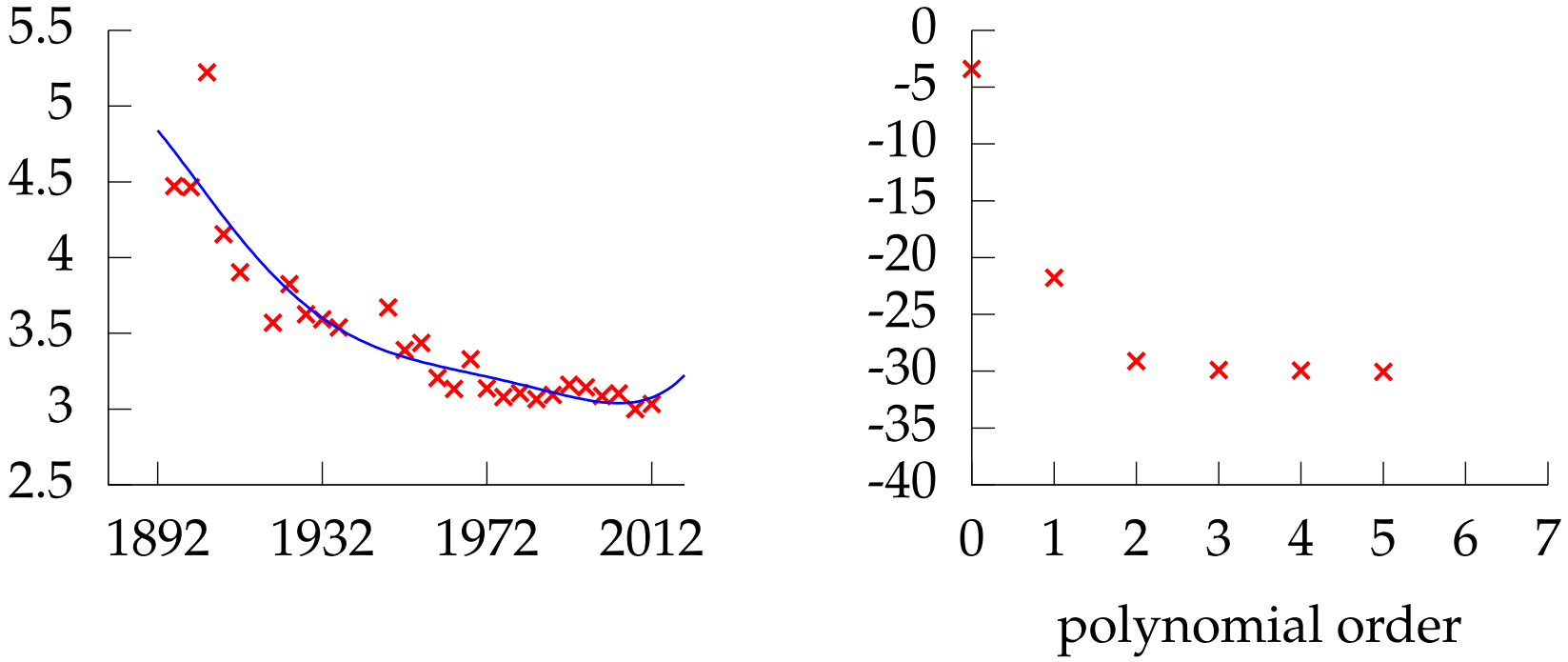
Left: fit to data, *Right*: model error. Polynomial order 3, model error -29.907, $\sigma^2 = 0.0401$, $\sigma = 0.200$.

Polynomial Fits to Olympics Data



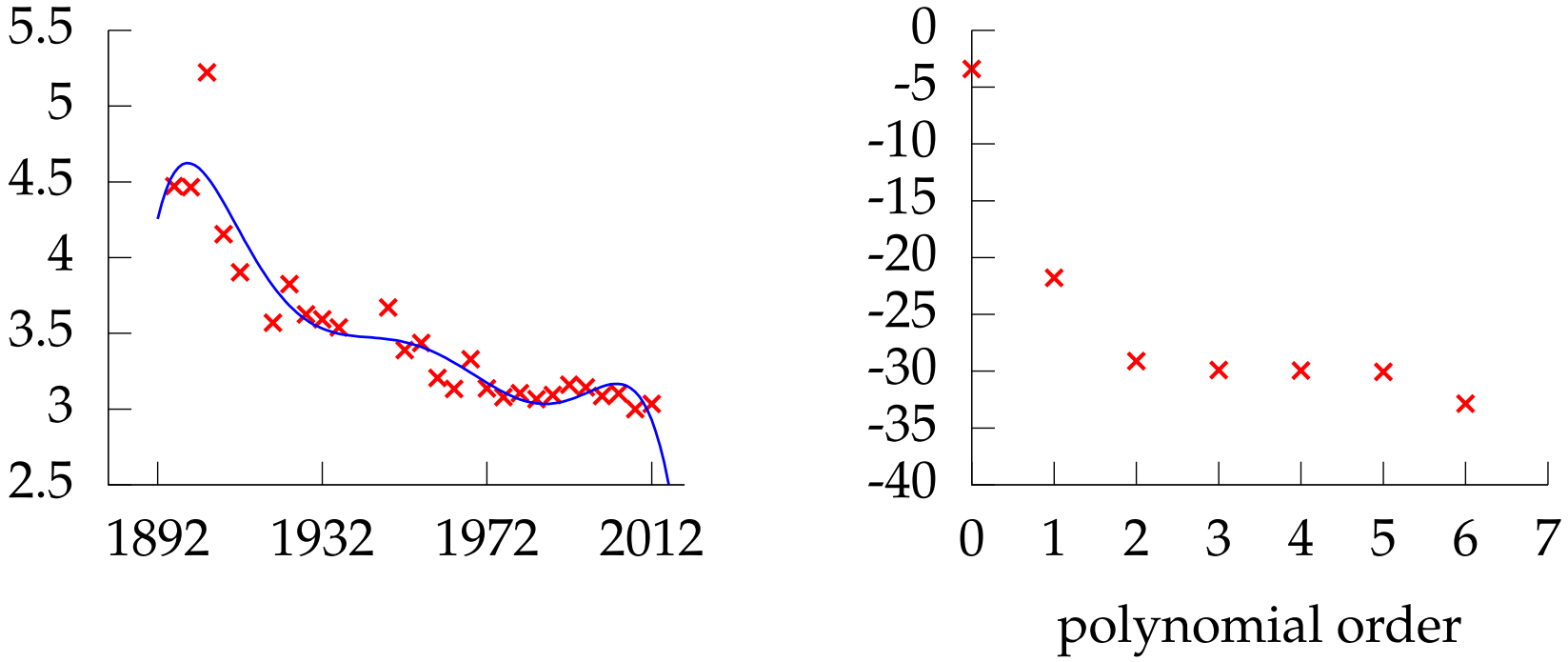
Left: fit to data, *Right*: model error. Polynomial order 4, model error -29.943, $\sigma^2 = 0.0400$, $\sigma = 0.200$.

Polynomial Fits to Olympics Data



Left: fit to data, *Right*: model error. Polynomial order 5, model error -30.056, $\sigma^2 = 0.0397$, $\sigma = 0.199$.

Polynomial Fits to Olympics Data



Left: fit to data, *Right*: model error. Polynomial order 6, model error -32.866, $\sigma^2 = 0.0322$, $\sigma = 0.180$.

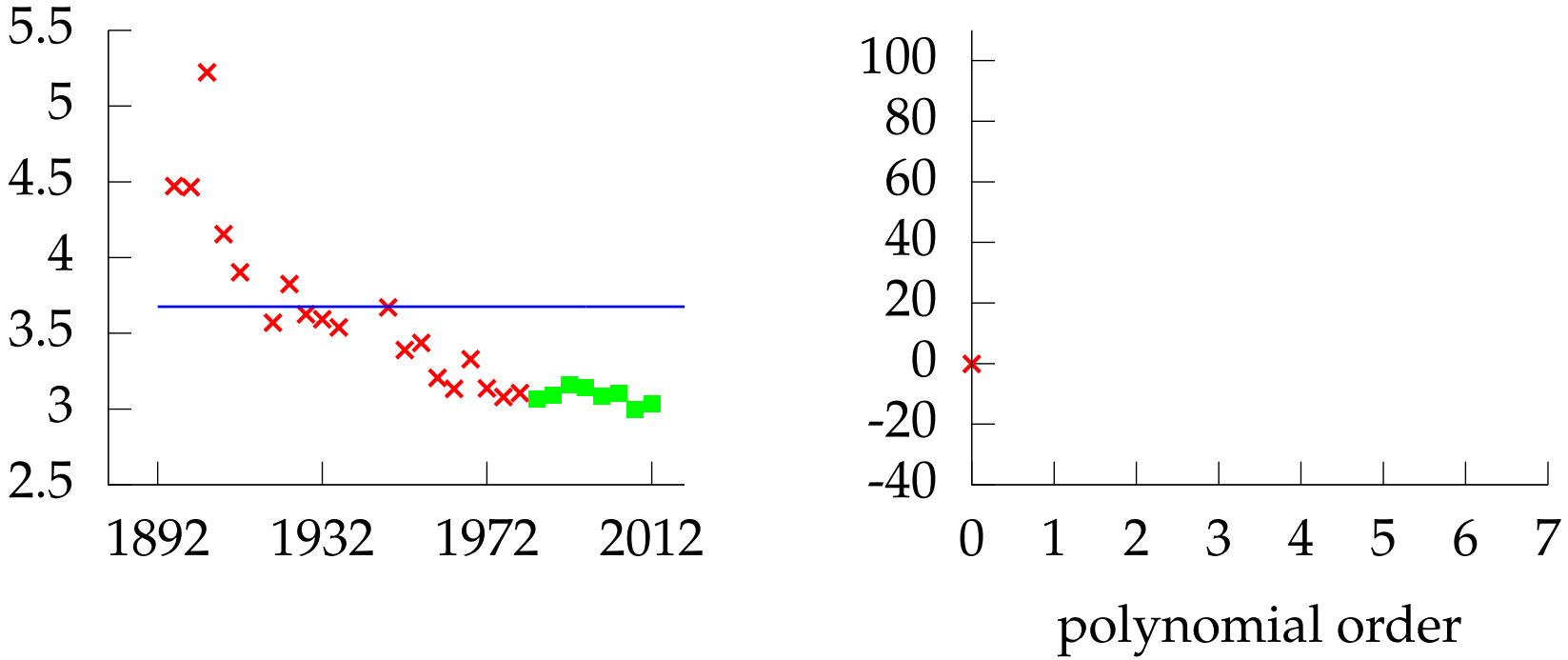
Overfitting

- Increase number of basis functions, we obtain a better ‘fit’ to the data.
- How will the model perform on previously unseen data?

Training and Test Sets

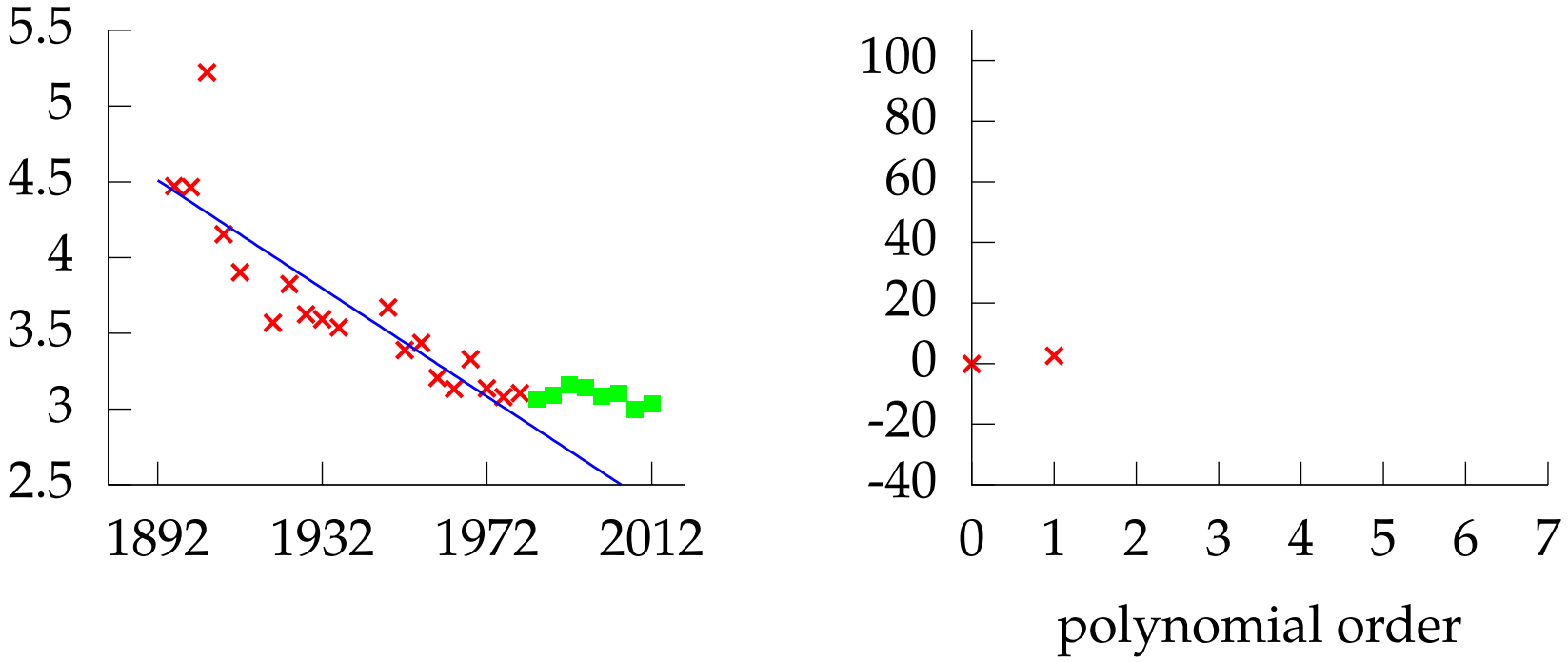
- We call the data used for fitting the model the ‘training set’.
- Data not used for training, but when the model is applied ‘in the field’ is called the ‘test data’.
- Challenge for generalization is to ensure a good performance on test data given only training data.

Validation Set



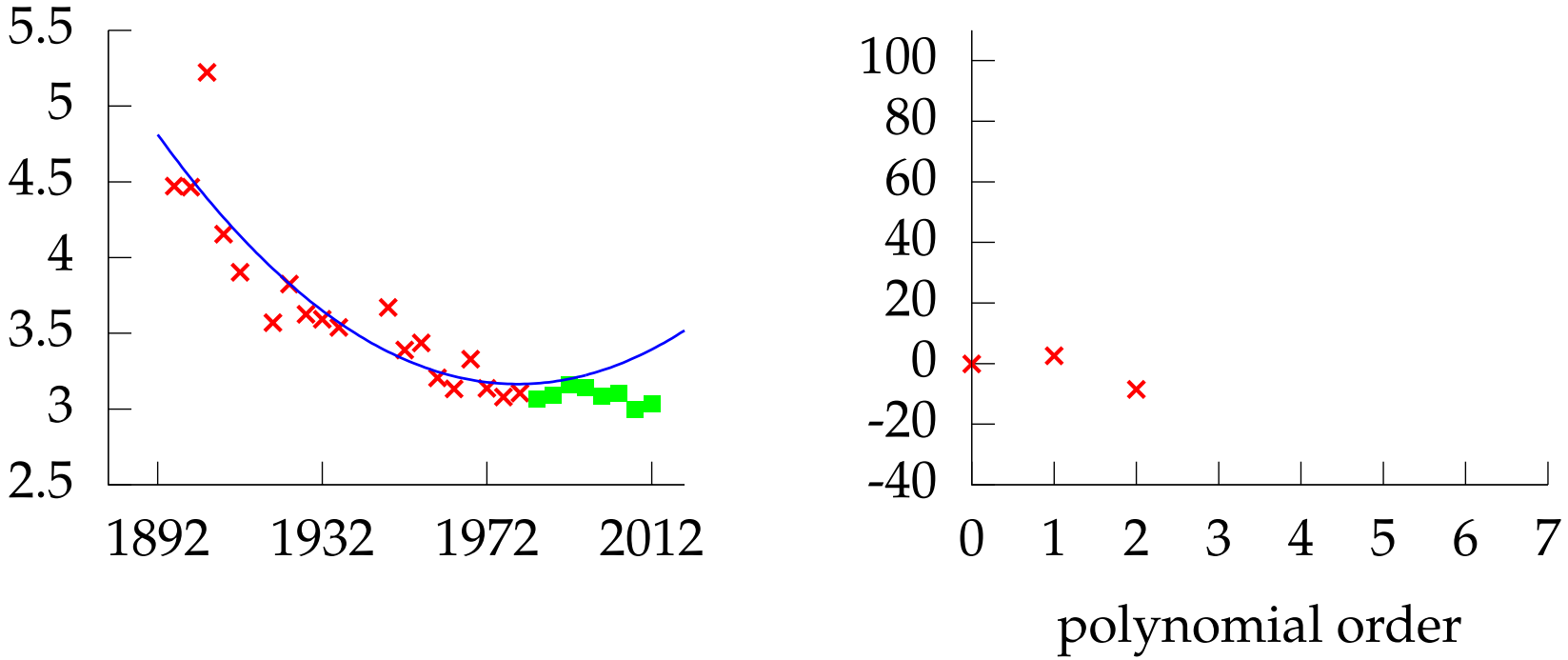
Left: fit to data, *Right*: model error. Polynomial order 0, training error -1.8774, validation error -0.13132, $\sigma^2 = 0.302$, $\sigma = 0.549$.

Validation Set



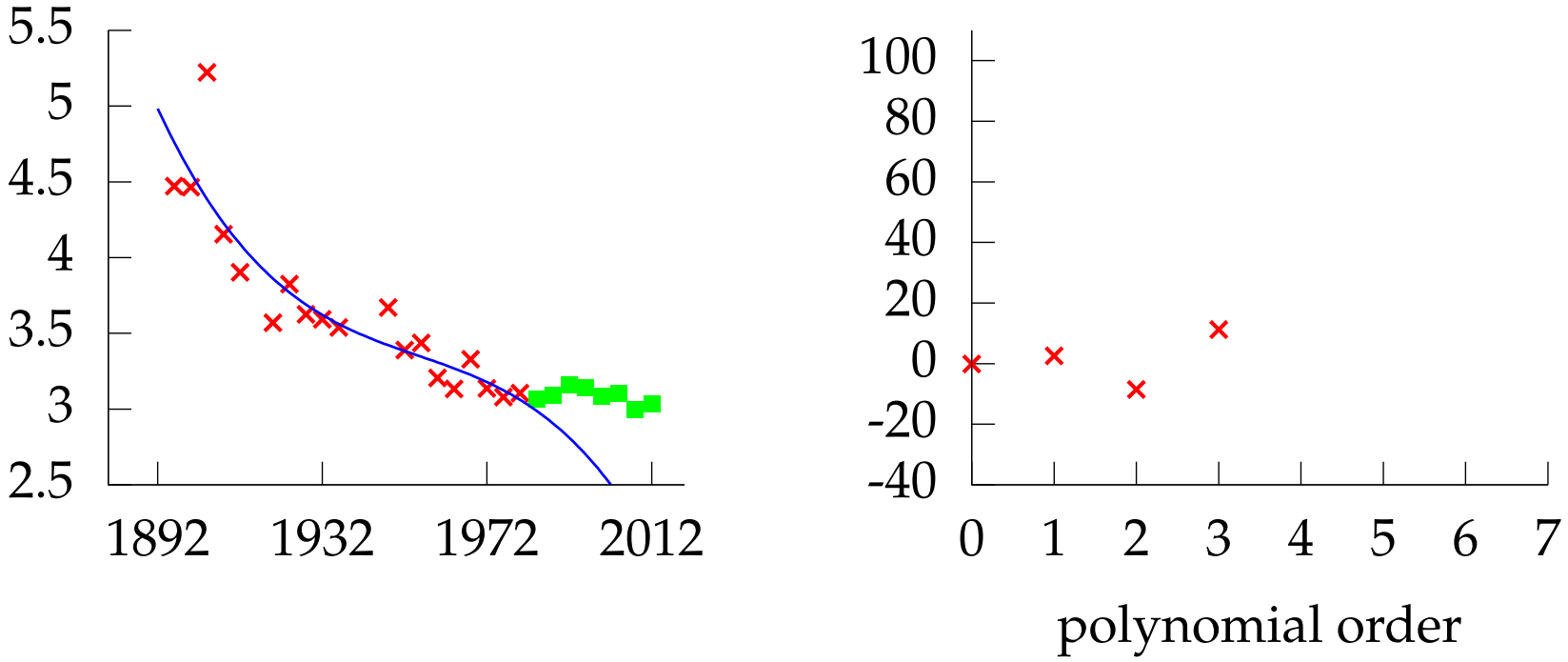
Left: fit to data, *Right*: model error. Polynomial order 1, training error -15.325, validation error 2.5863, $\sigma^2 = 0.0733$, $\sigma = 0.271$.

Validation Set



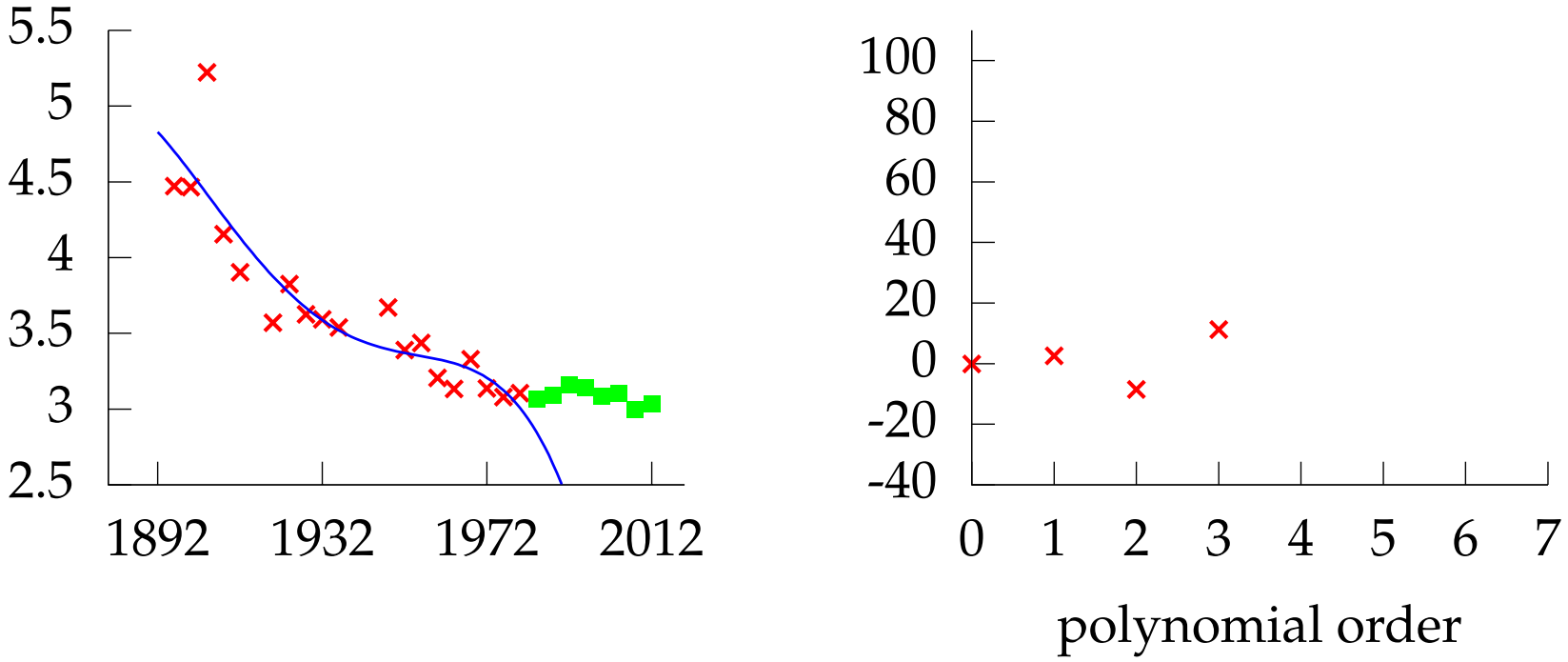
Left: fit to data, *Right*: model error. Polynomial order 2, training error -17.579, validation error -8.4831, $\sigma^2 = 0.0578$, $\sigma = 0.240$.

Validation Set



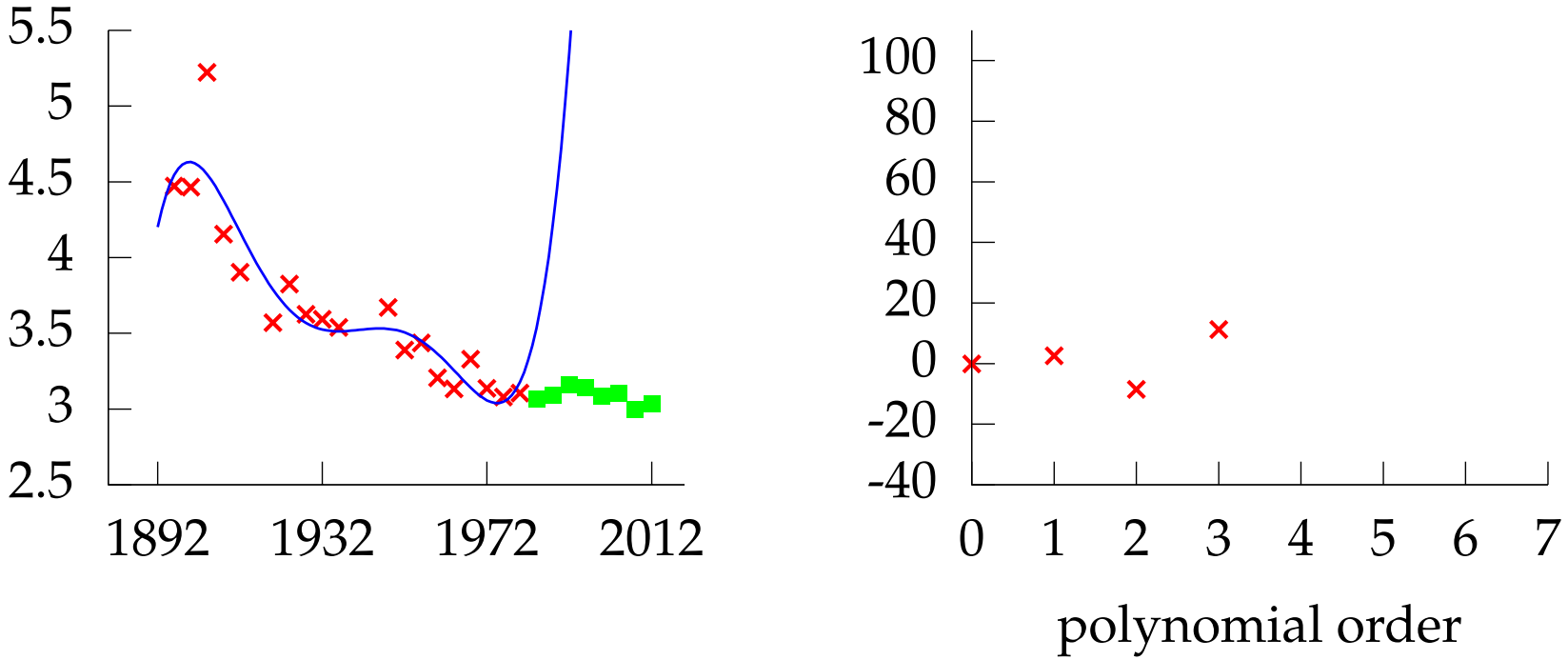
Left: fit to data, *Right*: model error. Polynomial order 3, training error -18.064, validation error 11.27, $\sigma^2 = 0.0549$, $\sigma = 0.234$.

Validation Set



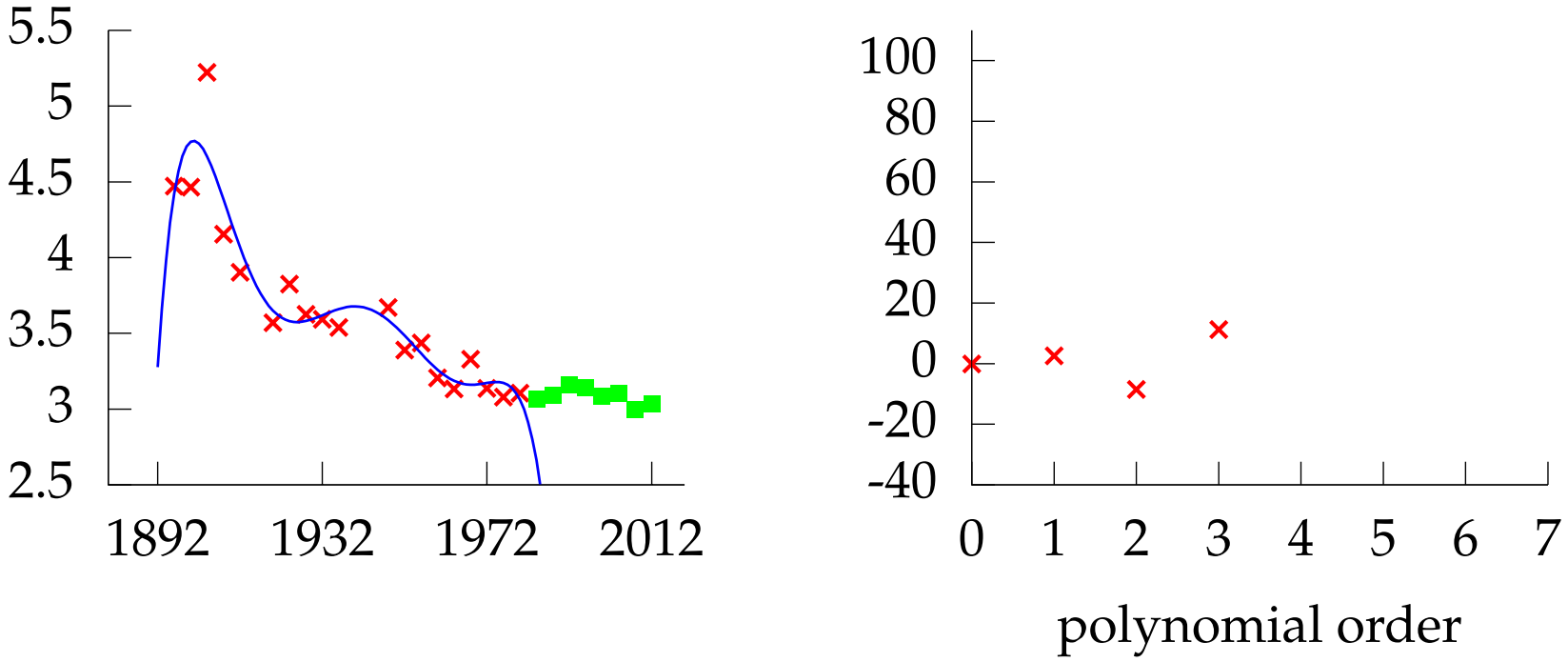
Left: fit to data, *Right*: model error. Polynomial order 4, training error -18.245, validation error 232.92, $\sigma^2 = 0.0539$, $\sigma = 0.232$.

Validation Set



Left: fit to data, *Right*: model error. Polynomial order 5, training error -20.471, validation error 9898.1, $\sigma^2 = 0.0426$, $\sigma = 0.207$.

Validation Set

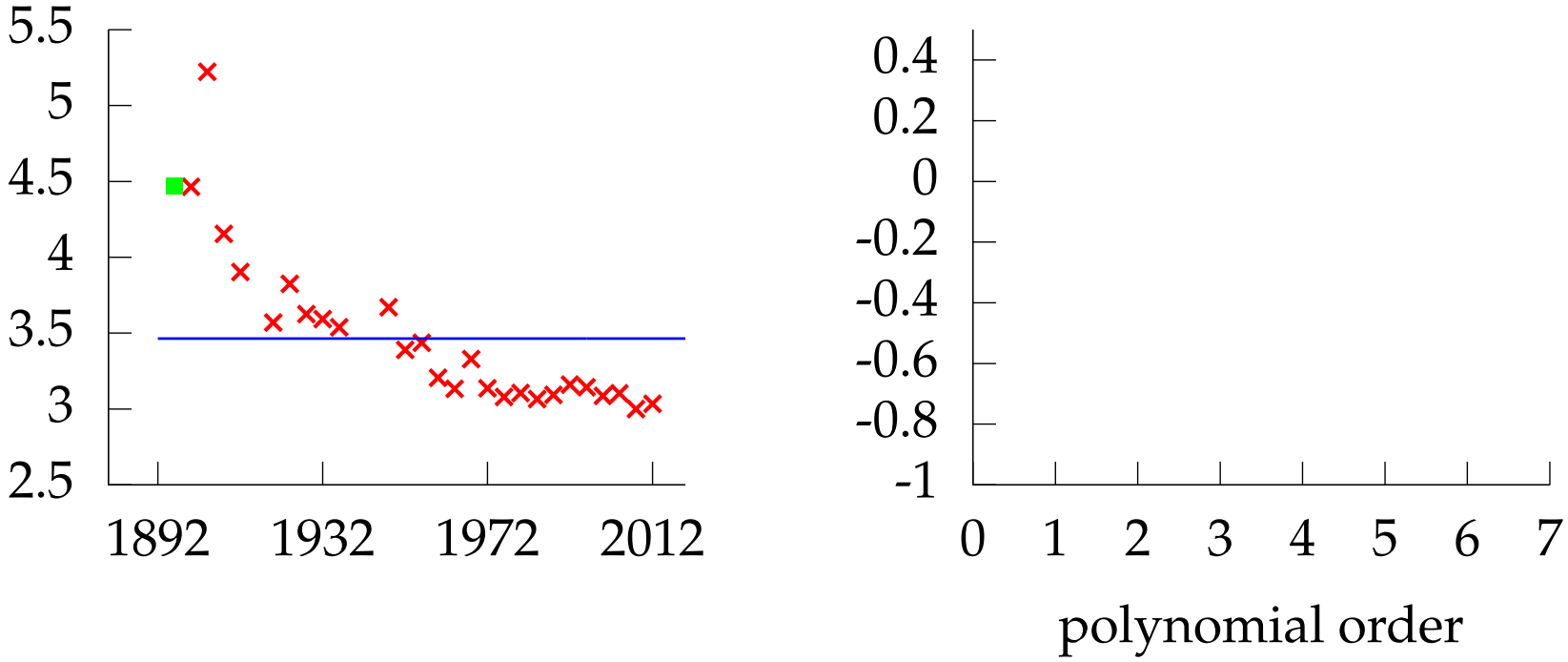


Left: fit to data, *Right*: model error. Polynomial order 6, training error -22.881, validation error 67775, $\sigma^2 = 0.0331$, $\sigma = 0.182$.

Leave One Out Error

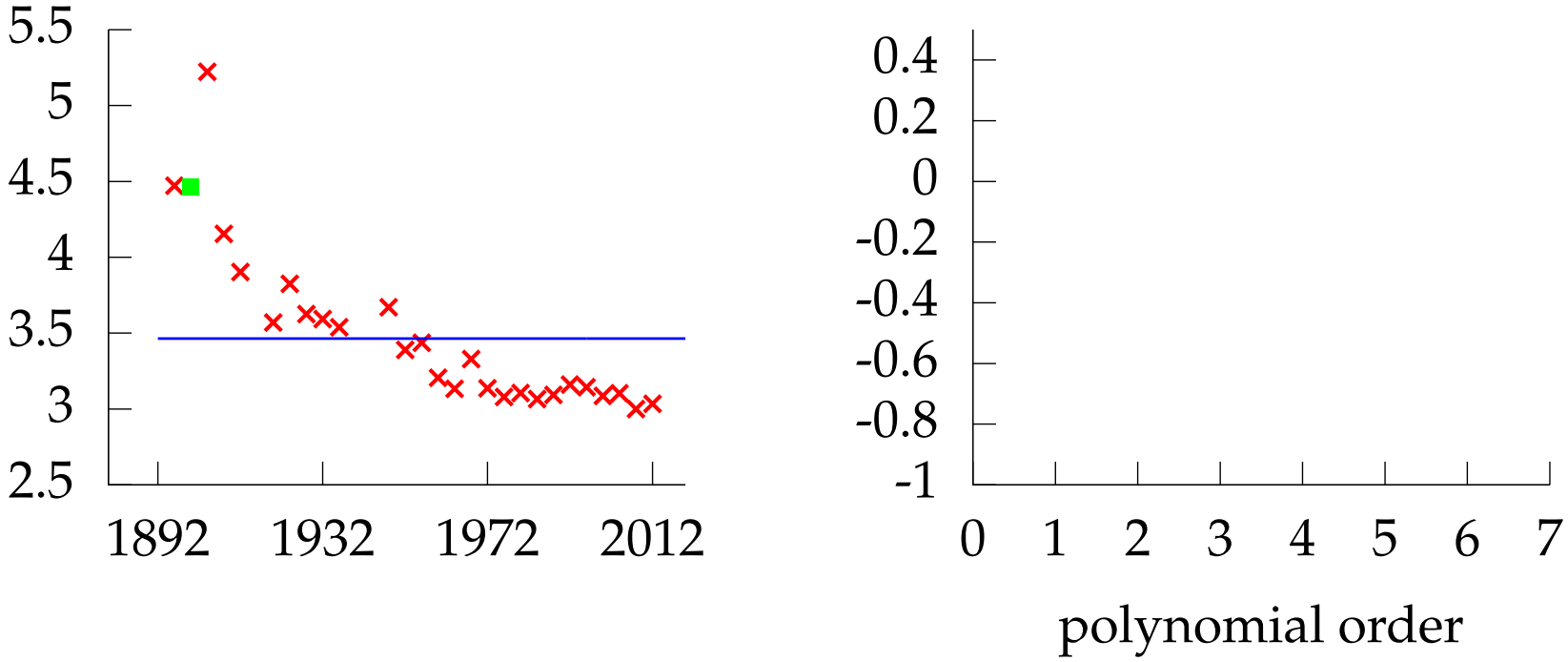
- Take training set and remove one point.
- Train on the remaining data.
- Compute the error on the point you removed (which wasn't in the training data).
- Do this for each point in the training set in turn.
- Average the resulting error. This is the leave one out error.

Leave One Out Error



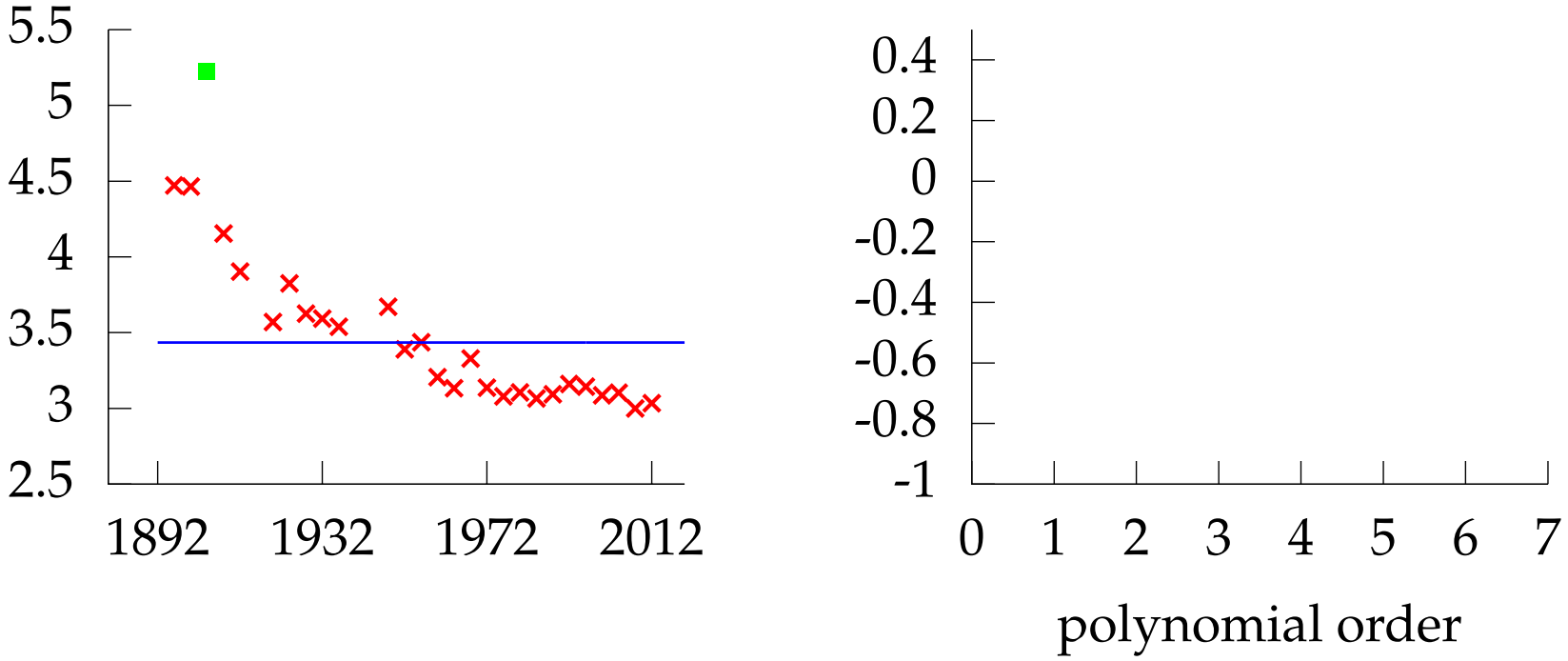
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



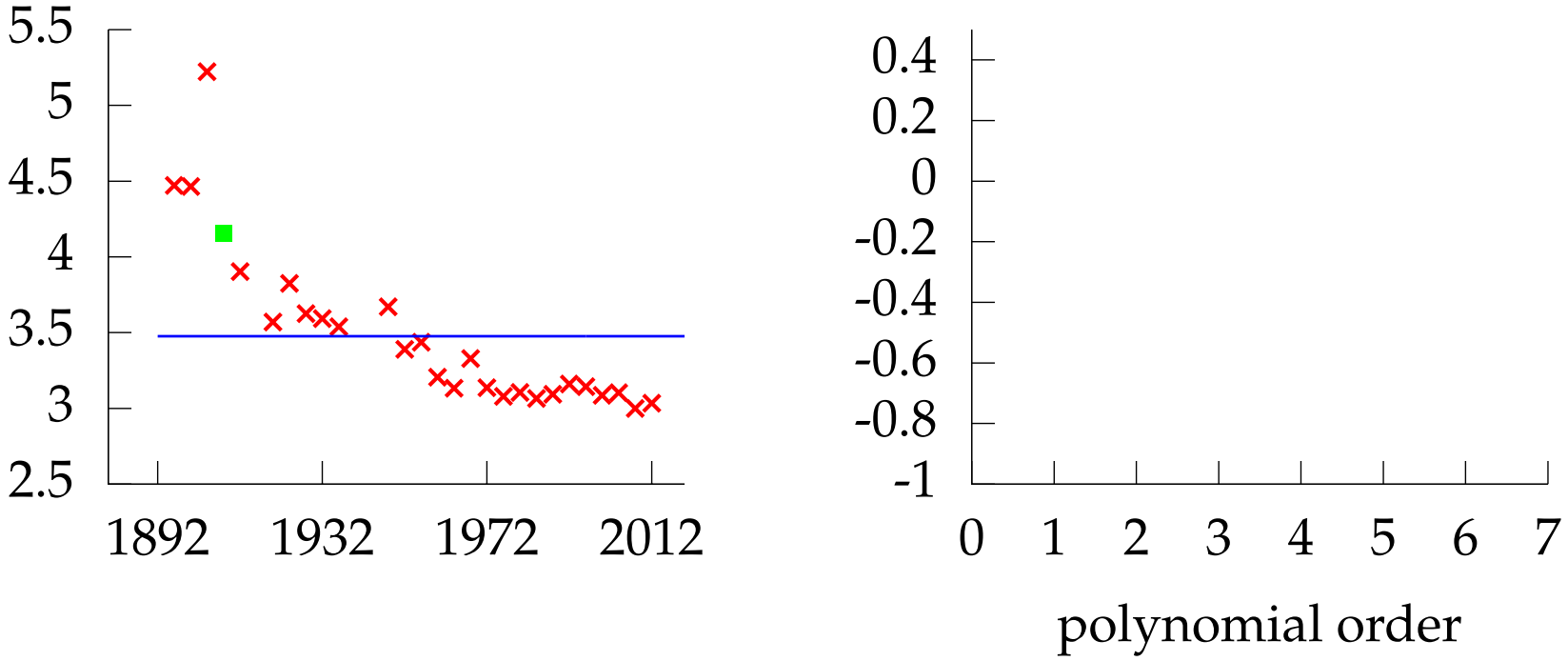
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



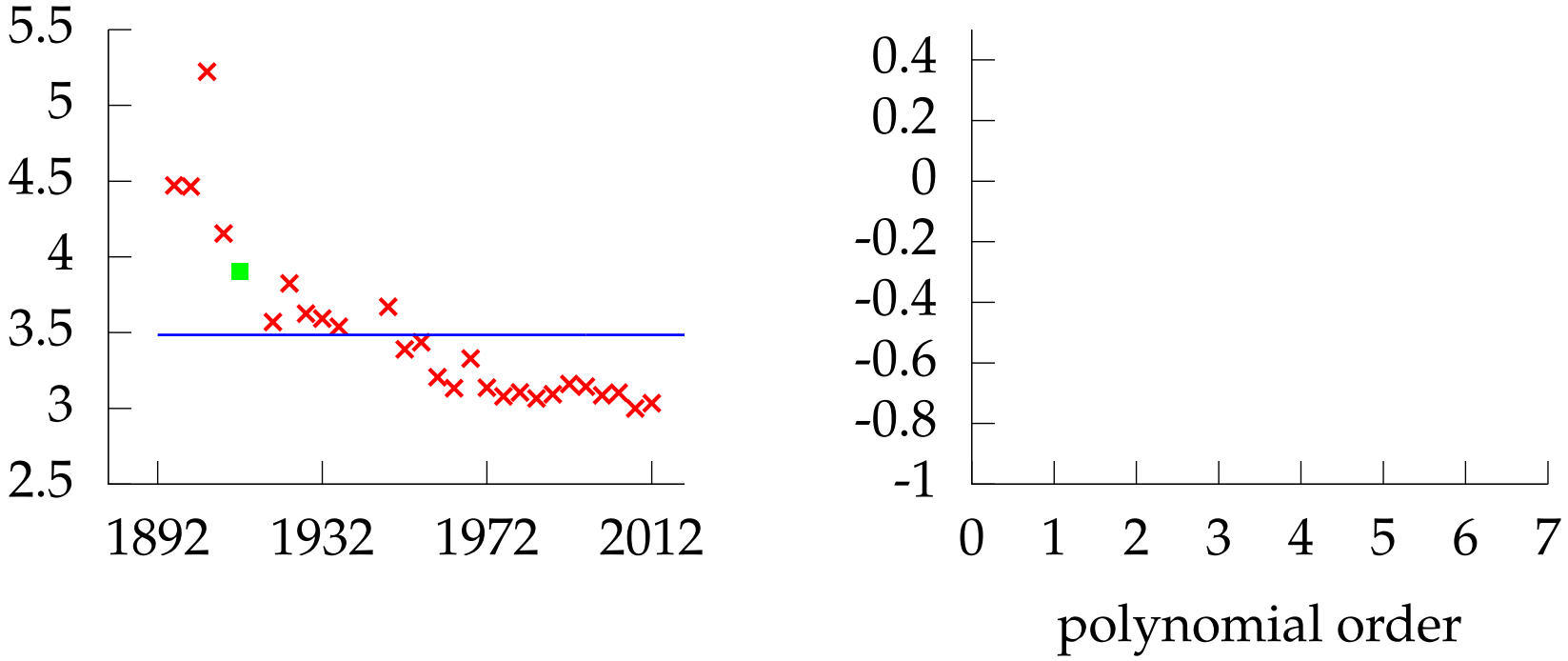
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



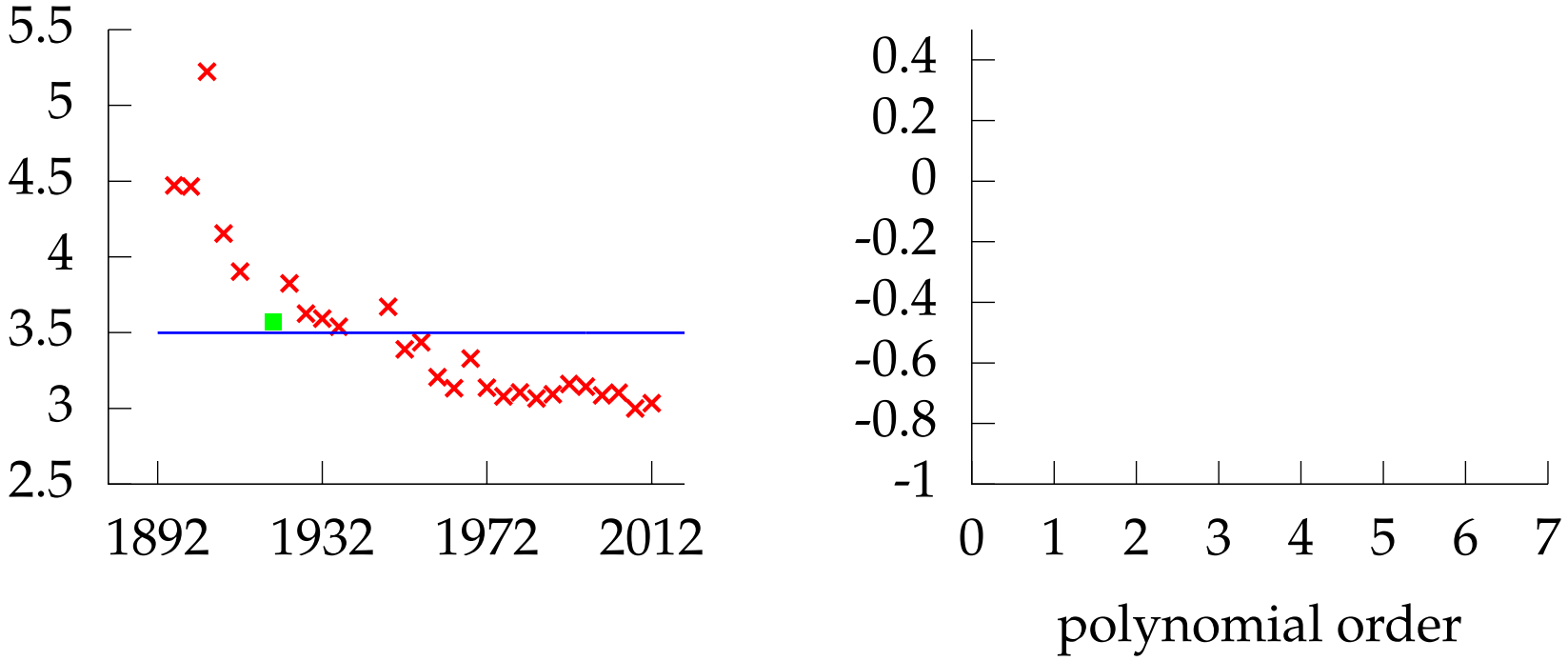
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



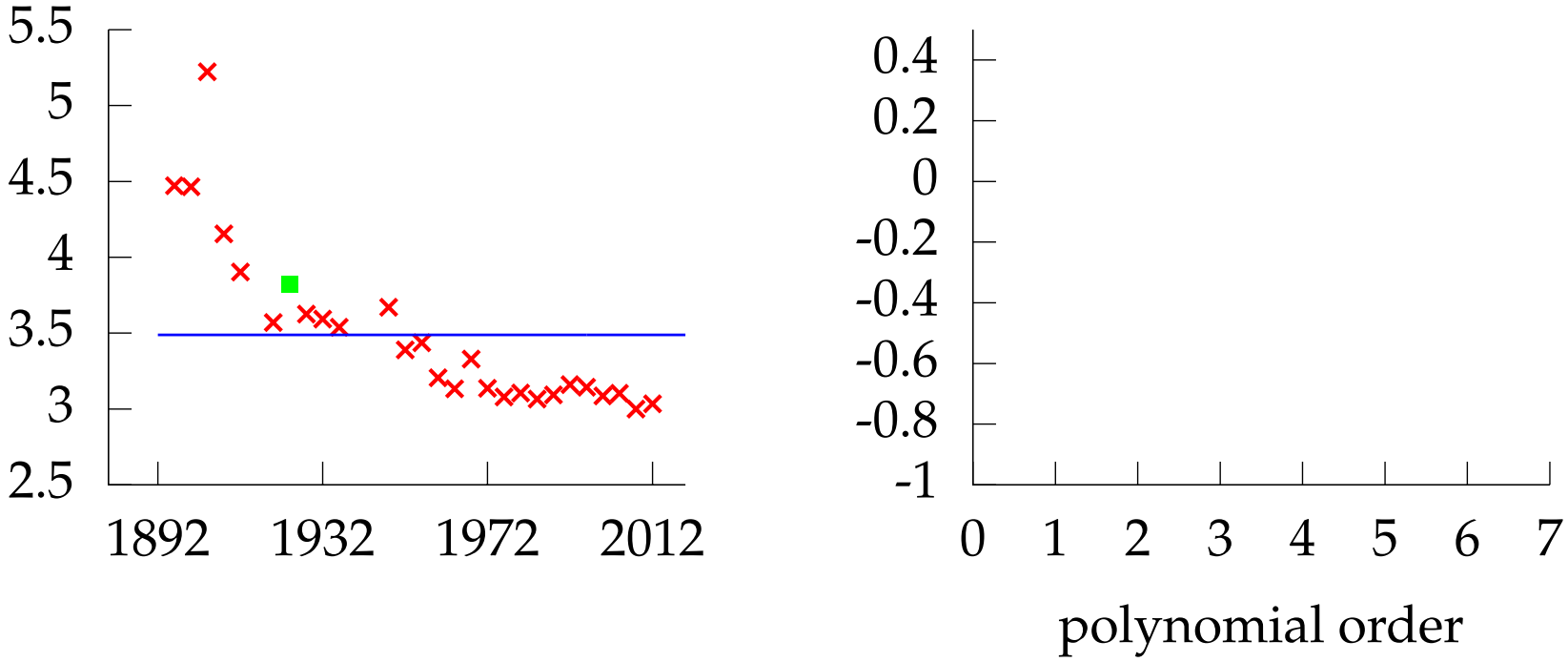
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



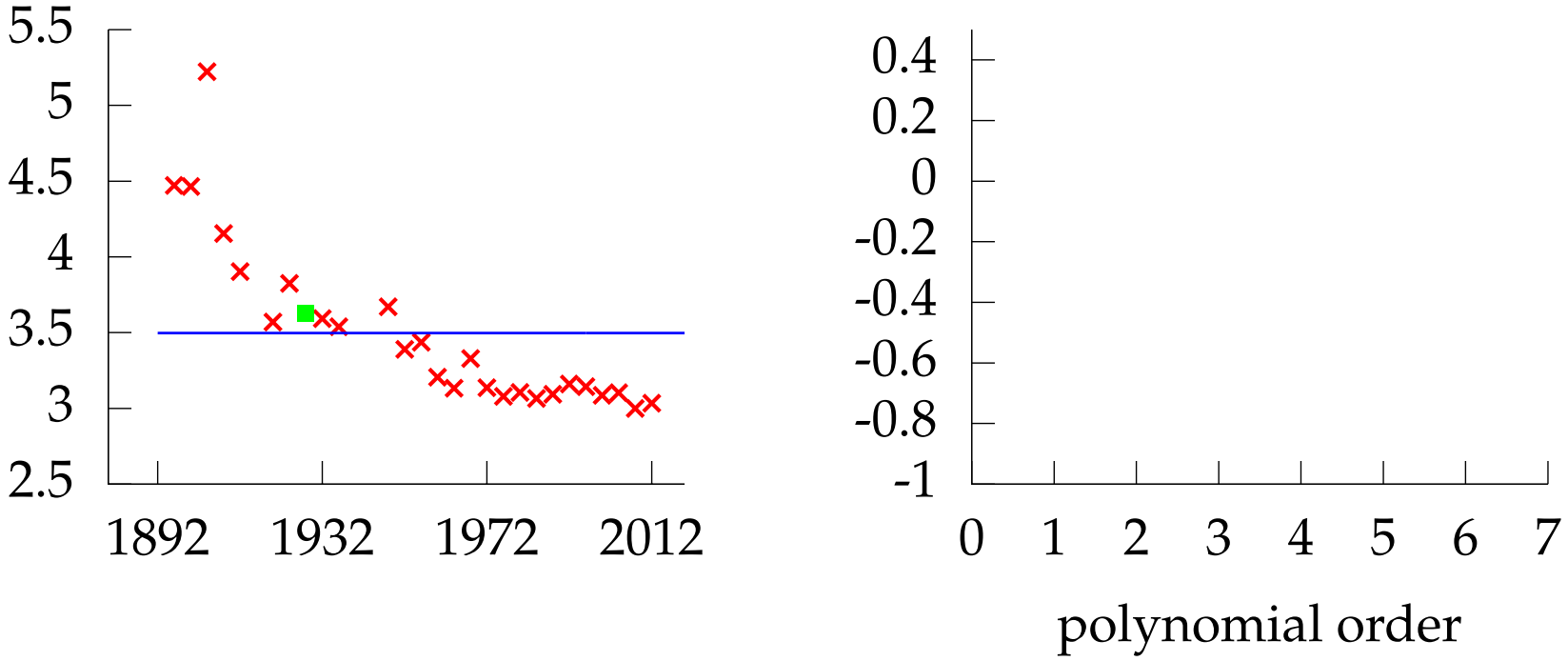
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



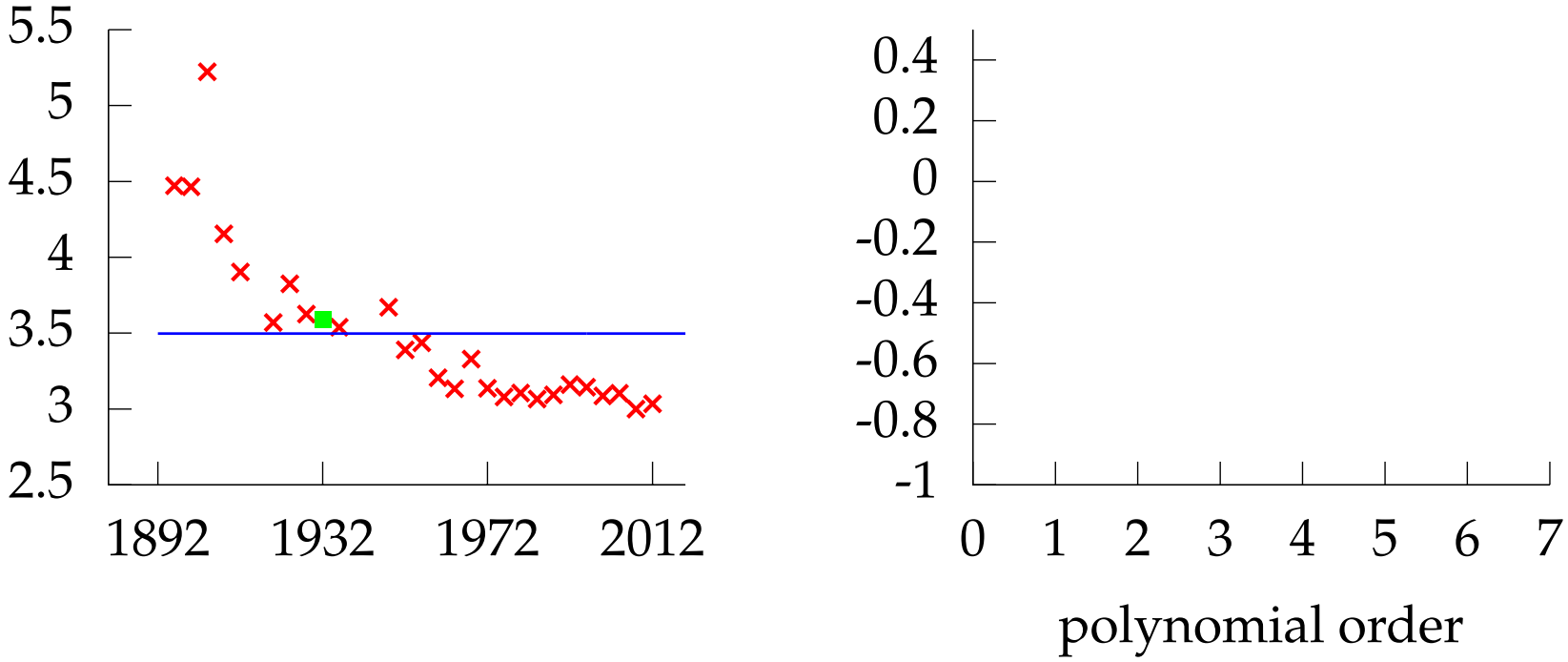
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



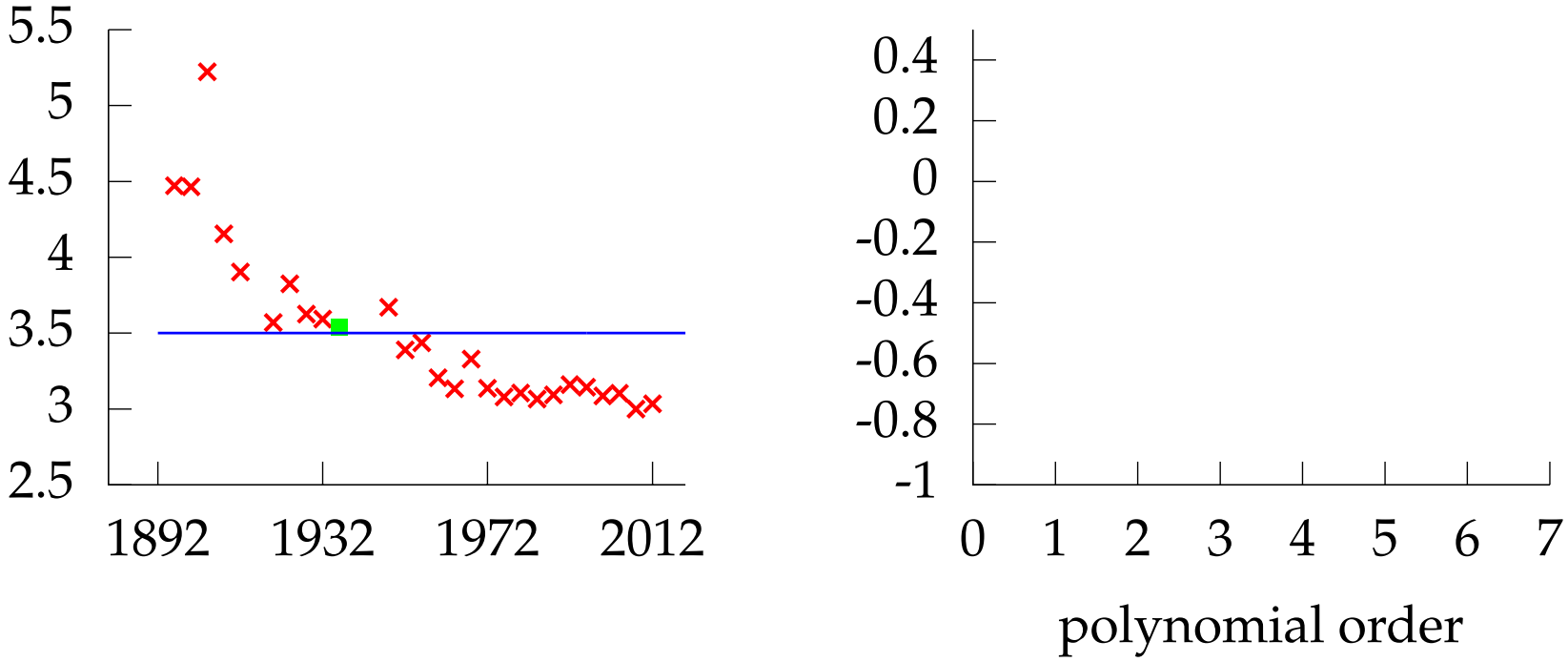
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



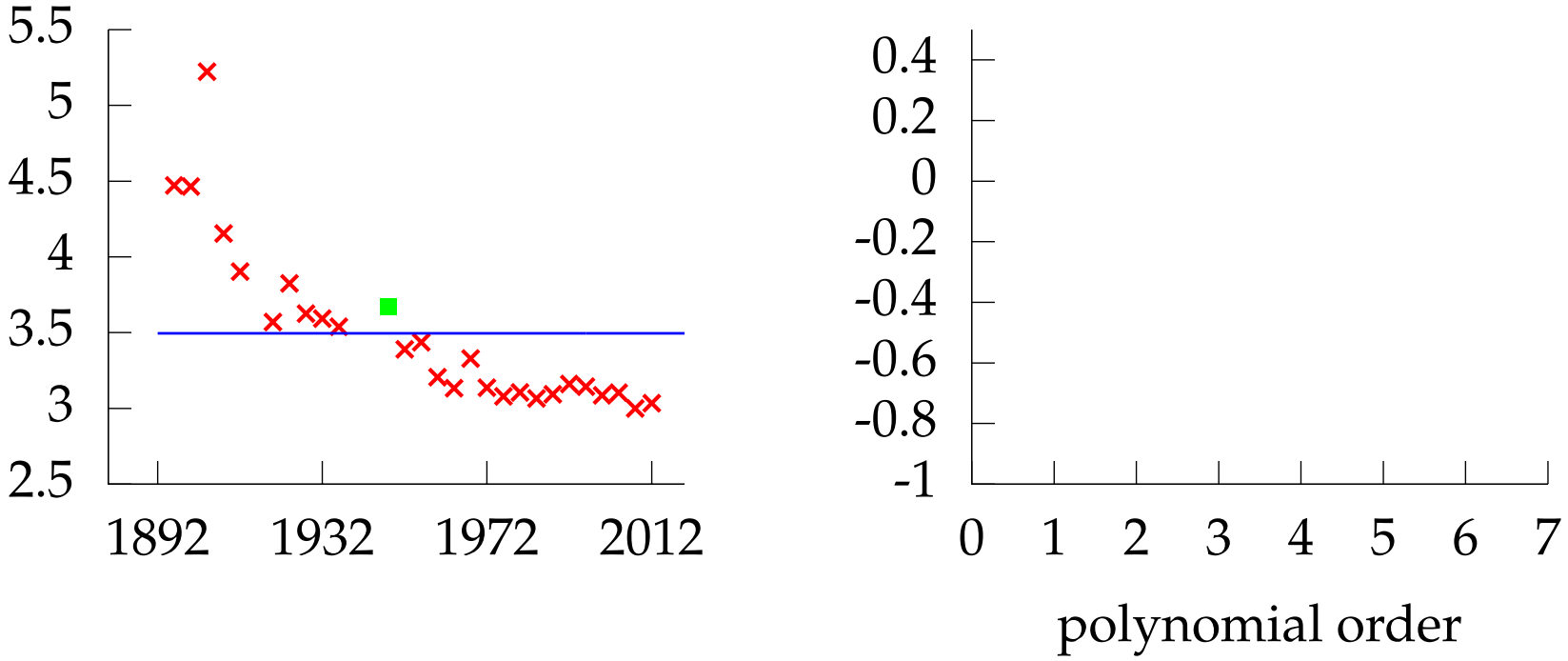
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



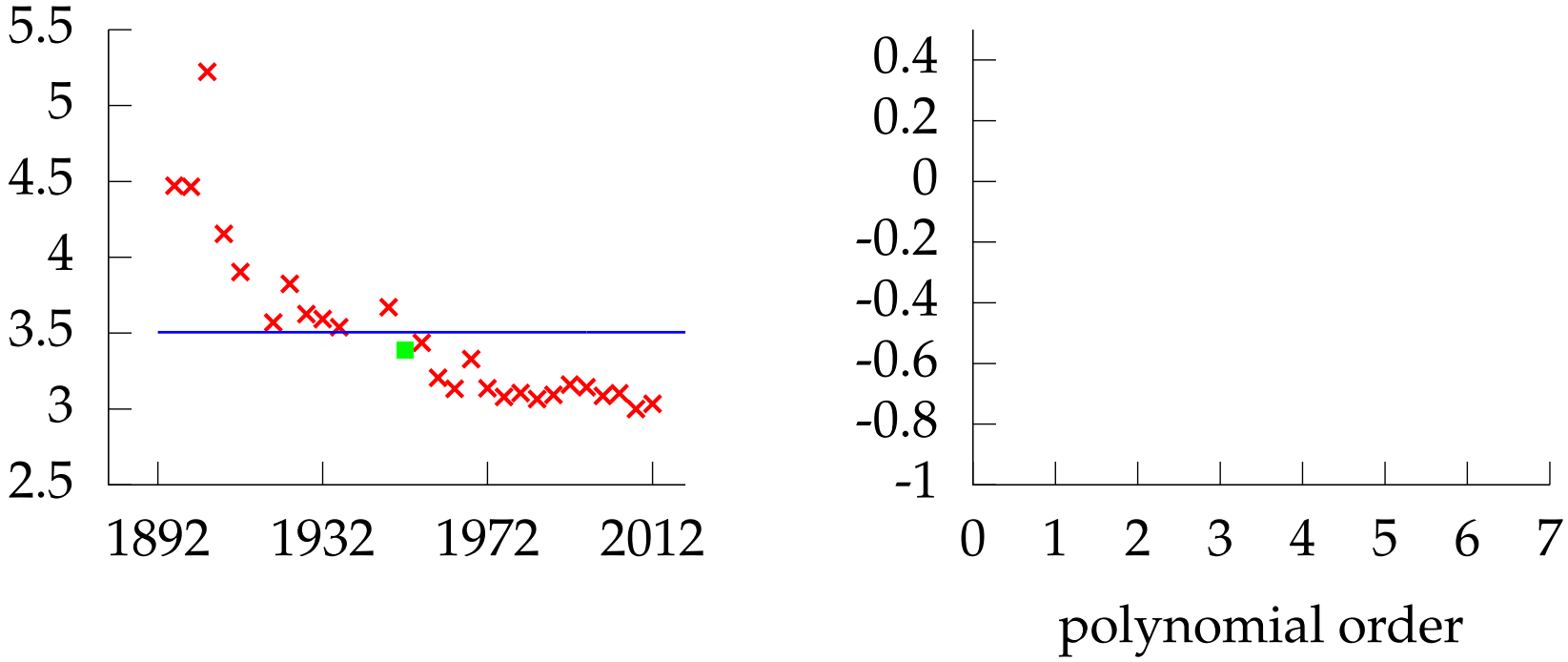
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



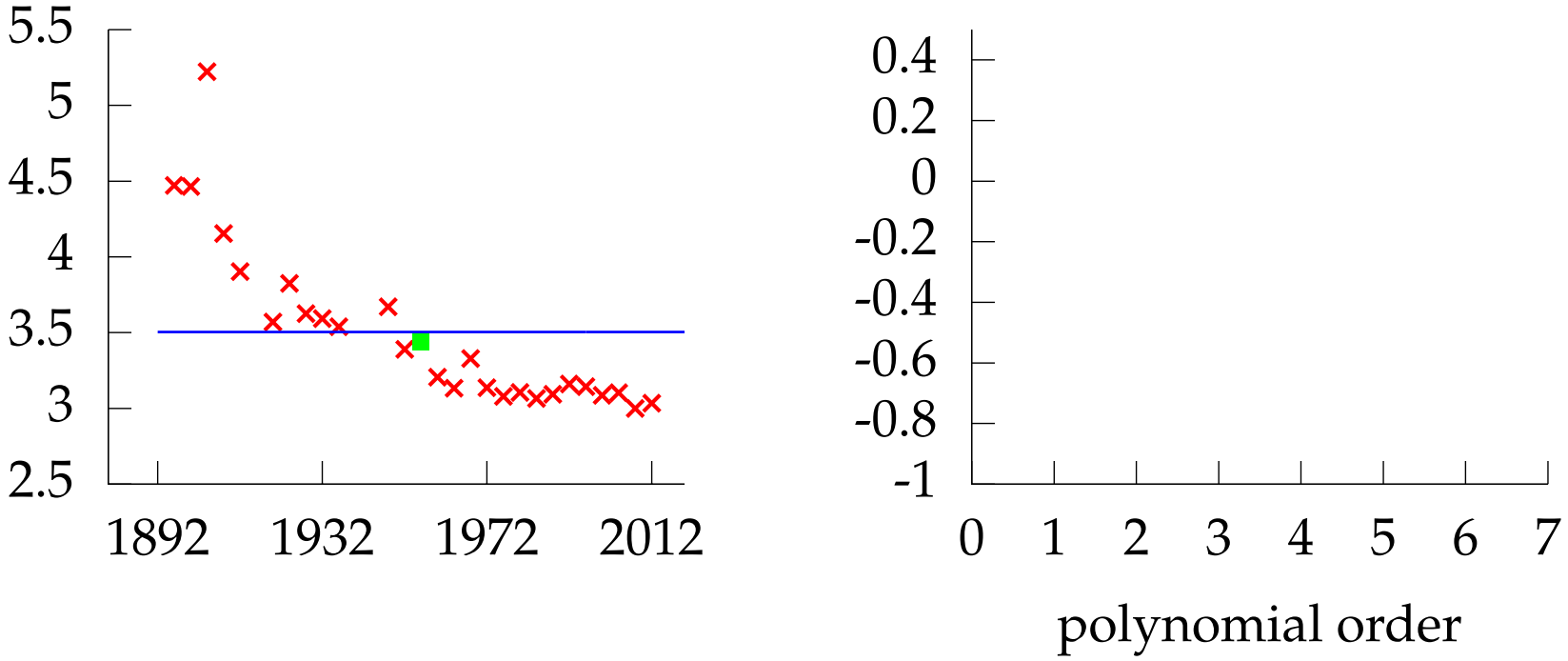
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



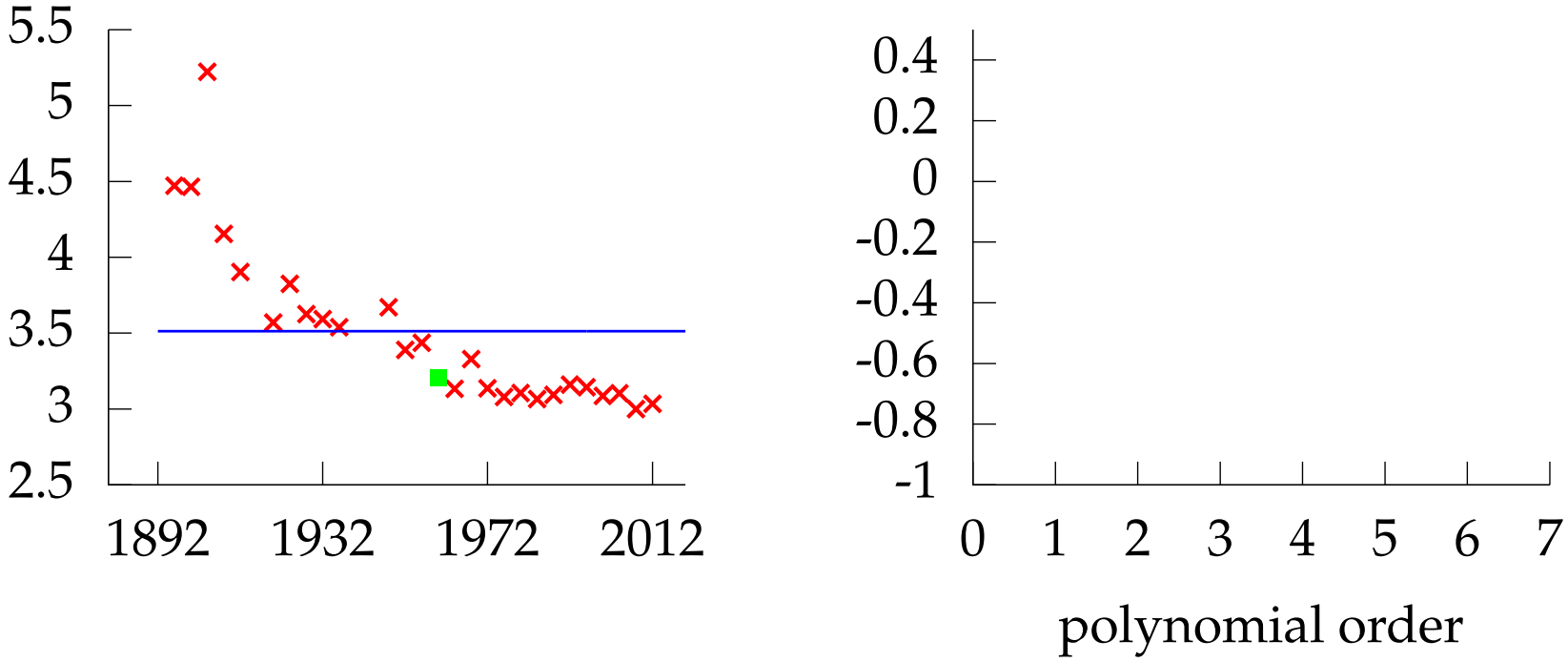
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



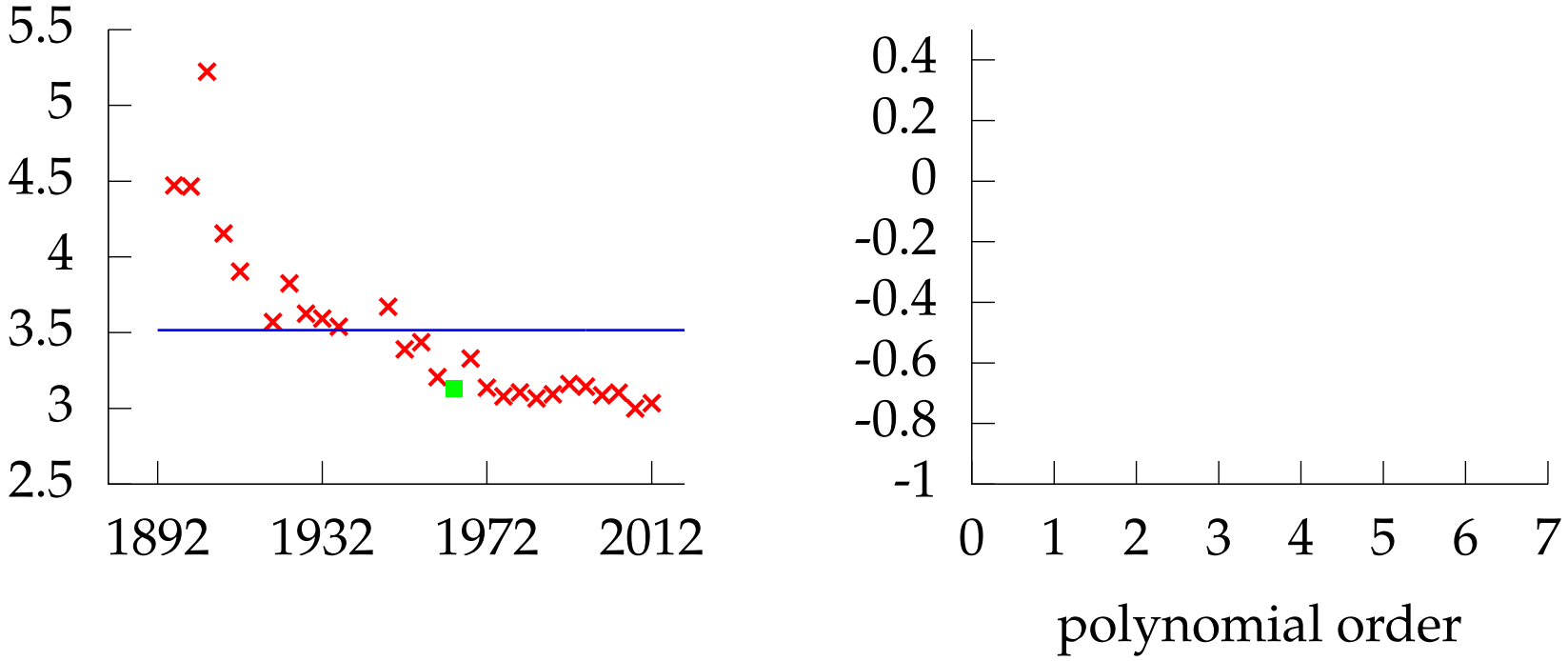
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



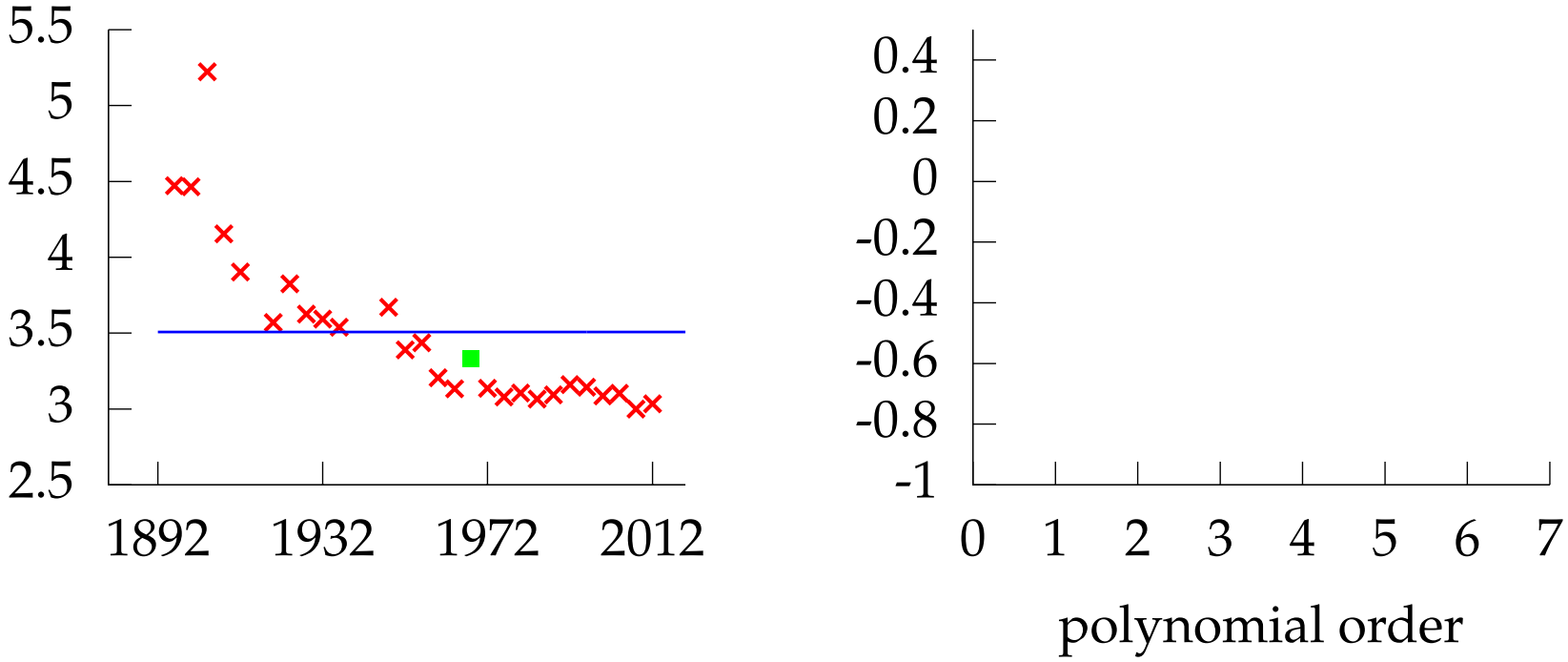
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



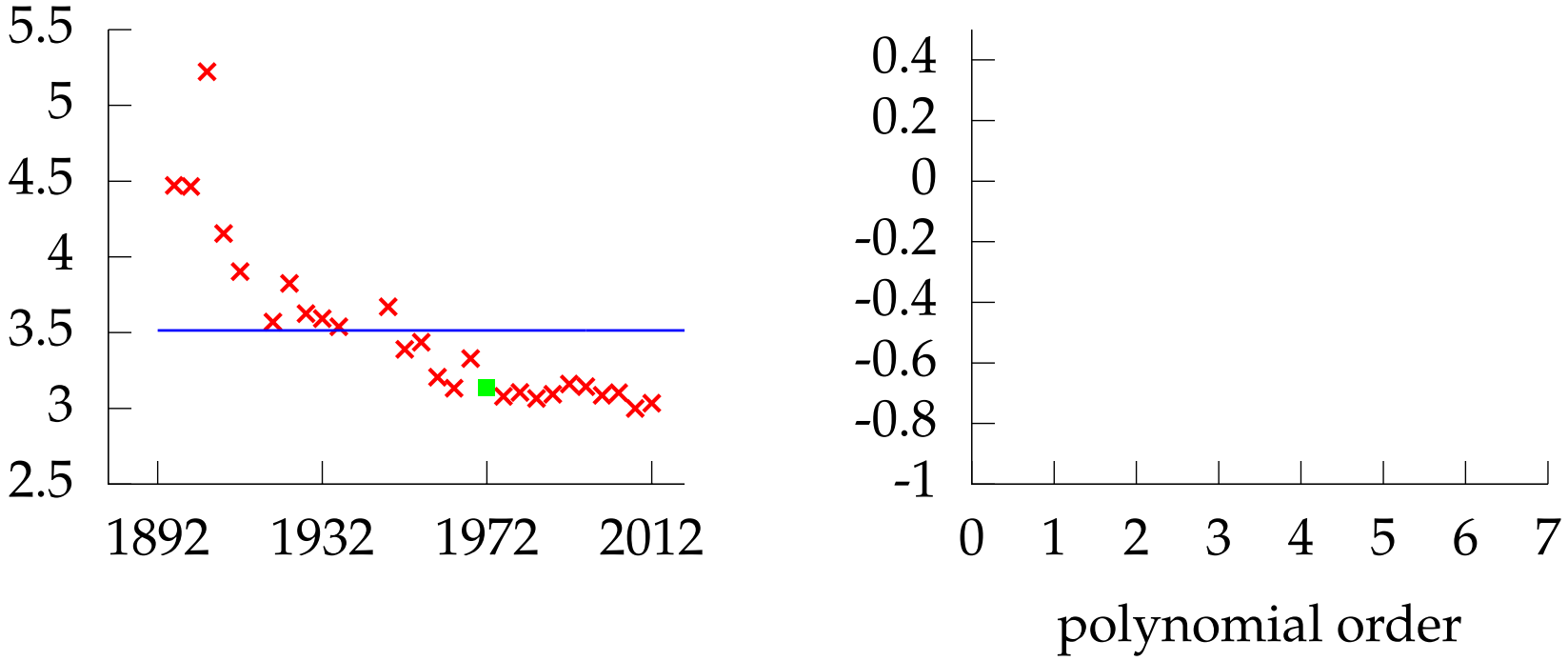
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



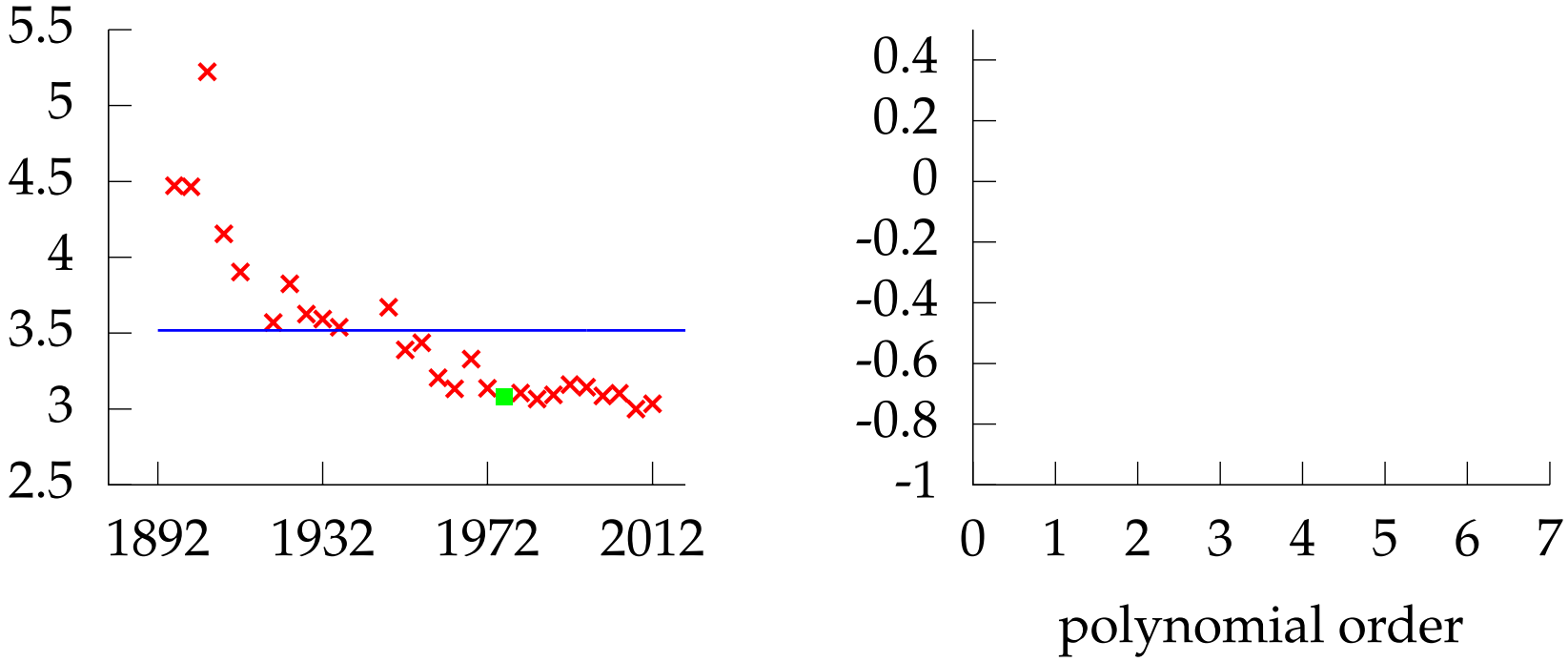
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



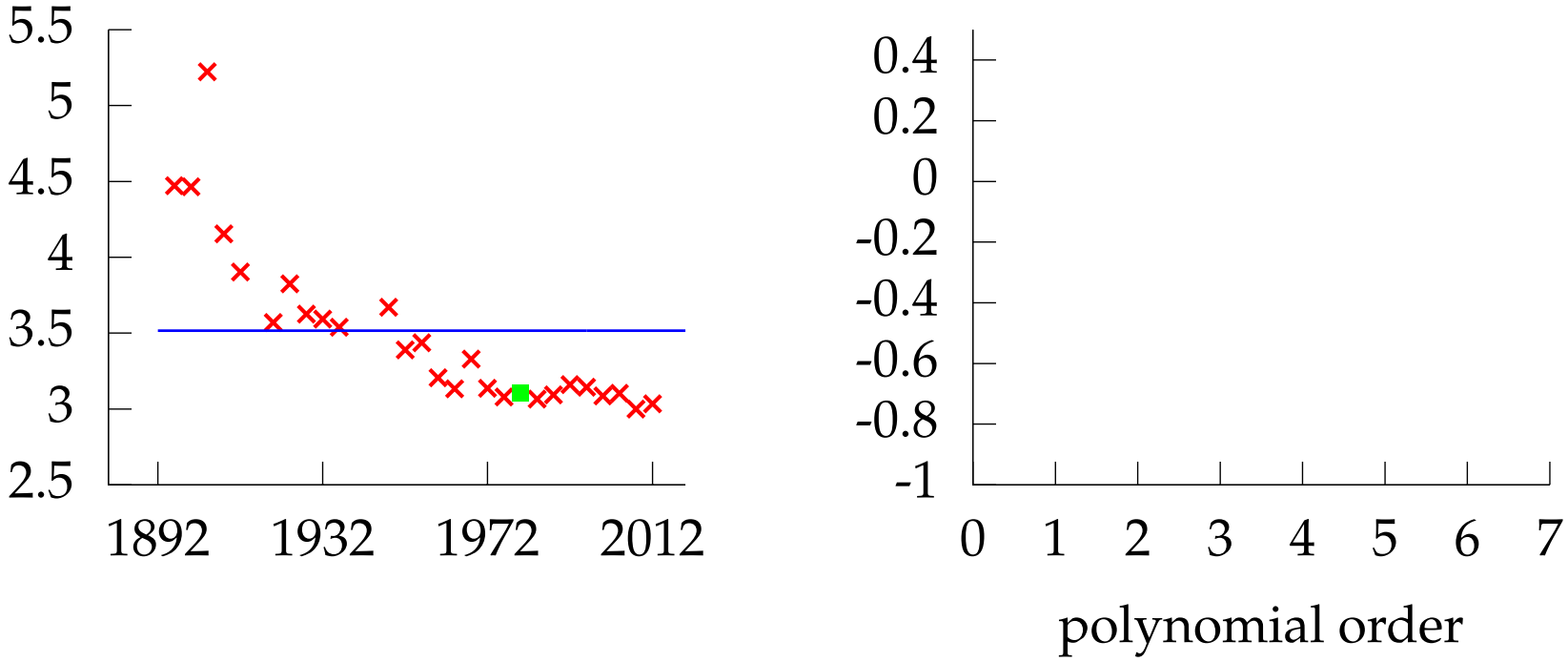
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



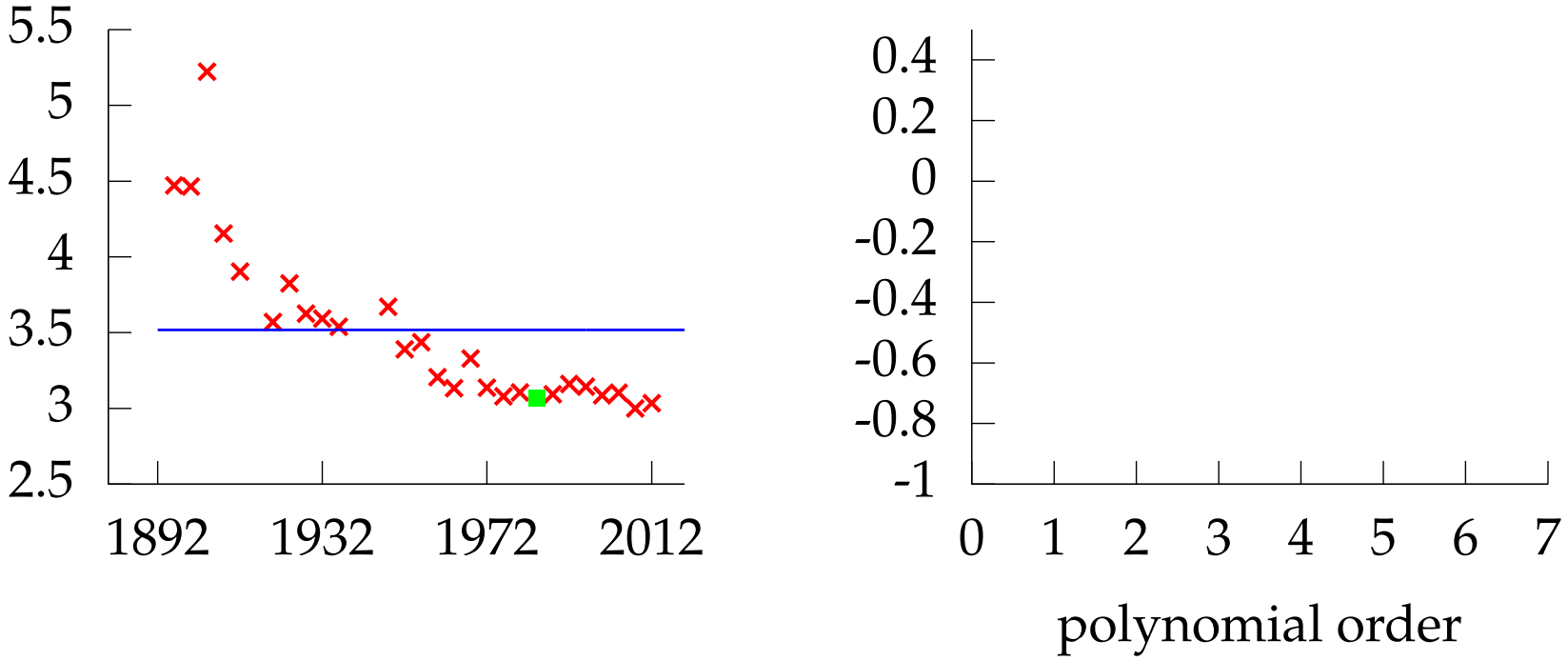
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



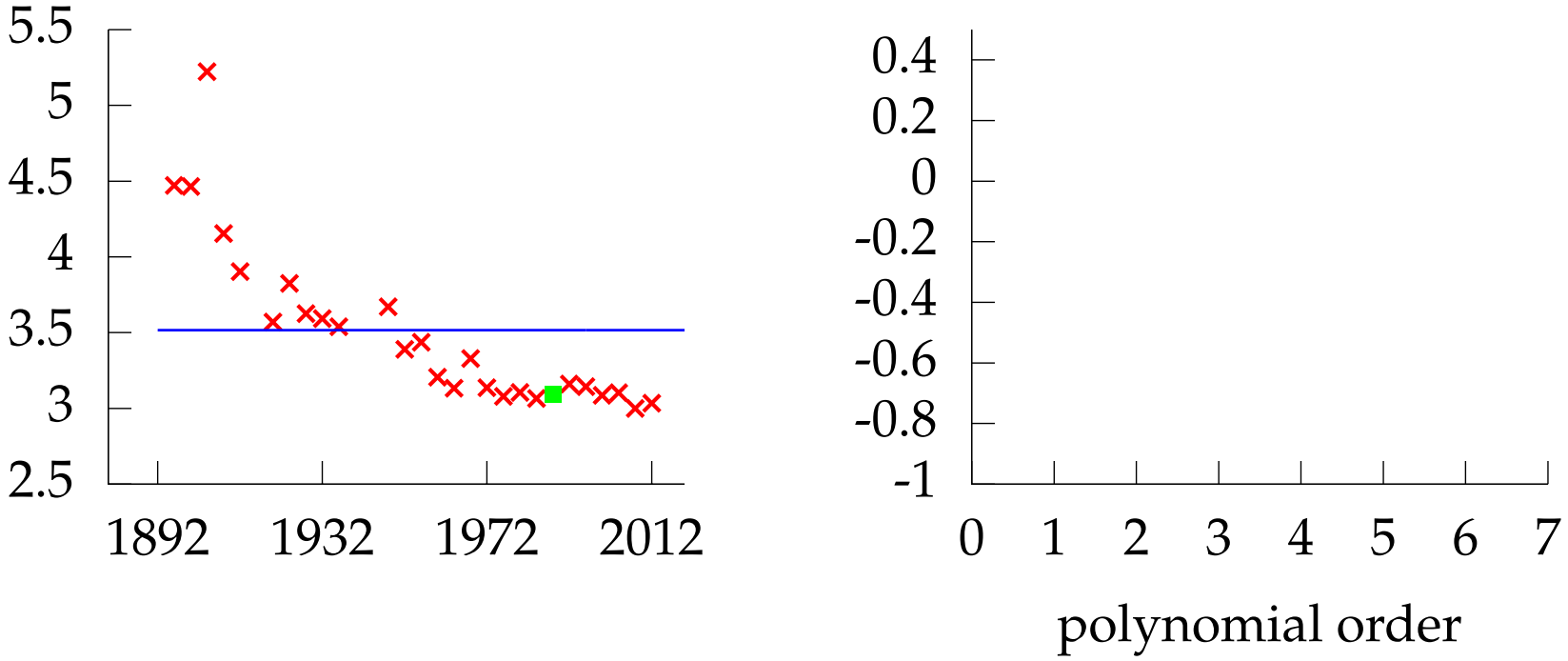
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



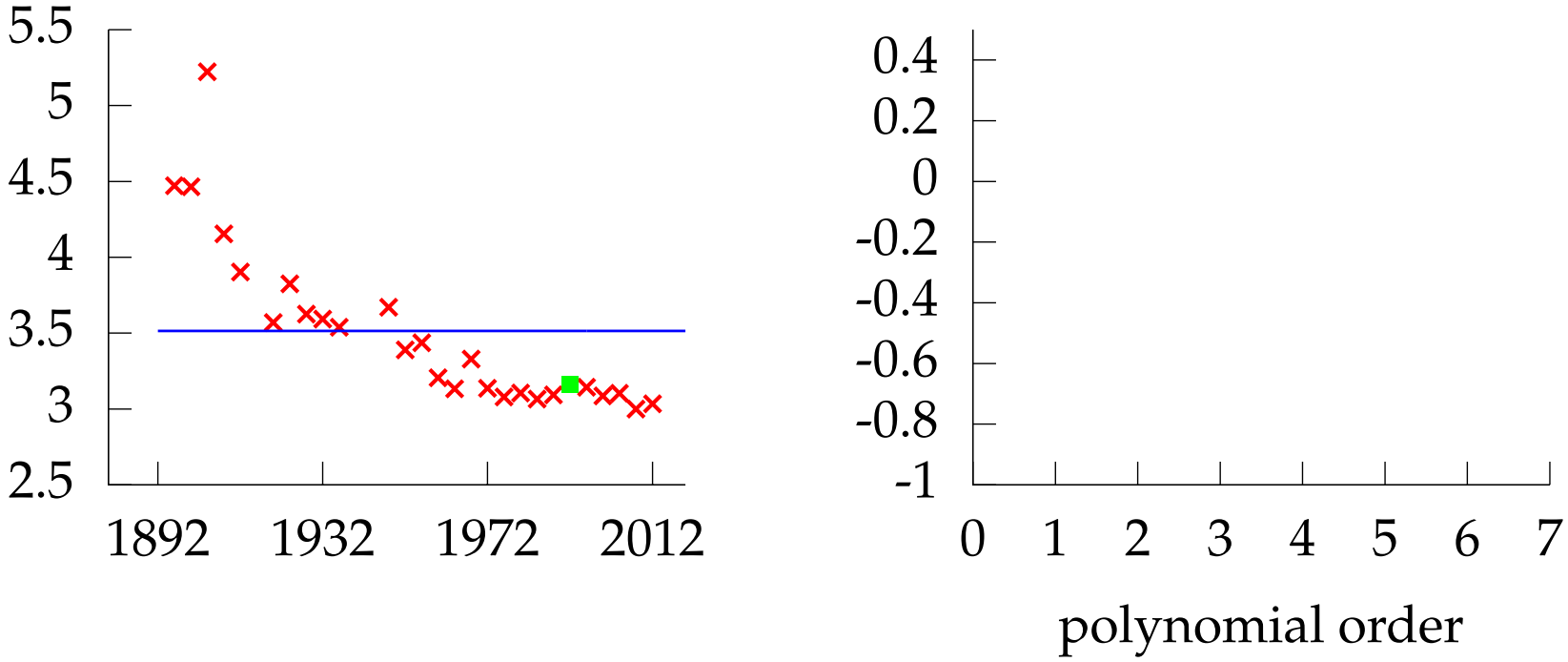
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



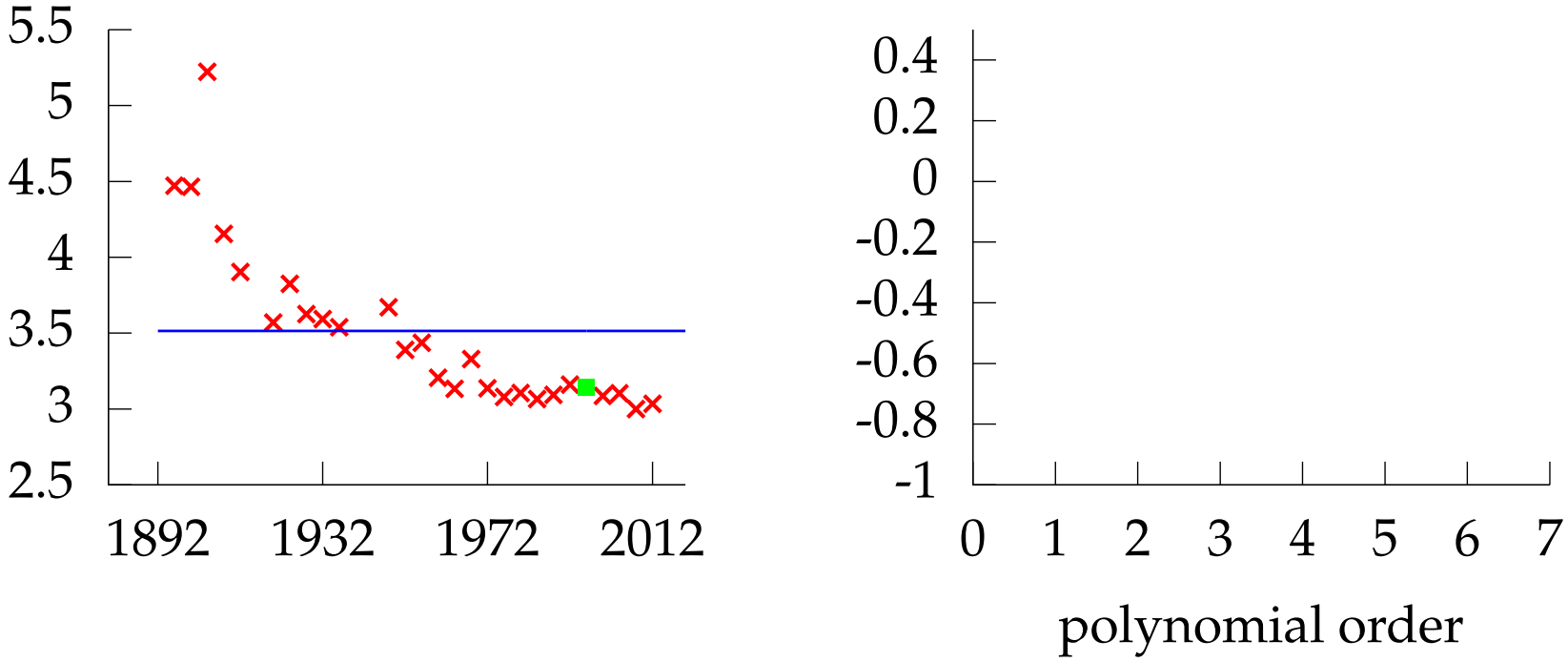
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



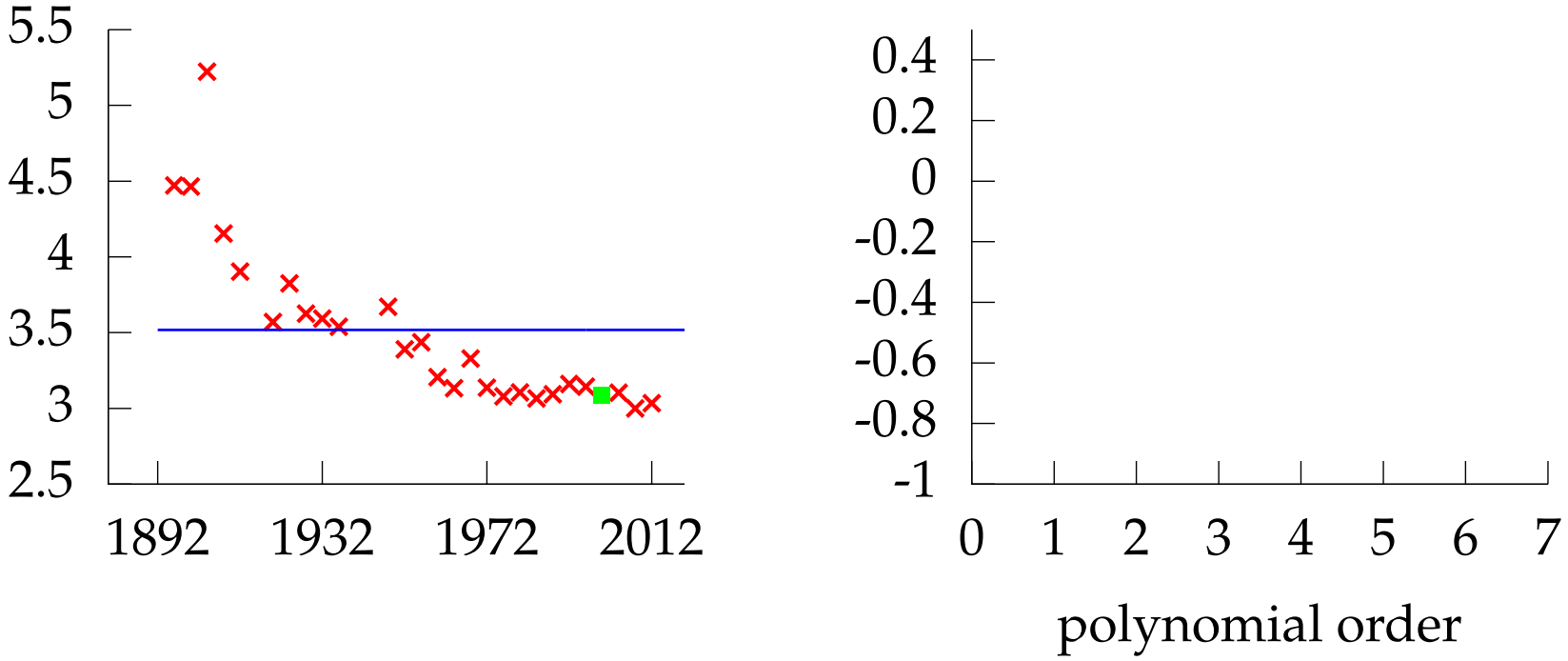
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



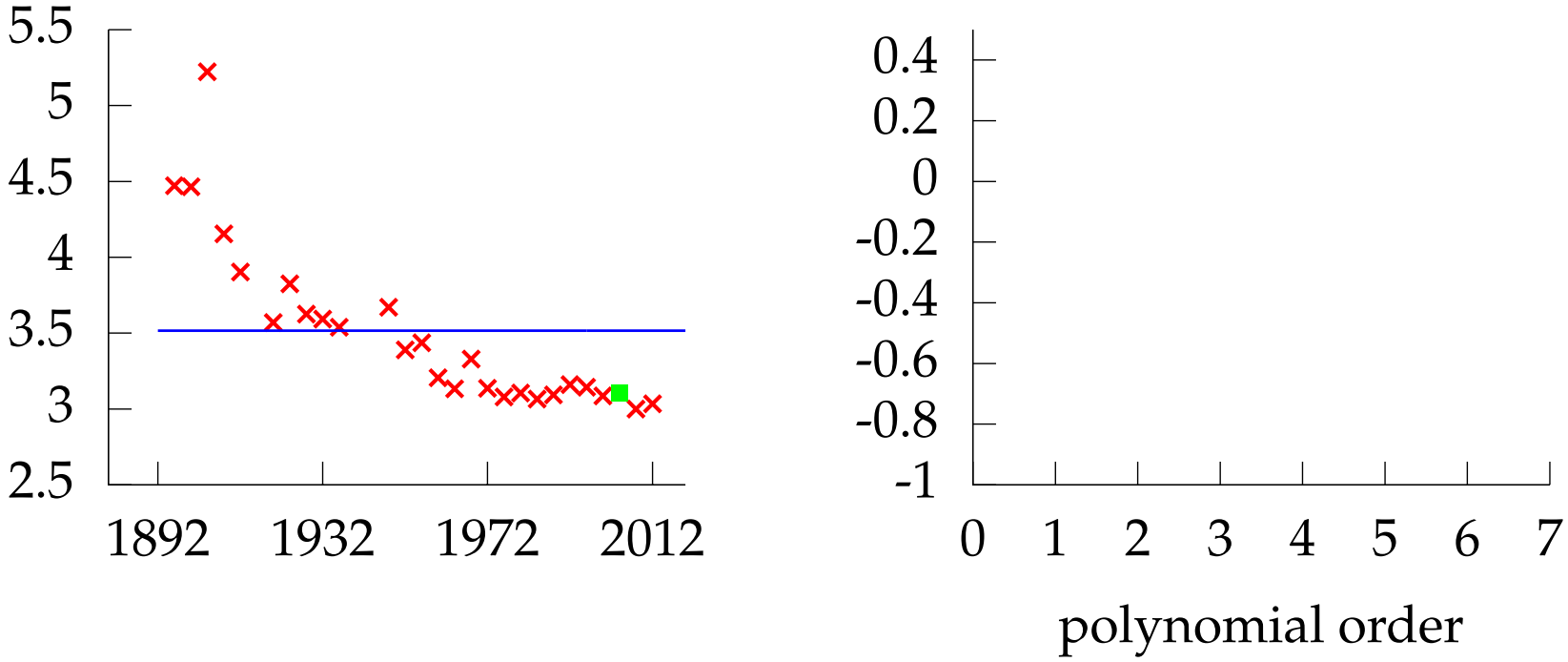
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



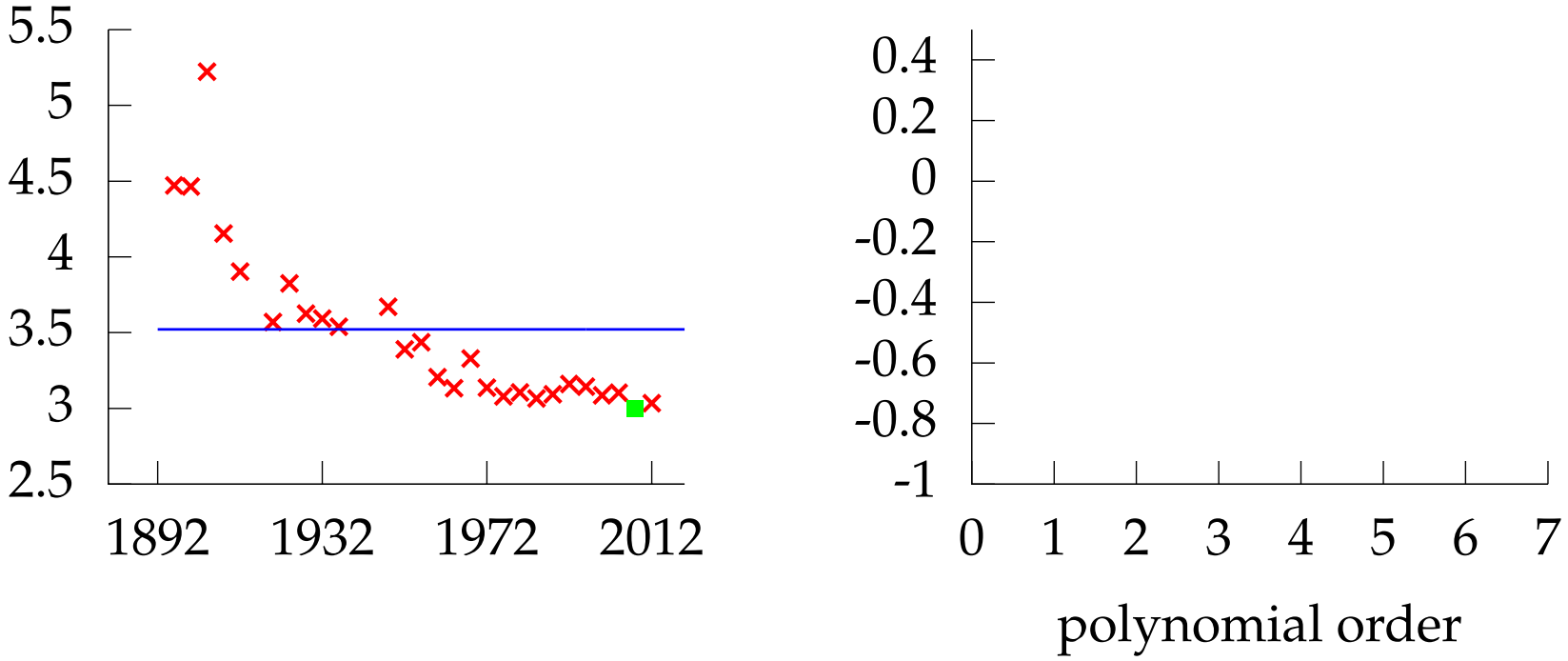
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



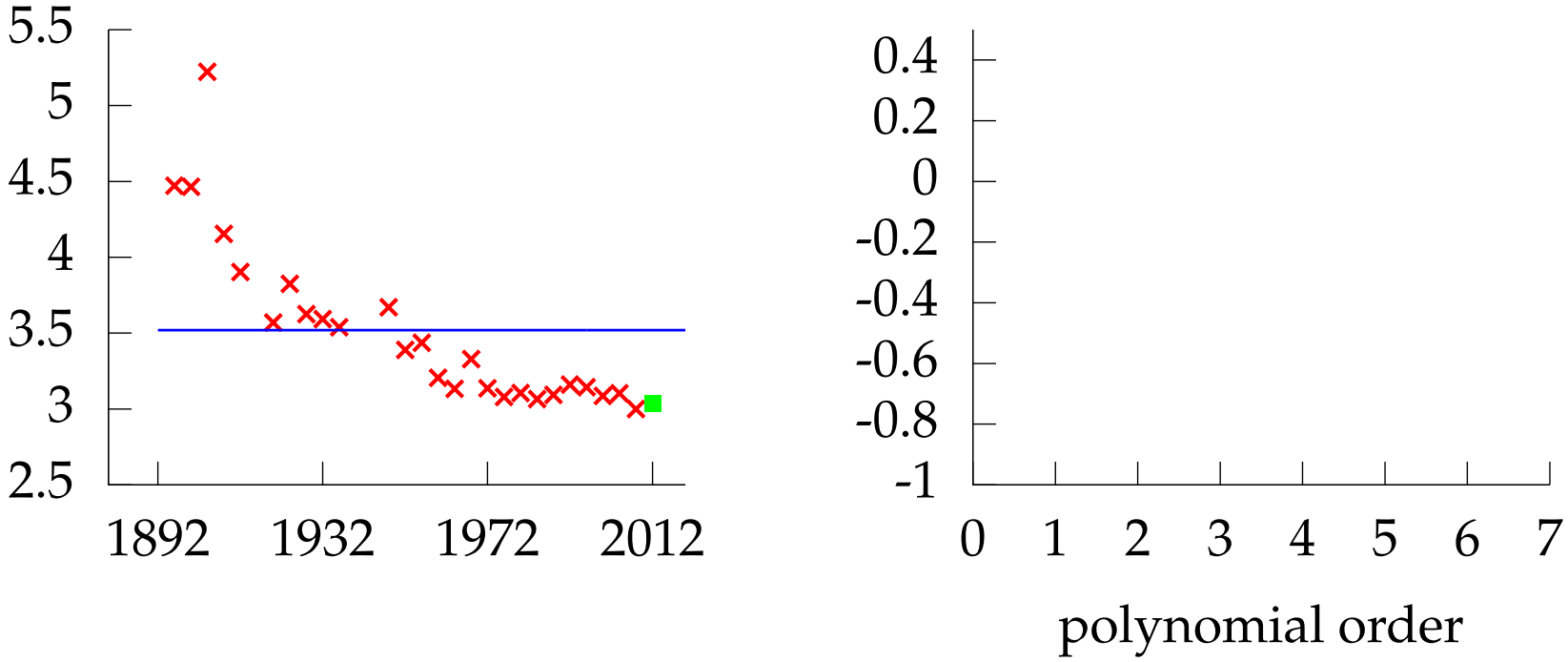
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



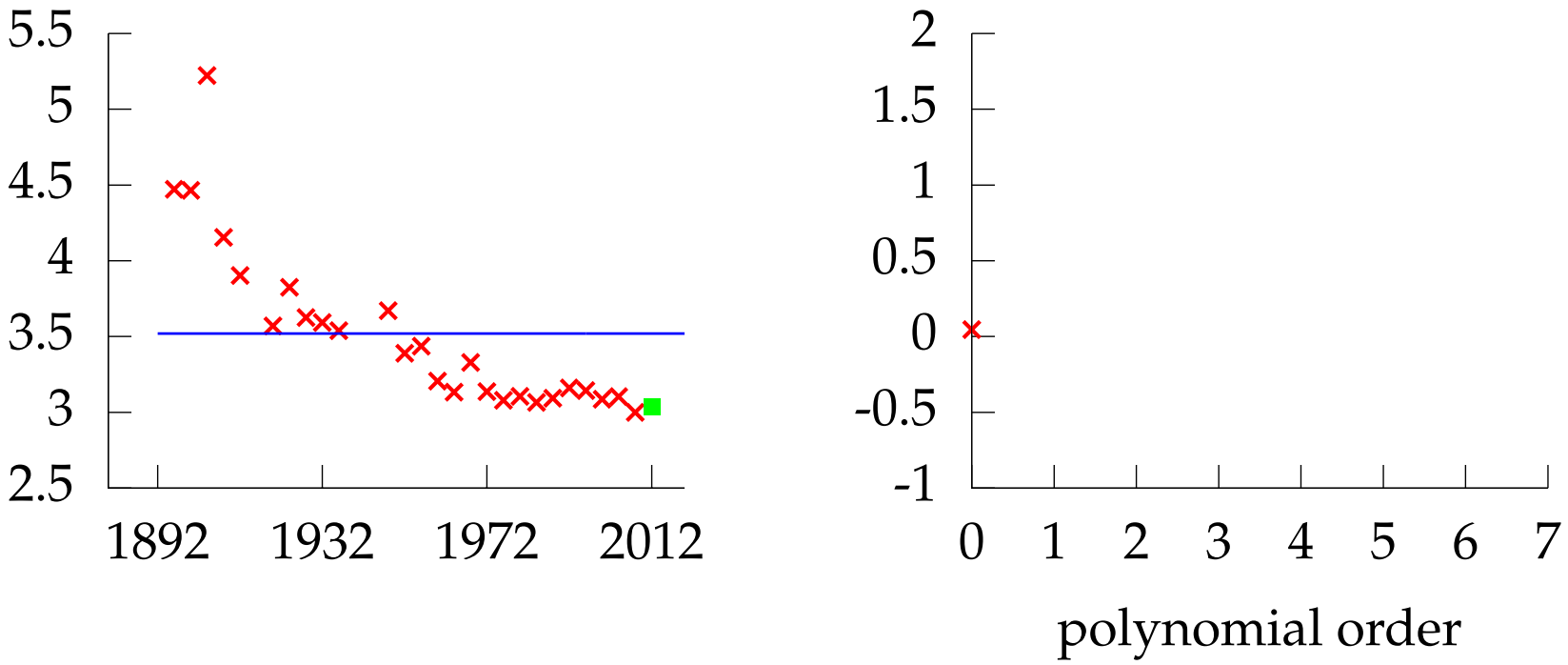
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



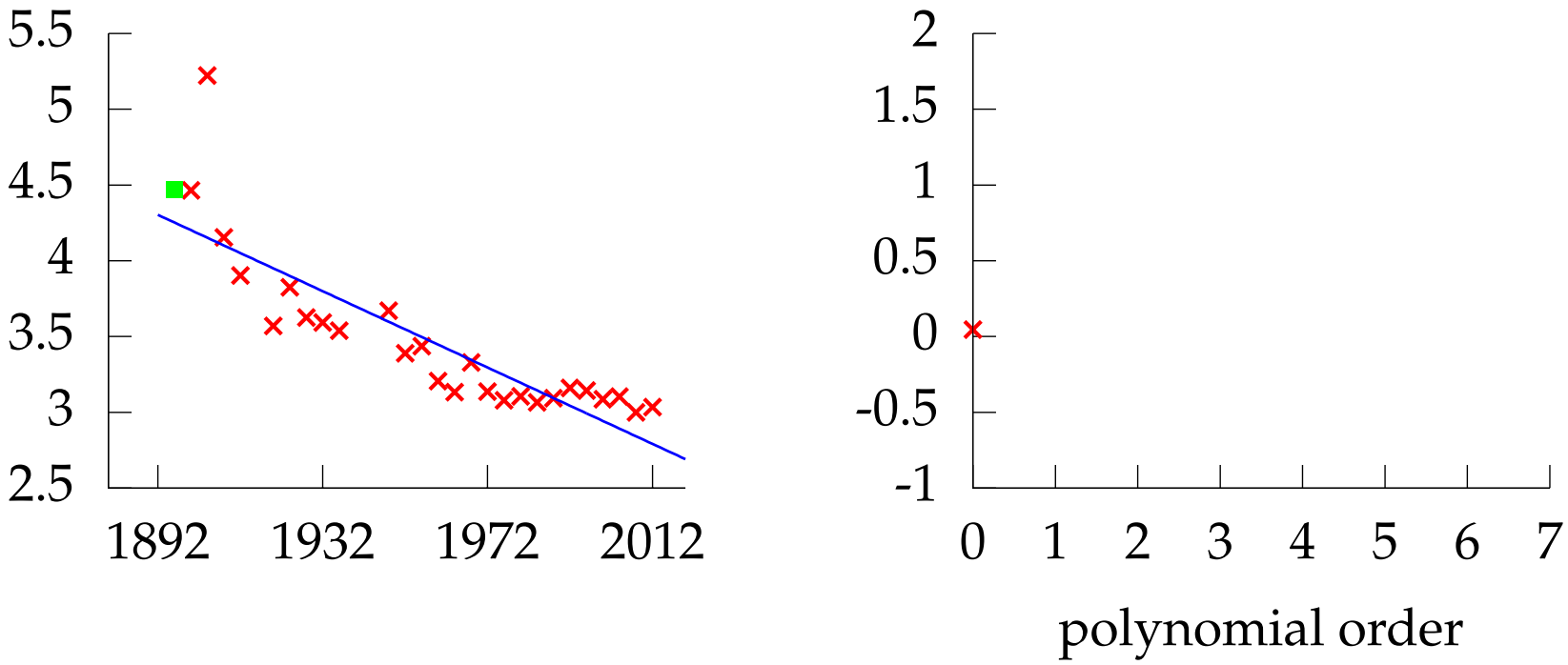
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



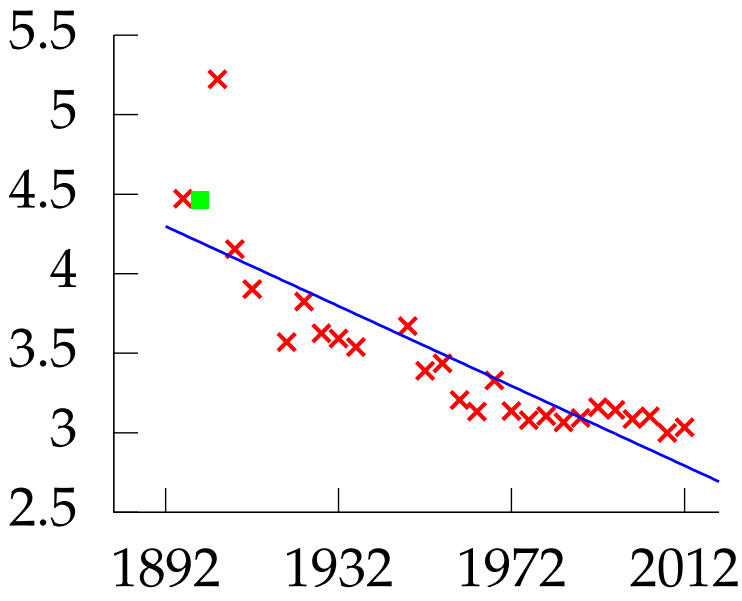
Polynomial order 0, training error -3.346, leave one out error 0.045811.

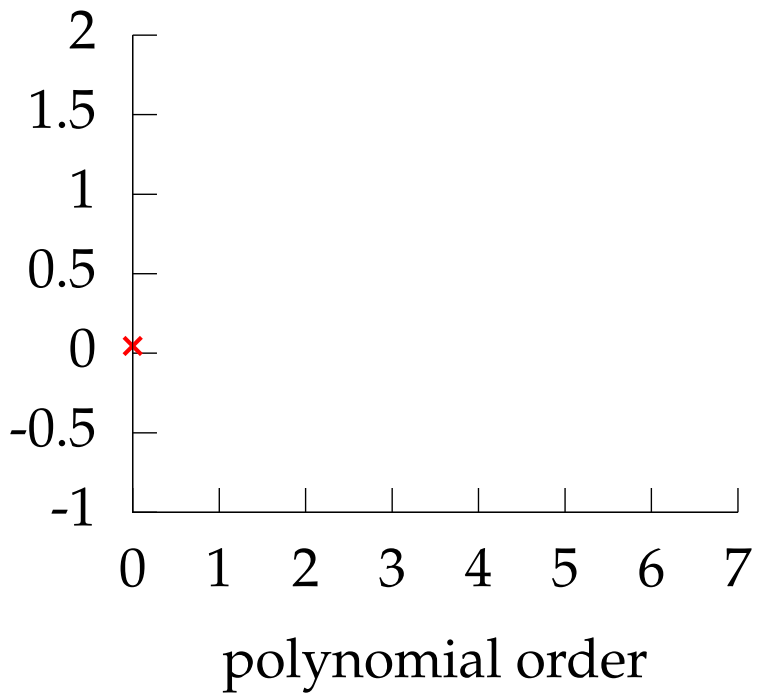
Leave One Out Error



Polynomial order 1, training error -21.183, leave one out error -0.15413.

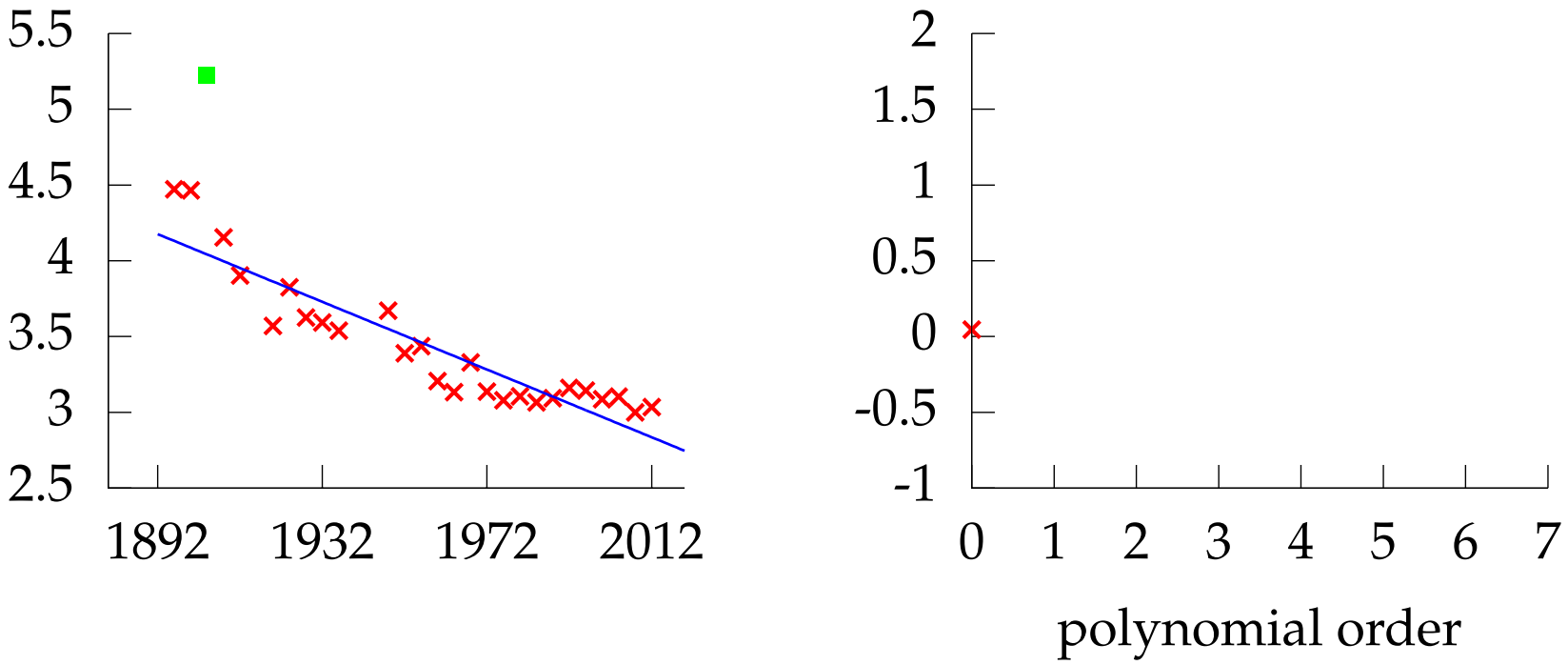
Leave One Out Error





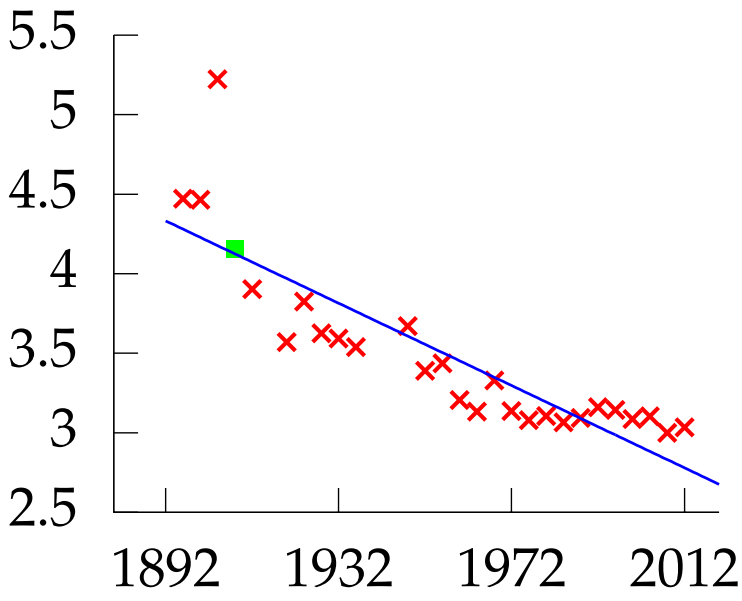
Polynomial order 1, training error -21.183, leave one out error -0.15413.

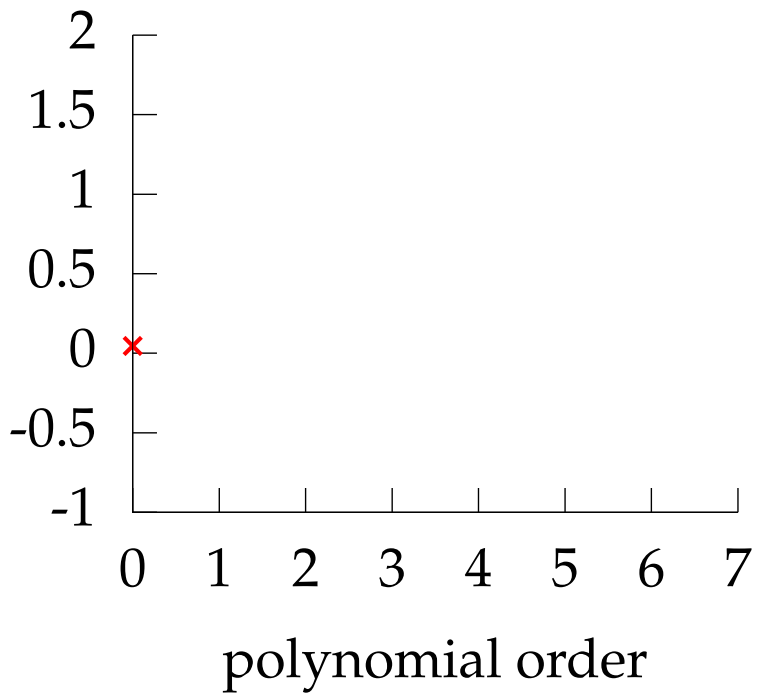
Leave One Out Error



Polynomial order 1, training error -21.183, leave one out error -0.15413.

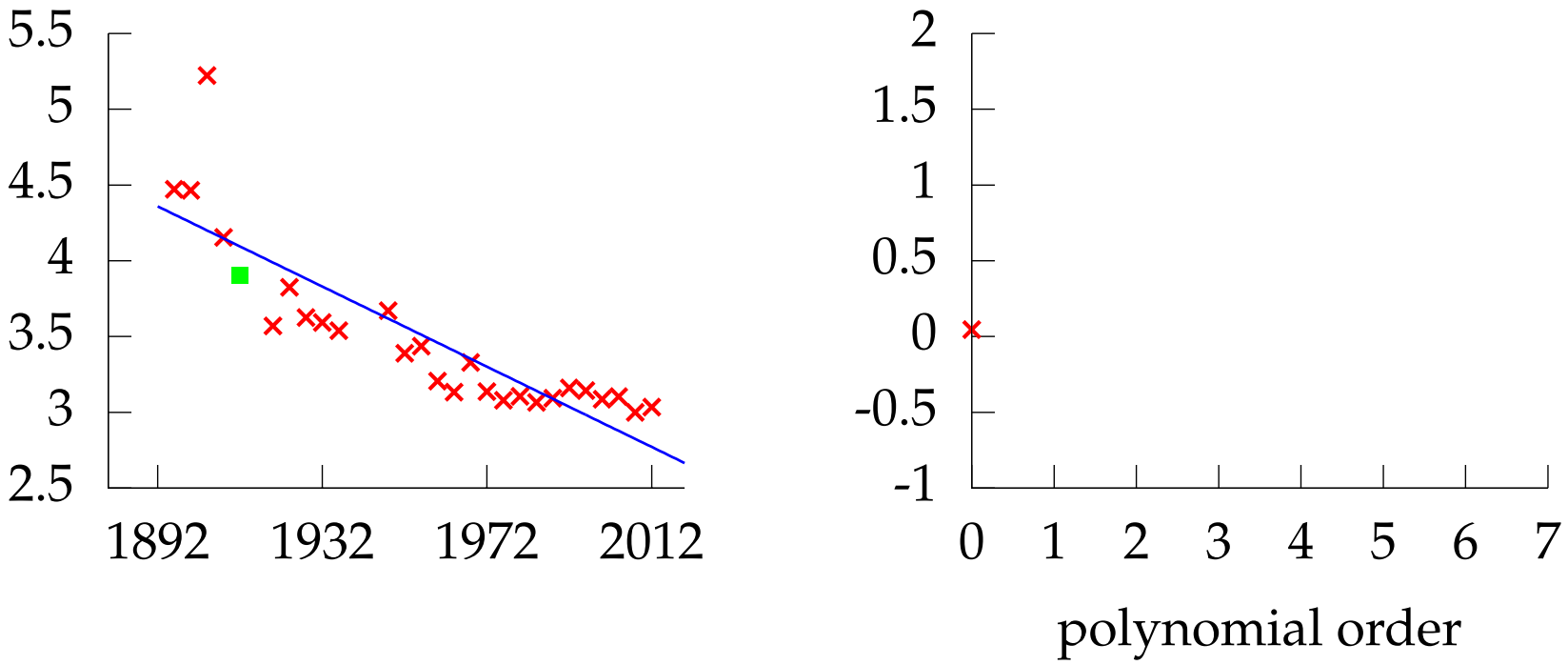
Leave One Out Error





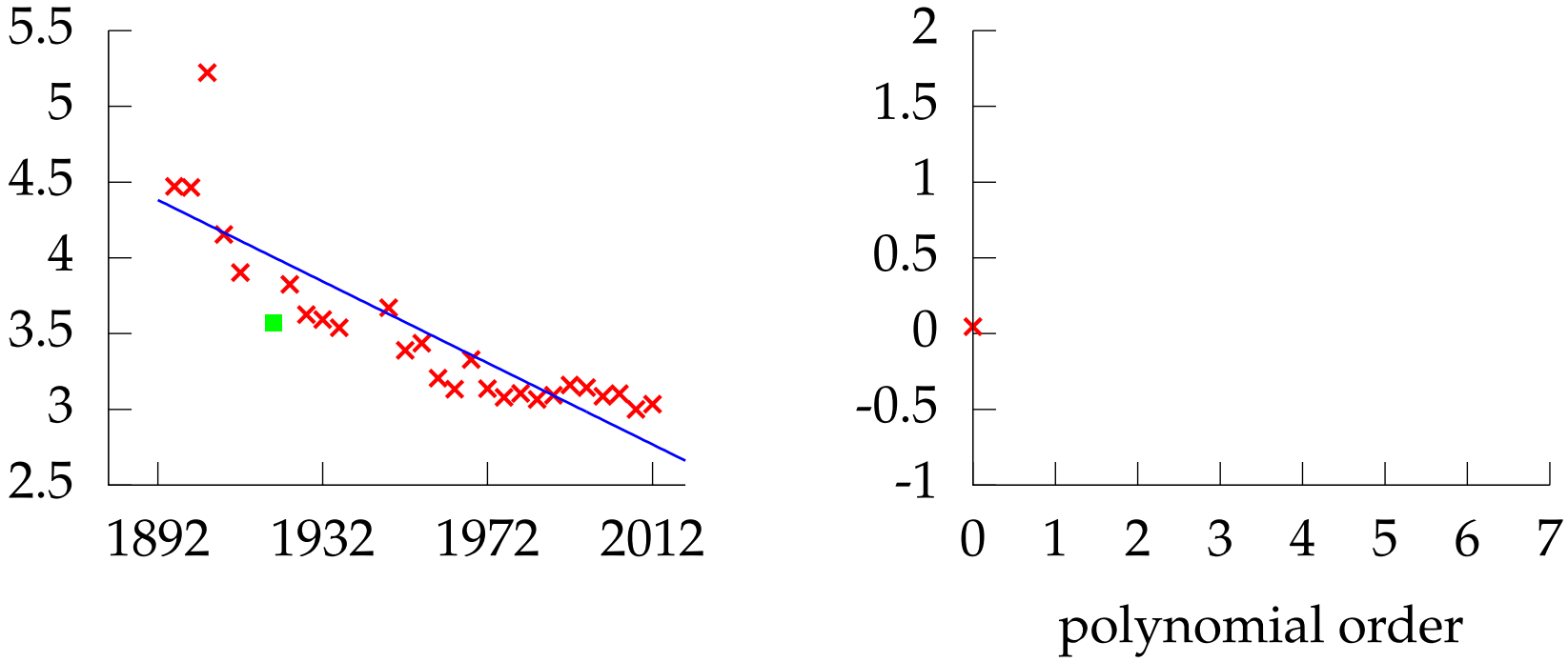
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



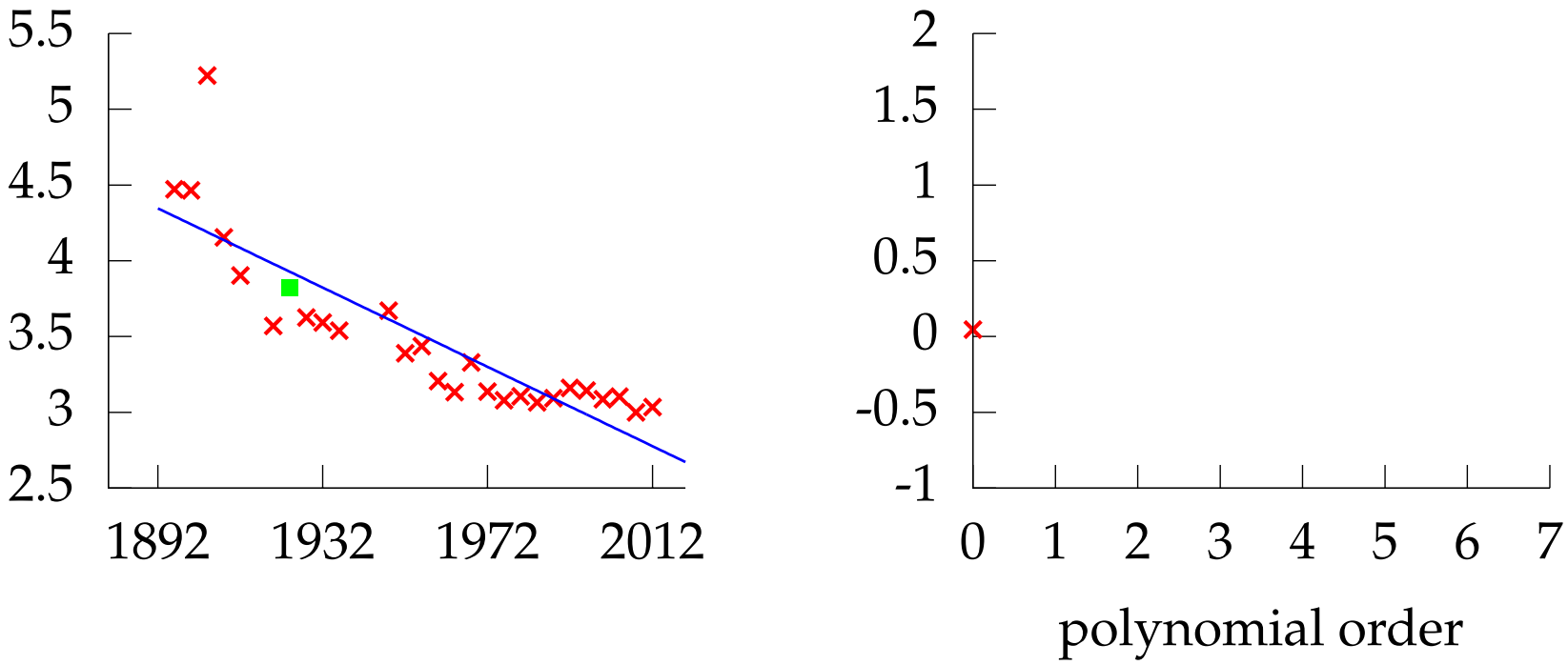
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



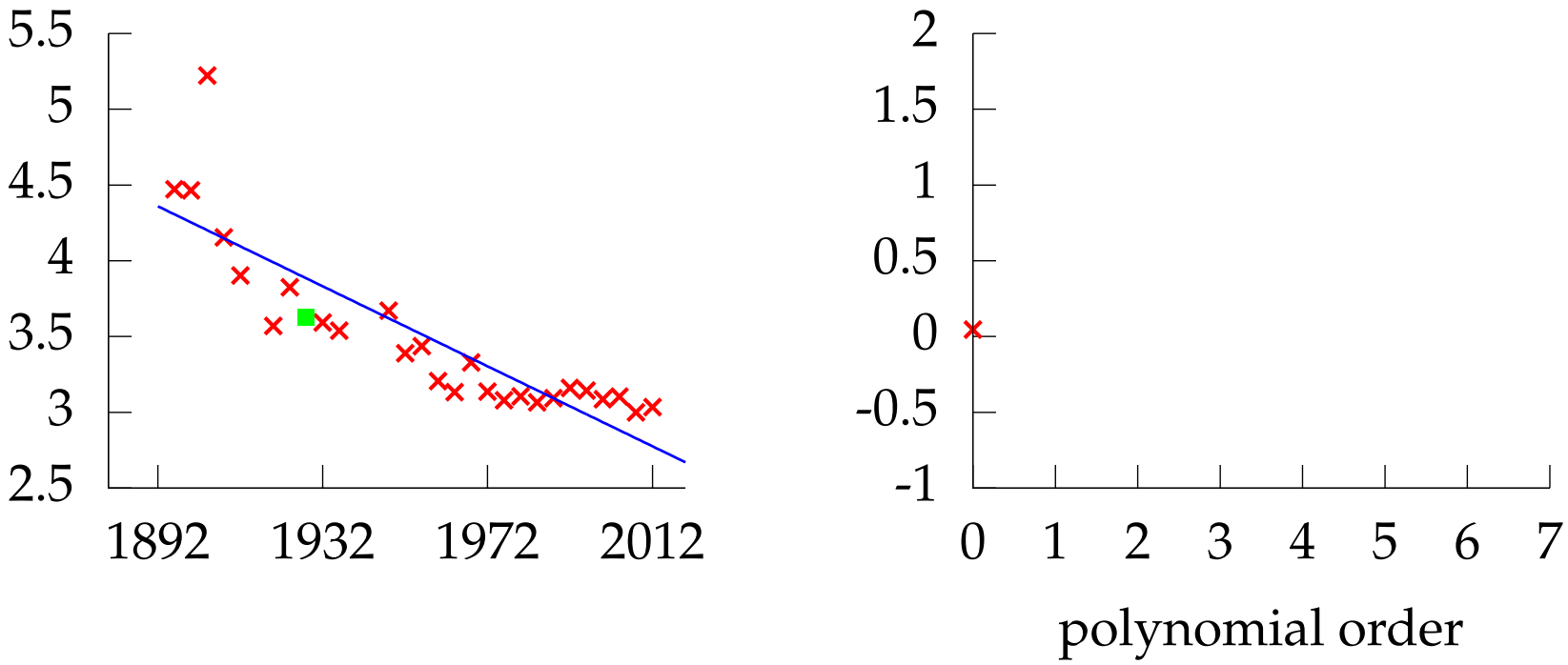
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



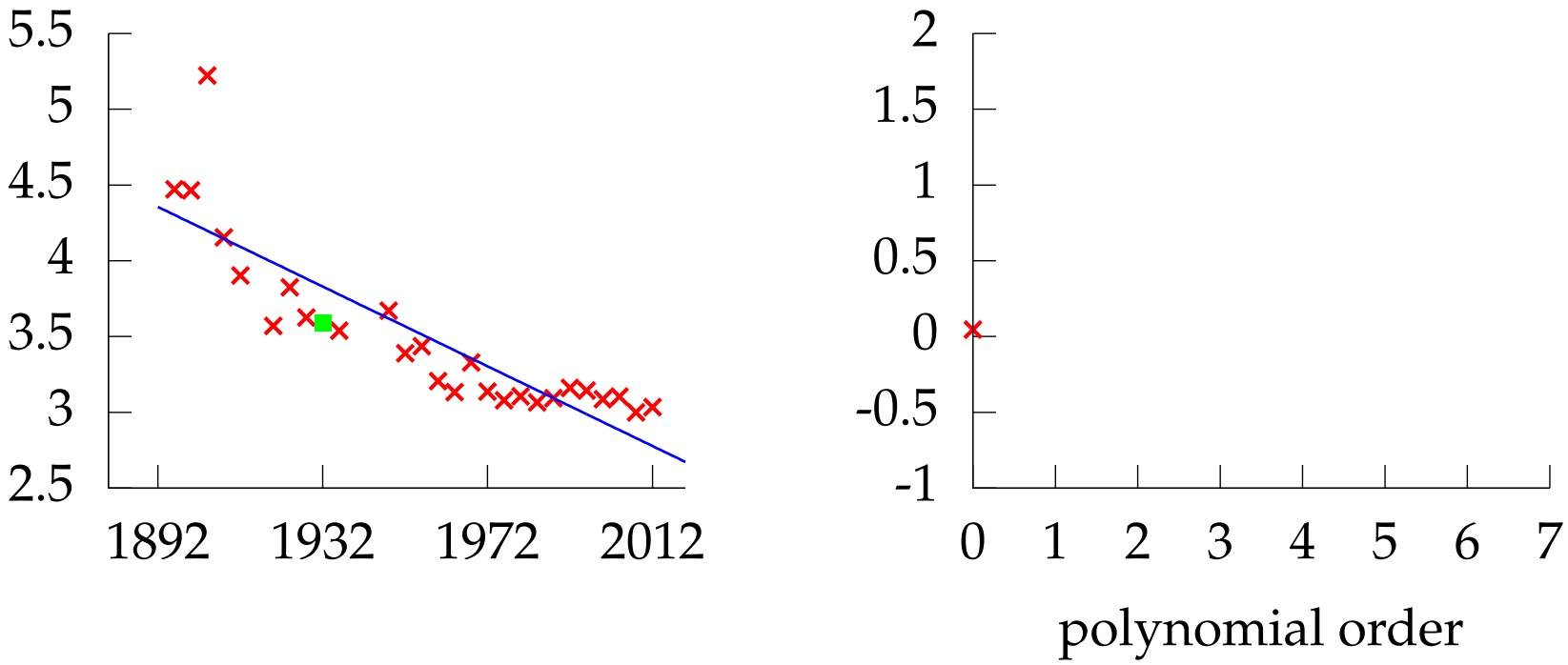
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



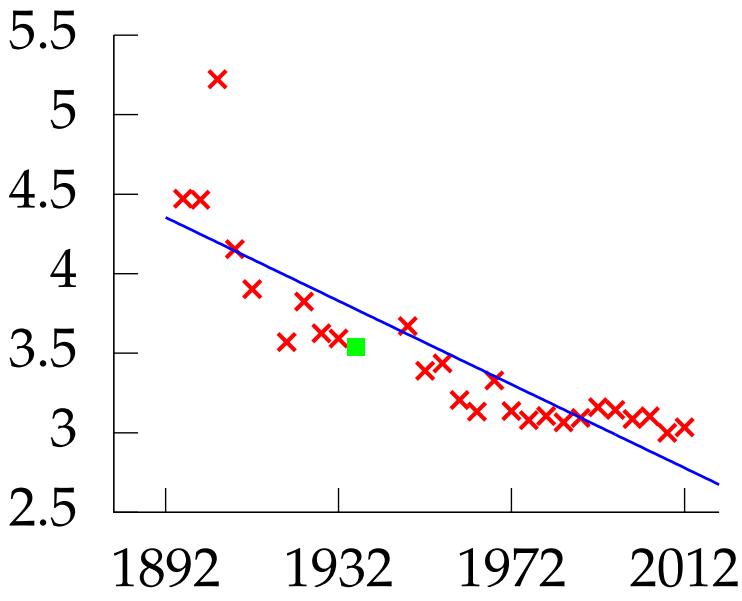
Polynomial order 1, training error -21.183, leave one out error -0.15413.

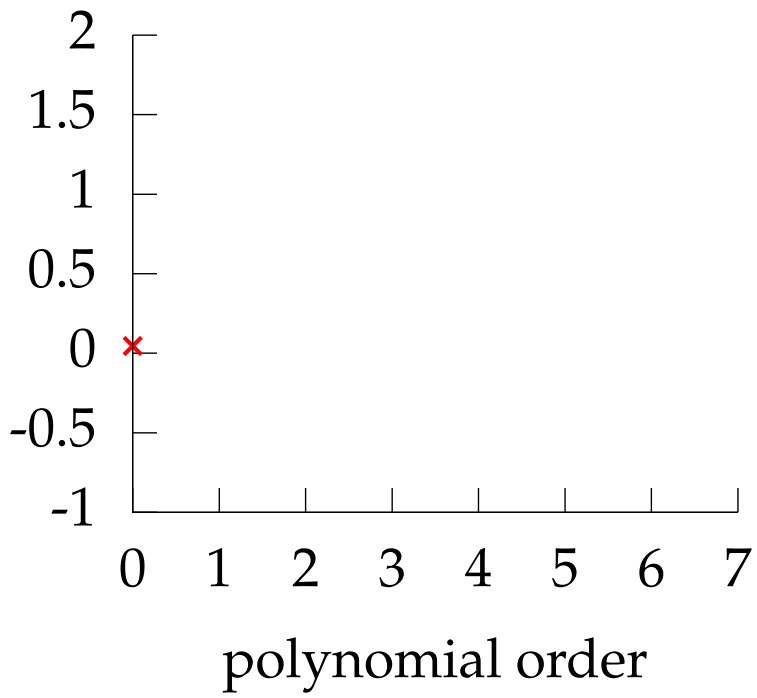
Leave One Out Error



Polynomial order 1, training error -21.183, leave one out error -0.15413.

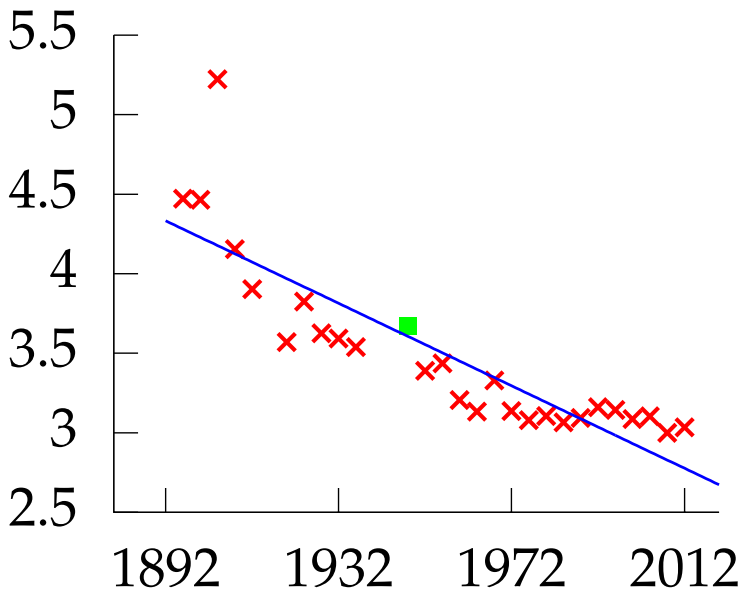
Leave One Out Error

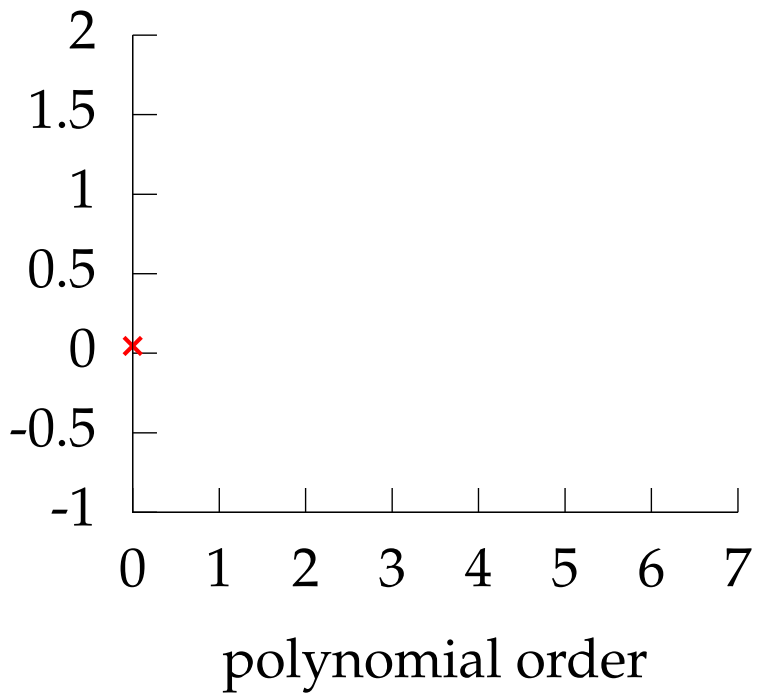




Polynomial order 1, training error -21.183, leave one out error -0.15413.

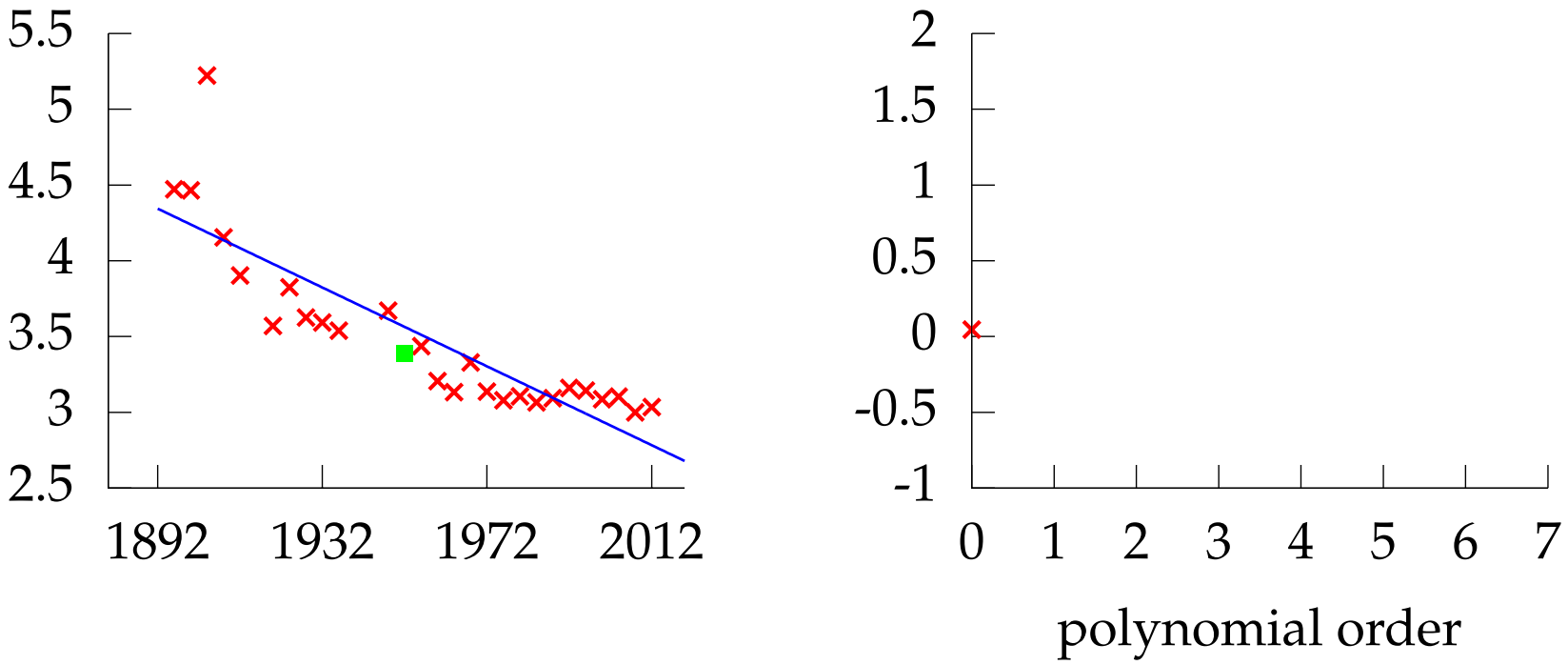
Leave One Out Error





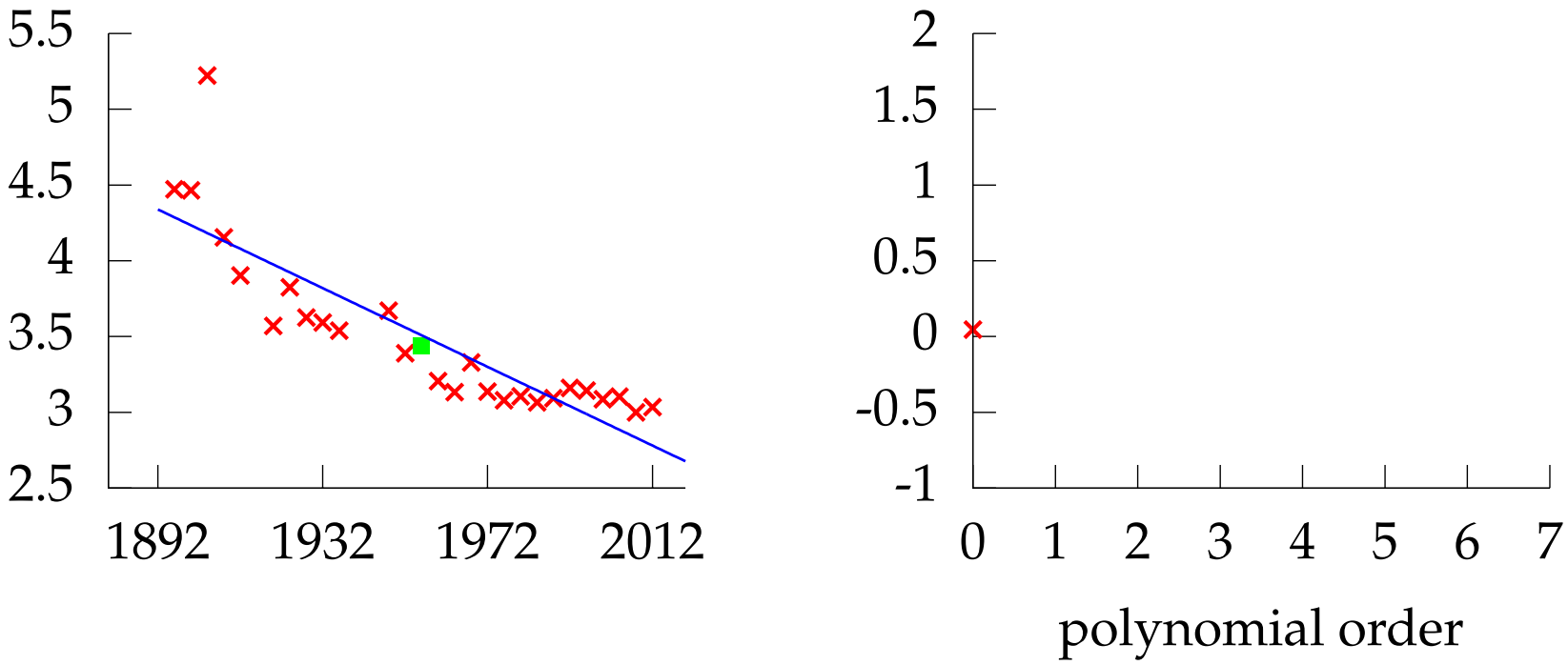
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



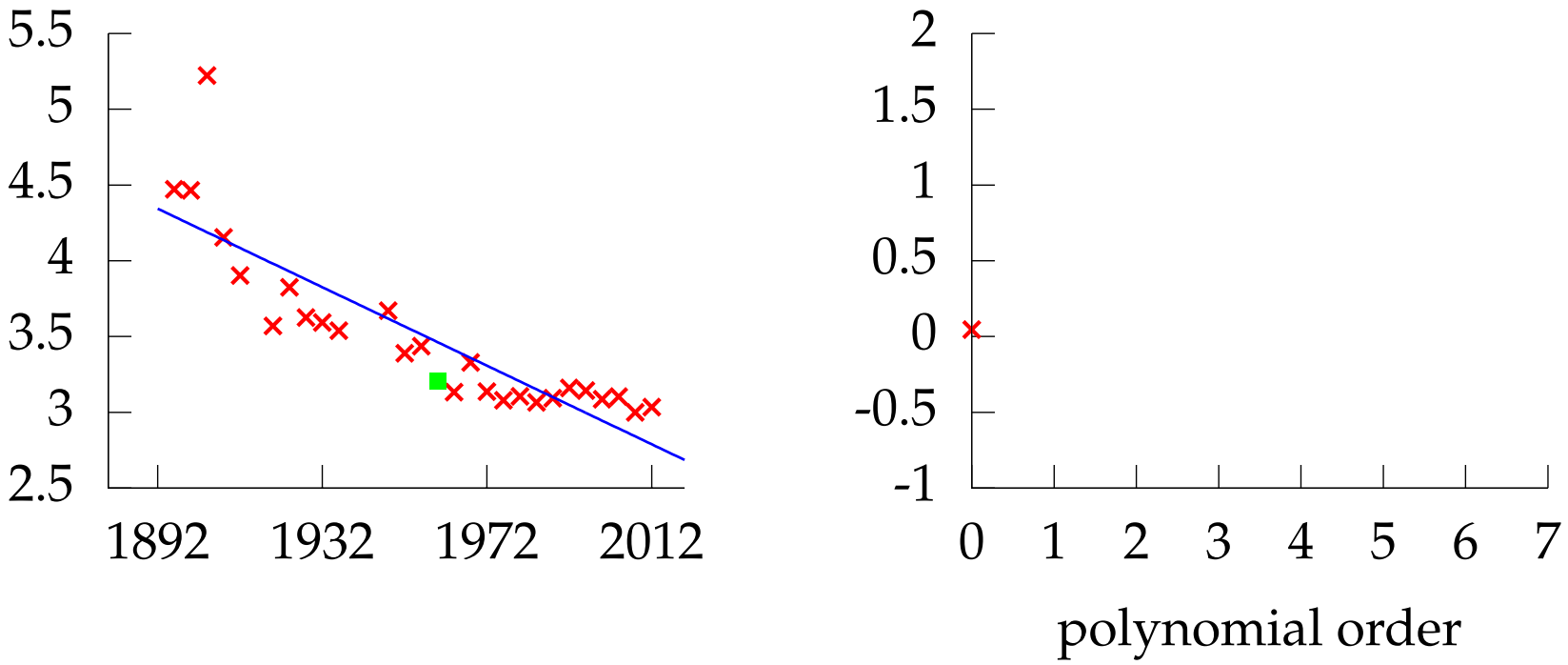
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



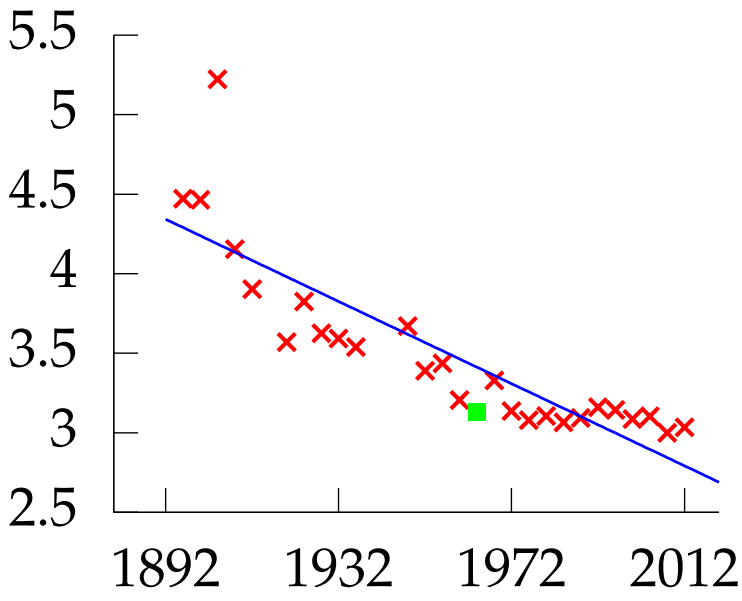
Polynomial order 1, training error -21.183, leave one out error -0.15413.

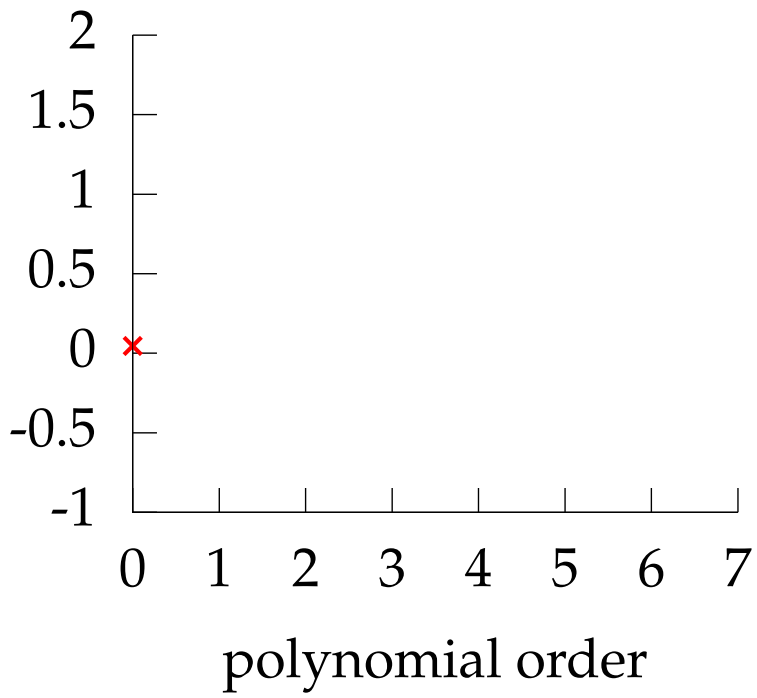
Leave One Out Error



Polynomial order 1, training error -21.183, leave one out error -0.15413.

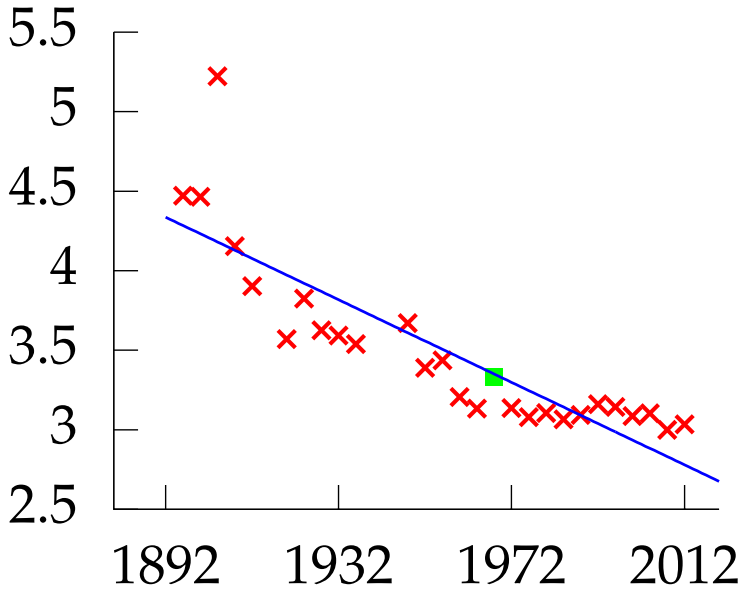
Leave One Out Error

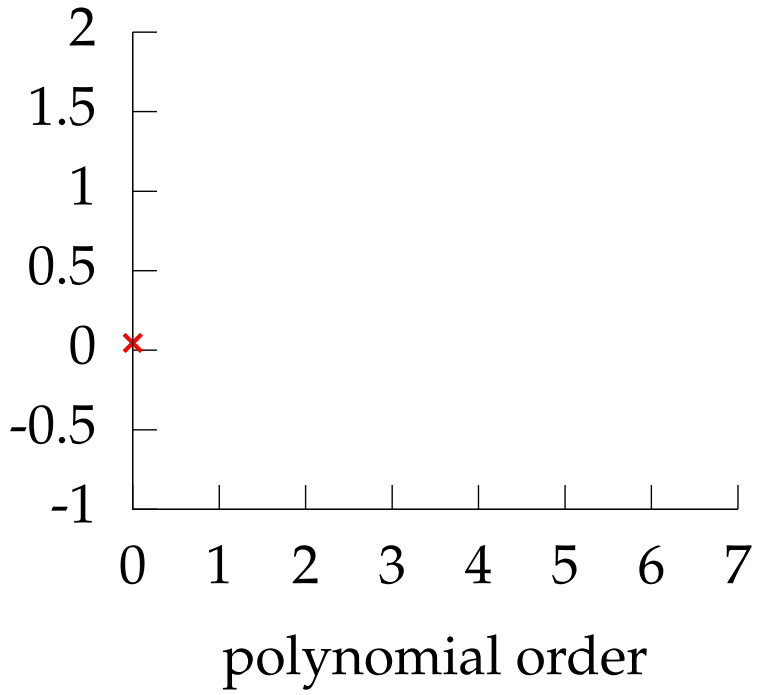




Polynomial order 1, training error -21.183, leave one out error -0.15413.

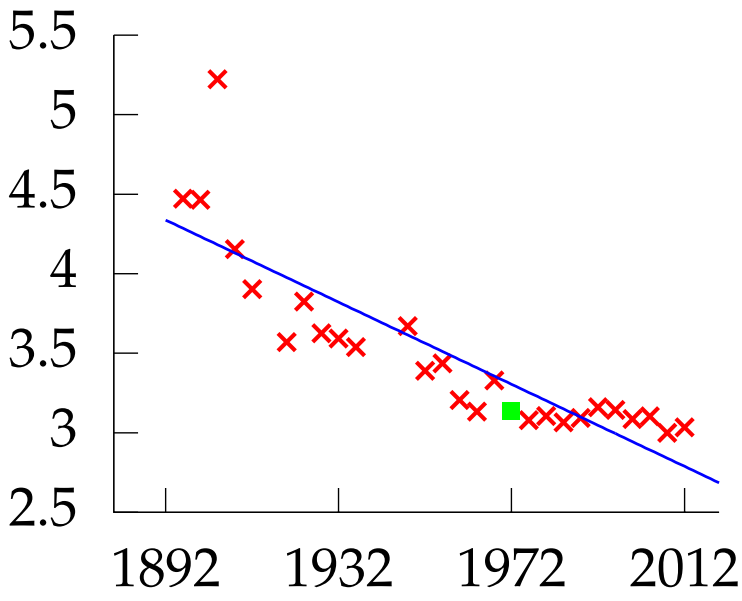
Leave One Out Error

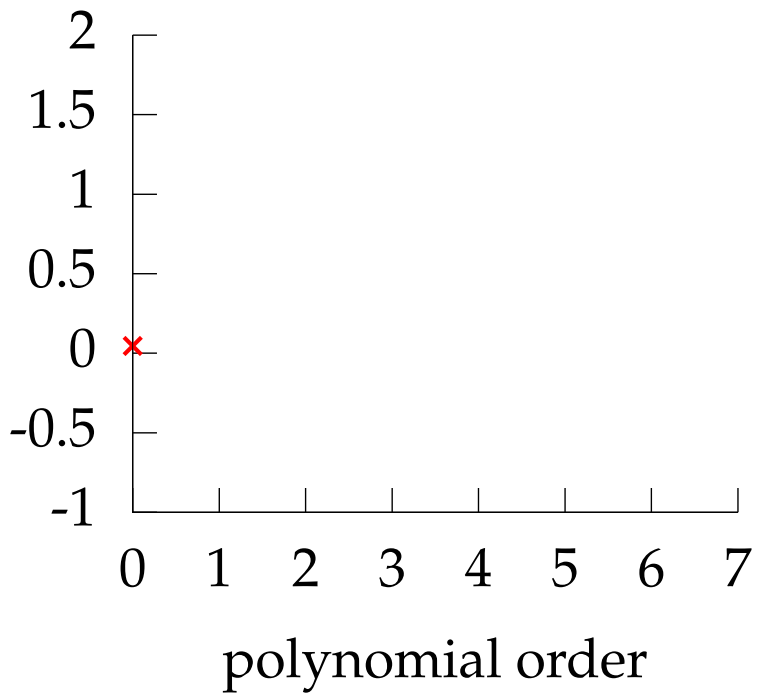




Polynomial order 1, training error -21.183, leave one out error -0.15413.

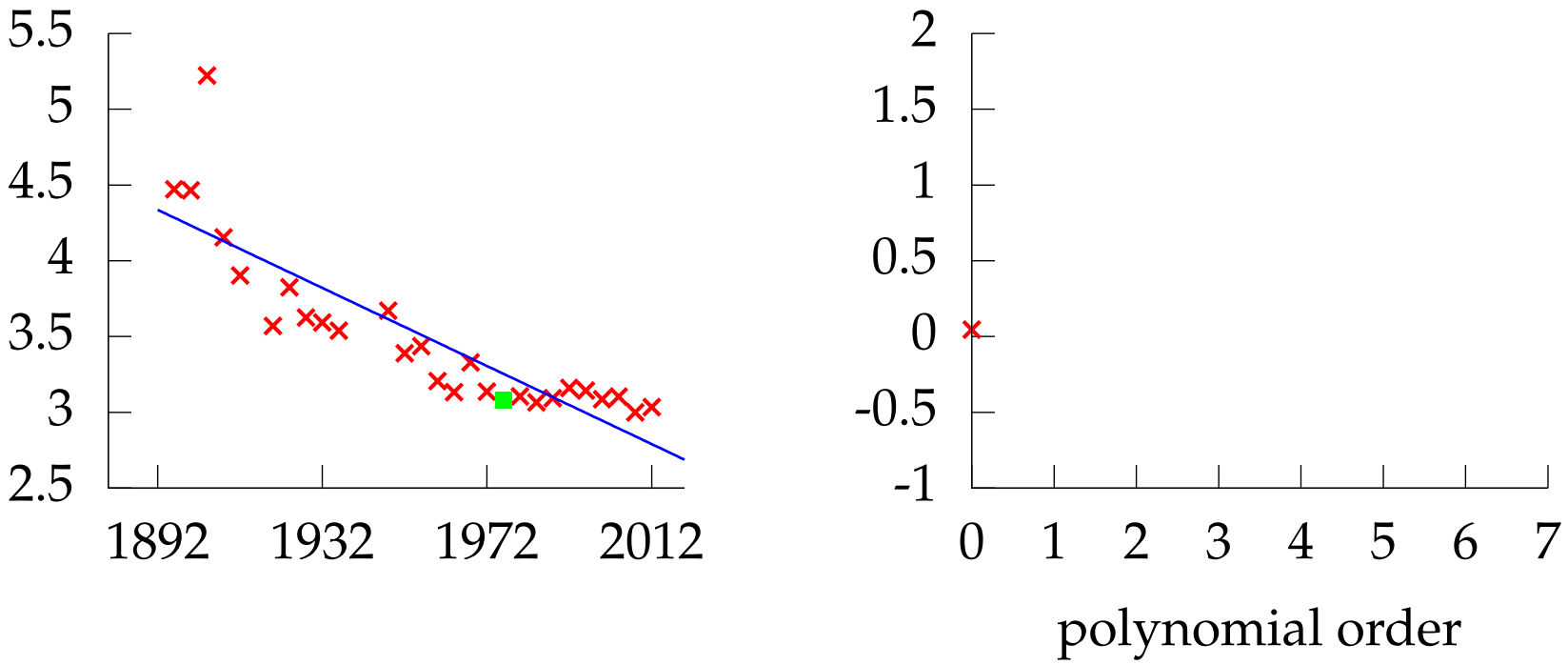
Leave One Out Error





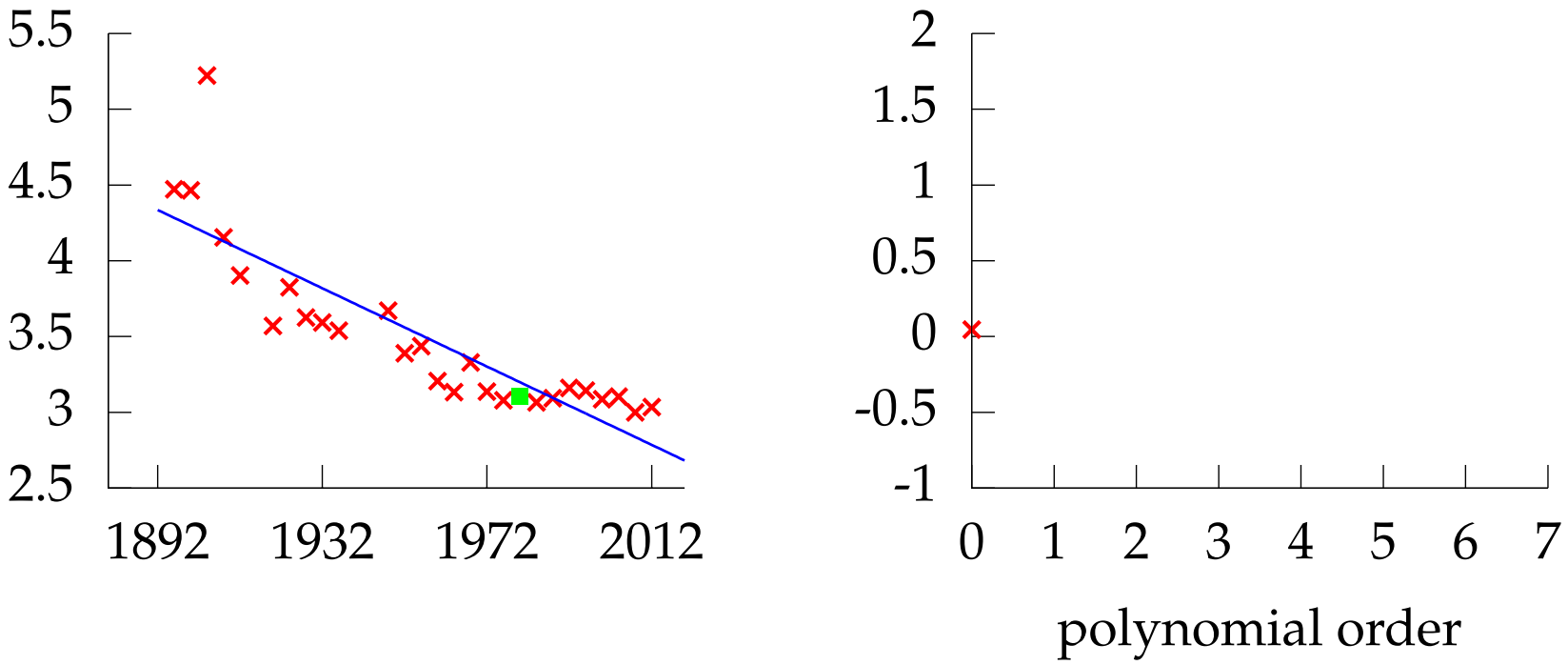
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



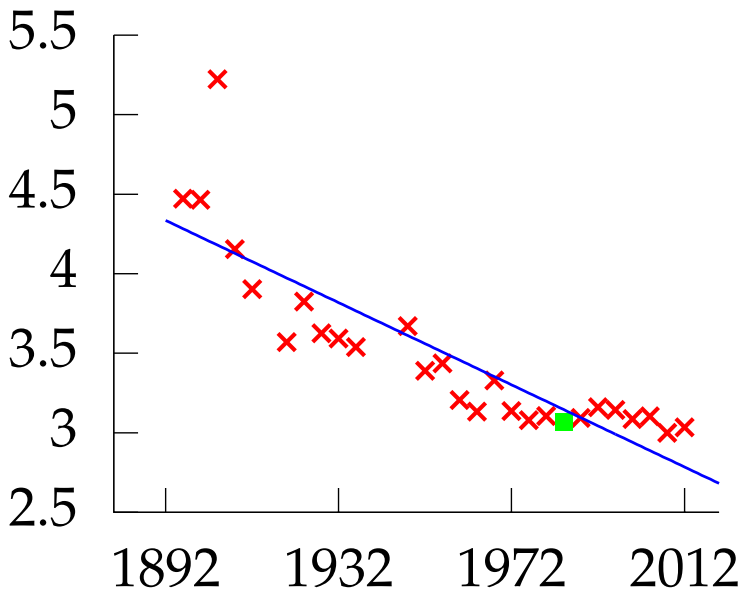
Polynomial order 1, training error -21.183, leave one out error -0.15413.

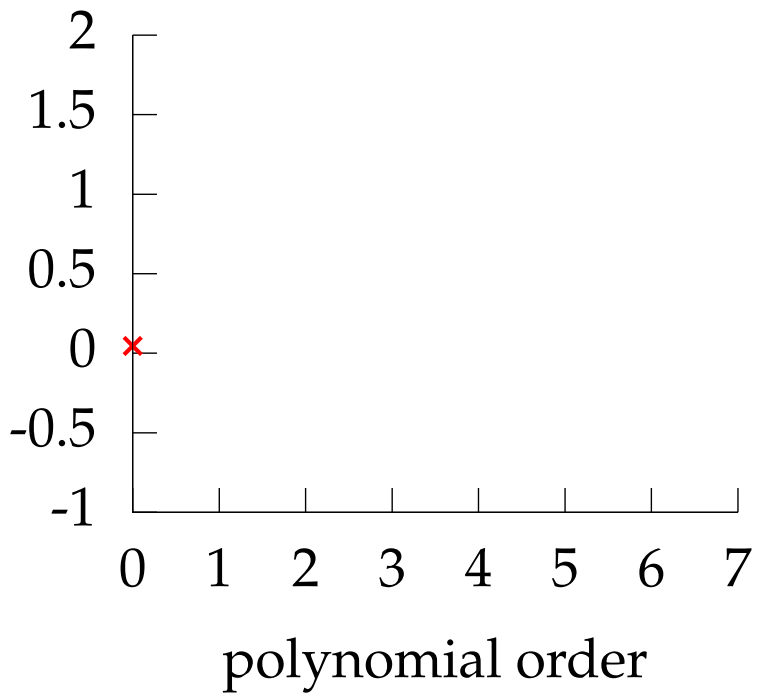
Leave One Out Error



Polynomial order 1, training error -21.183, leave one out error -0.15413.

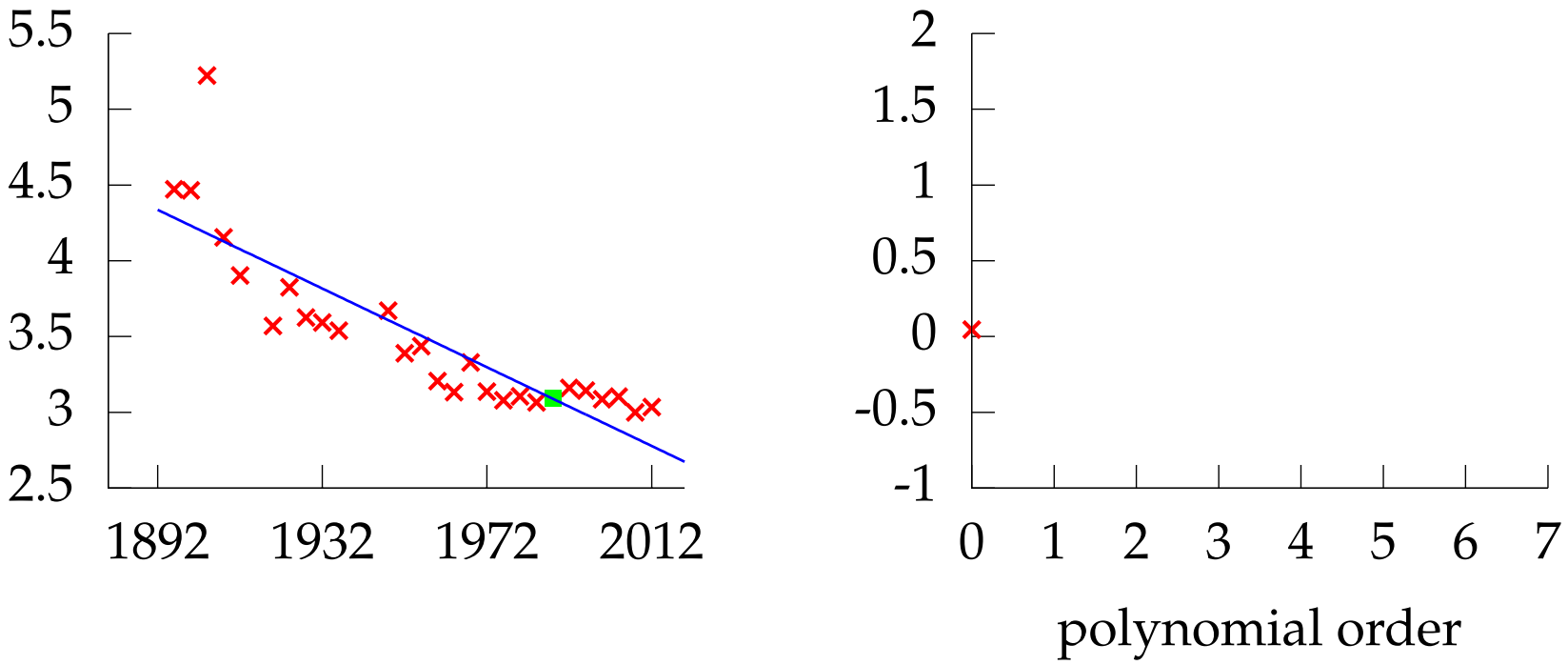
Leave One Out Error





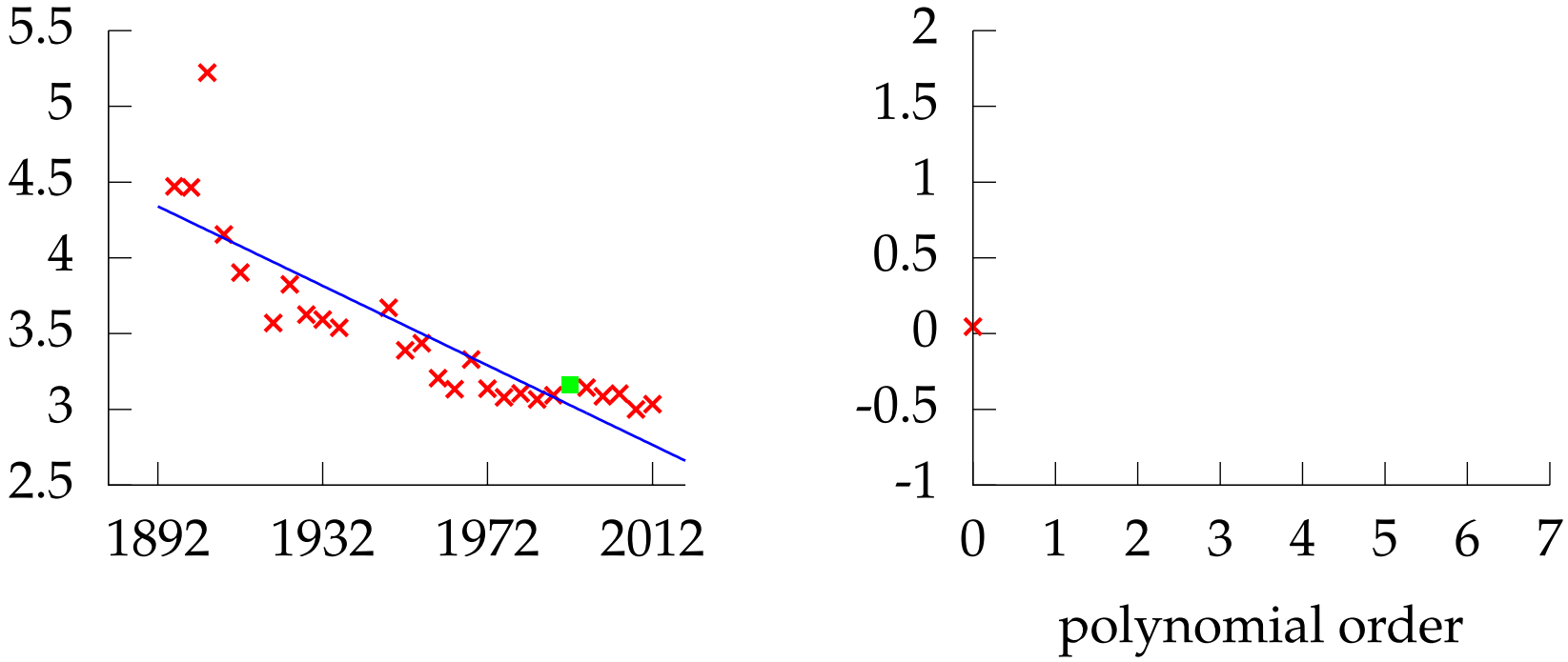
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



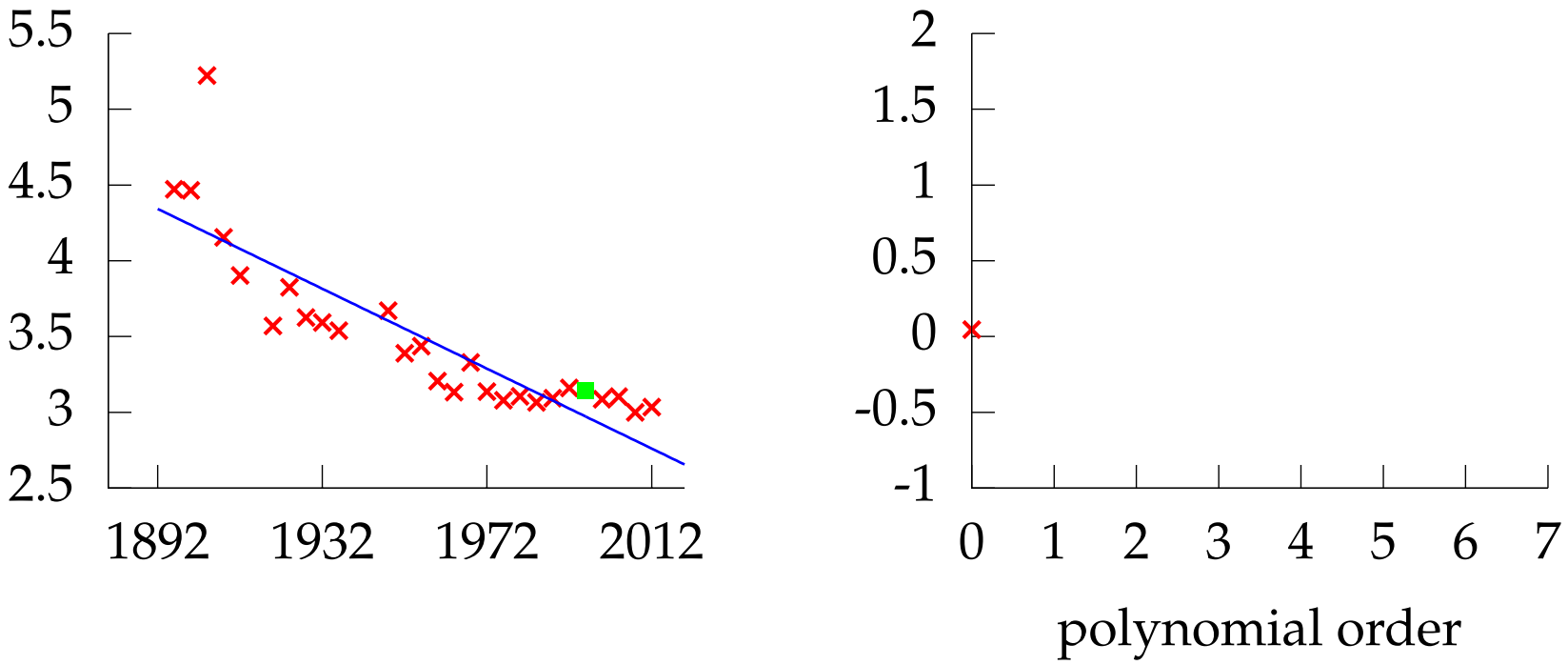
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



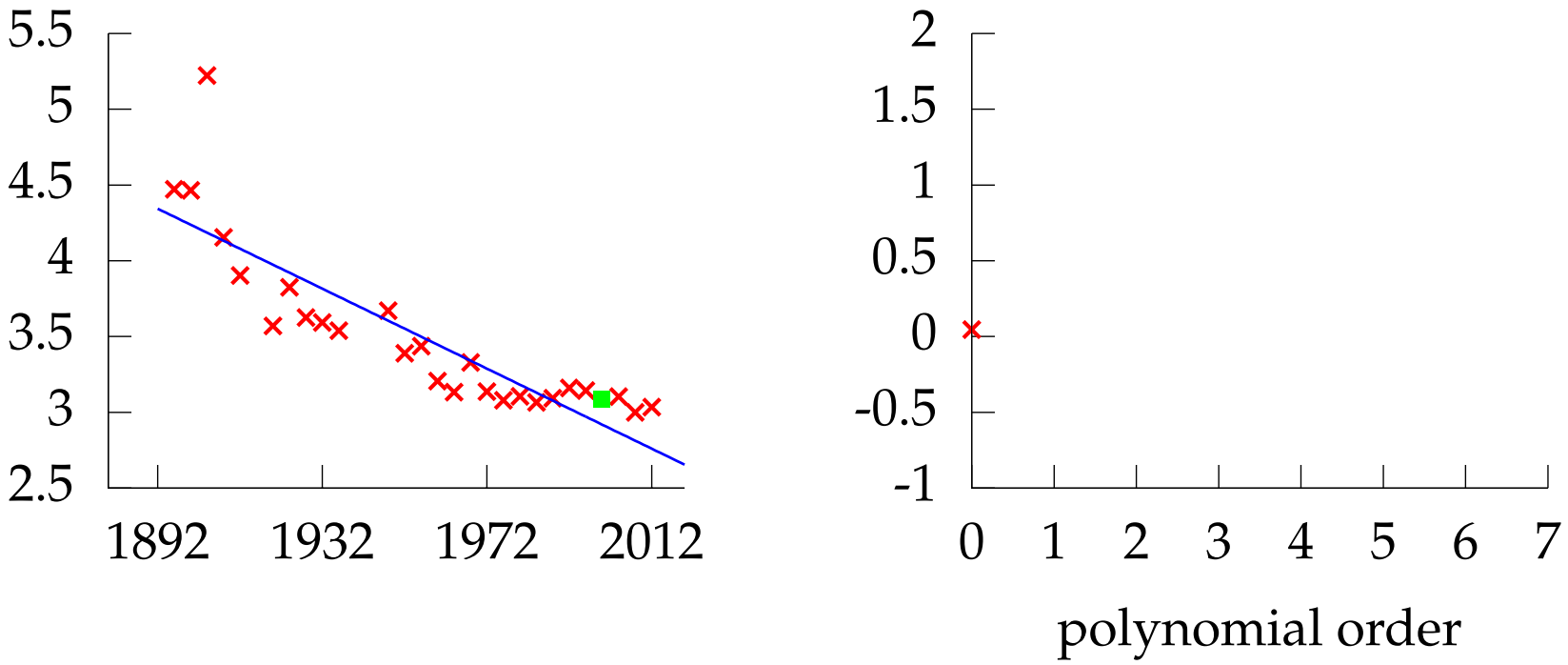
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



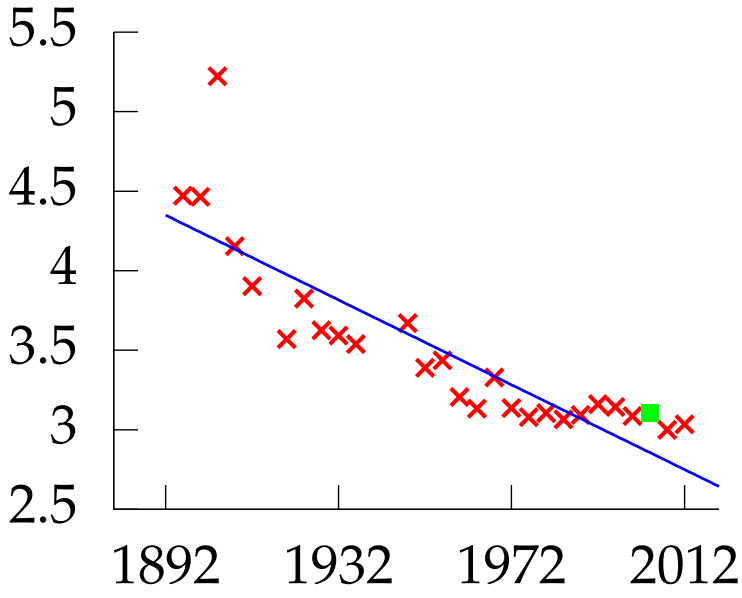
Polynomial order 1, training error -21.183, leave one out error -0.15413.

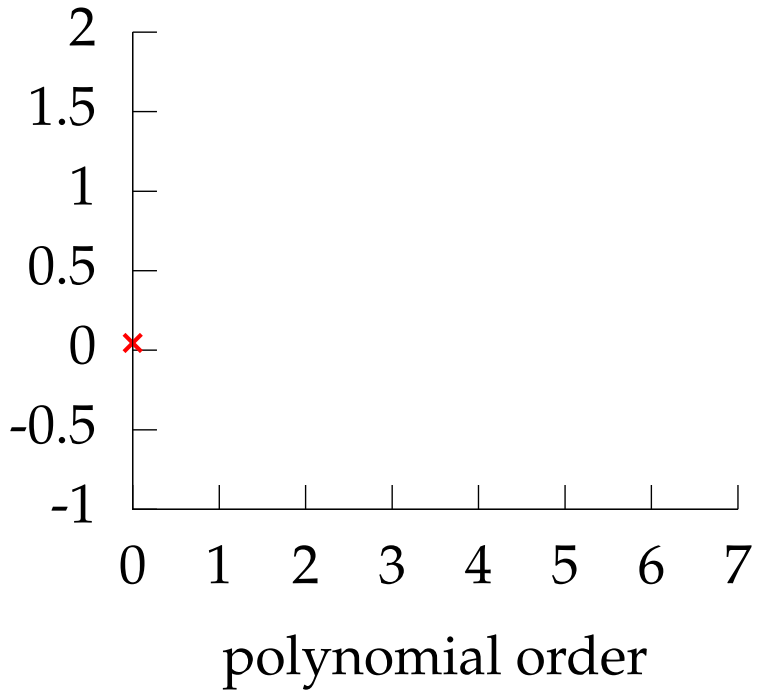
Leave One Out Error



Polynomial order 1, training error -21.183, leave one out error -0.15413.

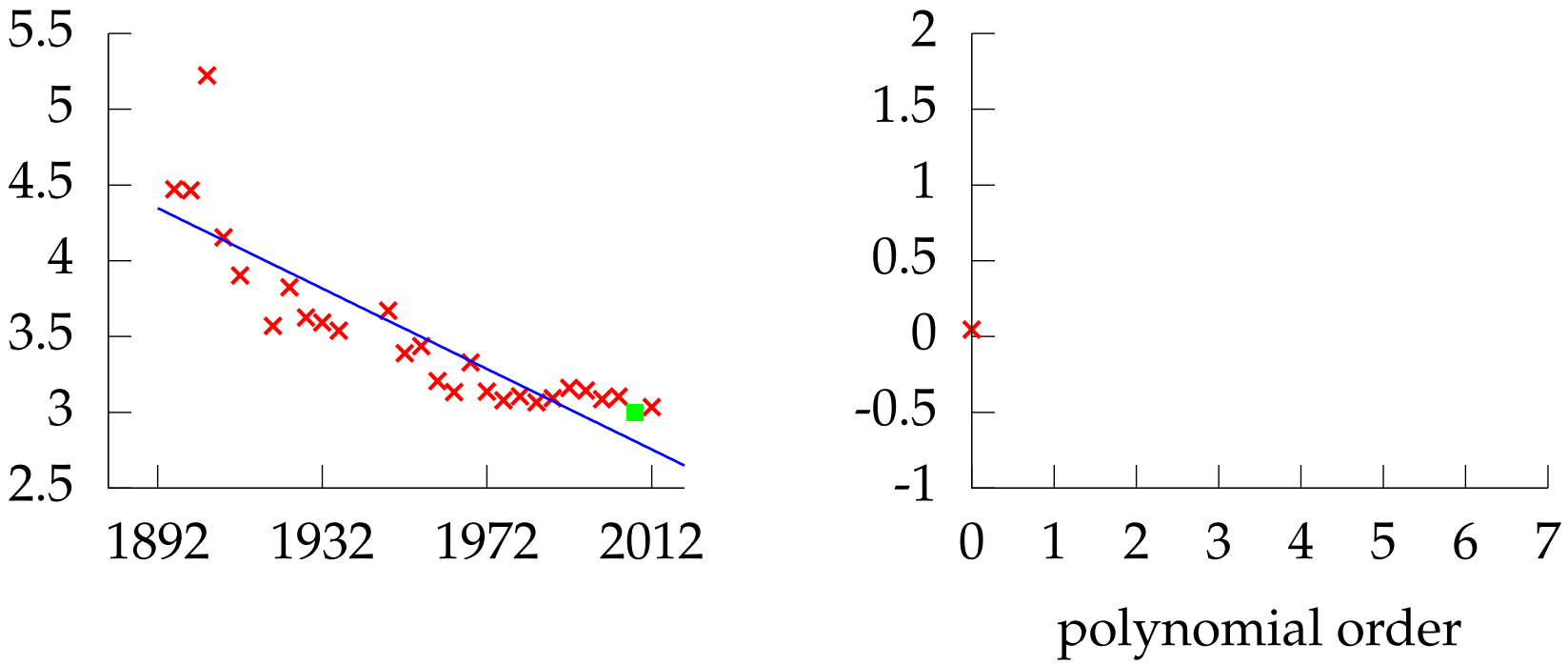
Leave One Out Error





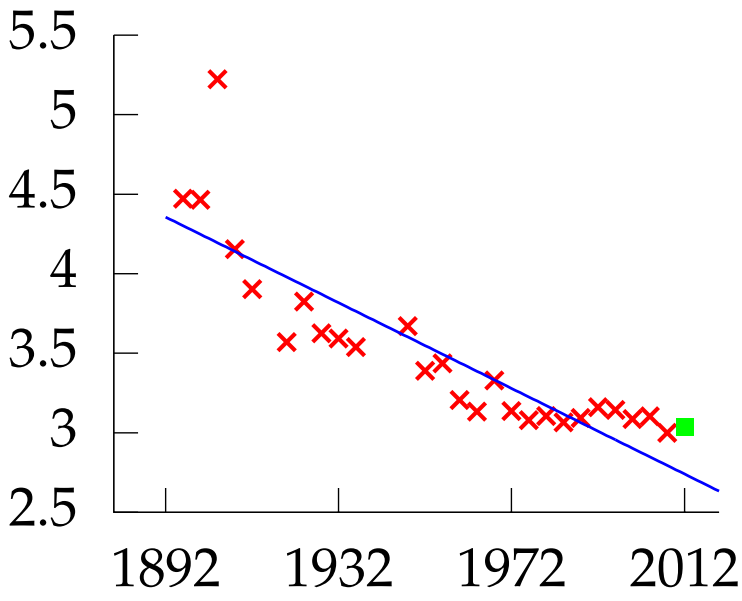
Polynomial order 1, training error -21.183, leave one out error -0.15413.

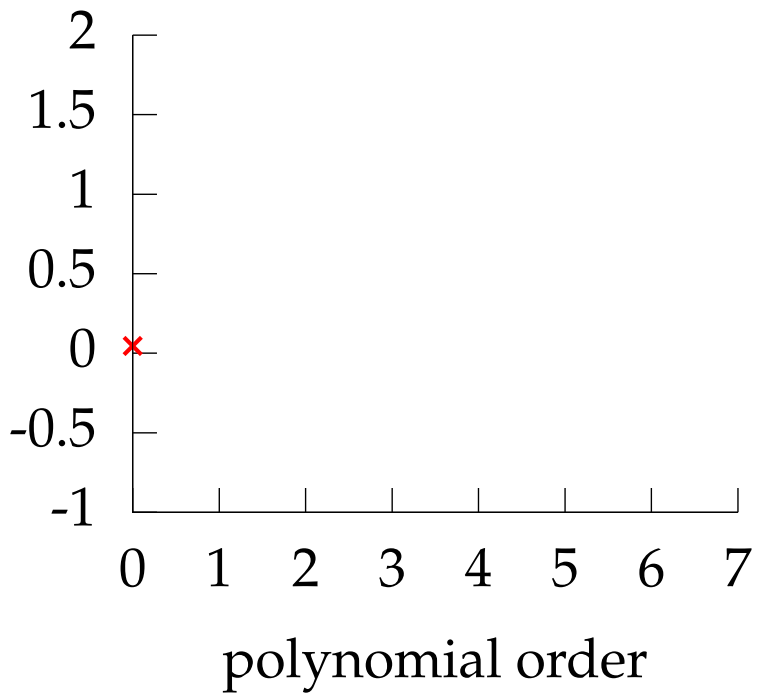
Leave One Out Error



Polynomial order 1, training error -21.183, leave one out error -0.15413.

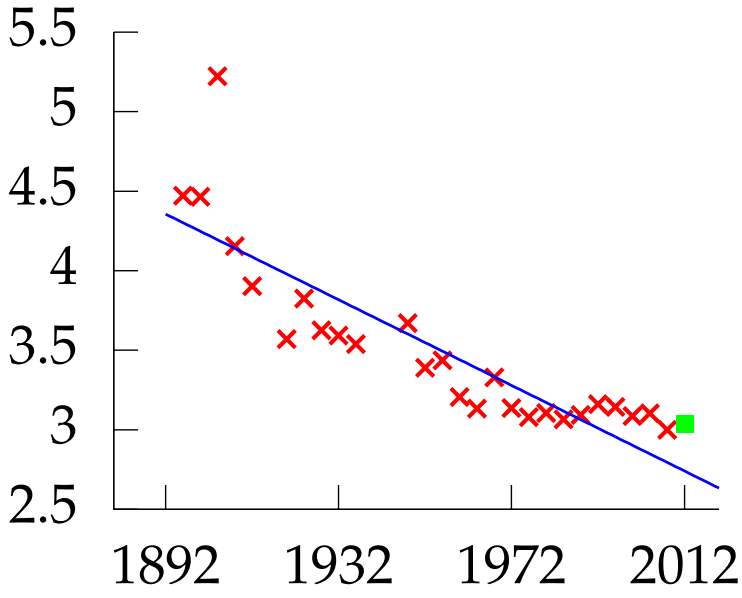
Leave One Out Error

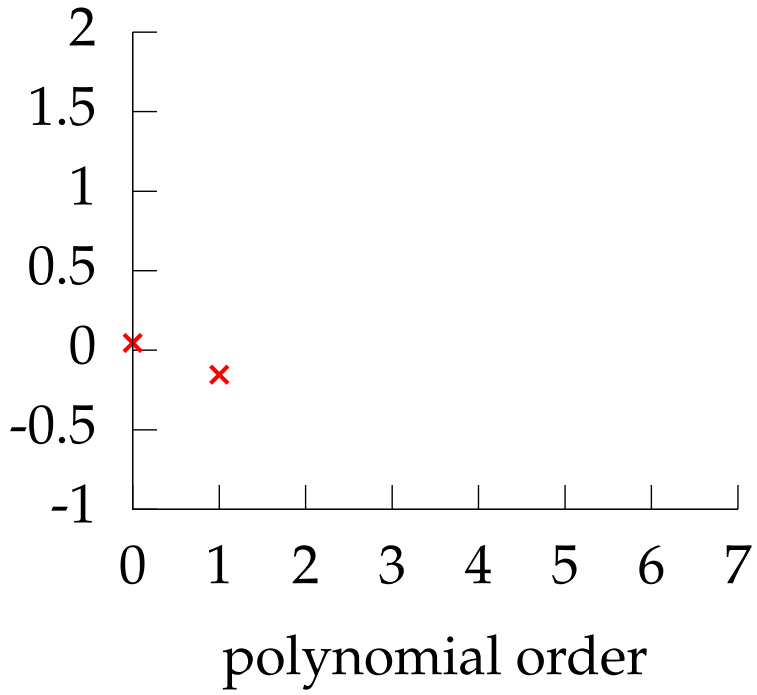




Polynomial order 1, training error -21.183, leave one out error -0.15413.

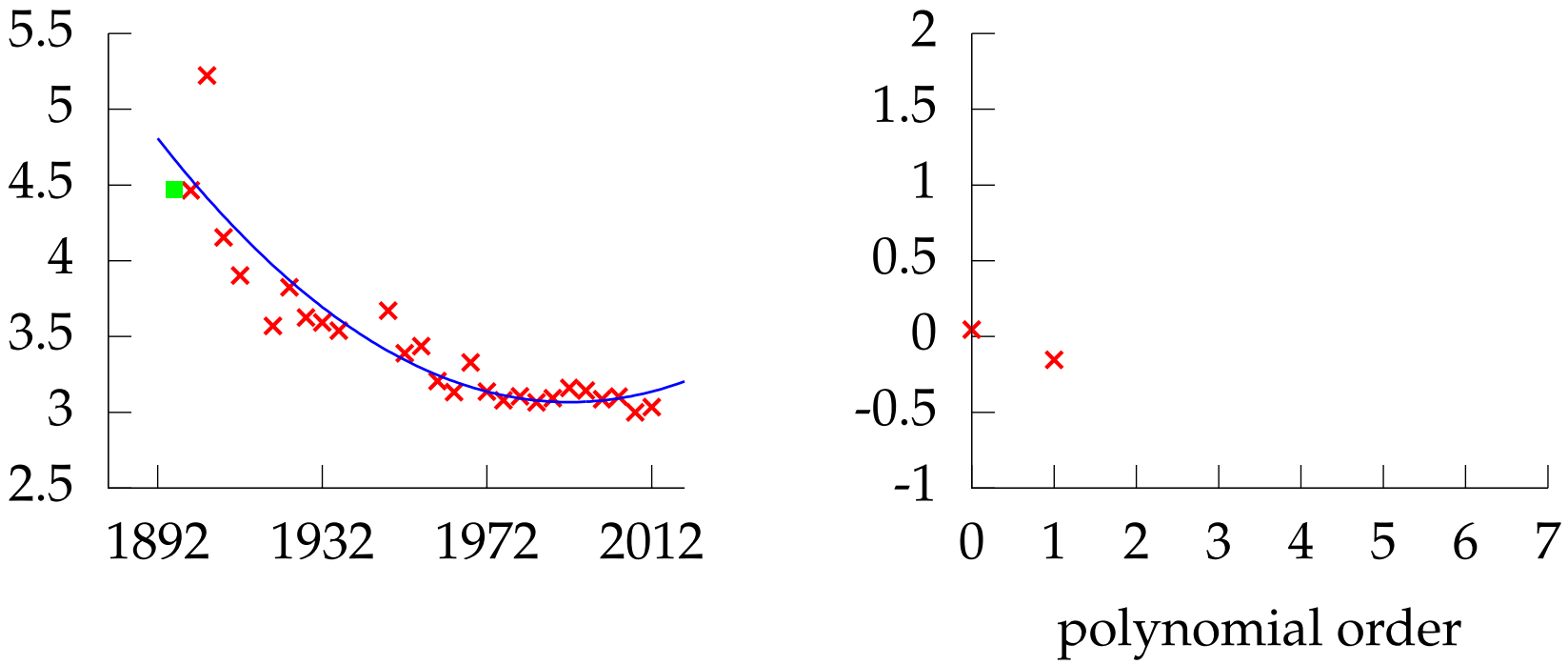
Leave One Out Error





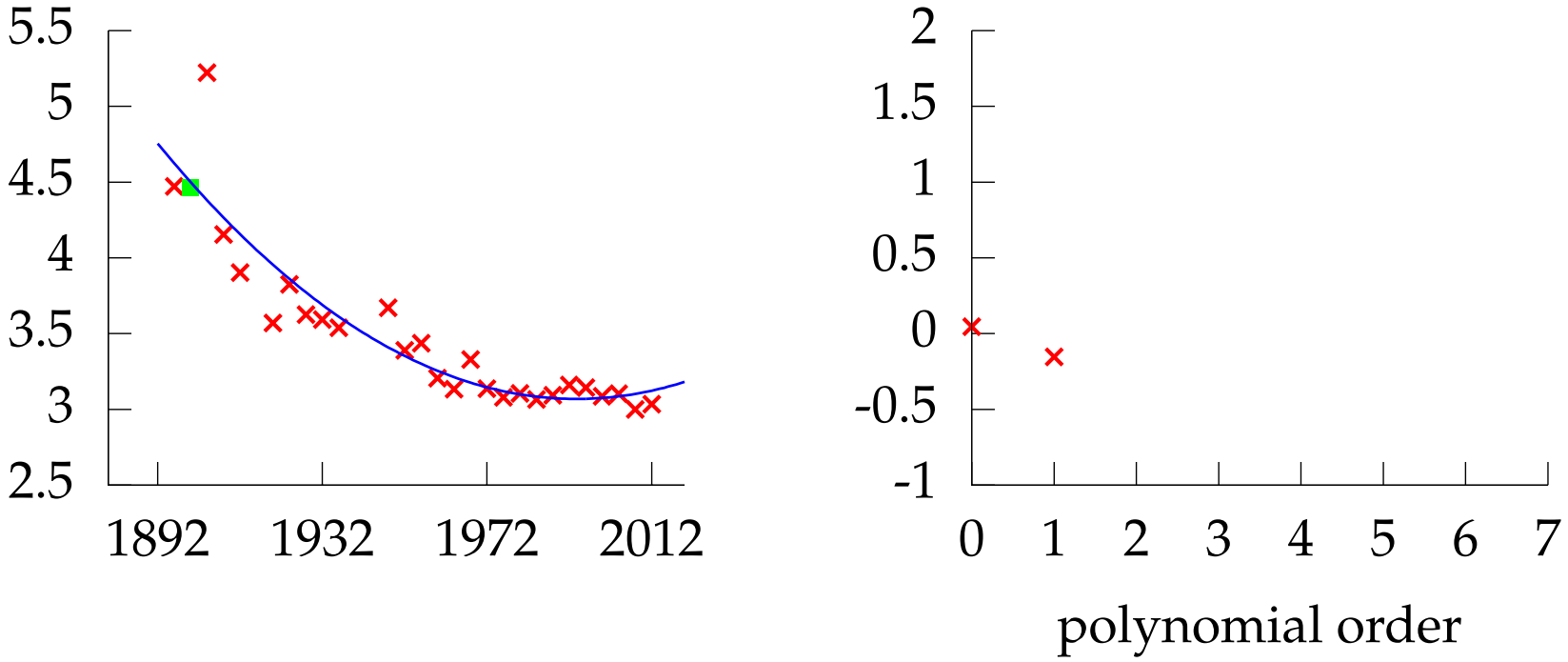
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



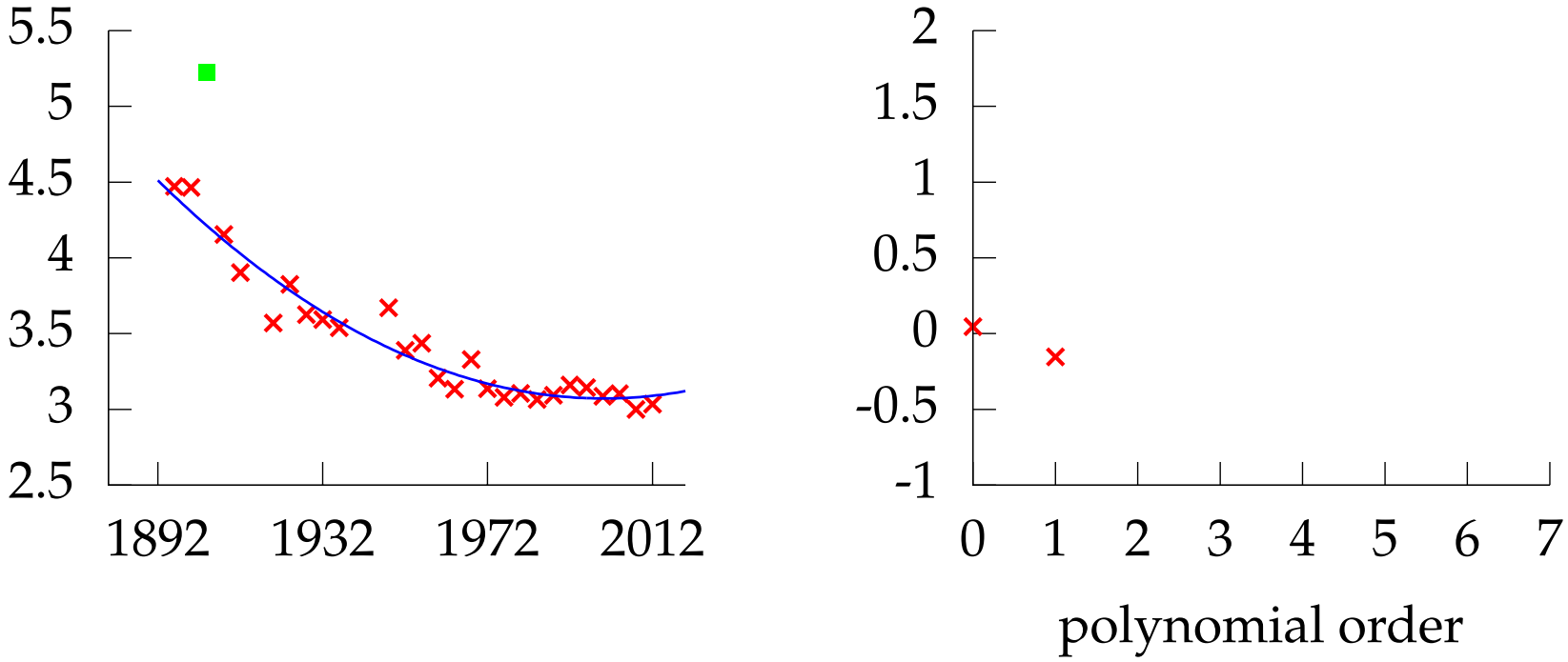
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



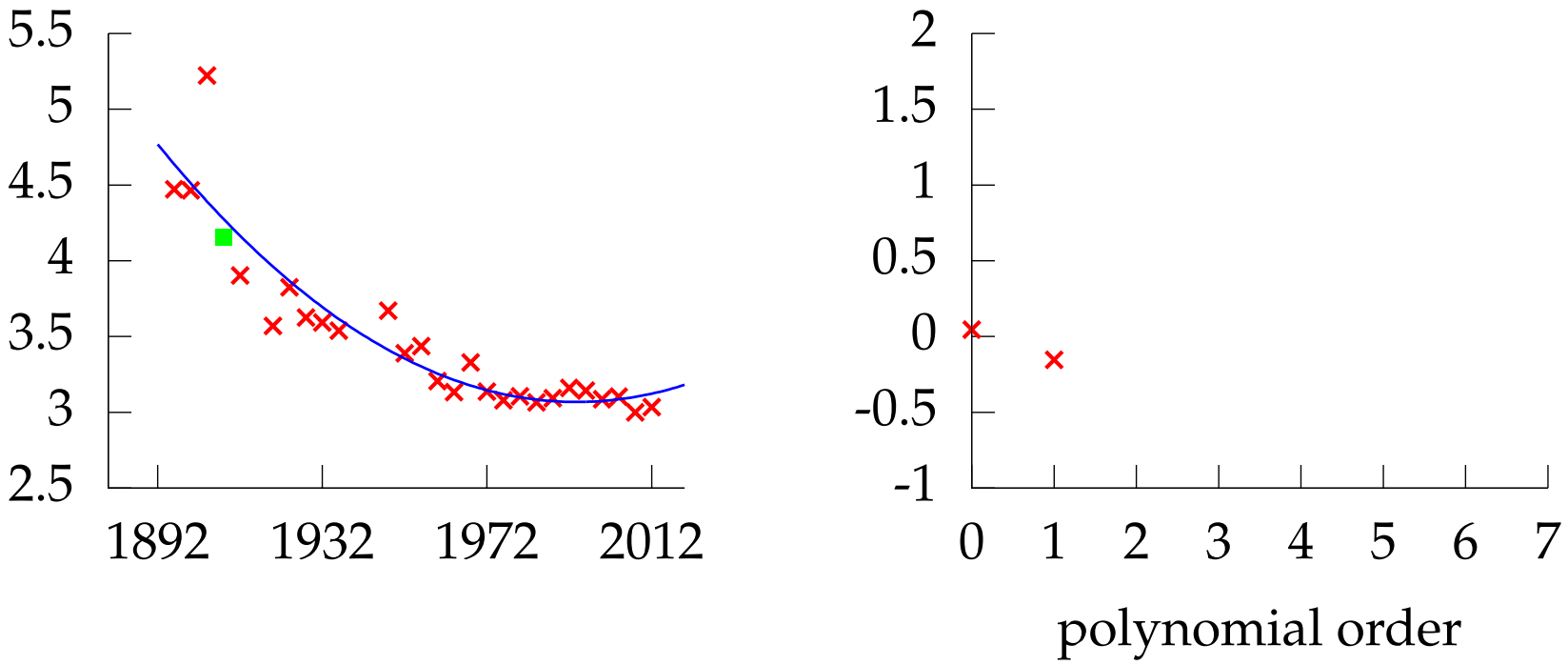
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



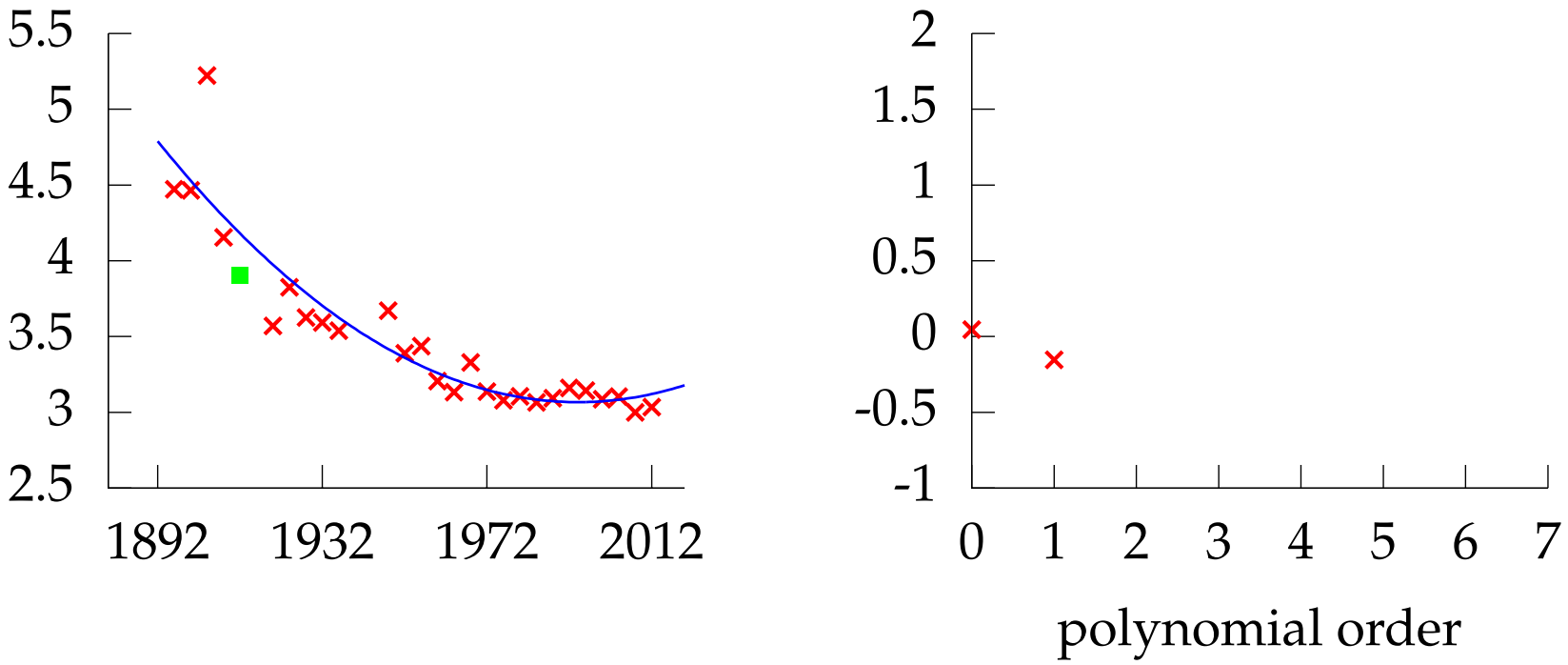
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



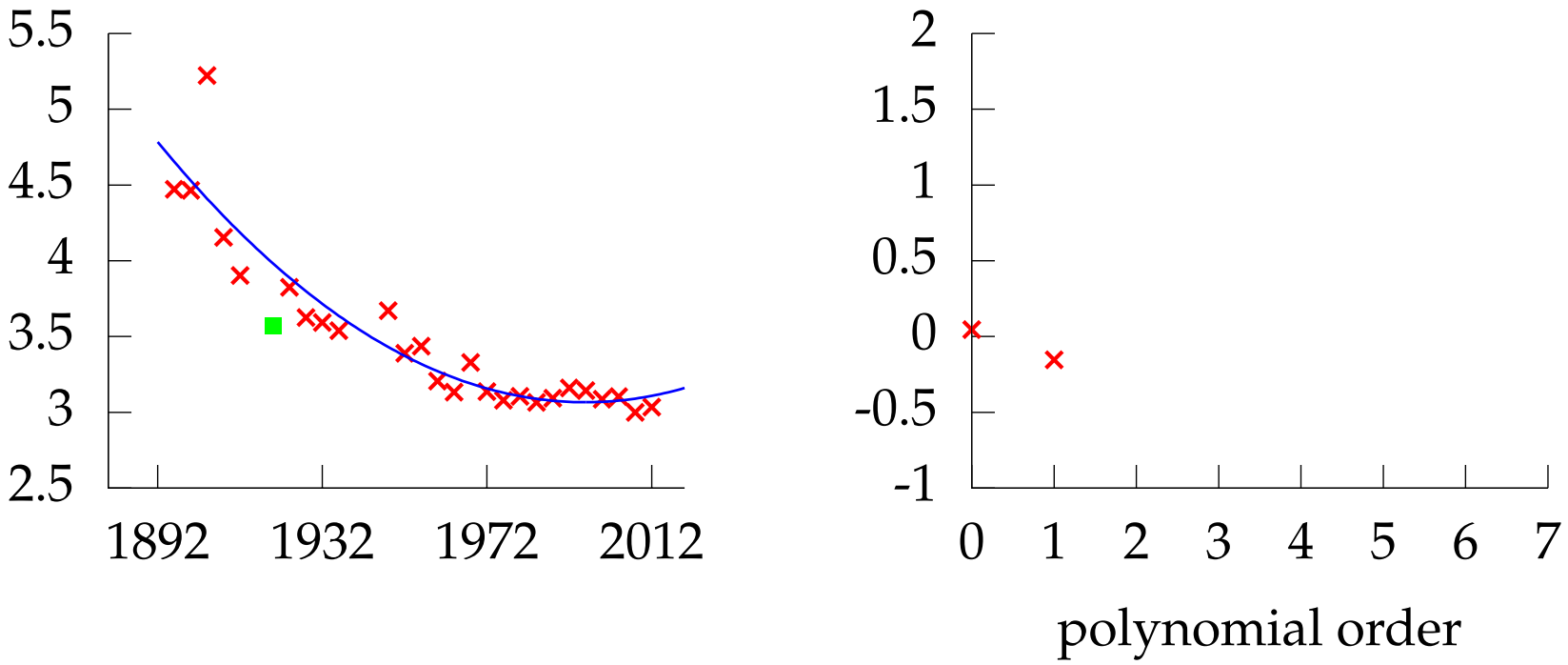
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



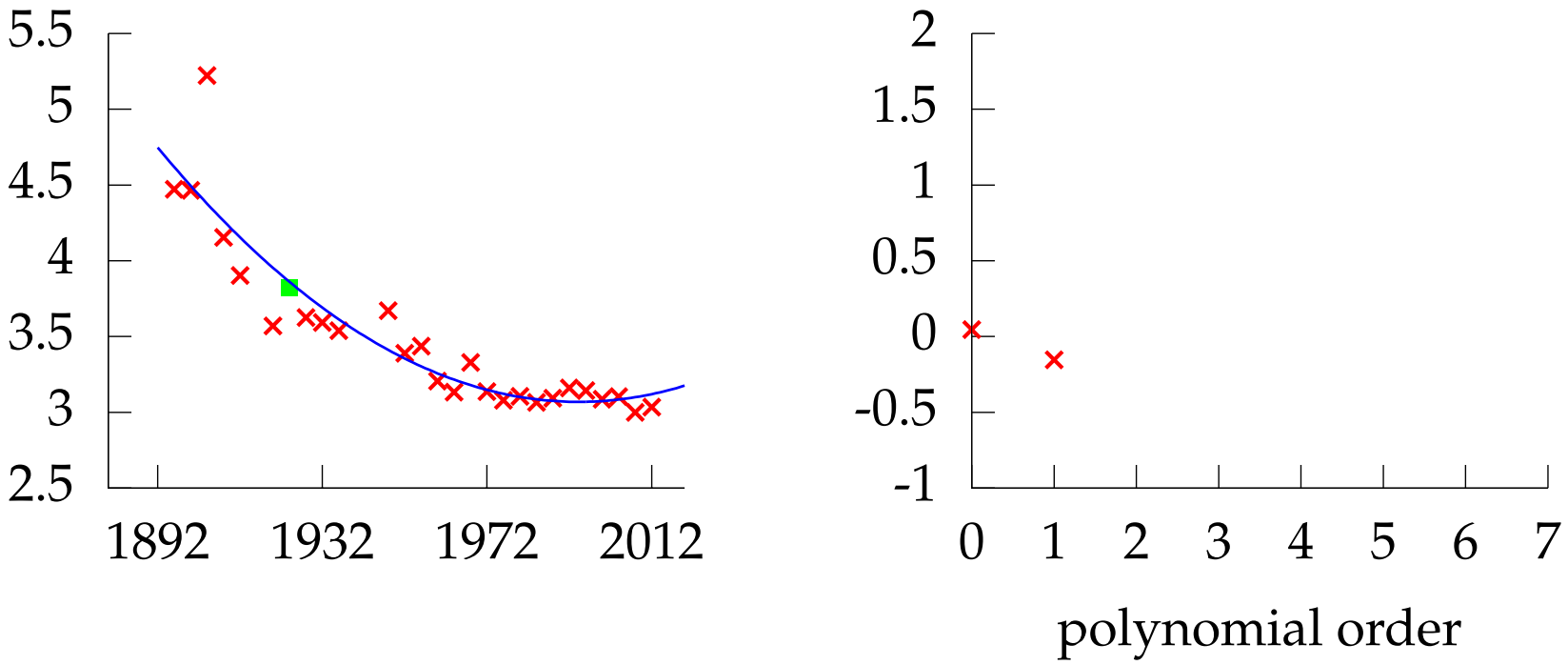
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



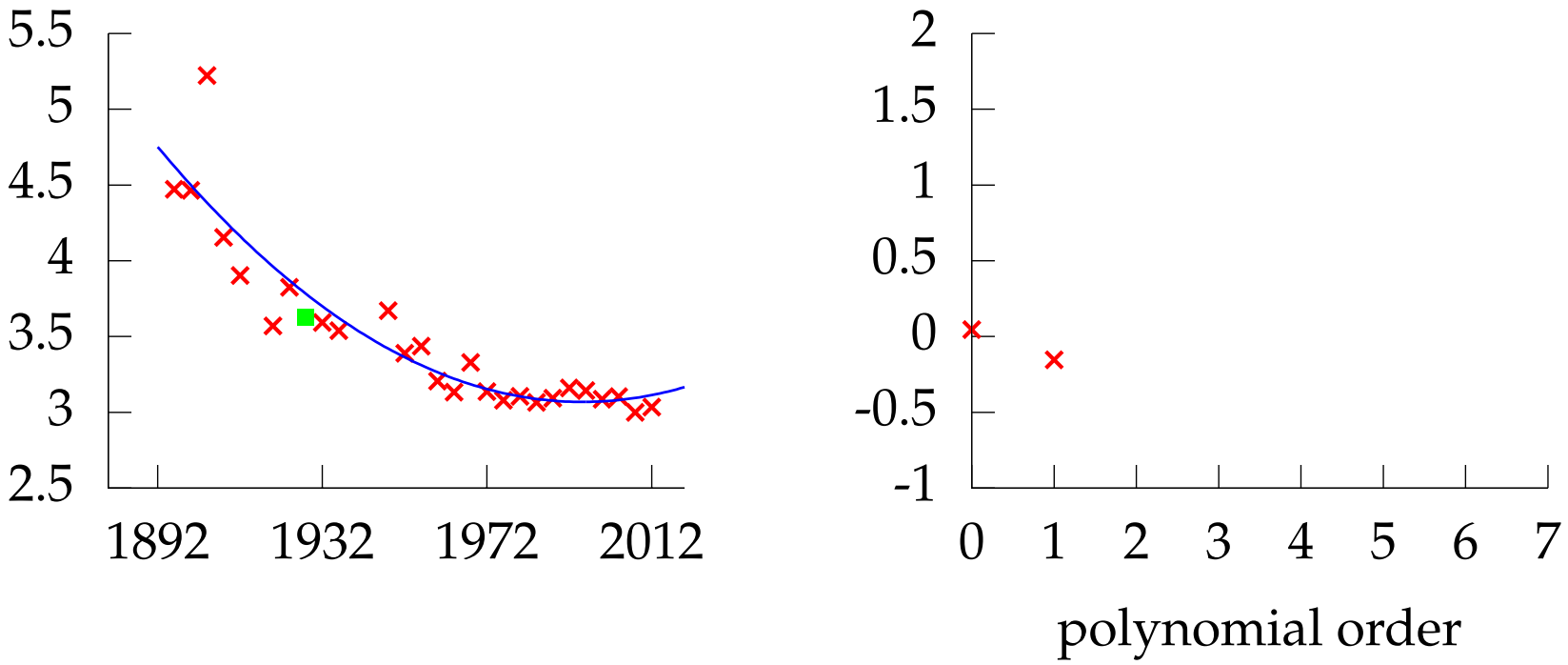
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



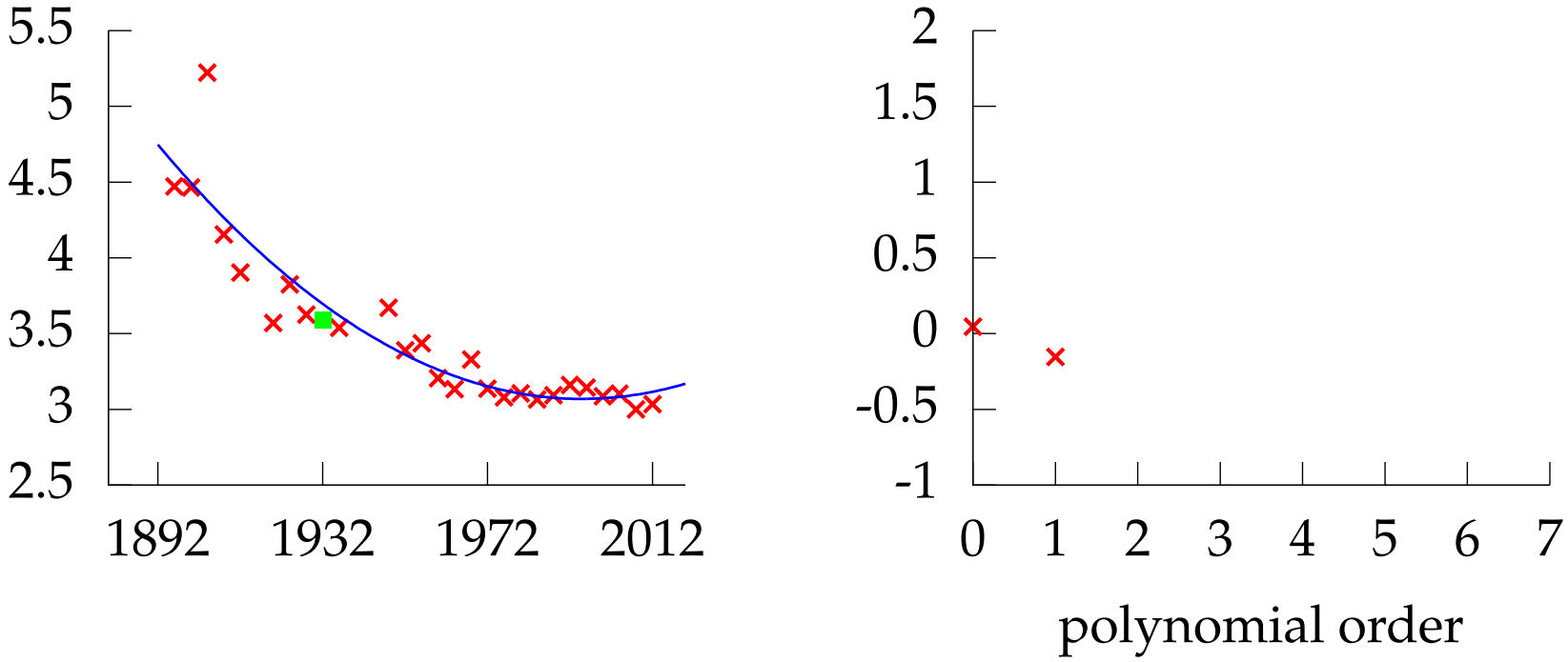
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



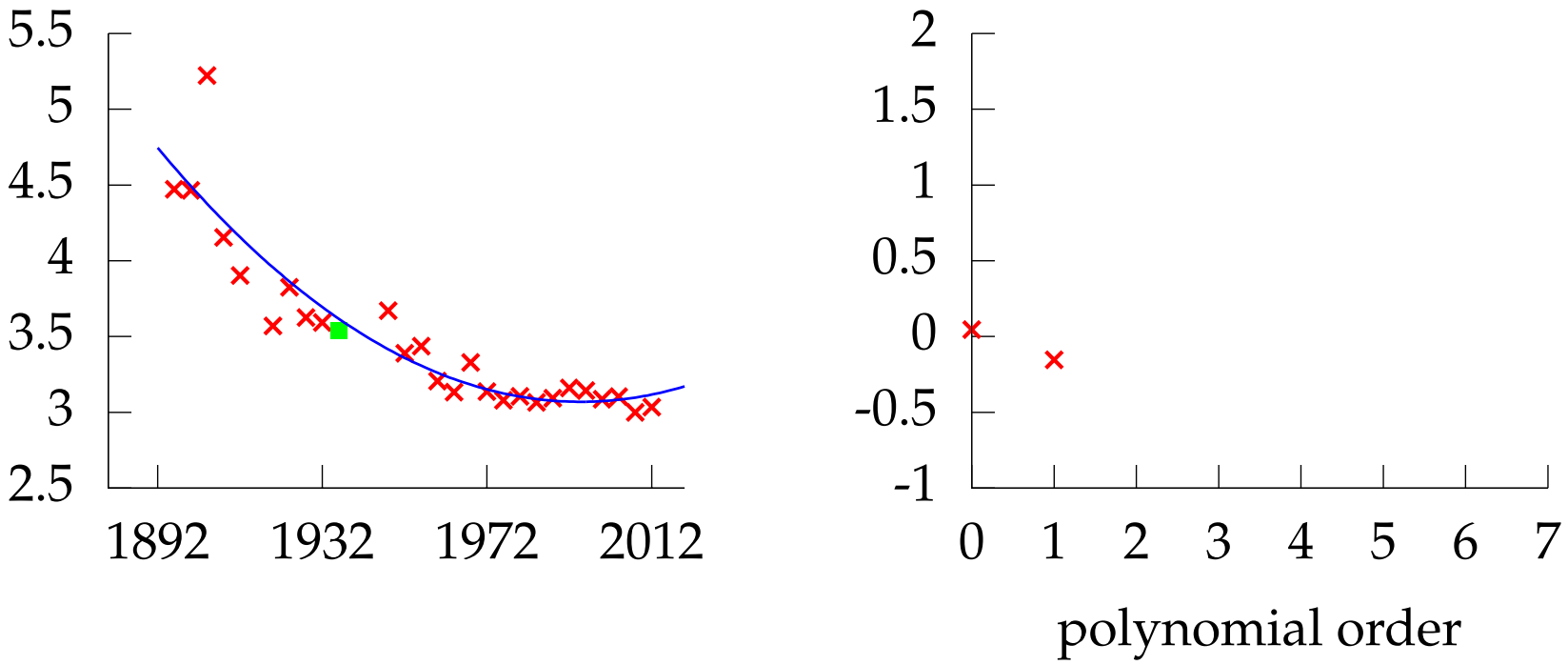
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



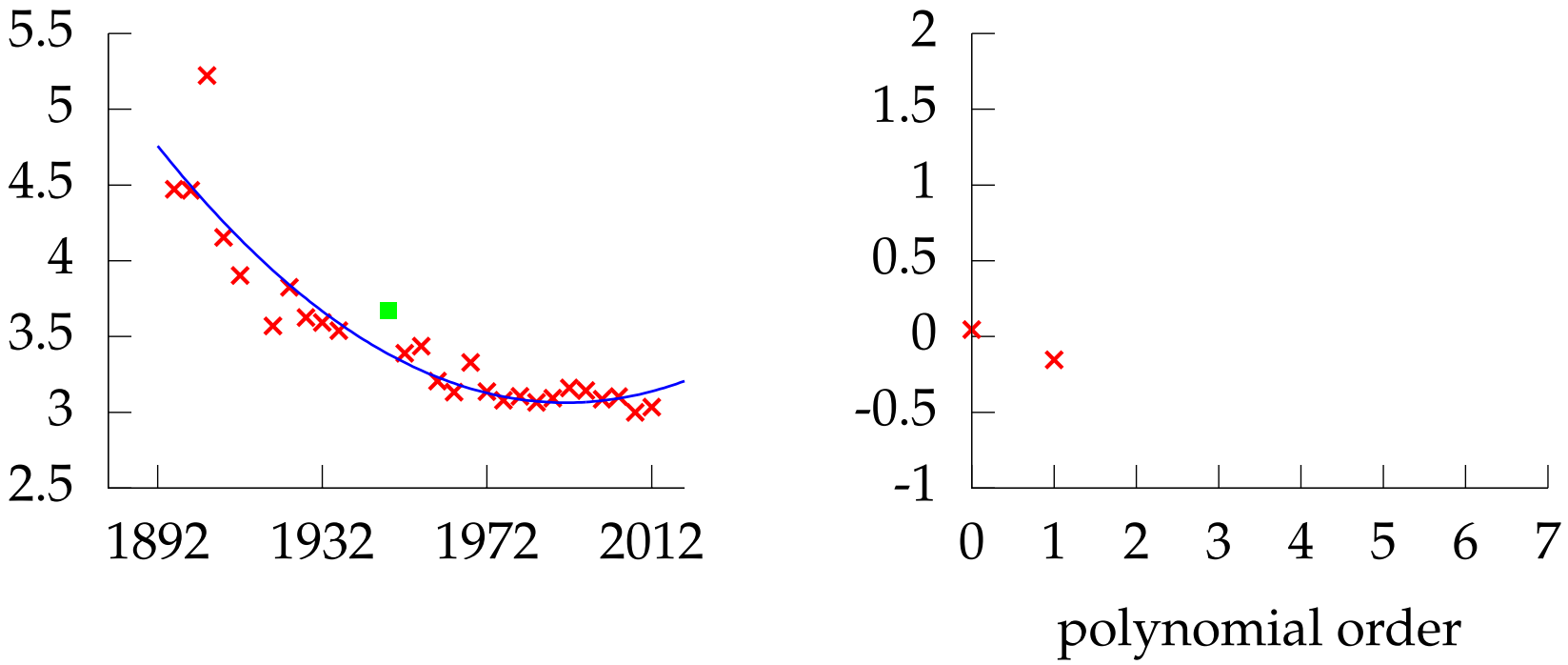
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



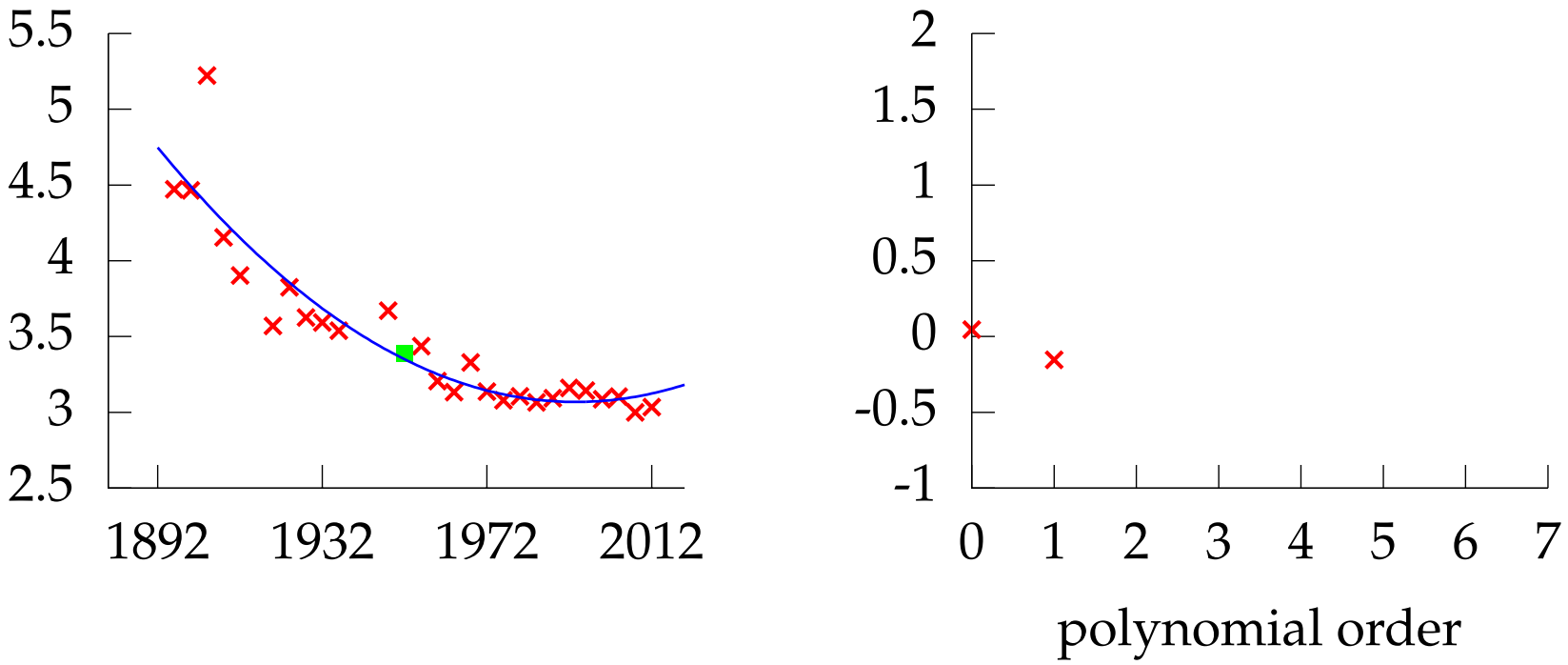
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



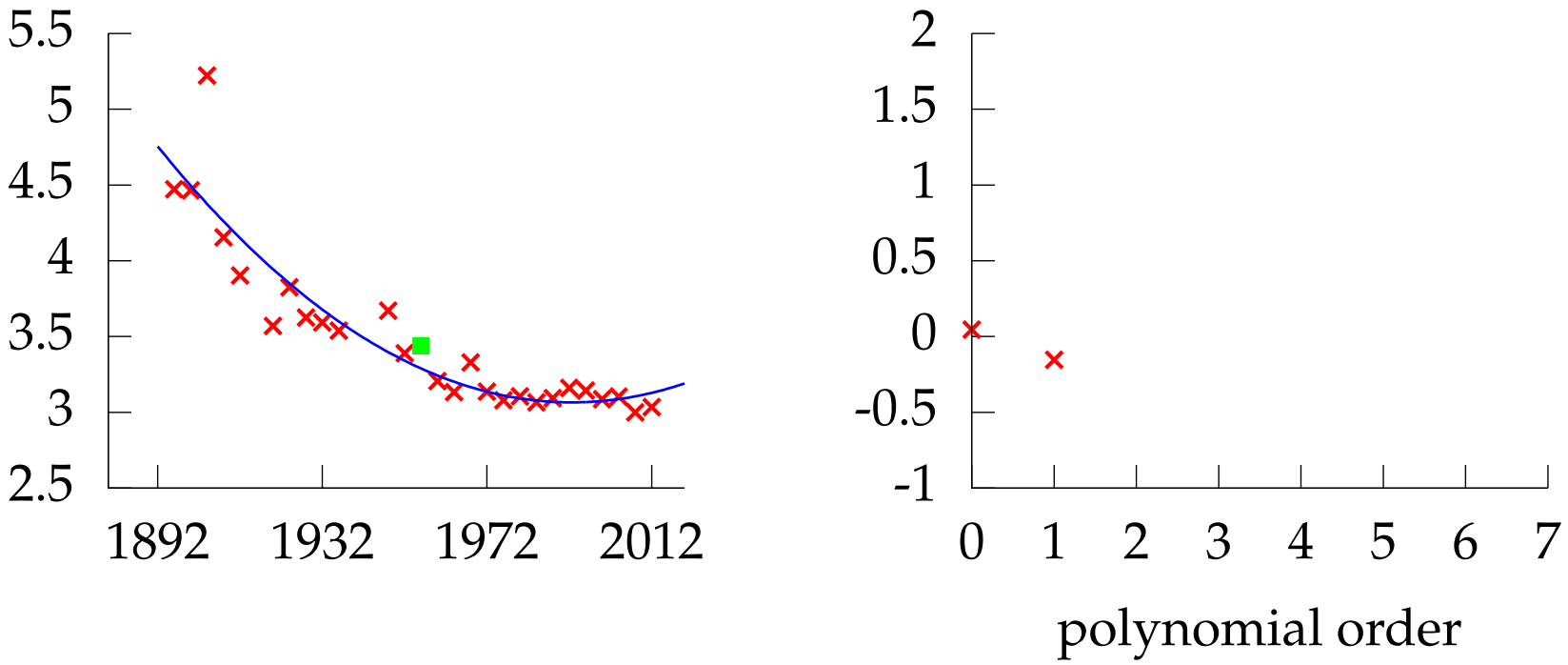
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



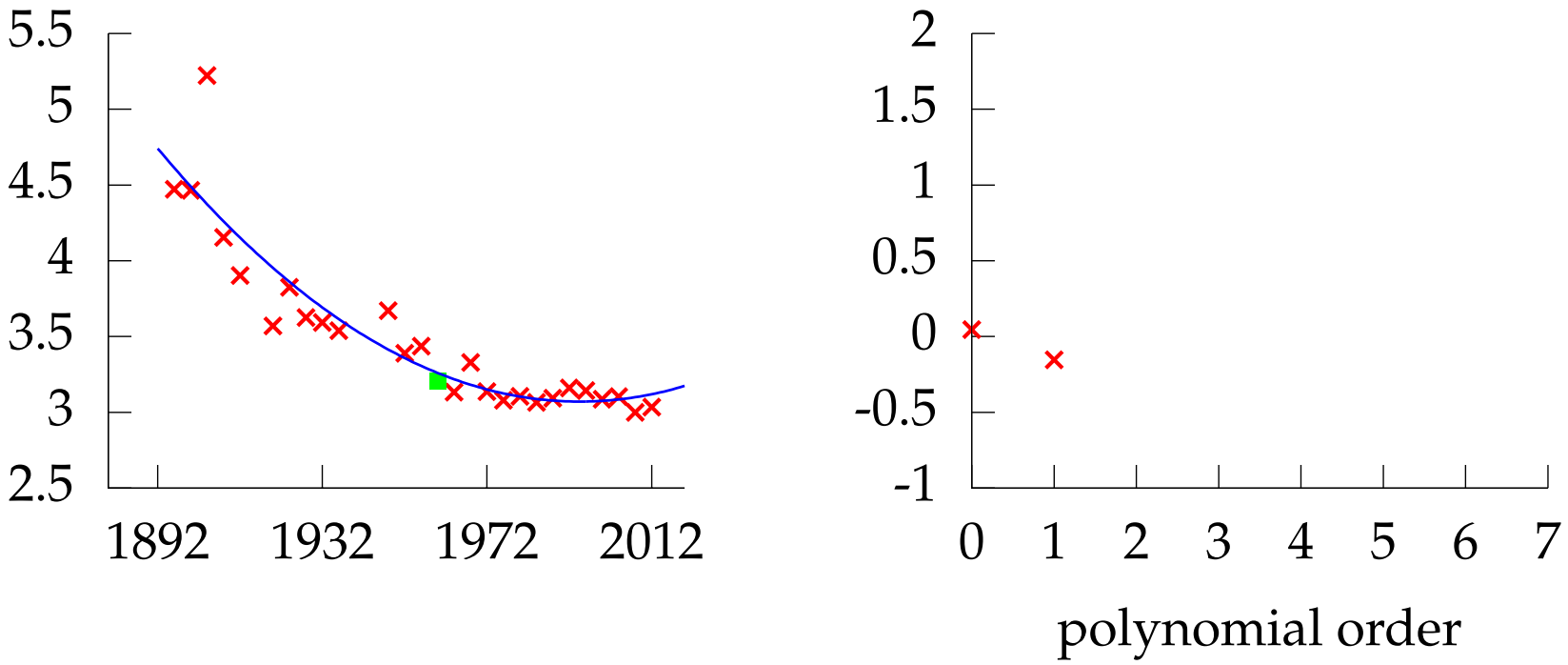
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



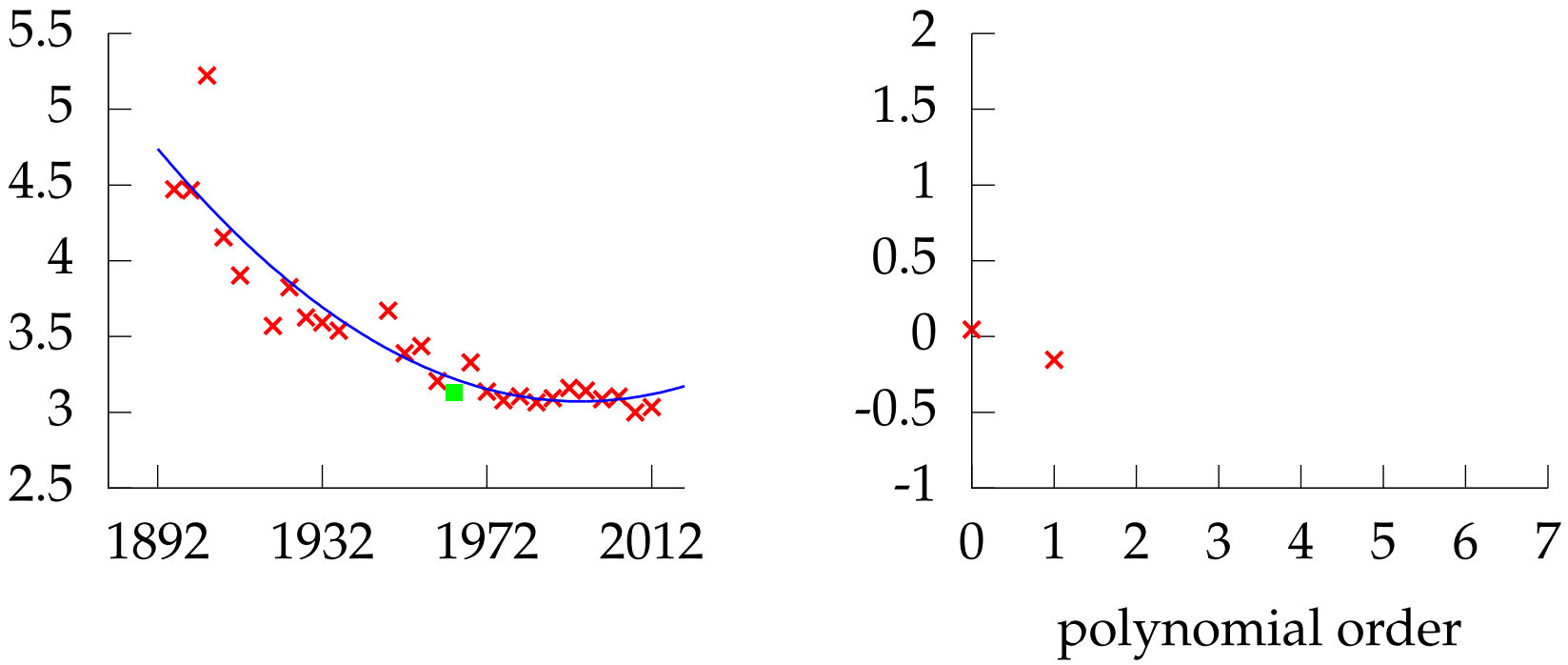
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



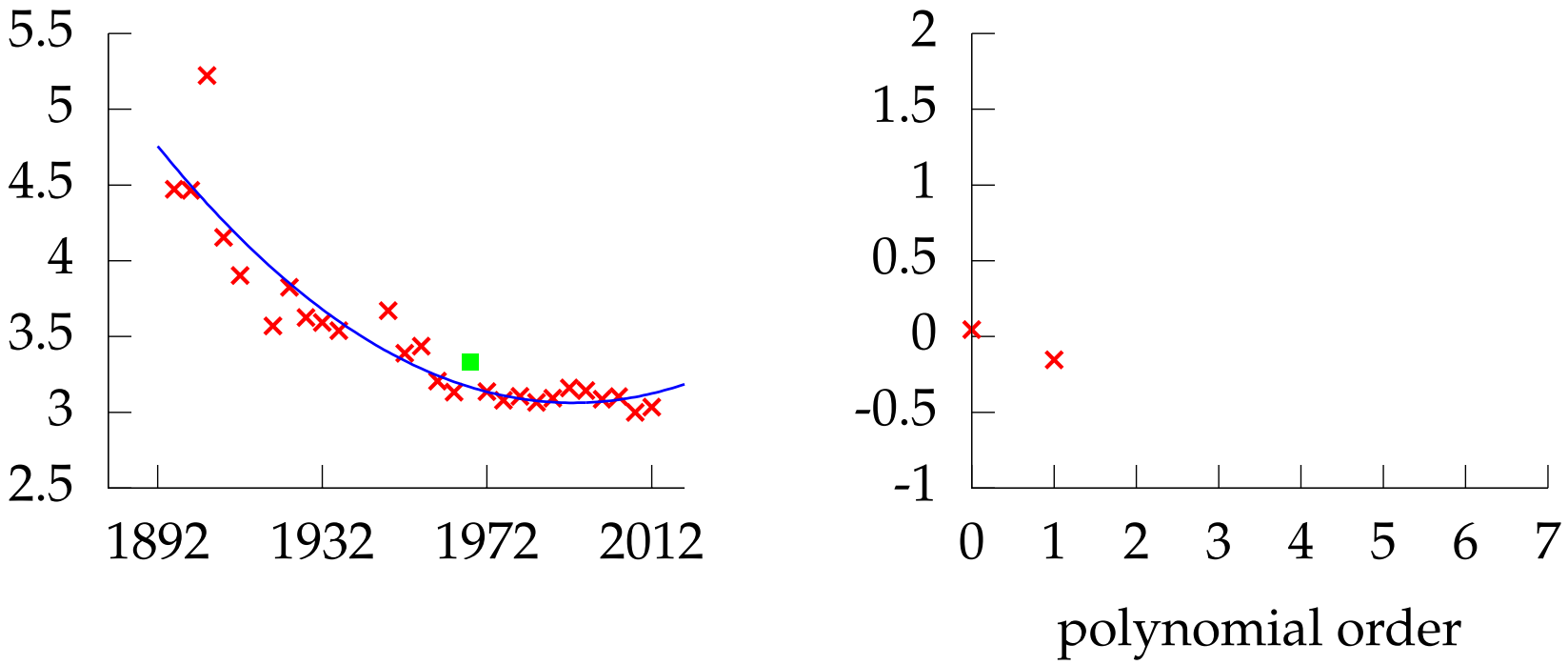
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



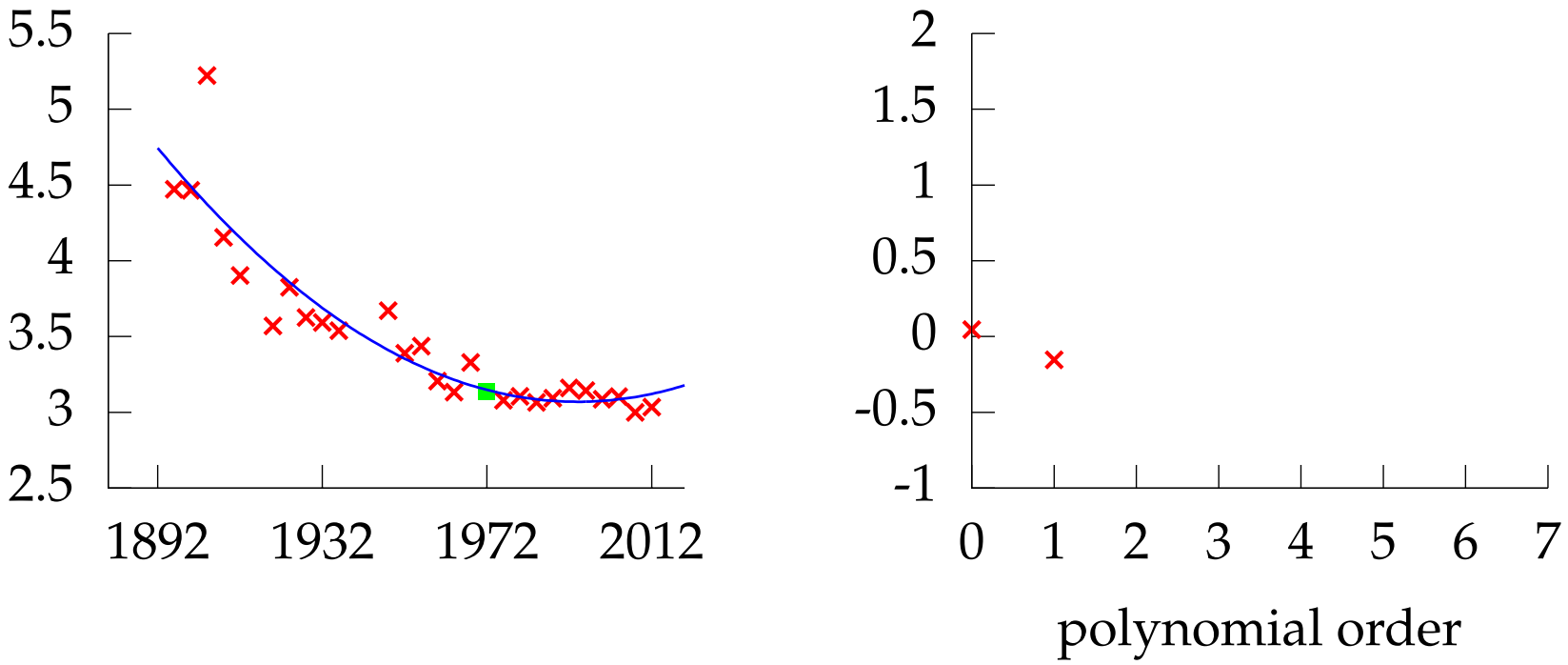
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



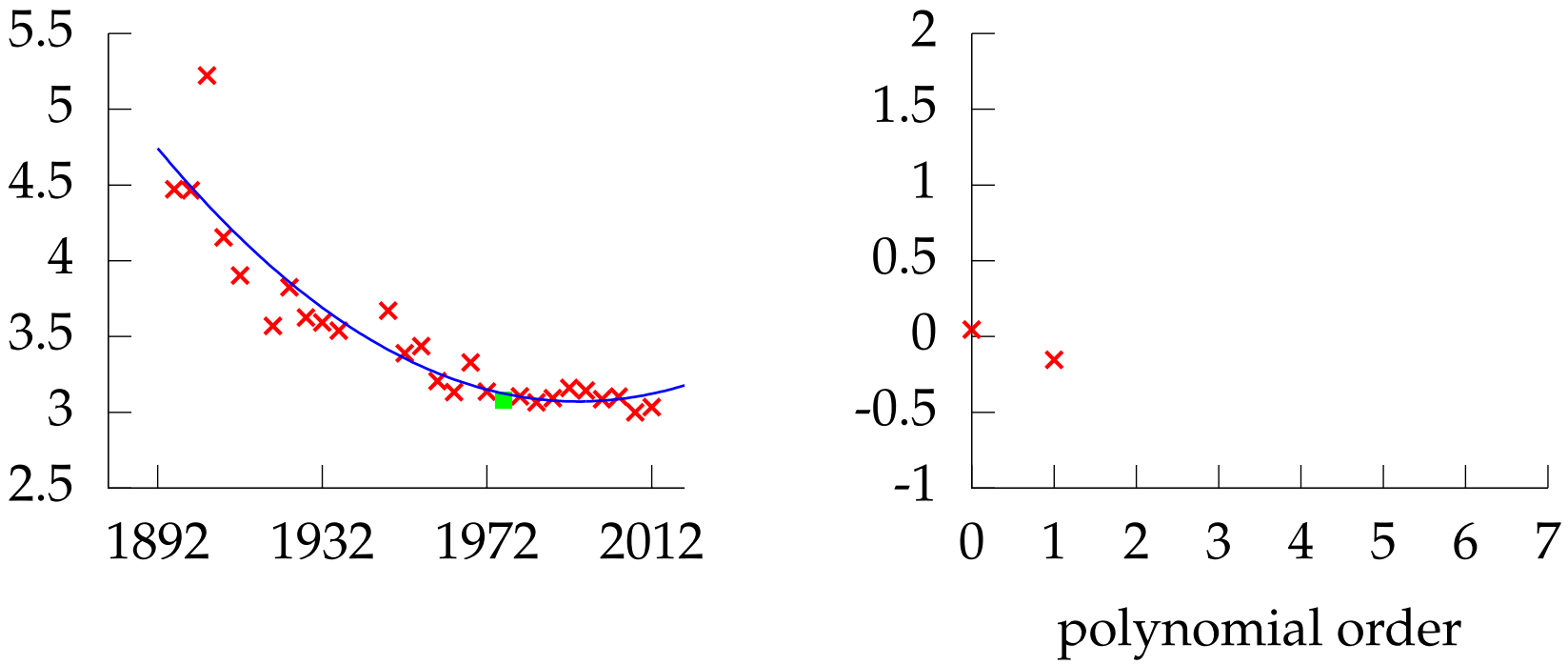
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



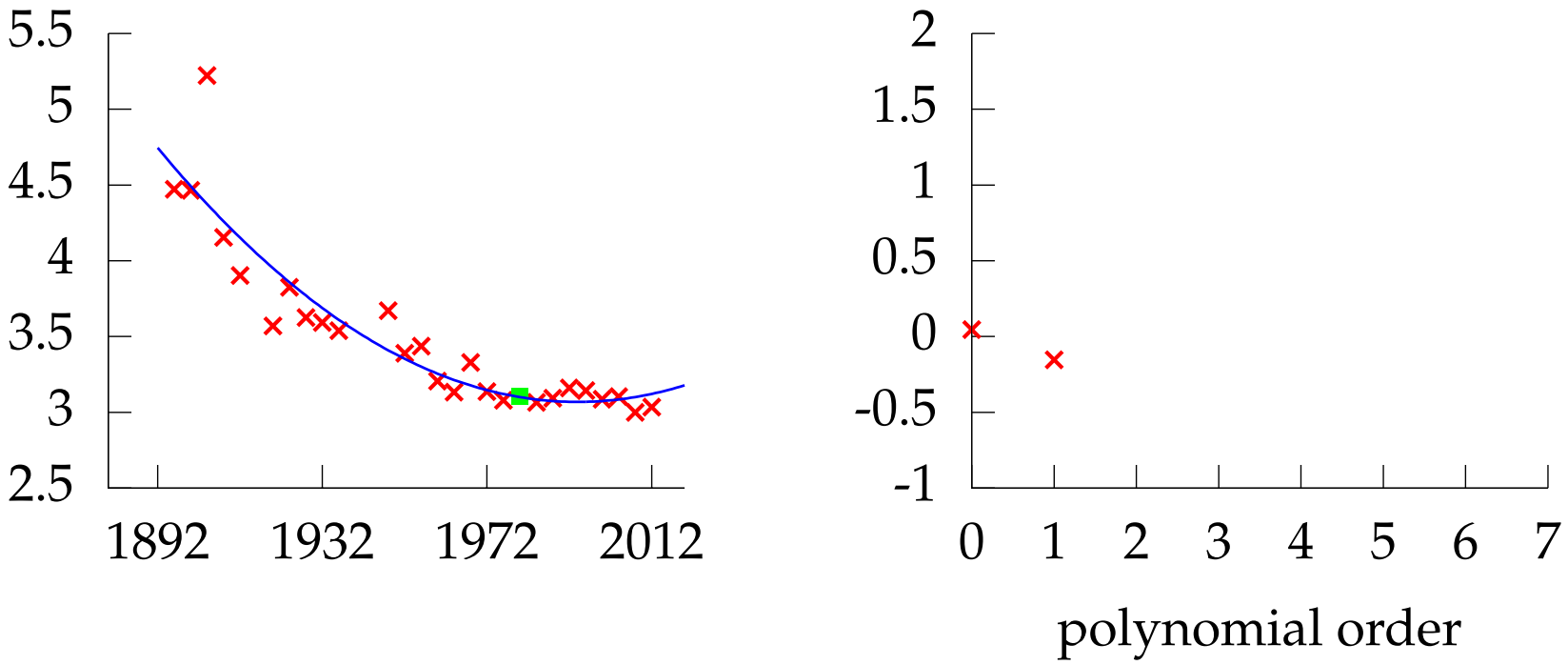
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



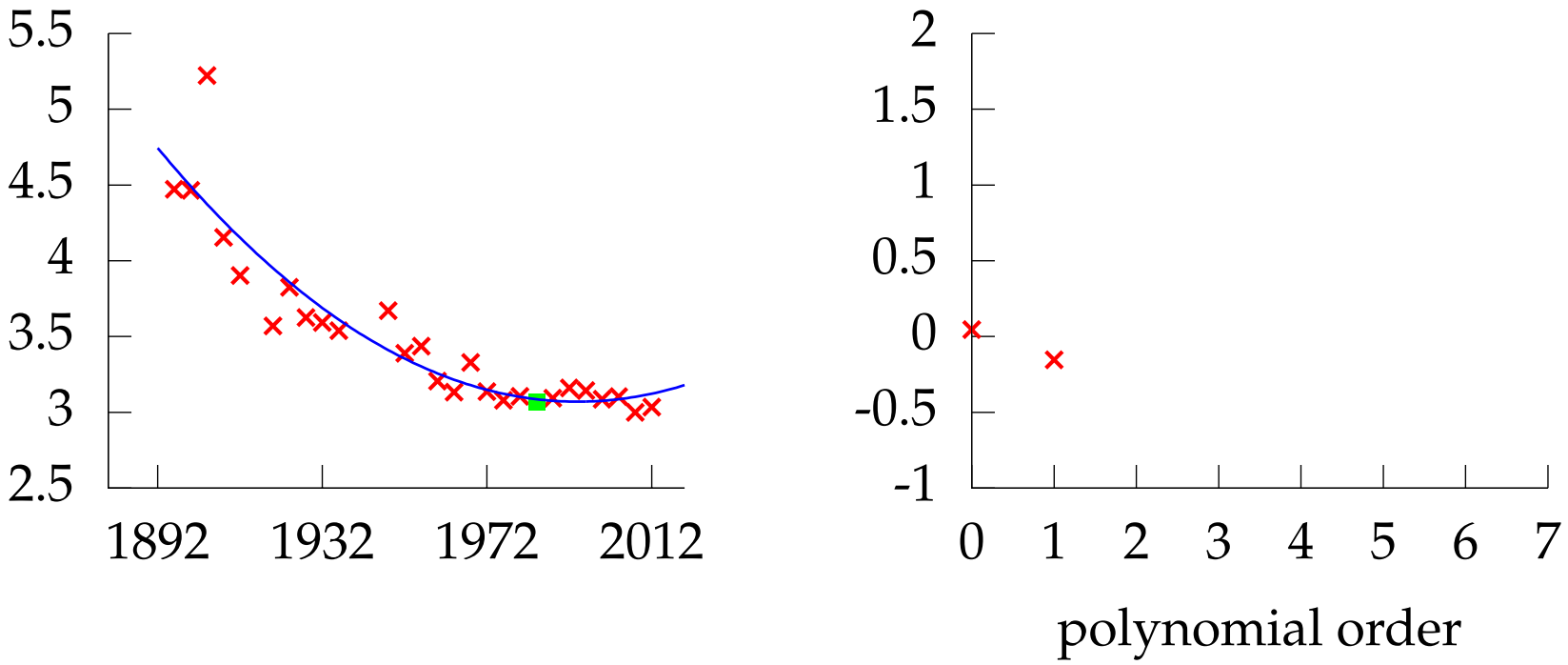
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



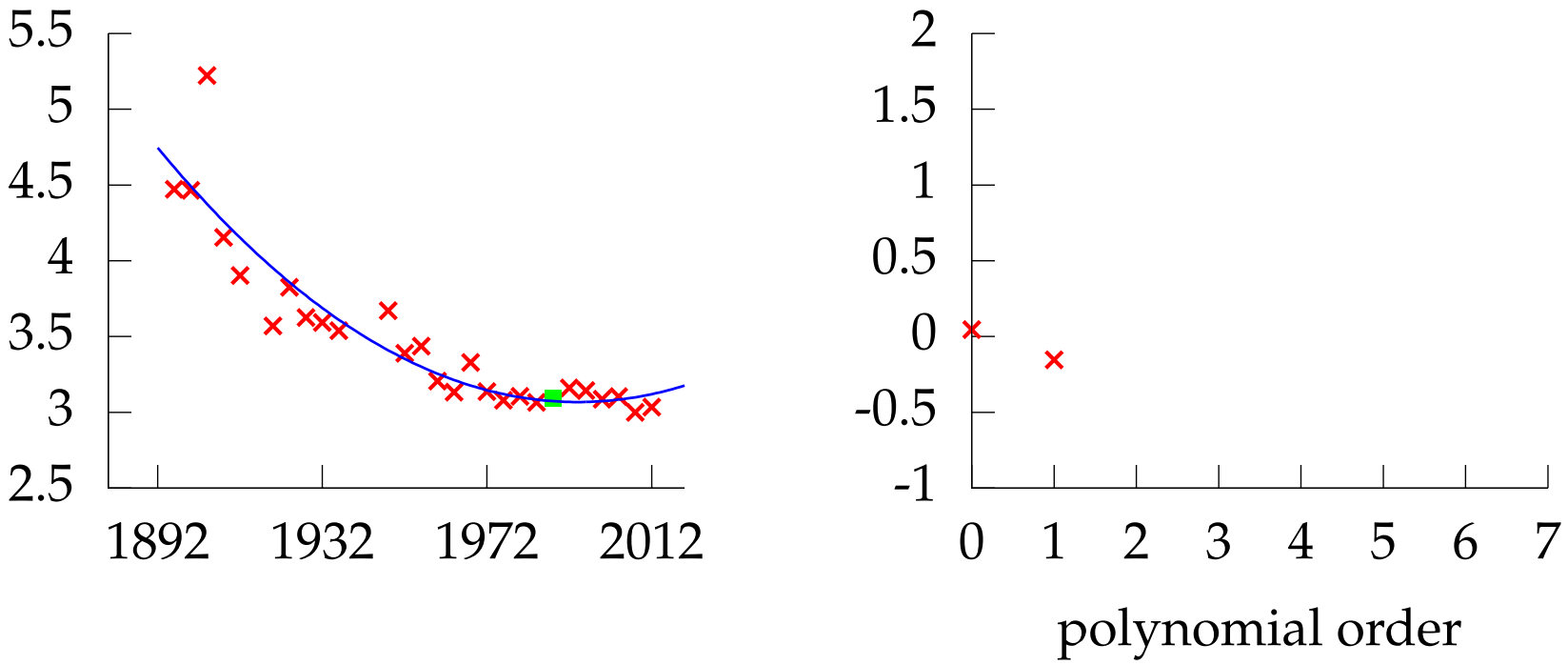
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



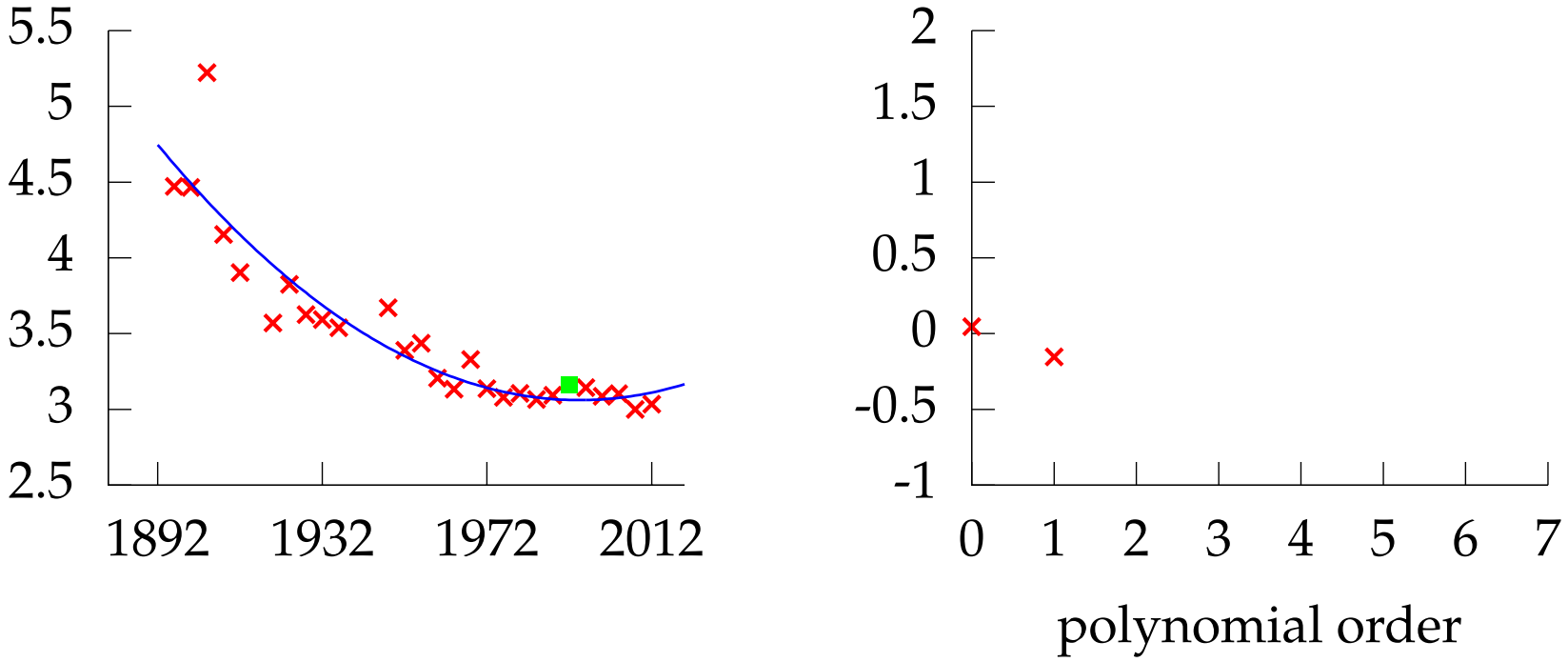
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



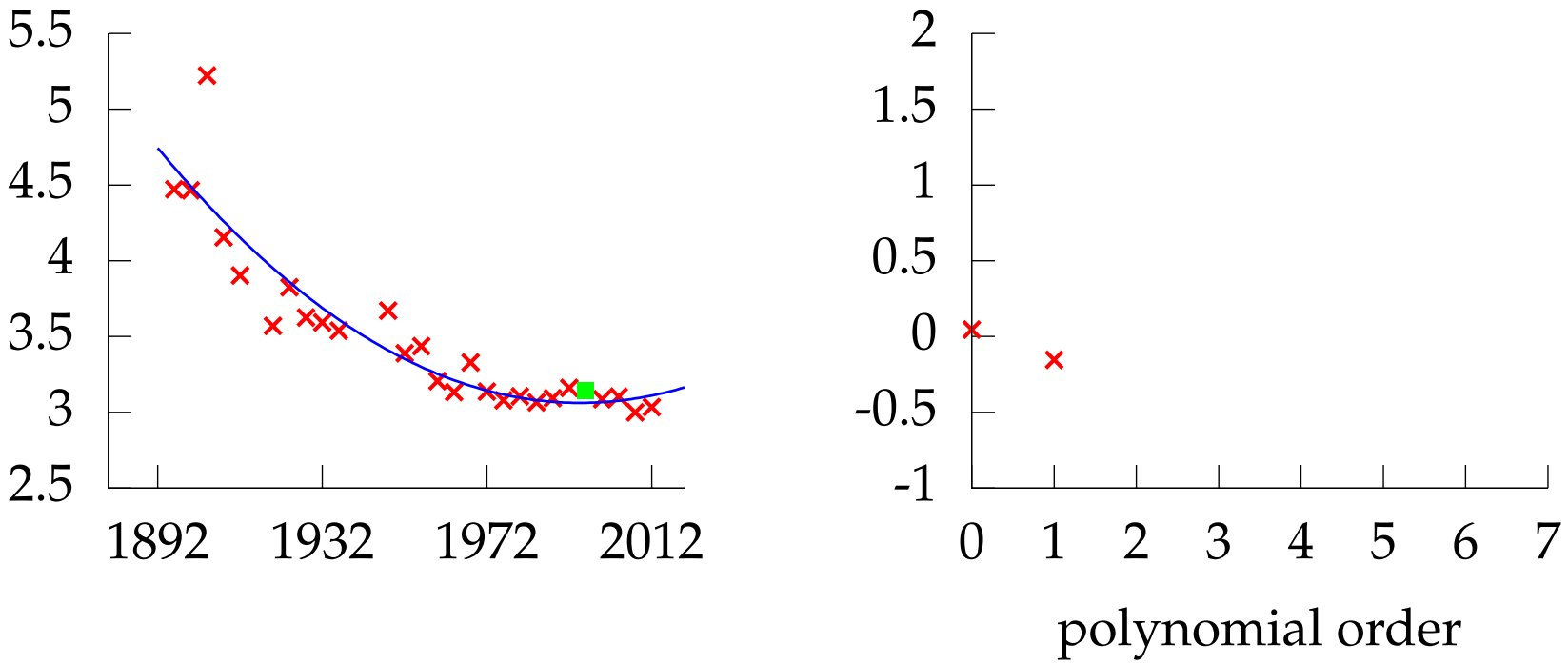
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



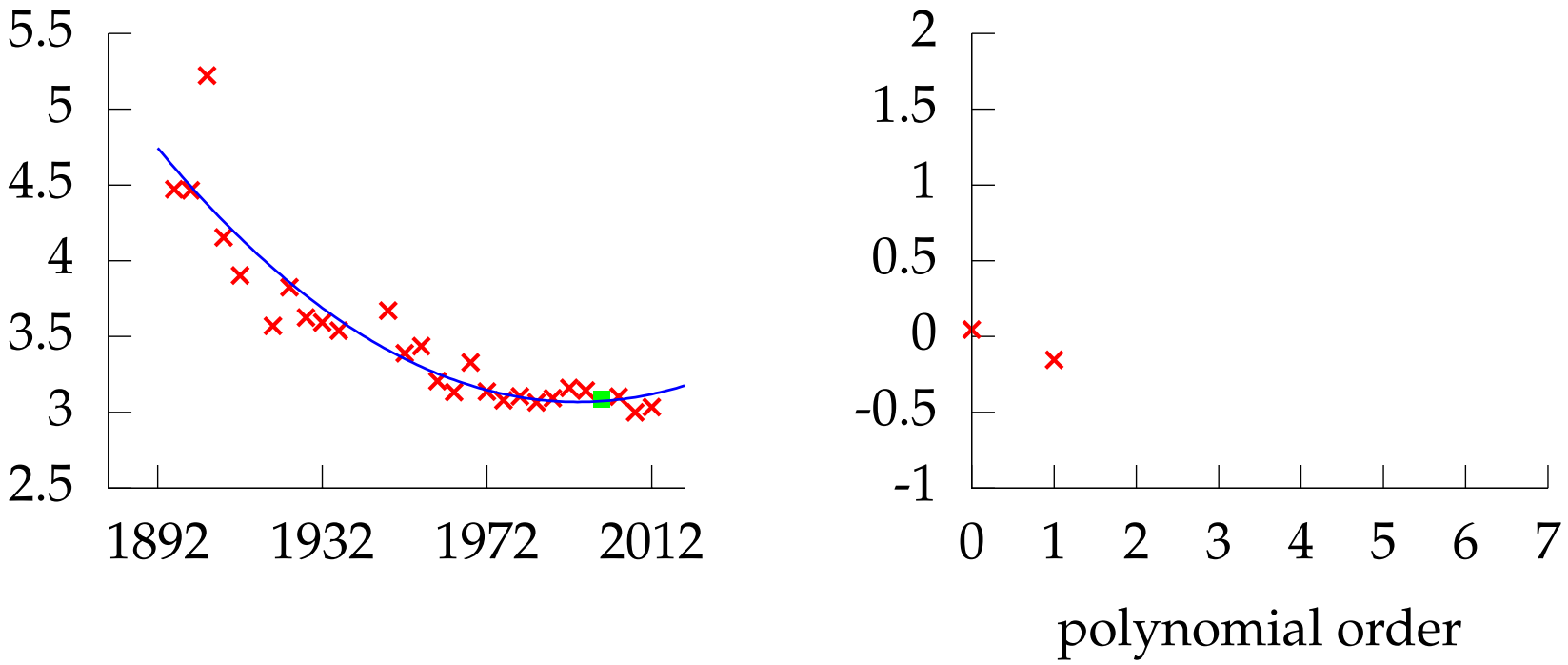
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



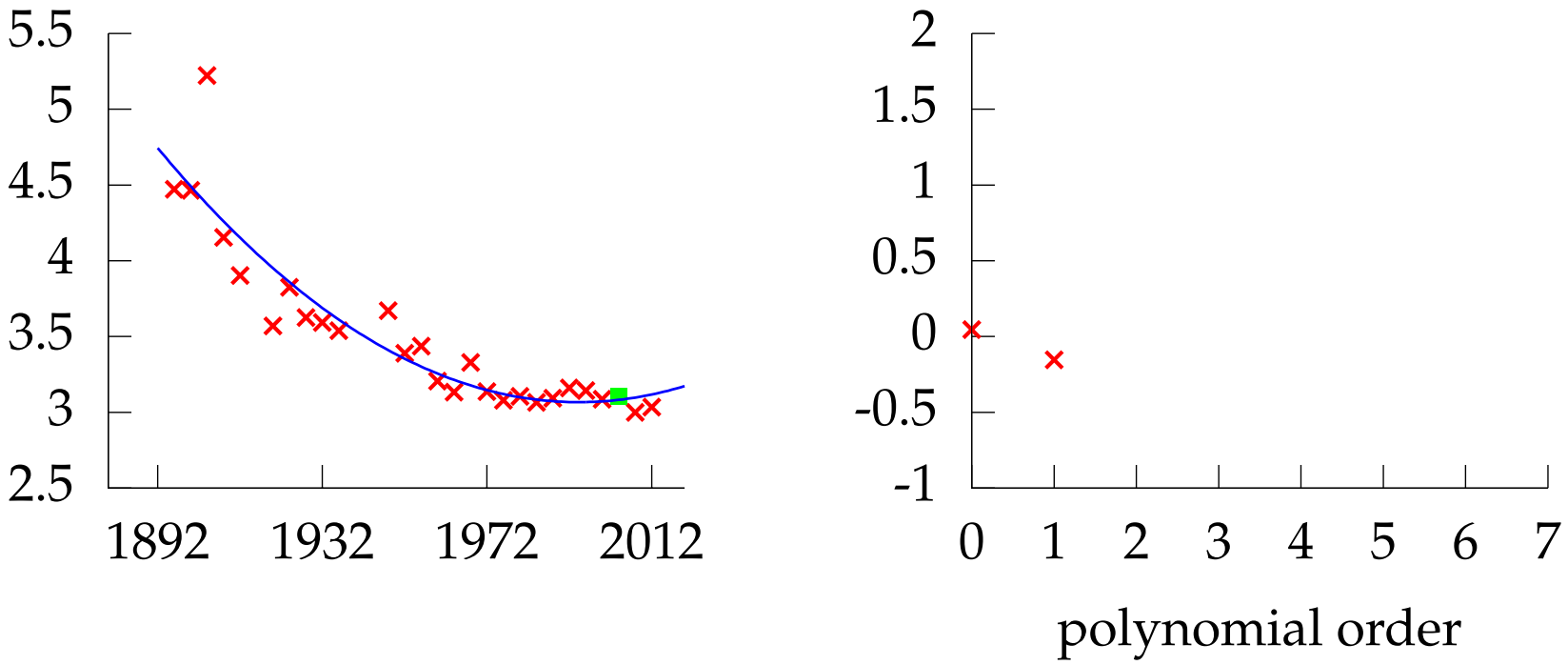
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



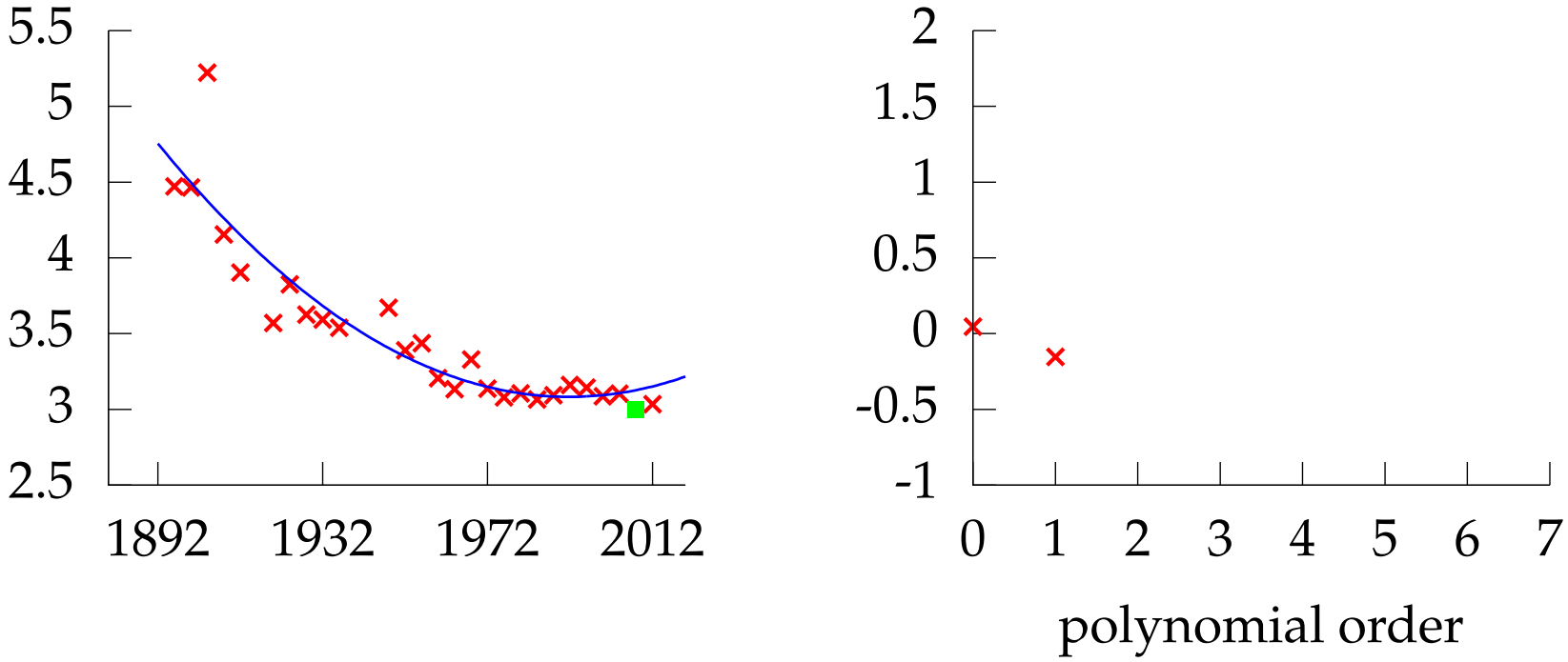
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



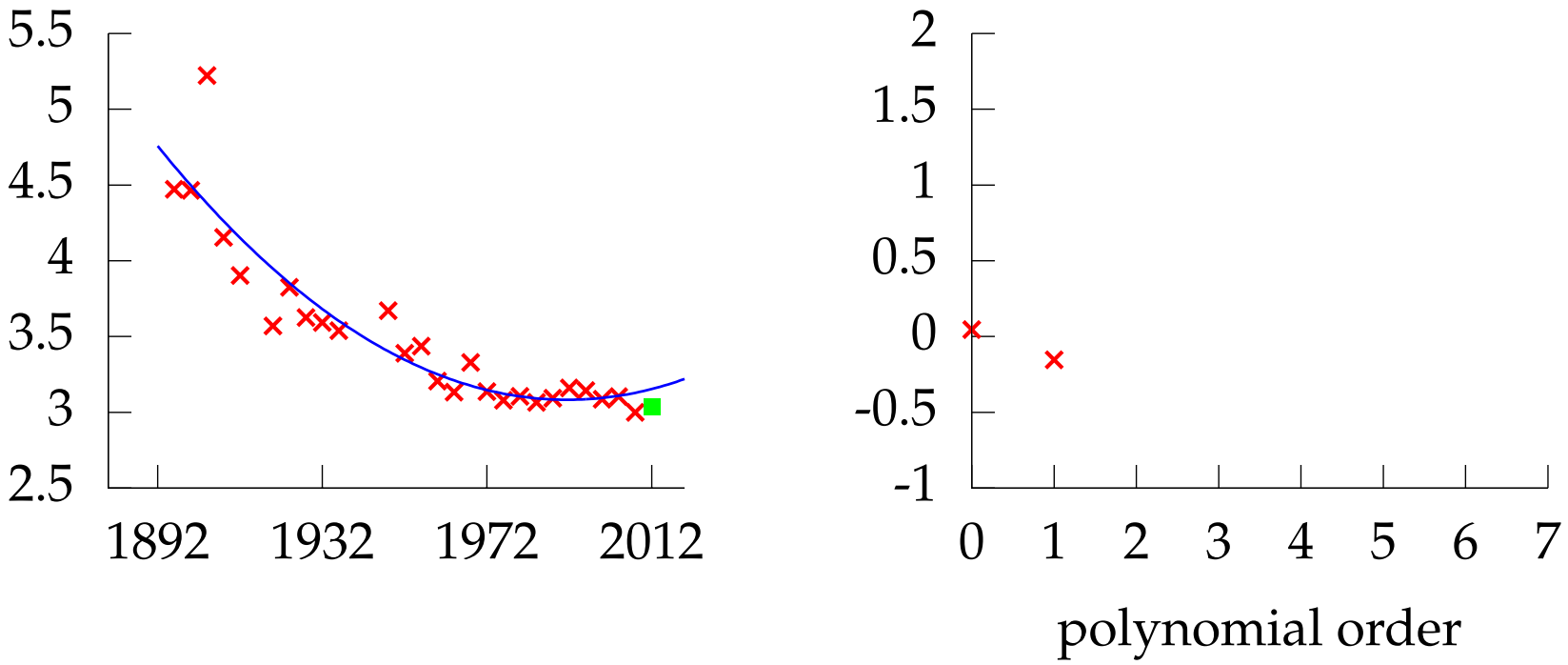
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



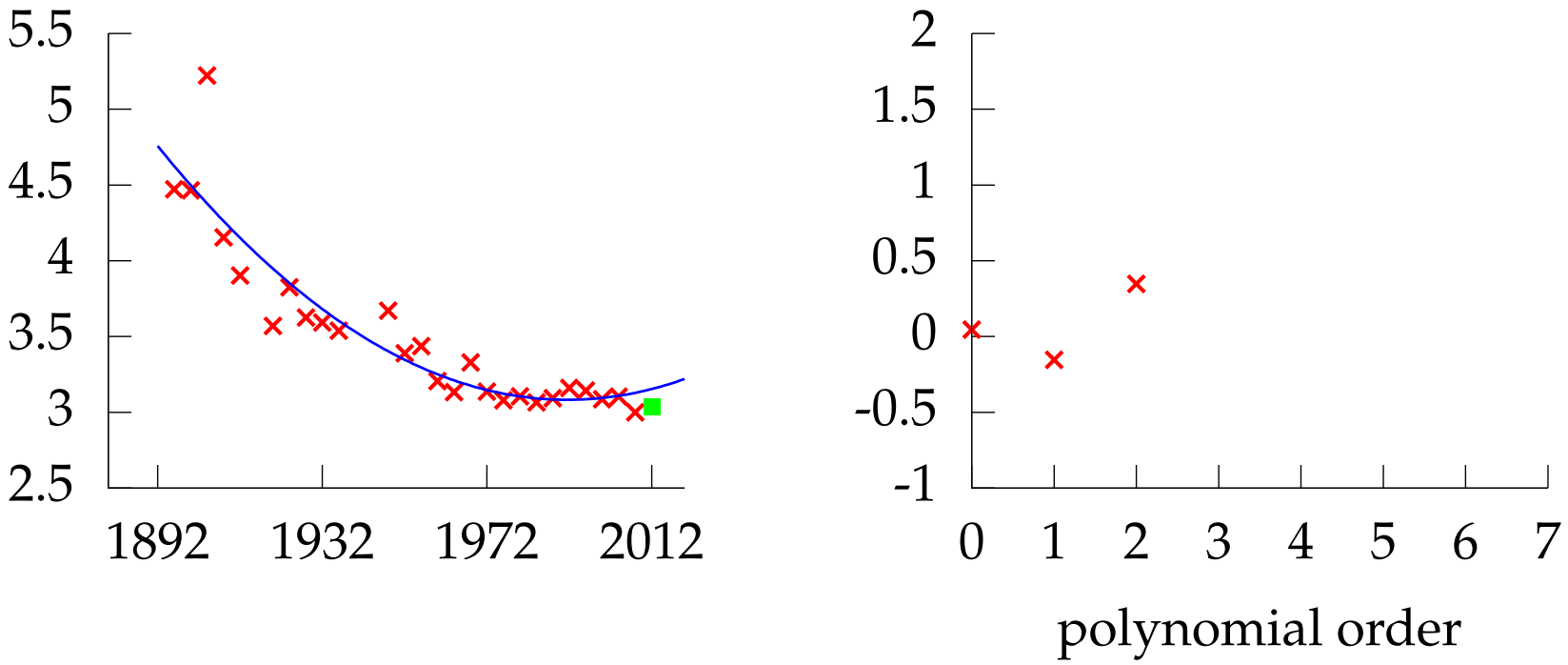
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



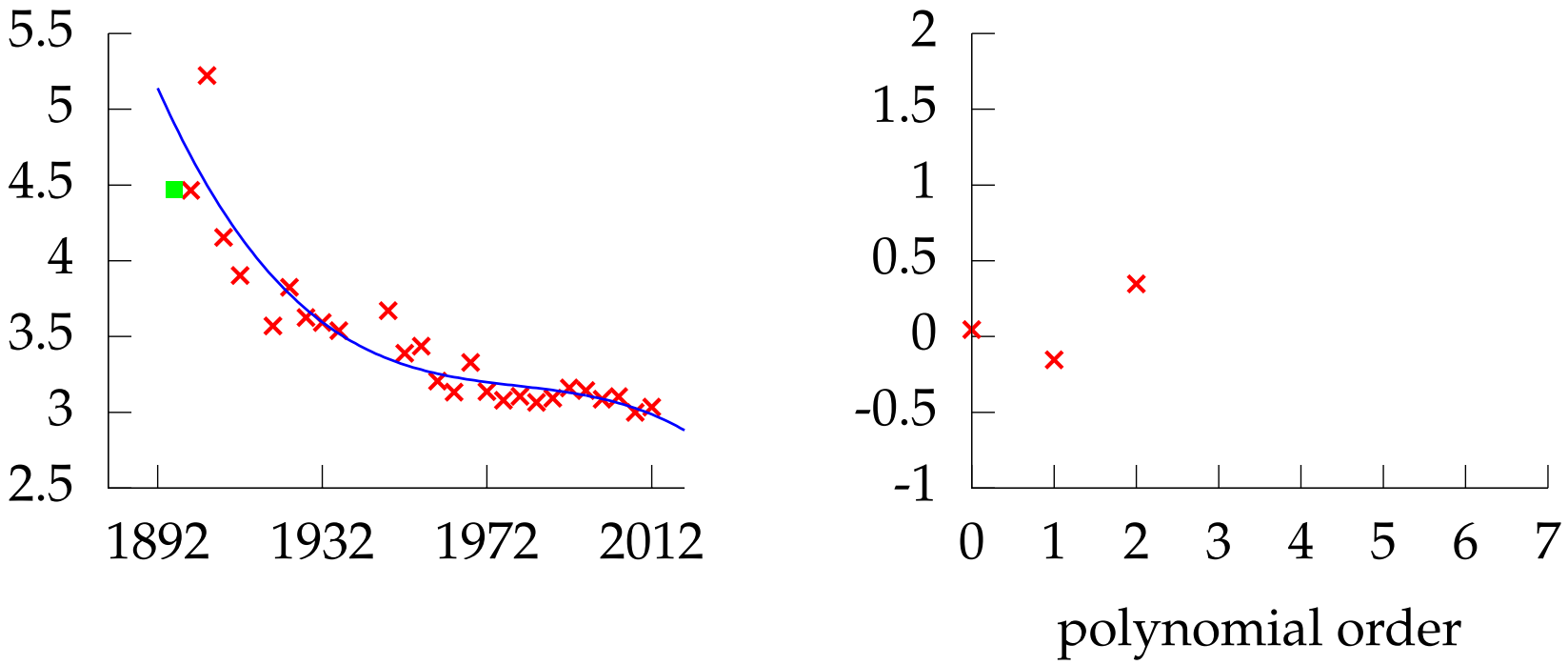
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



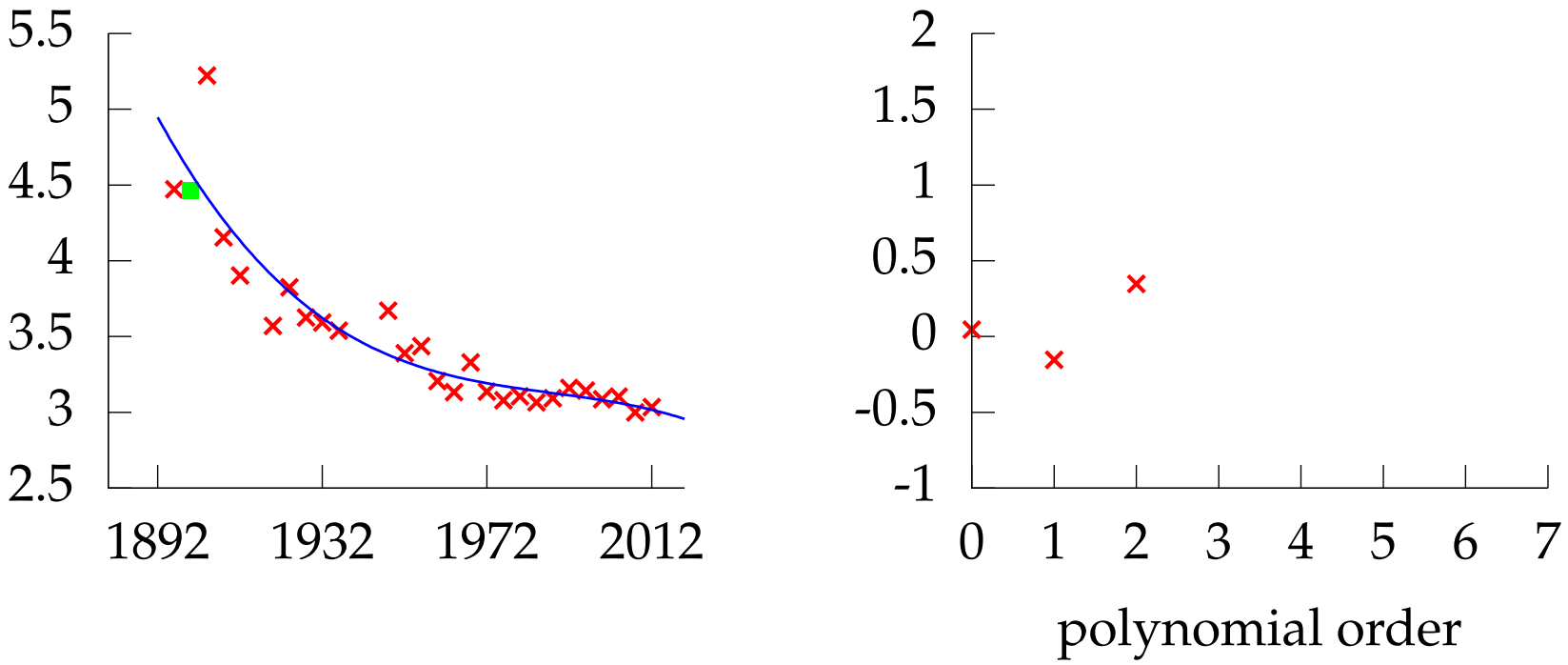
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



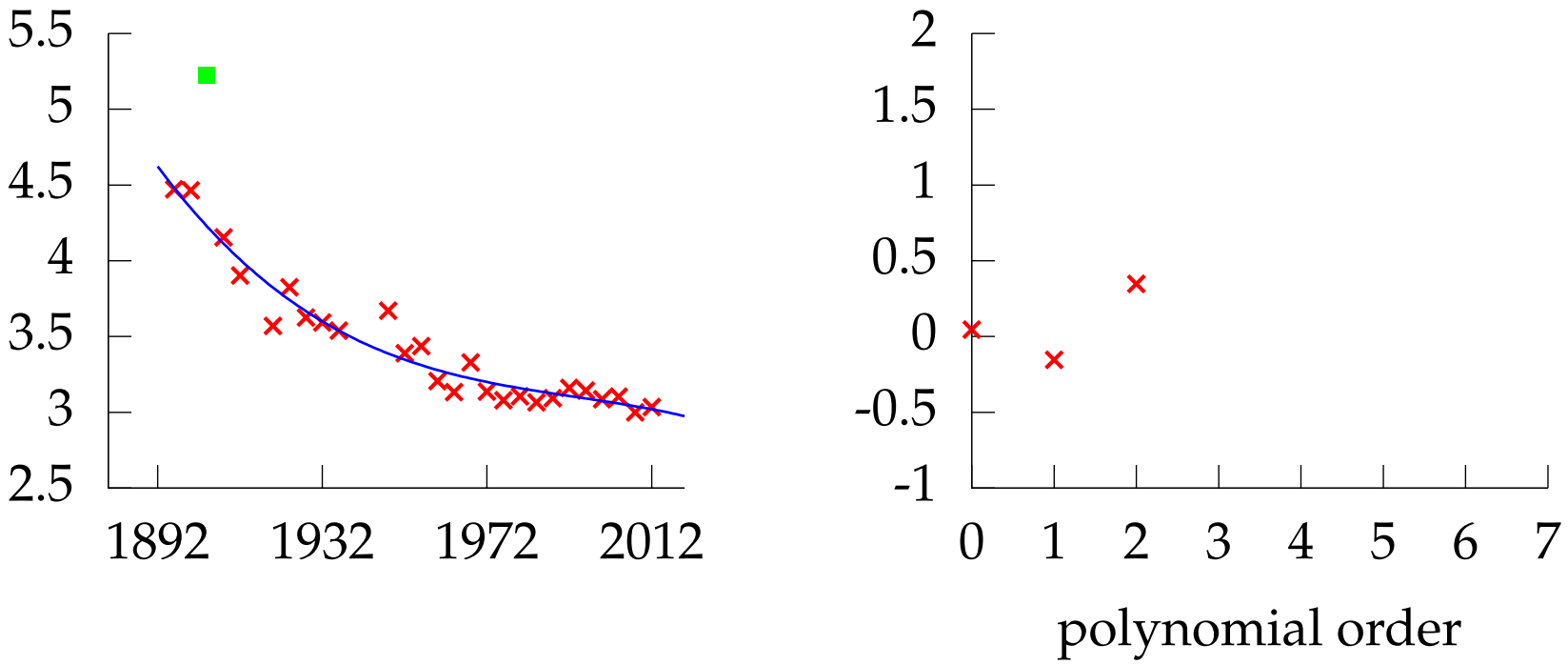
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



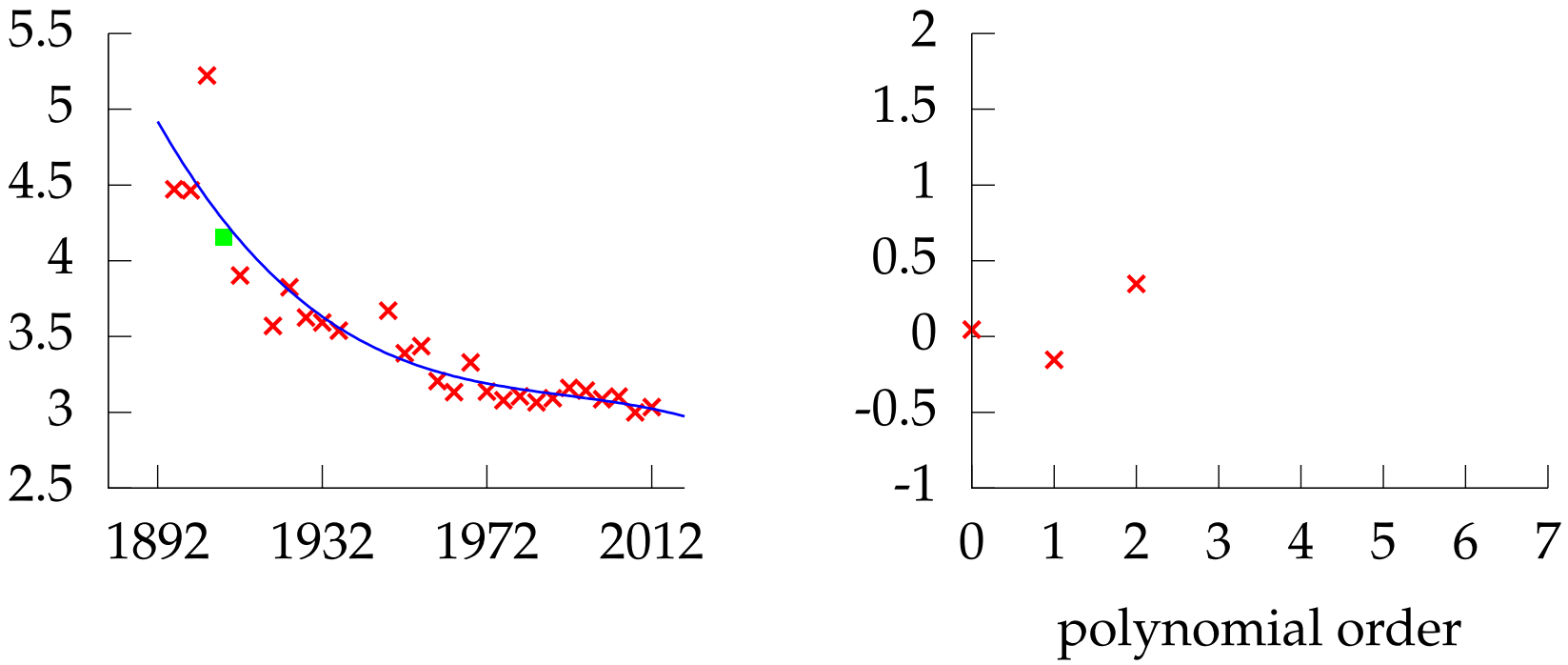
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



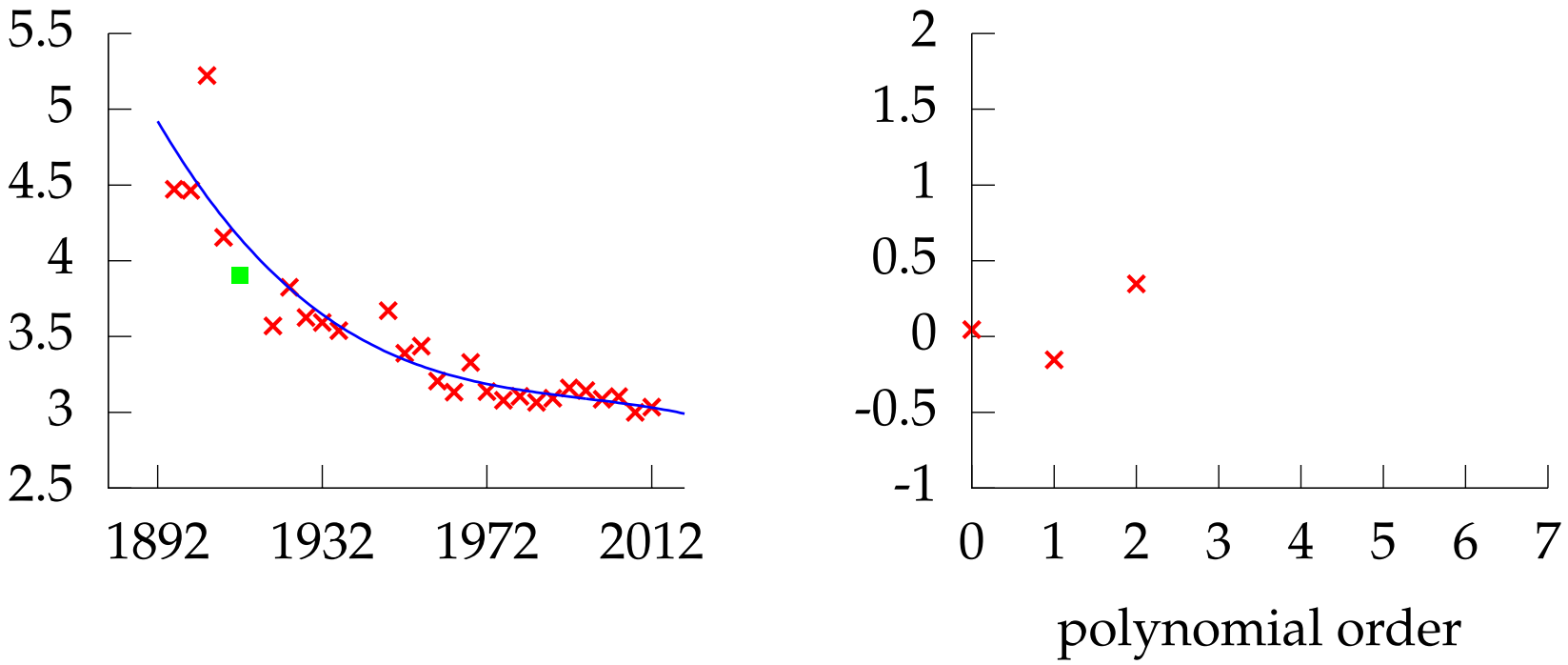
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



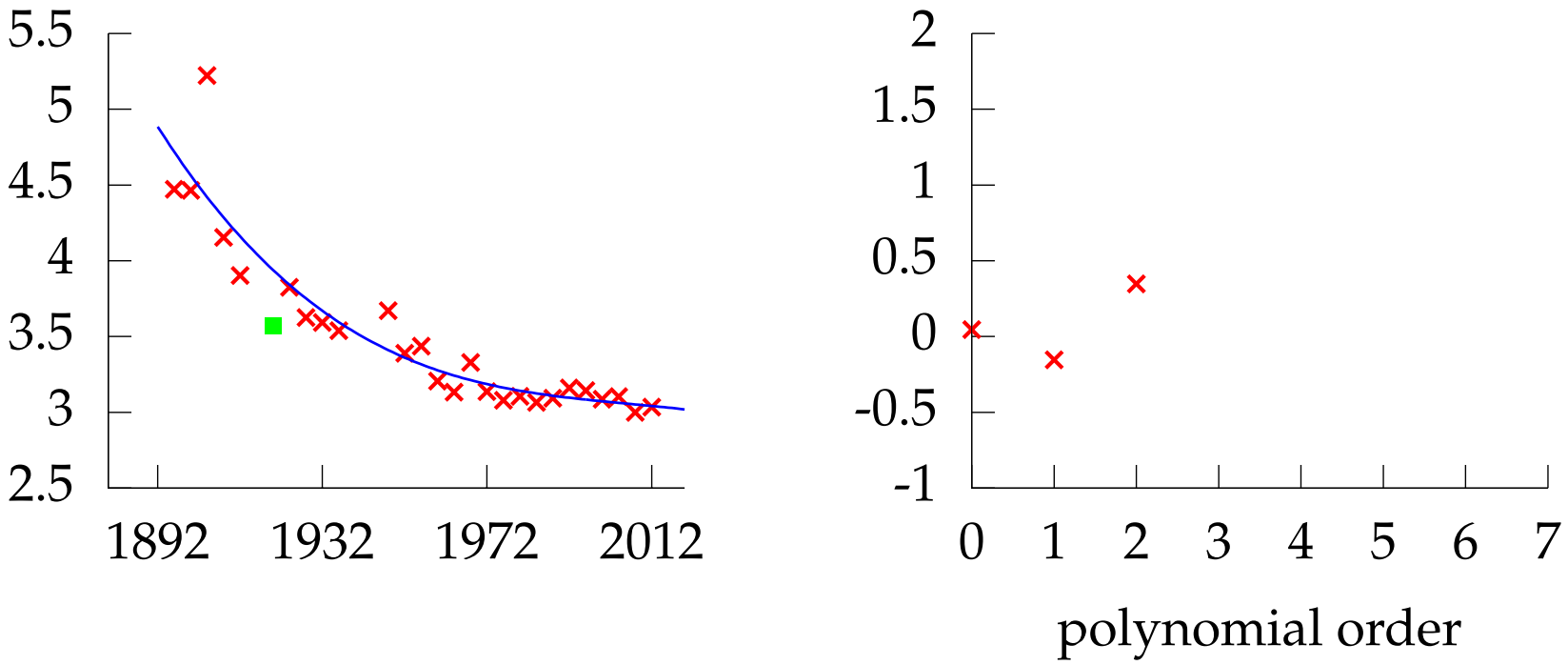
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



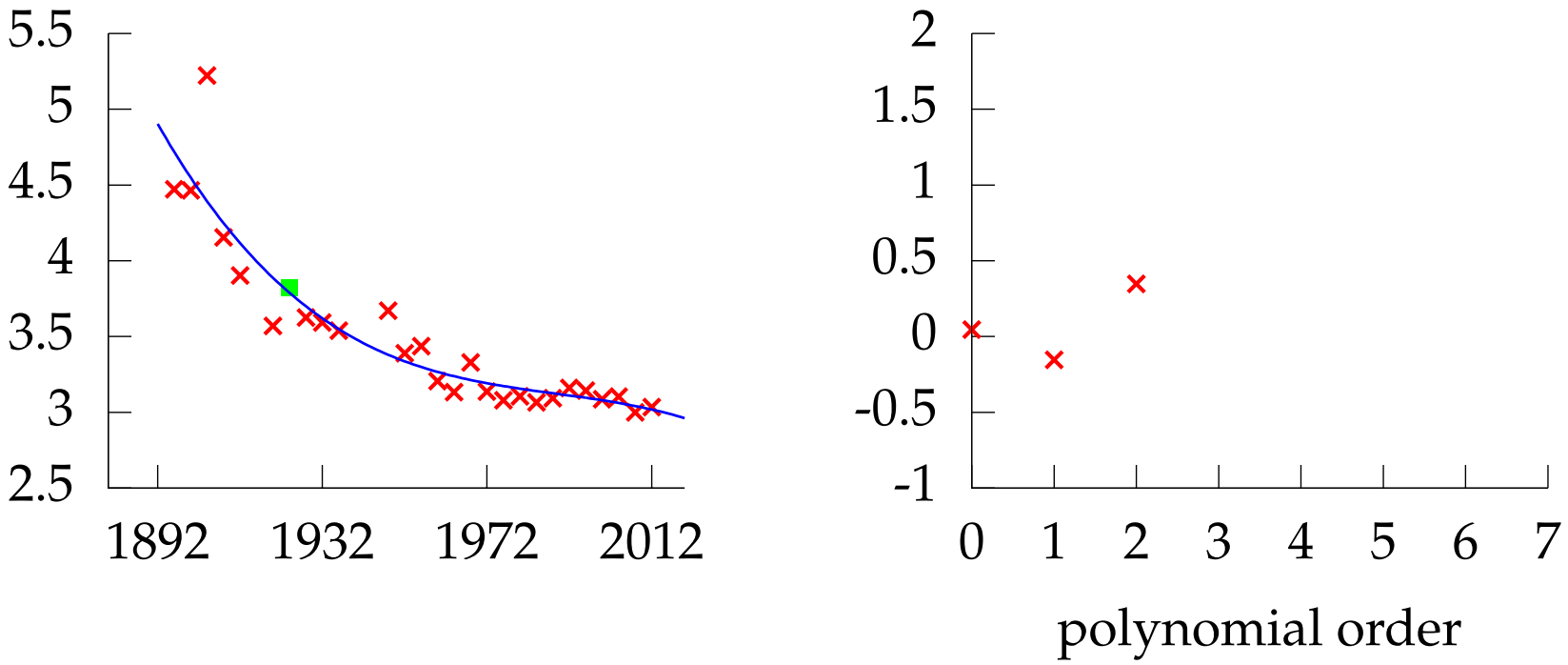
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



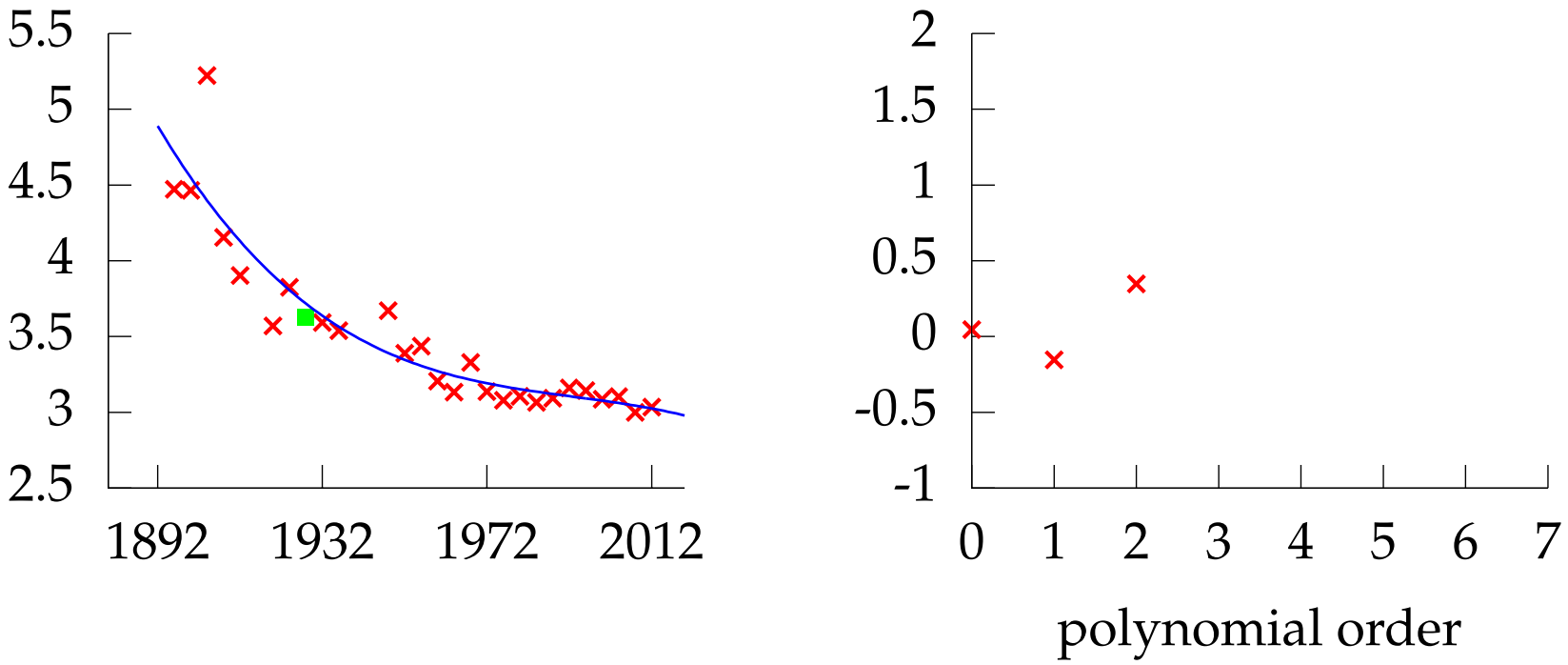
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



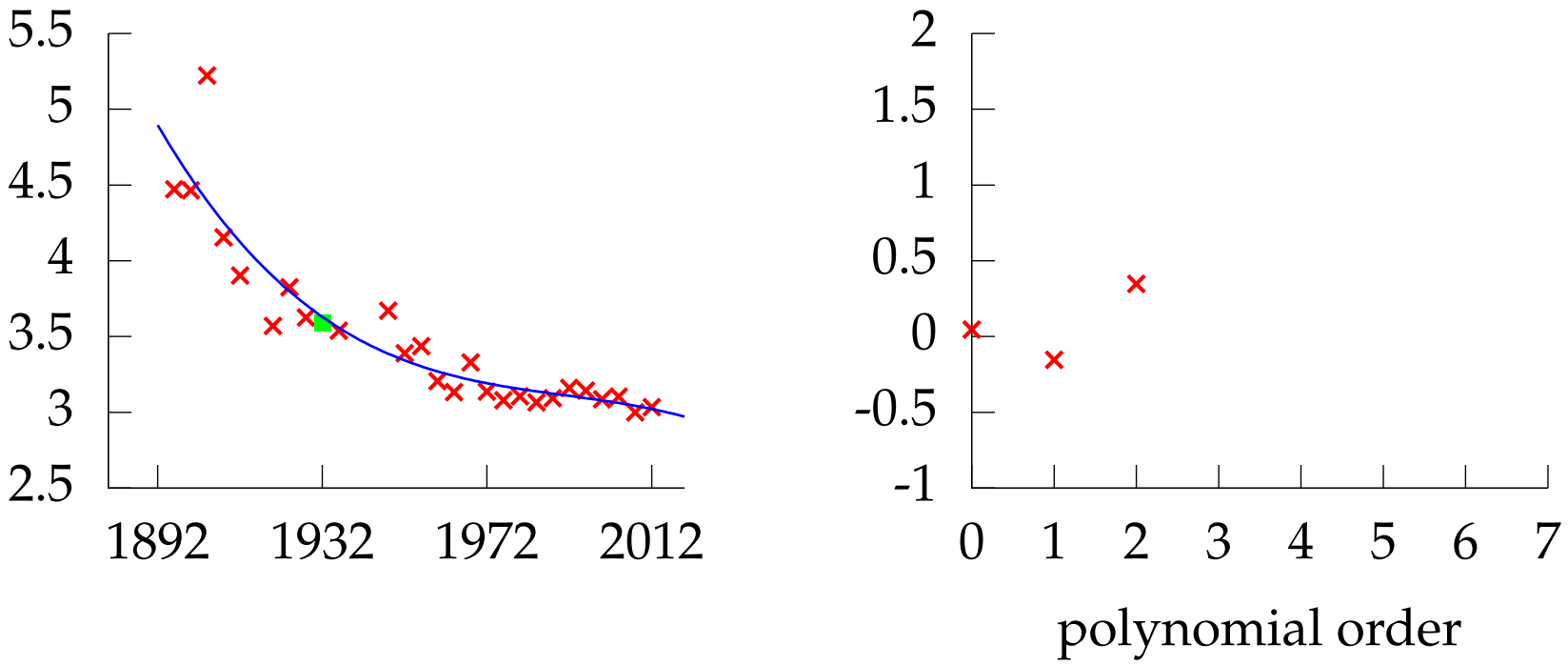
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



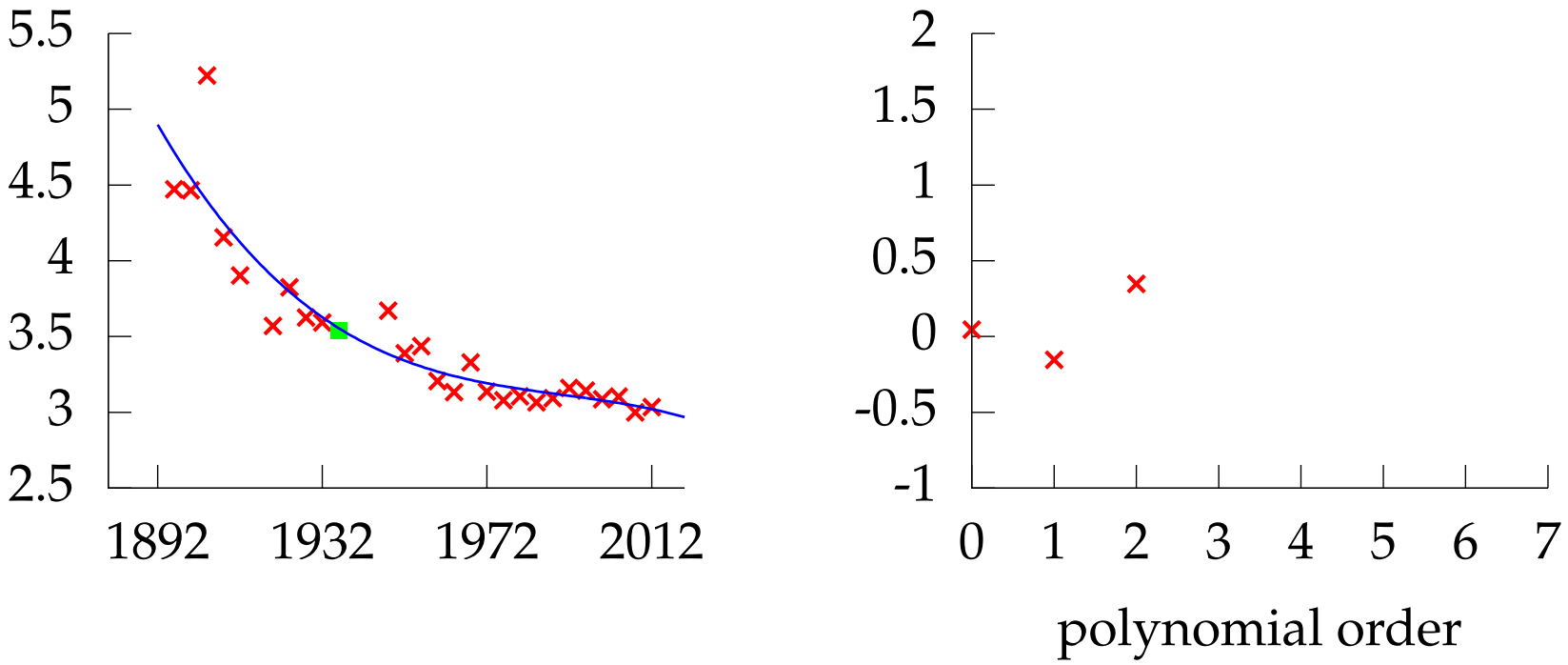
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



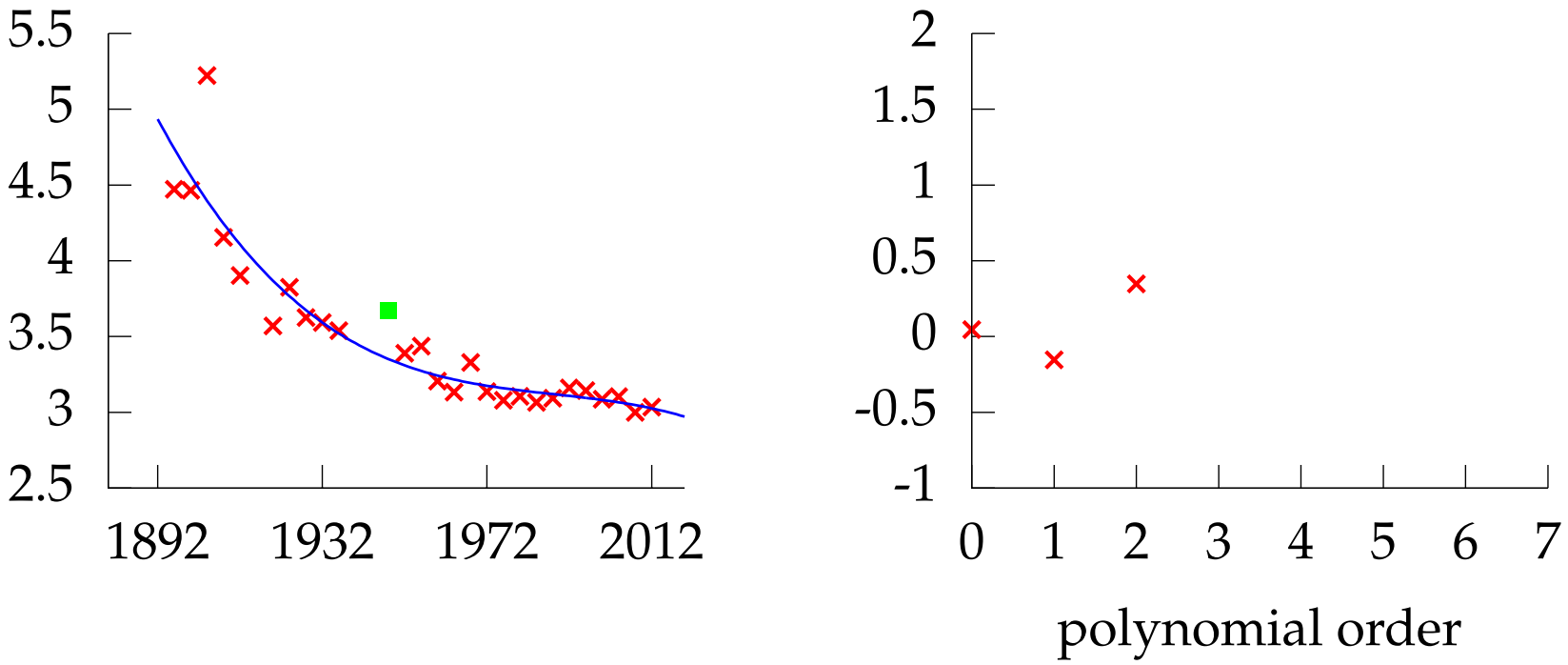
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



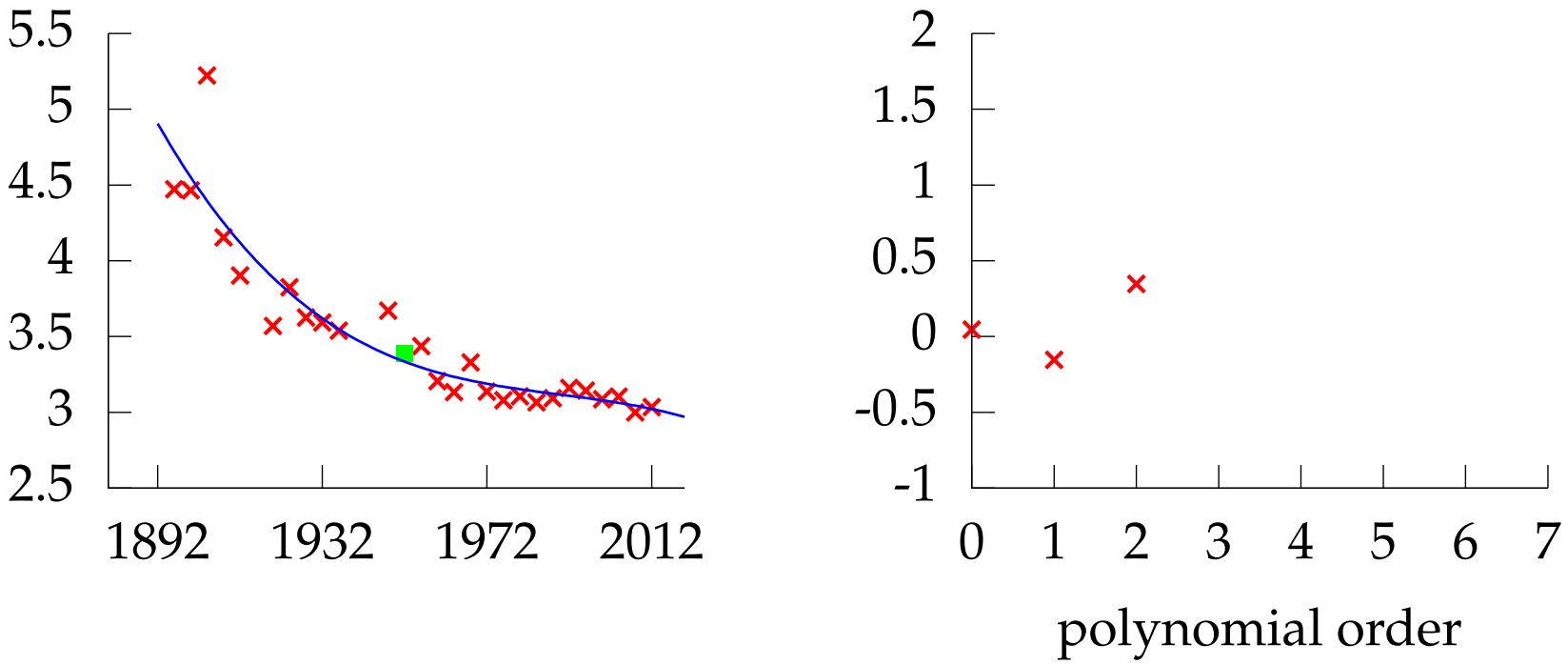
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



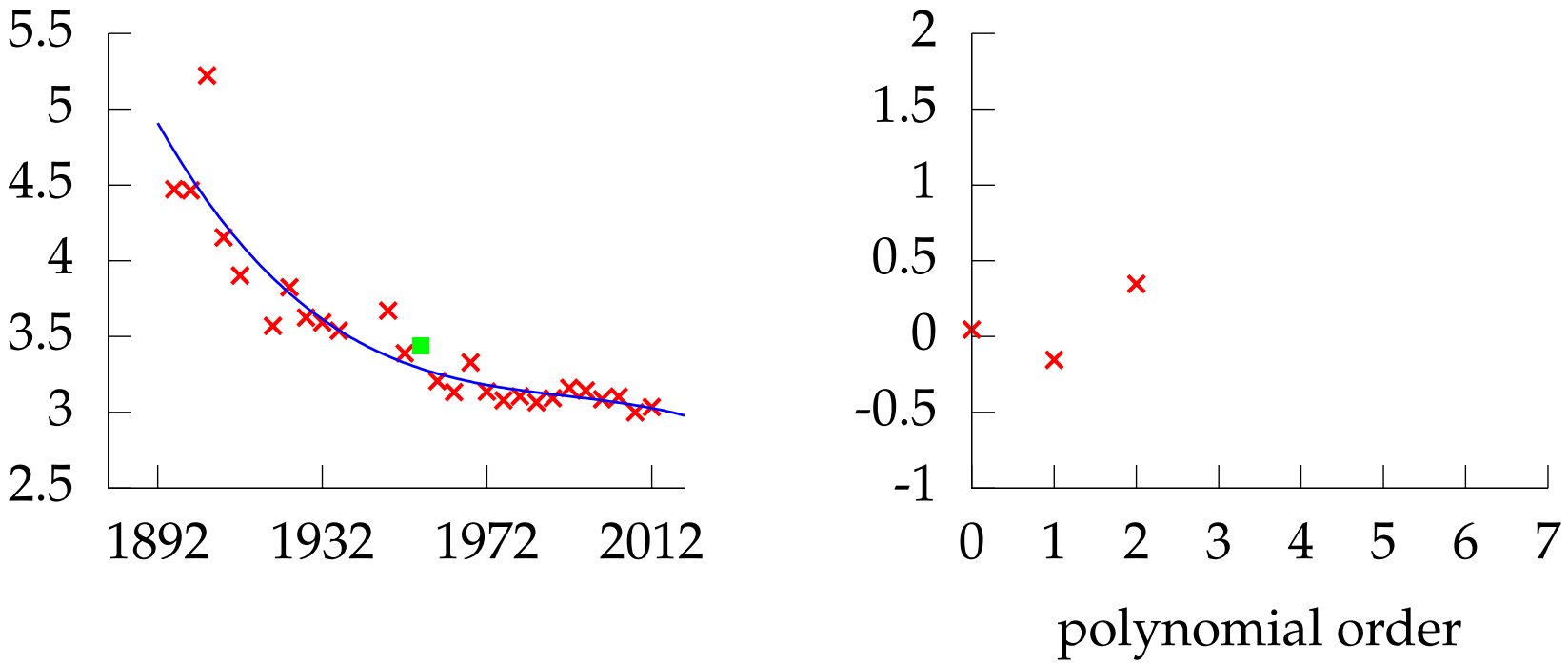
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



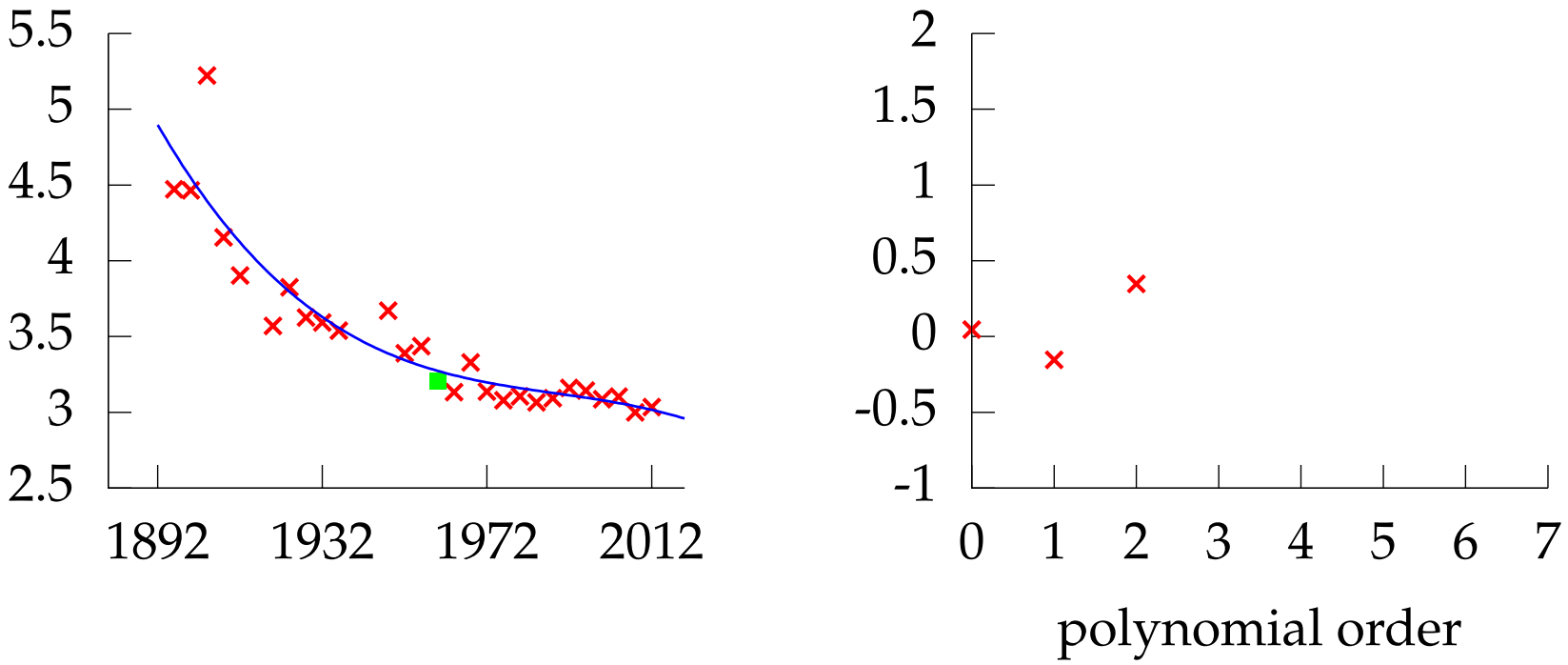
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



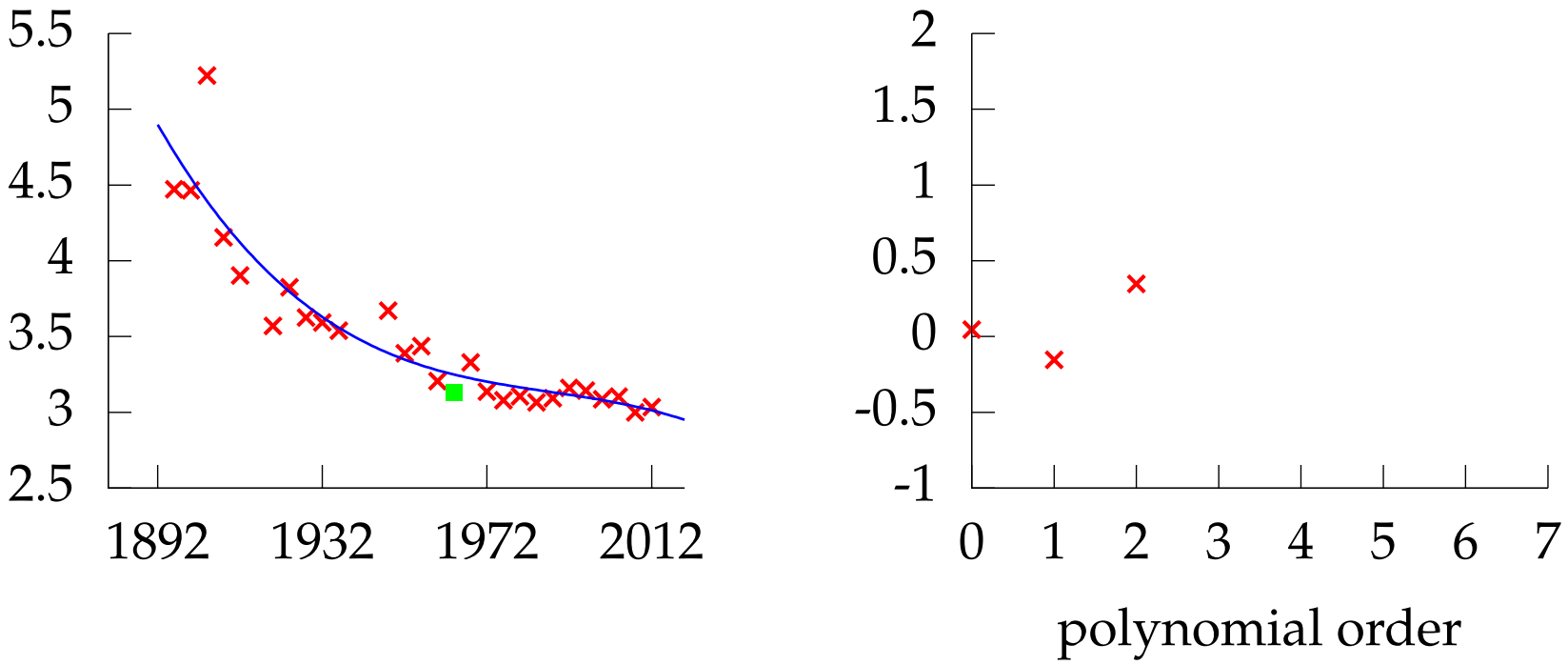
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



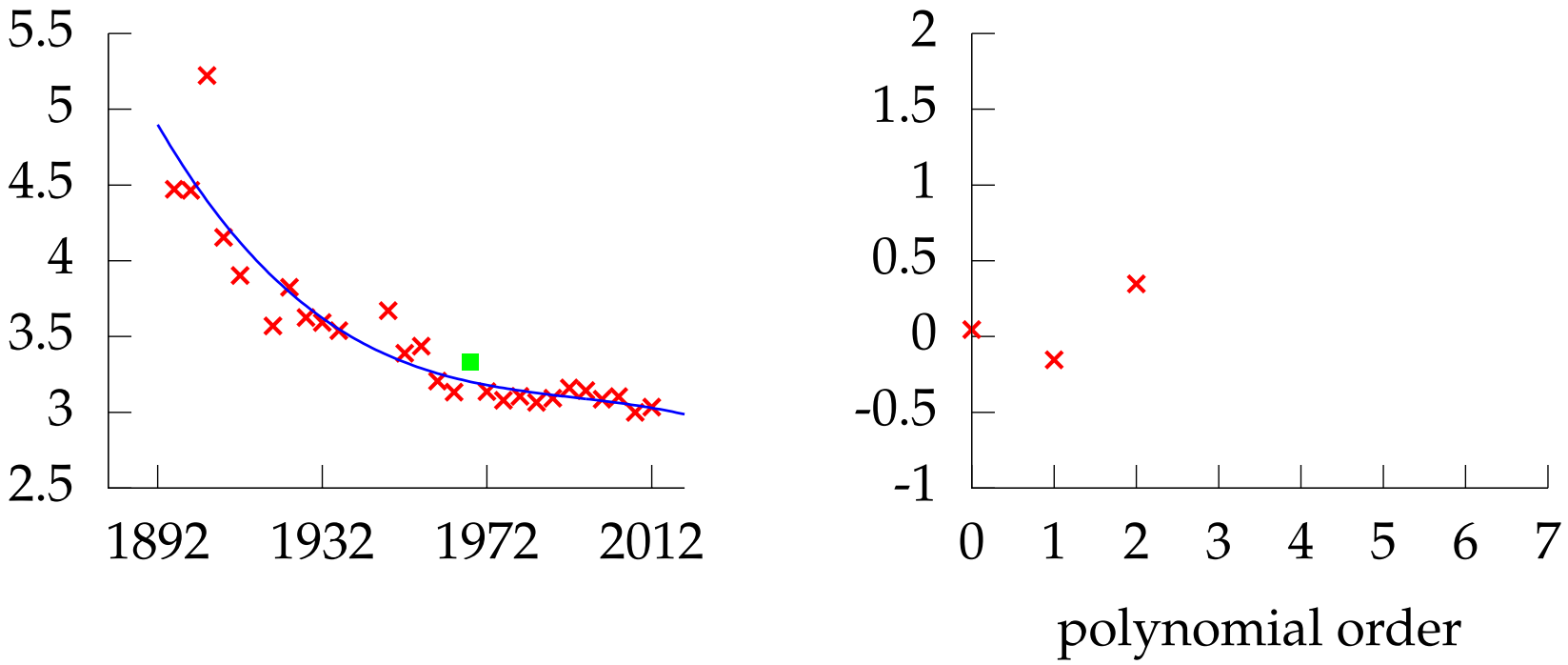
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



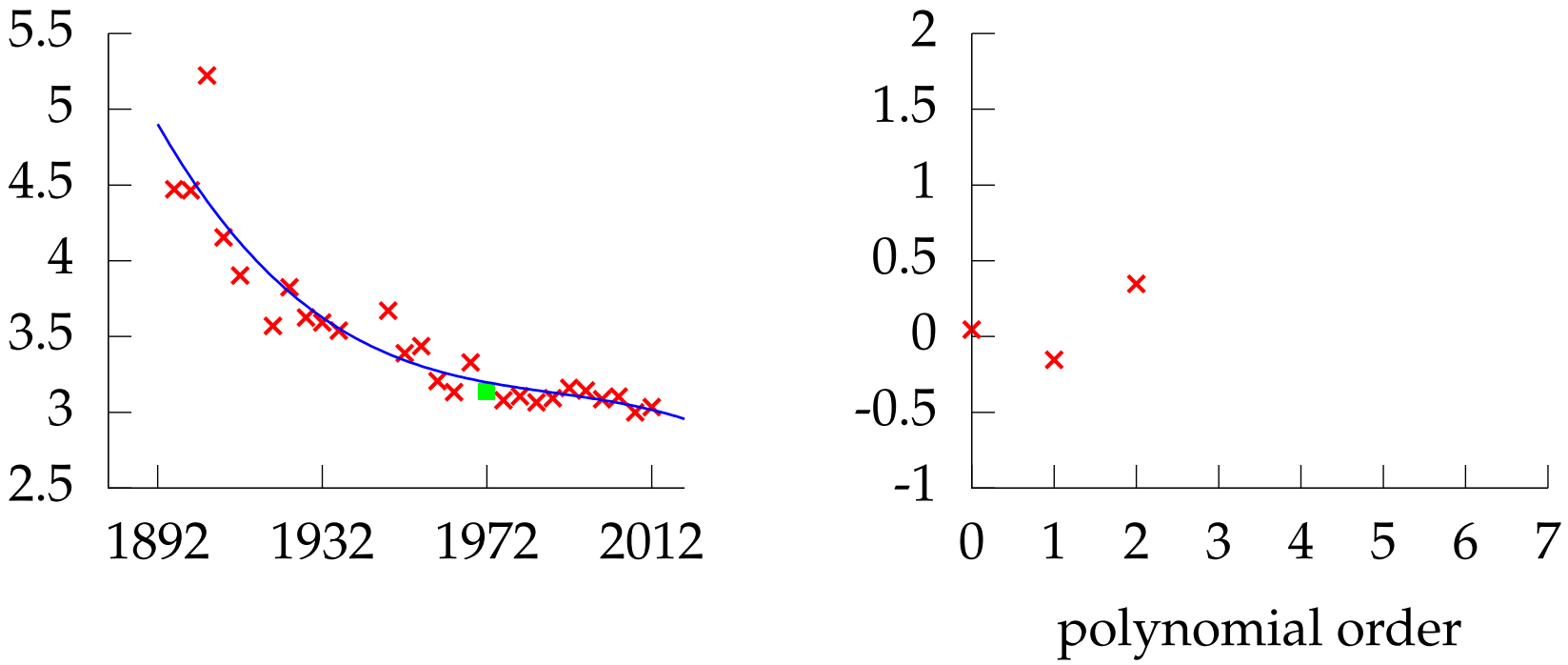
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



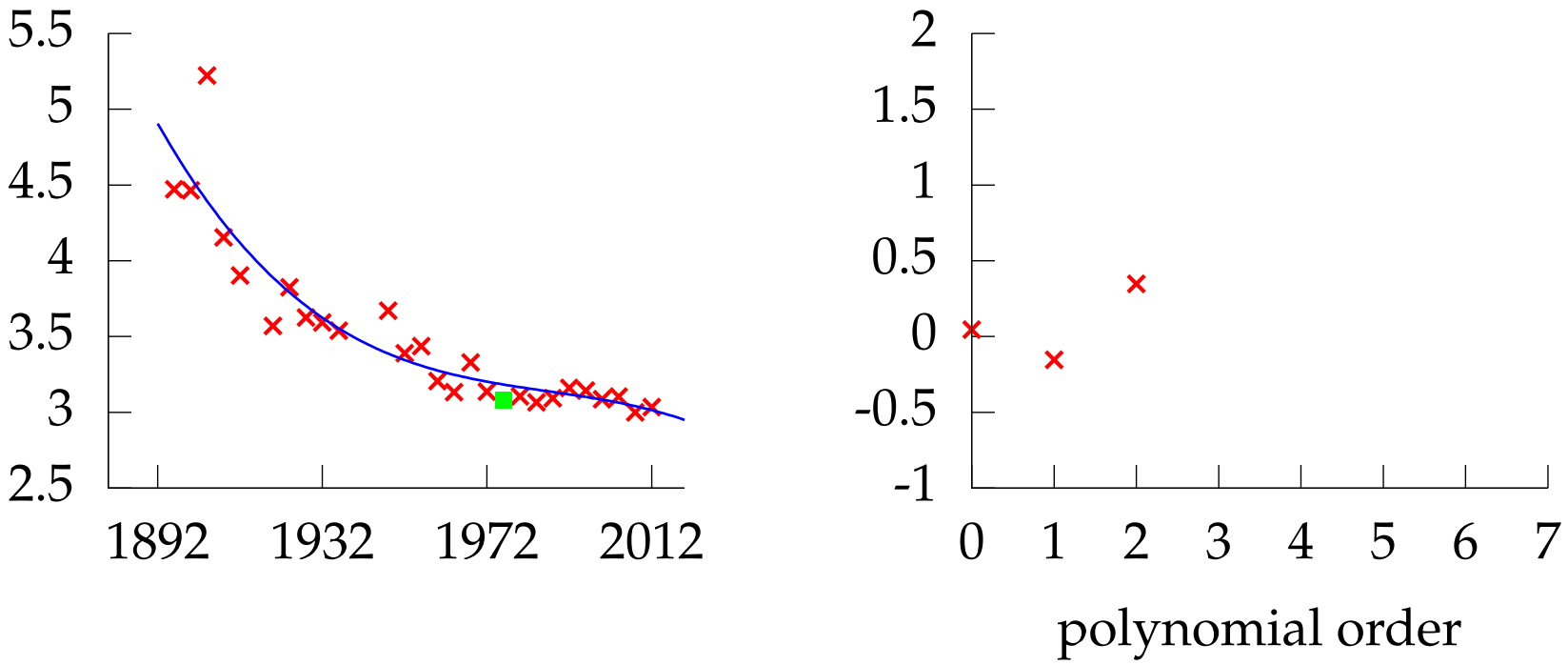
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



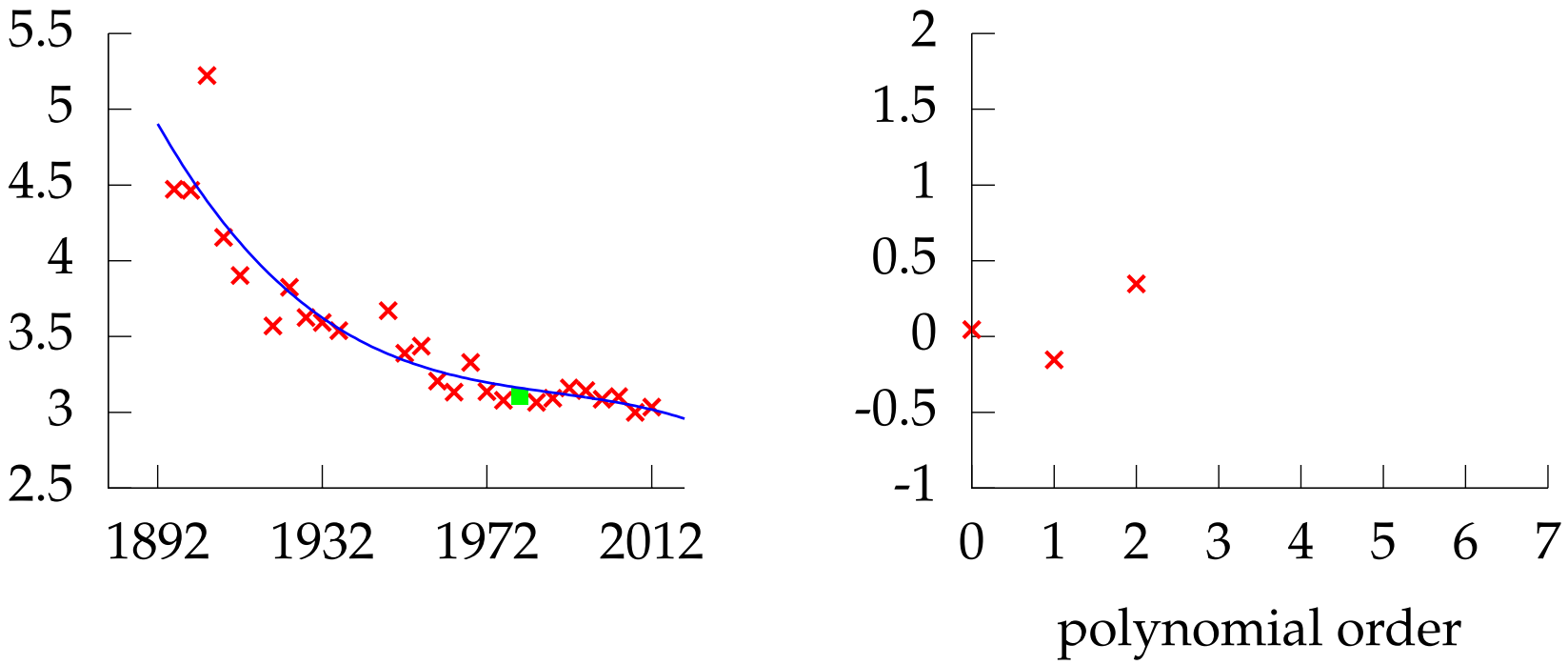
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



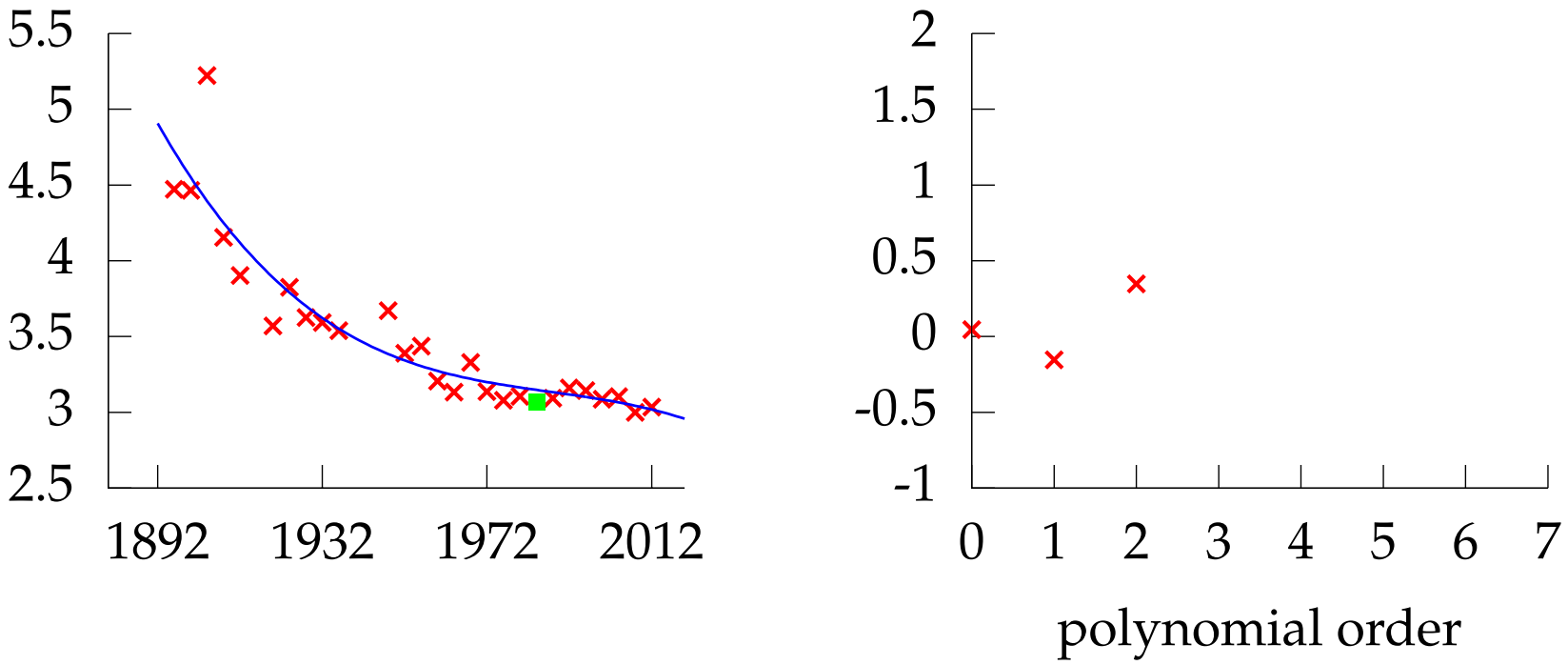
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



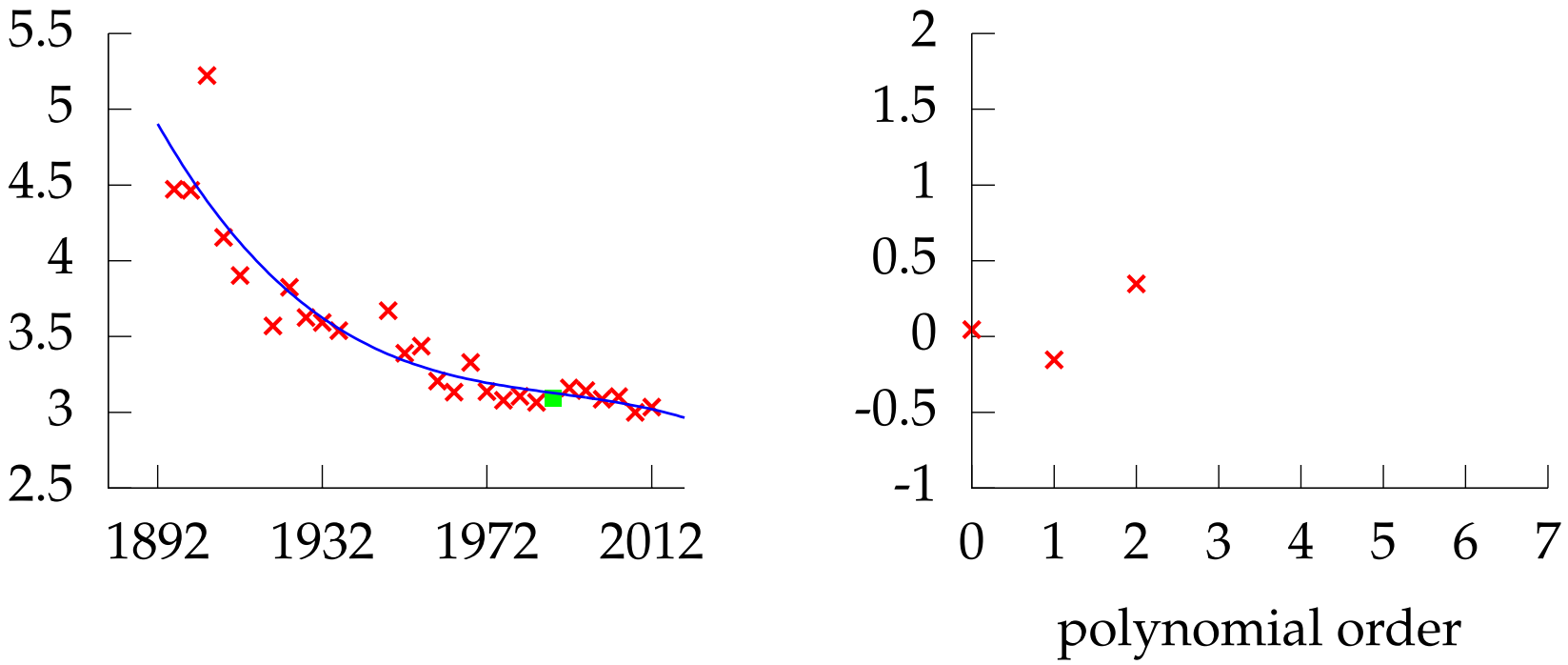
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



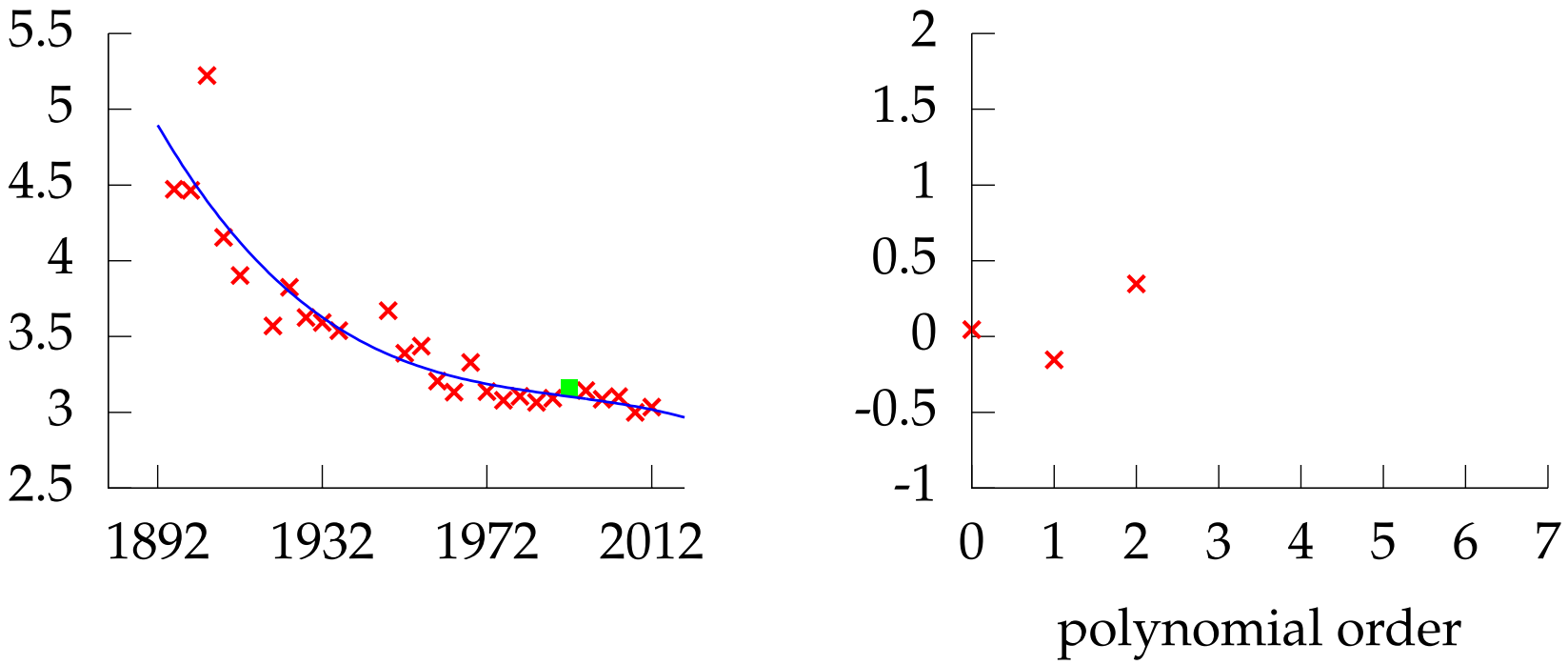
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



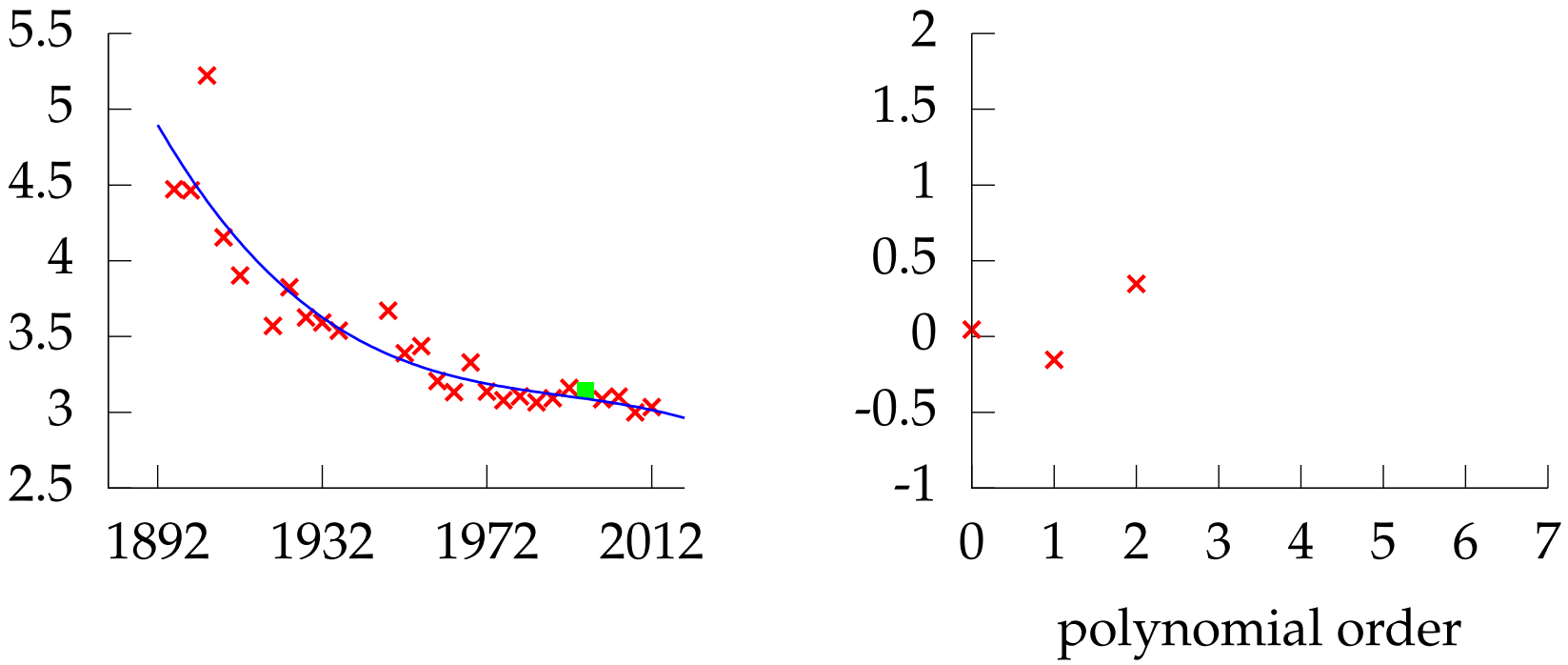
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



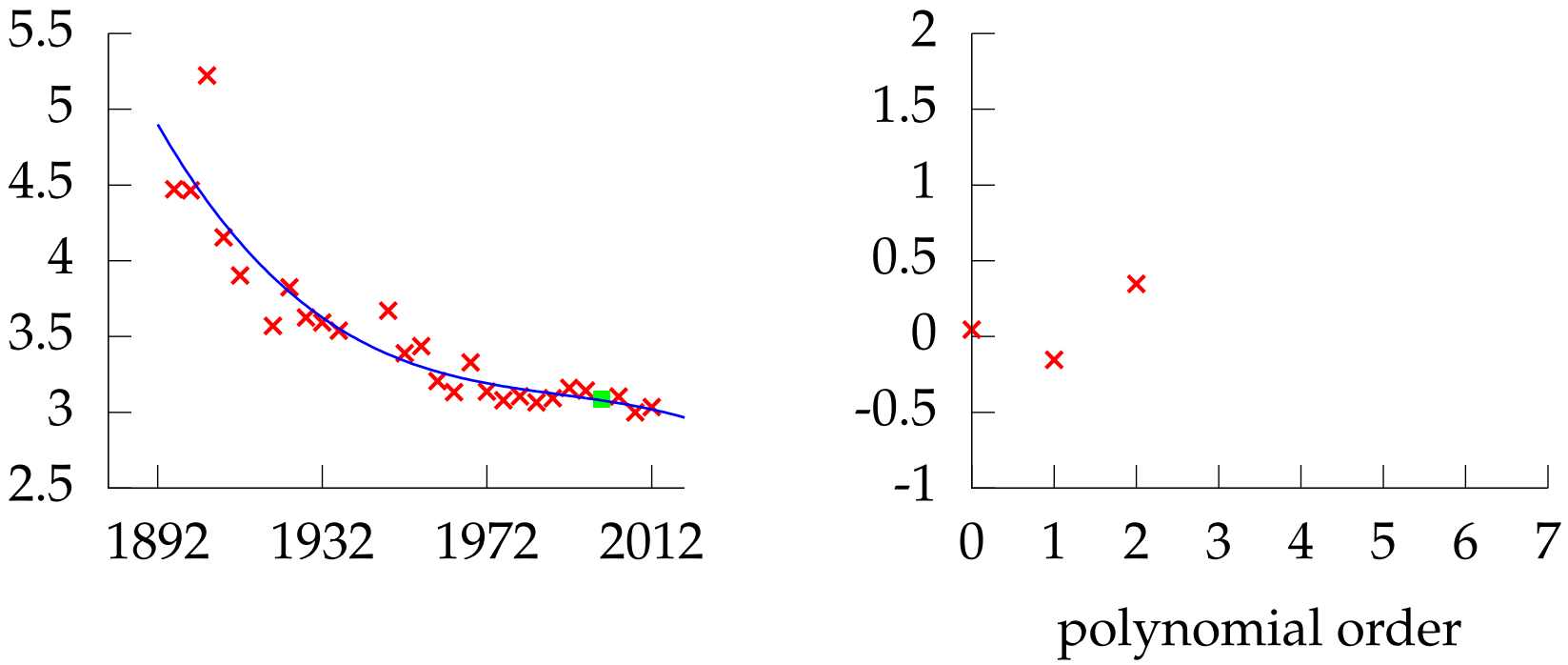
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



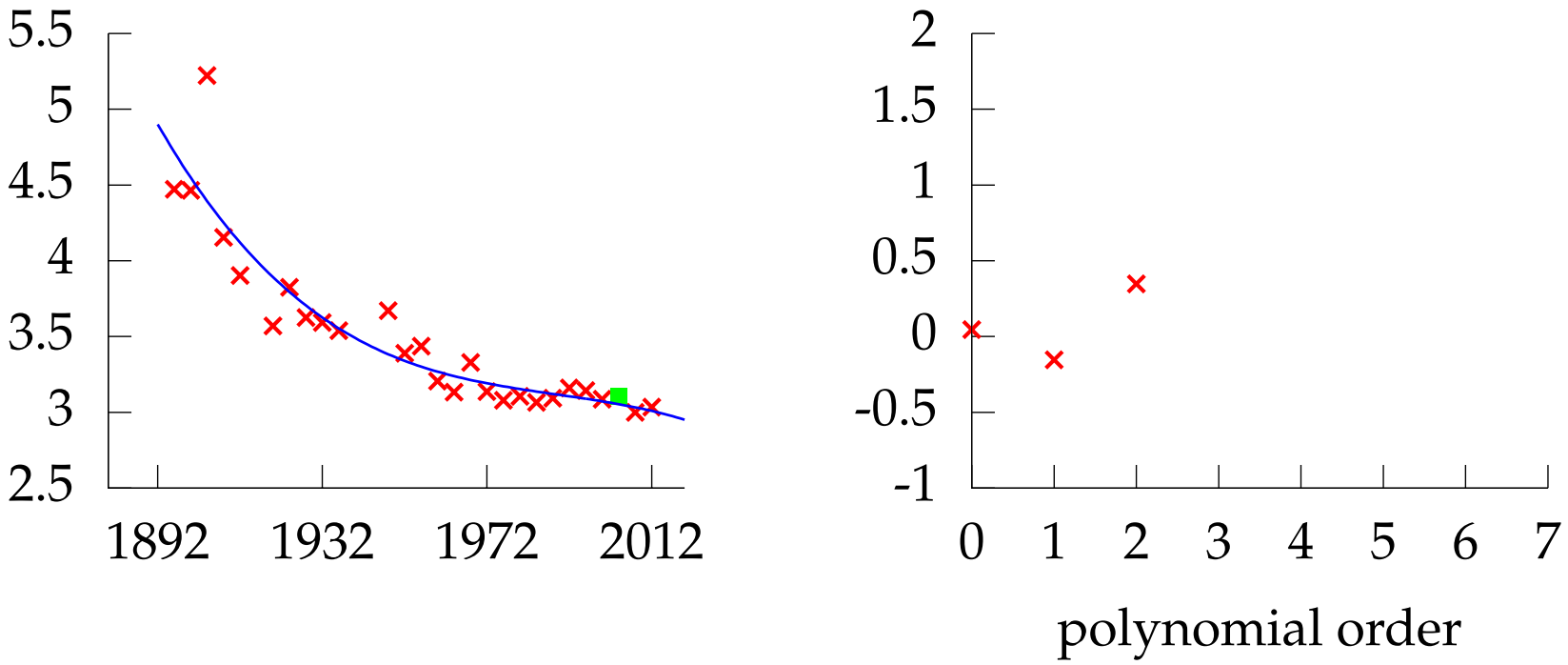
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



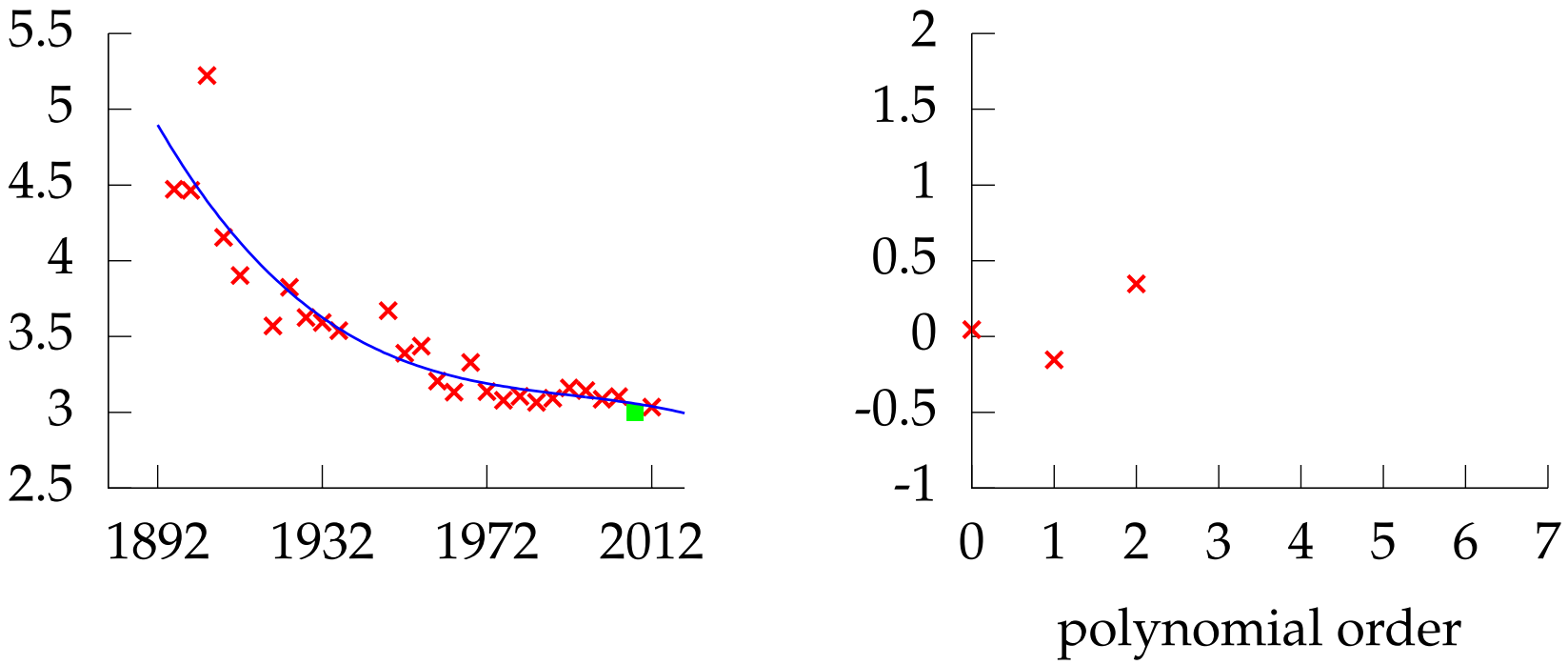
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



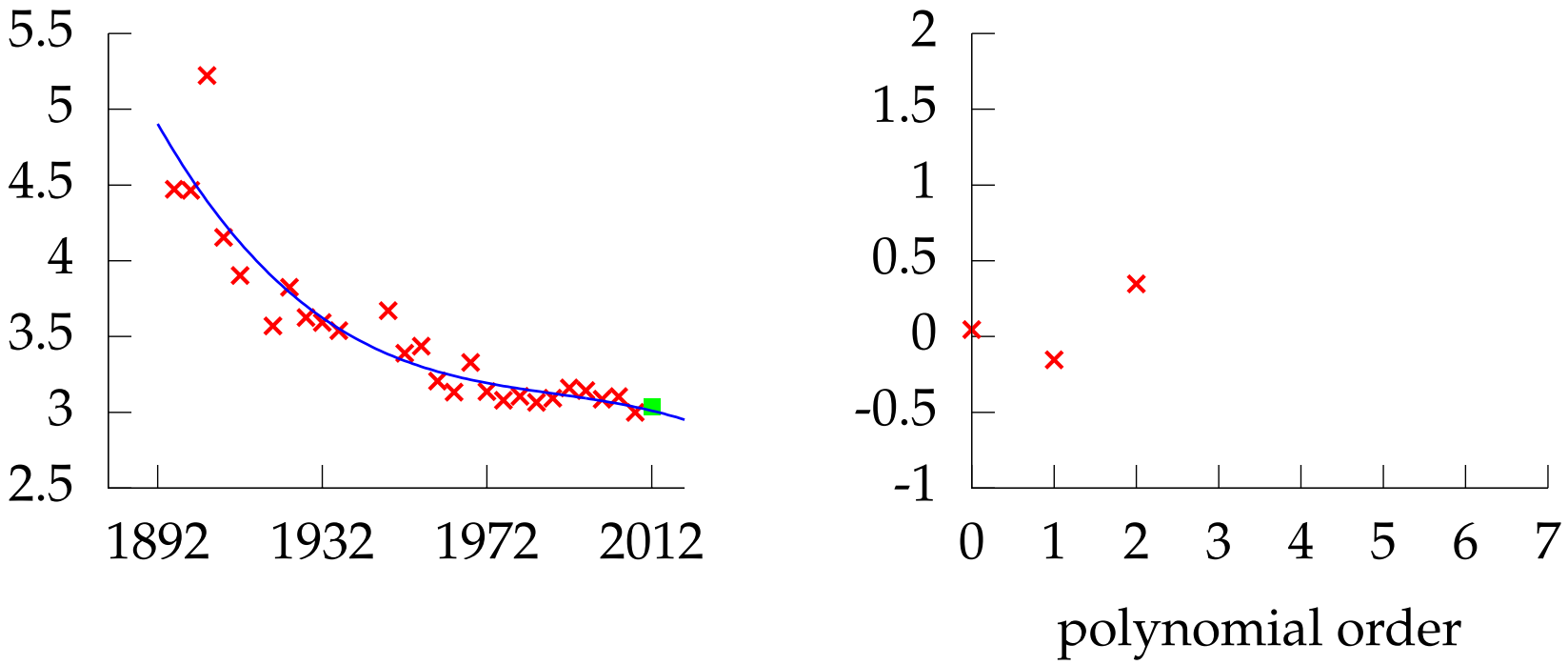
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



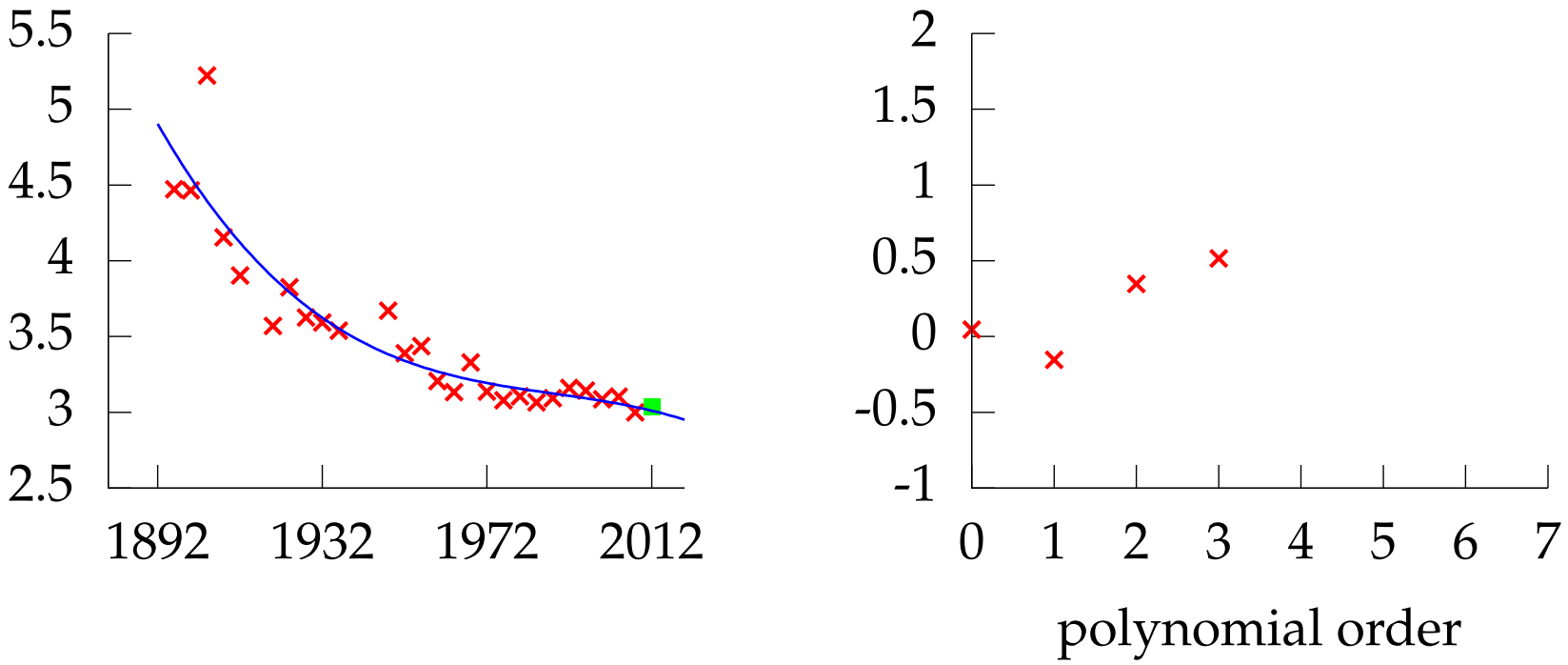
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



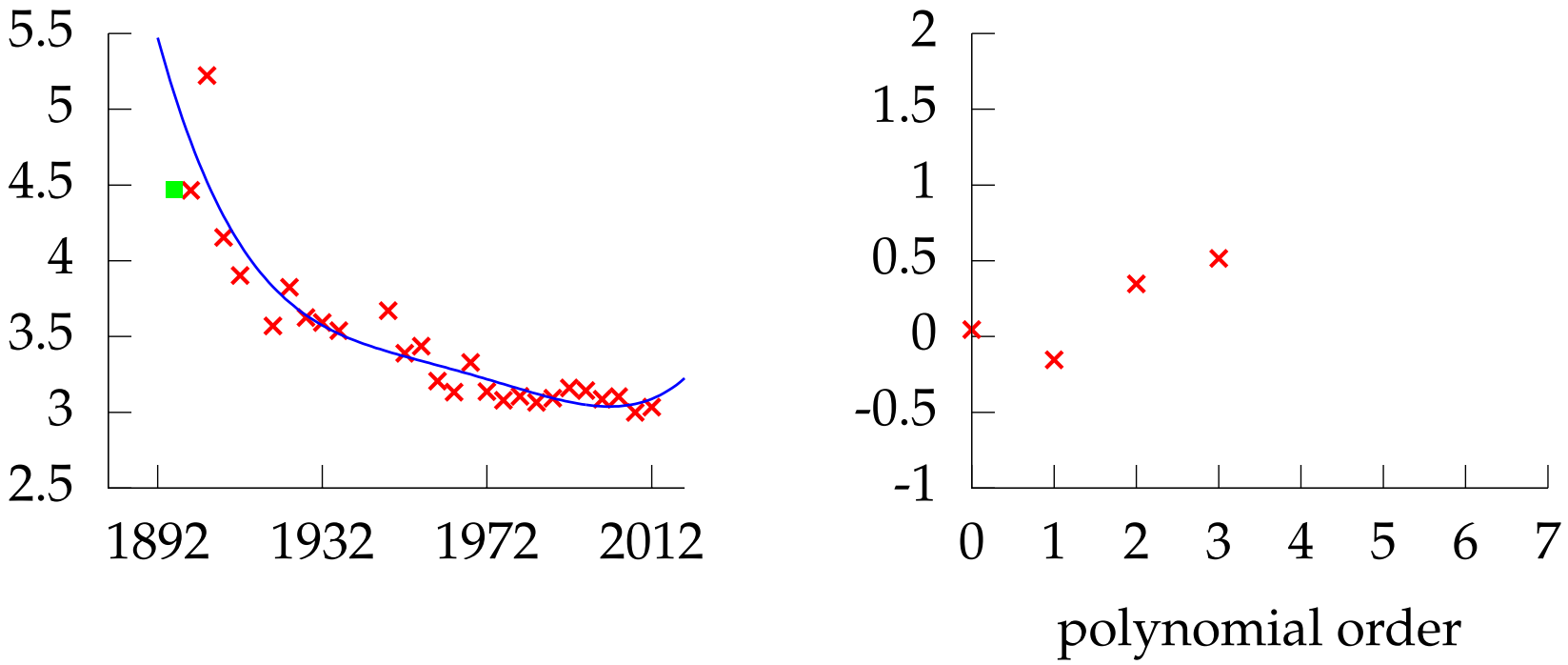
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



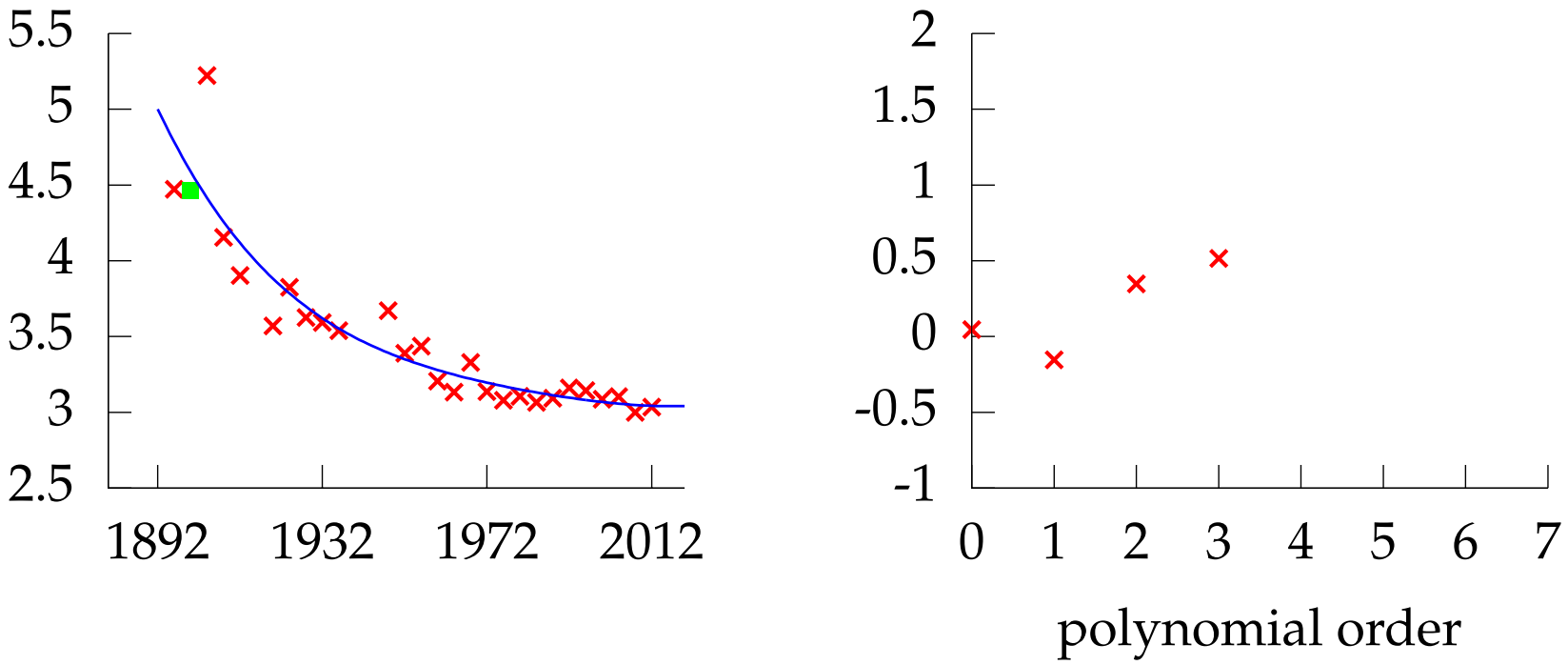
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



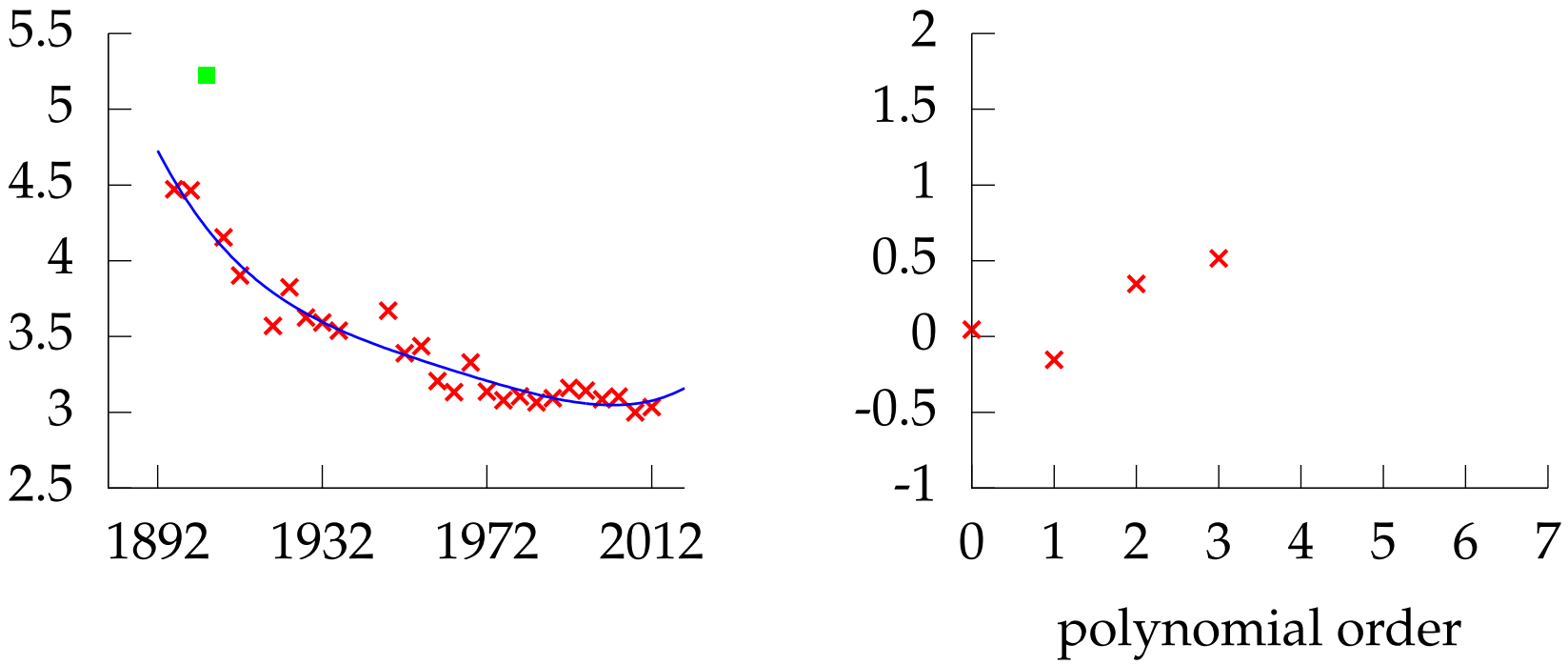
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



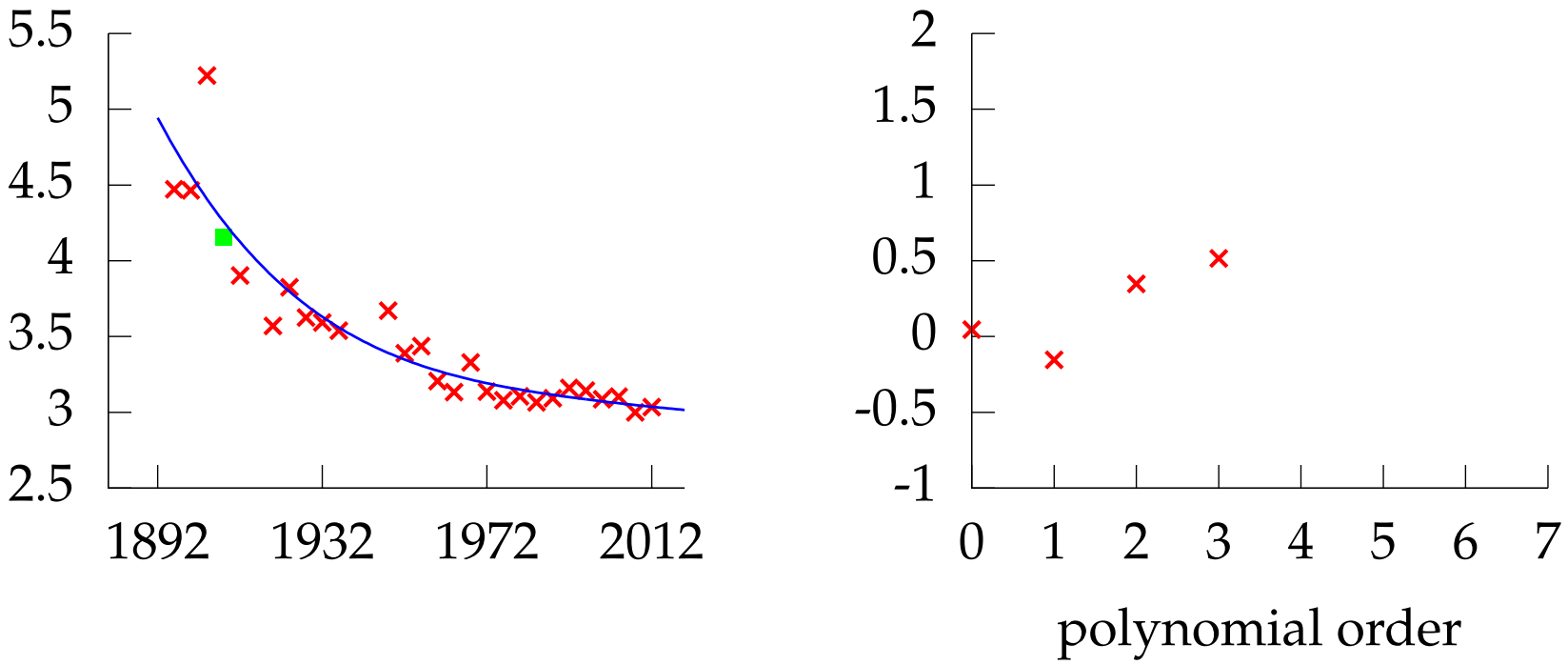
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



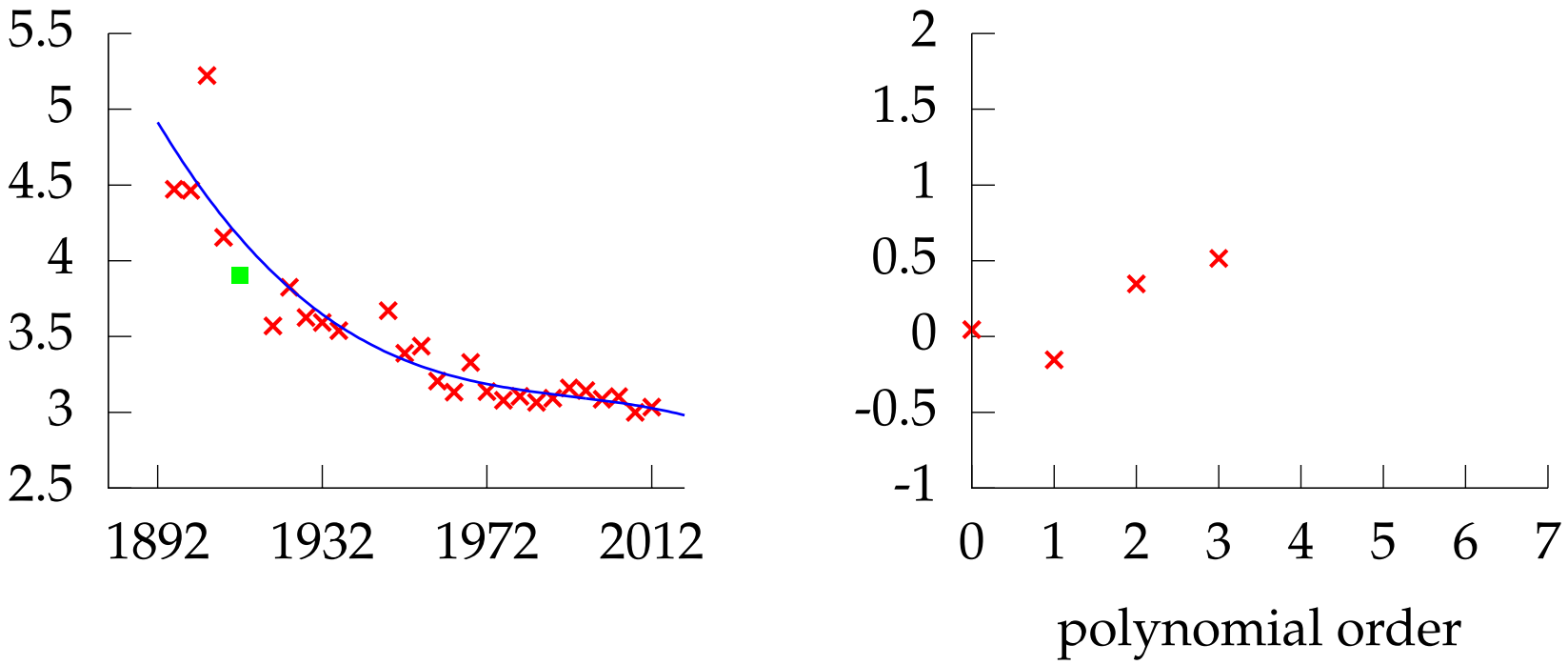
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



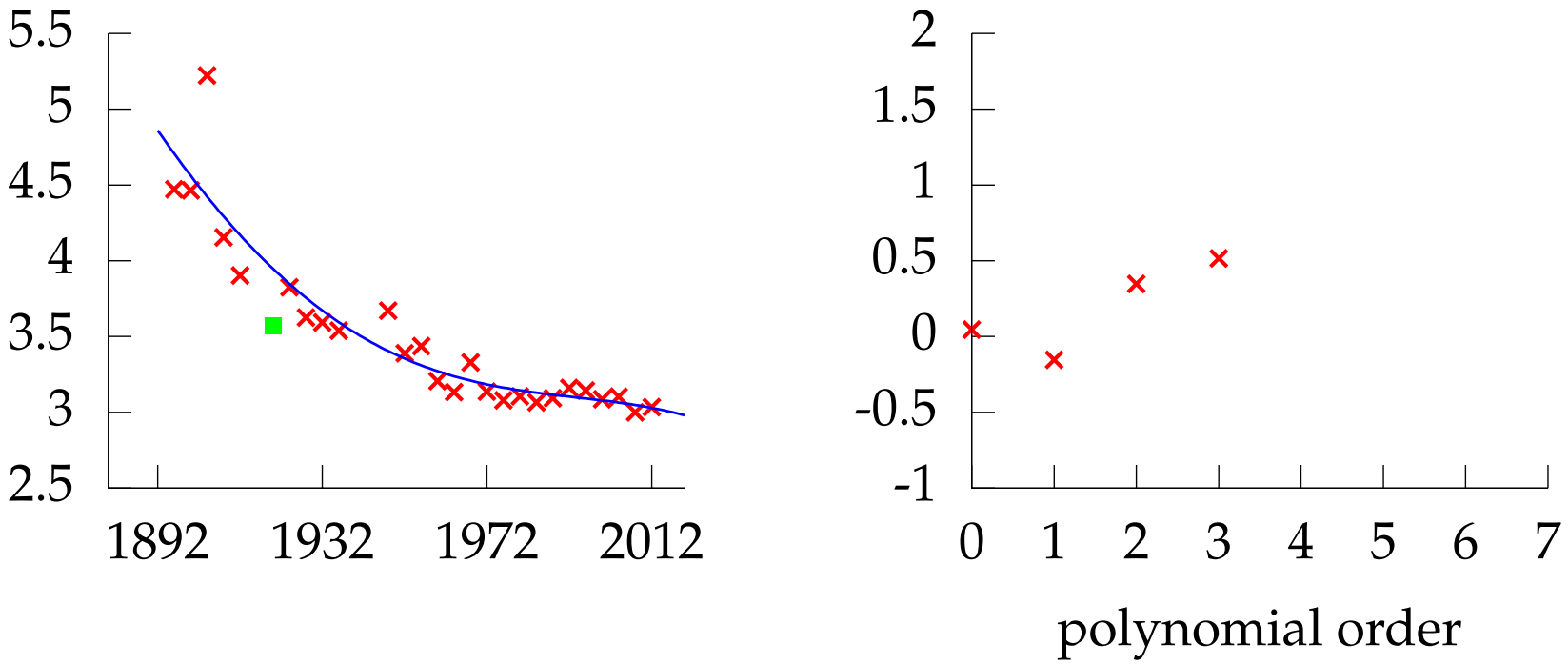
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



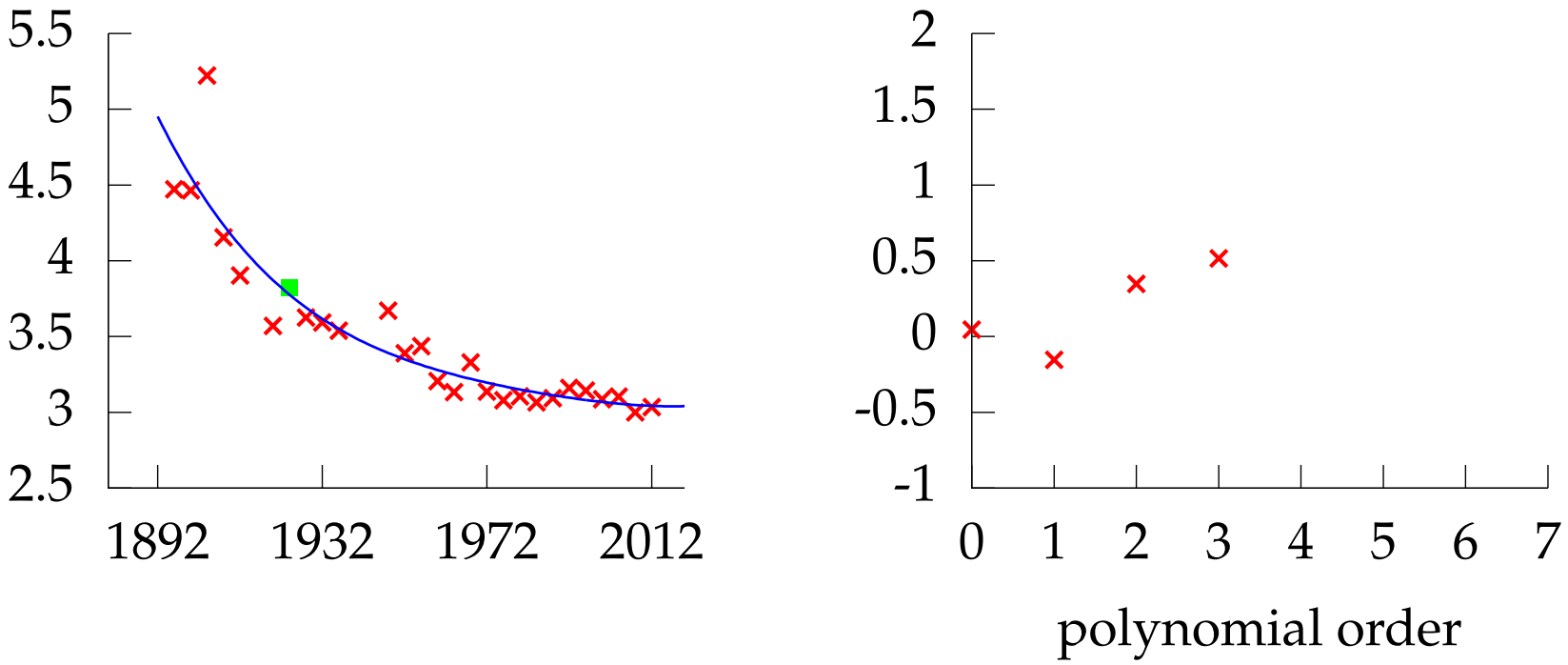
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



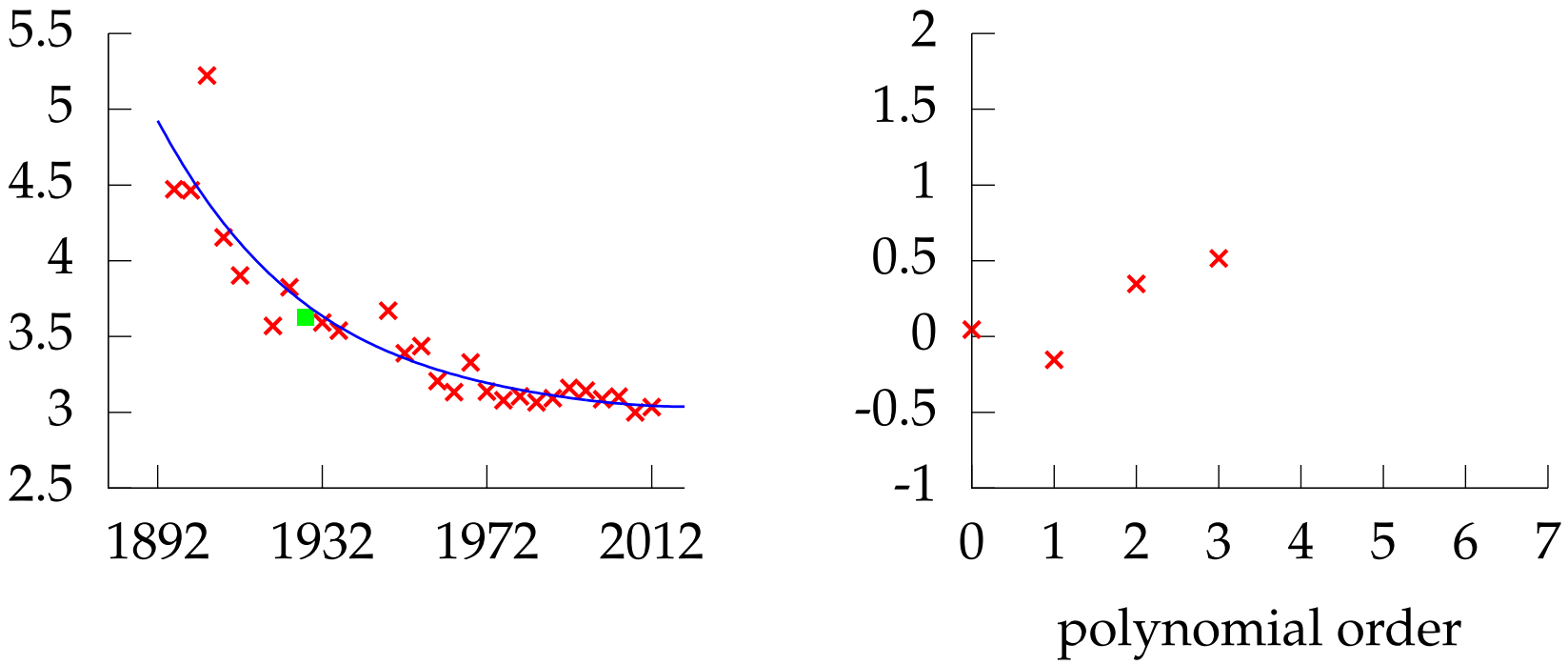
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



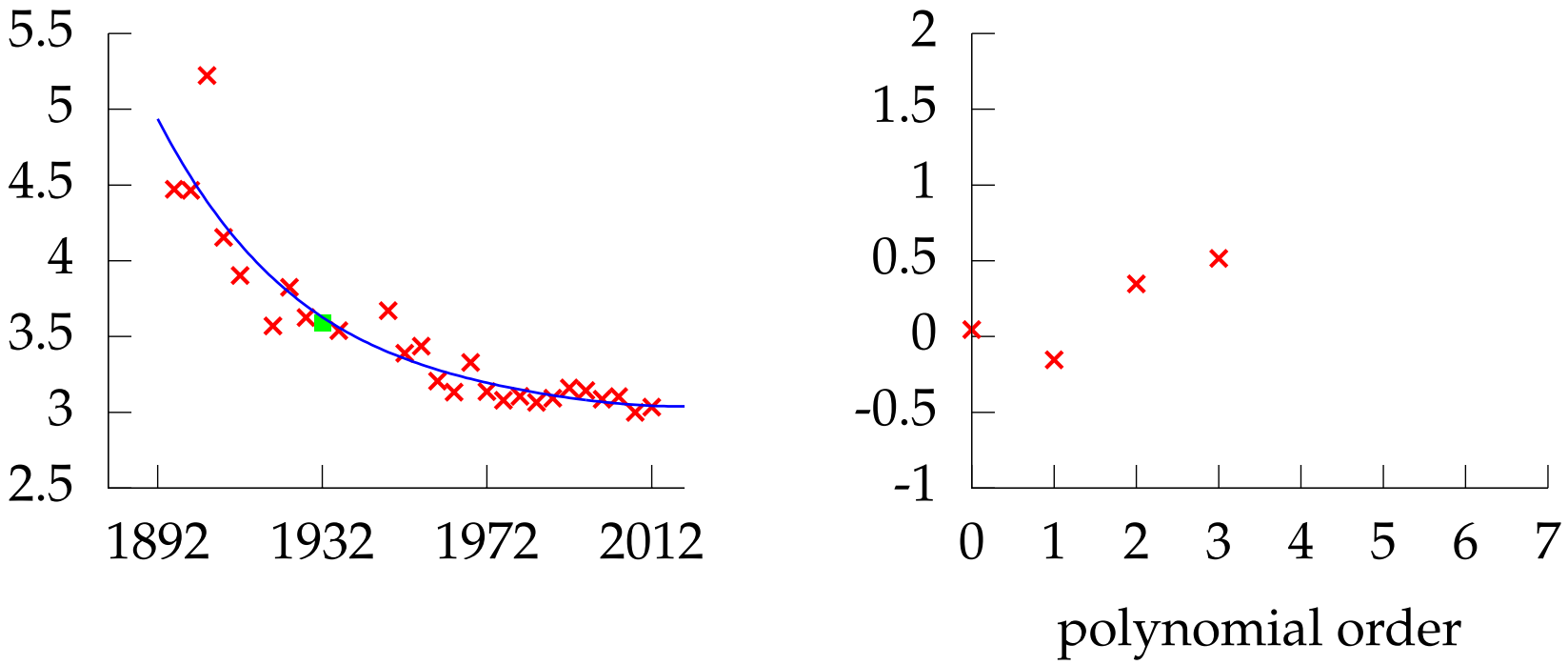
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



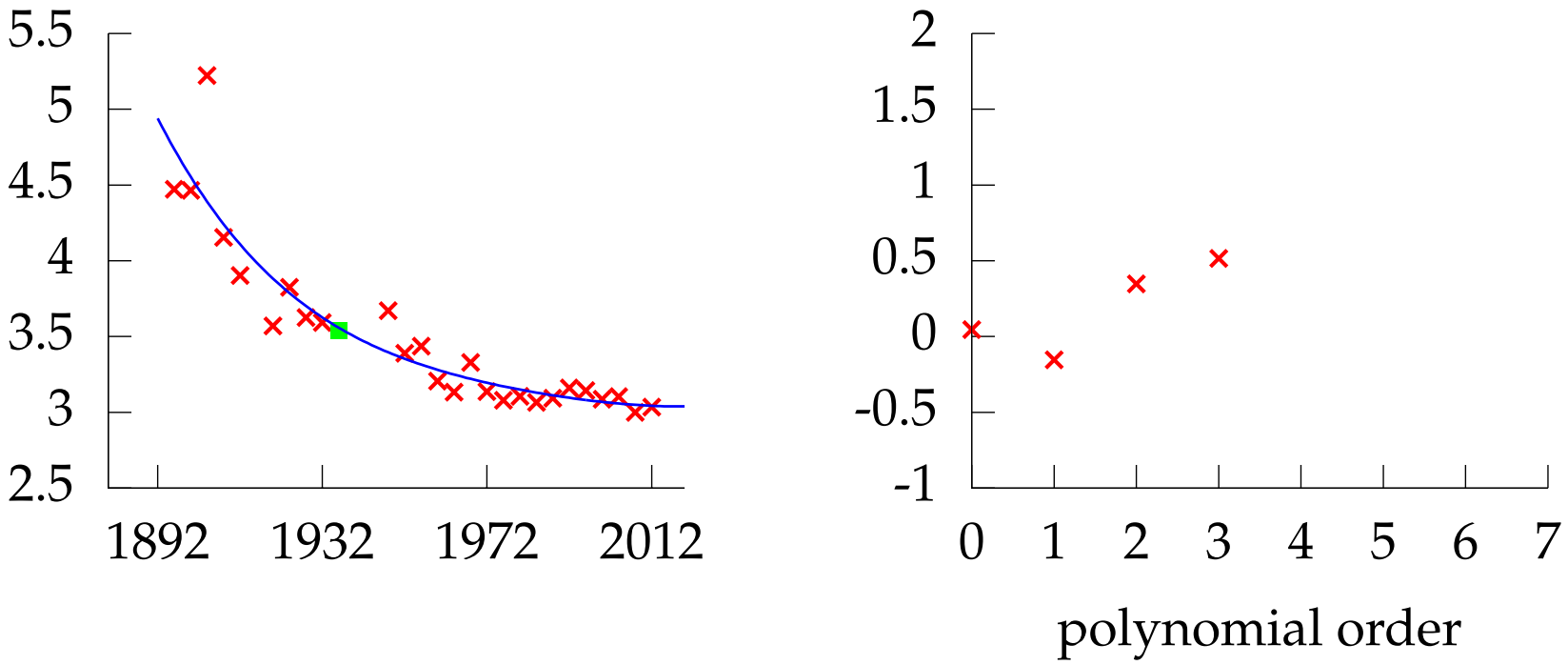
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



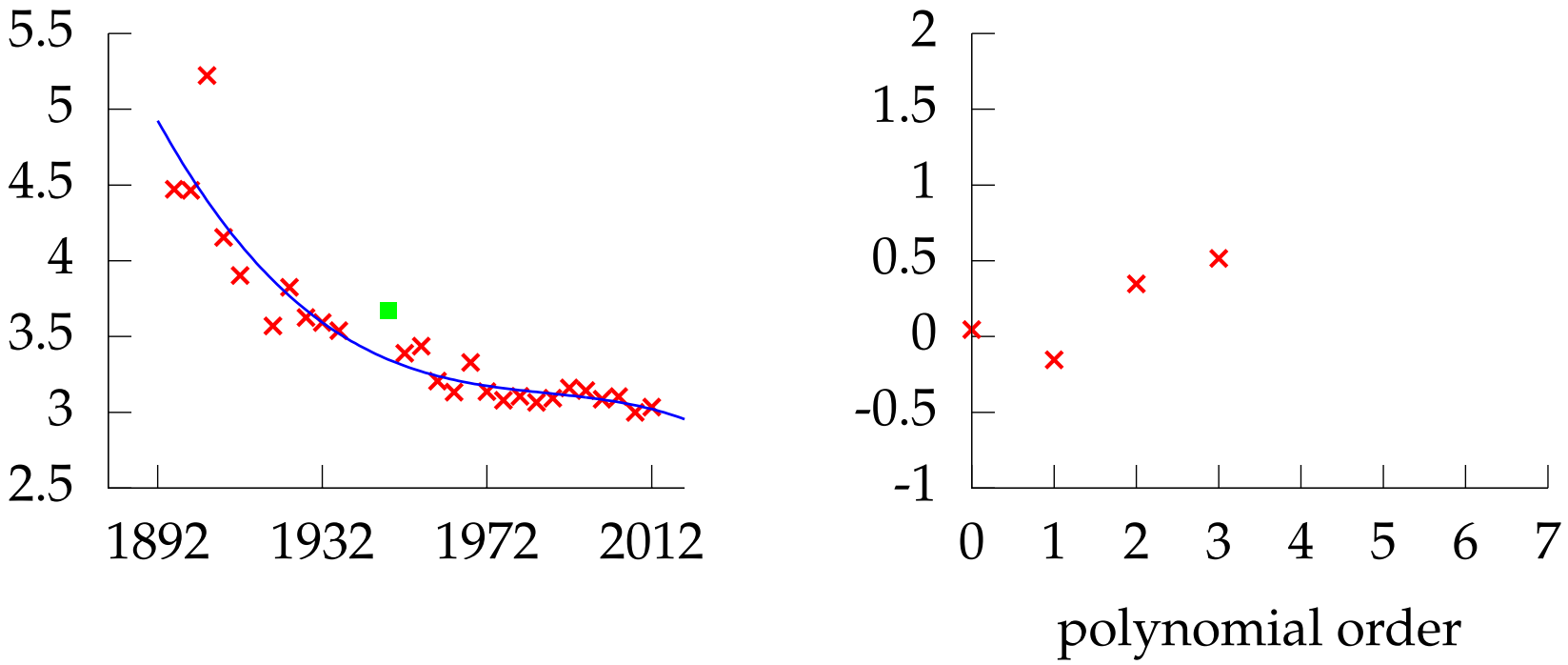
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



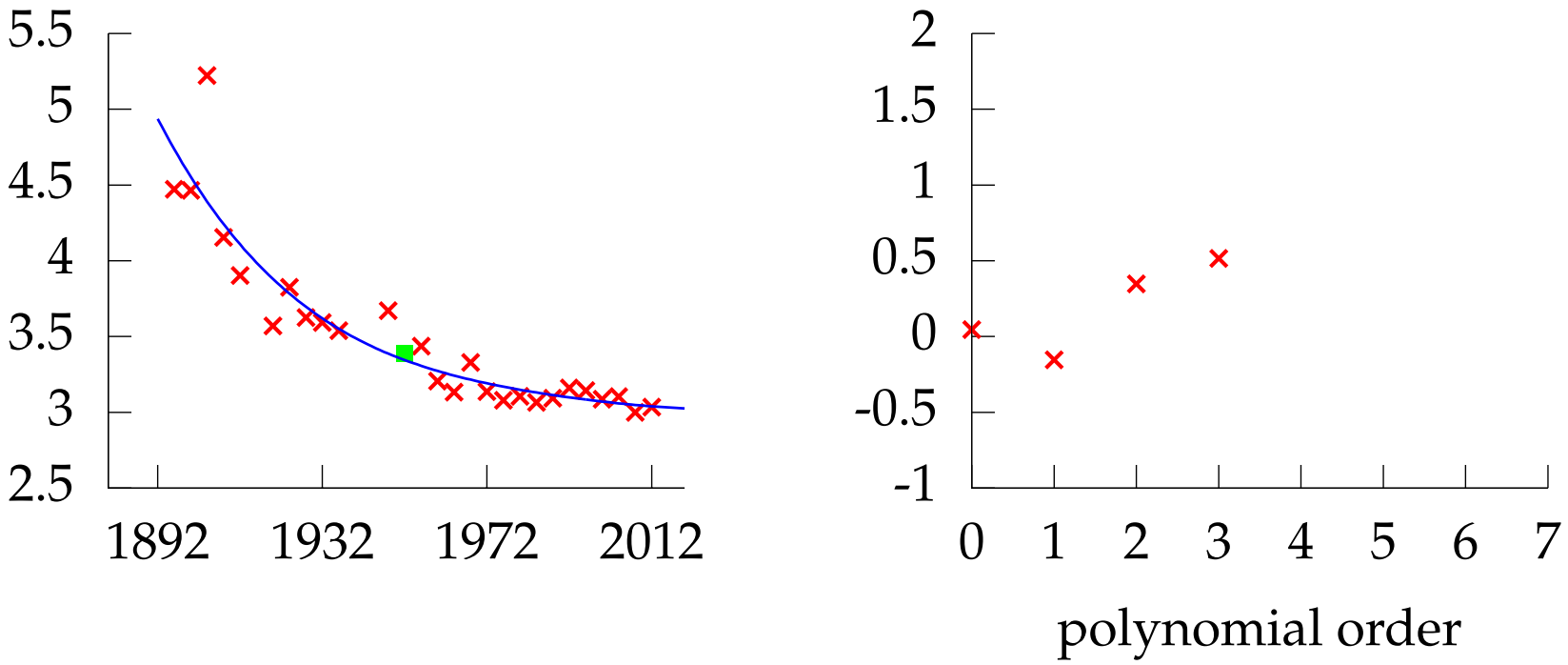
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



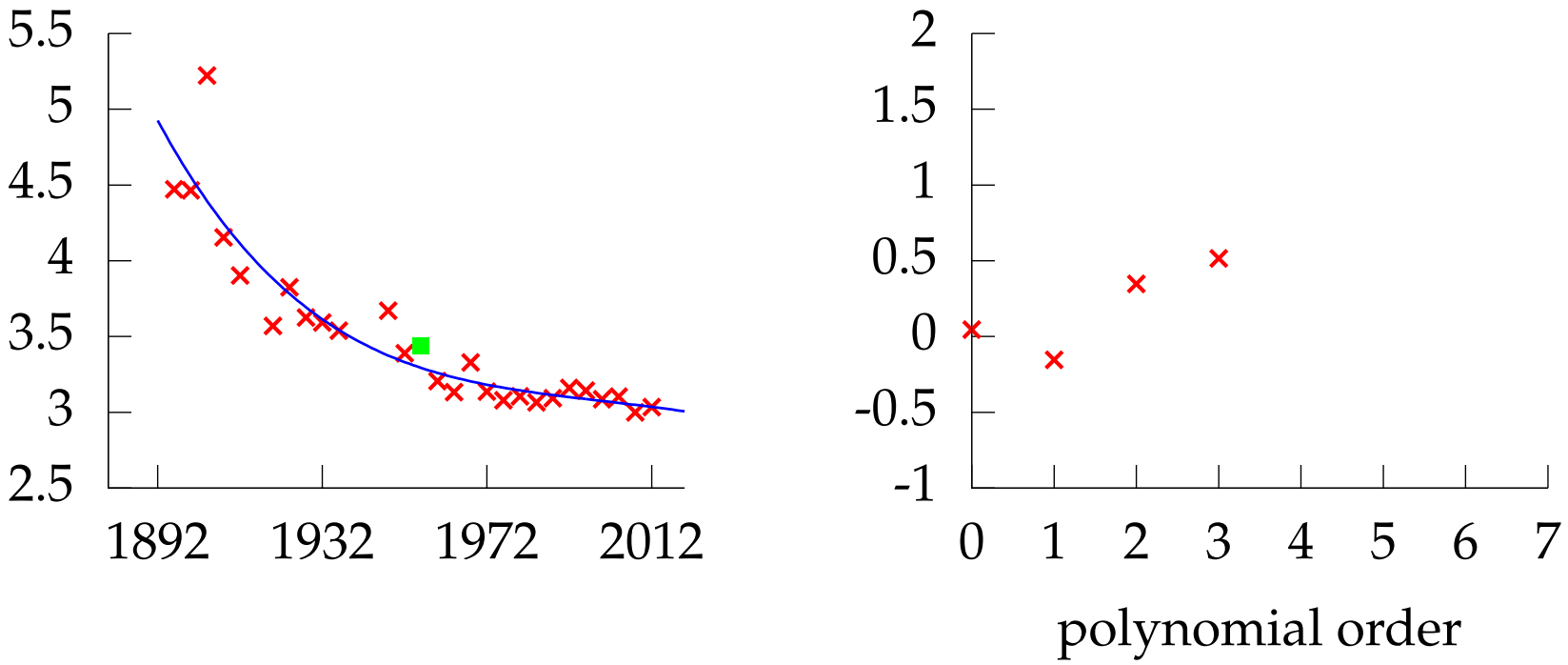
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



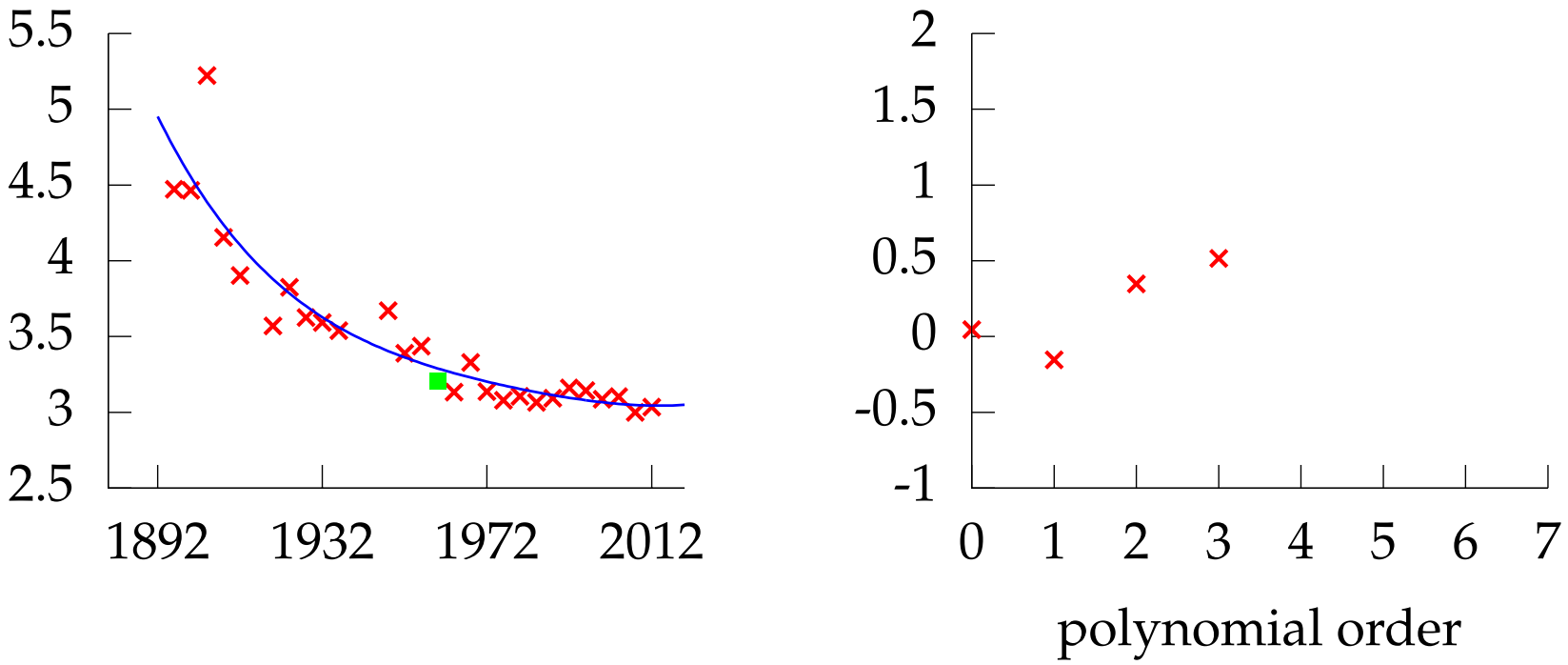
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



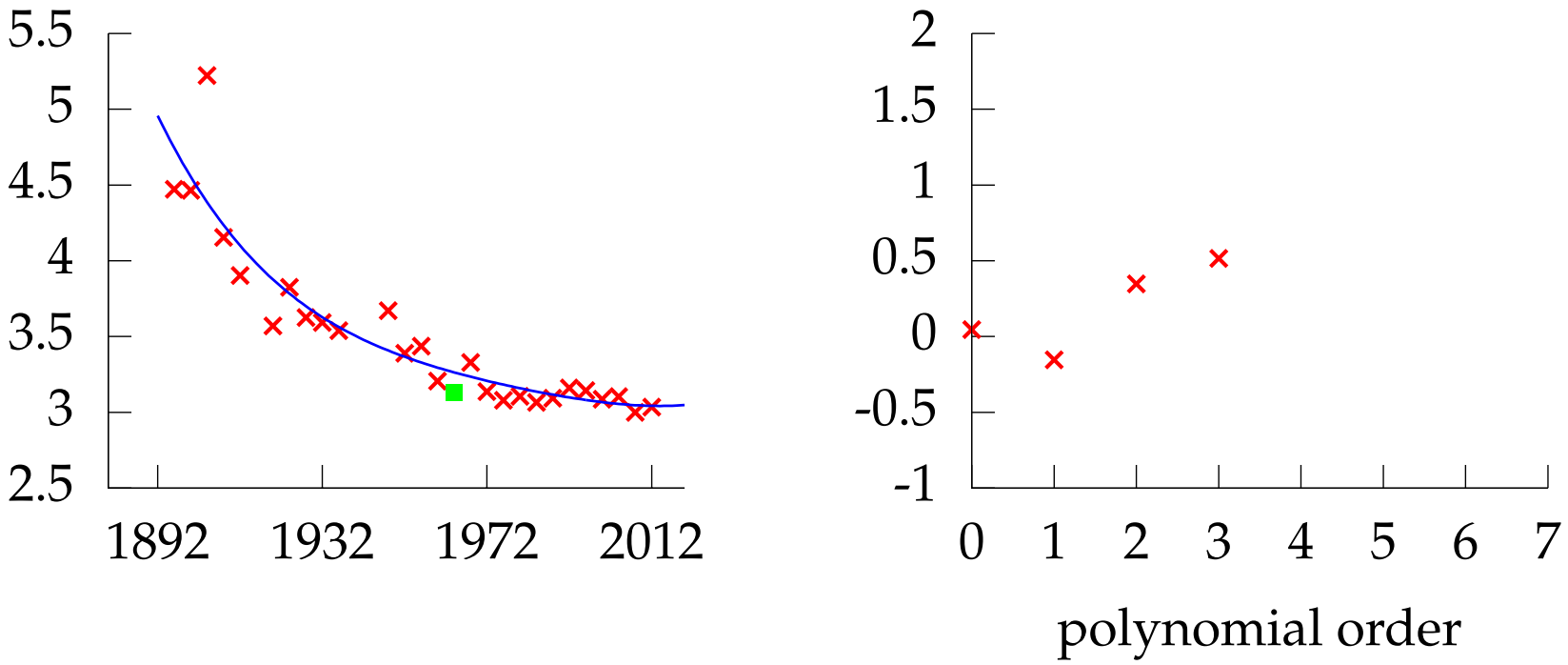
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



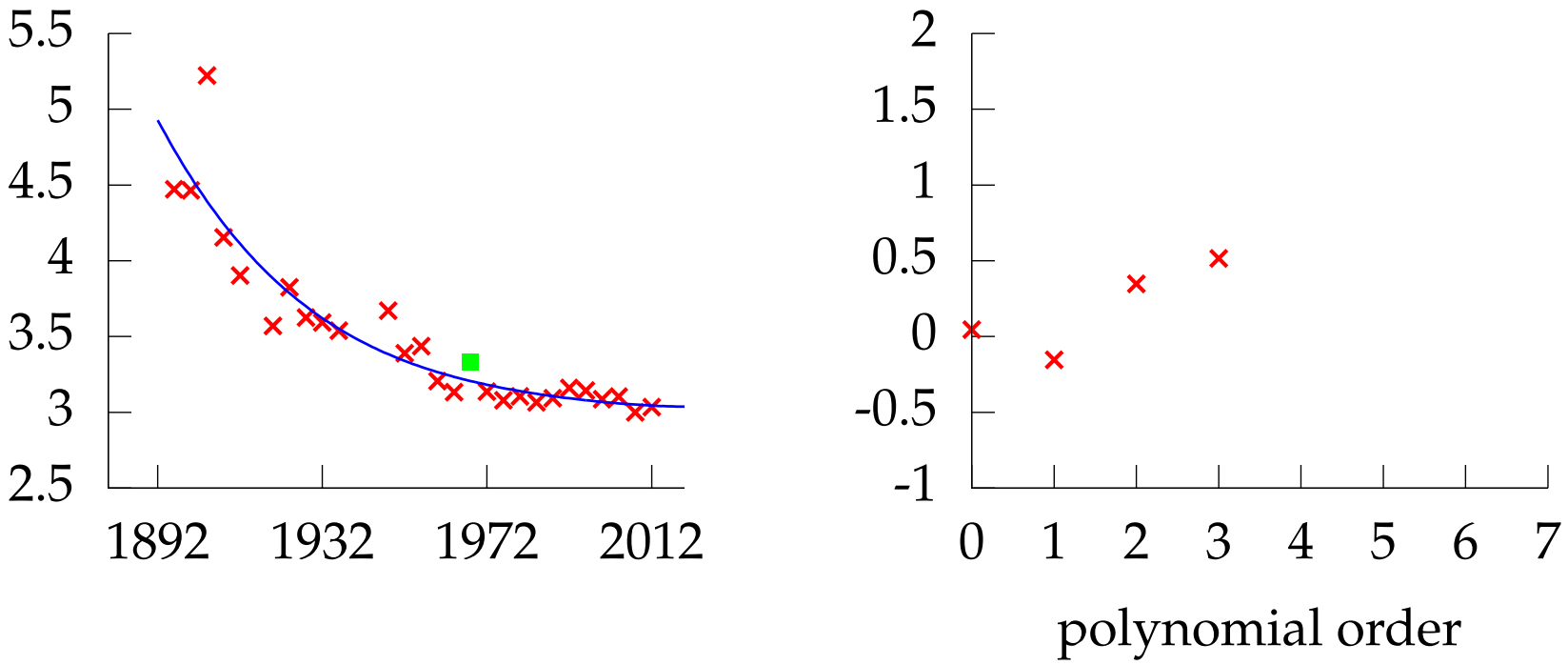
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



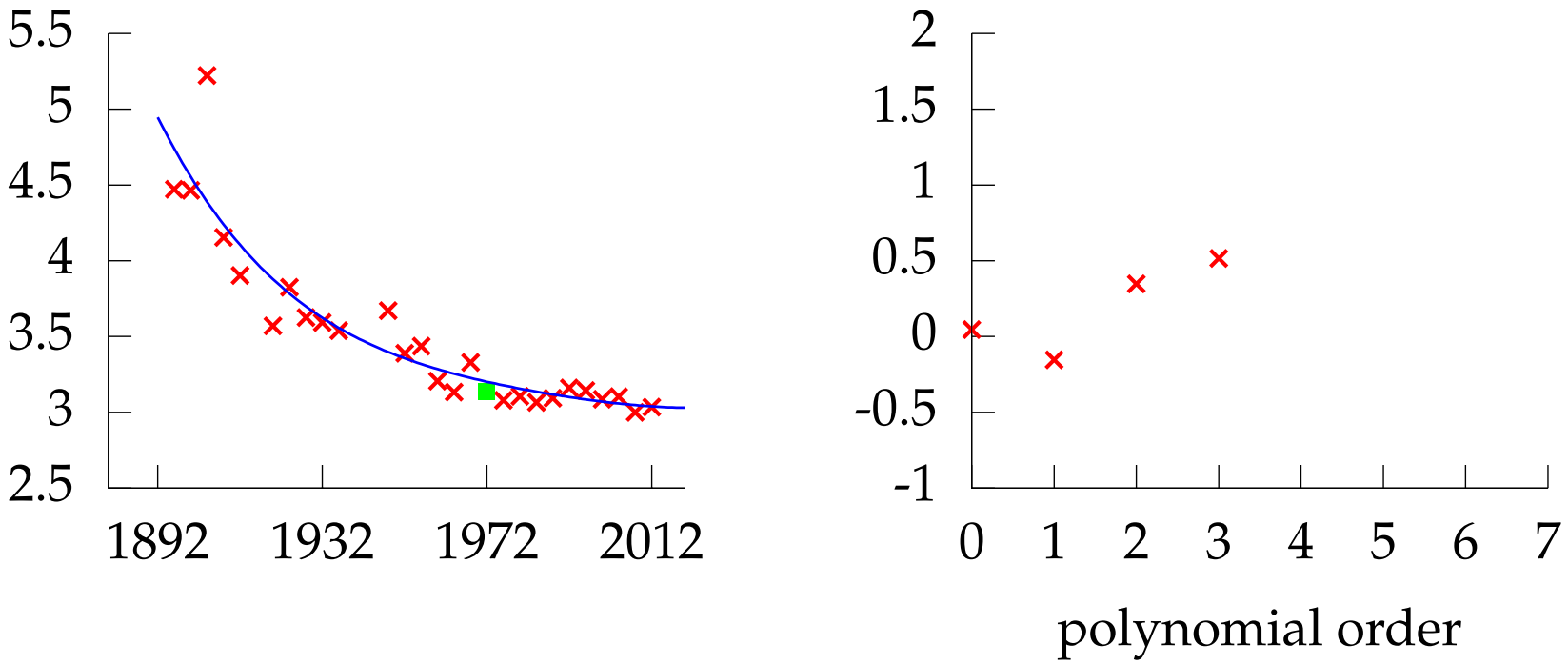
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



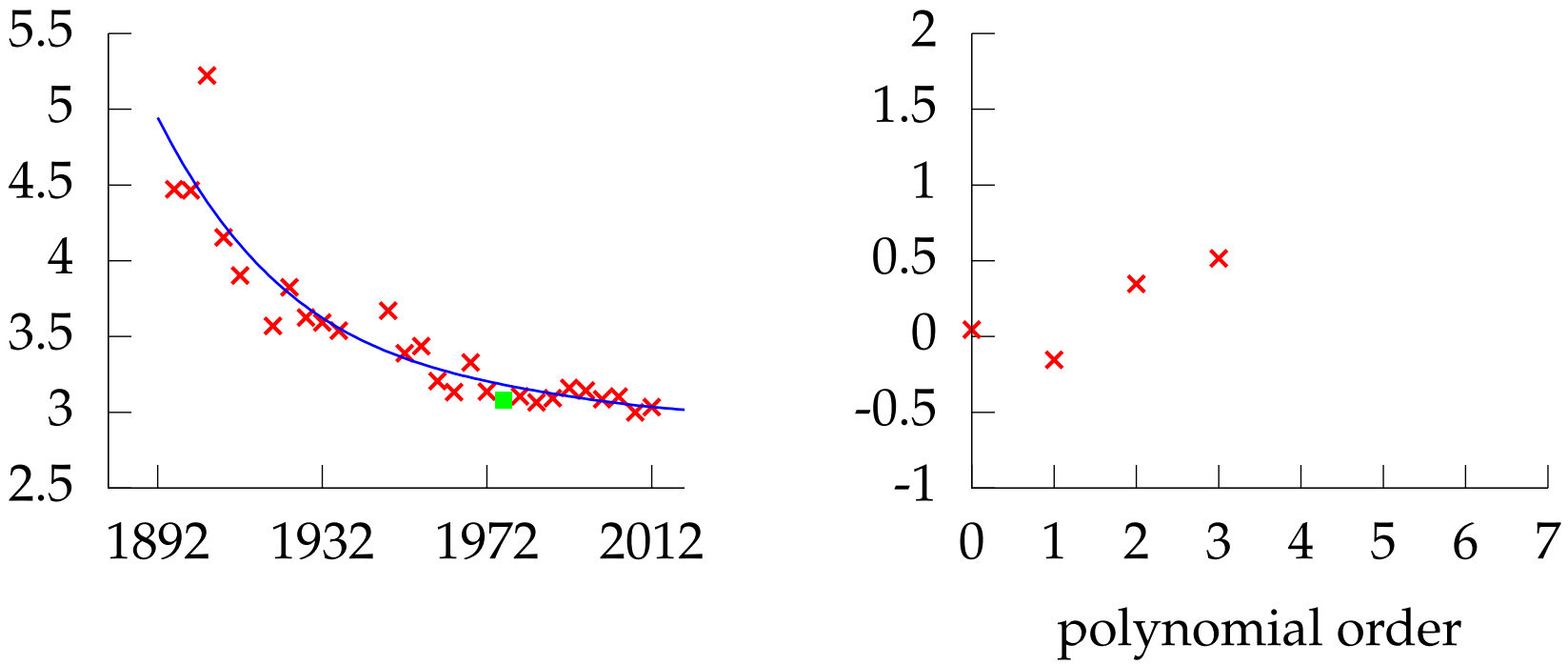
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



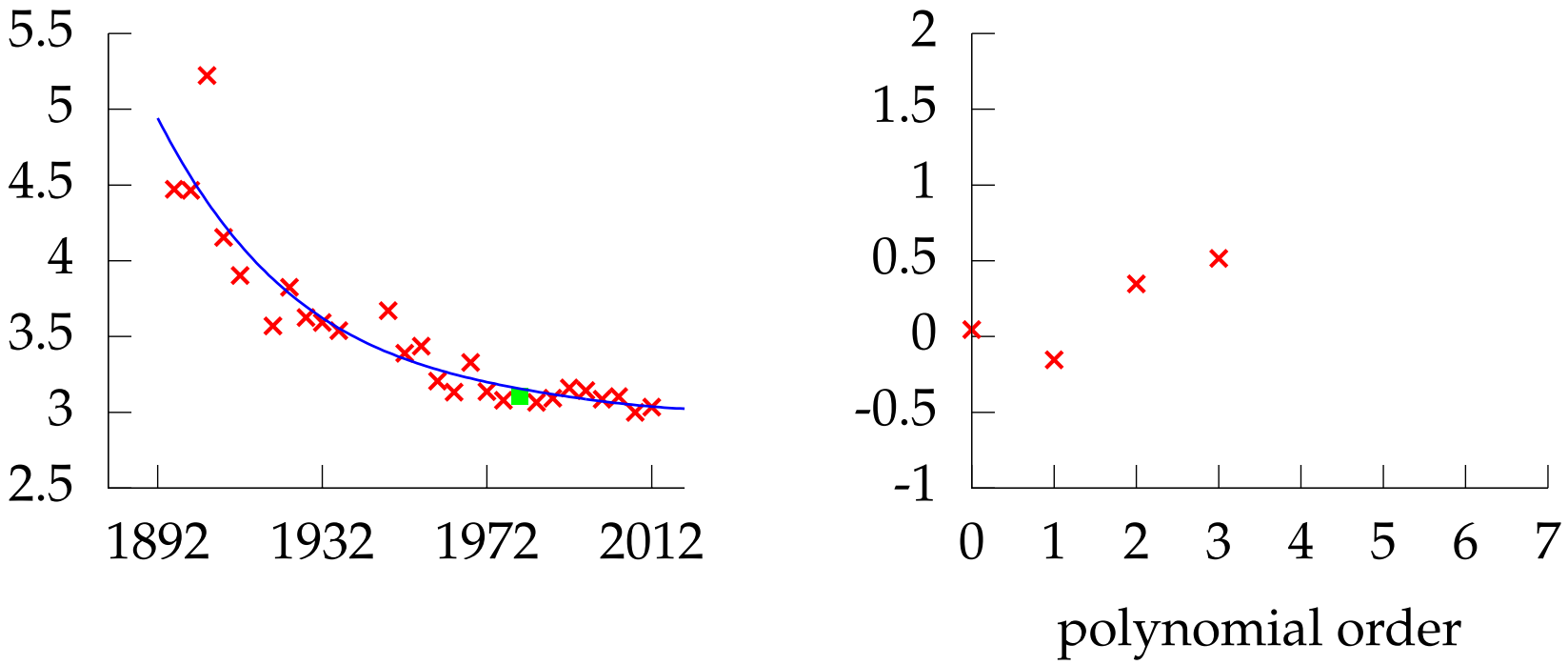
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



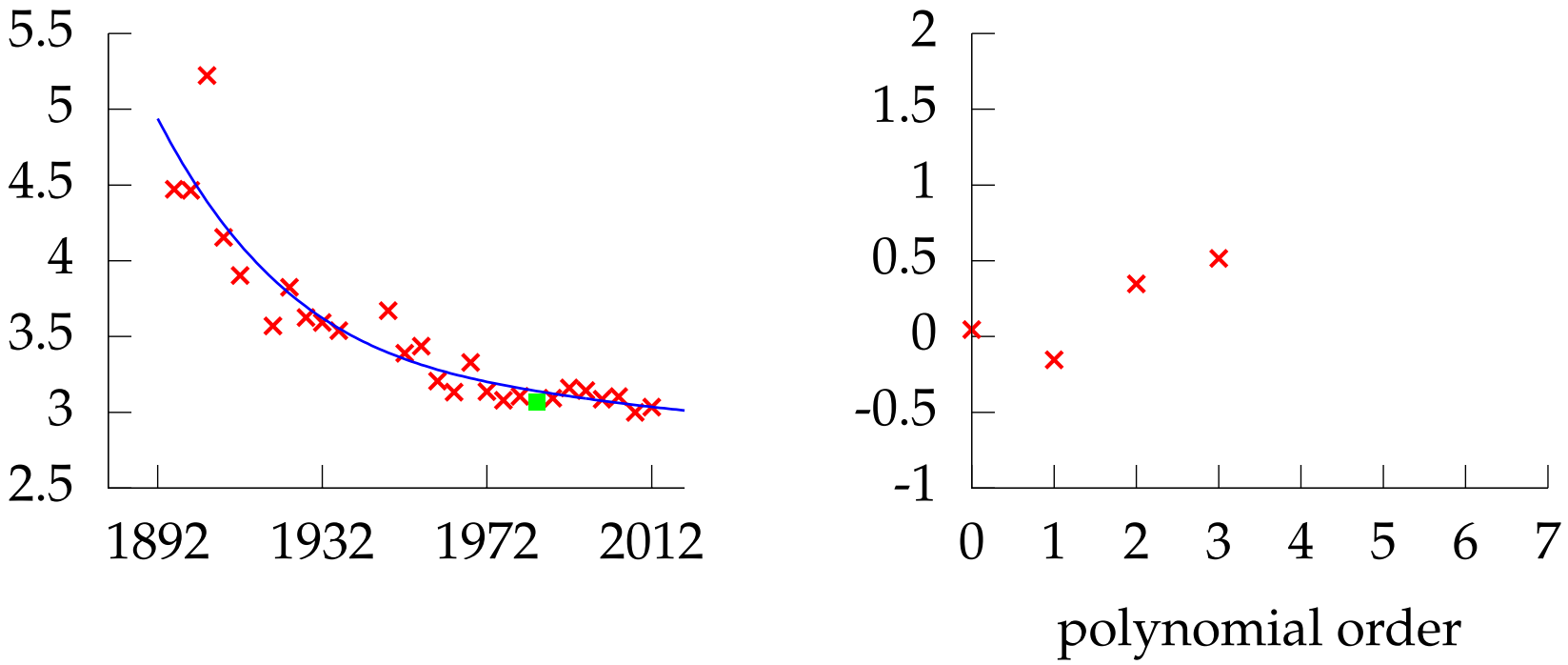
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



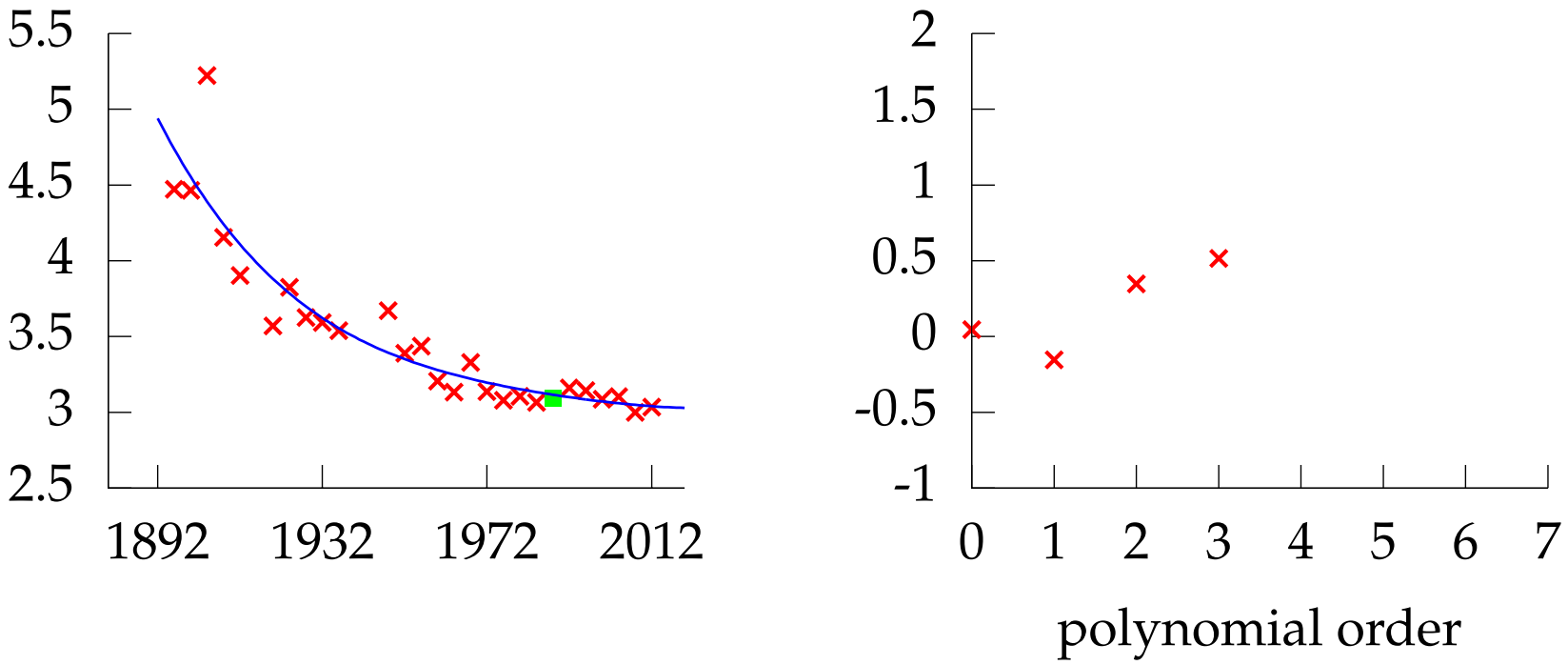
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



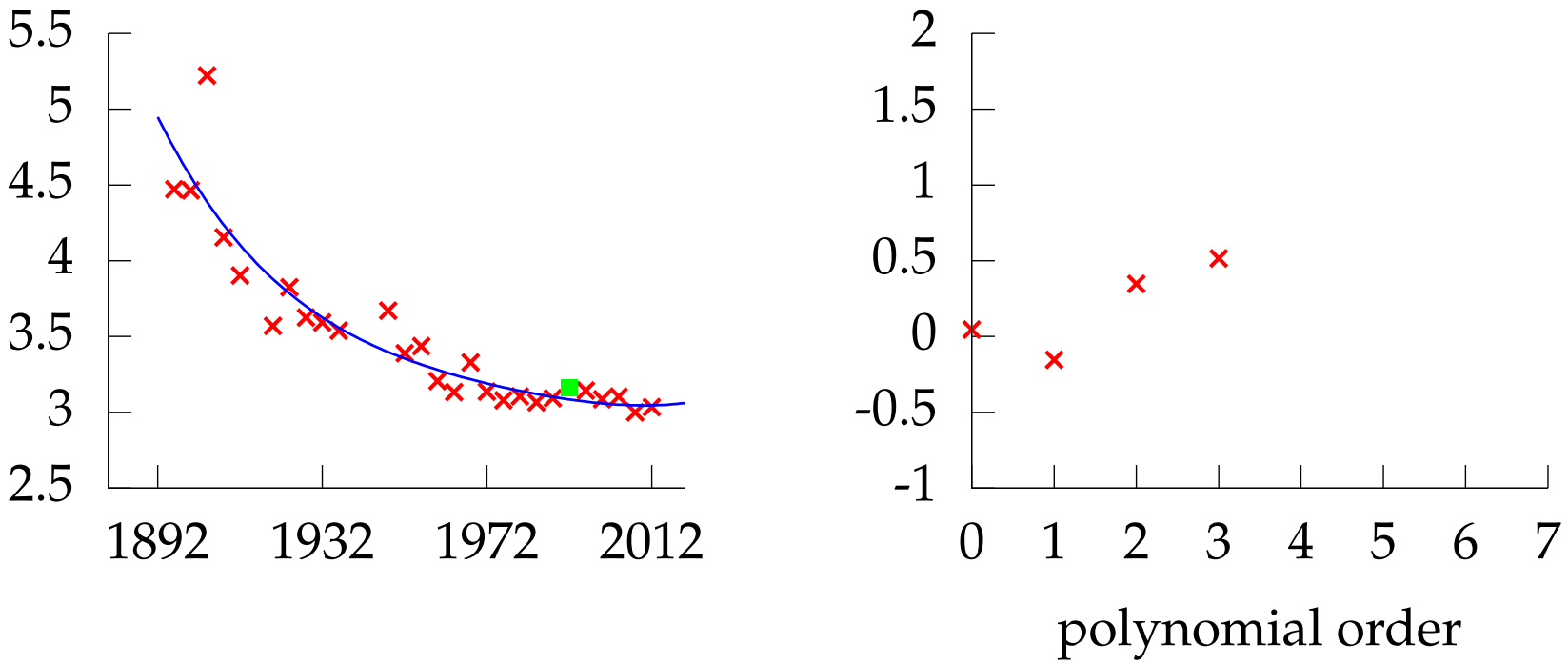
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



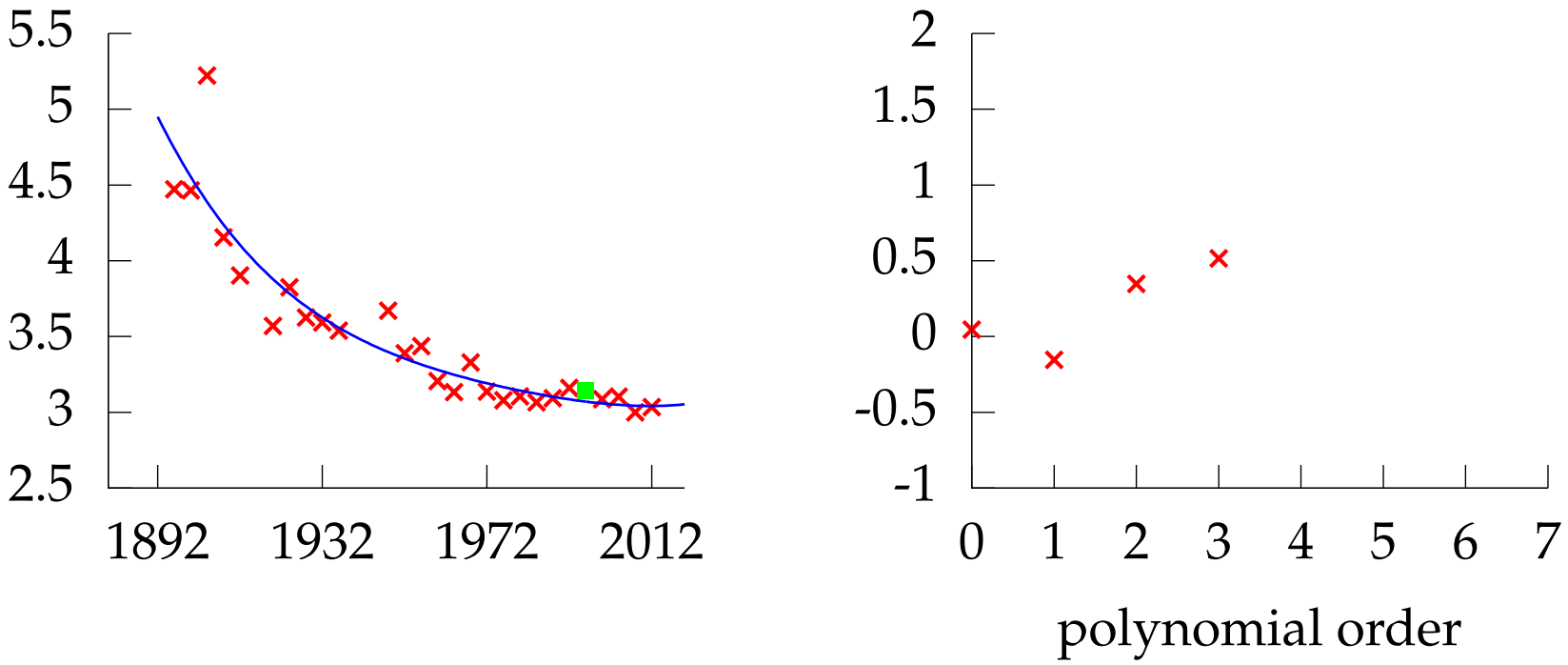
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



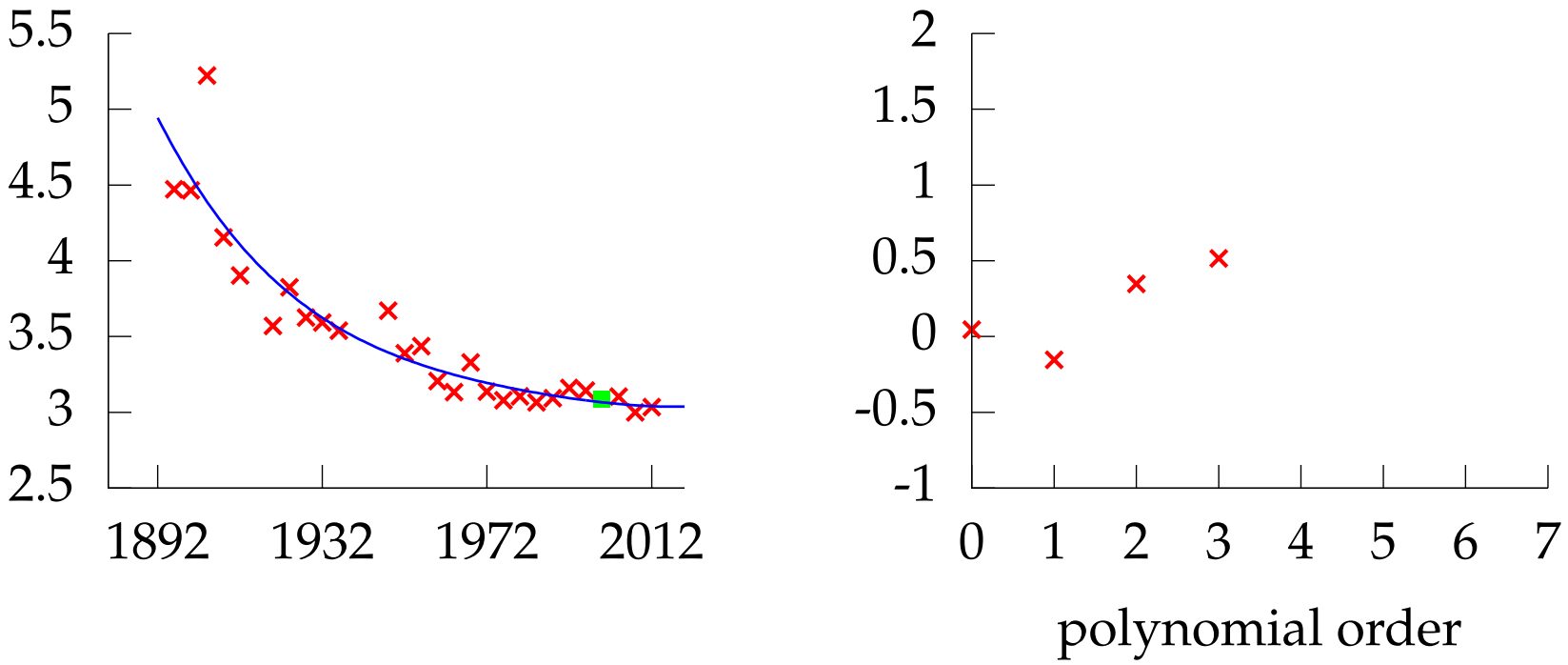
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



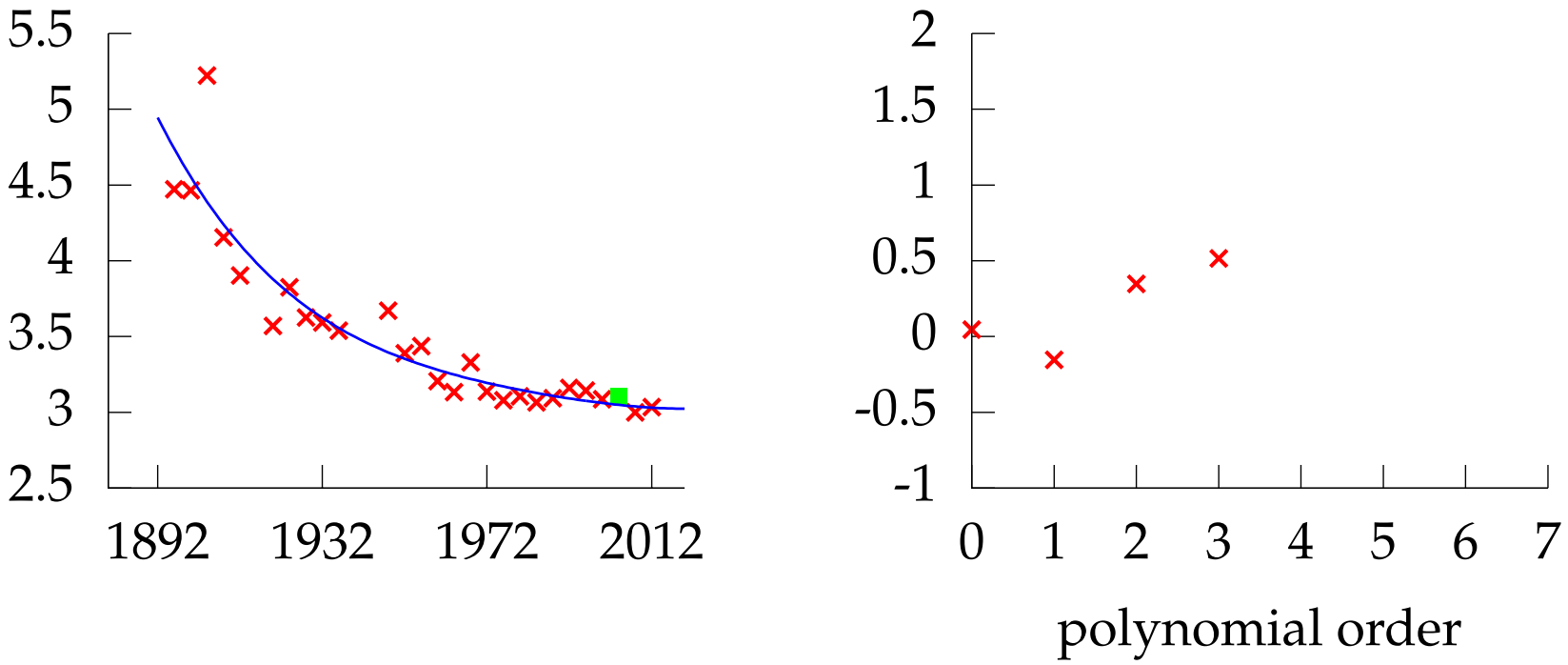
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



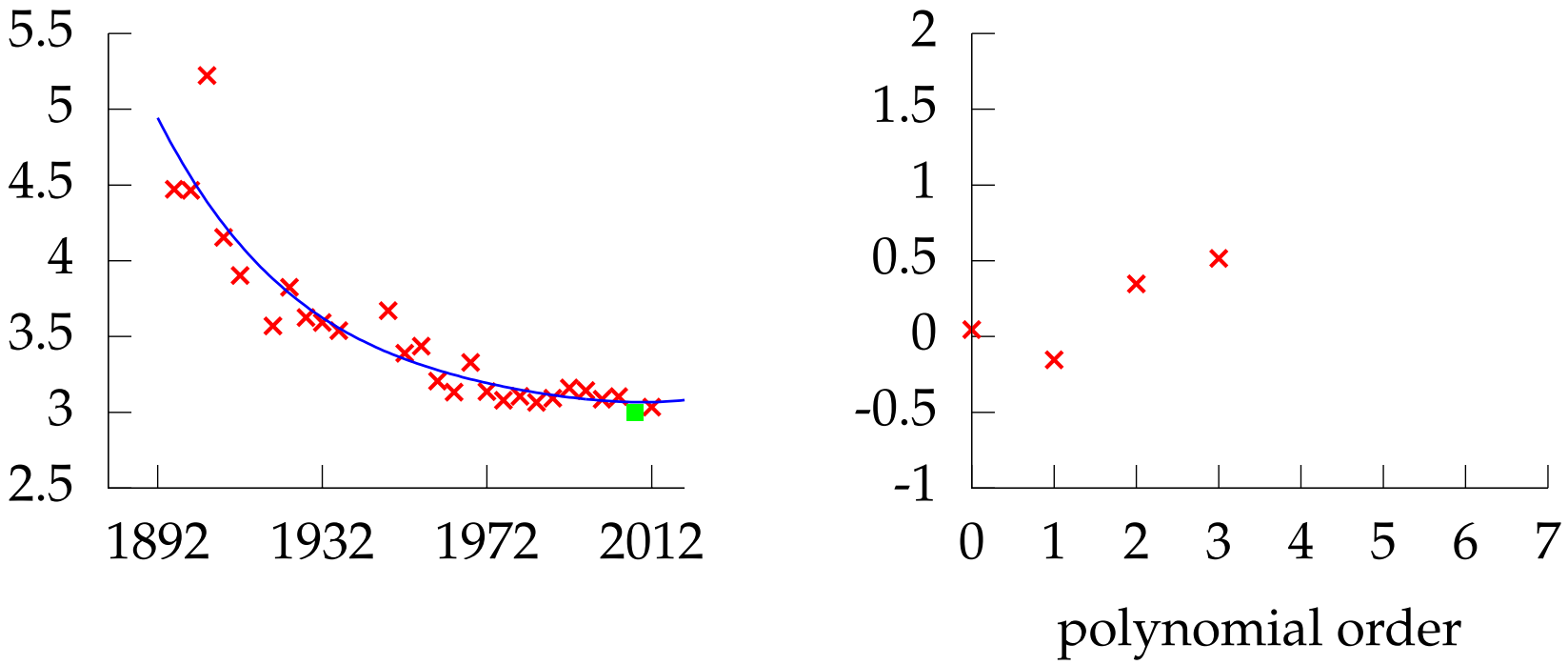
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



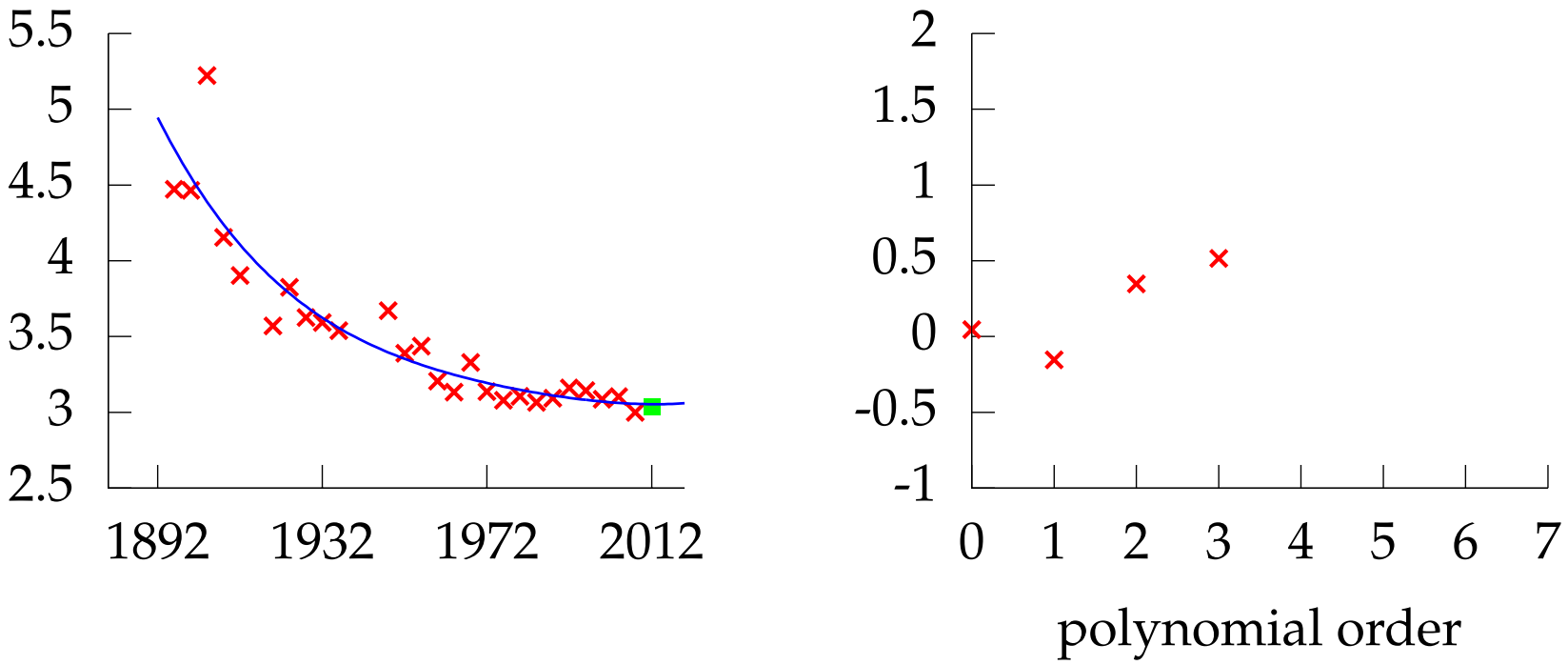
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



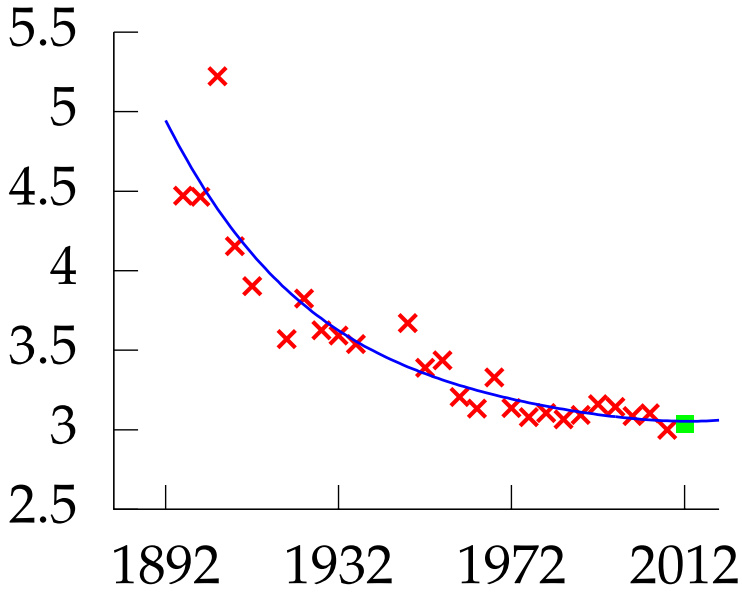
Polynomial order 4, training error -29.324, leave one out error 0.84844.

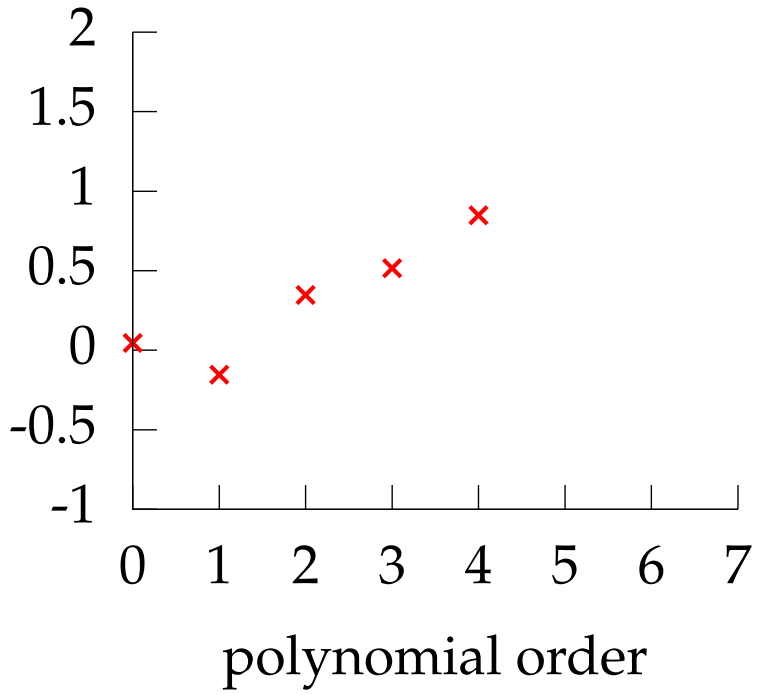
Leave One Out Error



Polynomial order 4, training error -29.324, leave one out error 0.84844.

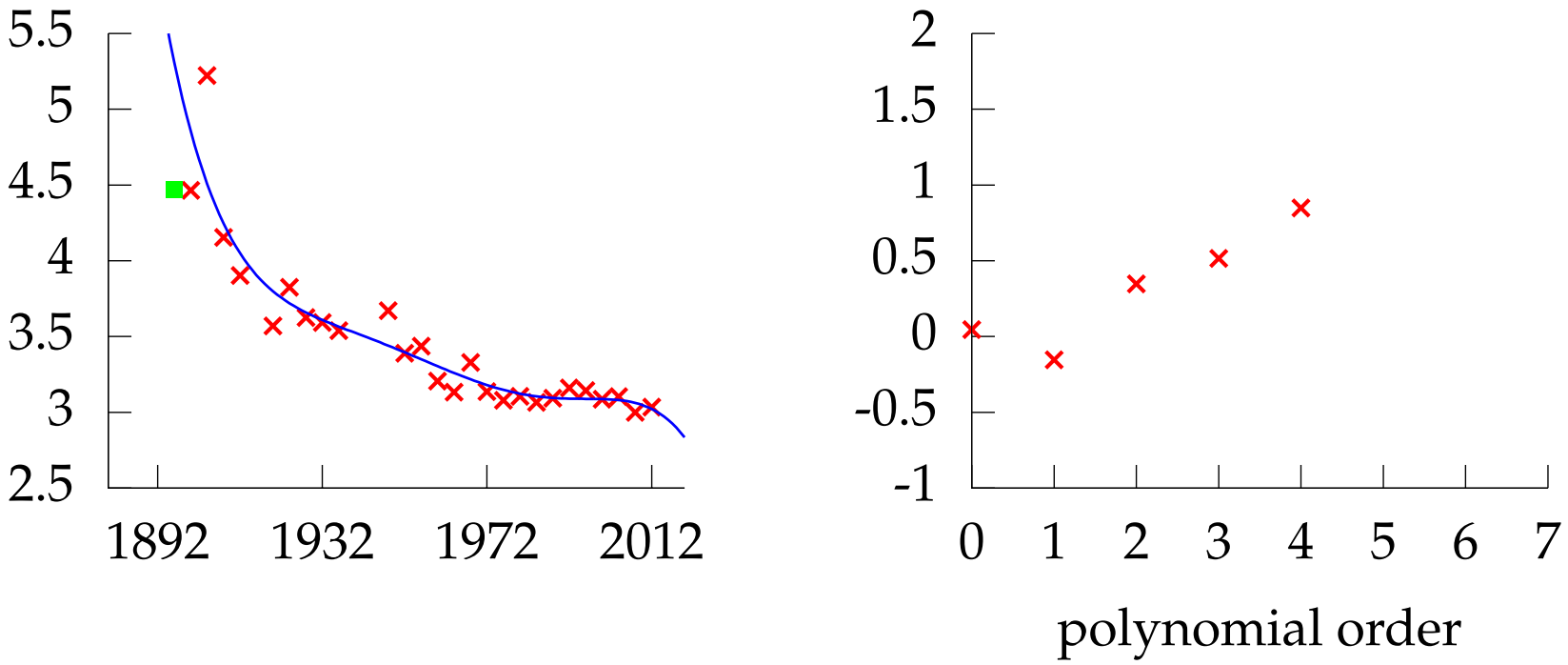
Leave One Out Error





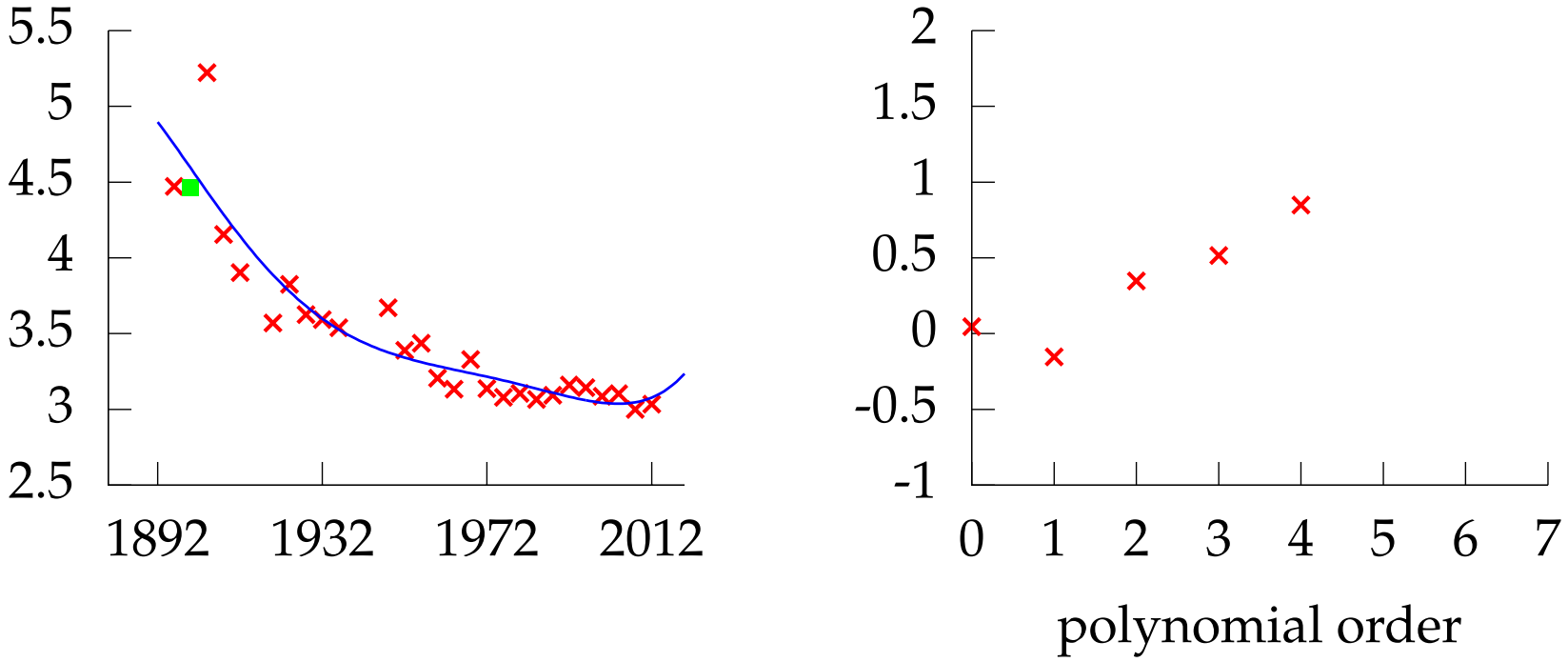
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



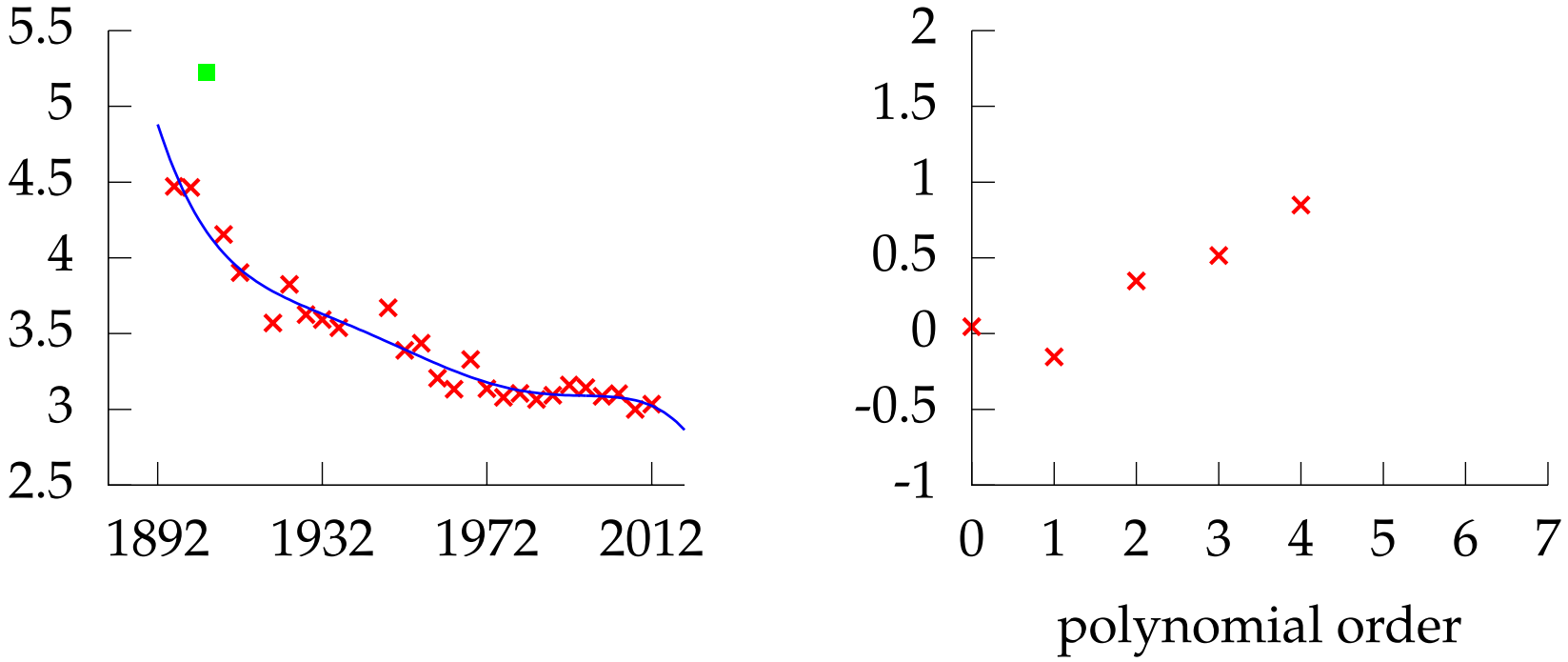
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



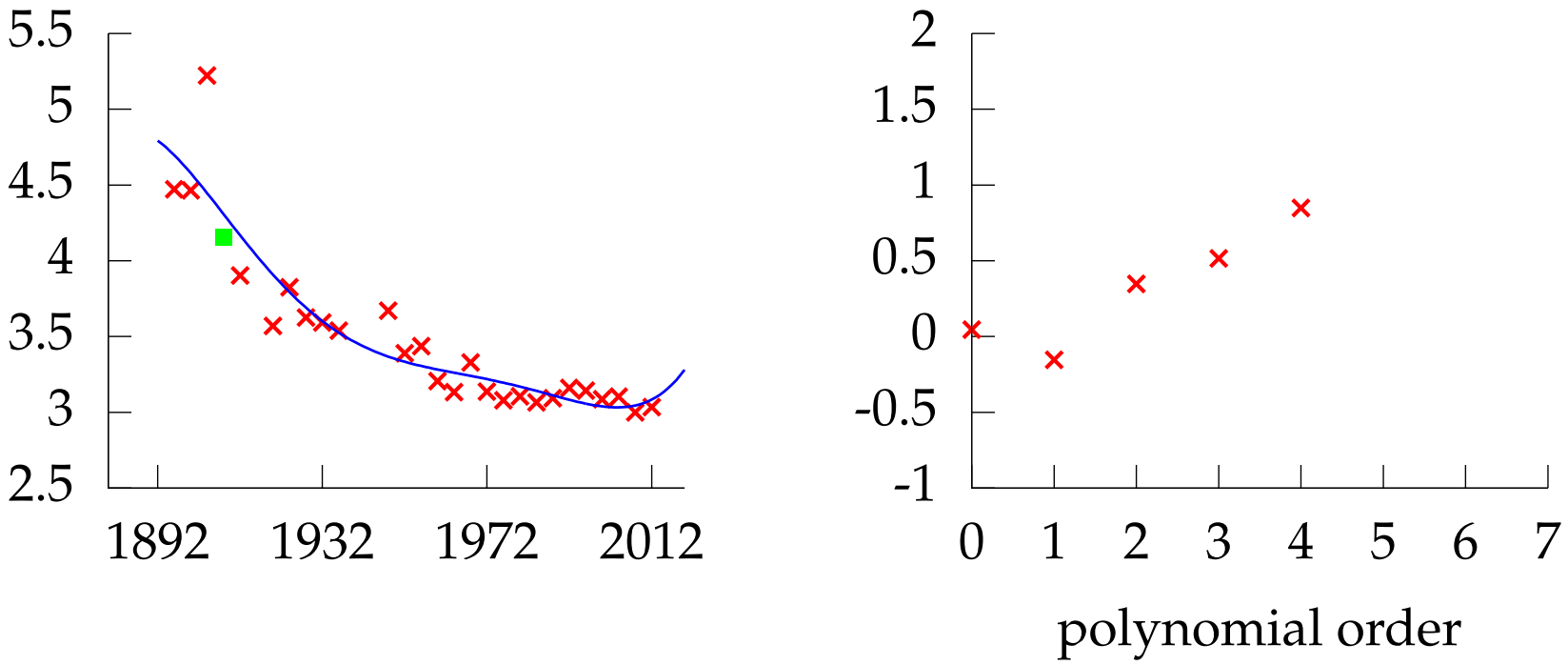
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



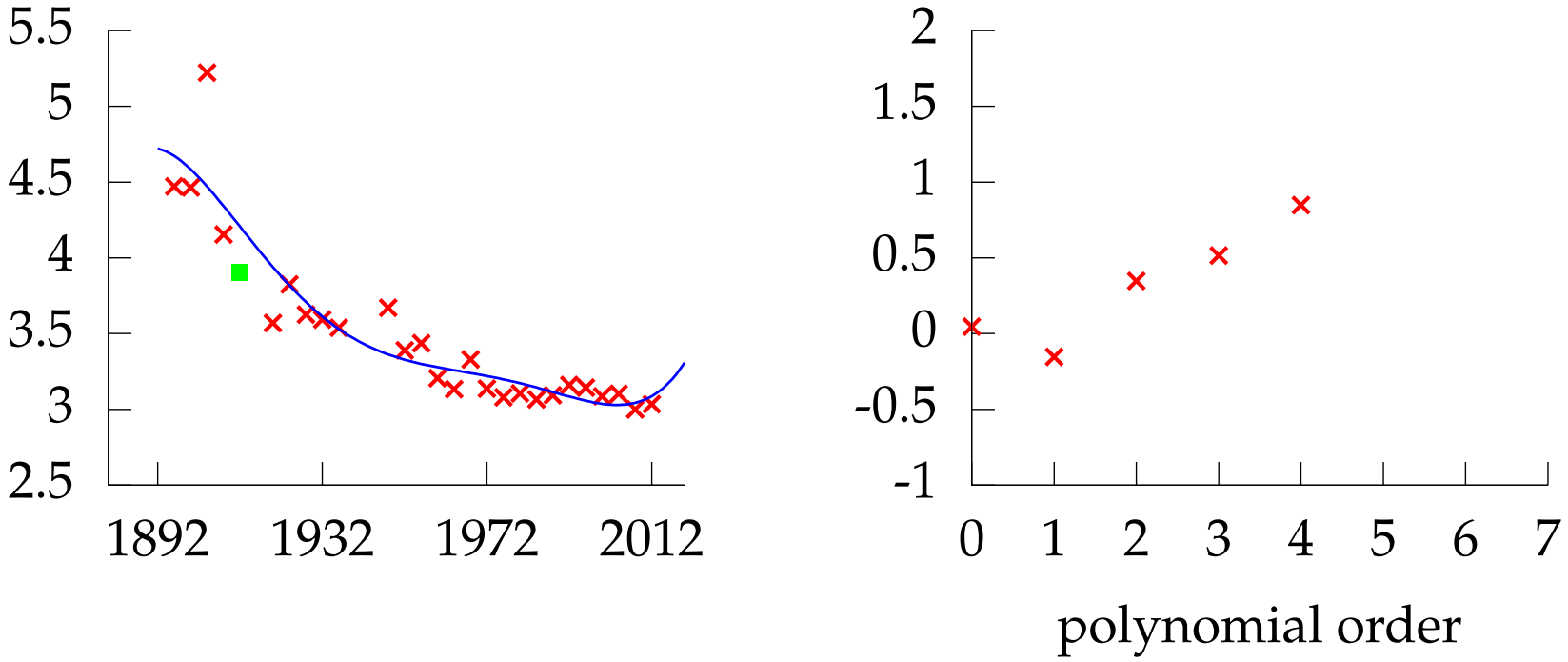
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



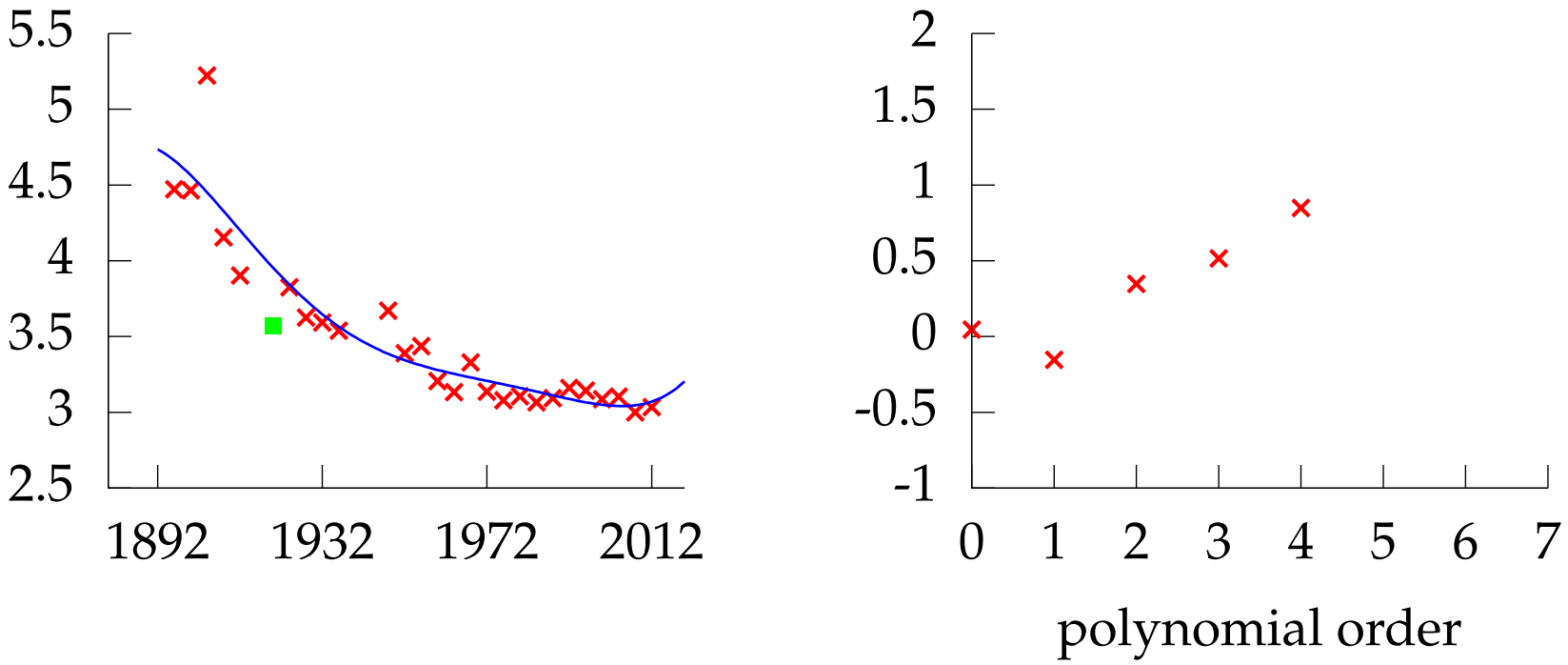
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



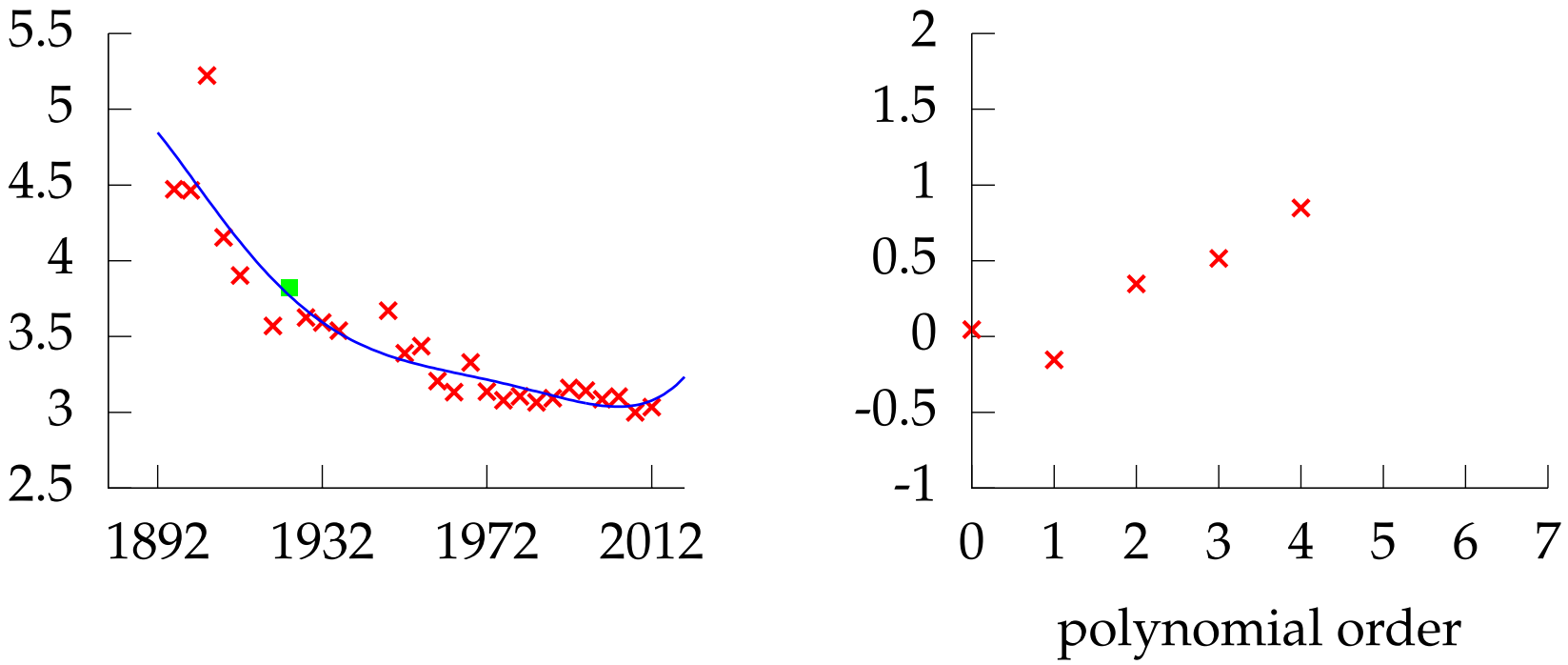
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



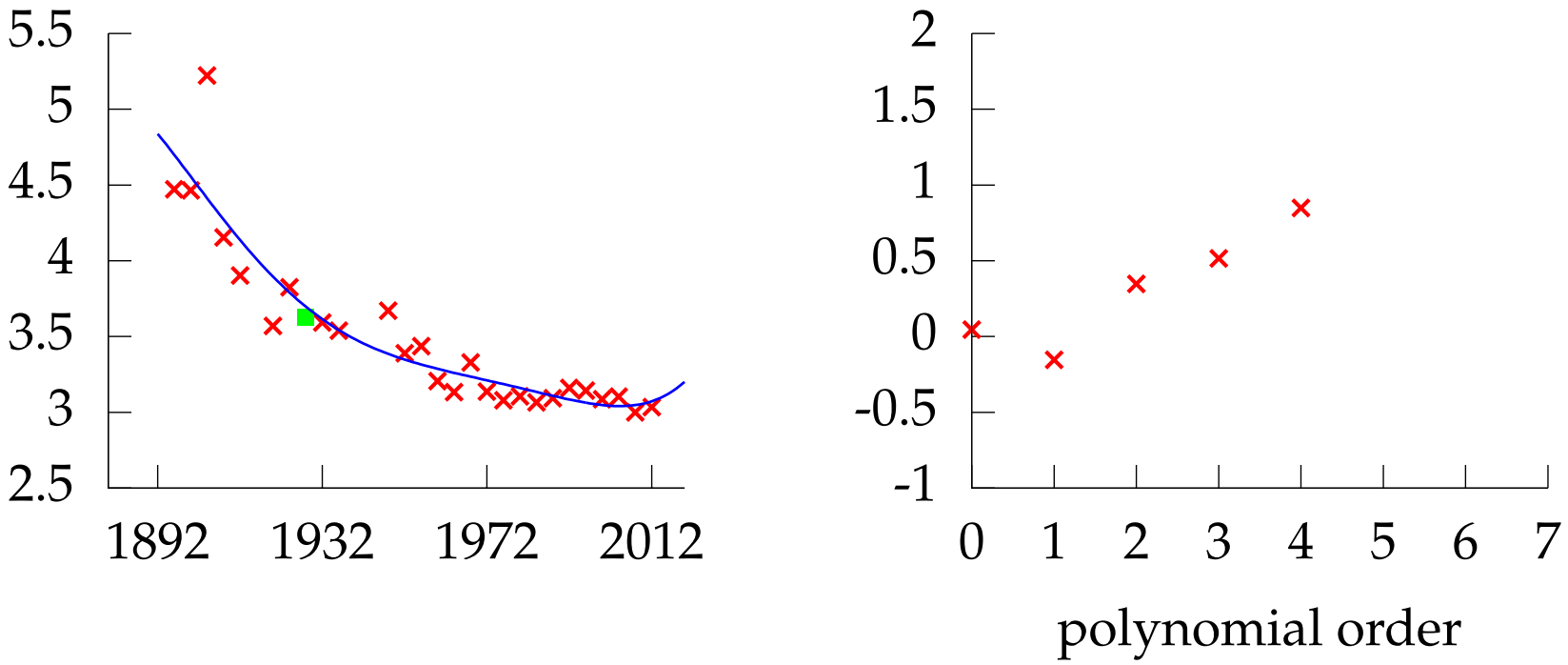
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



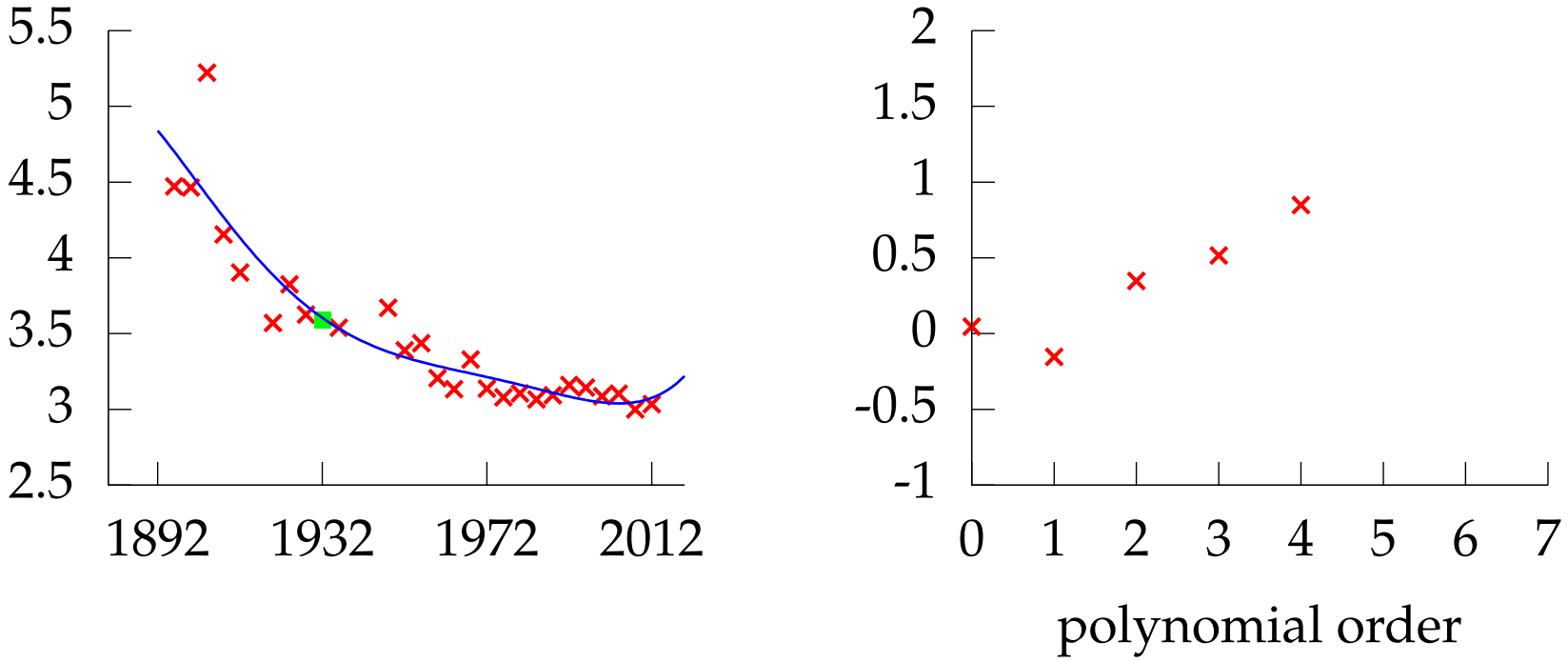
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



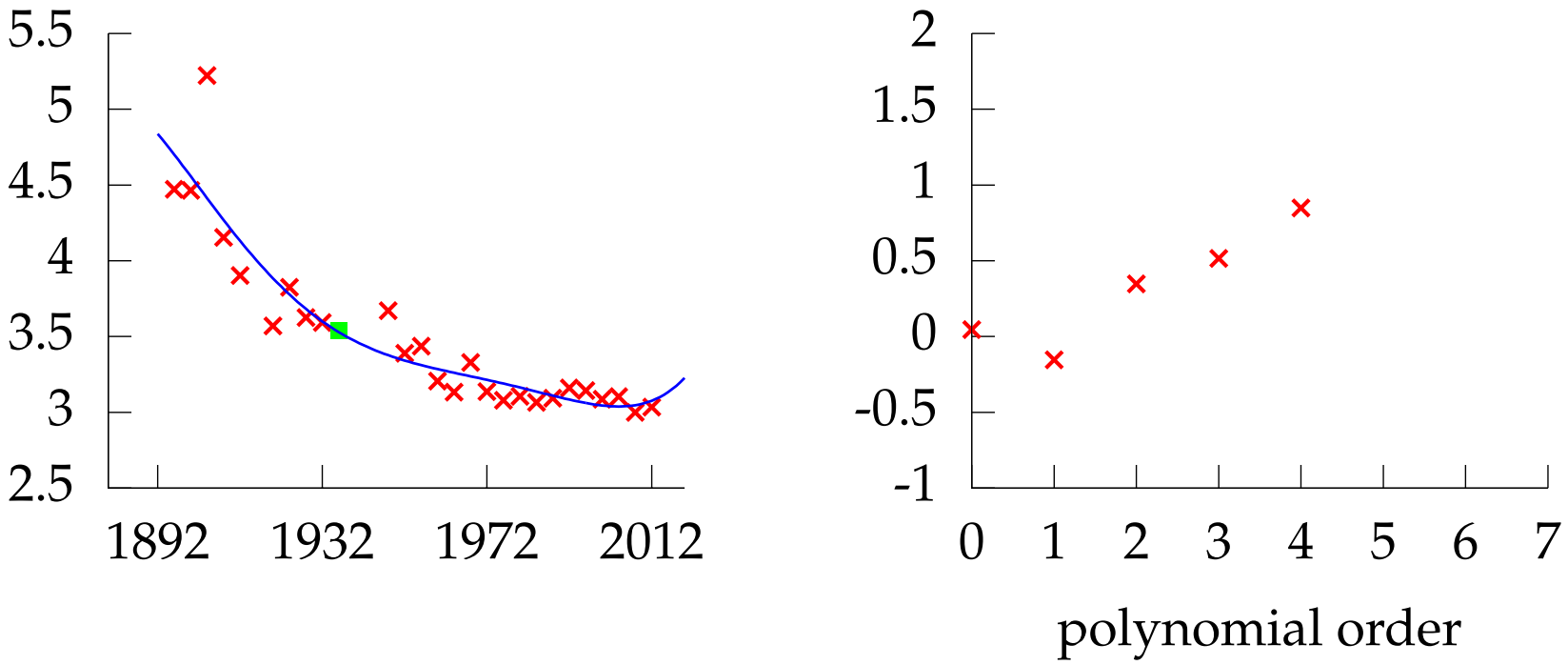
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



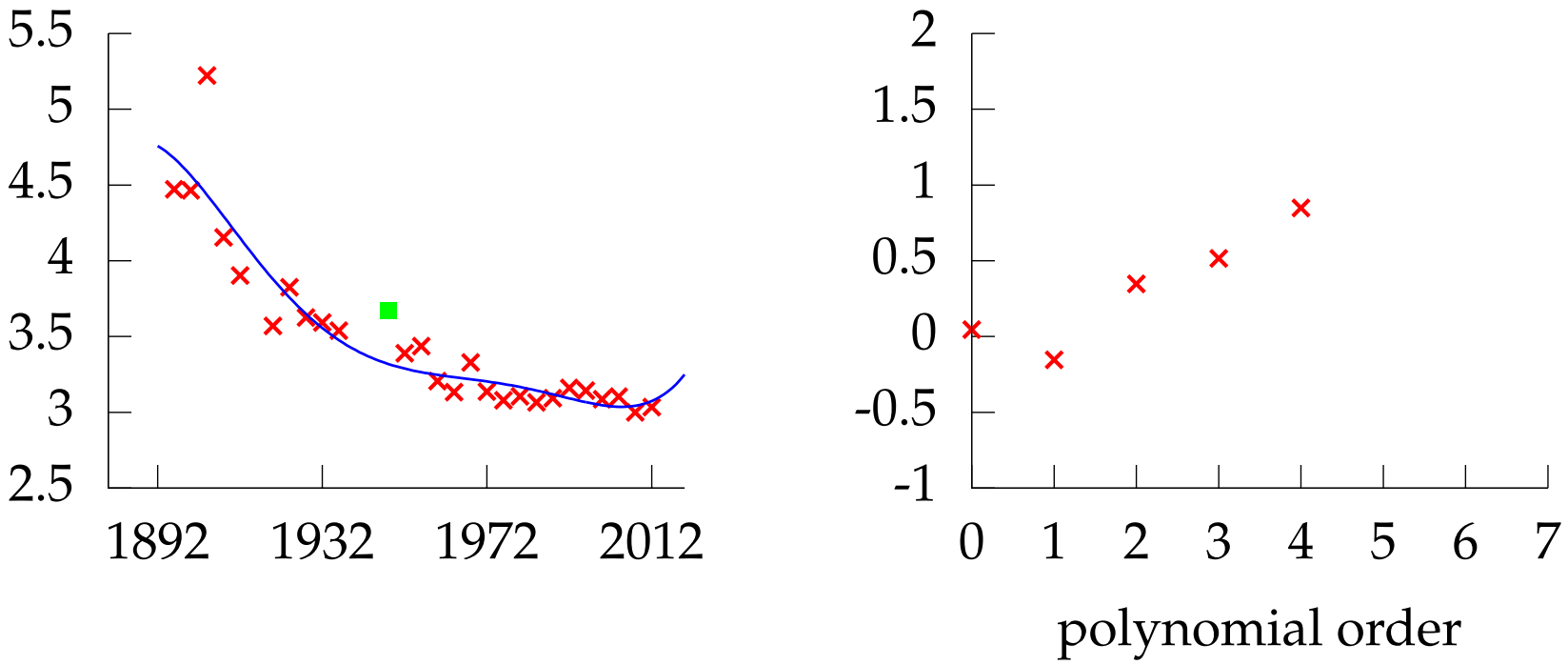
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



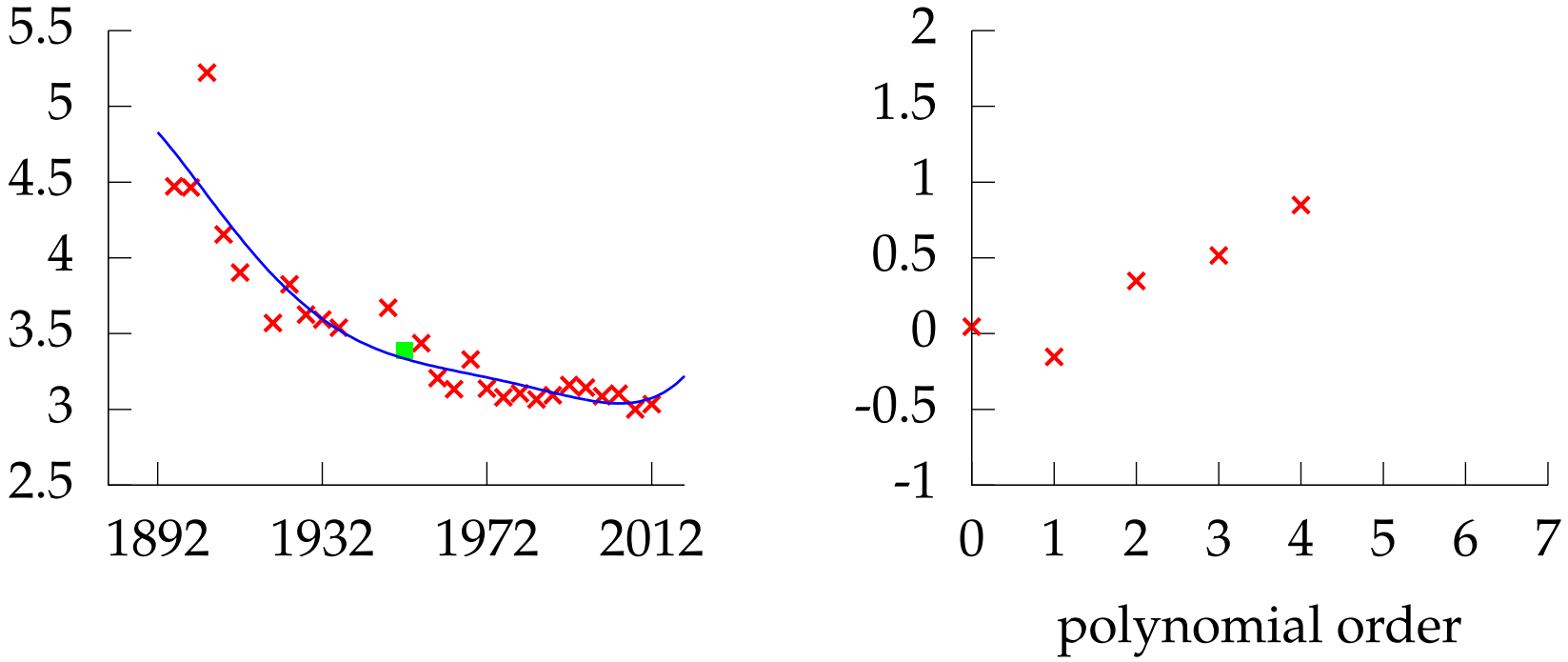
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



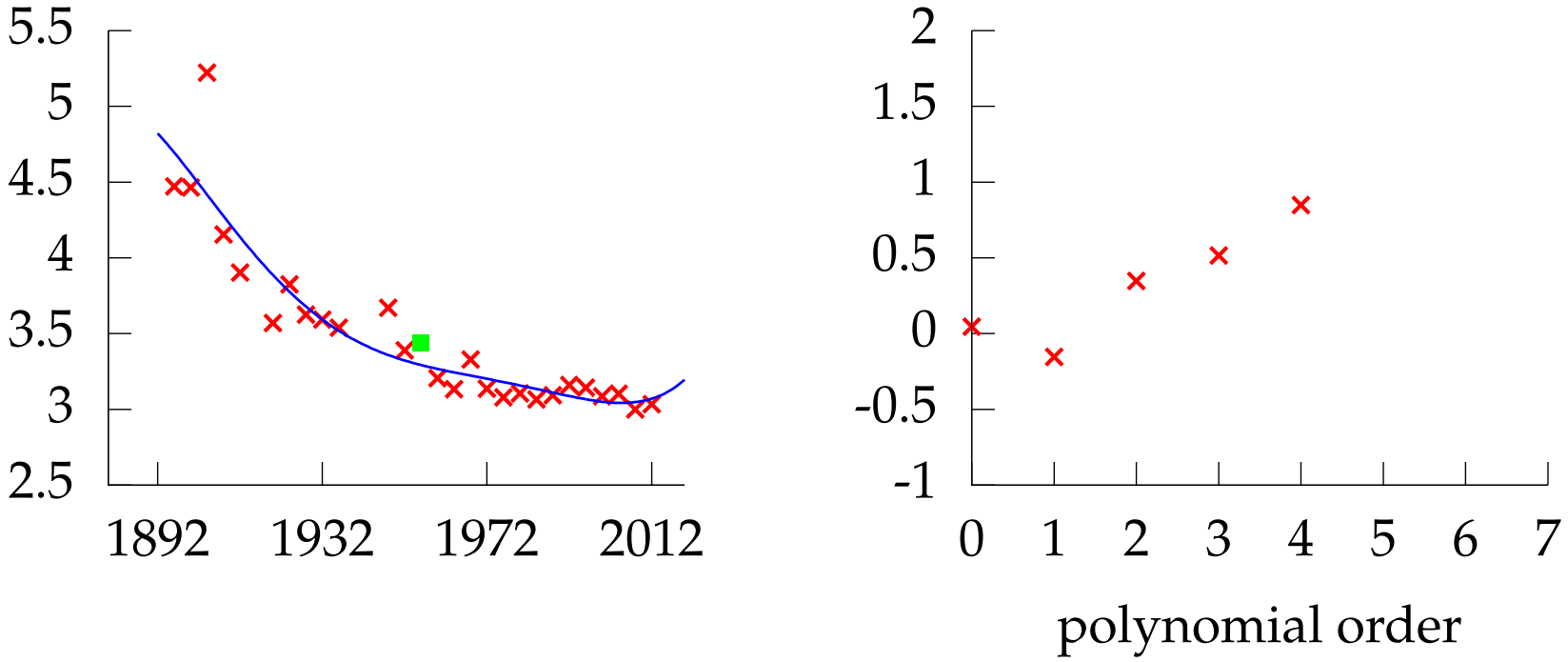
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



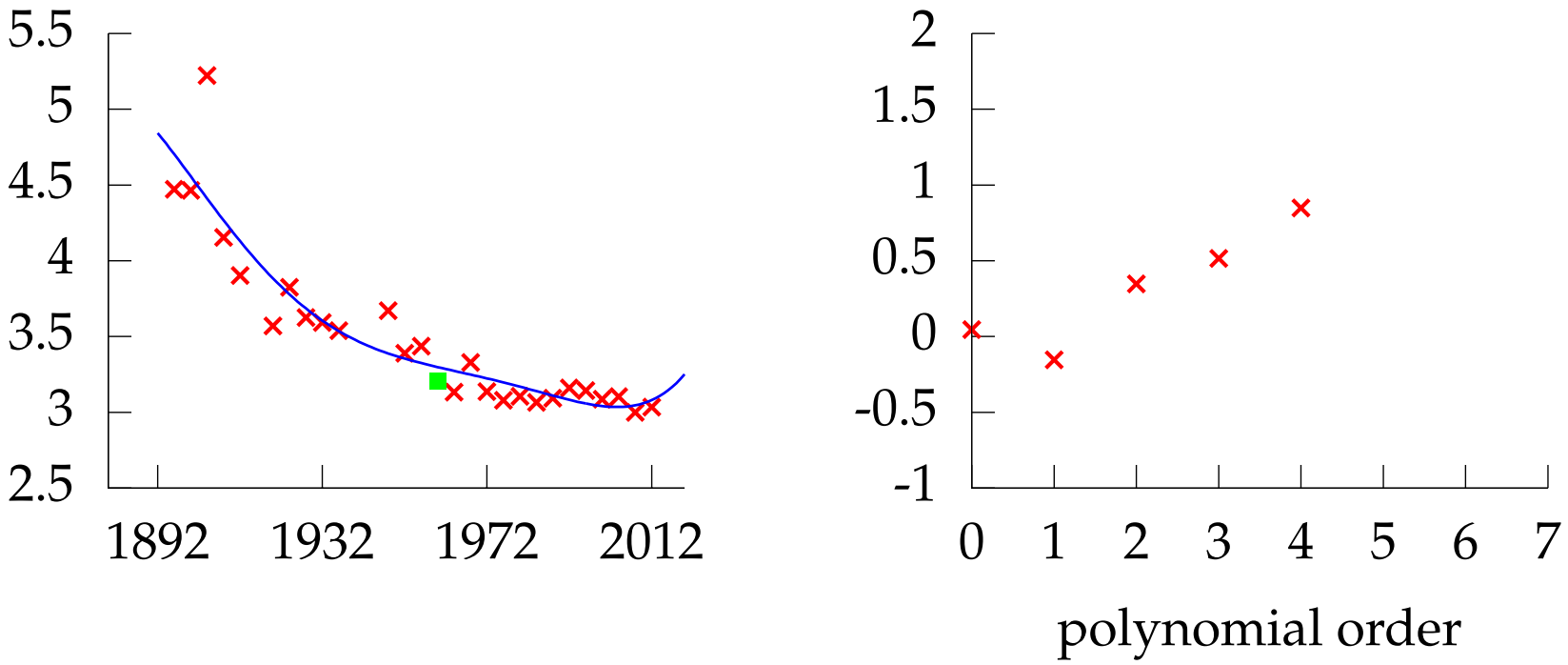
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



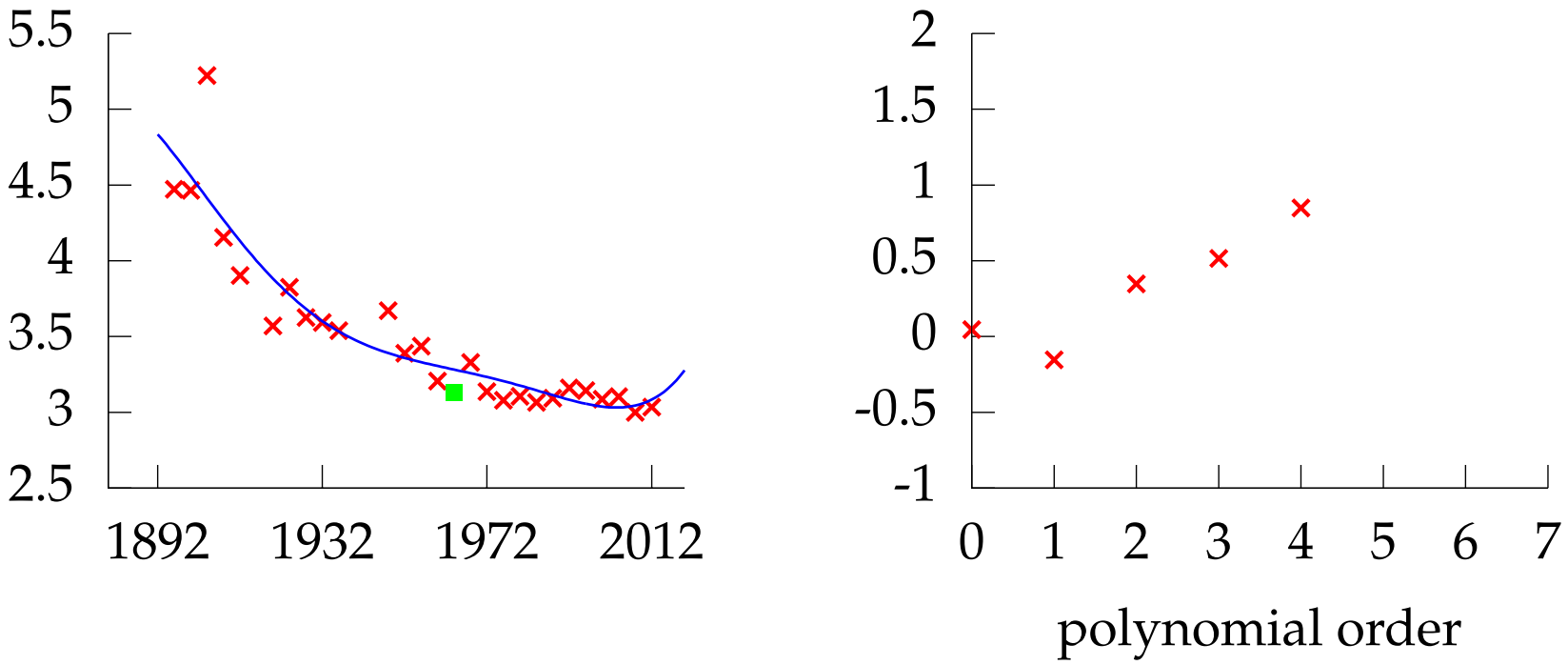
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



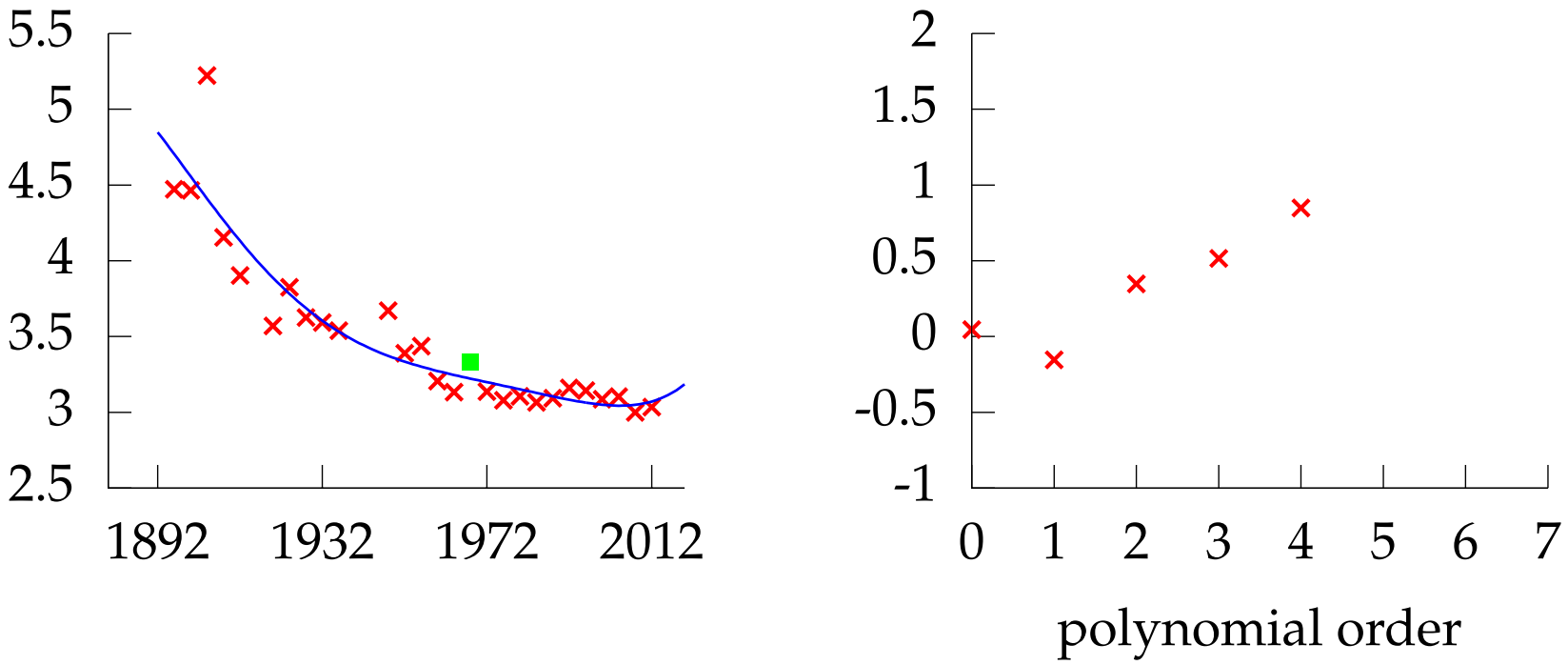
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



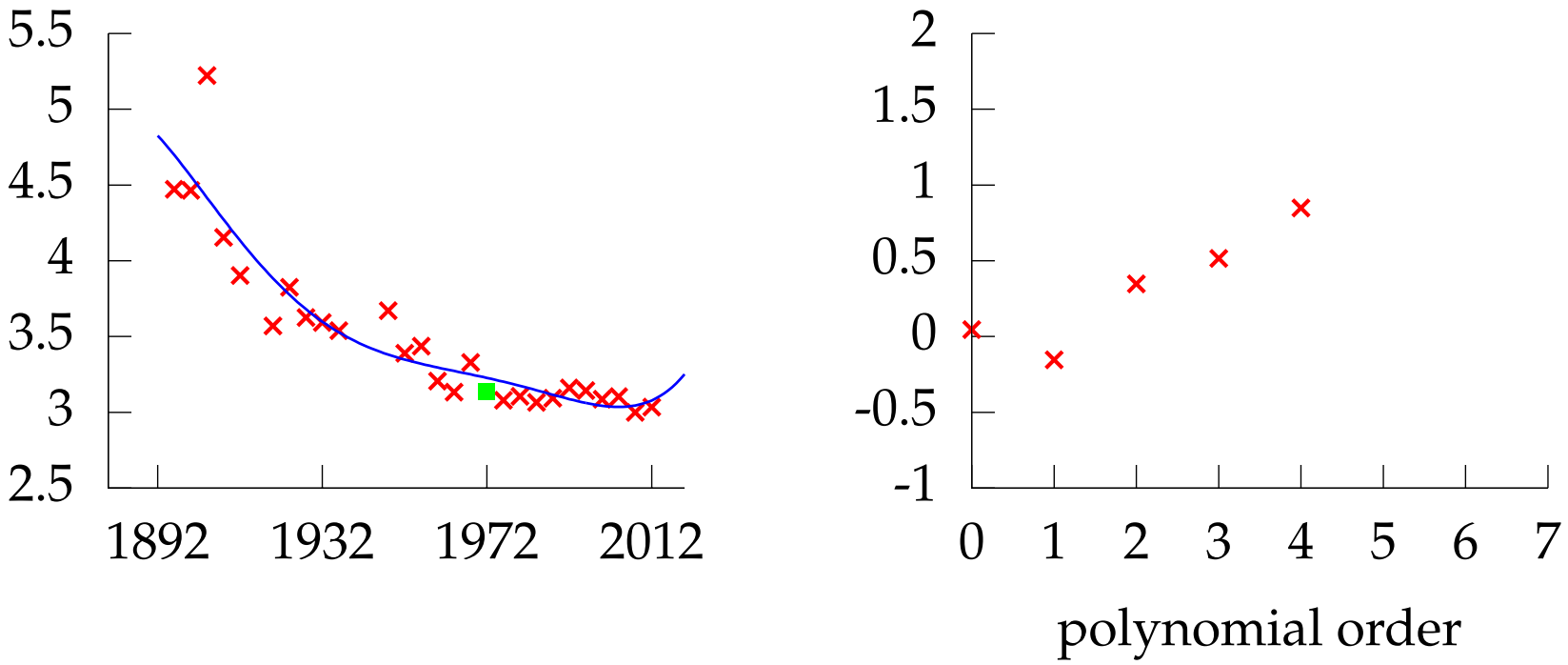
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



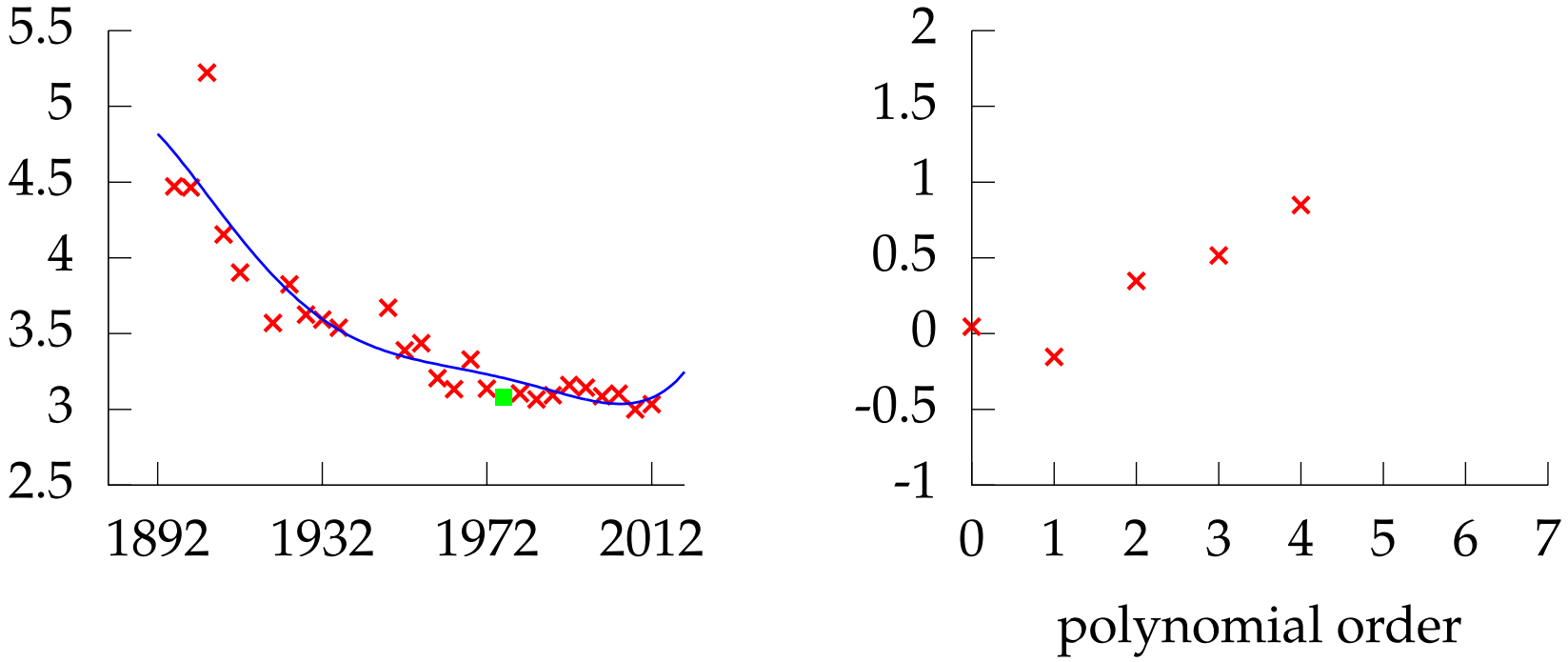
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



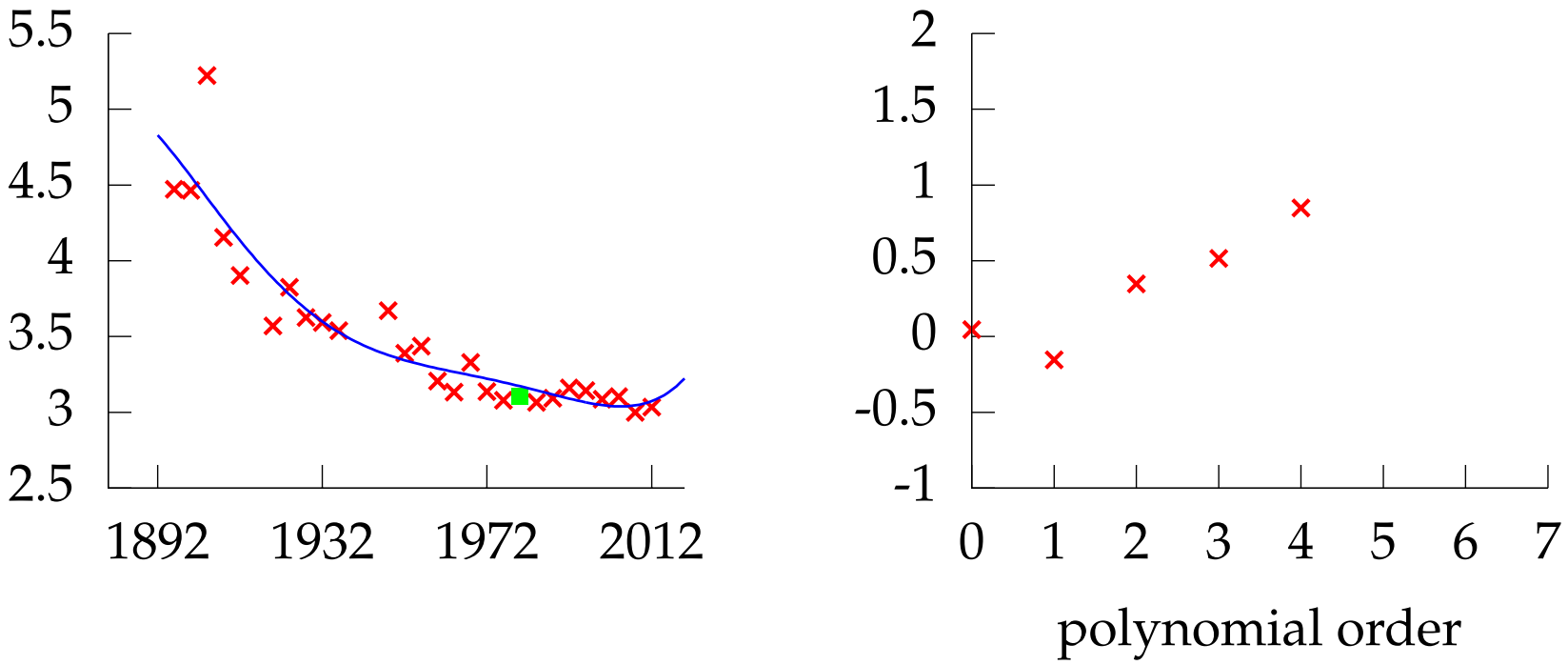
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



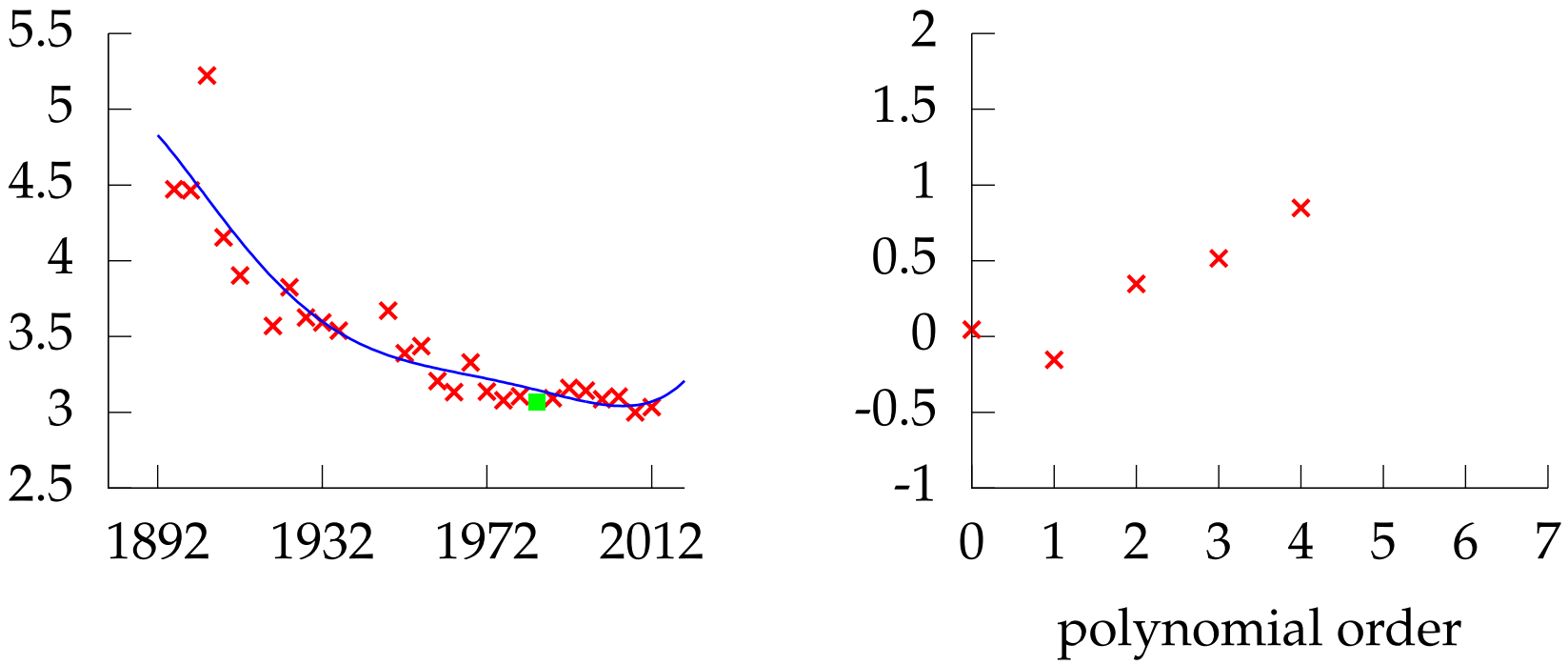
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



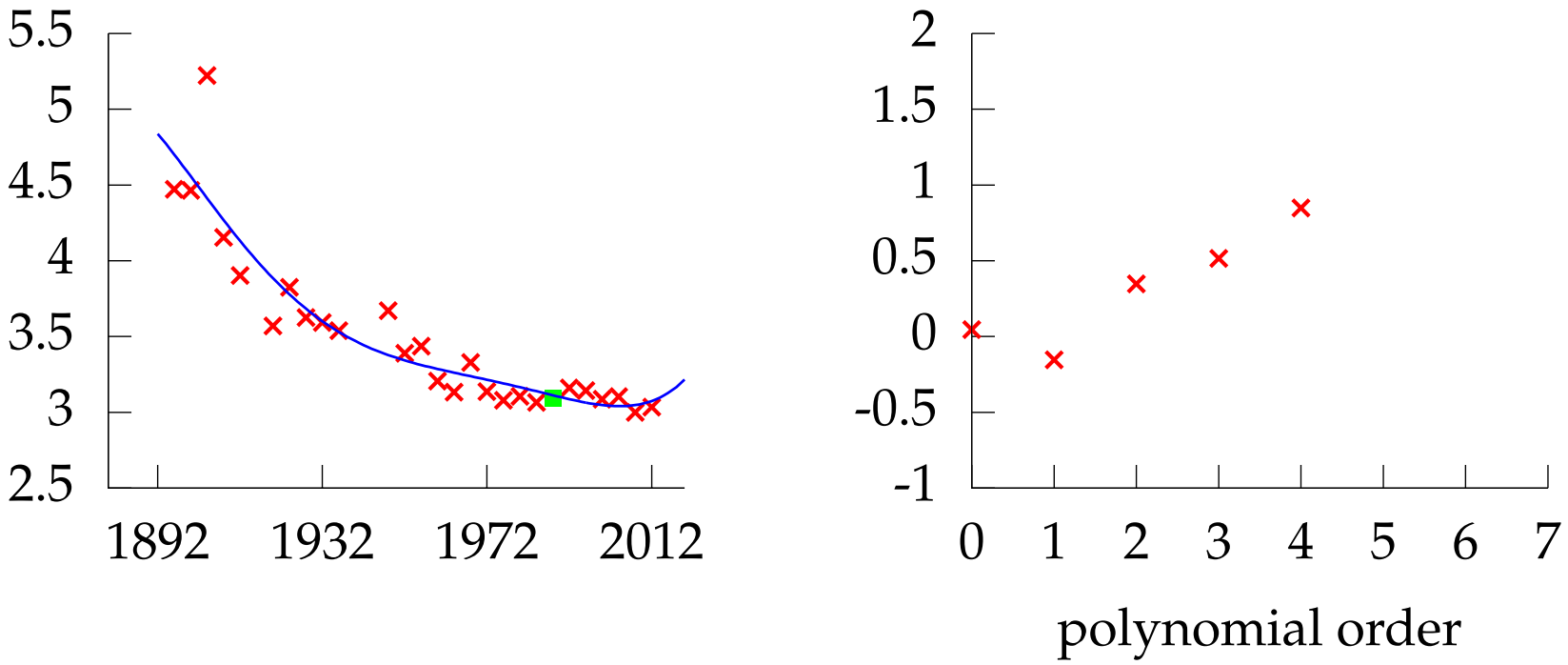
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



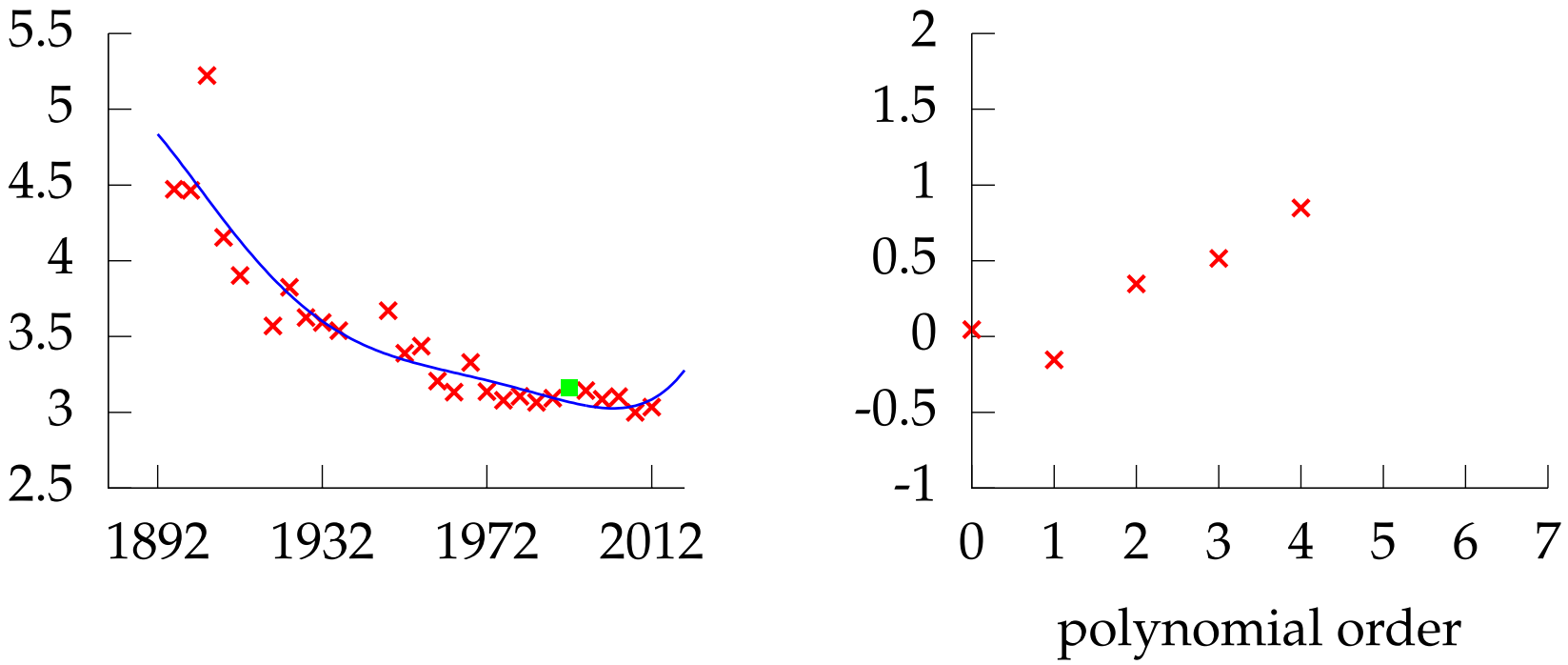
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



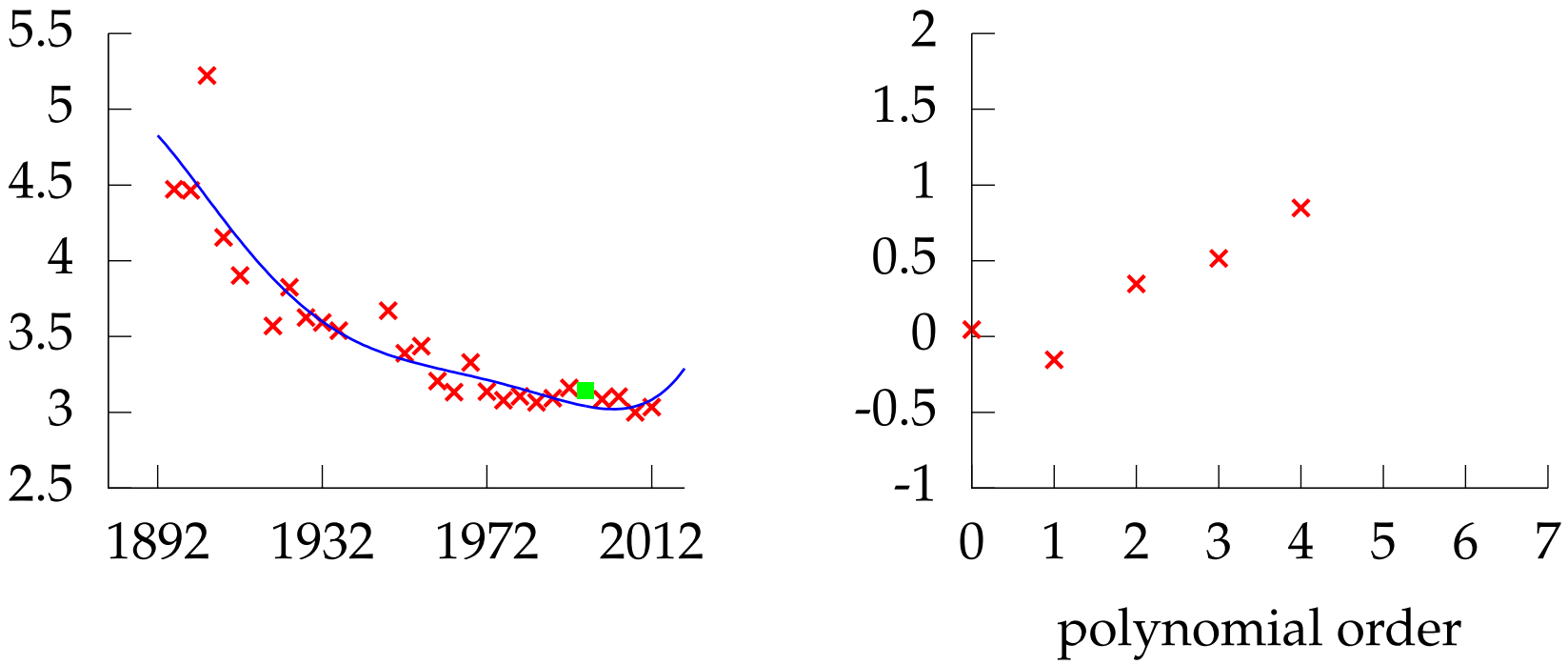
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



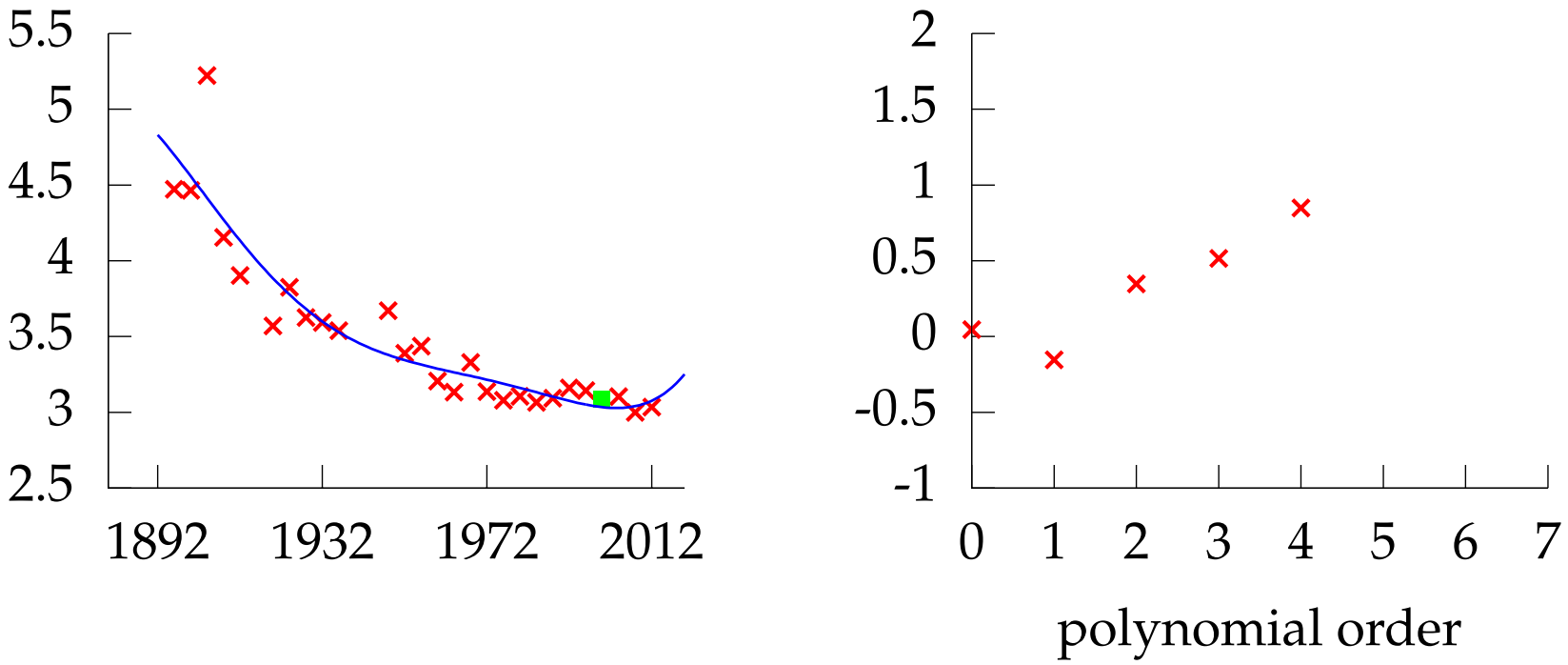
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



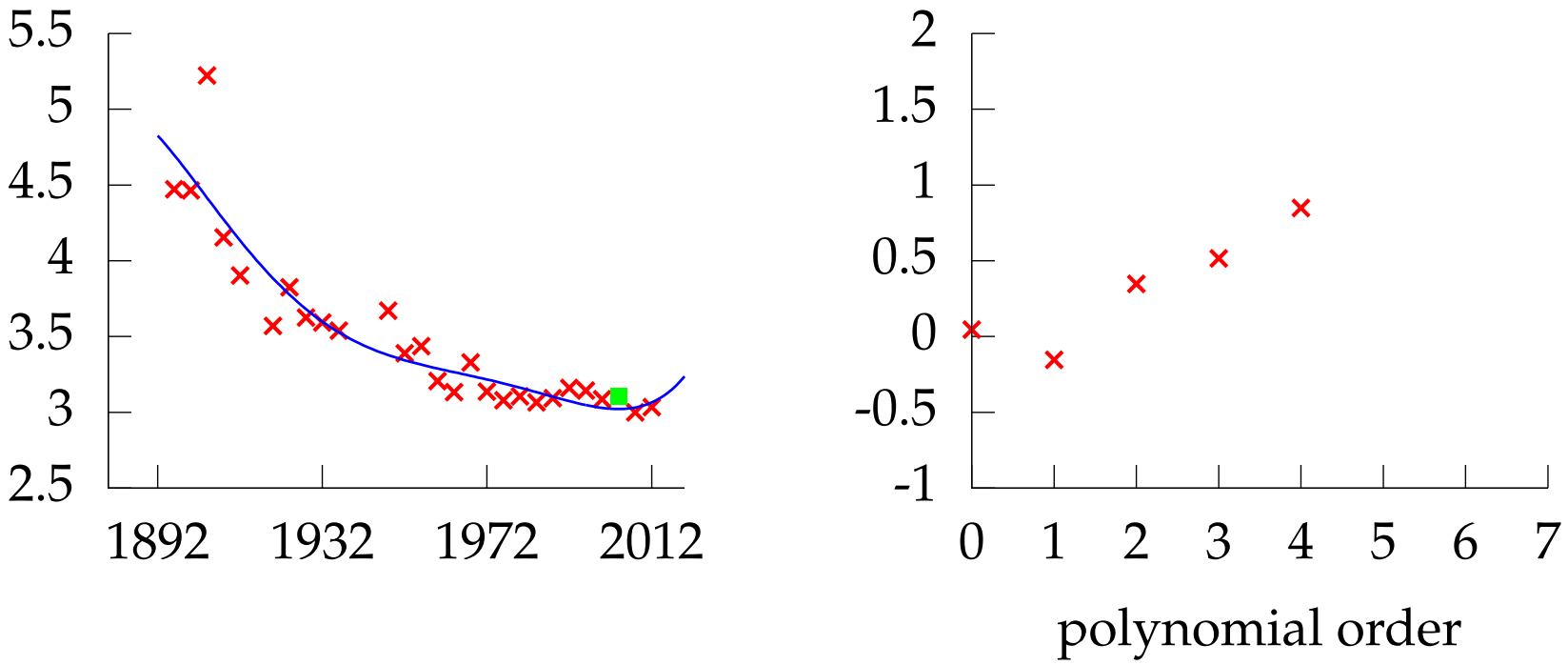
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



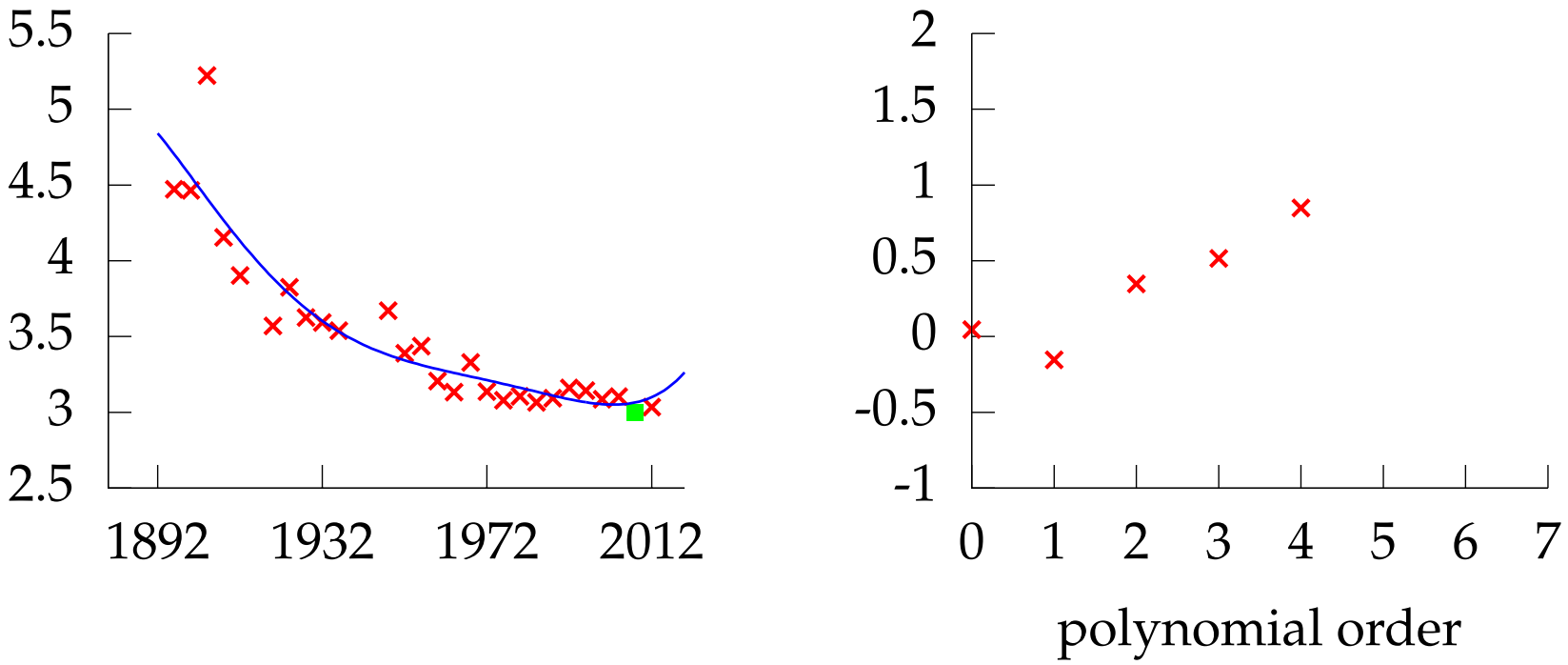
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



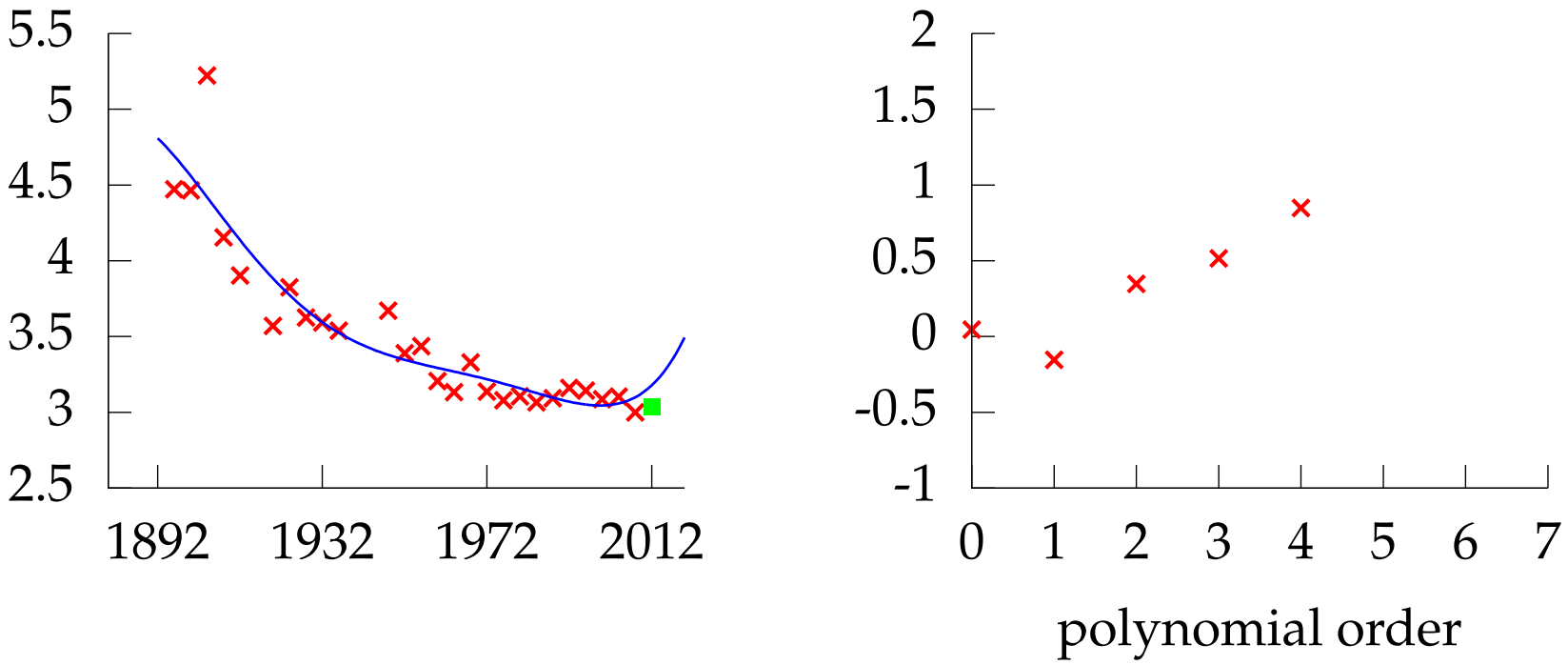
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



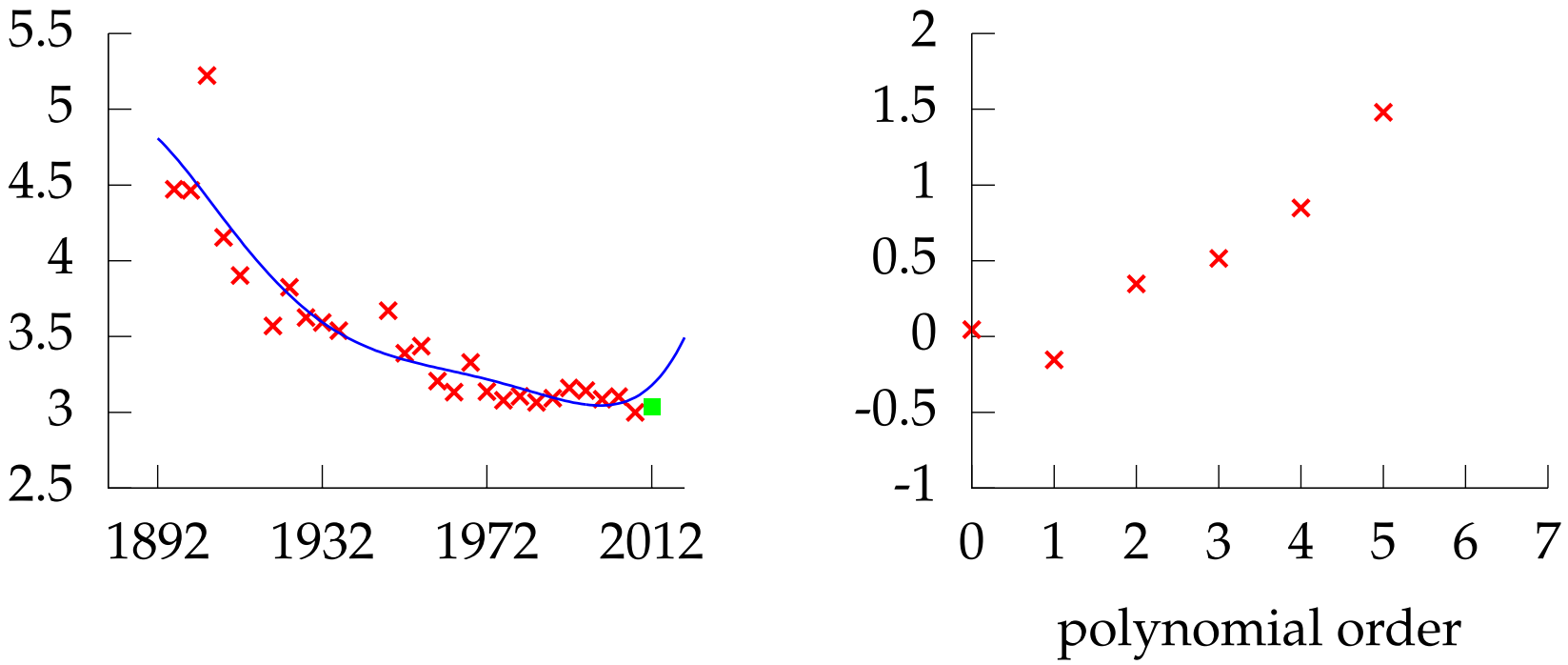
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



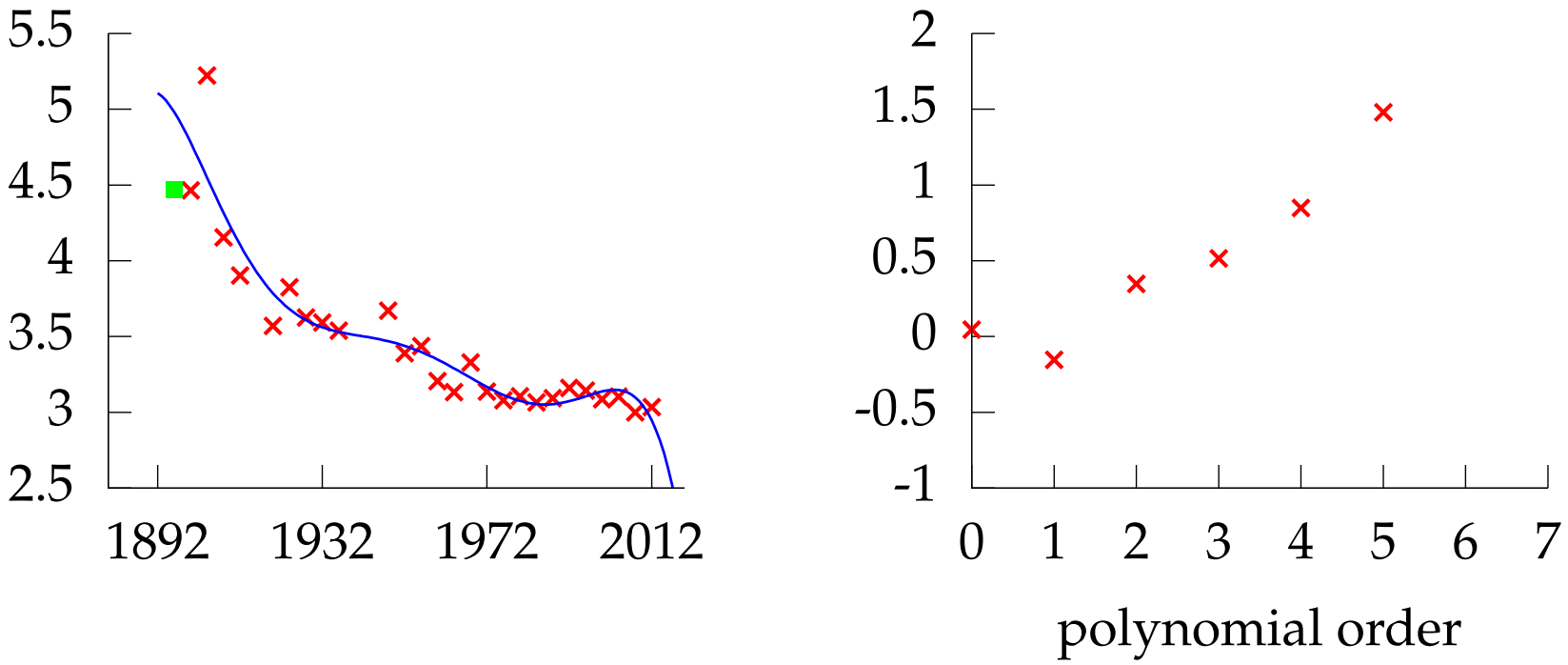
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



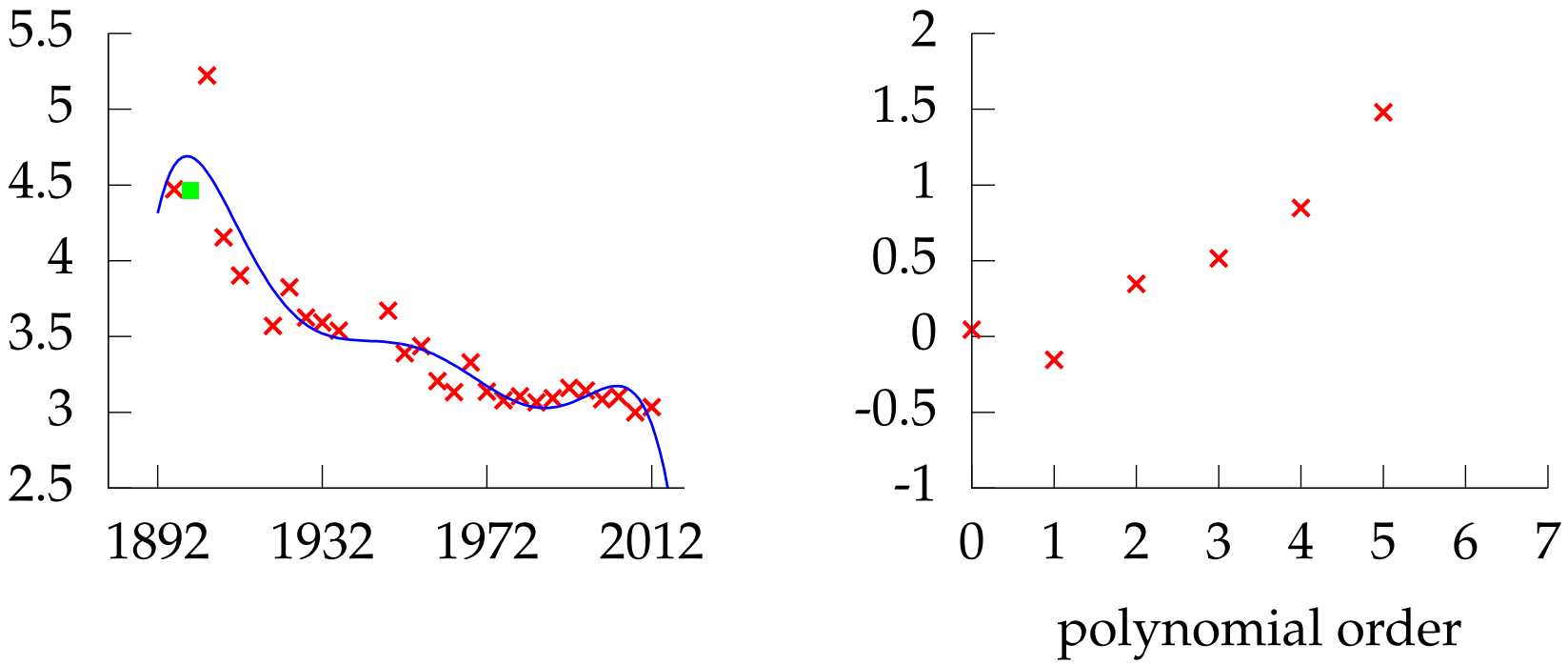
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



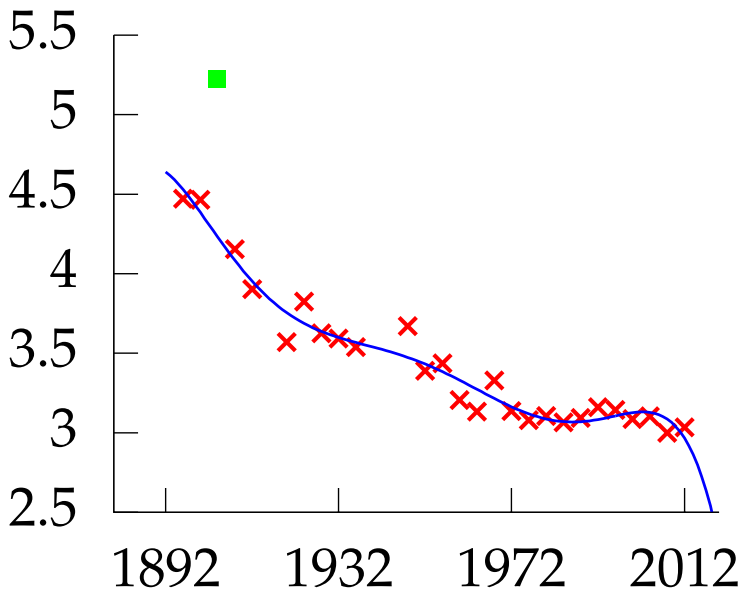
Polynomial order 6, training error -32.237, leave one out error 1.5047.

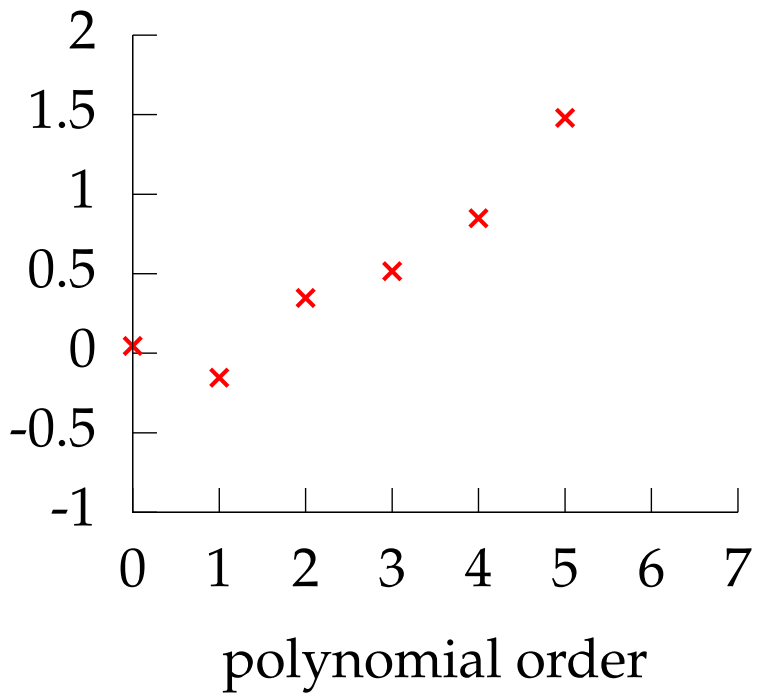
Leave One Out Error



Polynomial order 6, training error -32.237, leave one out error 1.5047.

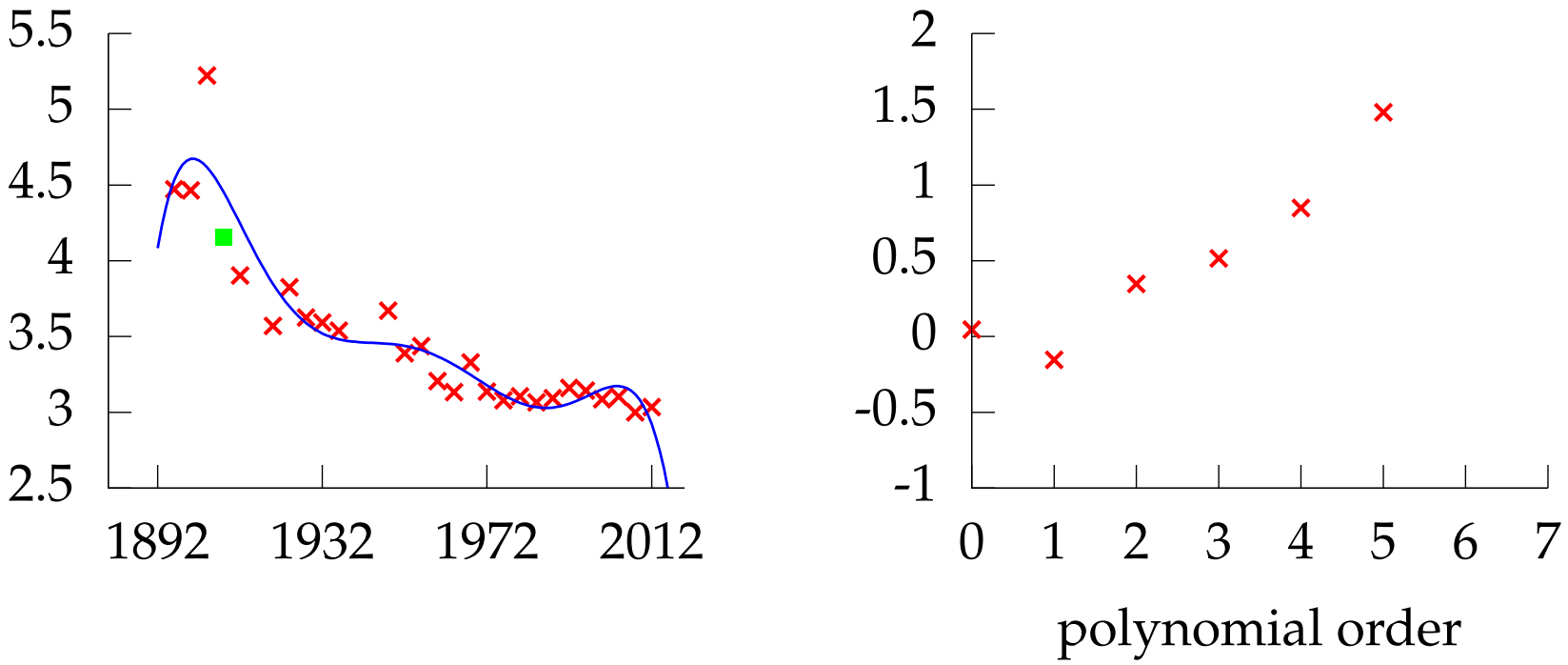
Leave One Out Error





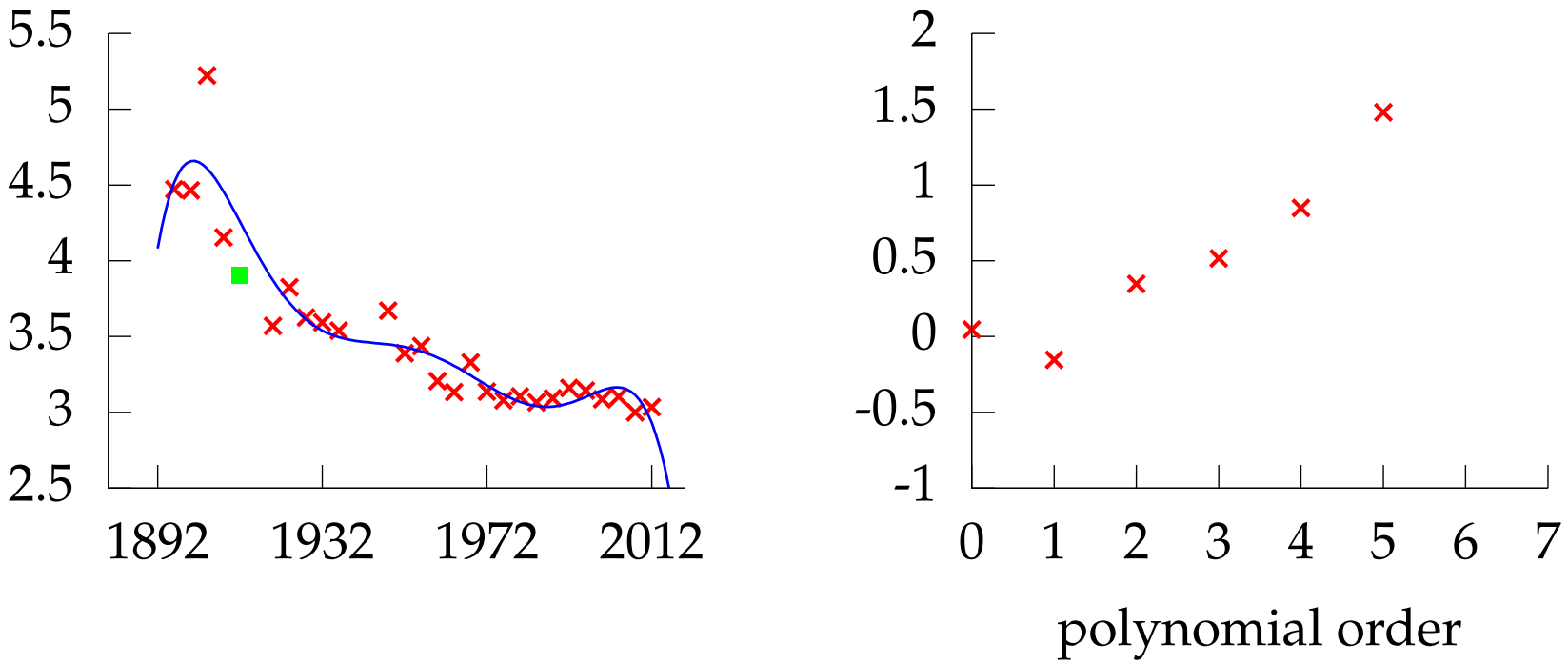
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



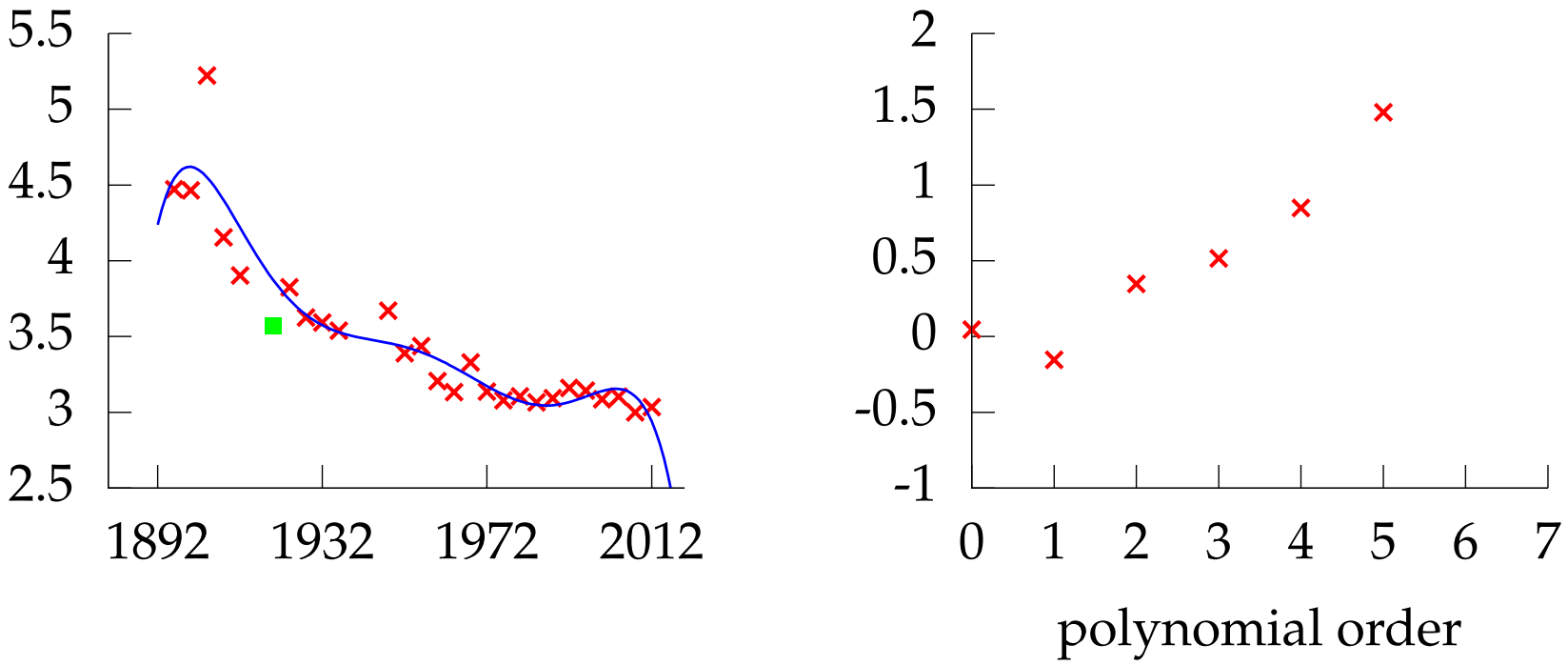
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



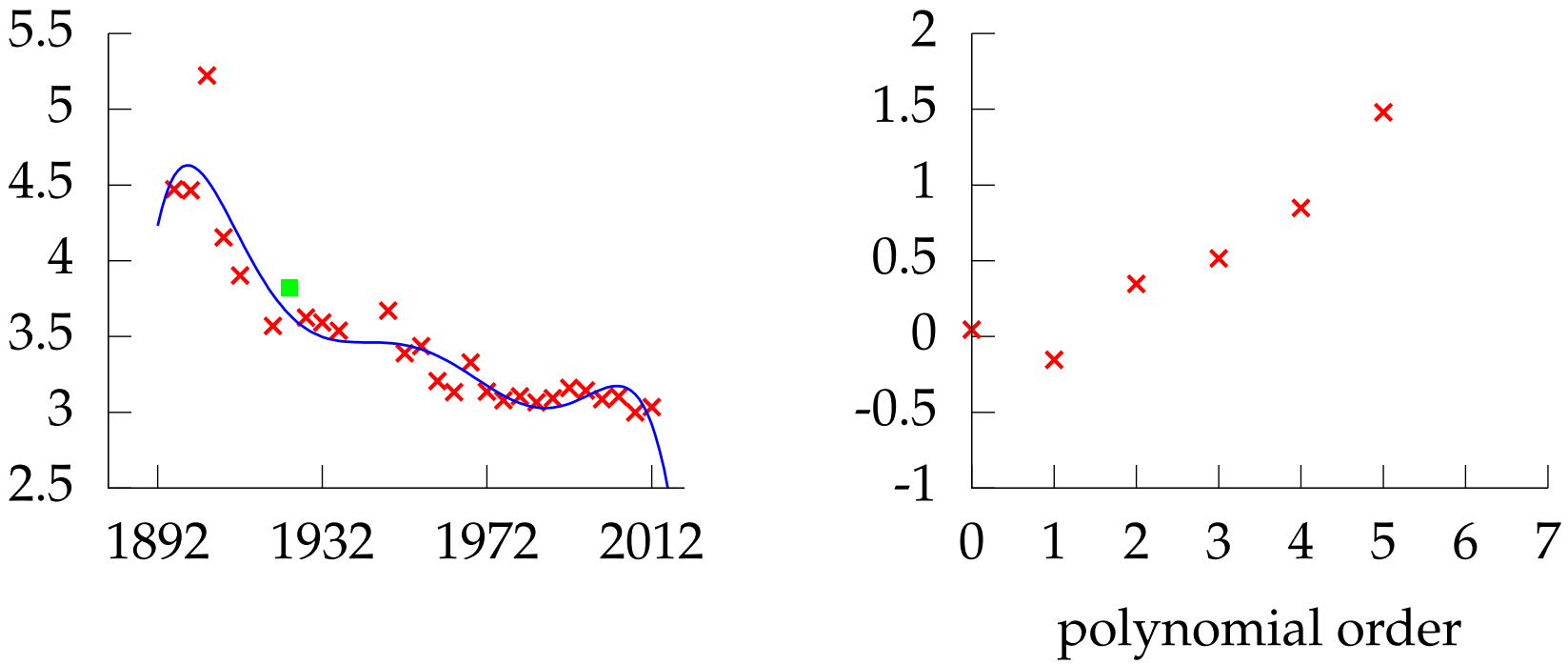
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



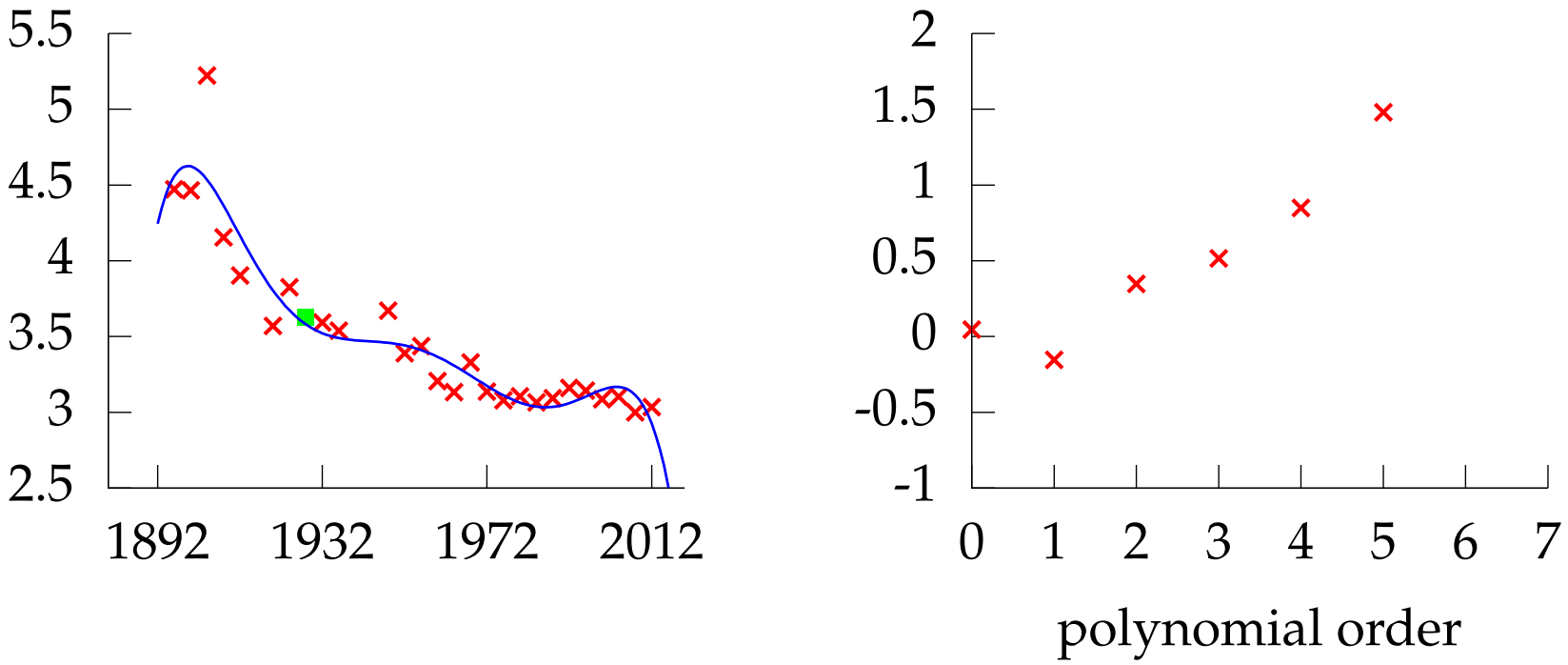
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



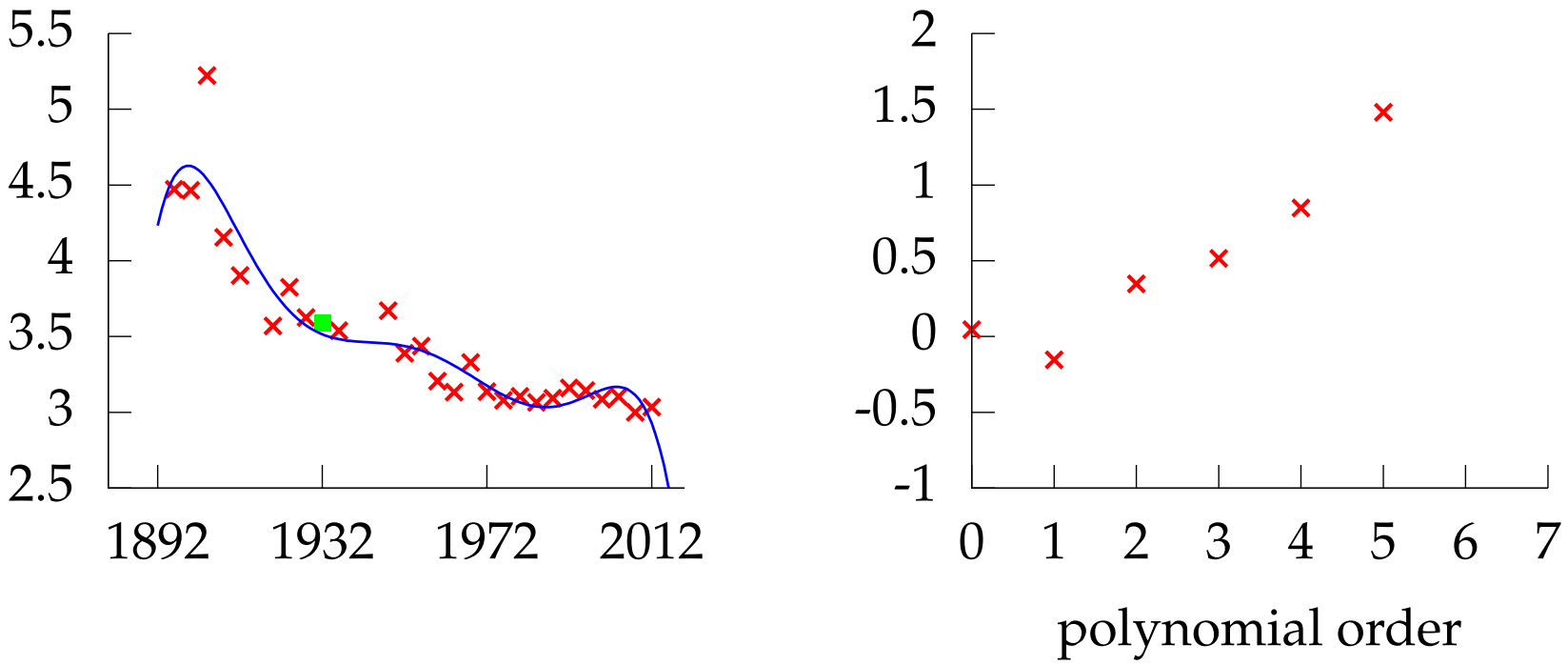
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



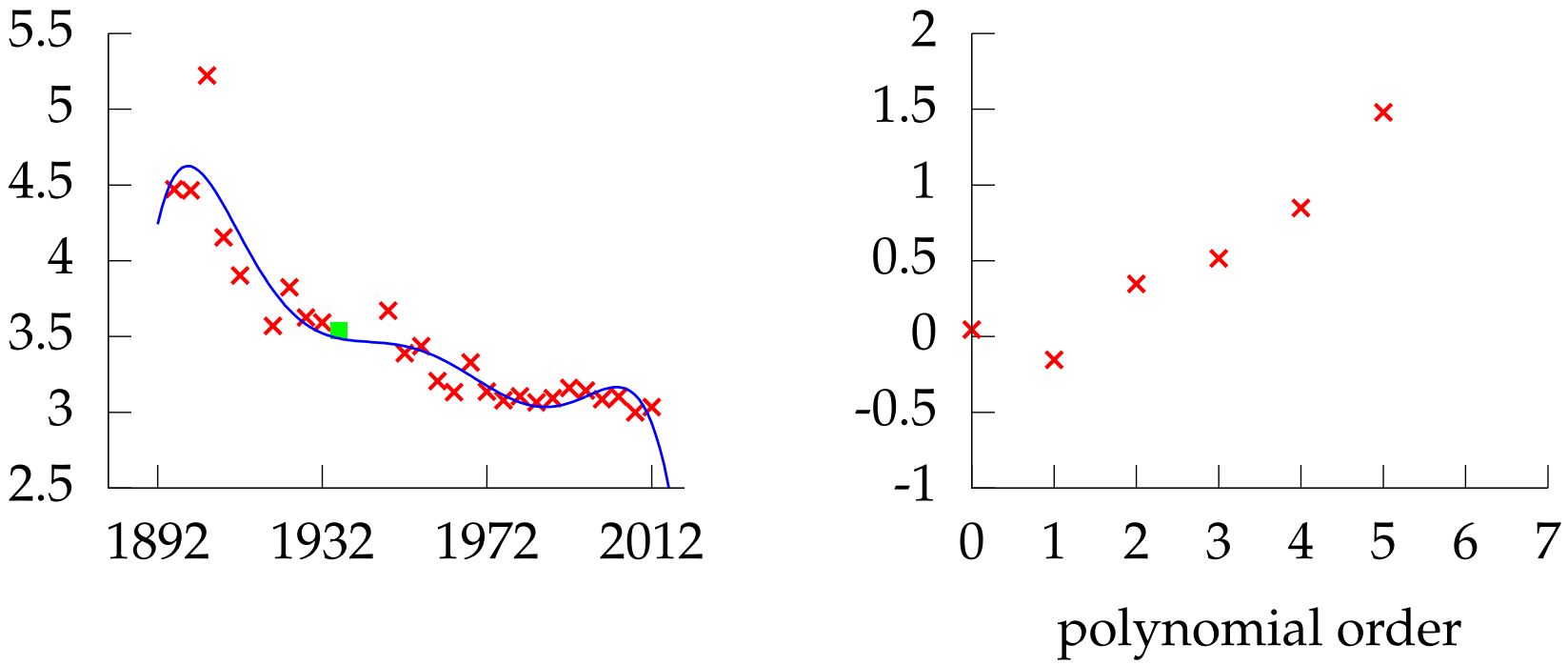
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



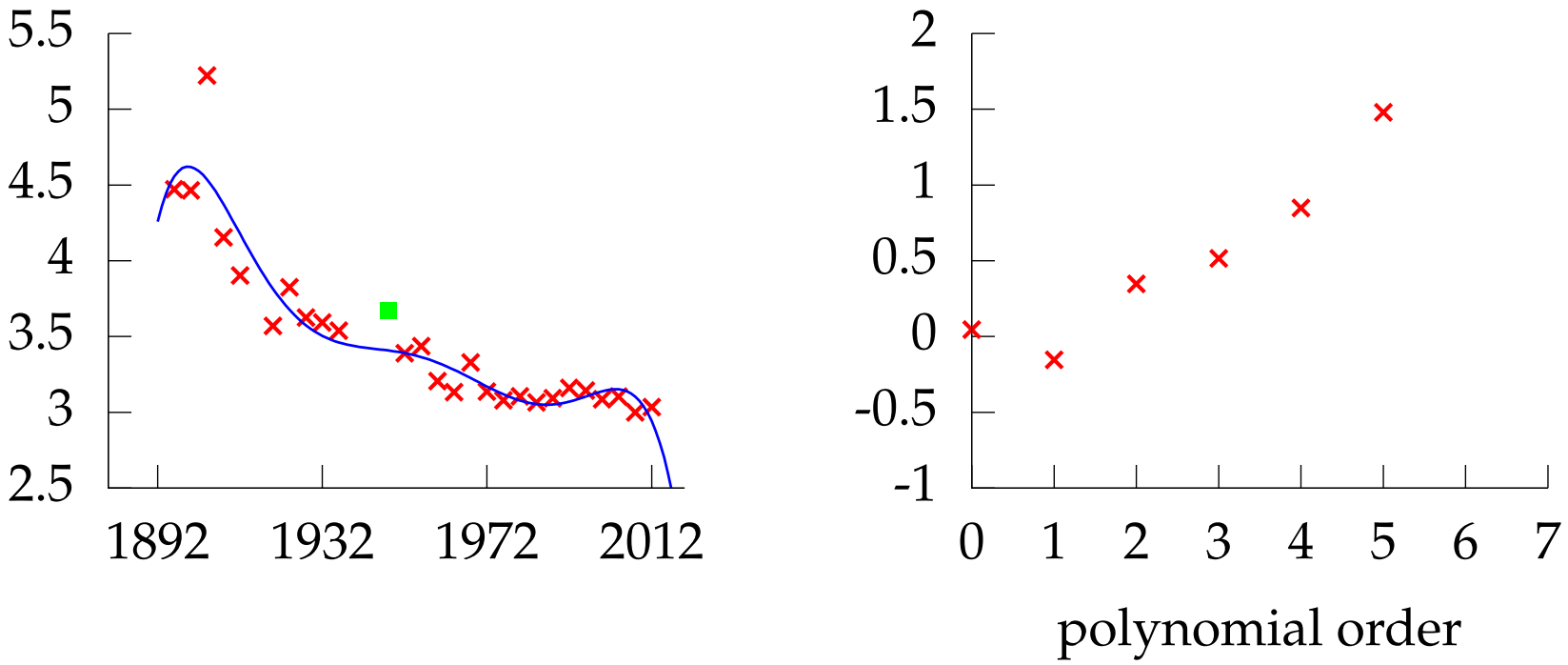
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



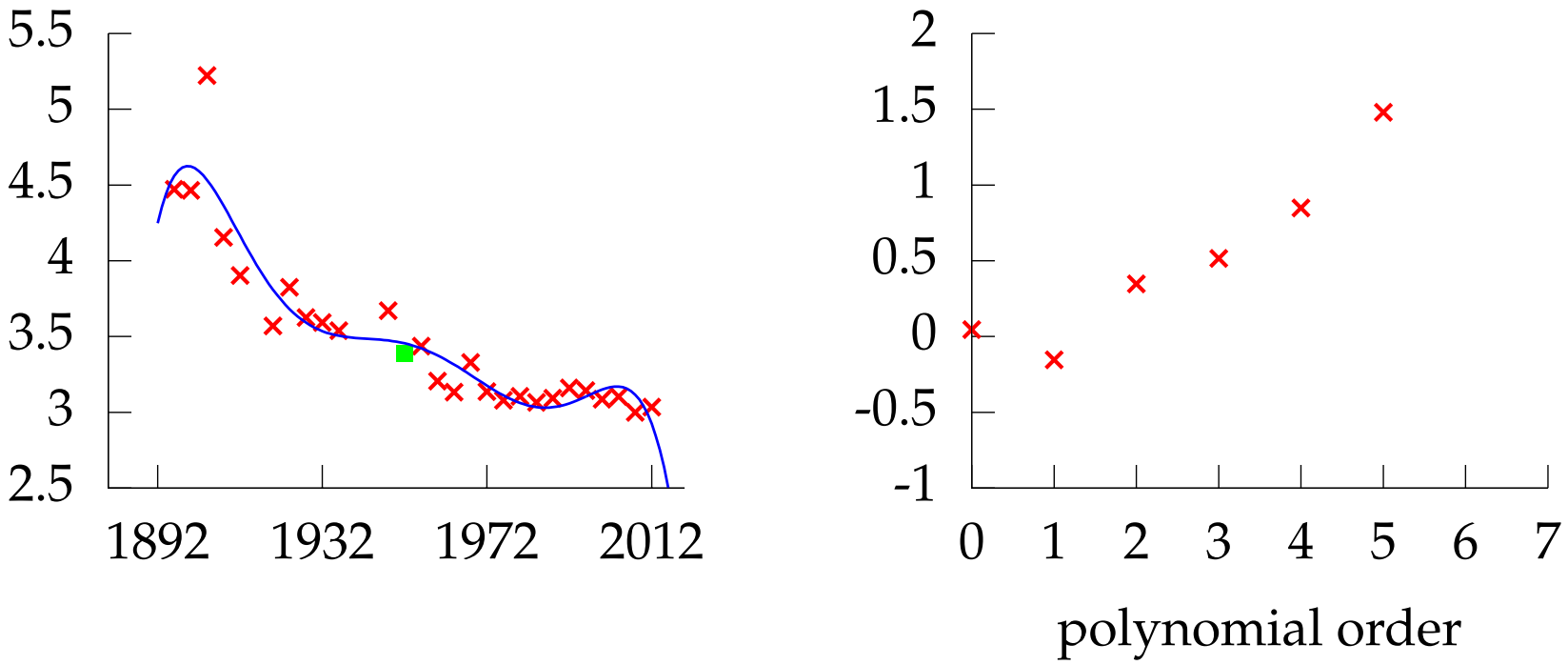
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



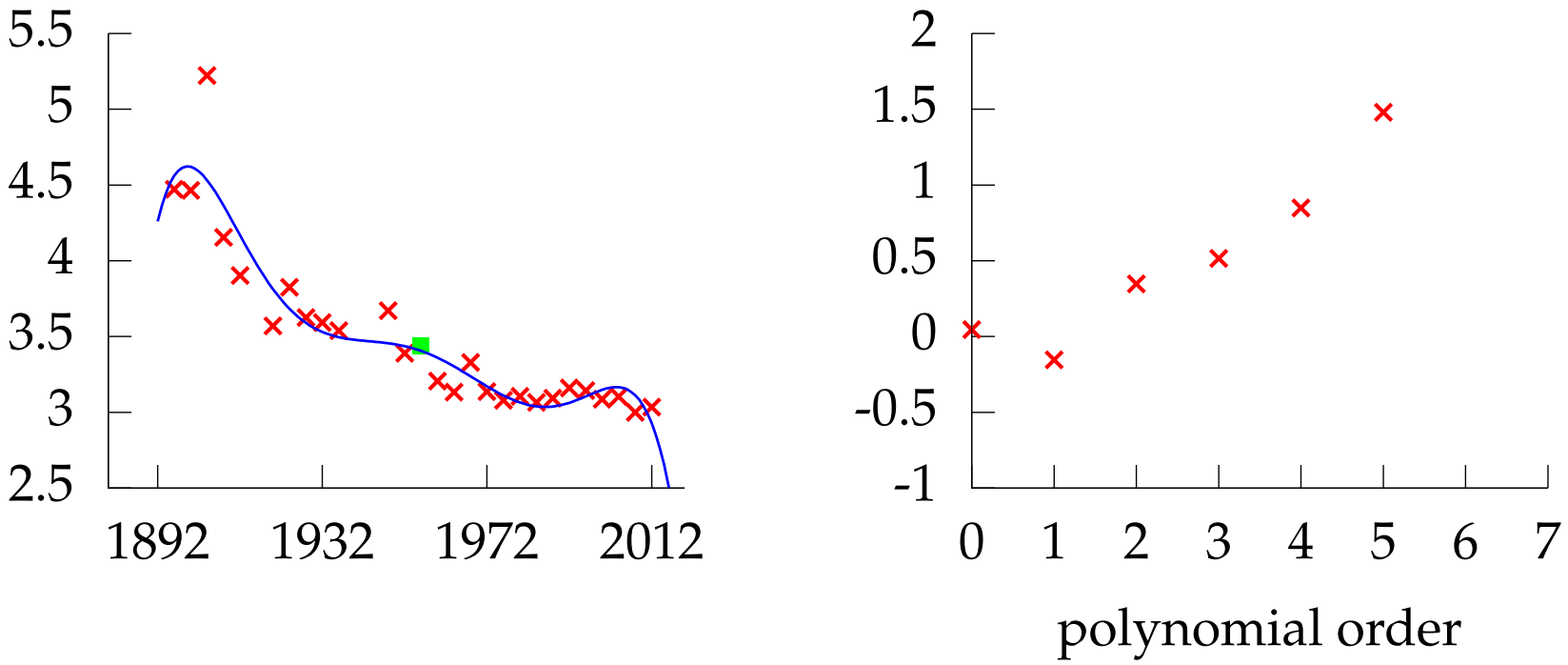
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



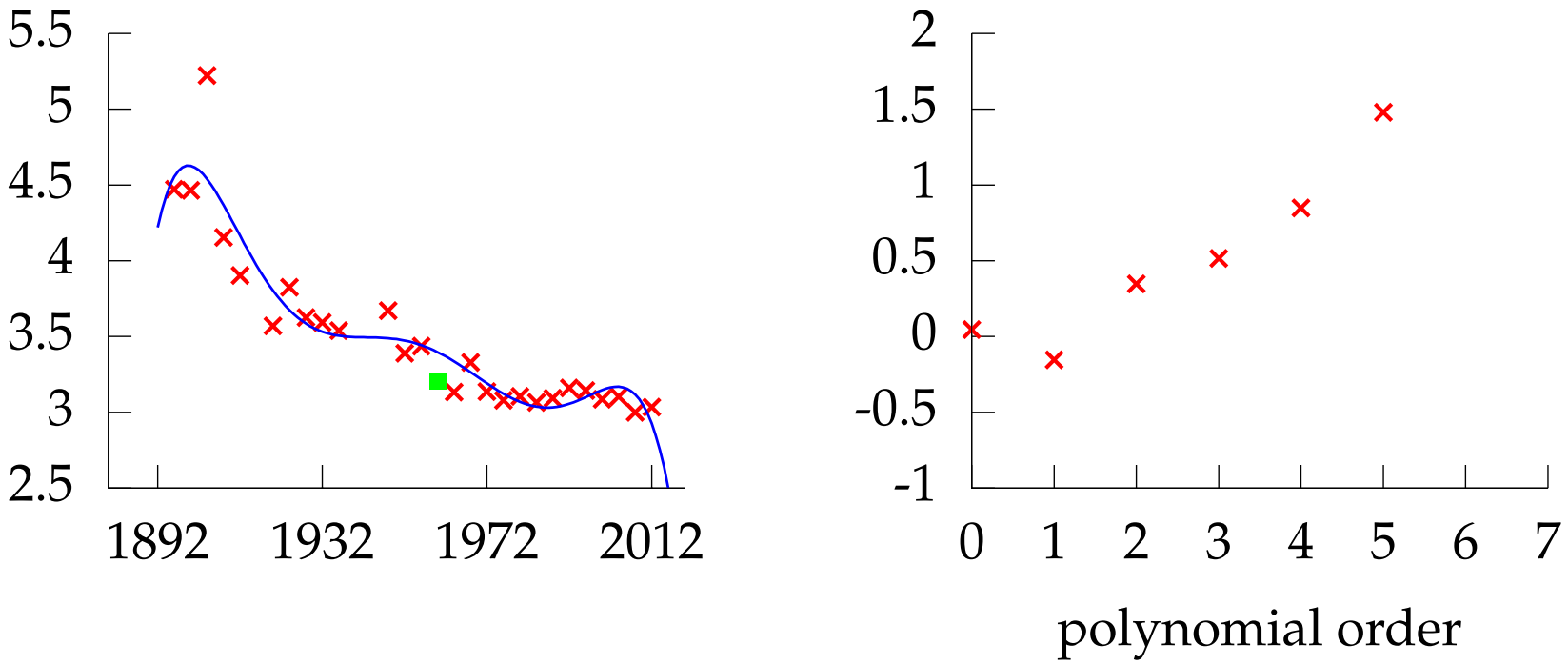
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



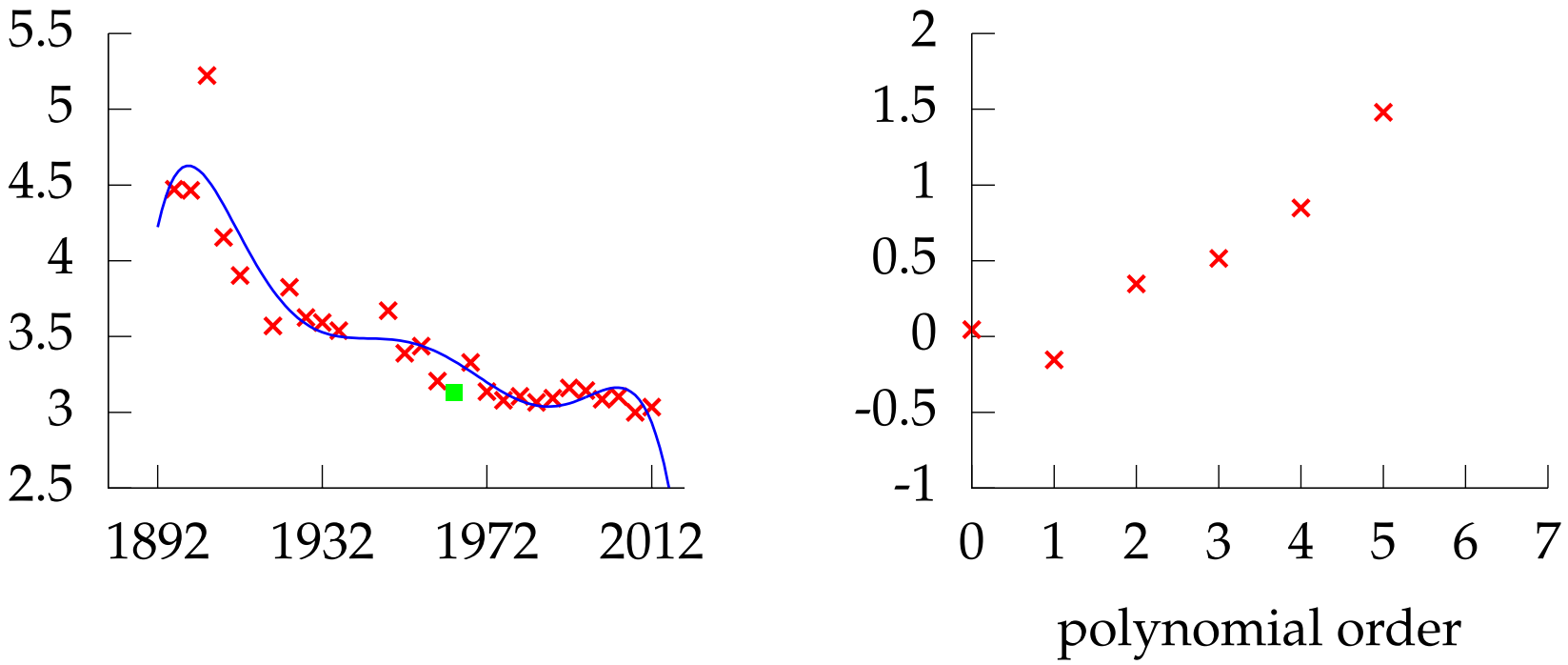
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



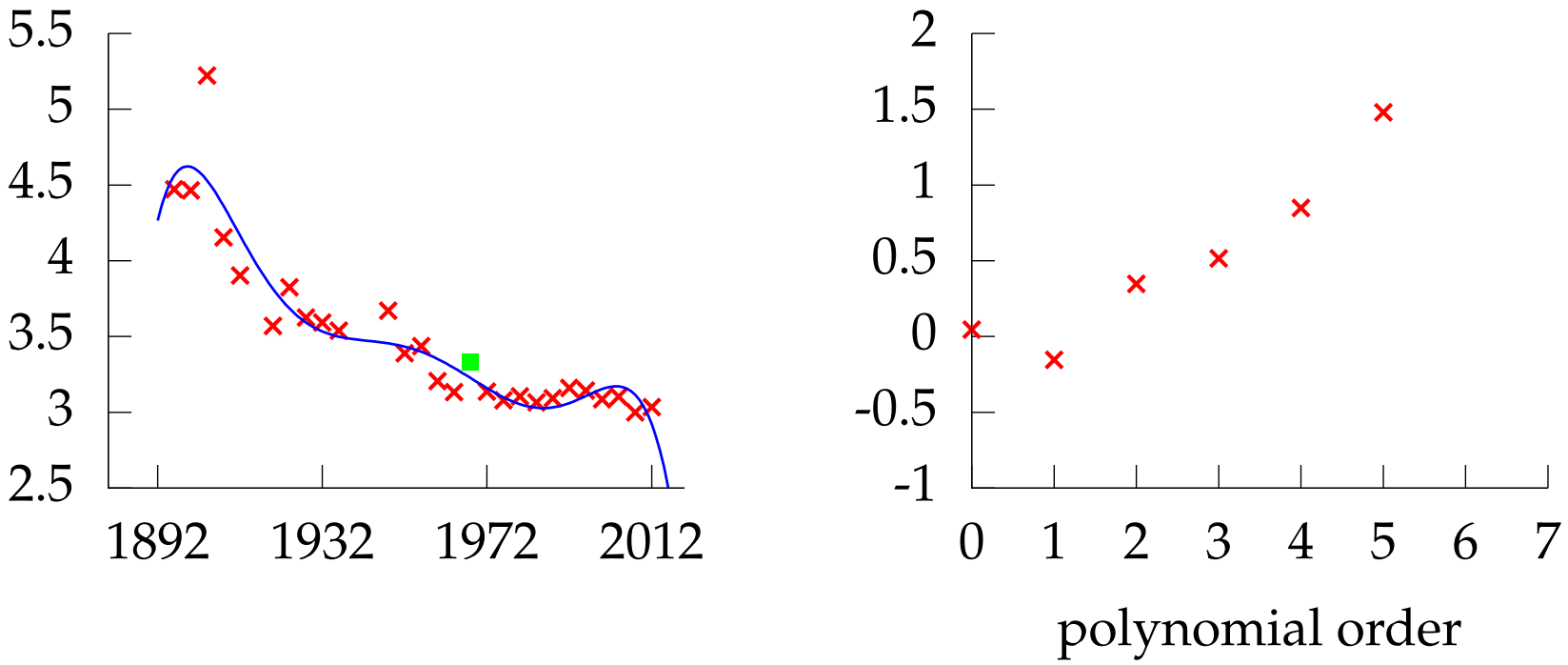
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



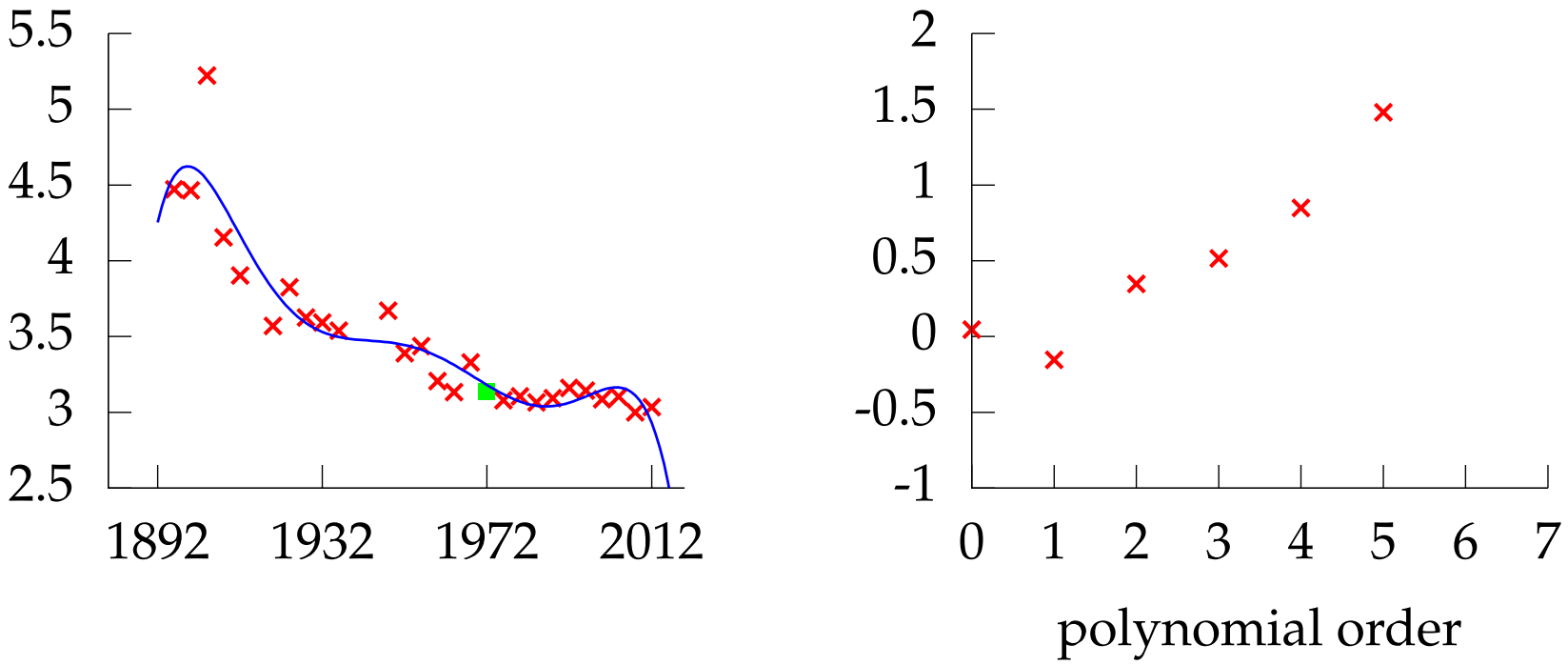
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



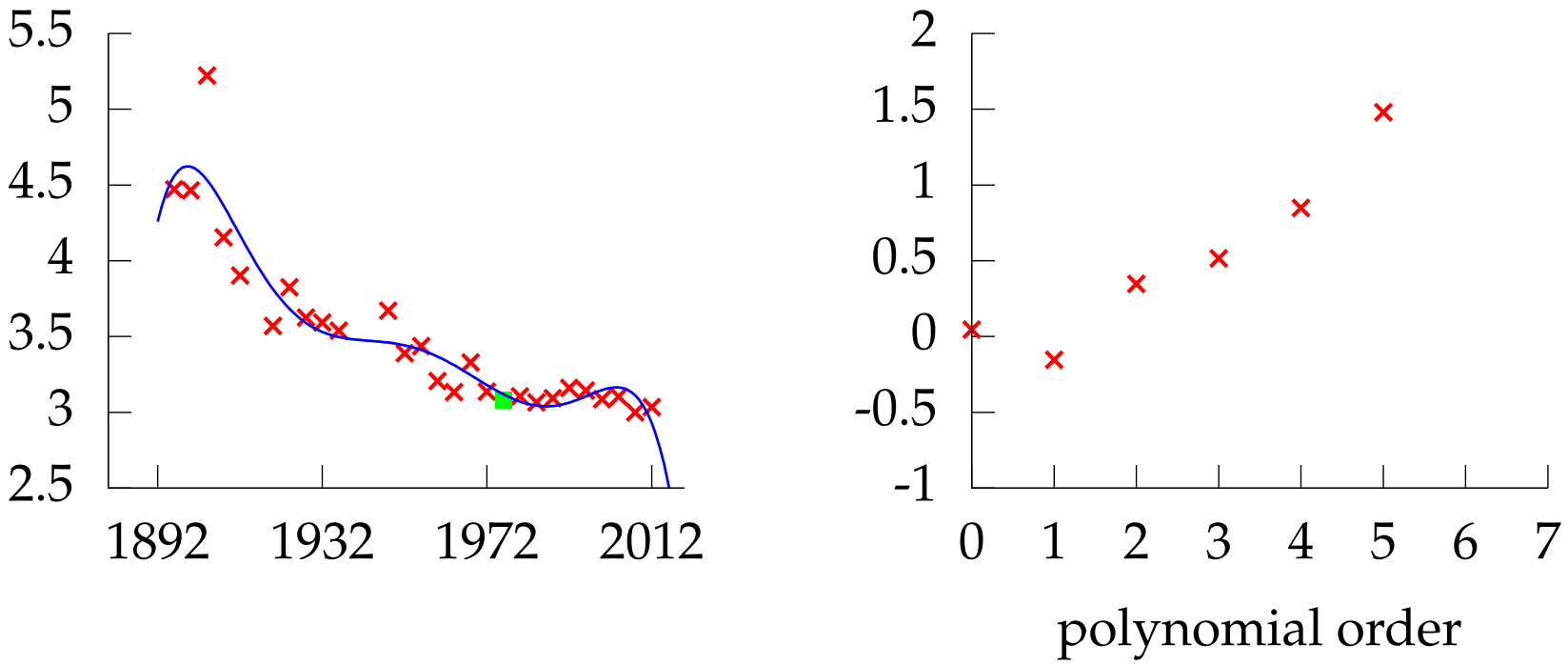
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



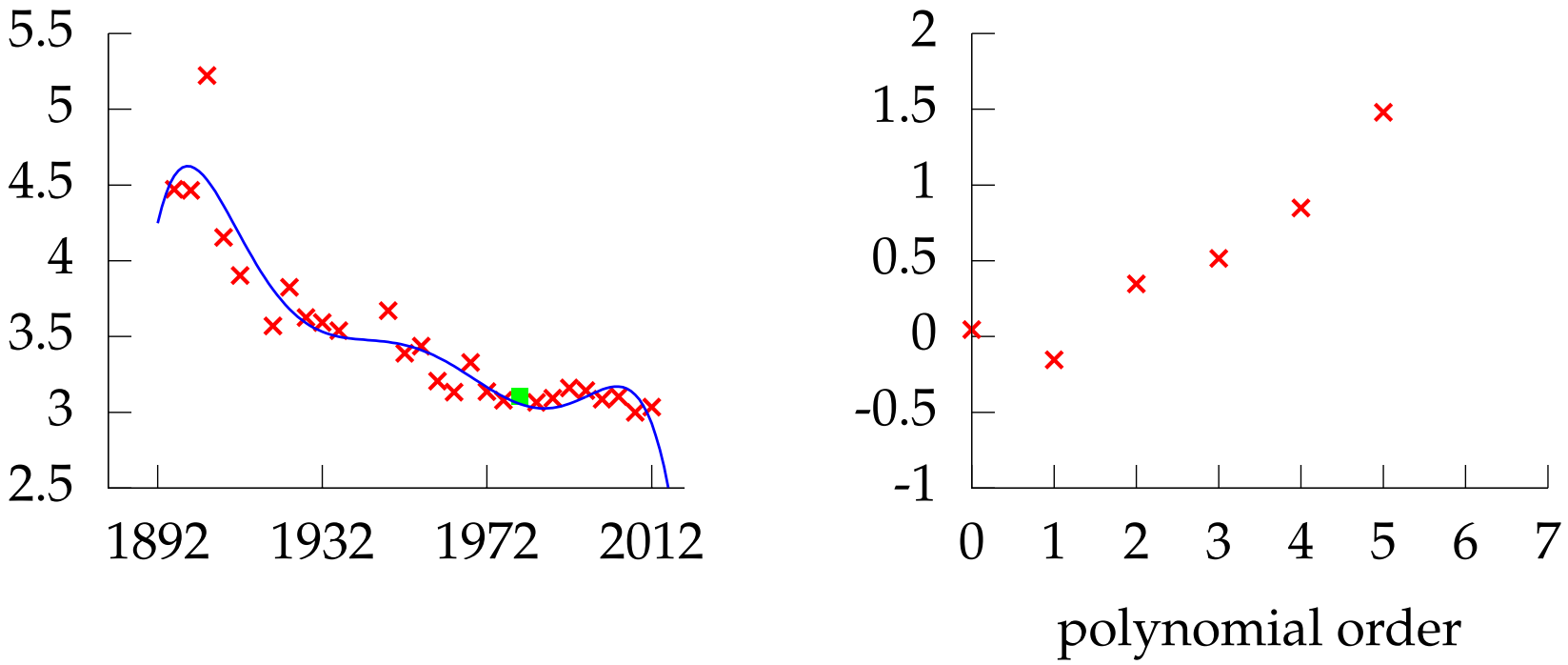
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



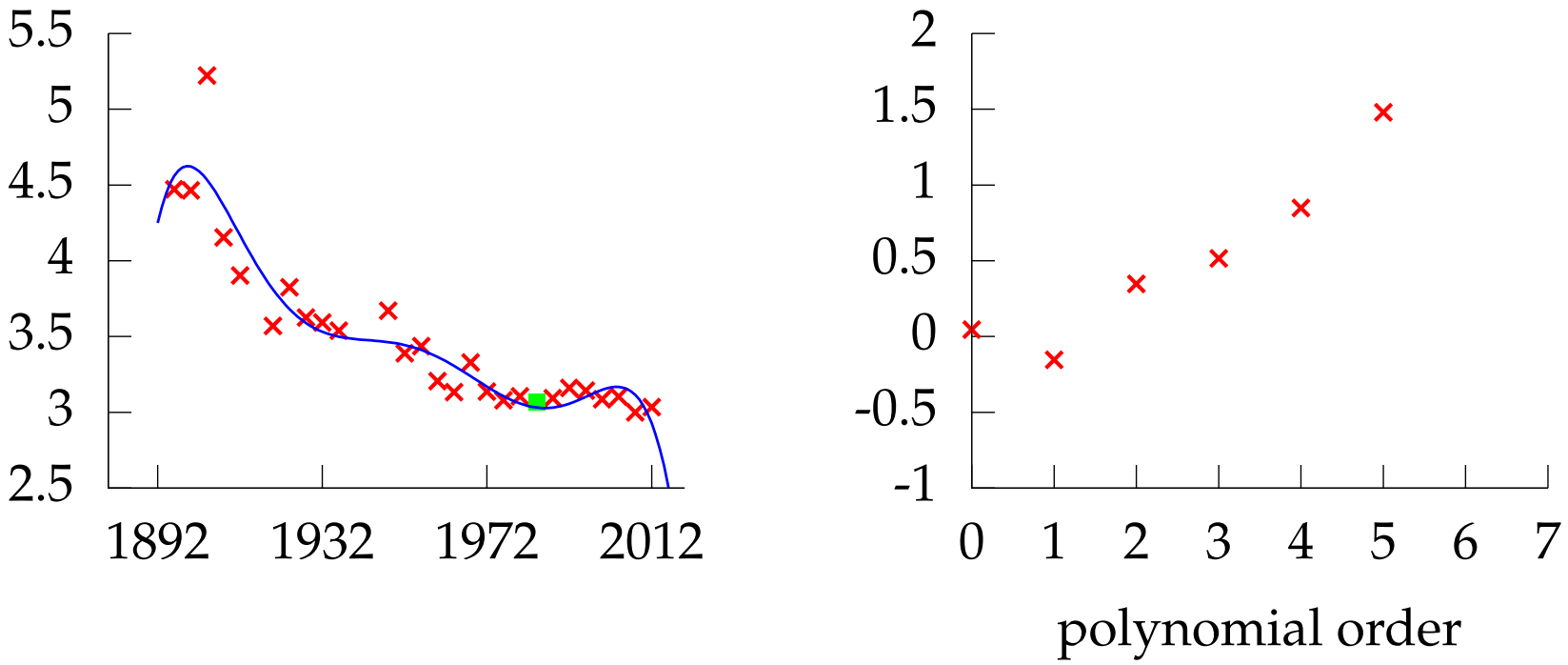
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



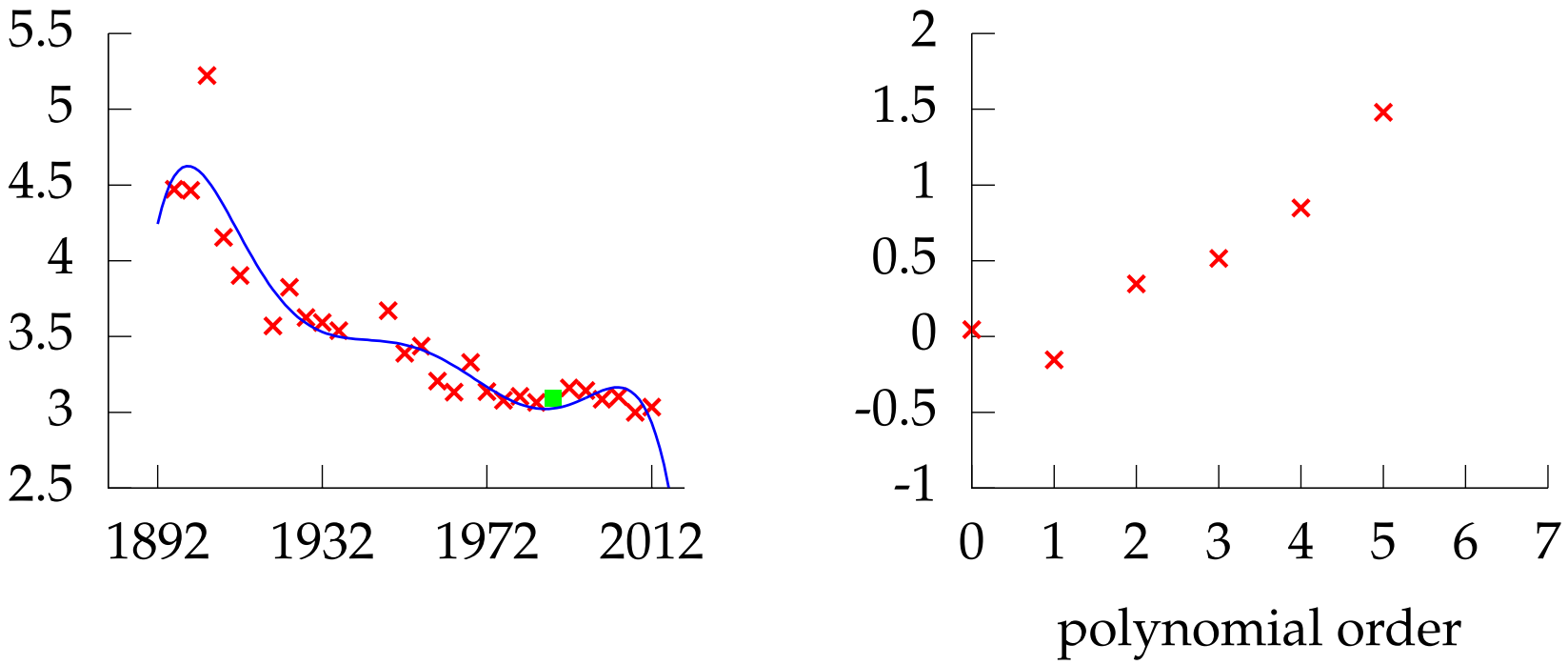
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



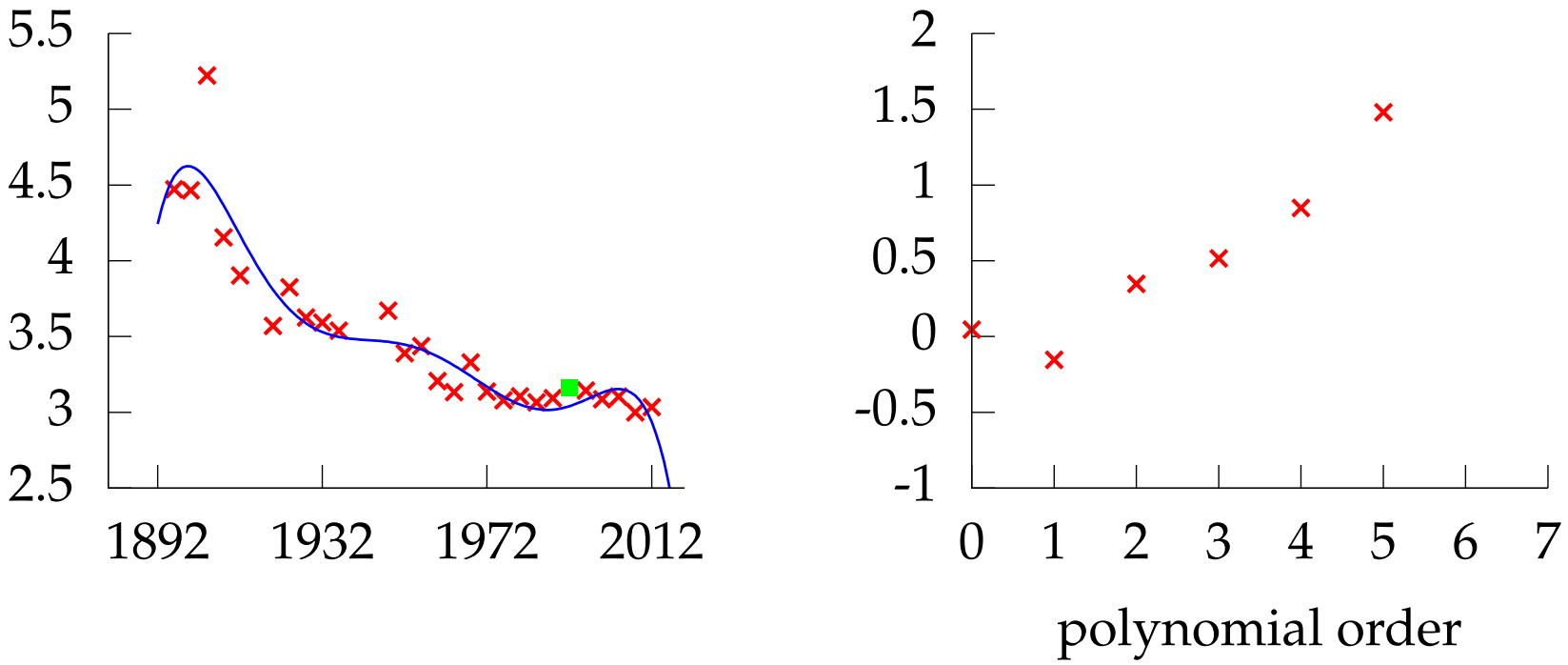
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



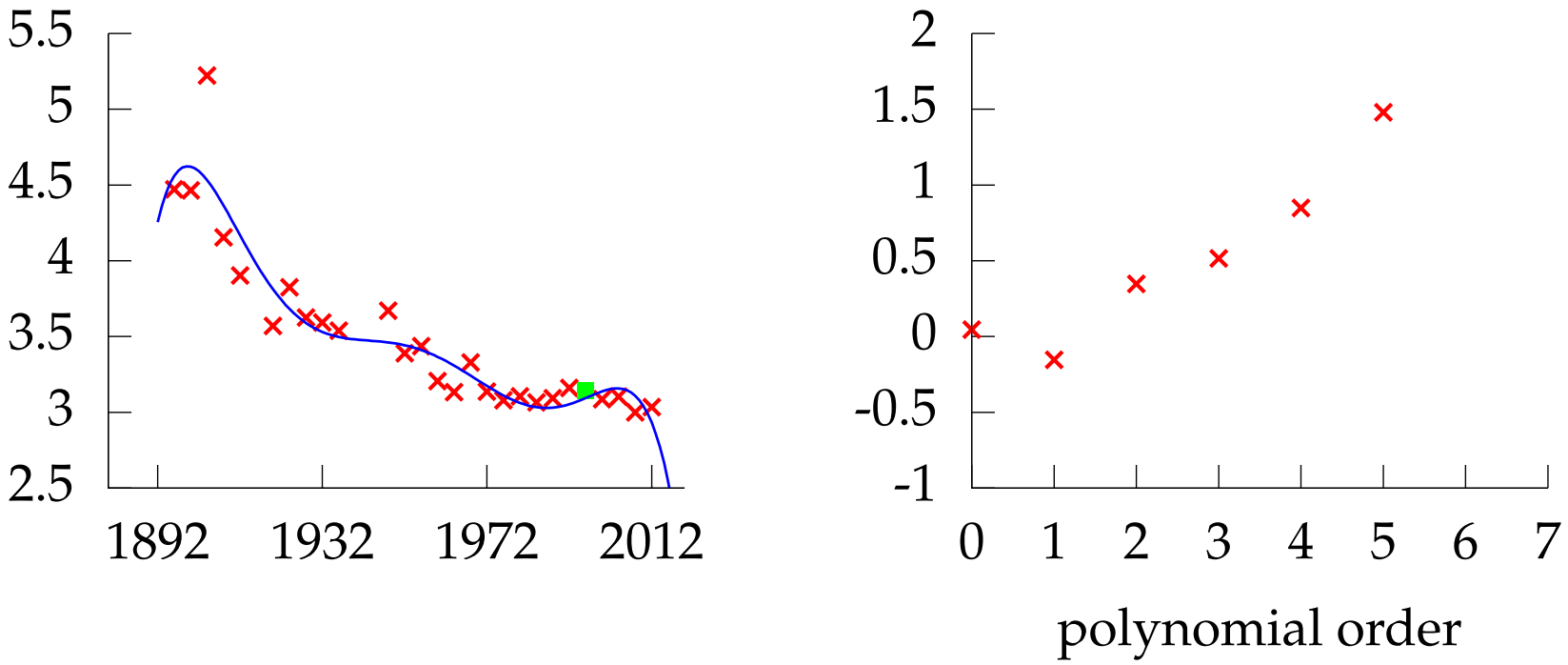
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



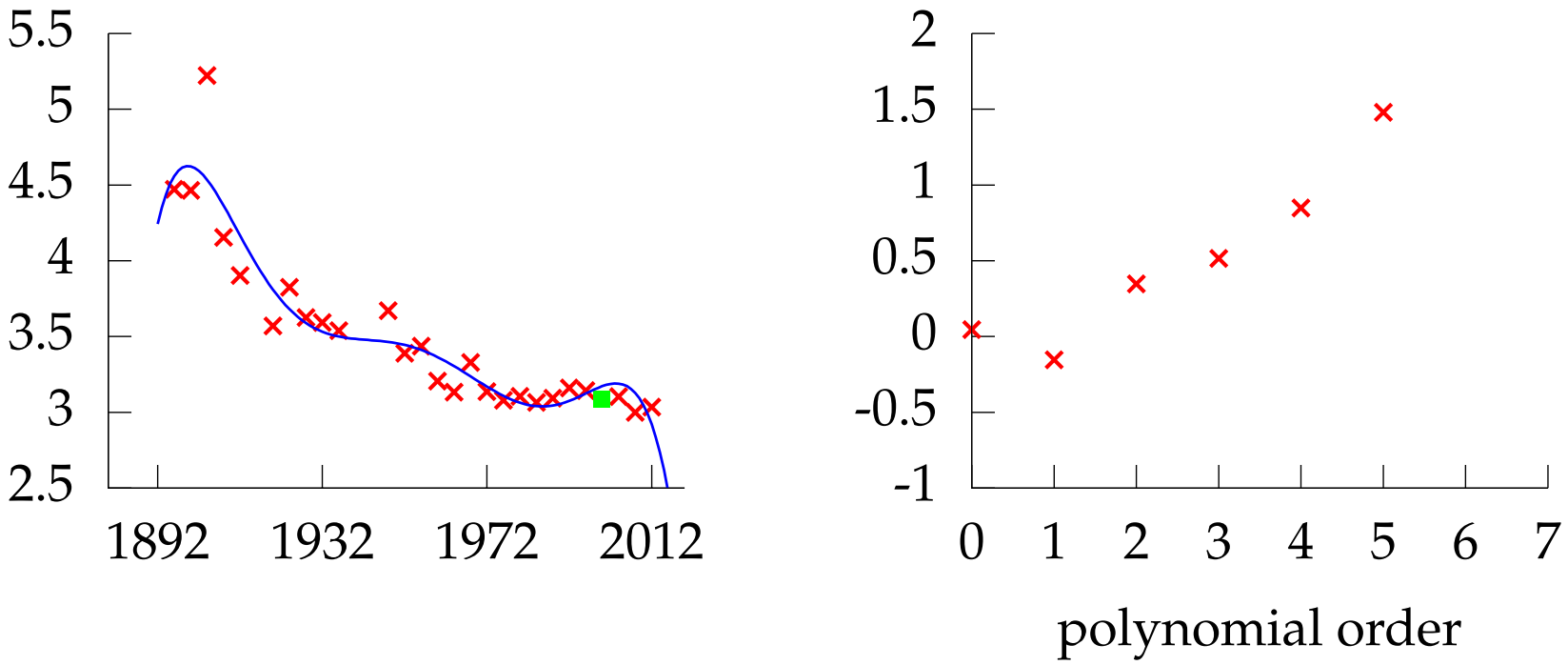
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



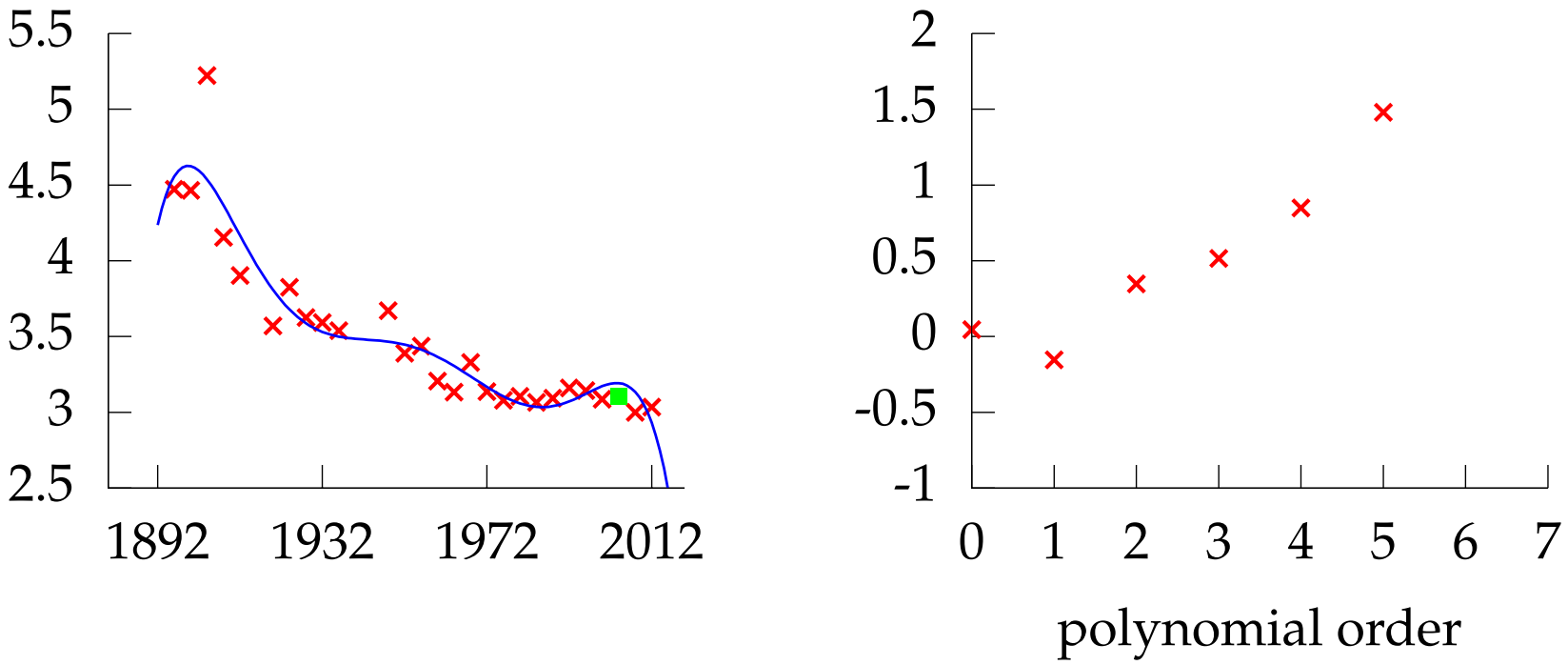
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



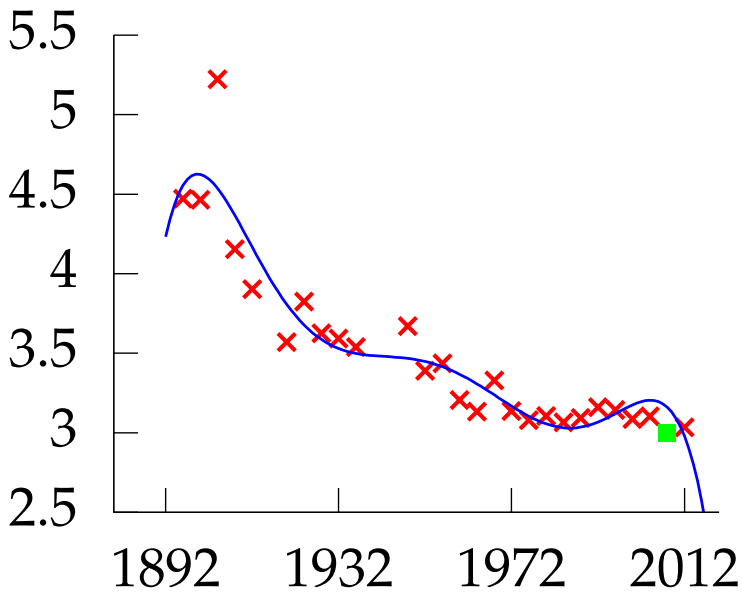
Polynomial order 6, training error -32.237, leave one out error 1.5047.

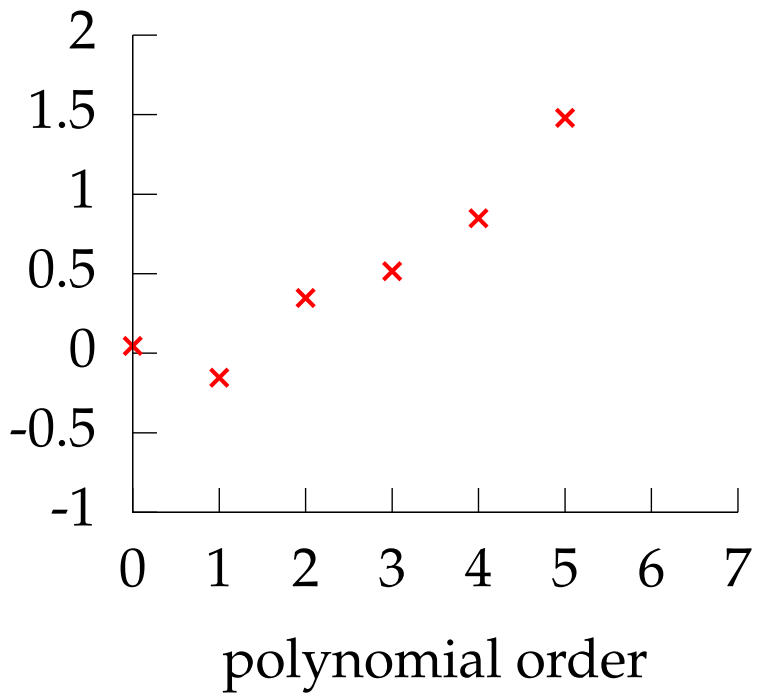
Leave One Out Error



Polynomial order 6, training error -32.237, leave one out error 1.5047.

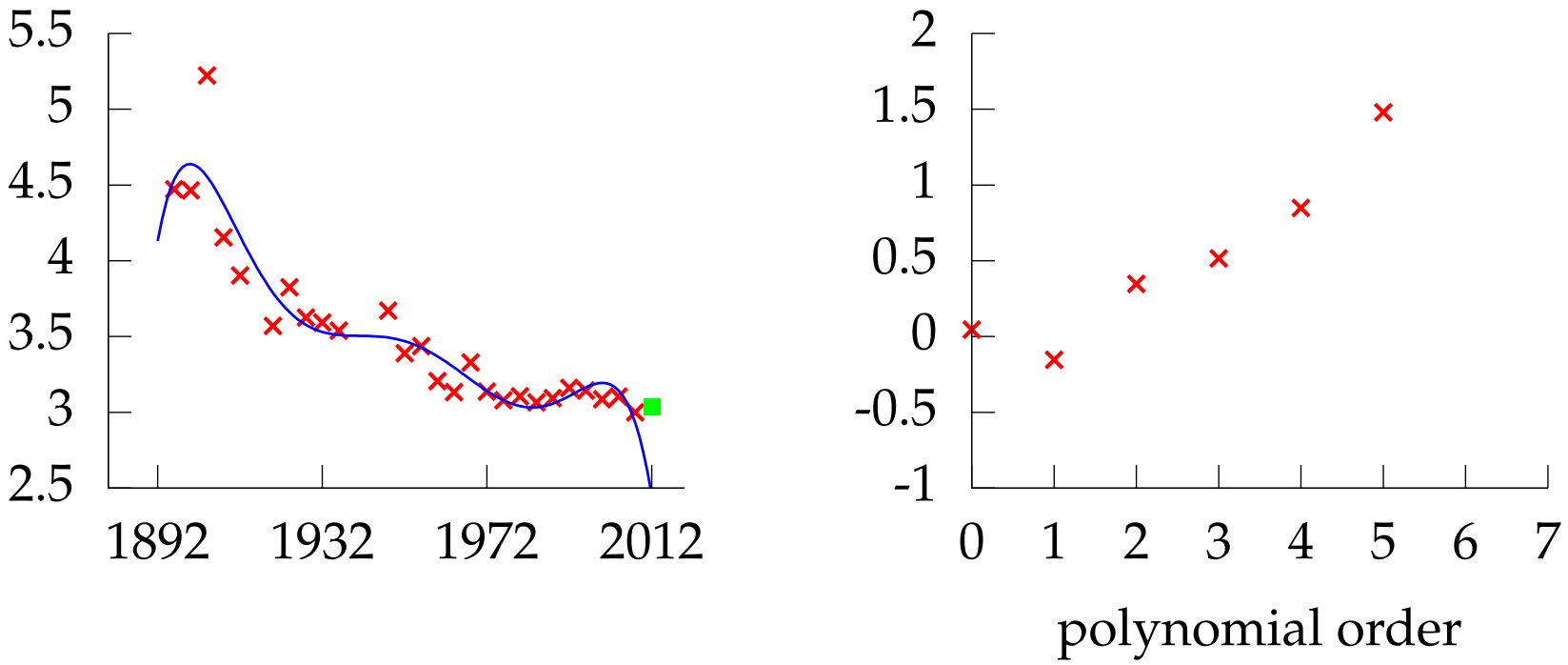
Leave One Out Error





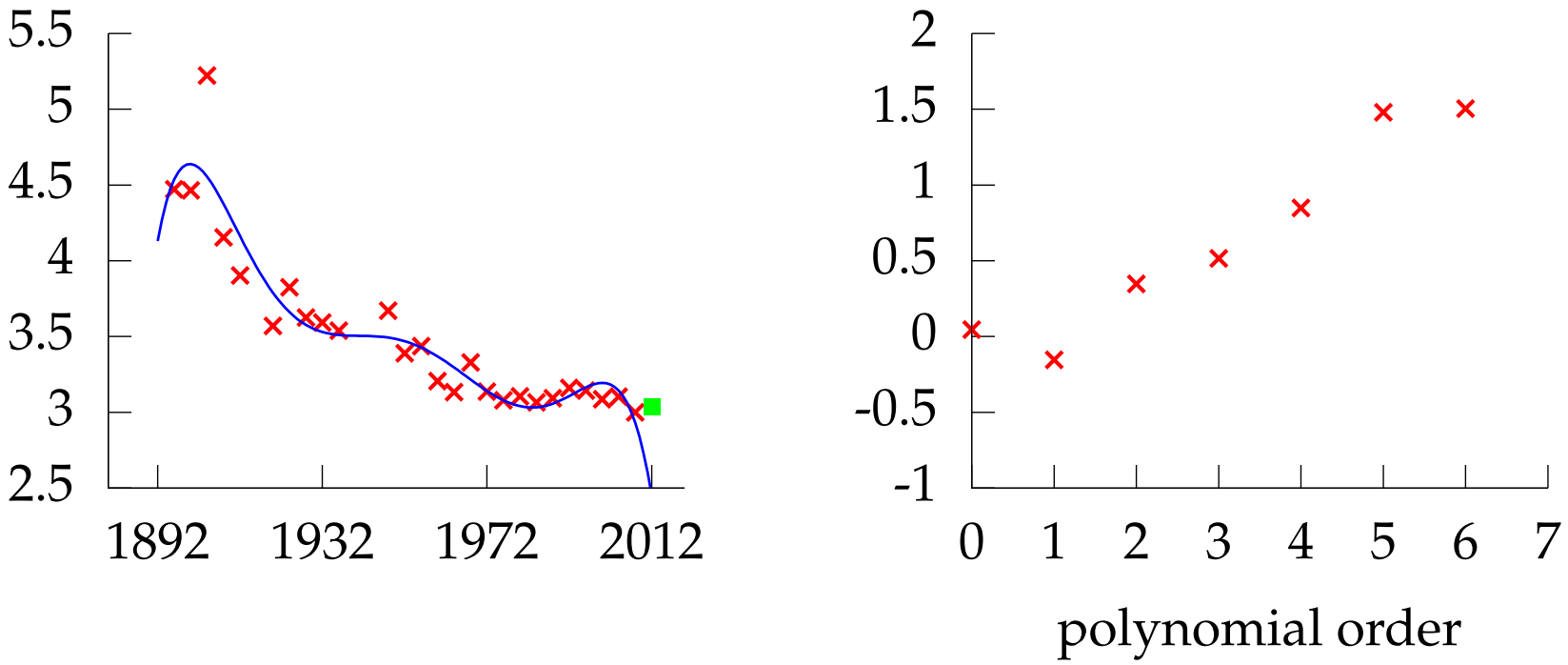
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error

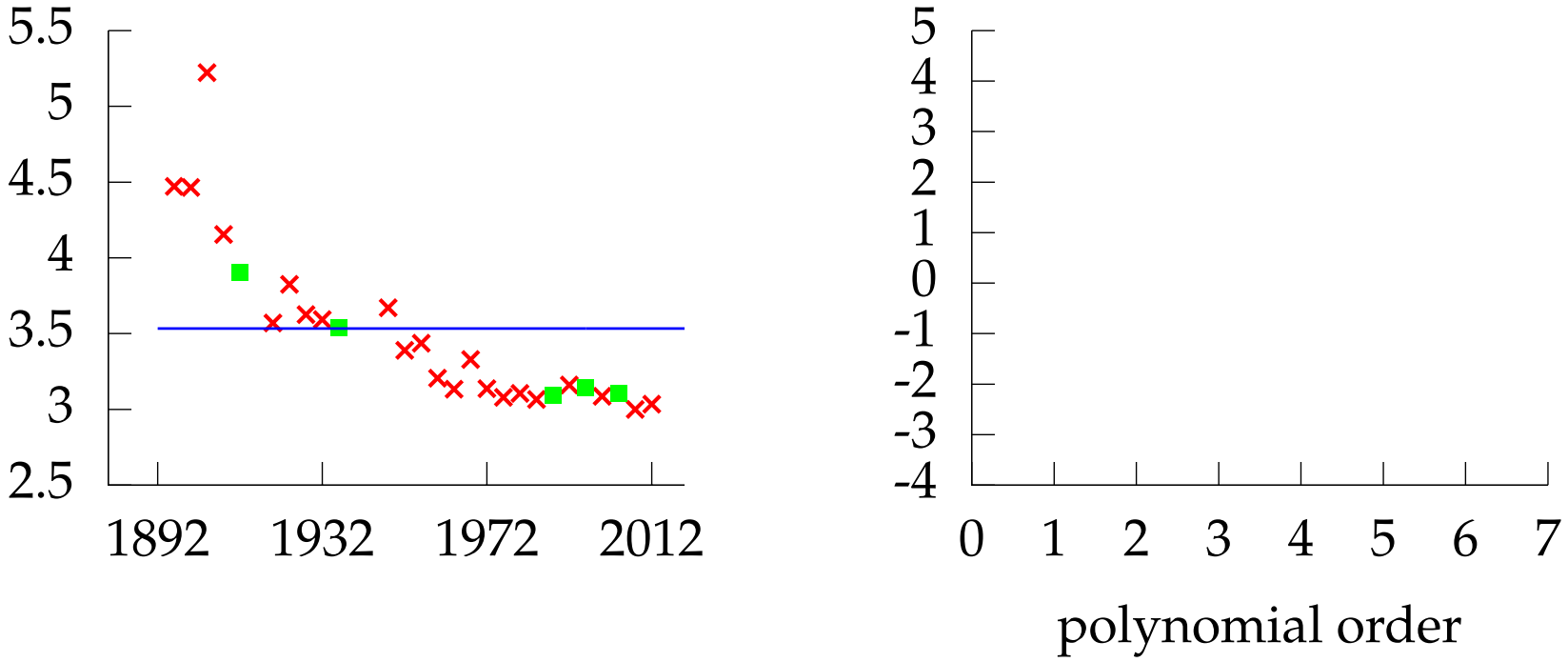


Polynomial order 6, training error -32.237, leave one out error 1.5047.

k Fold Cross Validation

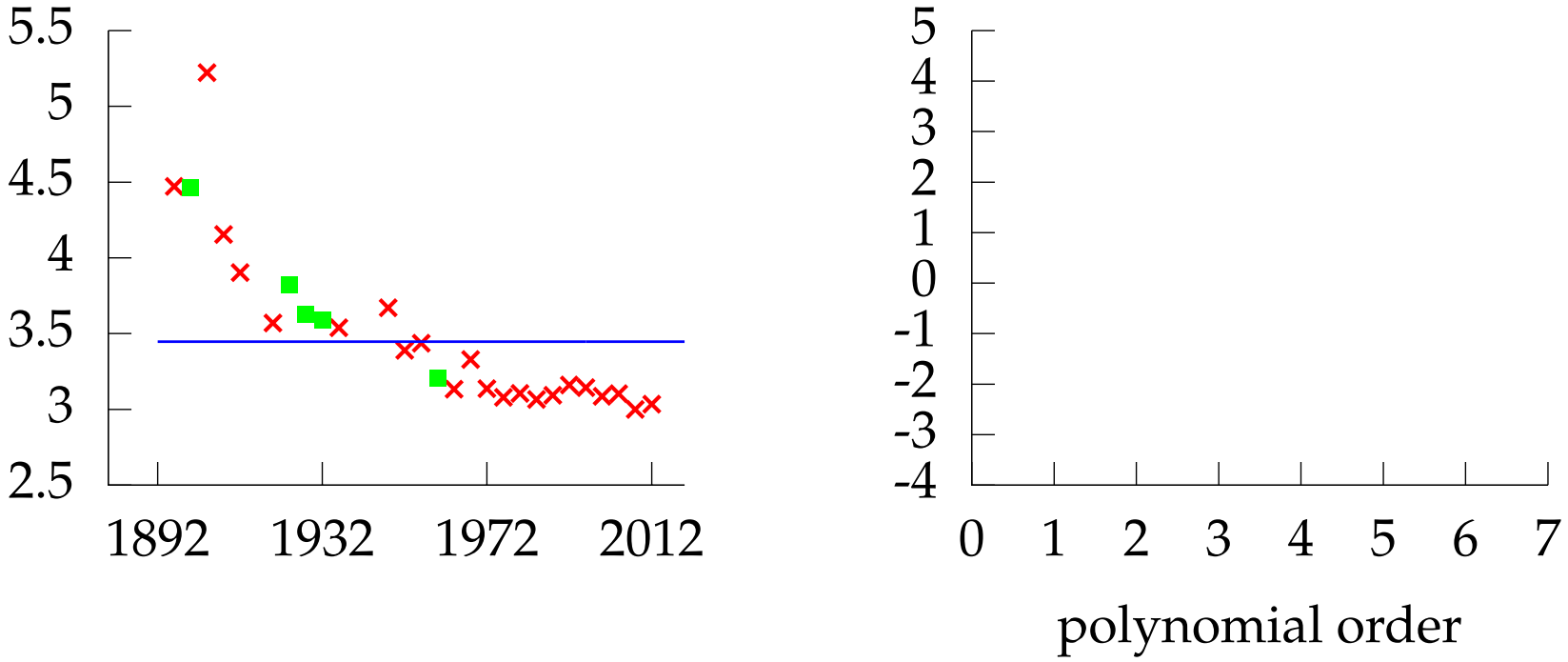
- Leave one out cross validation can be very time consuming!
- Need to train your algorithm n times.
- An alternative: k fold cross validation.

Cross Validation Error



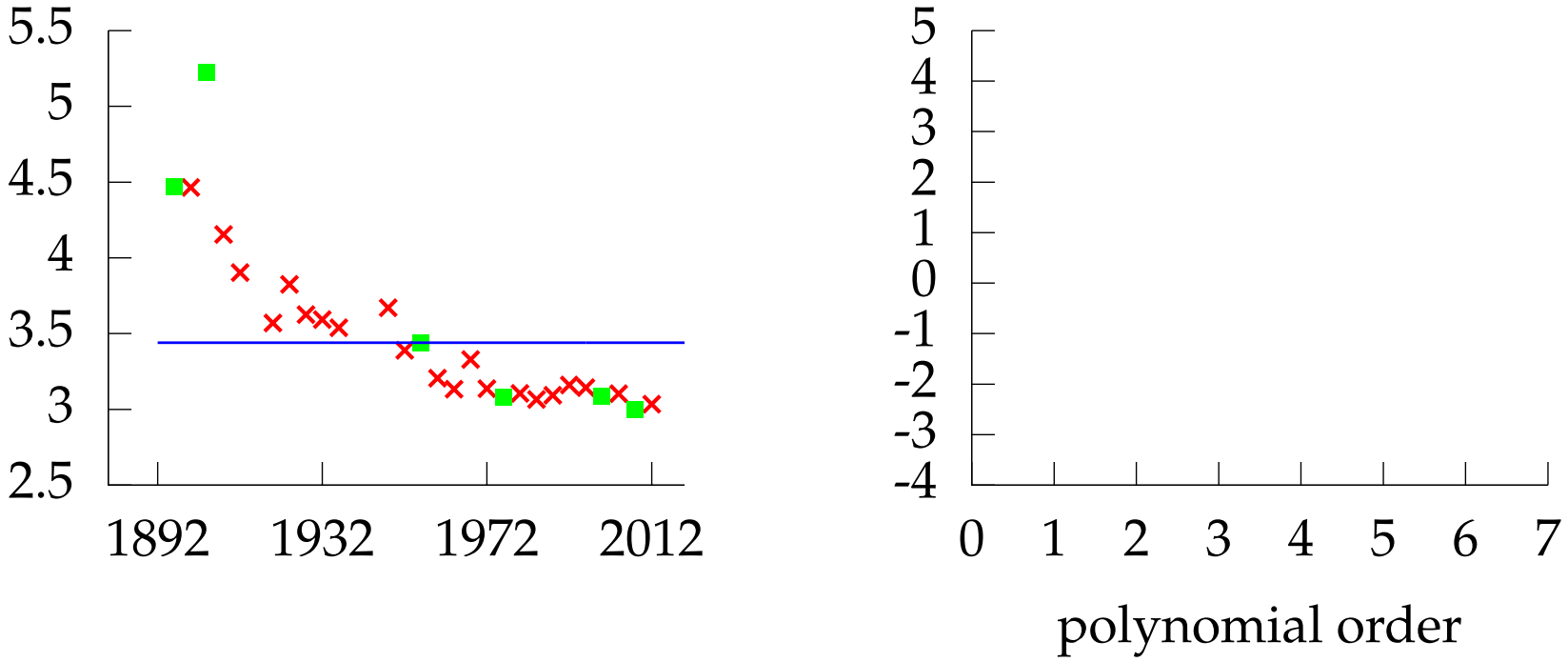
Polynomial order 0, training error -3.2644, leave one out error 0.045811.

Cross Validation Error



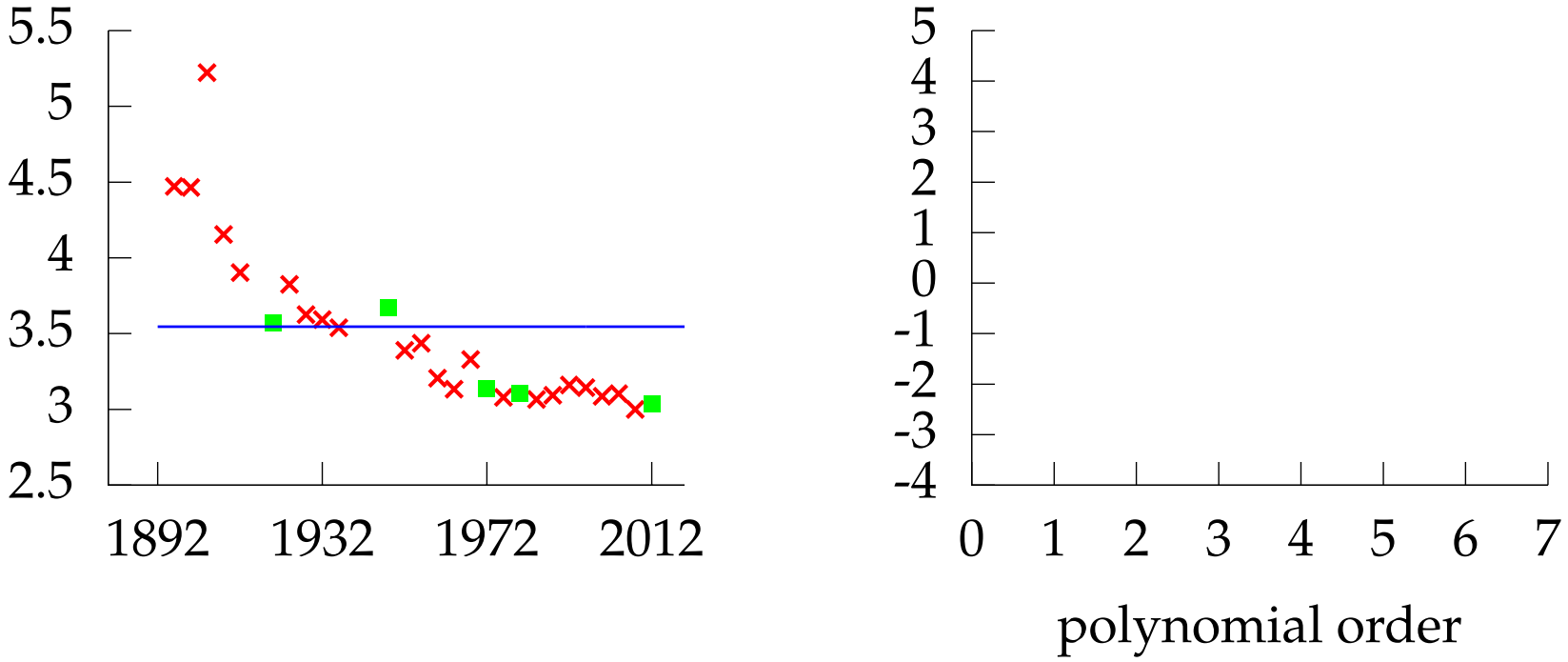
Polynomial order 0, training error -3.2644, leave one out error 0.045811.

Cross Validation Error



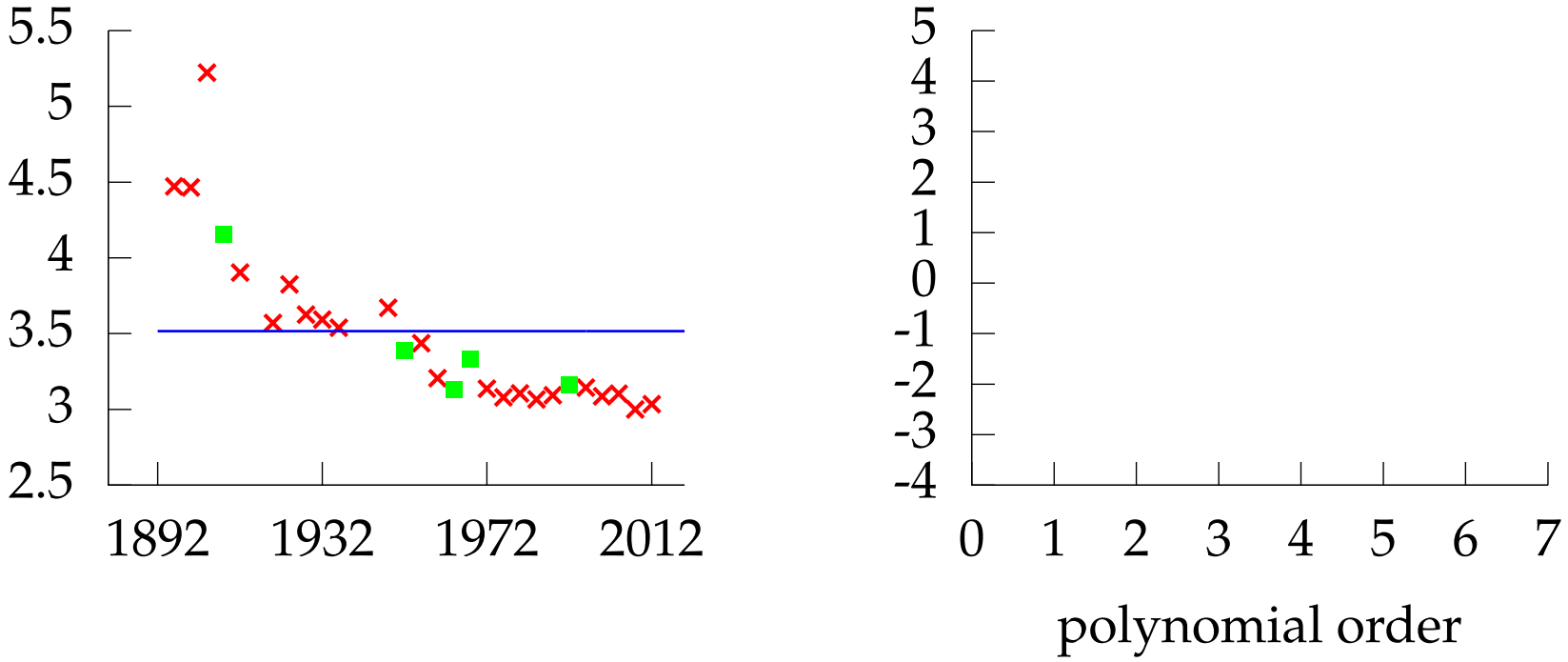
Polynomial order 0, training error -3.2644, leave one out error 0.045811.

Cross Validation Error



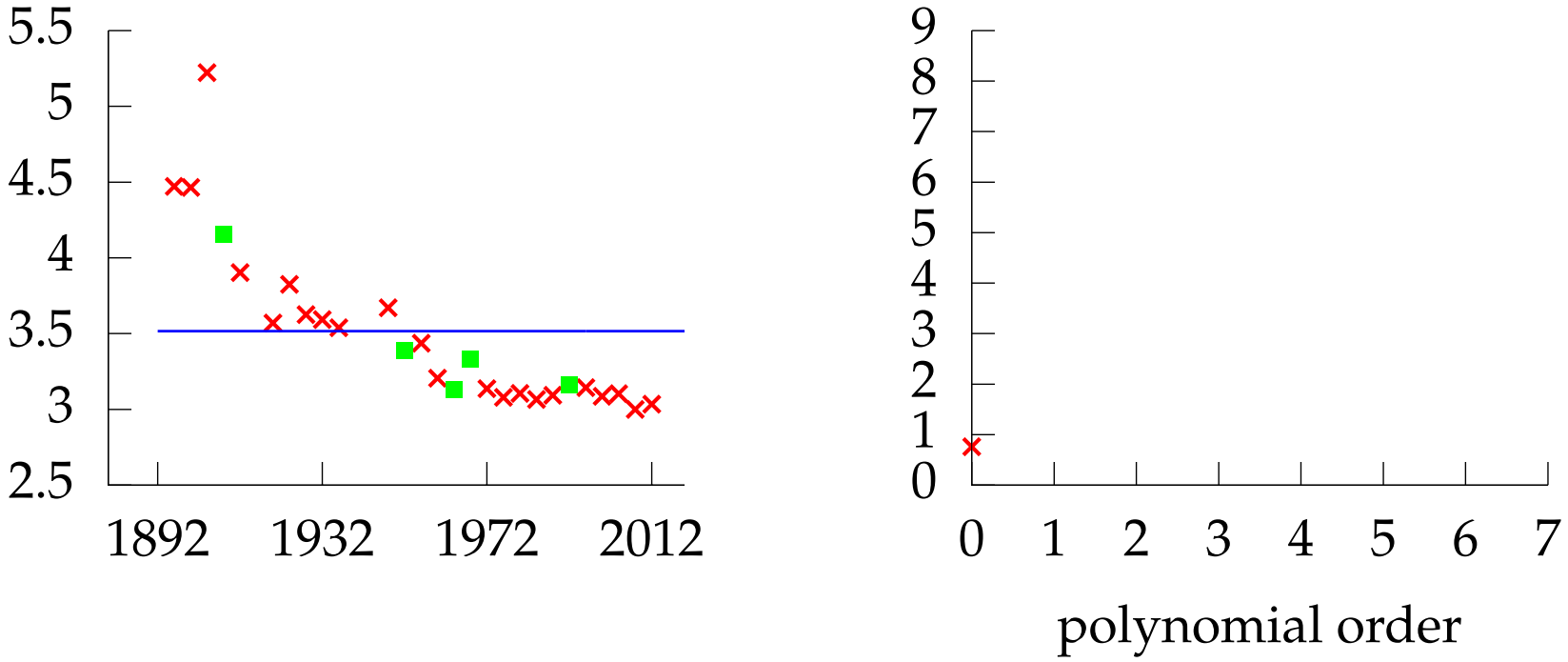
Polynomial order 0, training error -3.2644, leave one out error 0.045811.

Cross Validation Error



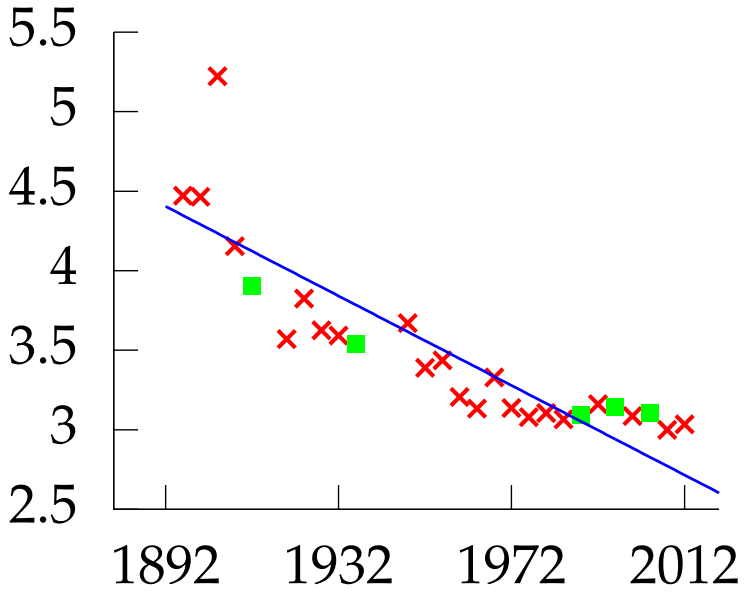
Polynomial order 0, training error -3.2644, leave one out error 0.045811.

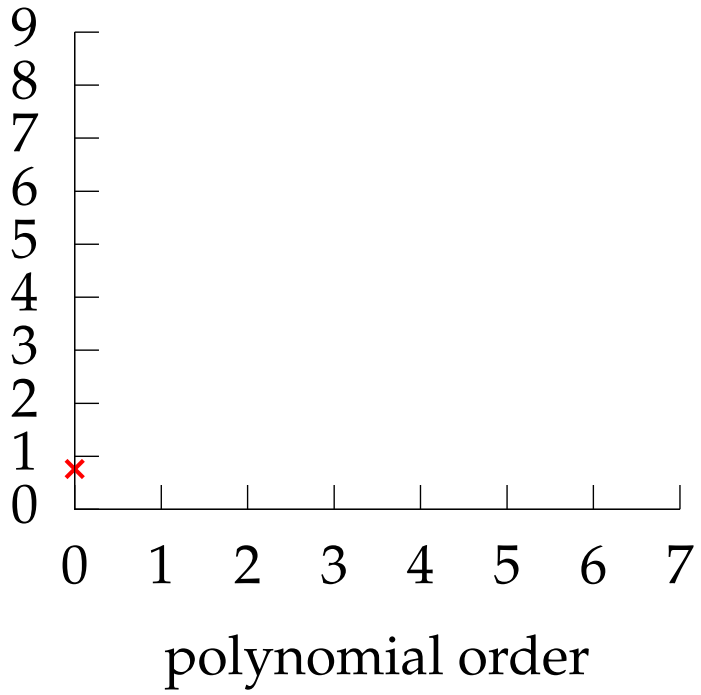
Cross Validation Error



Polynomial order 0, training error -3.2644, leave one out error 0.045811.

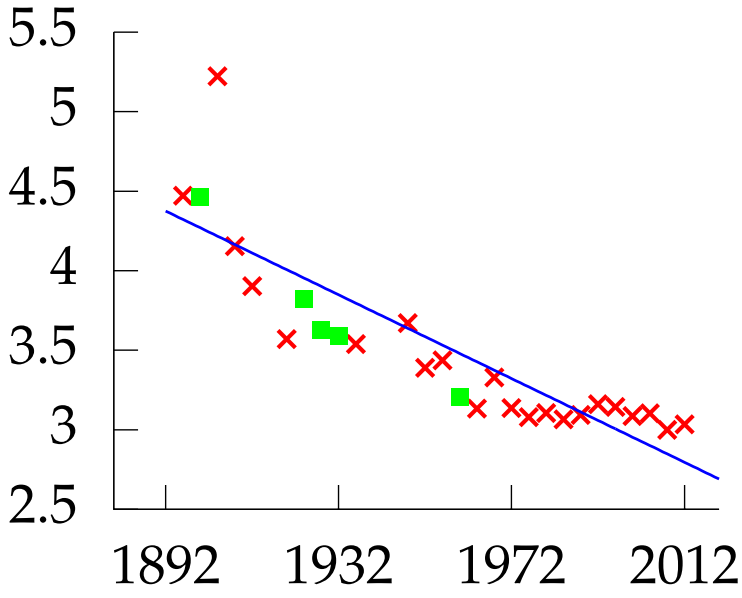
Cross Validation Error

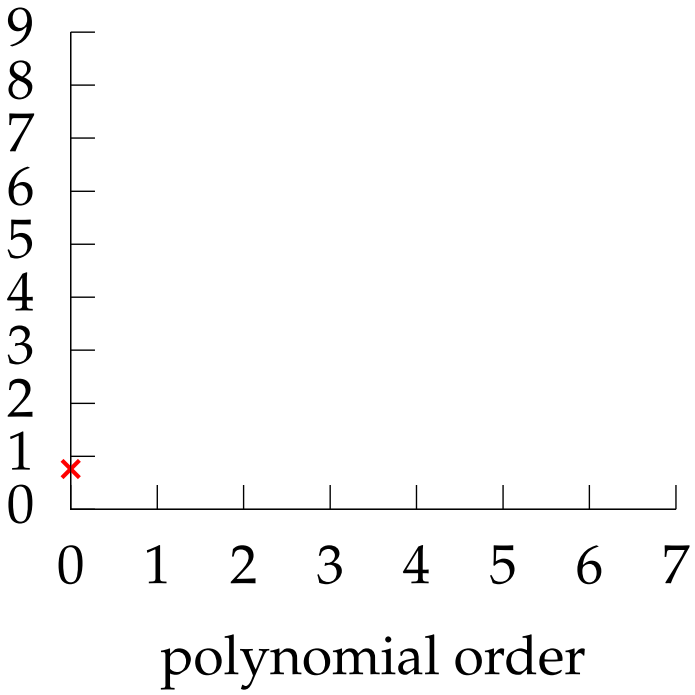




Polynomial order 1, training error -18.873, leave one out error -0.15413.

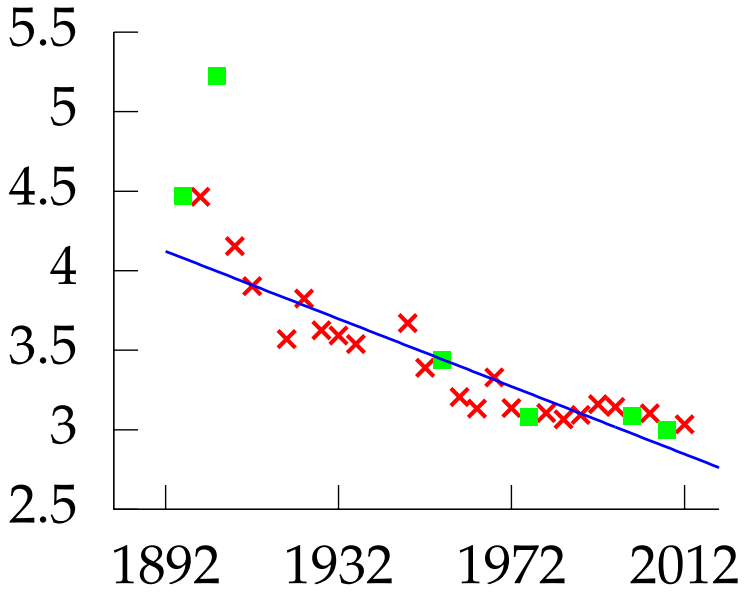
Cross Validation Error

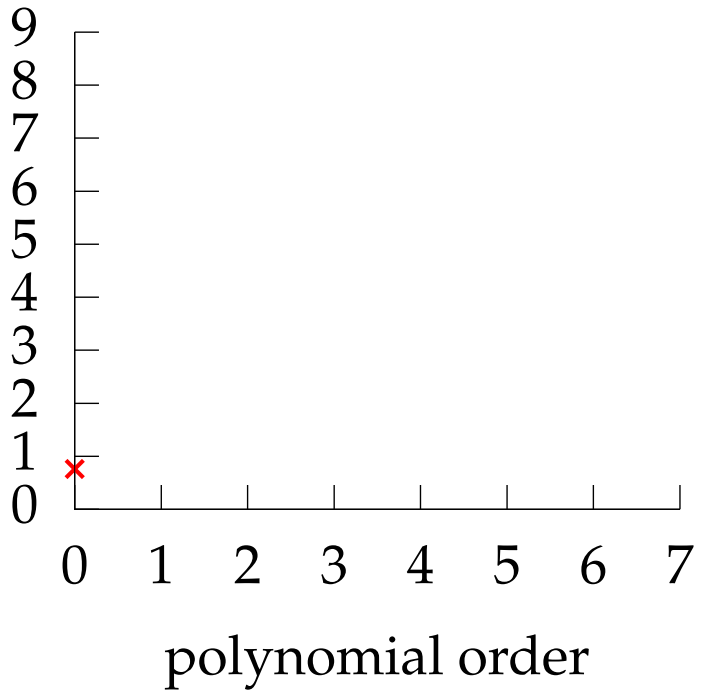




Polynomial order 1, training error -18.873, leave one out error -0.15413.

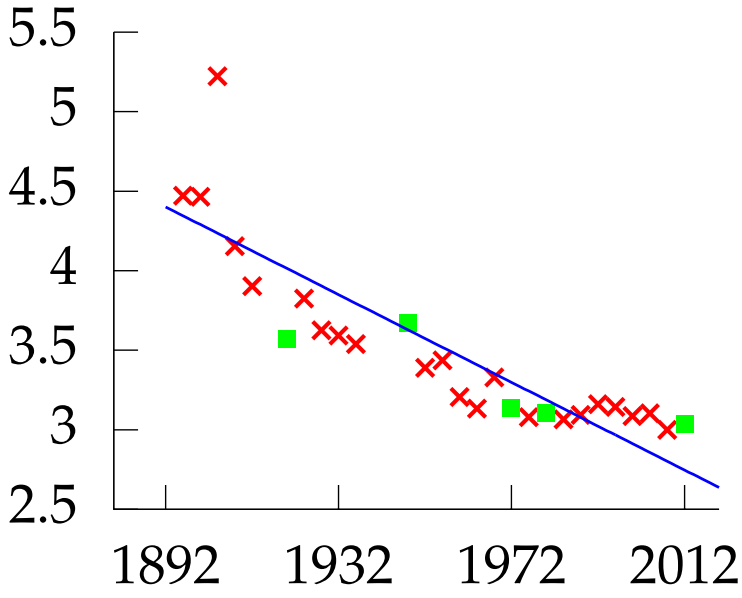
Cross Validation Error

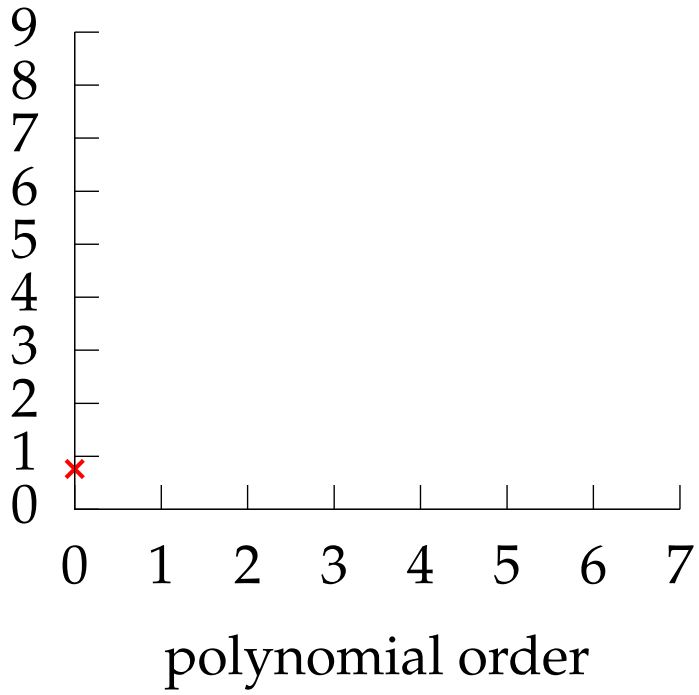




Polynomial order 1, training error -18.873, leave one out error -0.15413.

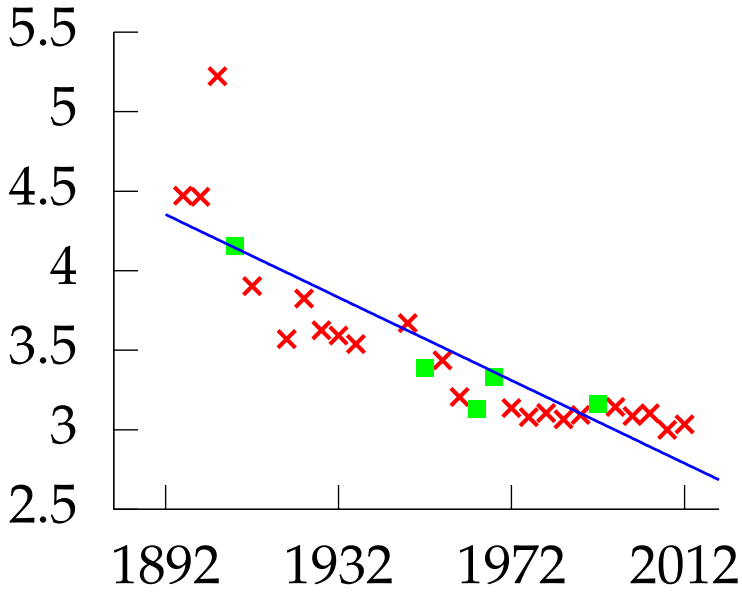
Cross Validation Error

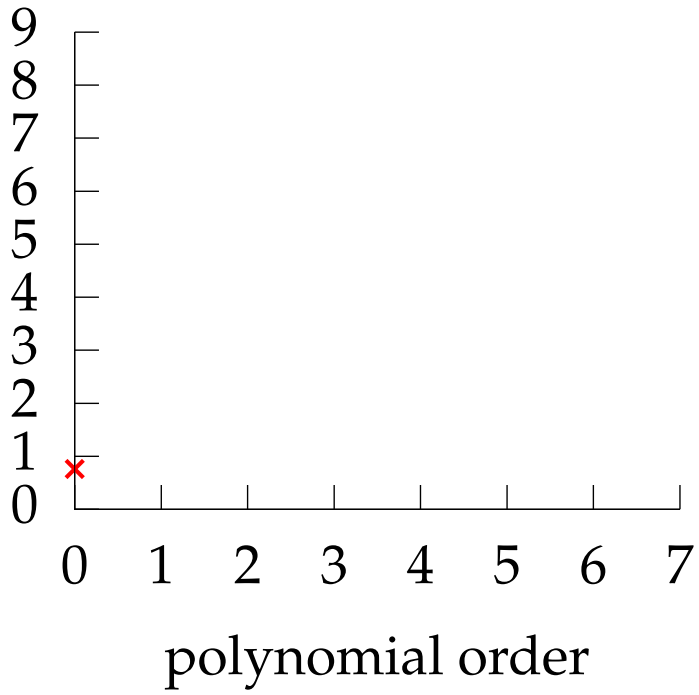




Polynomial order 1, training error -18.873, leave one out error -0.15413.

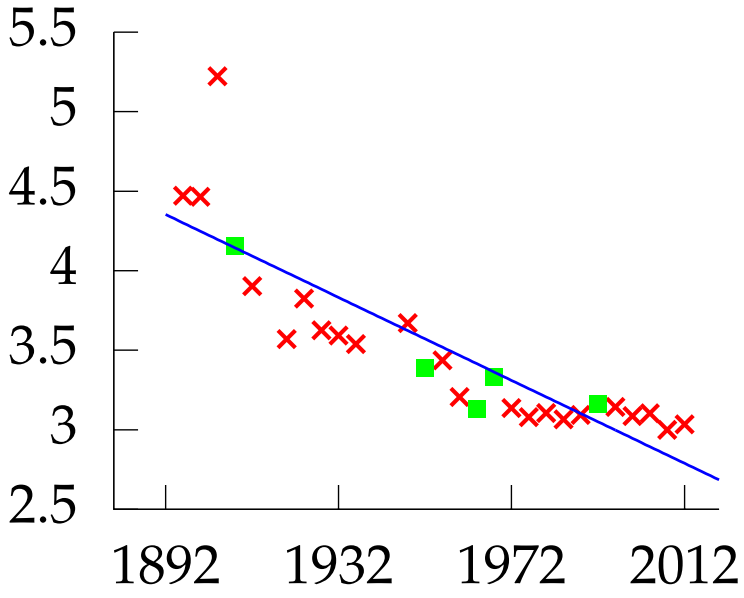
Cross Validation Error

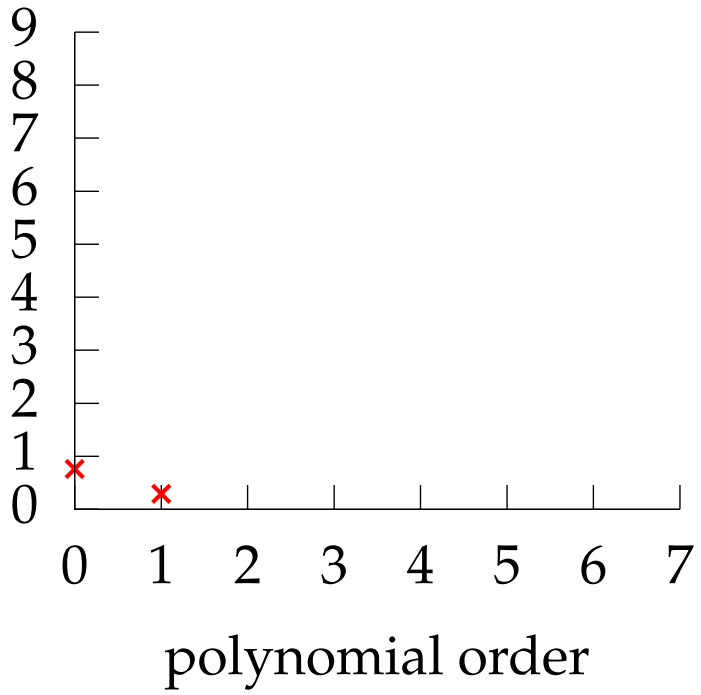




Polynomial order 1, training error -18.873, leave one out error -0.15413.

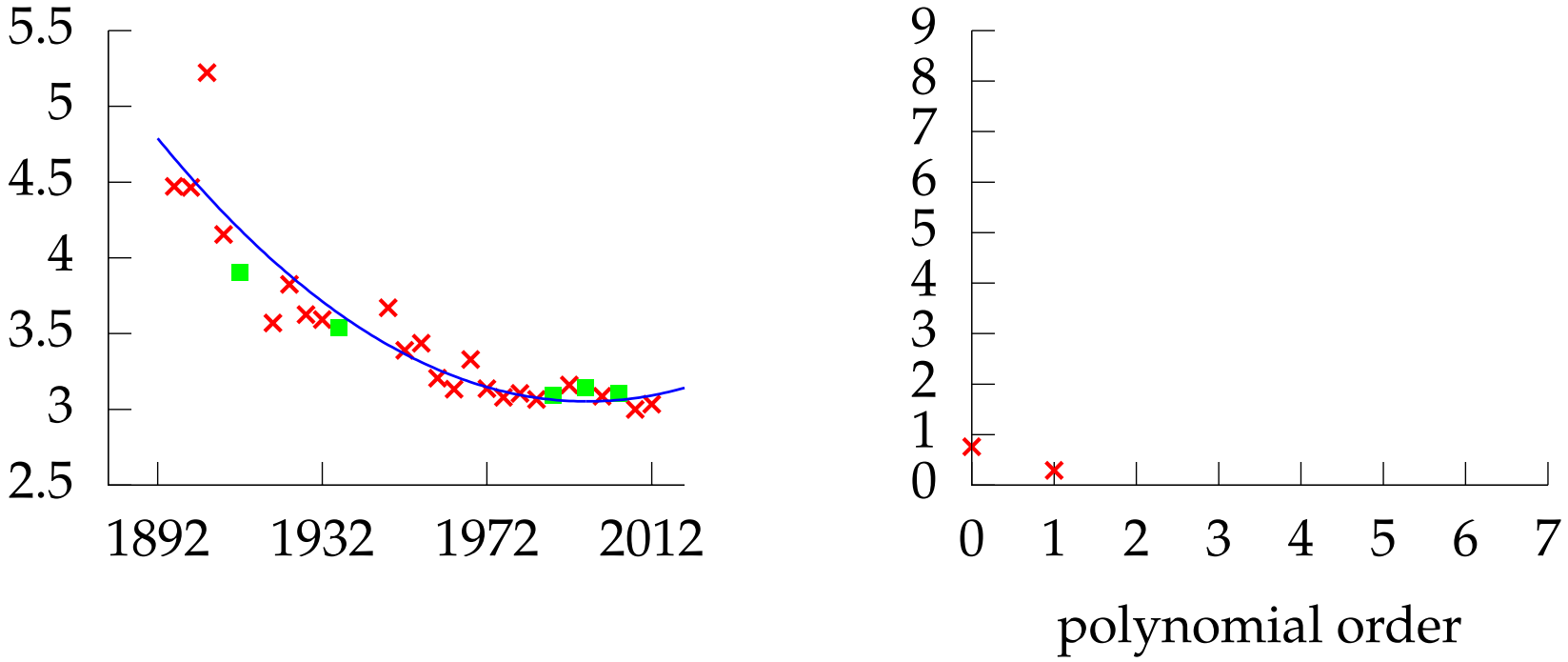
Cross Validation Error





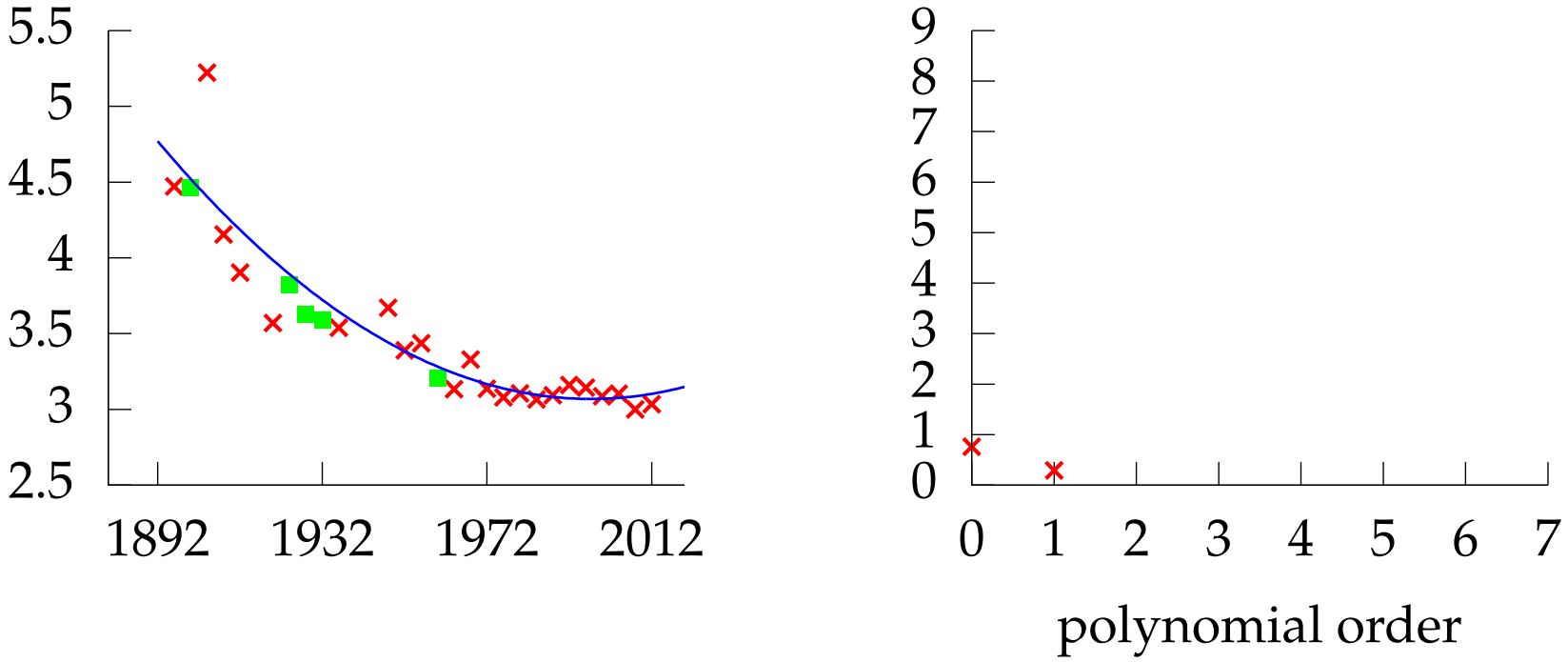
Polynomial order 1, training error -18.873, leave one out error -0.15413.

Cross Validation Error



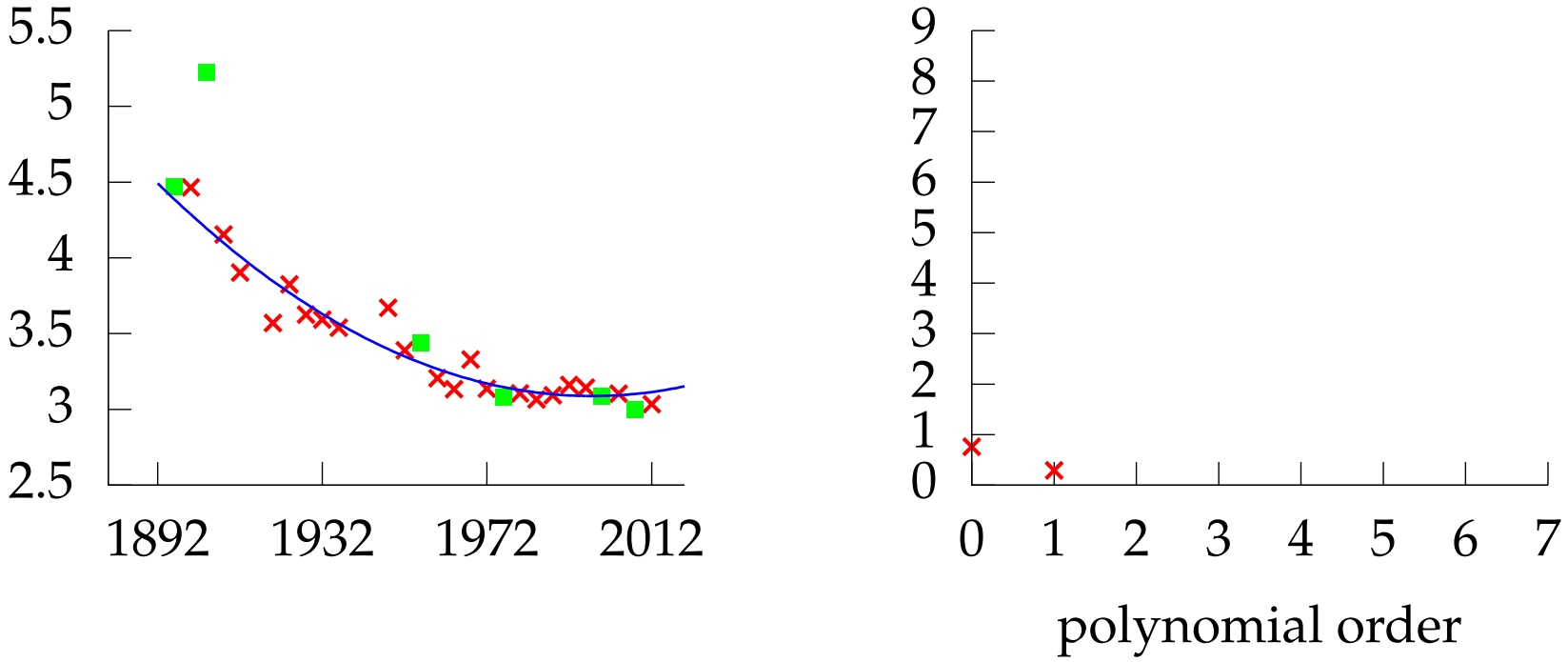
Polynomial order 2, training error -25.177, leave one out error 0.34669.

Cross Validation Error



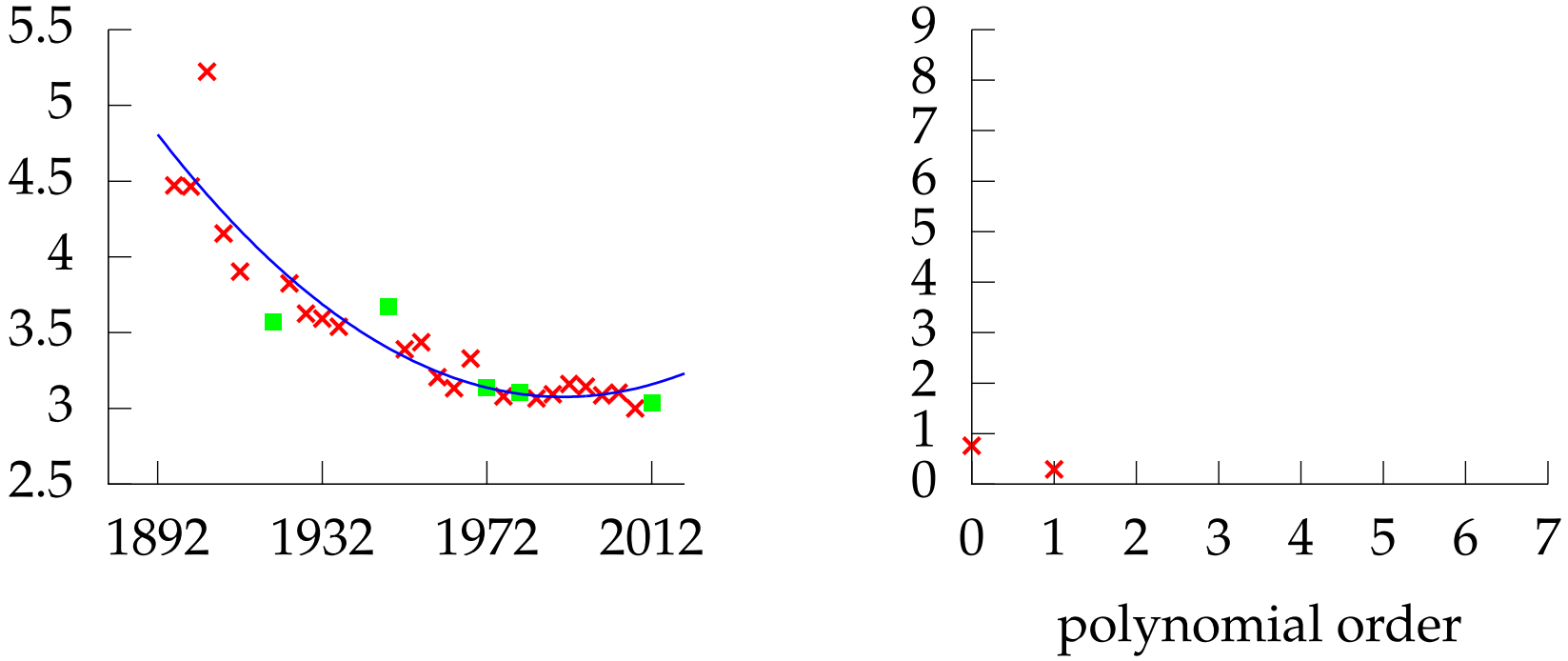
Polynomial order 2, training error -25.177, leave one out error 0.34669.

Cross Validation Error



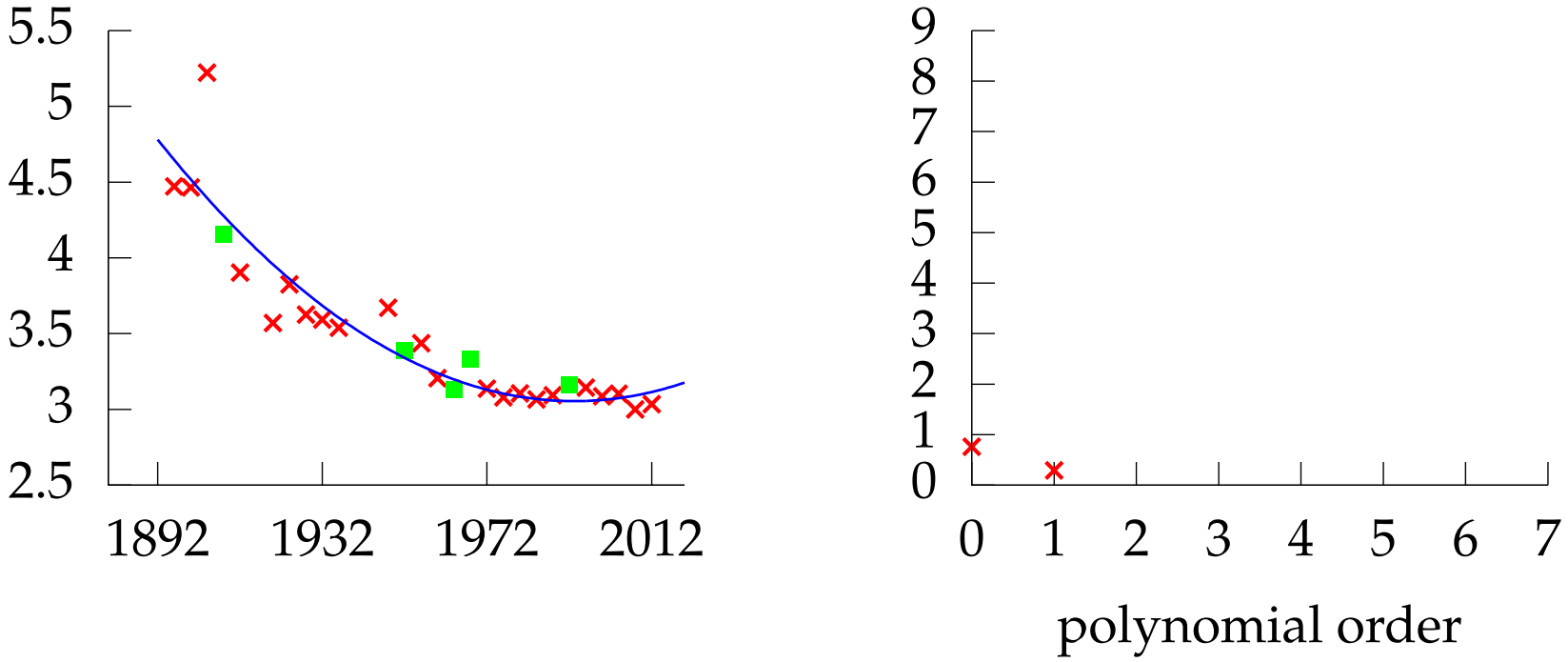
Polynomial order 2, training error -25.177, leave one out error 0.34669.

Cross Validation Error



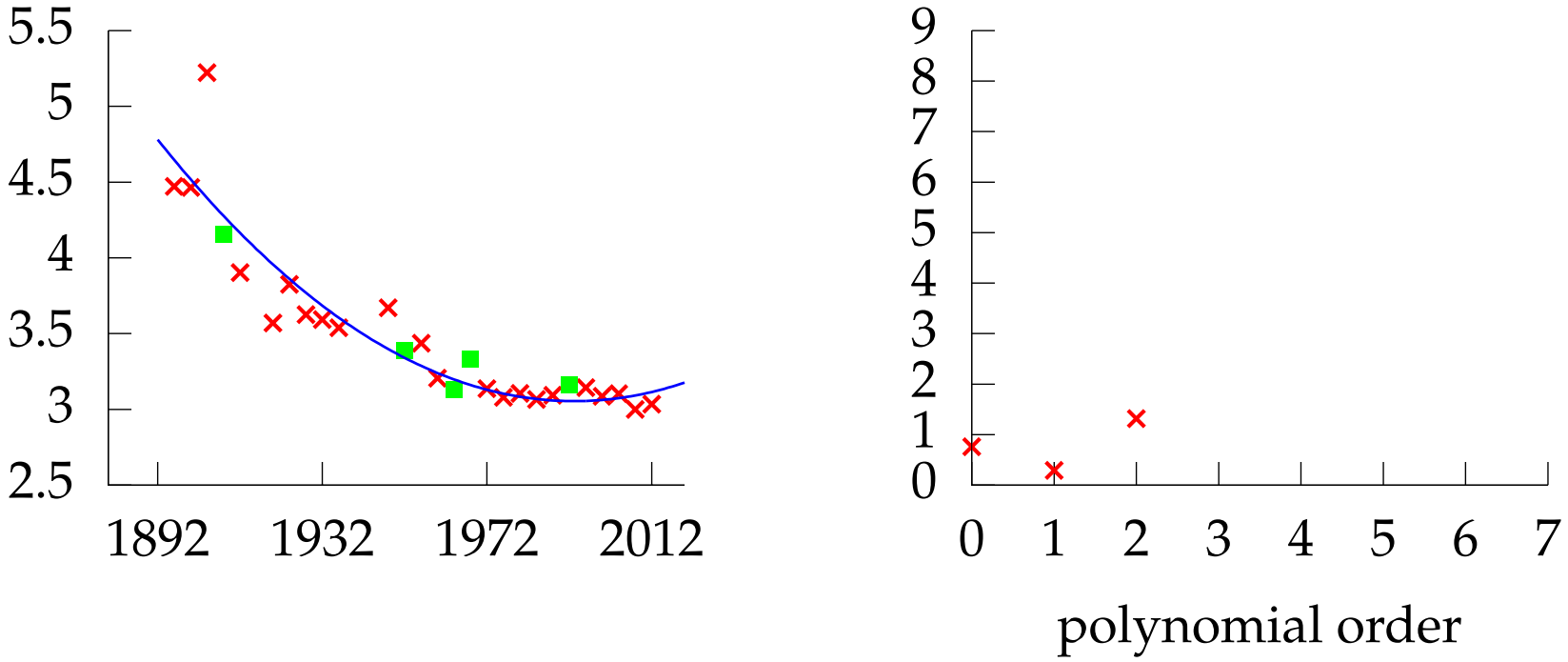
Polynomial order 2, training error -25.177, leave one out error 0.34669.

Cross Validation Error



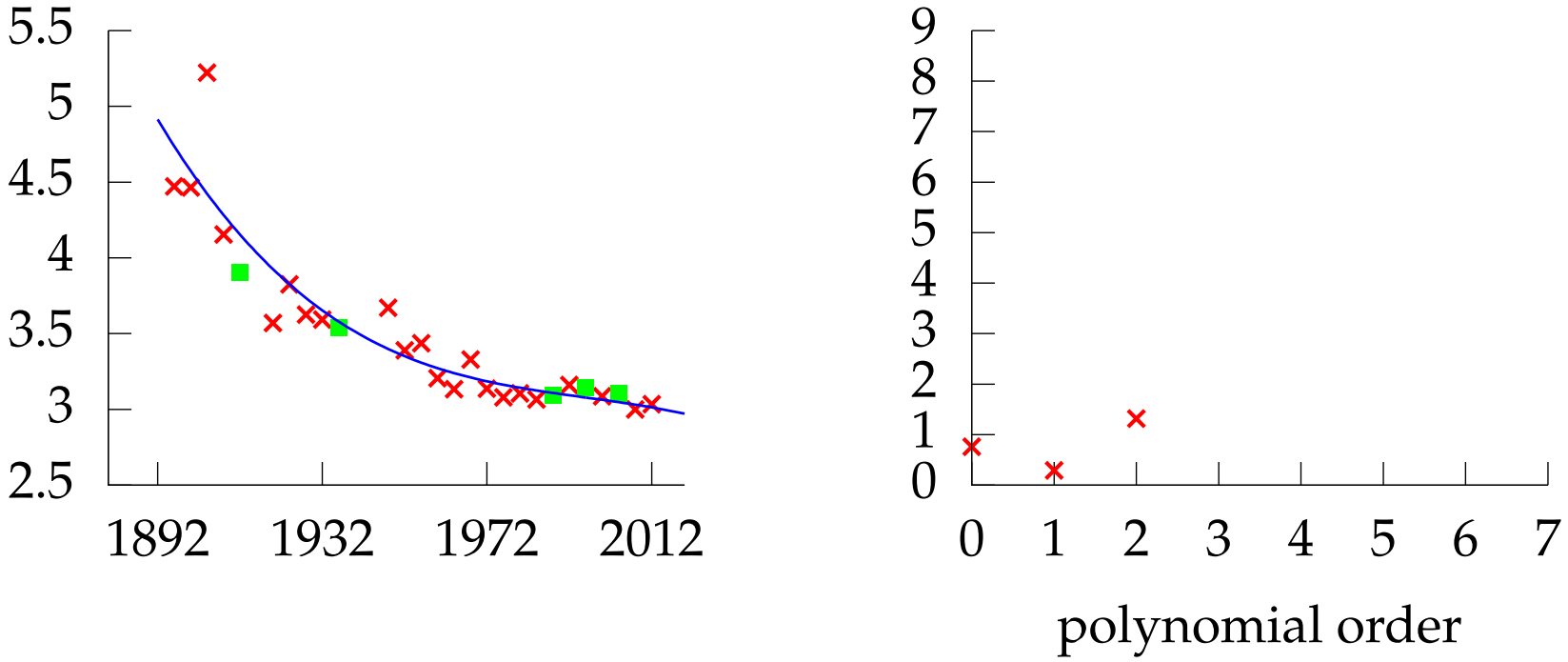
Polynomial order 2, training error -25.177, leave one out error 0.34669.

Cross Validation Error



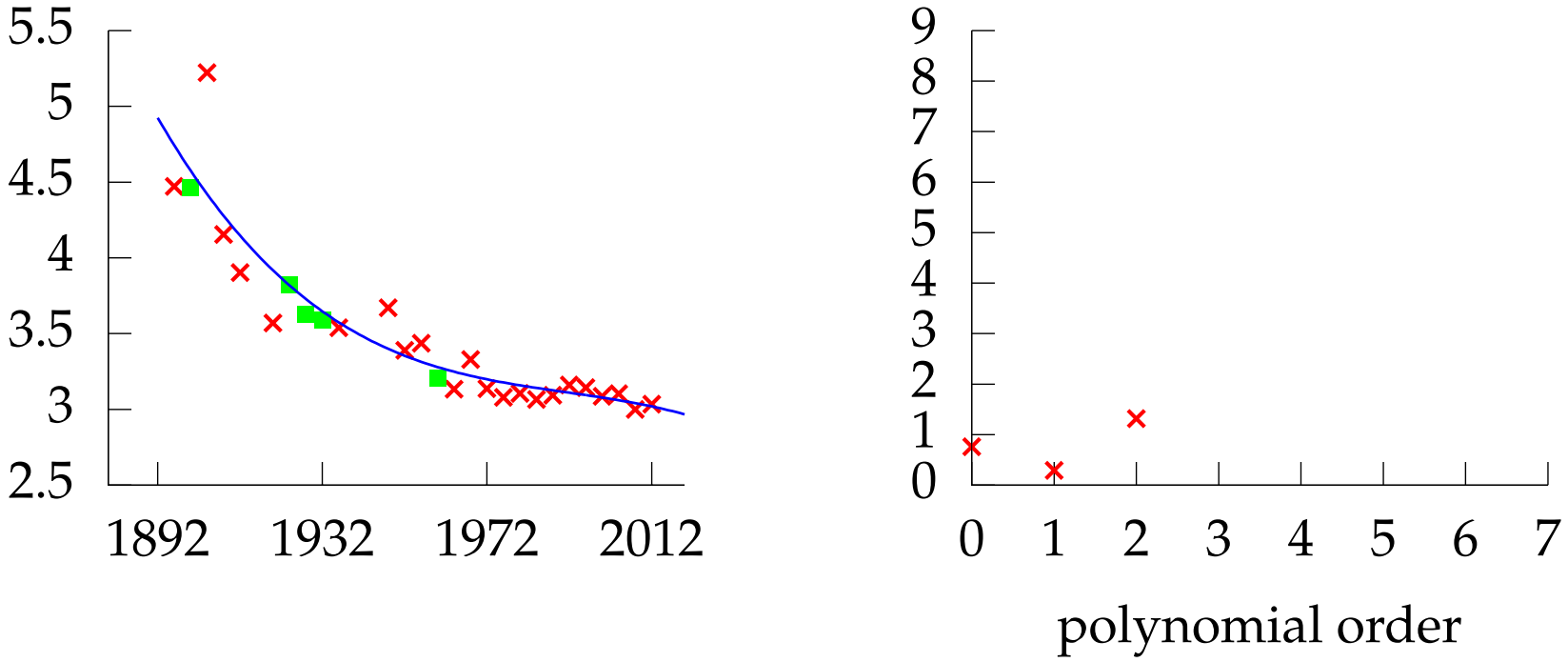
Polynomial order 2, training error -25.177, leave one out error 0.34669.

Cross Validation Error



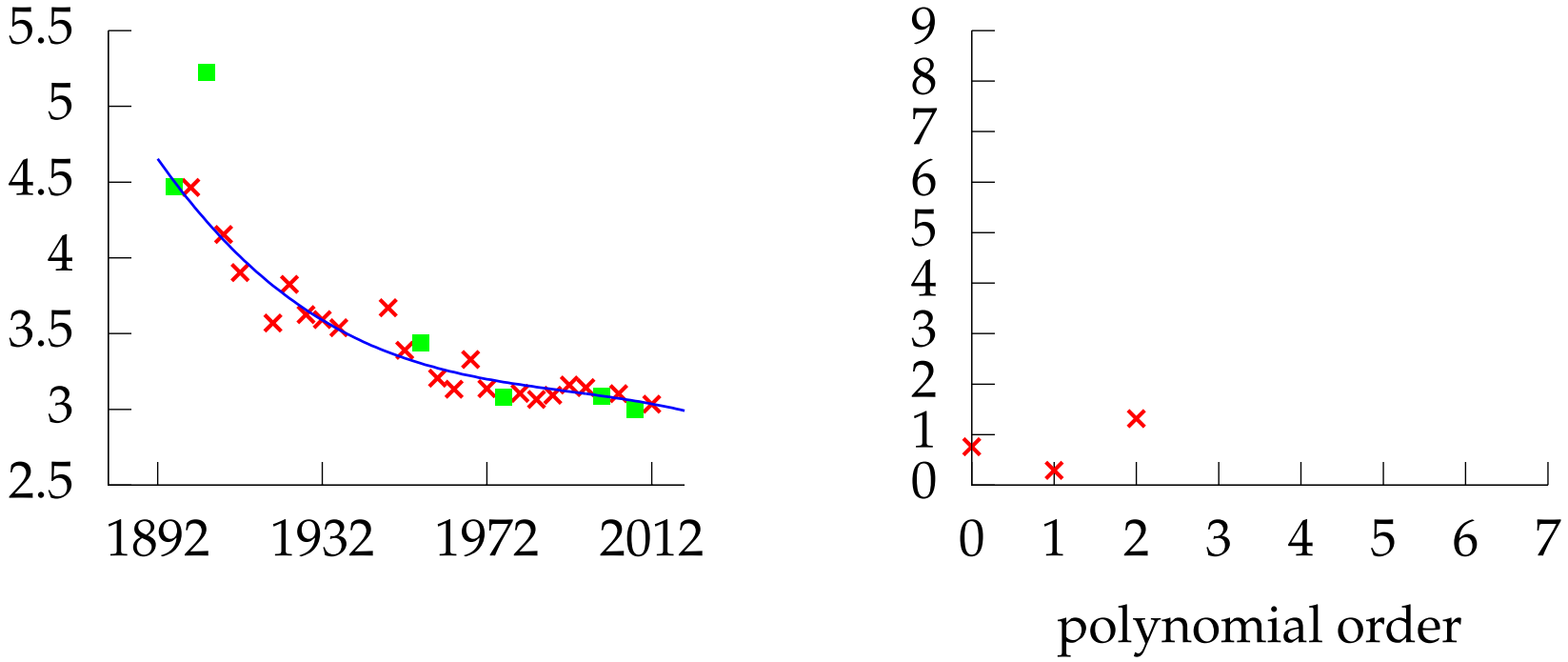
Polynomial order 3, training error -25.777, leave one out error 0.51621.

Cross Validation Error



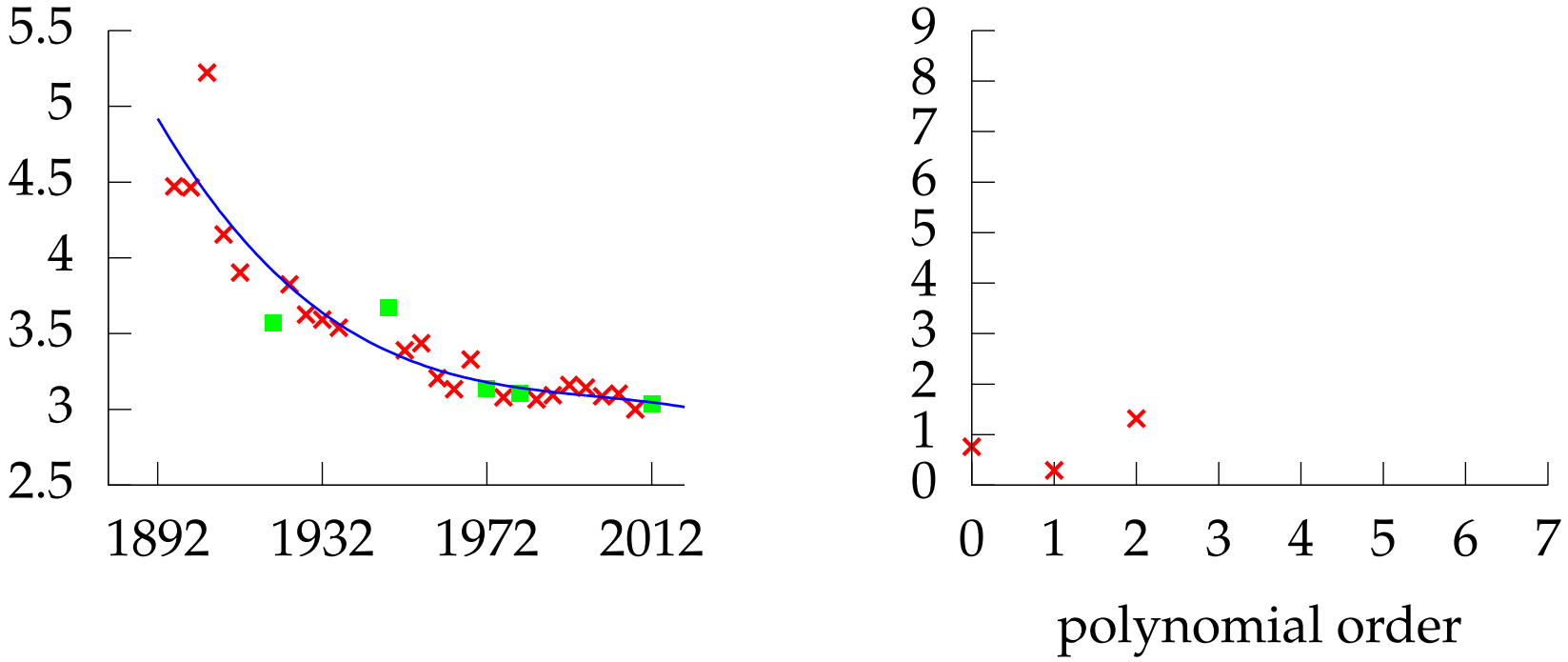
Polynomial order 3, training error -25.777, leave one out error 0.51621.

Cross Validation Error



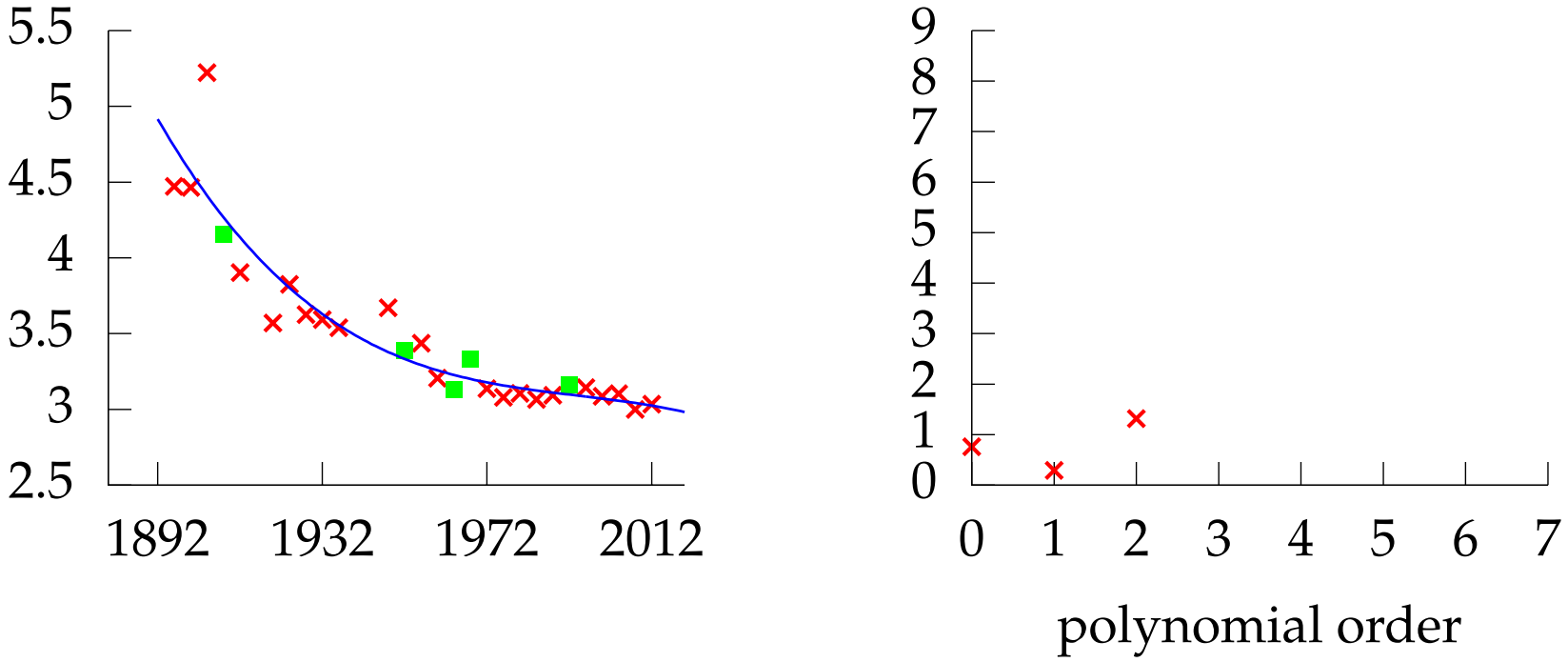
Polynomial order 3, training error -25.777, leave one out error 0.51621.

Cross Validation Error



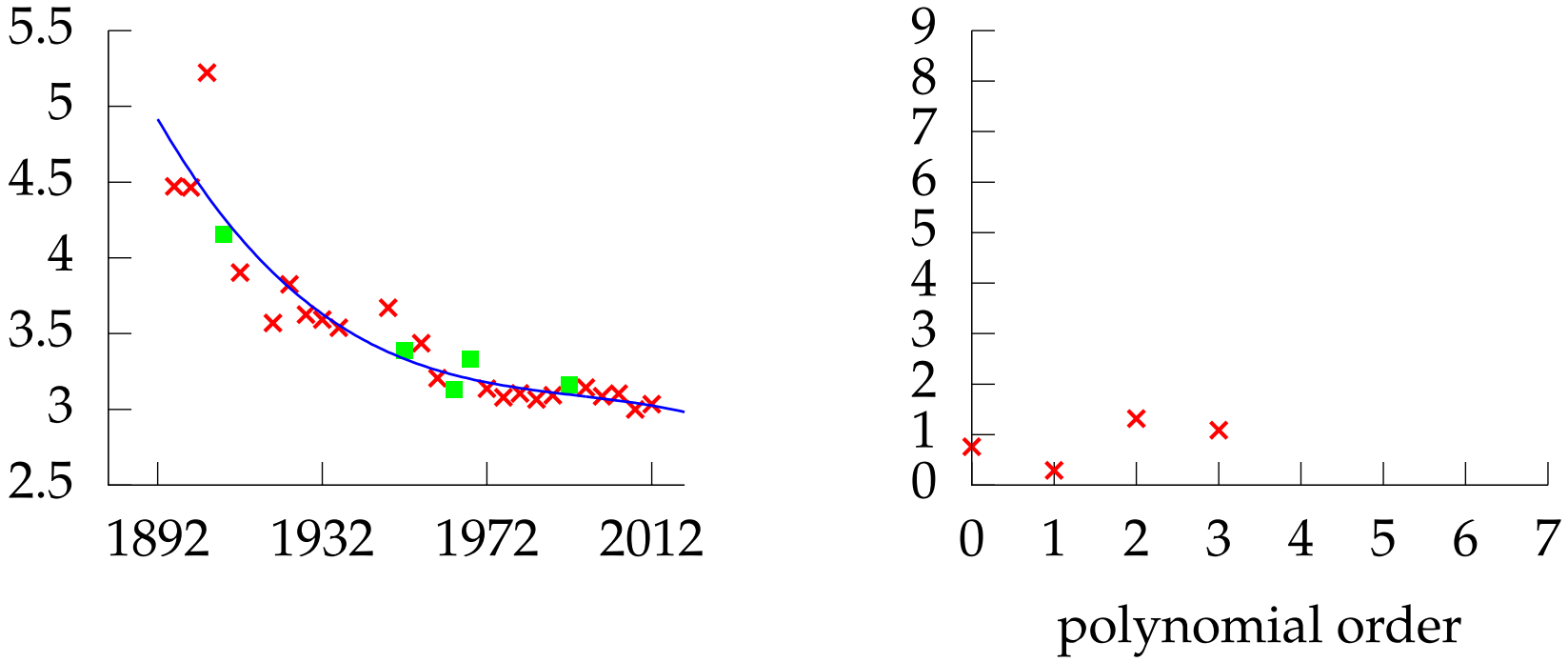
Polynomial order 3, training error -25.777, leave one out error 0.51621.

Cross Validation Error



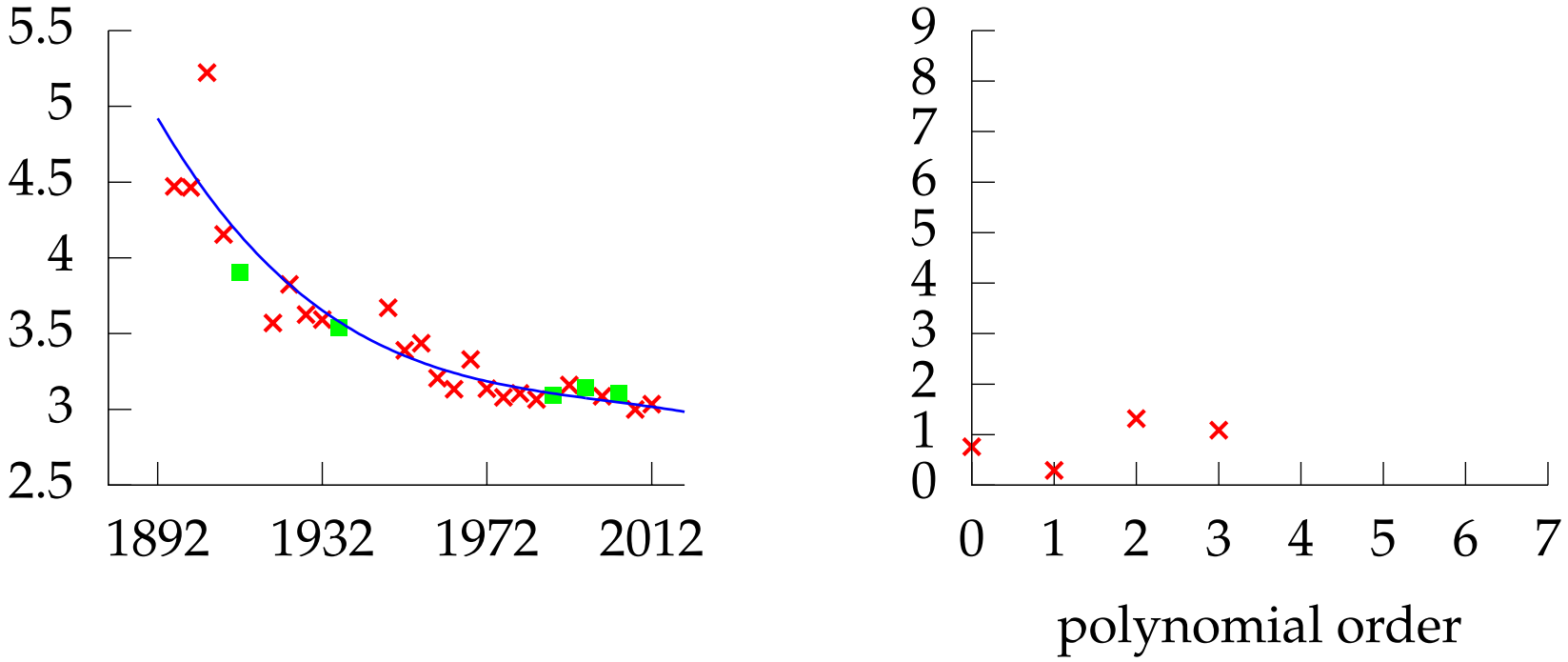
Polynomial order 3, training error -25.777, leave one out error 0.51621.

Cross Validation Error



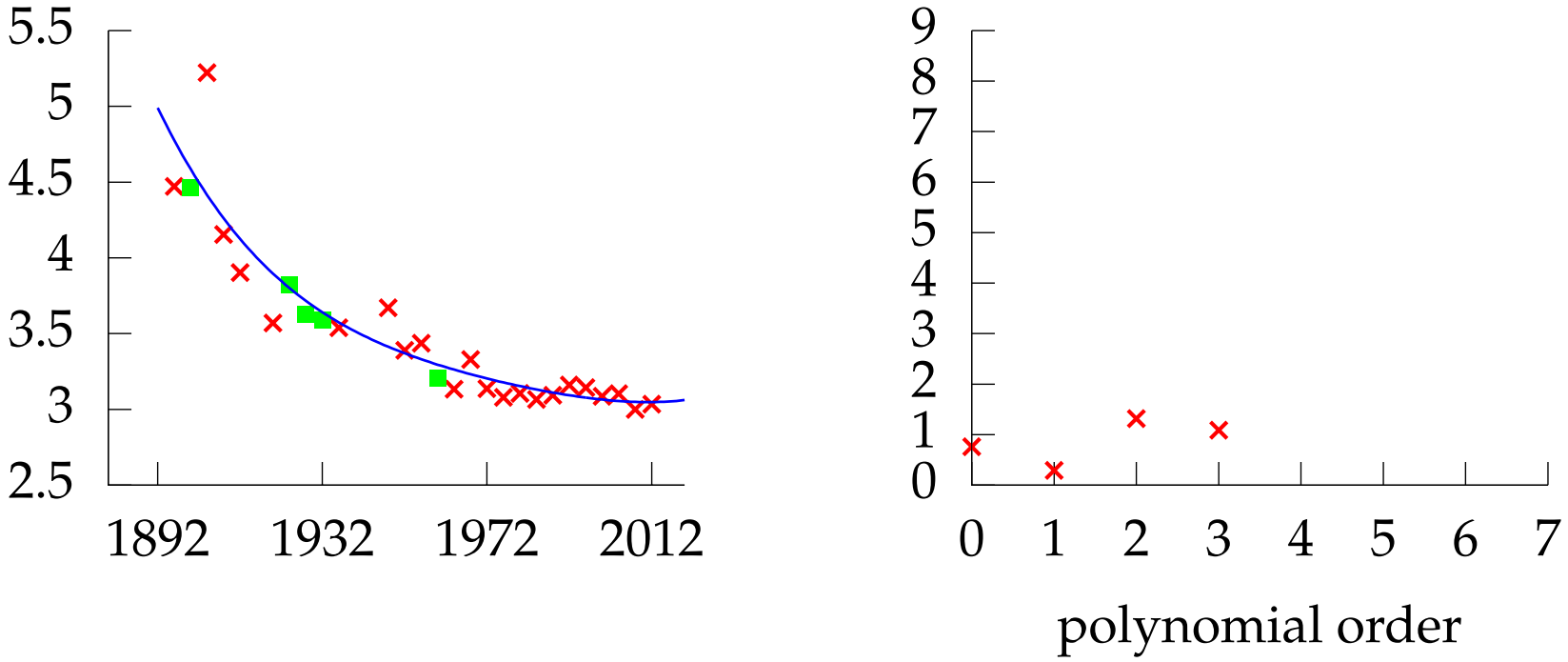
Polynomial order 3, training error -25.777, leave one out error 0.51621.

Cross Validation Error



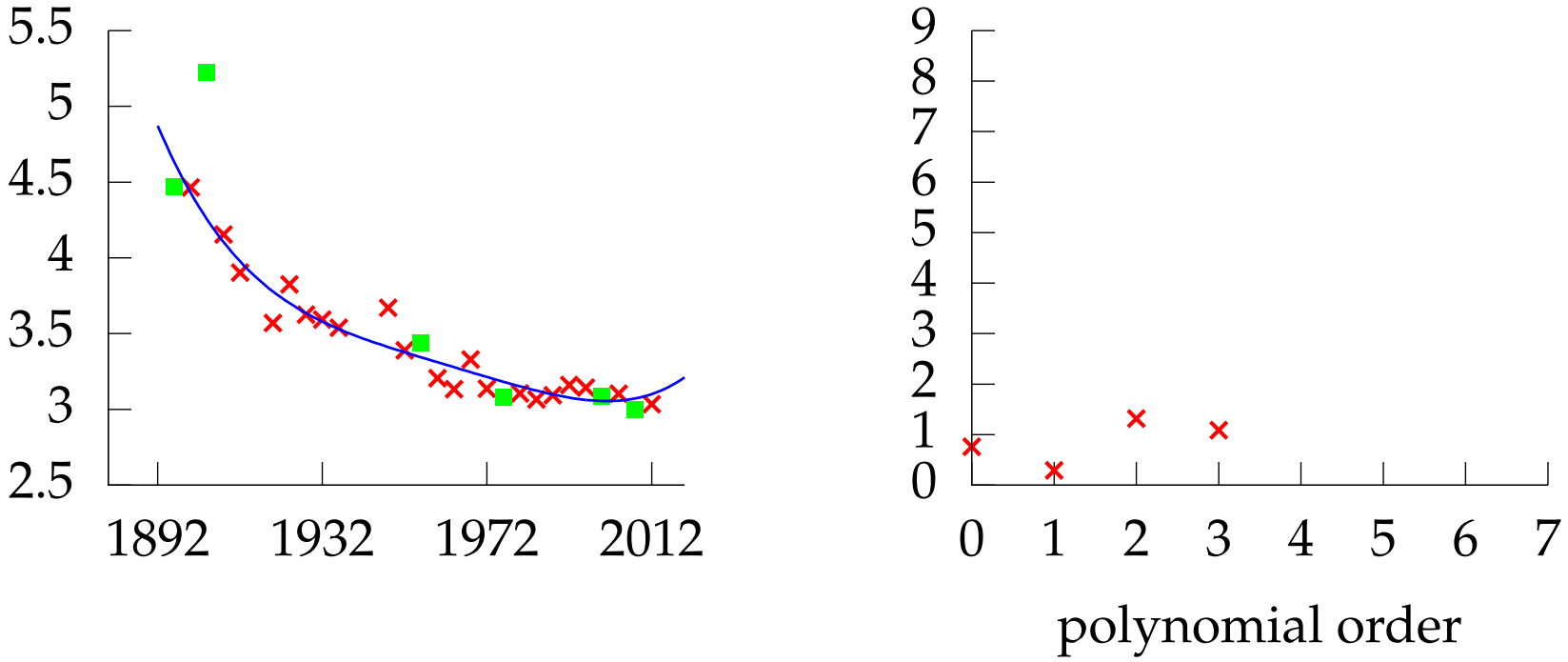
Polynomial order 4, training error -26.048, leave one out error 0.84844.

Cross Validation Error



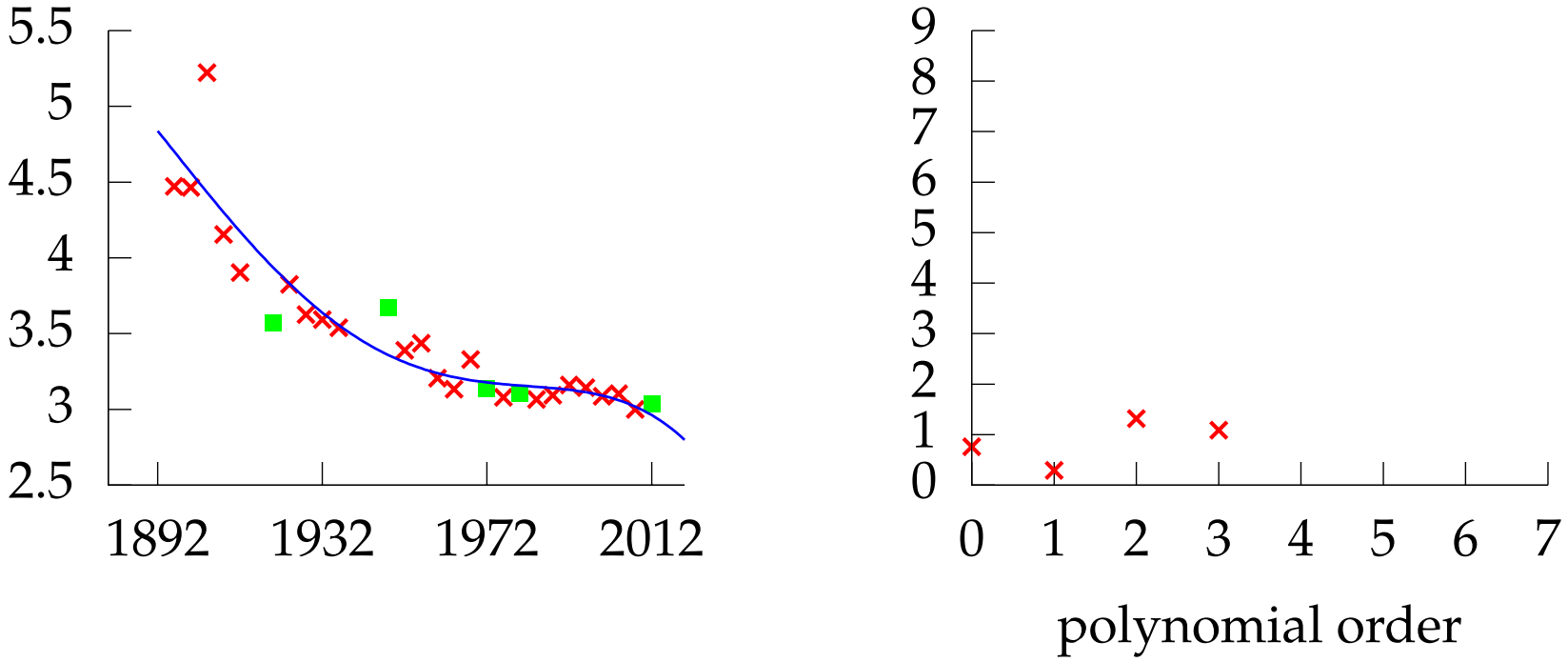
Polynomial order 4, training error -26.048, leave one out error 0.84844.

Cross Validation Error



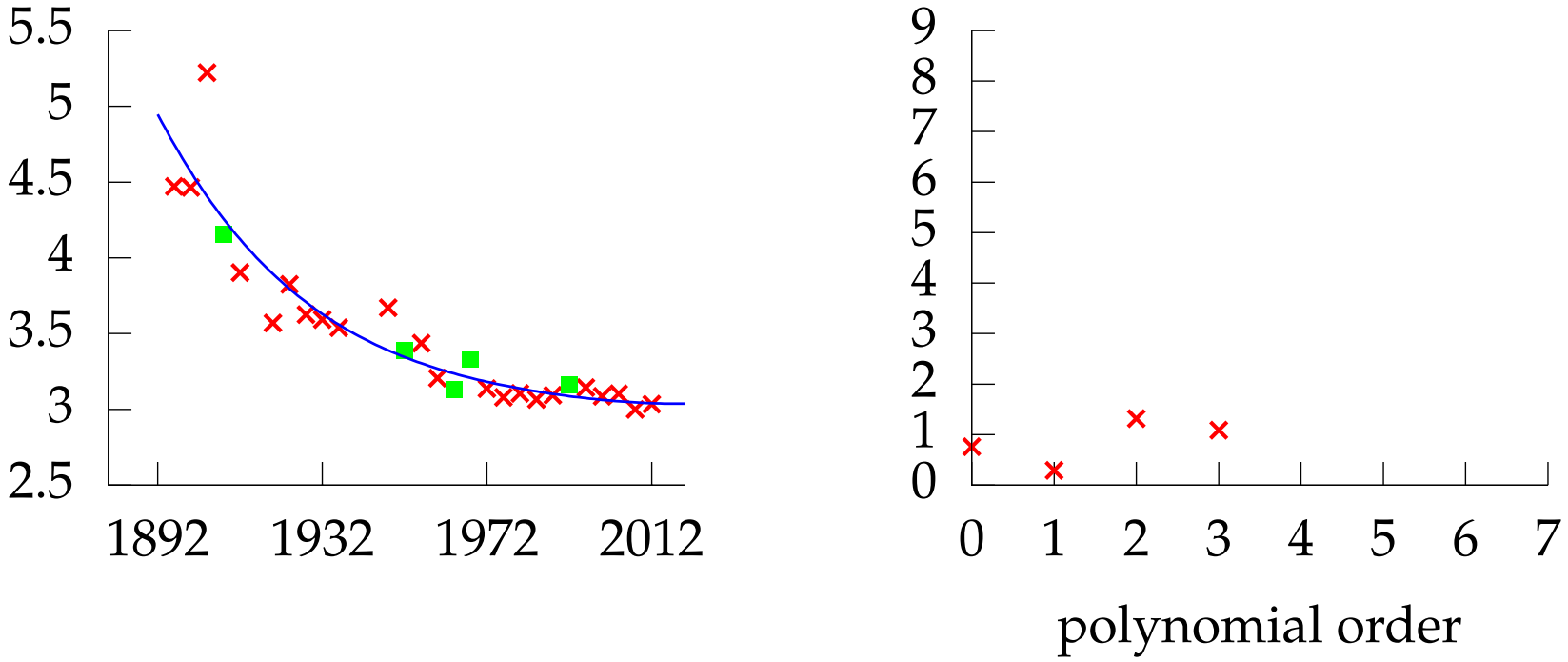
Polynomial order 4, training error -26.048, leave one out error 0.84844.

Cross Validation Error



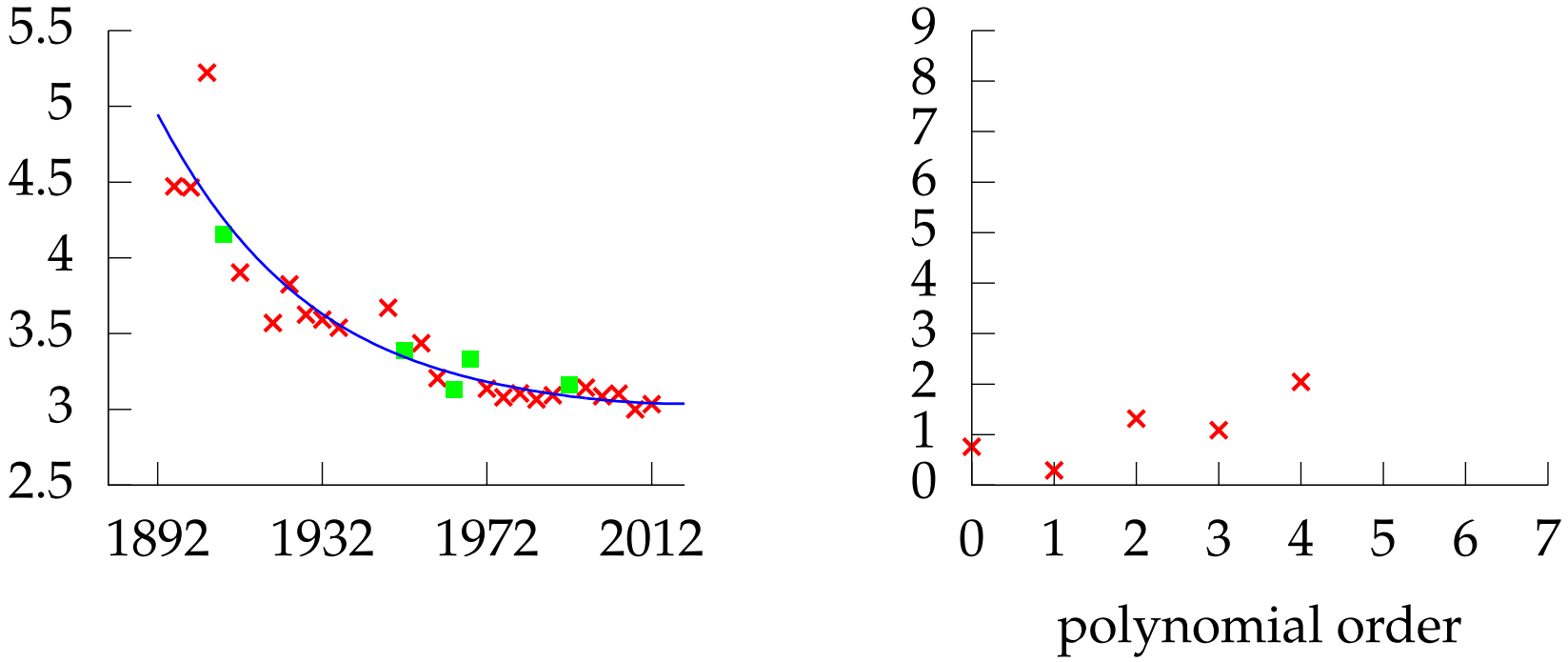
Polynomial order 4, training error -26.048, leave one out error 0.84844.

Cross Validation Error



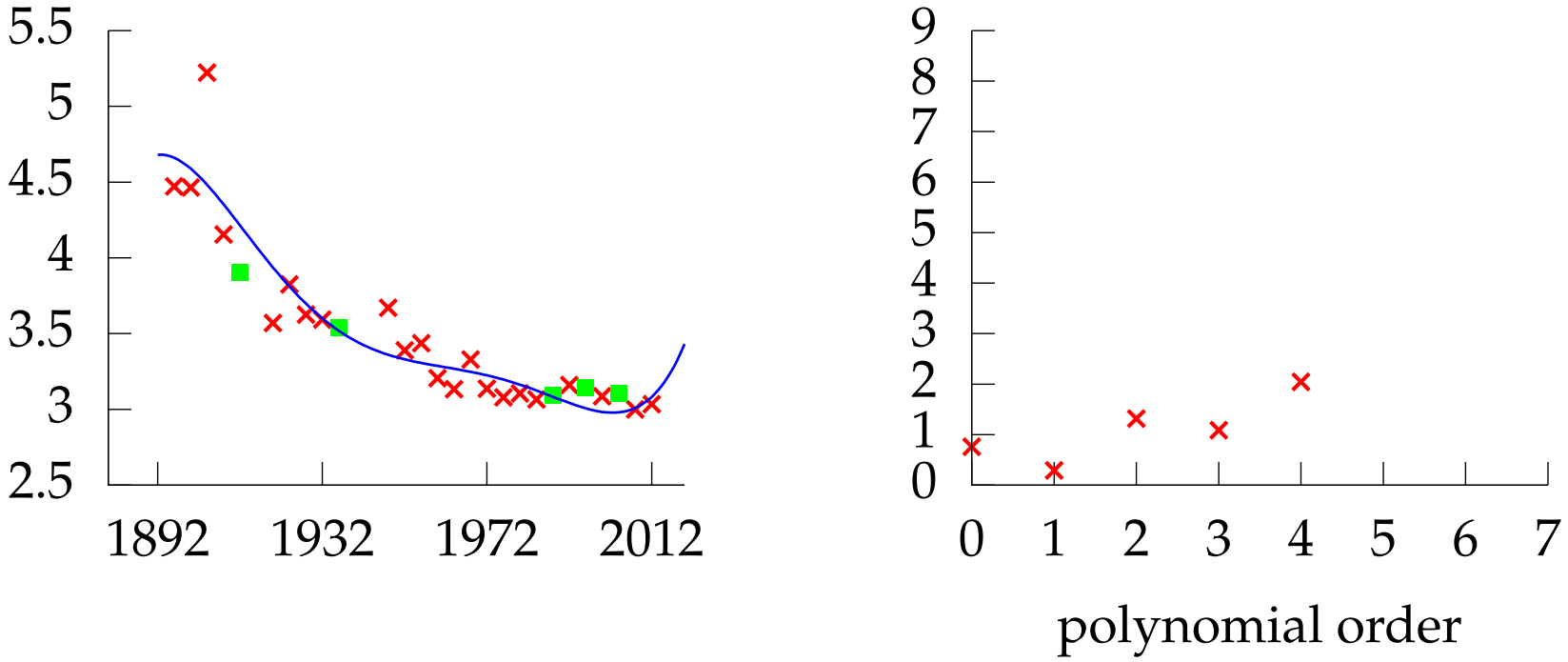
Polynomial order 4, training error -26.048, leave one out error 0.84844.

Cross Validation Error



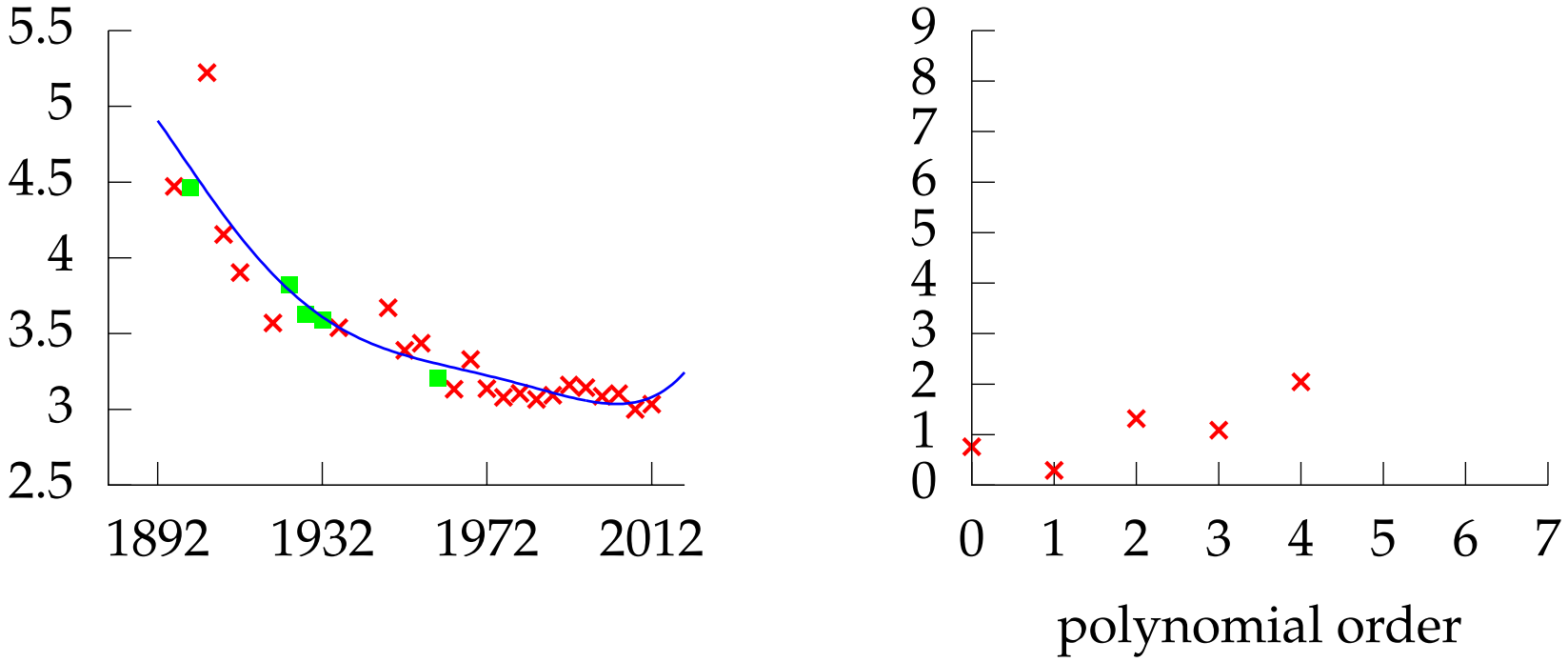
Polynomial order 4, training error -26.048, leave one out error 0.84844.

Cross Validation Error



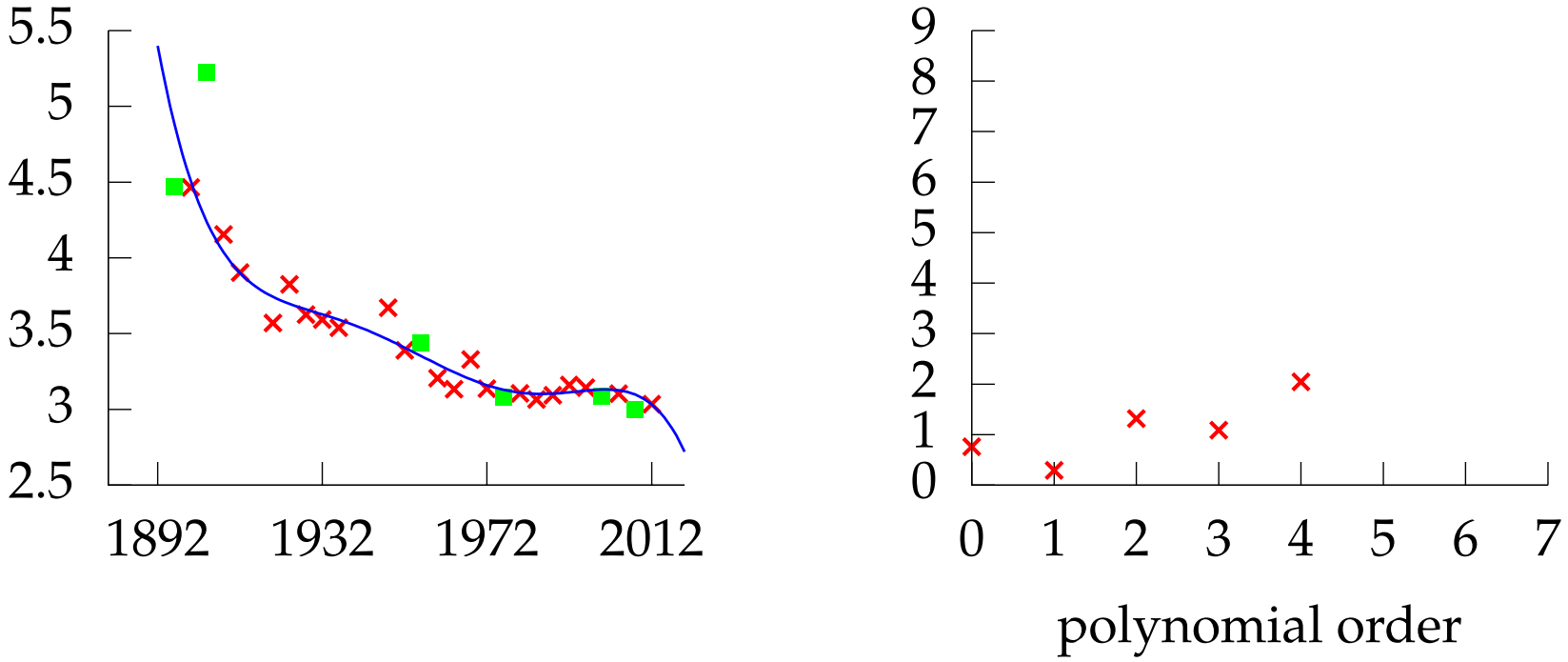
Polynomial order 5, training error -26.892, leave one out error 1.48.

Cross Validation Error



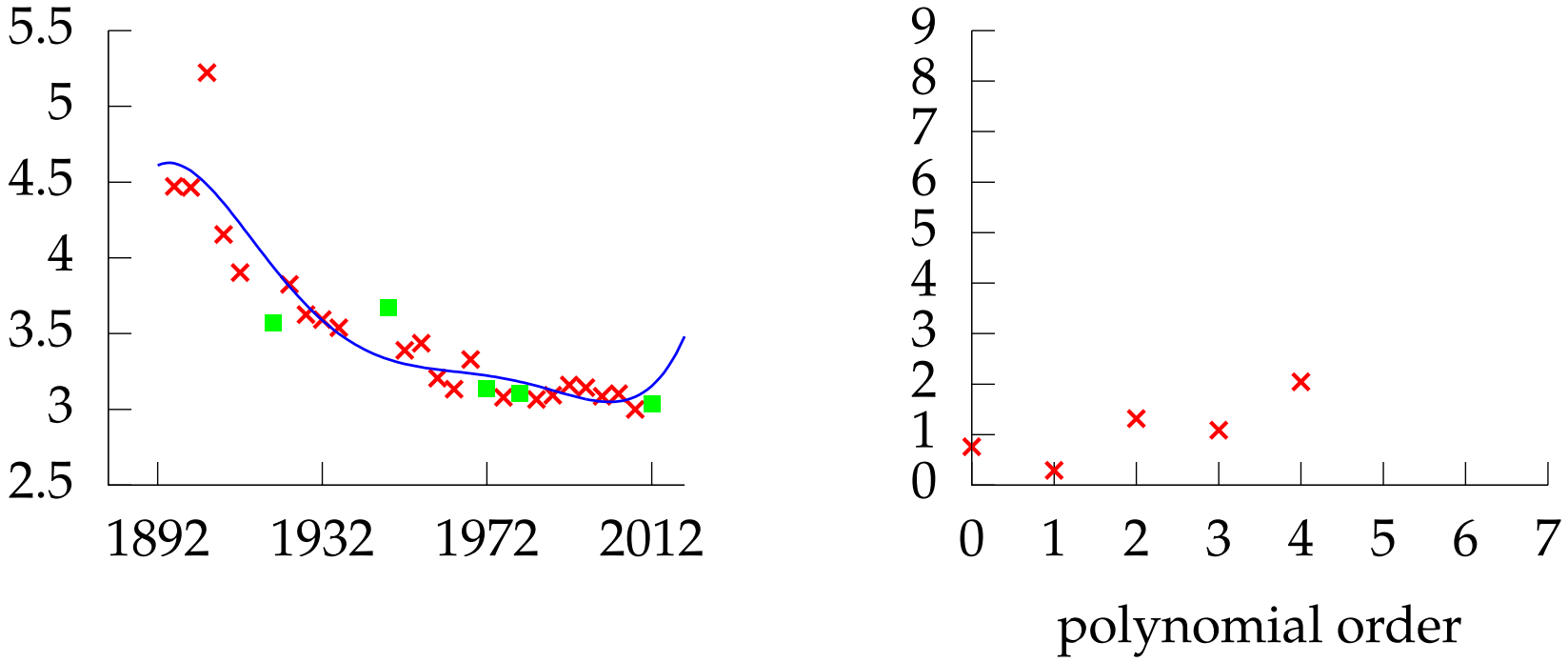
Polynomial order 5, training error -26.892, leave one out error 1.48.

Cross Validation Error



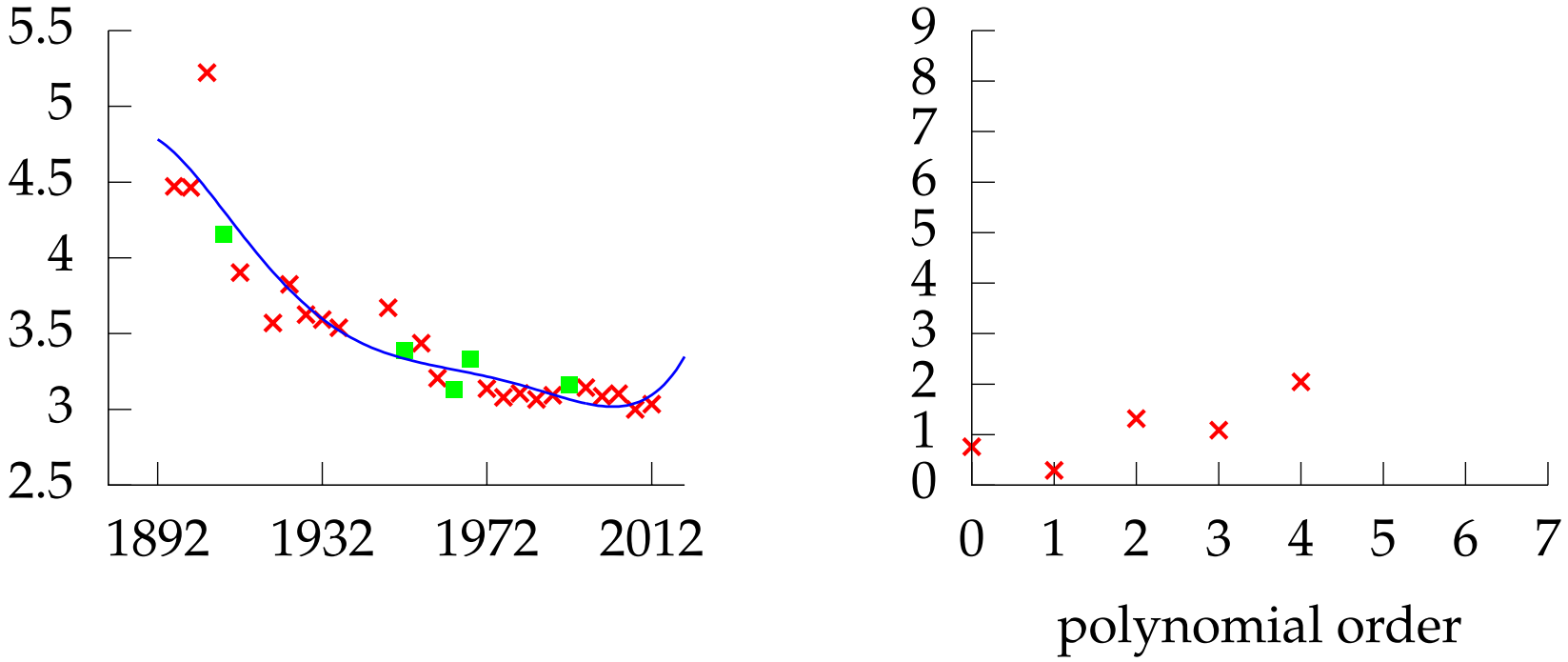
Polynomial order 5, training error -26.892, leave one out error 1.48.

Cross Validation Error



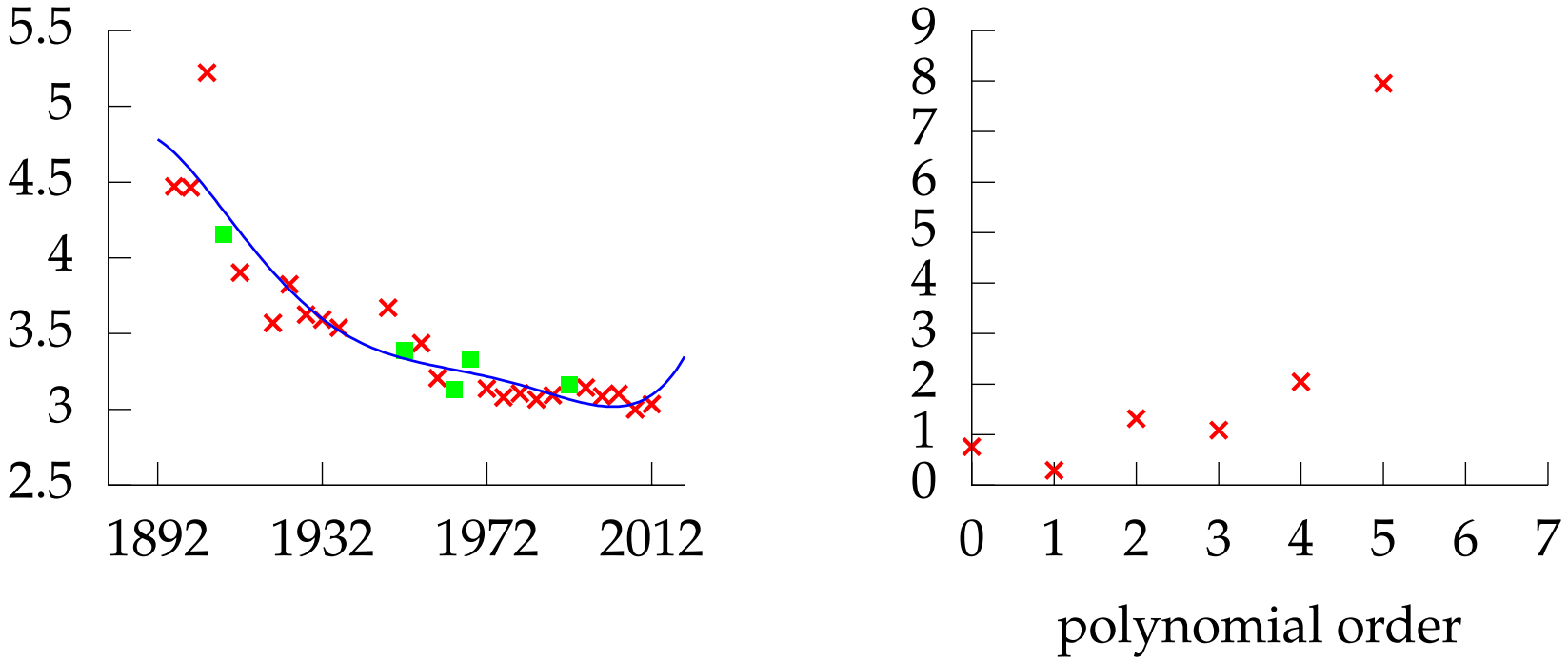
Polynomial order 5, training error -26.892, leave one out error 1.48.

Cross Validation Error



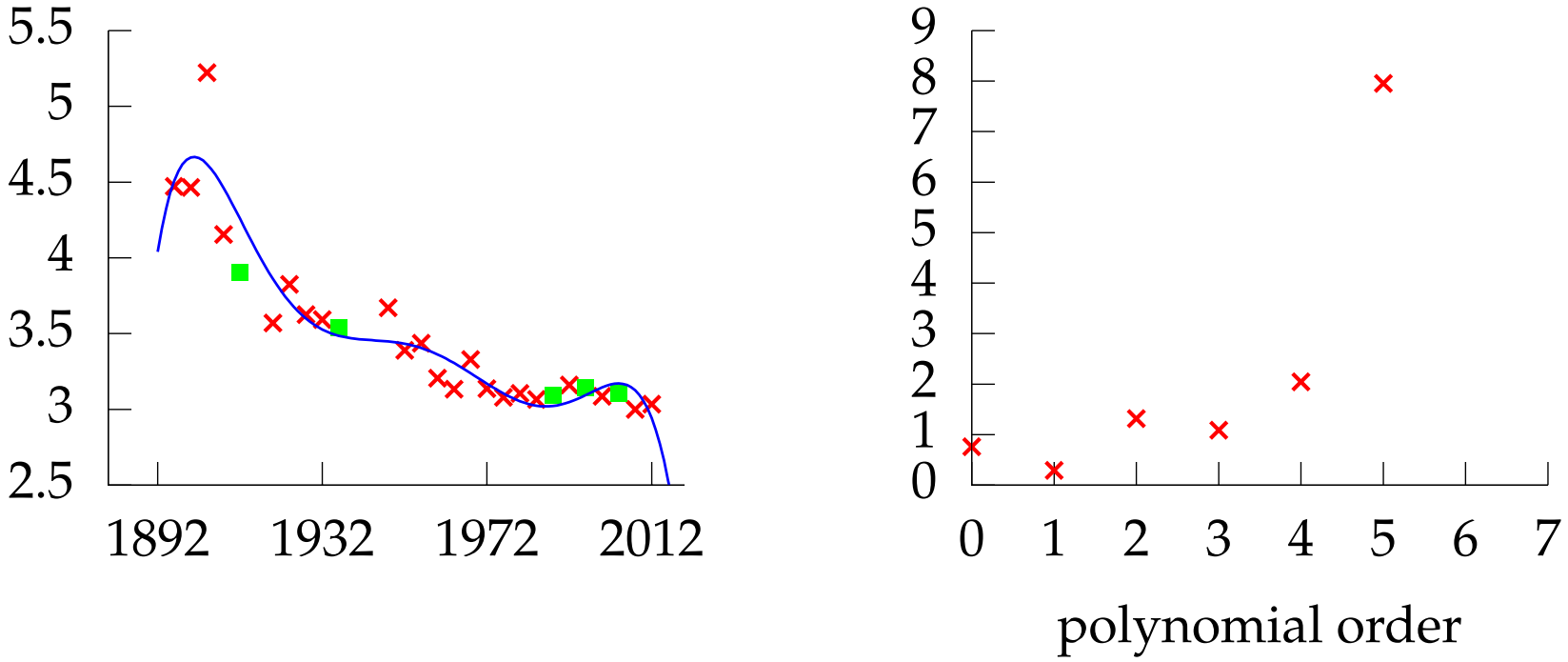
Polynomial order 5, training error -26.892, leave one out error 1.48.

Cross Validation Error



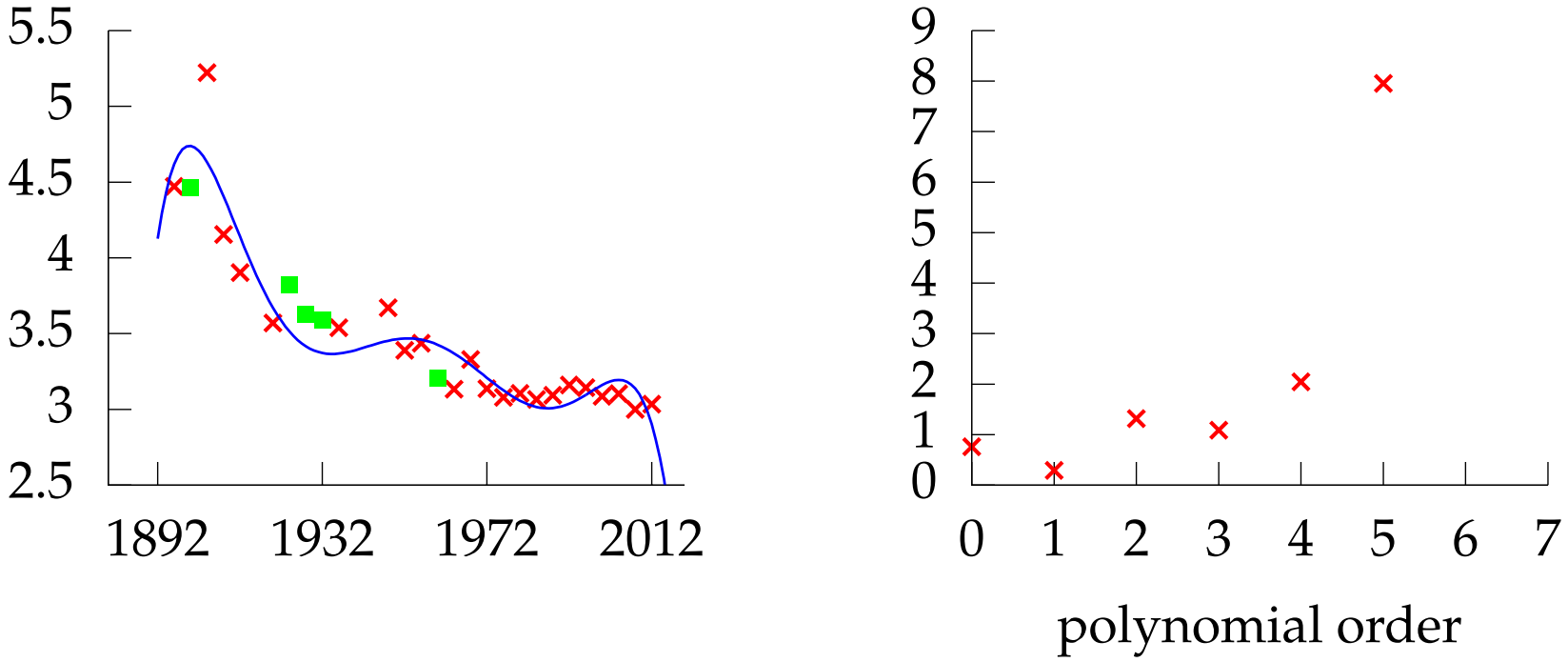
Polynomial order 5, training error -26.892, leave one out error 1.48.

Cross Validation Error



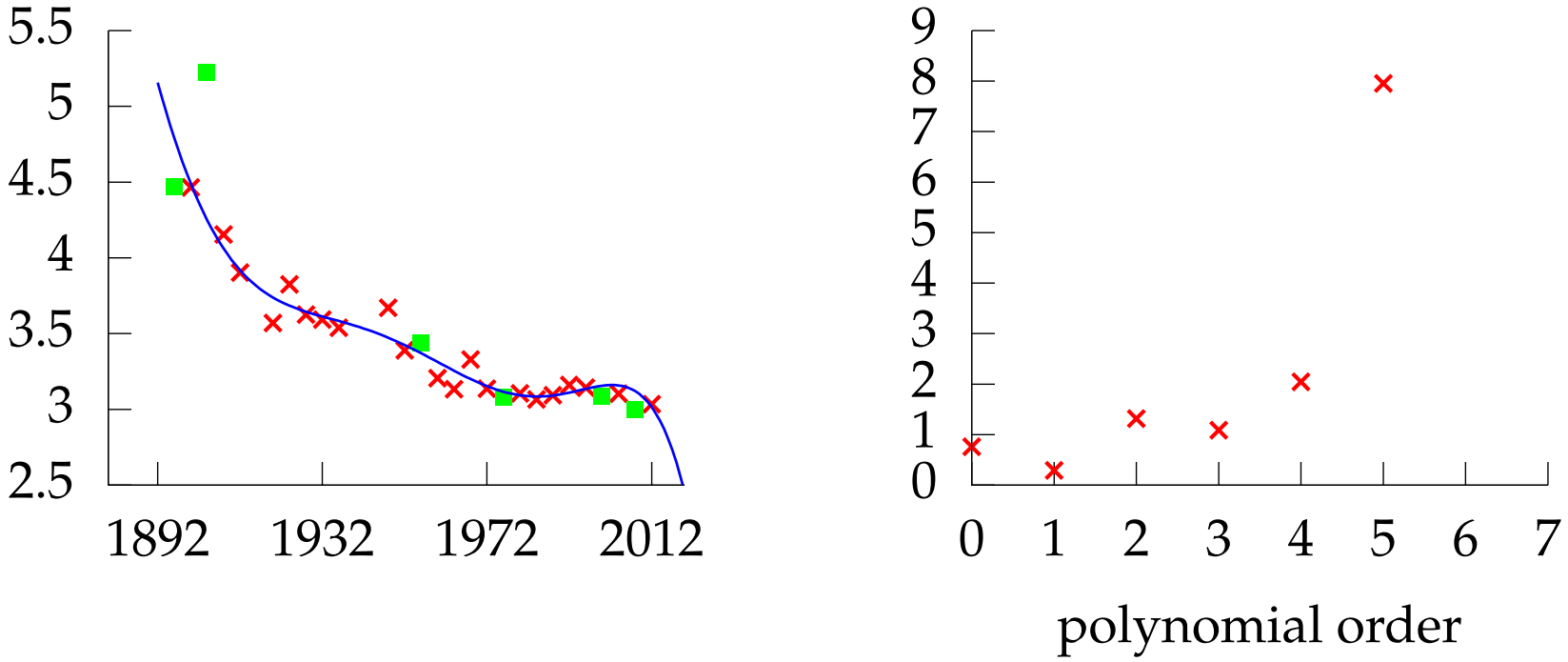
Polynomial order 6, training error -29.395, leave one out error 1.5047.

Cross Validation Error



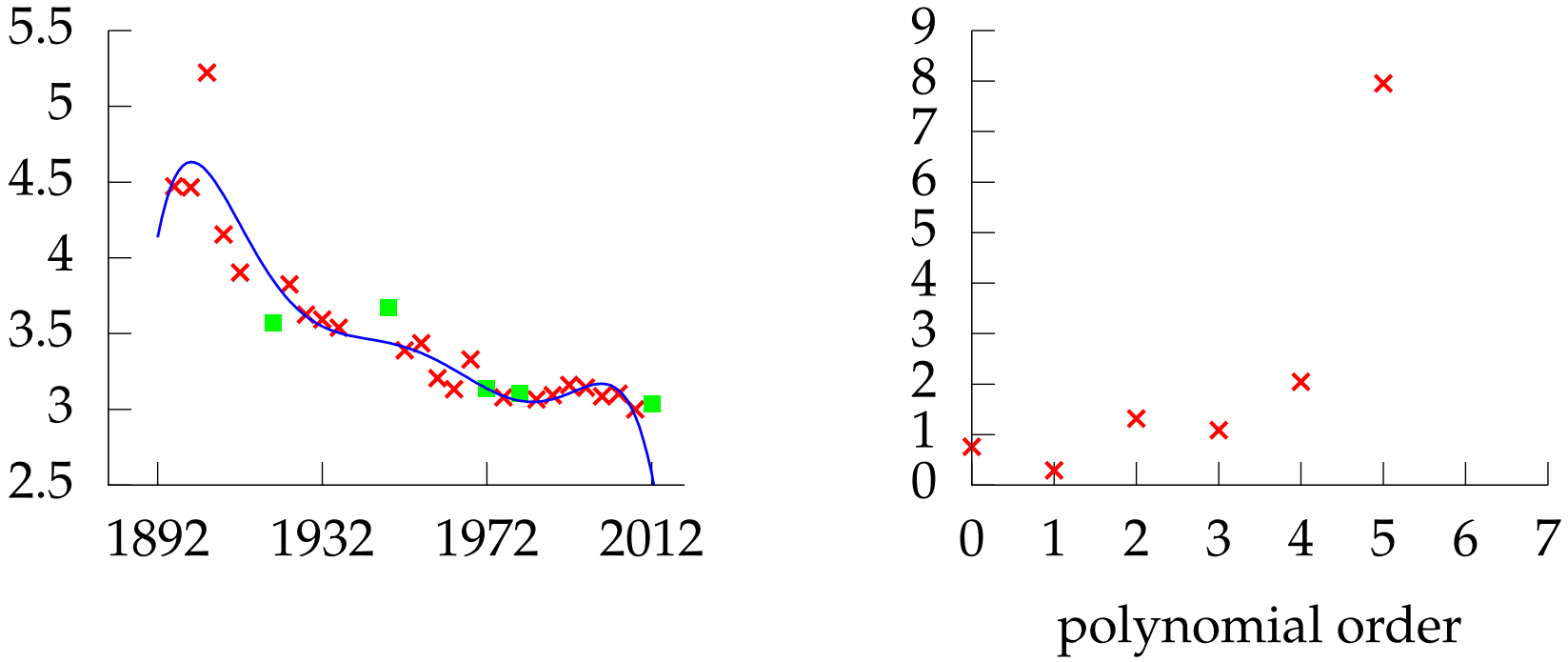
Polynomial order 6, training error -29.395, leave one out error 1.5047.

Cross Validation Error



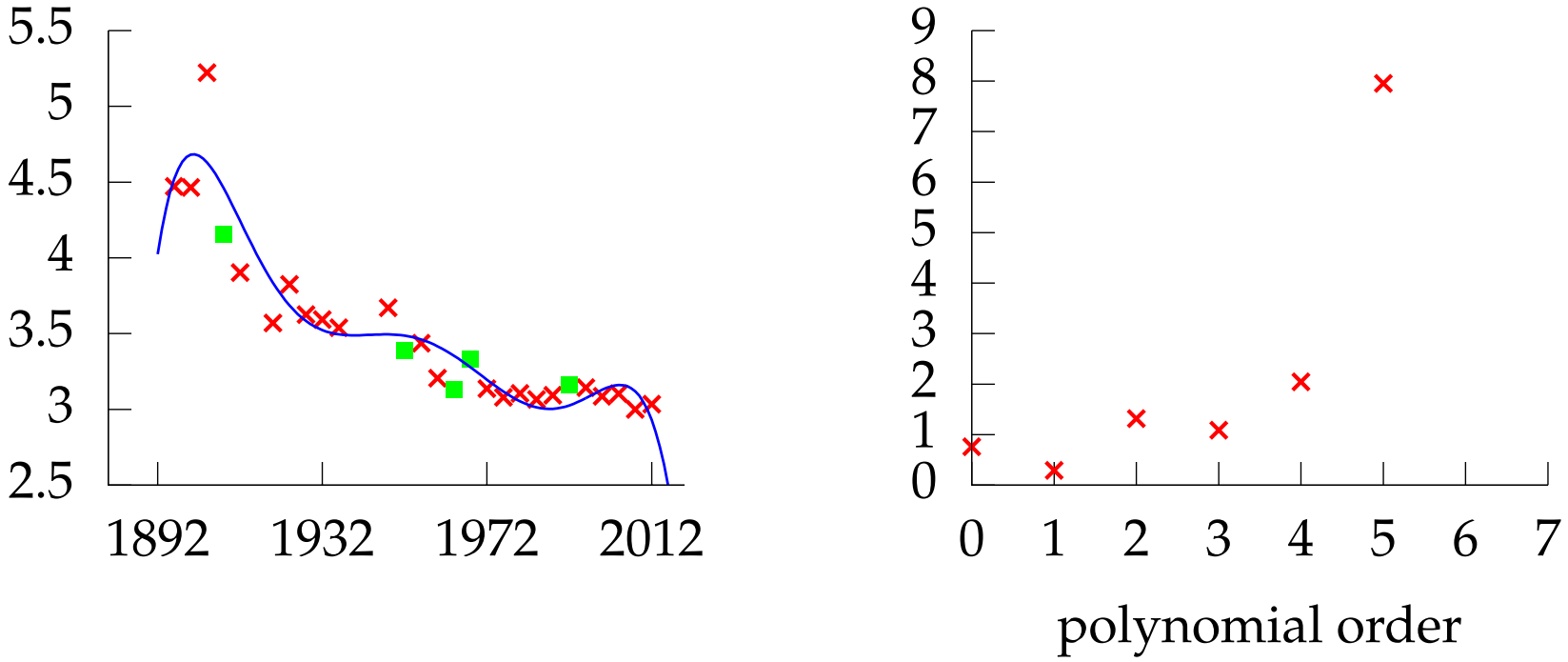
Polynomial order 6, training error -29.395, leave one out error 1.5047.

Cross Validation Error



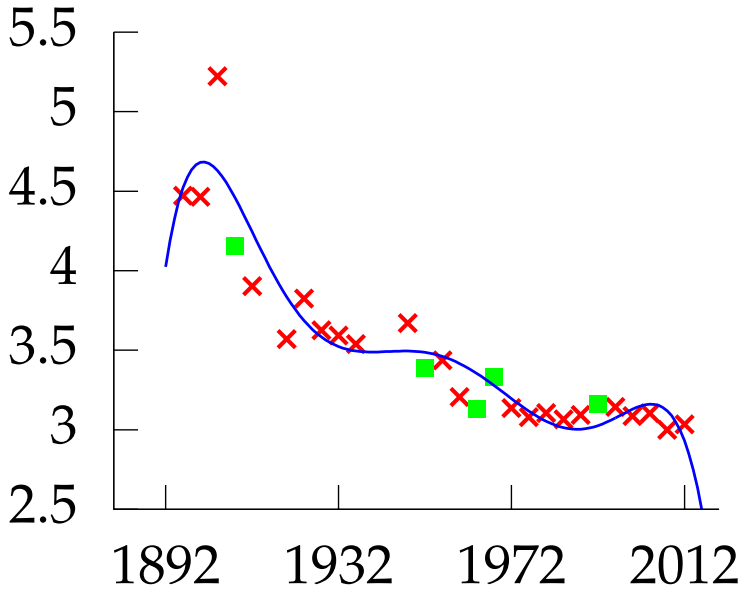
Polynomial order 6, training error -29.395, leave one out error 1.5047.

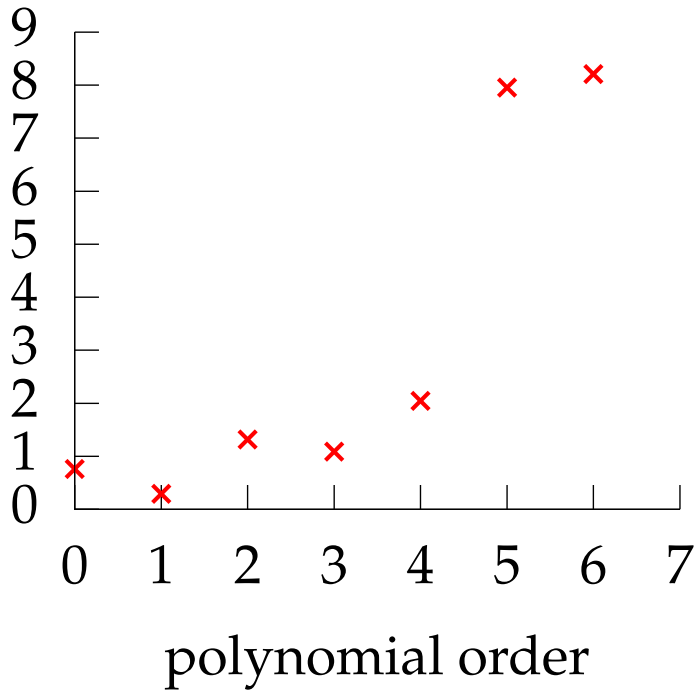
Cross Validation Error



Polynomial order 6, training error -29.395, leave one out error 1.5047.

Cross Validation Error



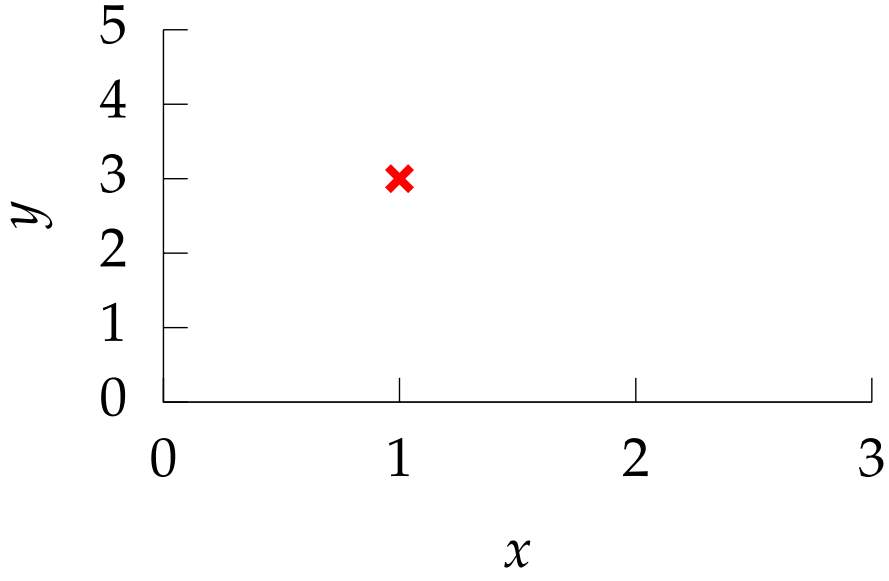


Polynomial order 6, training error -29.395, leave one out error 1.5047.

Underdetermined System

What about two unknowns and *one* observation?

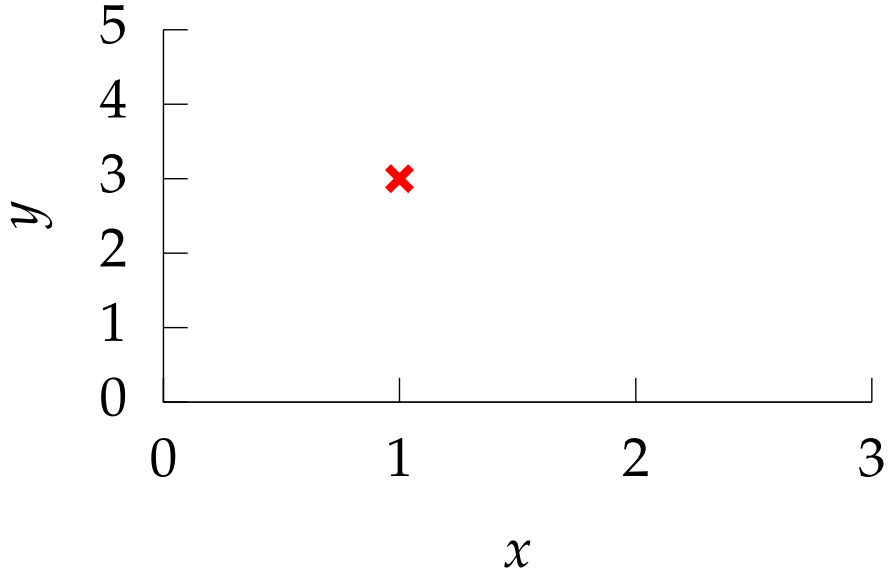
$$y_1 = mx_1 + c$$



Underdetermined System

Can compute m given c.

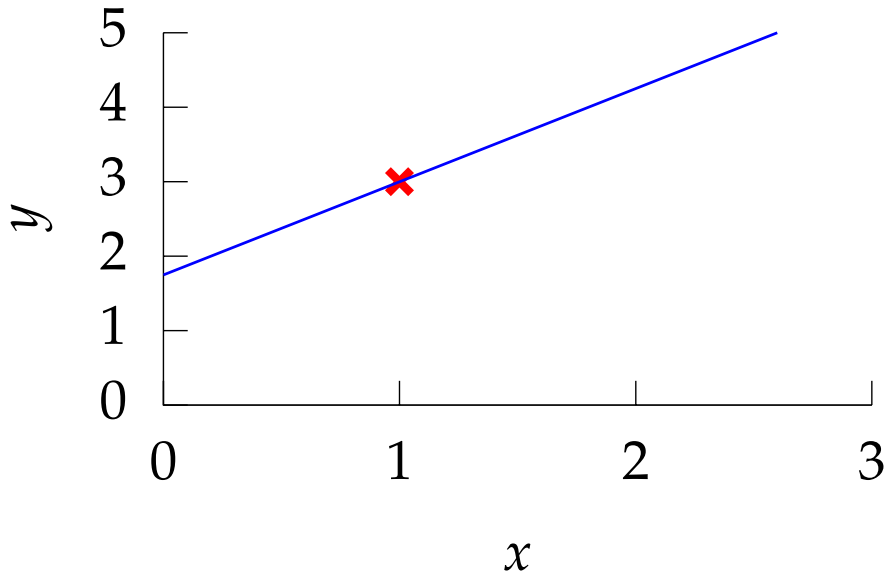
$$m = \frac{y_1 - c}{x}$$



Underdetermined System

Can compute m given c.

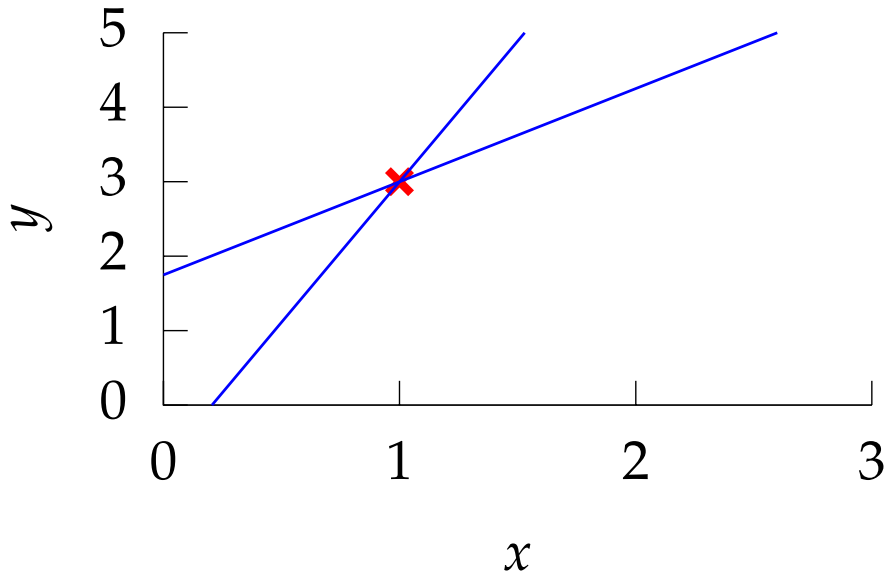
$$c = 1.75 \implies m = 1.25$$



Underdetermined System

Can compute m given c.

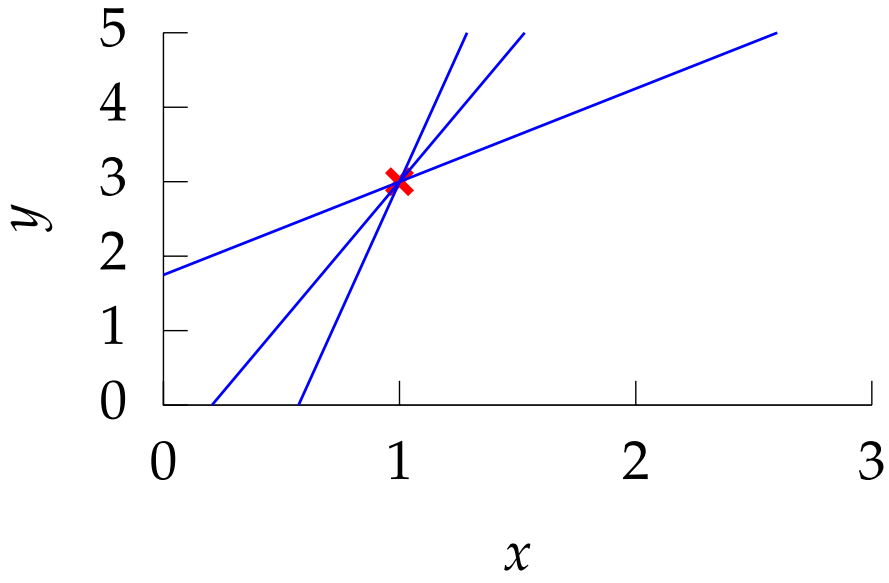
$$c = -0.777 \implies m = 3.78$$



Underdetermined System

Can compute m given c.

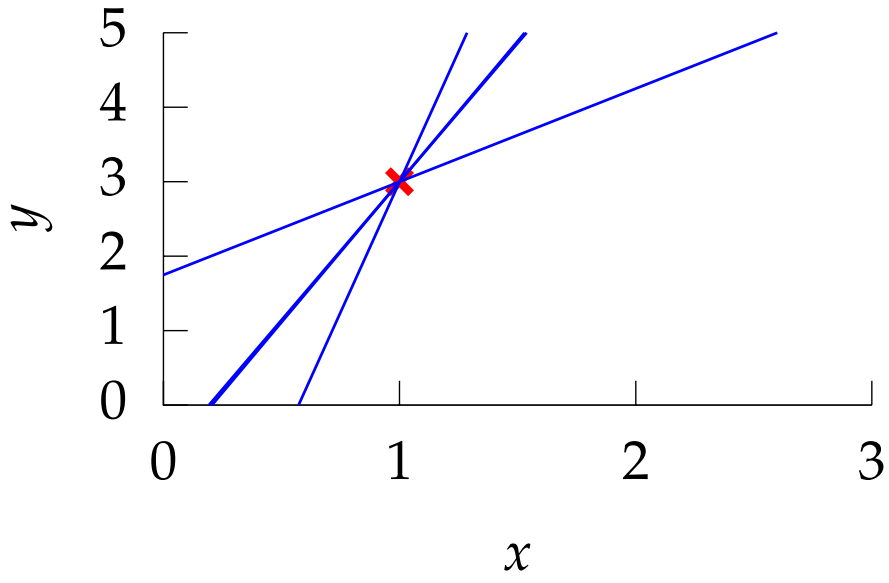
$$c = -4.01 \implies m = 7.01$$



Underdetermined System

Can compute m given c.

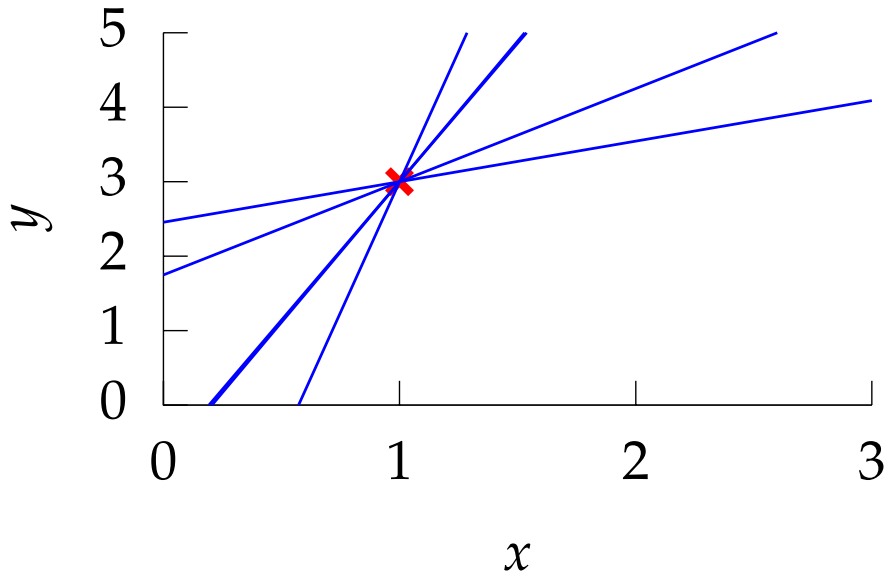
$$c = -0.718 \implies m = 3.72$$



Underdetermined System

Can compute m given c.

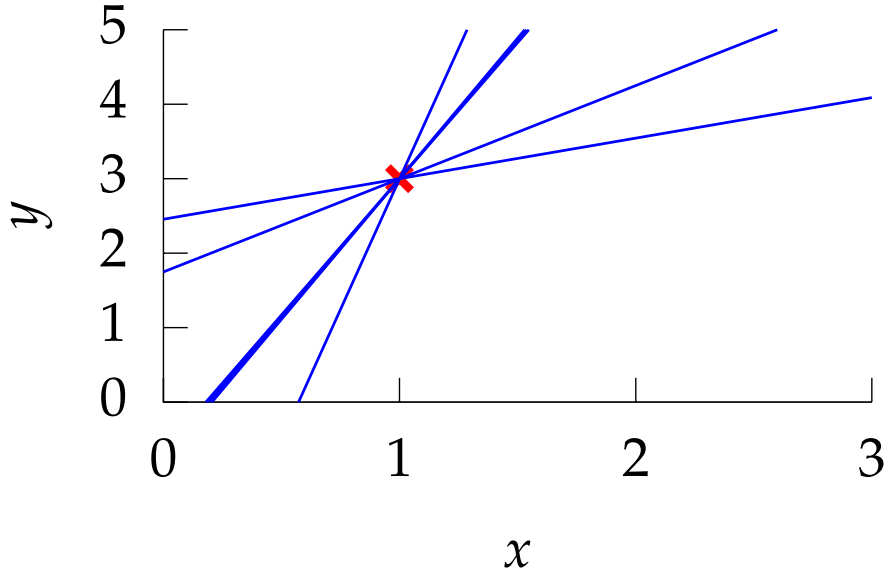
$$c = 2.45 \implies m = 0.545$$



Underdetermined System

Can compute m given c.

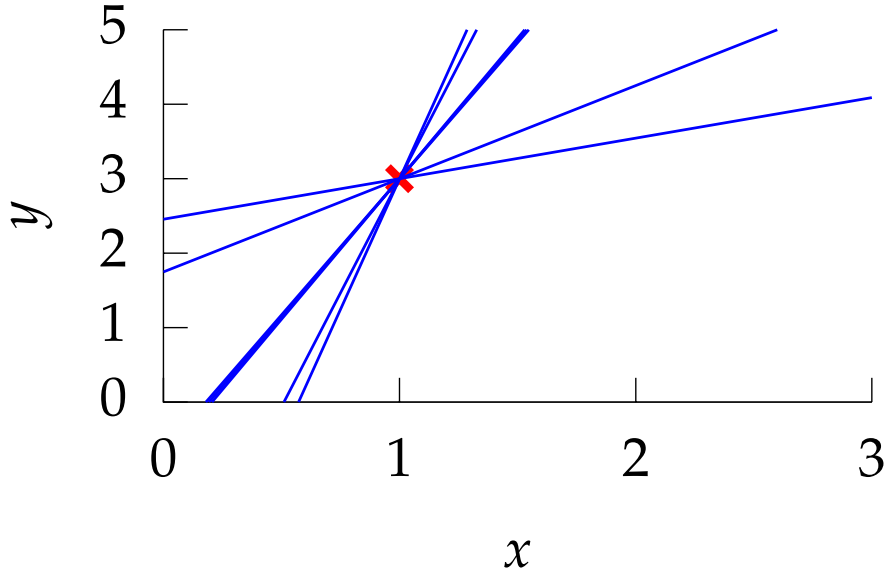
$$c = -0.657 \implies m = 3.66$$



Underdetermined System

Can compute m given c.

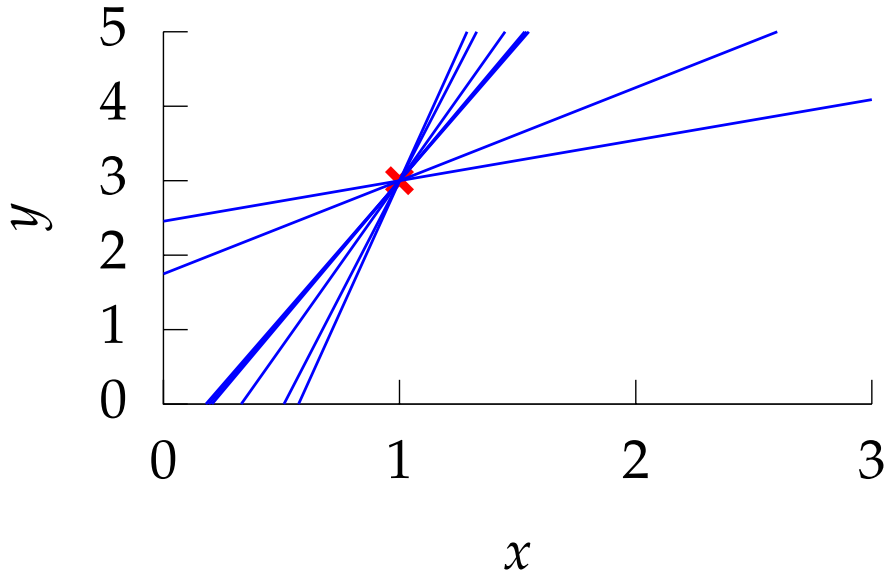
$$c = -3.13 \implies m = 6.13$$



Underdetermined System

Can compute m given c.

$$c = -1.47 \implies m = 4.47$$

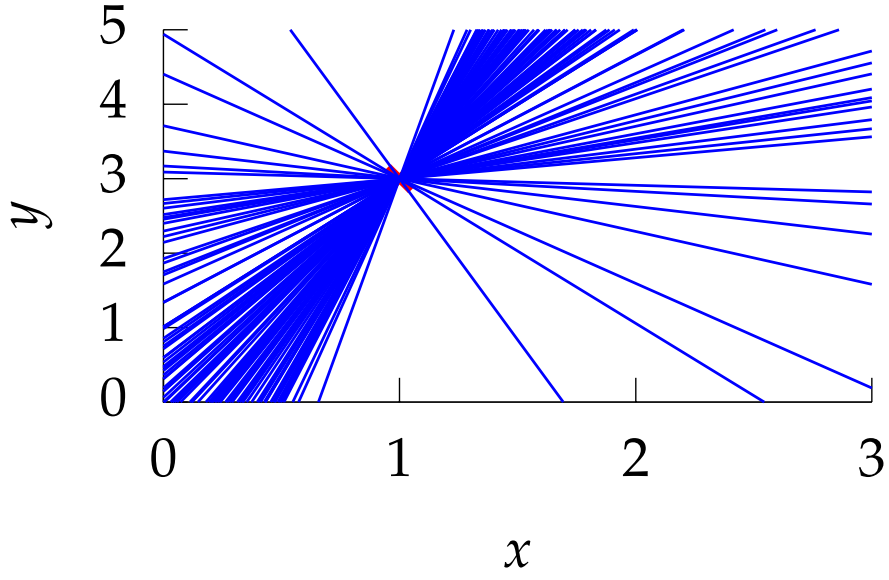


Underdetermined System

Can compute m given c.
Assume

$$c \sim \mathcal{N}(0, 4),$$

we find a distribution of solutions.



Probability for Under- and Overdetermined

- To deal with overdetermined introduced probability distribution for 'variable', ϵ_i.
- For underdetermined system introduced probability distribution for 'parameter', c.
- This is known as a Bayesian treatment.

Reading

- Bishop Section 1.2.3 (pg 21–24).
- Bishop Section 1.2.6 (start from just past eq 1.64 pg 30-32).
- Rogers and Girolami use an example of a coin toss for introducing Bayesian inference Chapter 3, Sections 3.1-3.4 (pg 95-117). Although you also need the beta density which we haven't yet discussed. This is also the example that Laplace used.

Prior Distribution

- Bayesian inference requires a prior on the parameters.
- The prior represents your belief *before* you see the data of the likely value of the parameters.
- For linear regression, consider a Gaussian prior on the intercept:

$$c \sim \mathcal{N}(0, \alpha_1)$$

Posterior Distribution

- Posterior distribution is found by combining the prior with the likelihood.
- Posterior distribution is your belief *after* you see the data of the likely value of the parameters.
- The posterior is found through **Bayes' Rule**

$$p(c|y) = \frac{p(y|c)p(c)}{p(y)}$$

Bayes Update

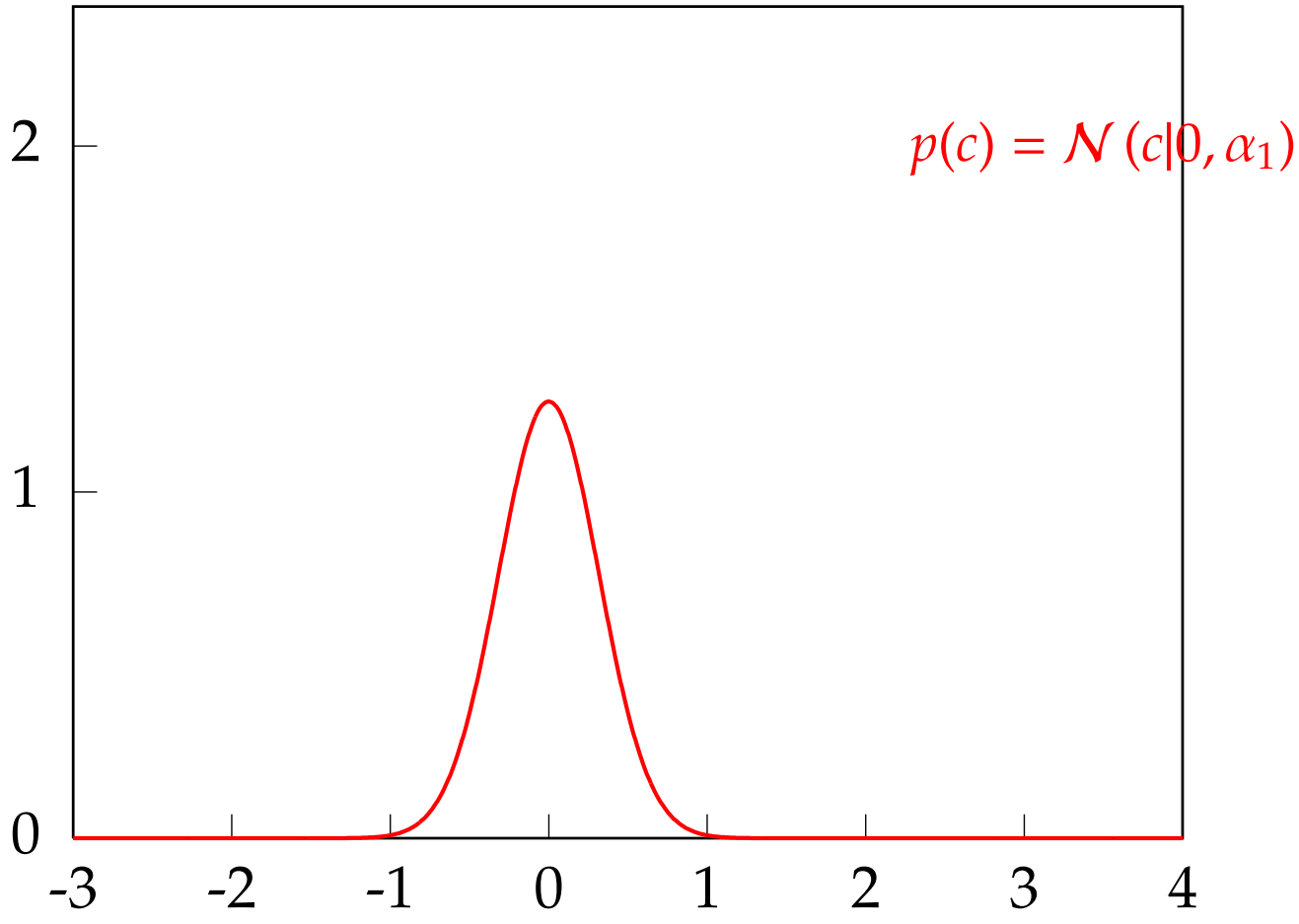


Figure: A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Bayes Update

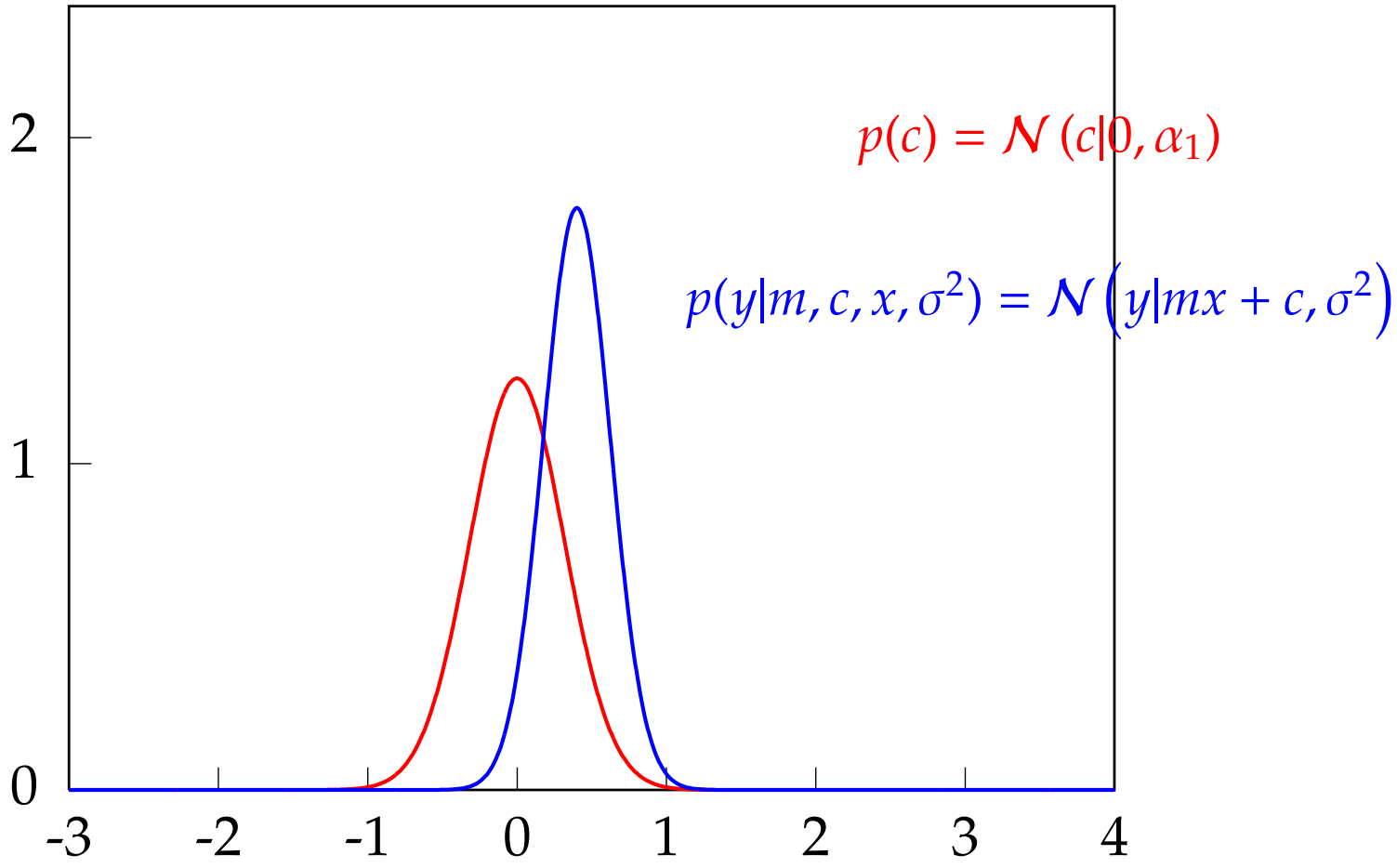


Figure: A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Bayes Update

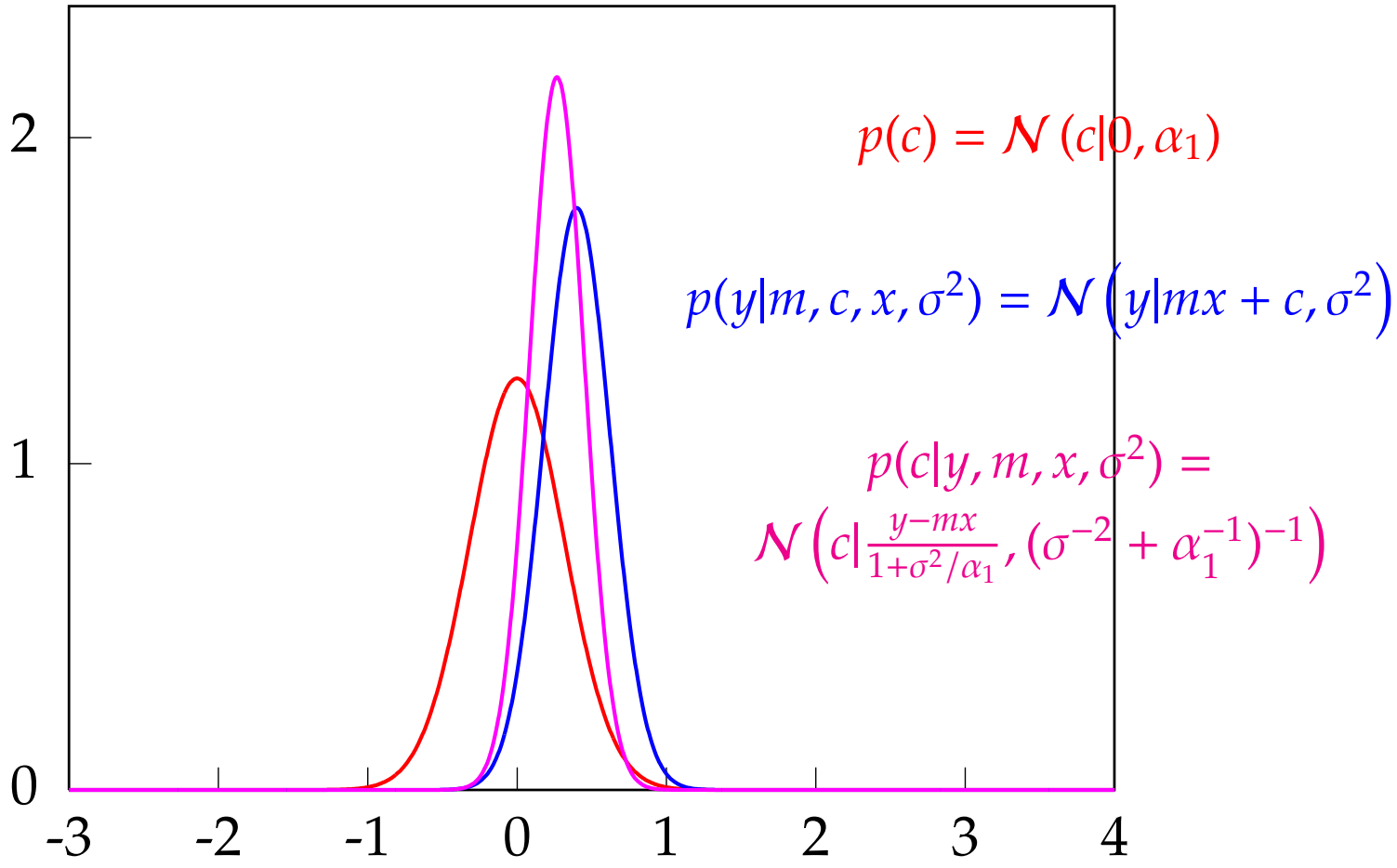


Figure: A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Stages to Derivation of the Posterior

- Multiply likelihood by prior
  - they are "exponentiated quadratics", the answer is always also an exponentiated quadratic because $\exp(a^2)\exp(b^2) = \exp(a^2 + b^2)$.
- Complete the square to get the resulting density in the form of a Gaussian.
- Recognise the mean and (co)variance of the Gaussian. This is the estimate of the posterior.

Multivariate Prior Distributions

- For general Bayesian inference need multivariate priors.
- E.g. for multivariate linear regression:

$$y_i = \sum_i w_j x_{i,j} + \epsilon_i$$

(where we've dropped c for convenience), we need a prior over $\mathbf{w}$.

- This motivates a *multivariate* Gaussian density.

Multivariate Prior Distributions

- For general Bayesian inference need multivariate priors.
- E.g. for multivariate linear regression:

$$y_i = \mathbf{w}^\top \mathbf{x}_{i,:} + \epsilon_i$$

(where we've dropped c for convenience), we need a prior over $\mathbf{w}$.

- This motivates a *multivariate* Gaussian density.

Two Dimensional Gaussian

- Consider height, h/m and weight, w/kg.
- Could sample height from a distribution:

$$p(h) \sim \mathcal{N}(1.7, 0.0225)$$

- And similarly weight:

$$p(w) \sim \mathcal{N}(75, 36)$$

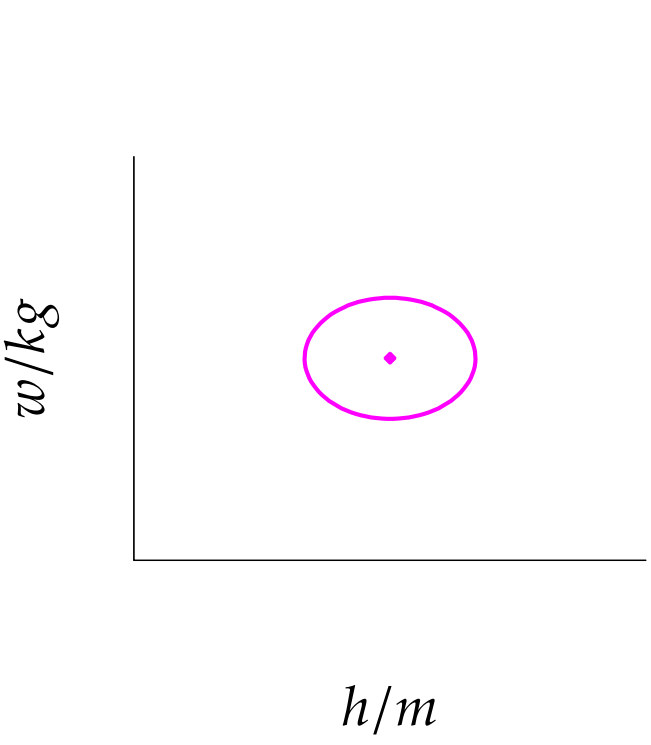
Height and Weight Models

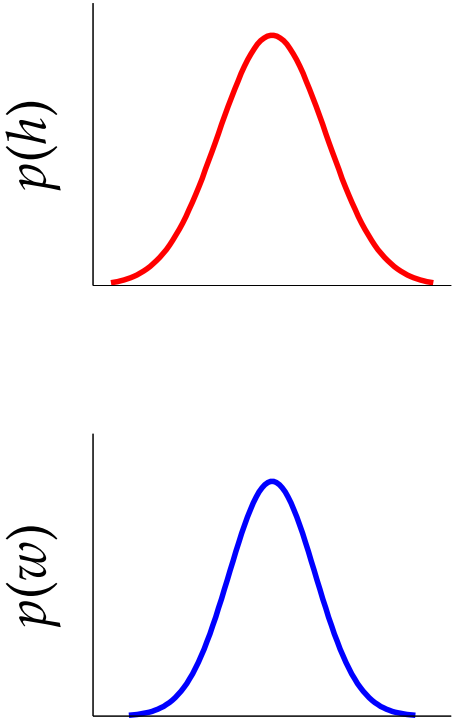
Gaussian distributions for height and weight.

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



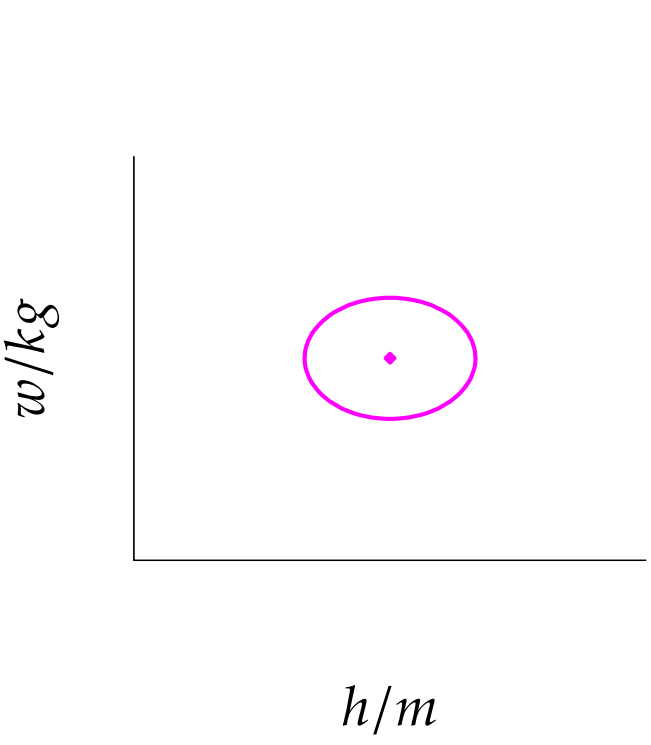


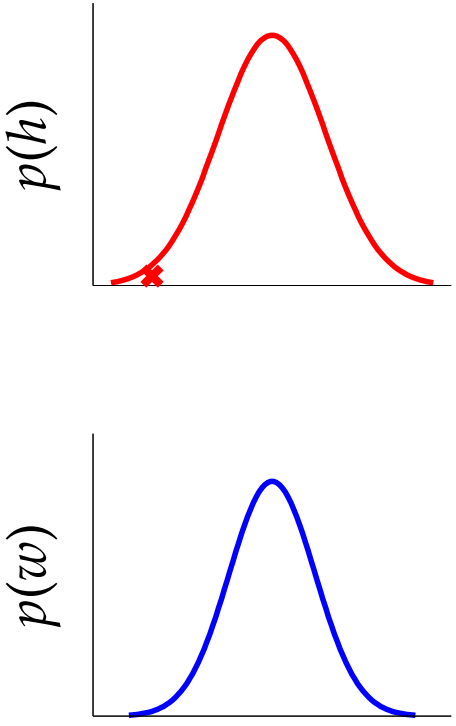
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



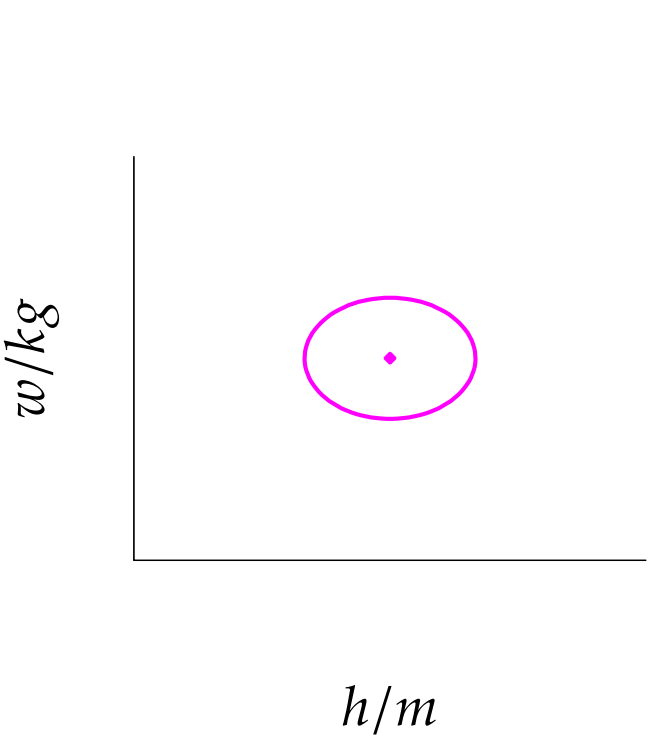


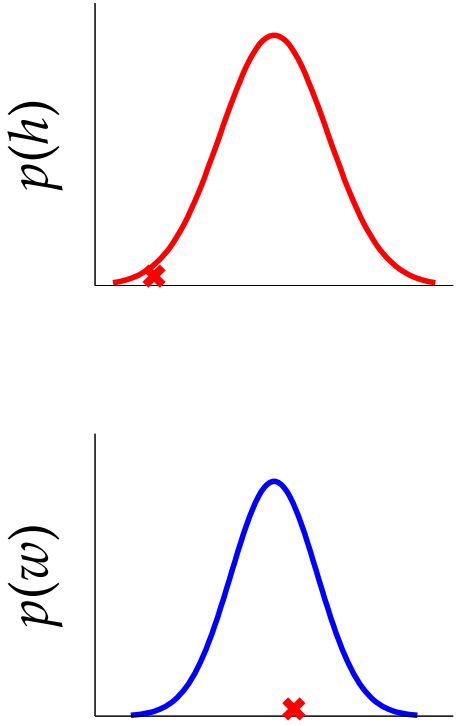
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



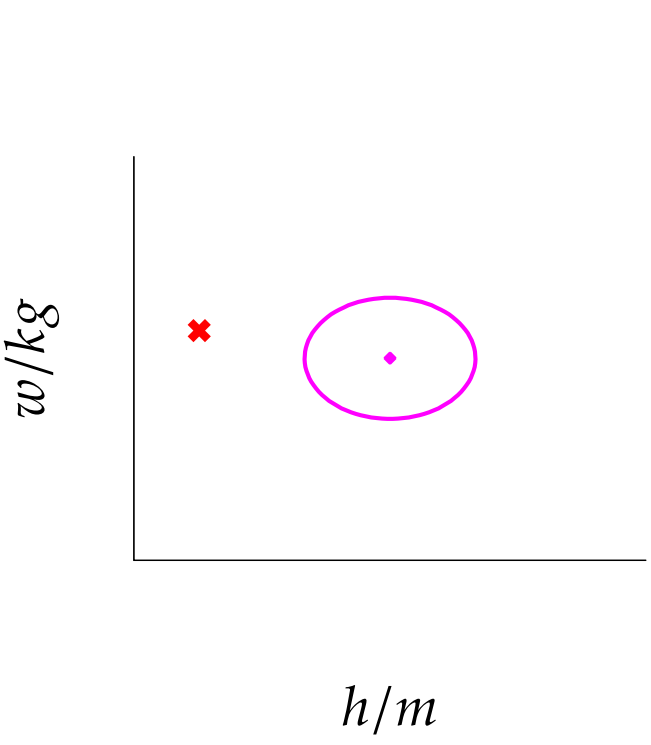


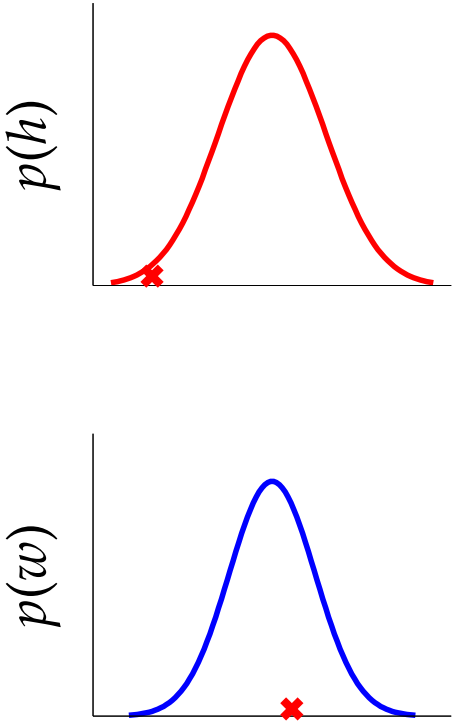
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



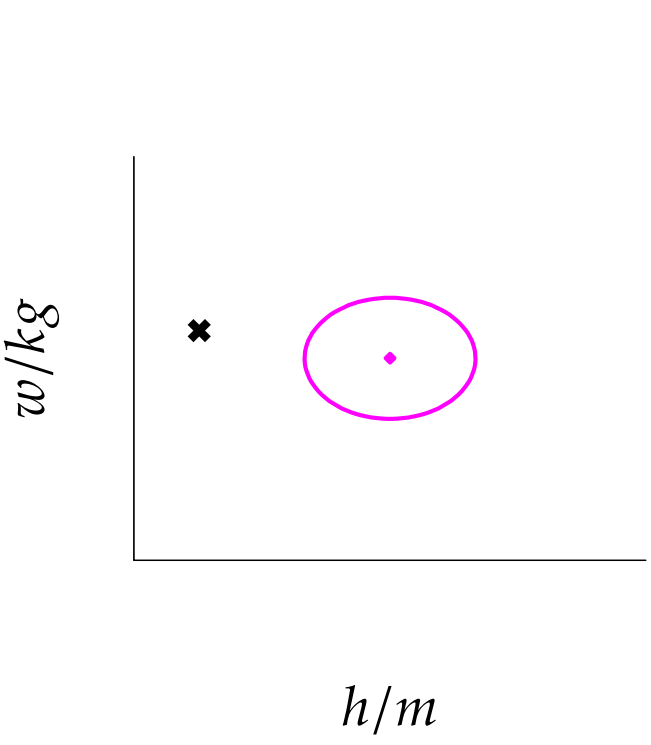


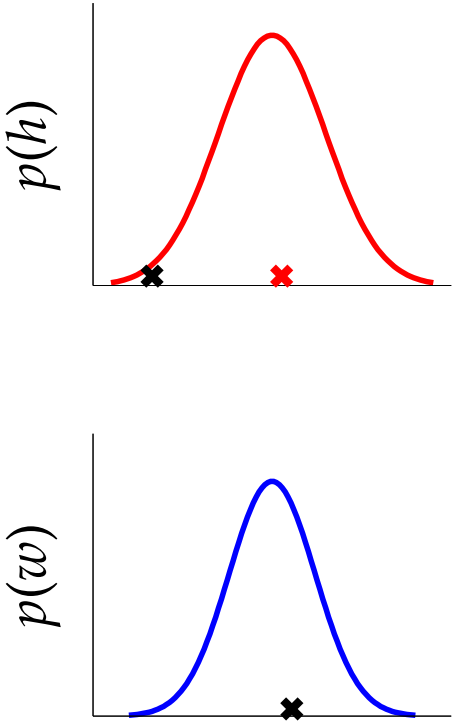
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



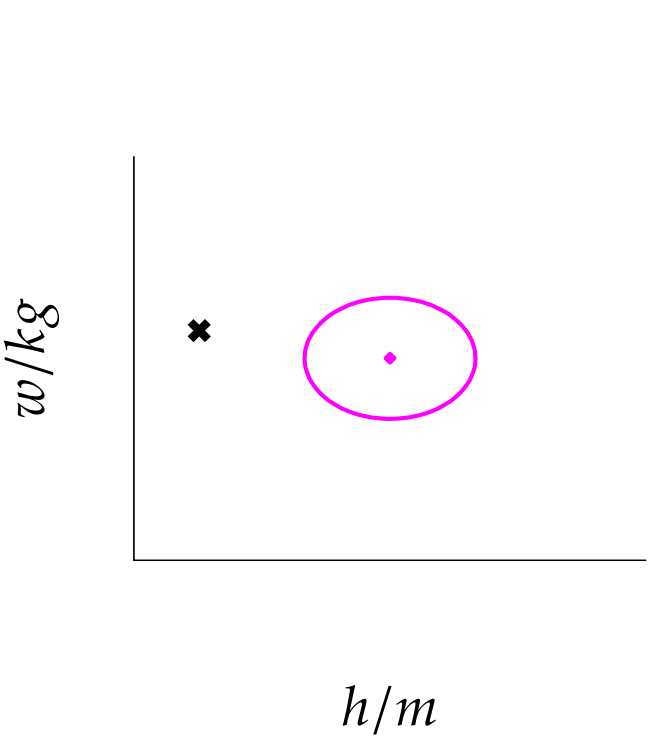


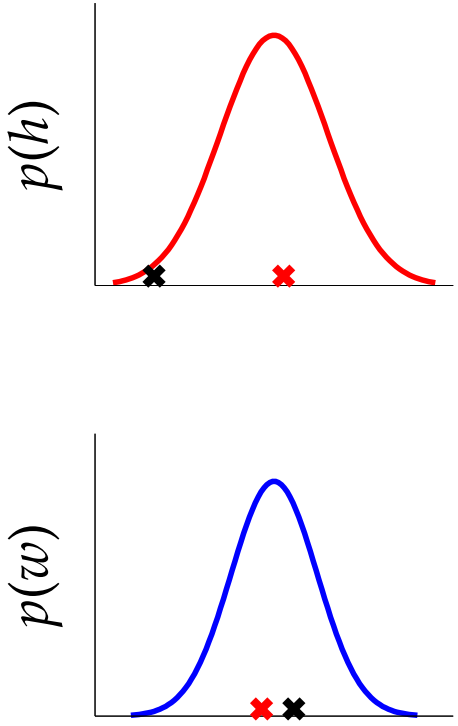
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



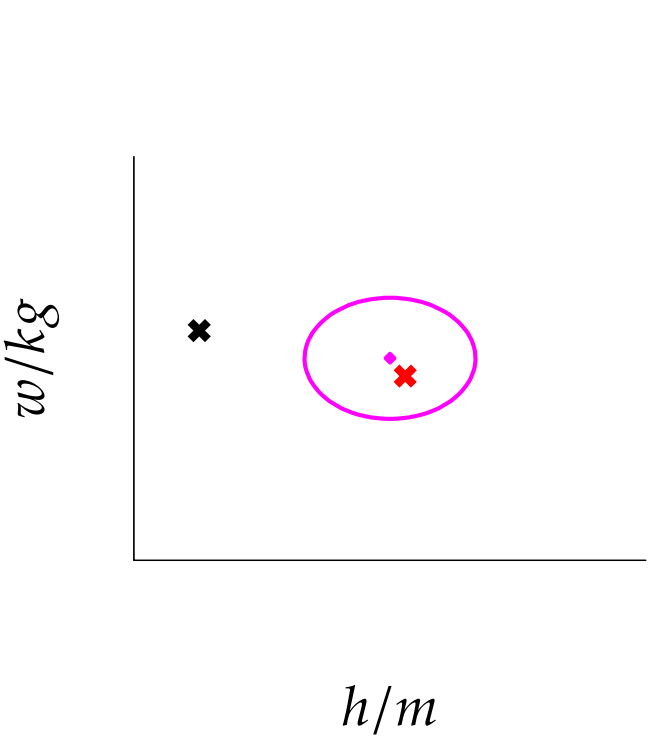


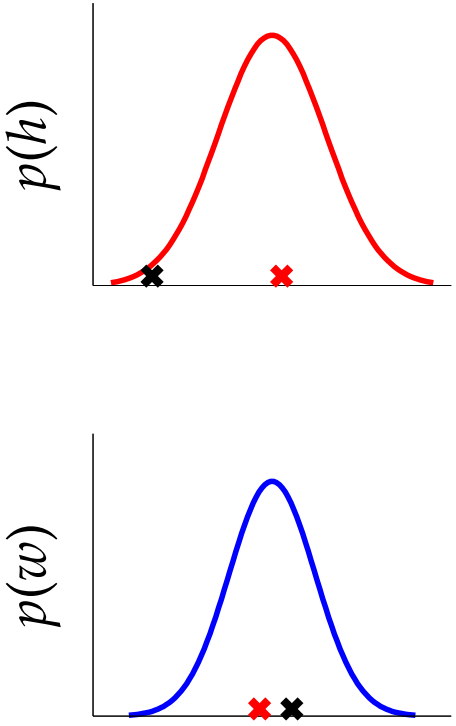
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



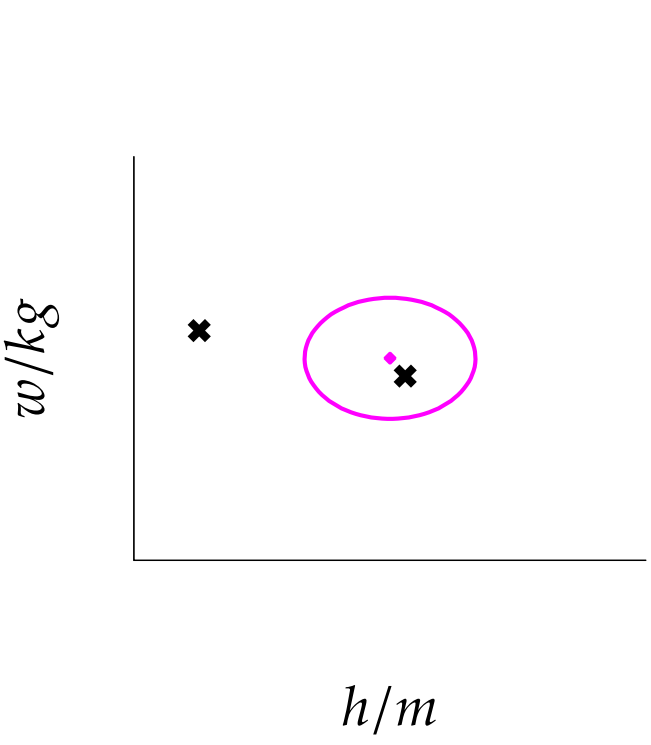


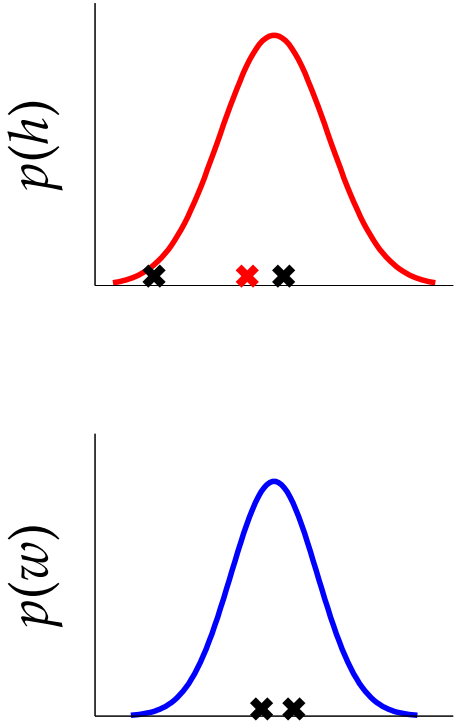
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



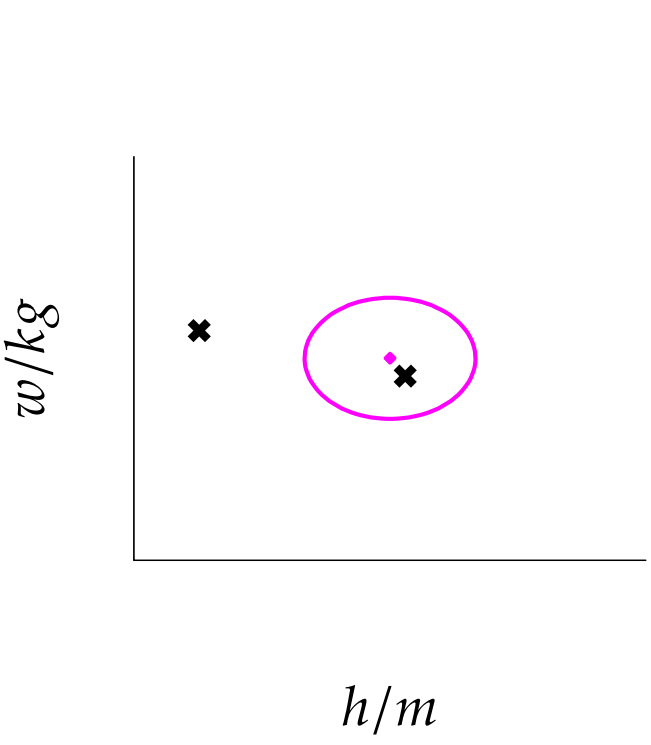


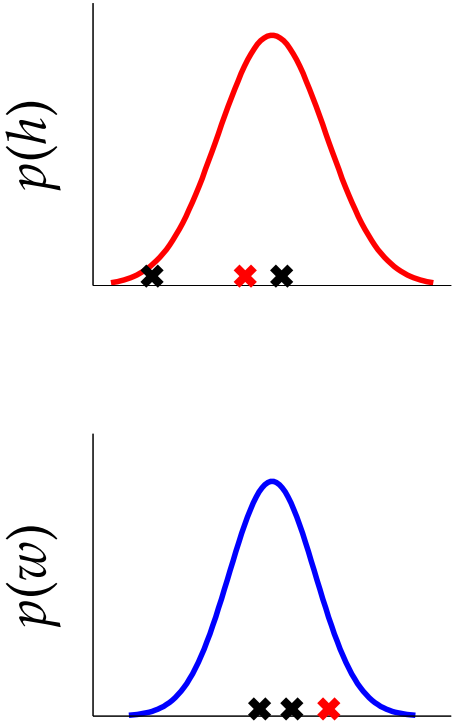
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



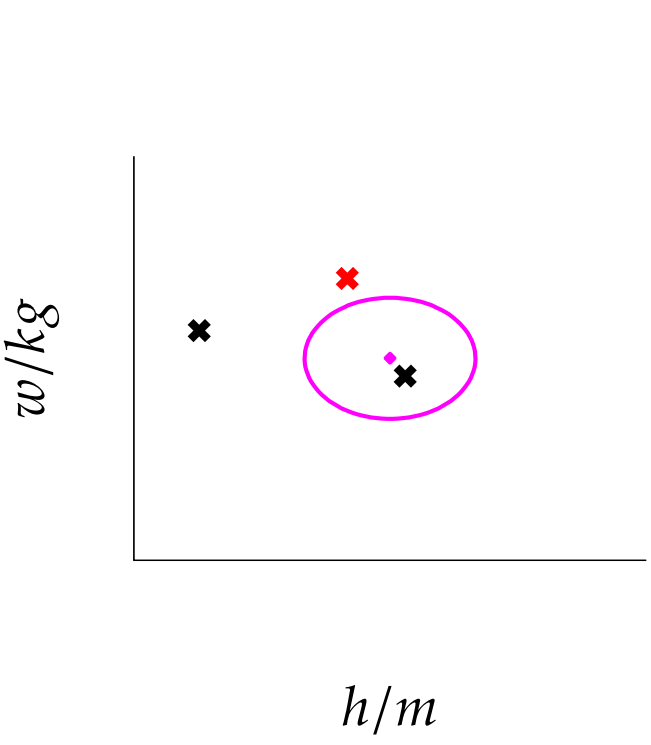


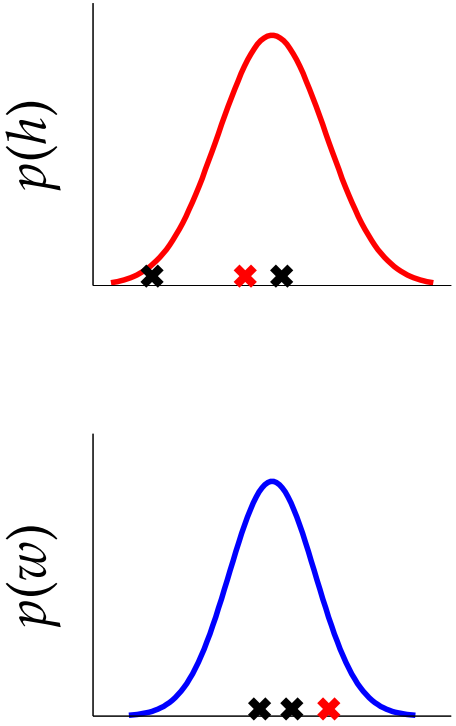
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



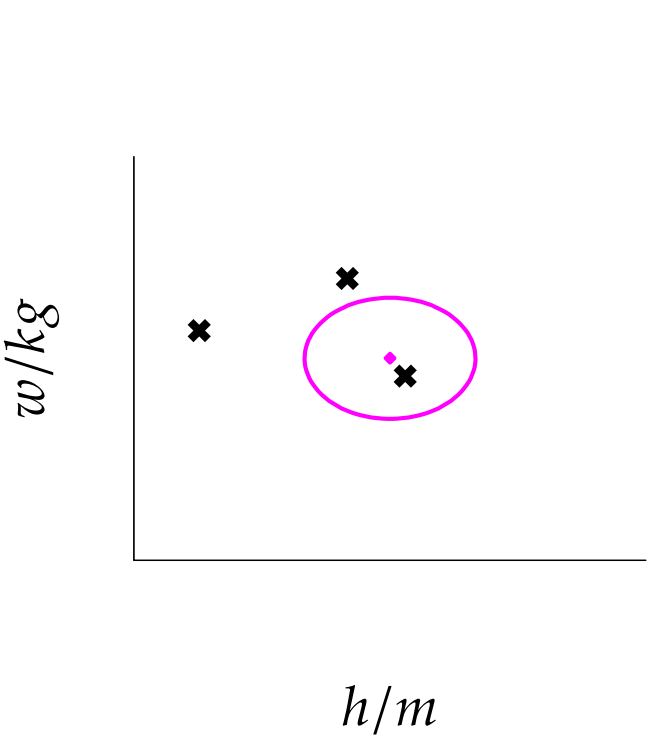


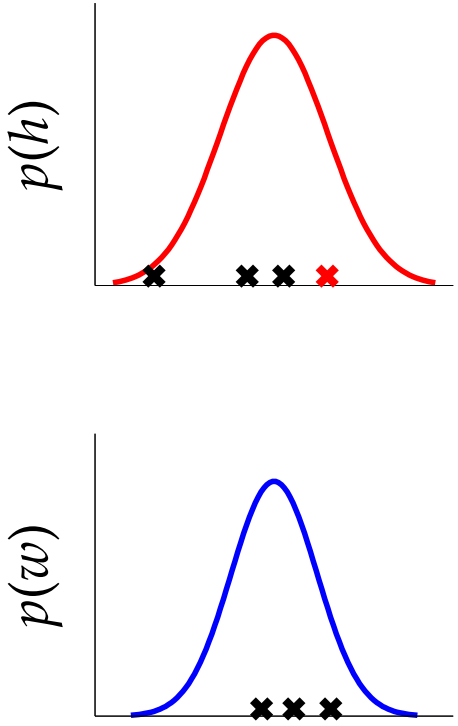
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



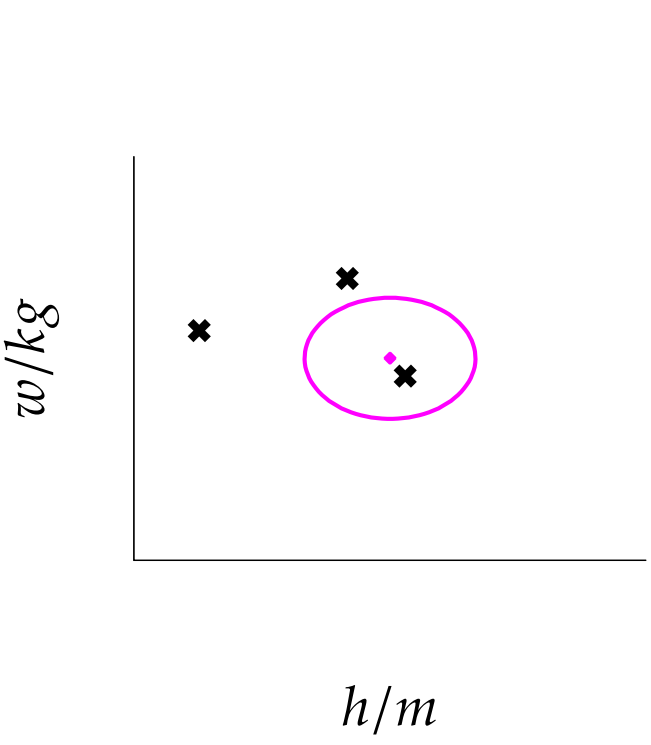


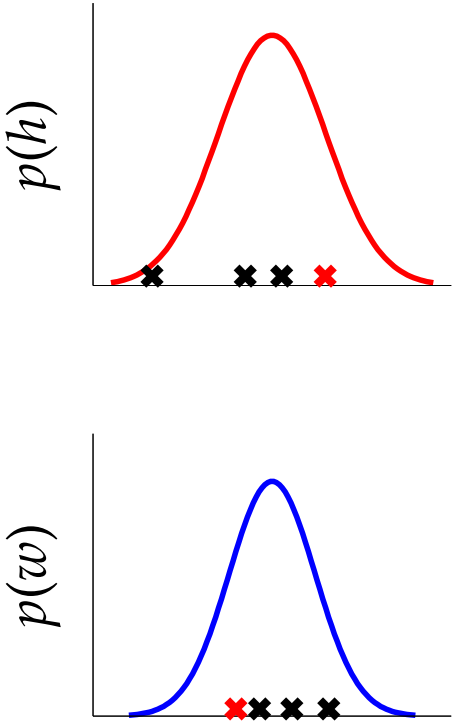
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



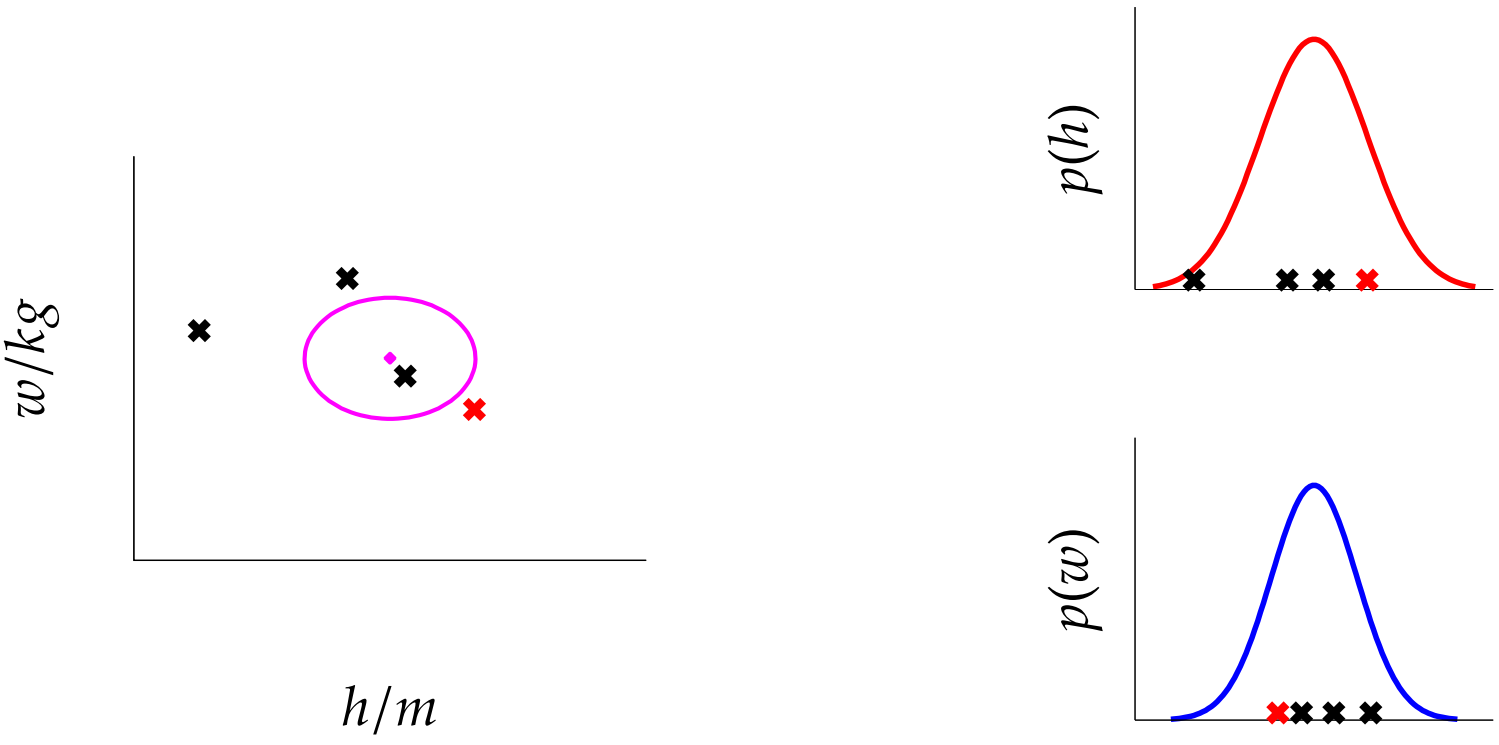


Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution

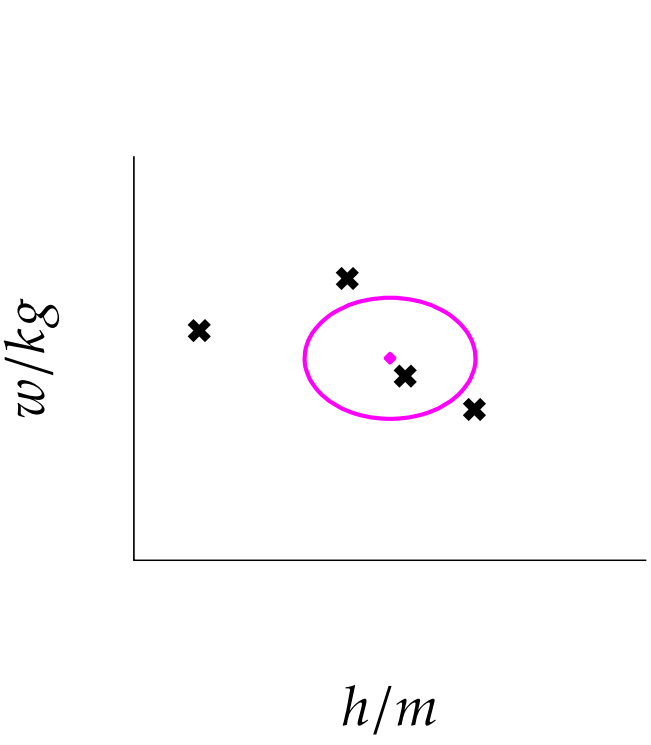


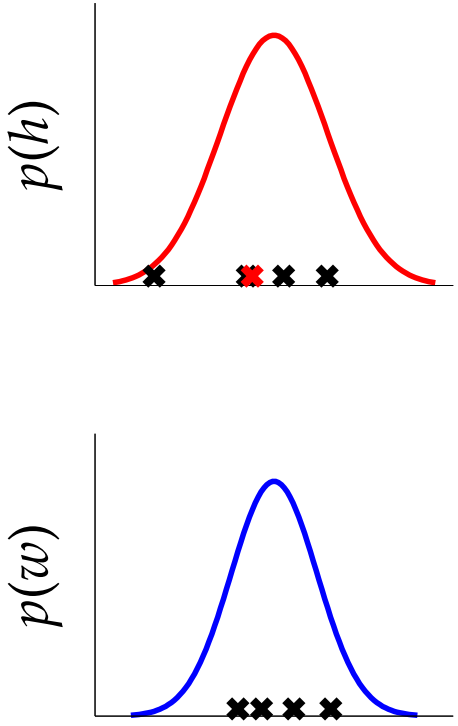
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



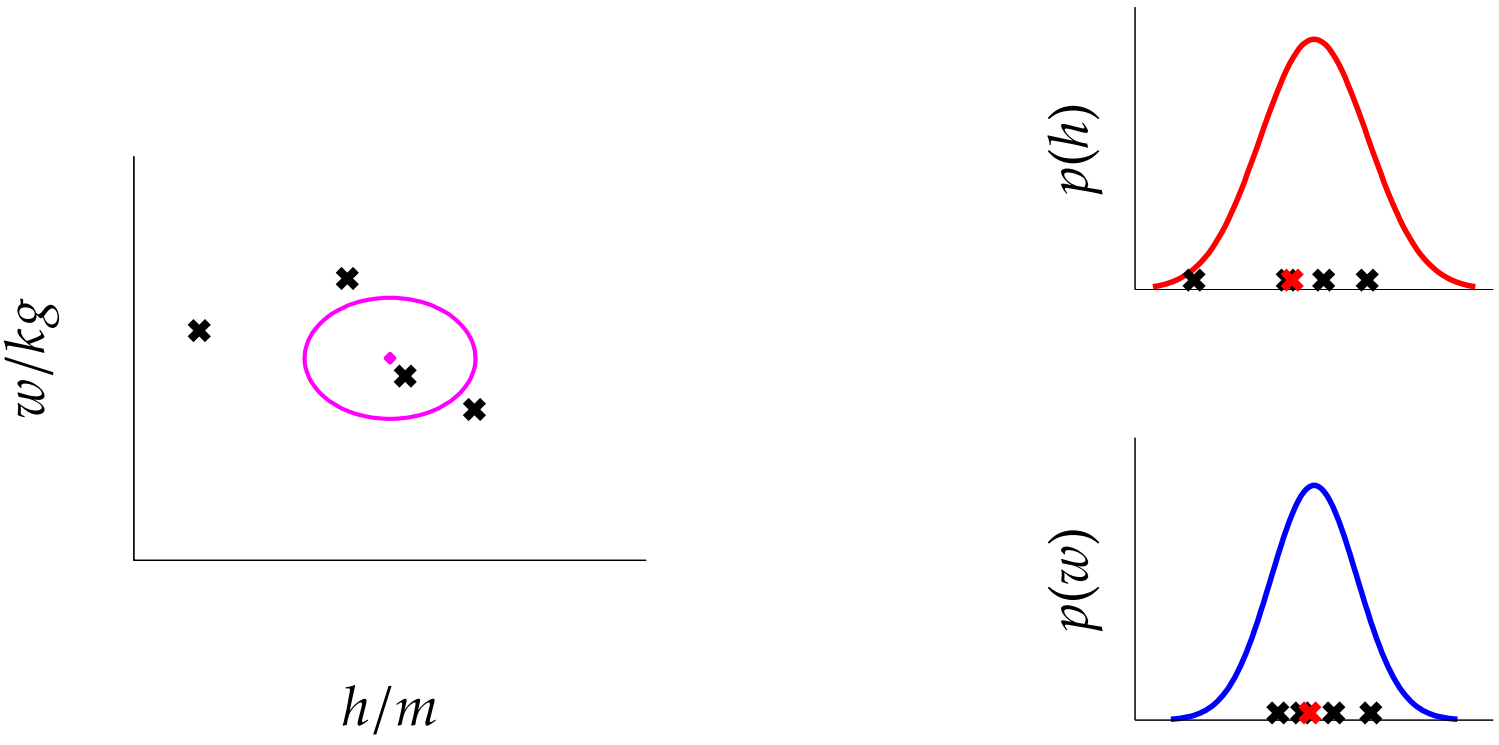


Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution

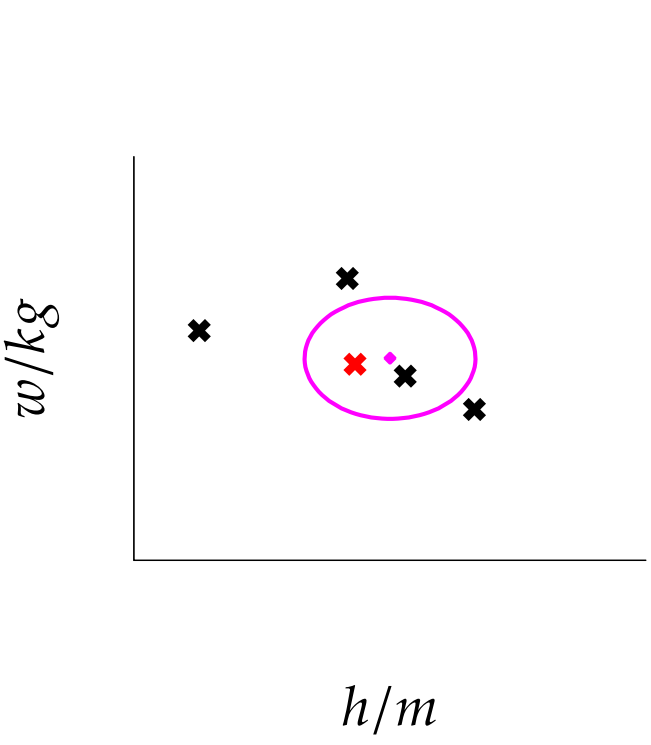


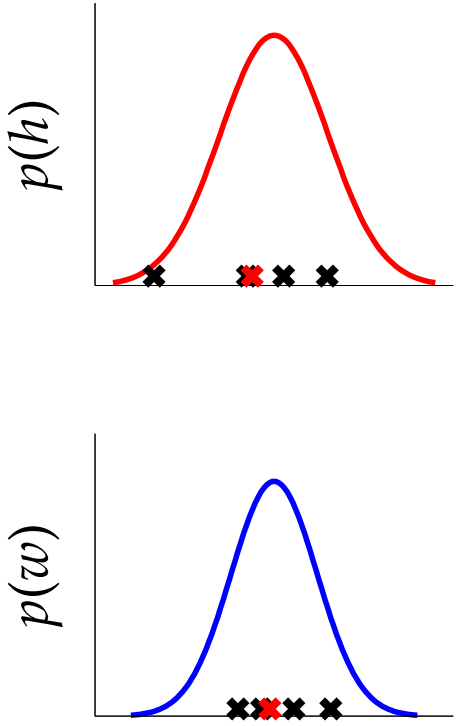
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



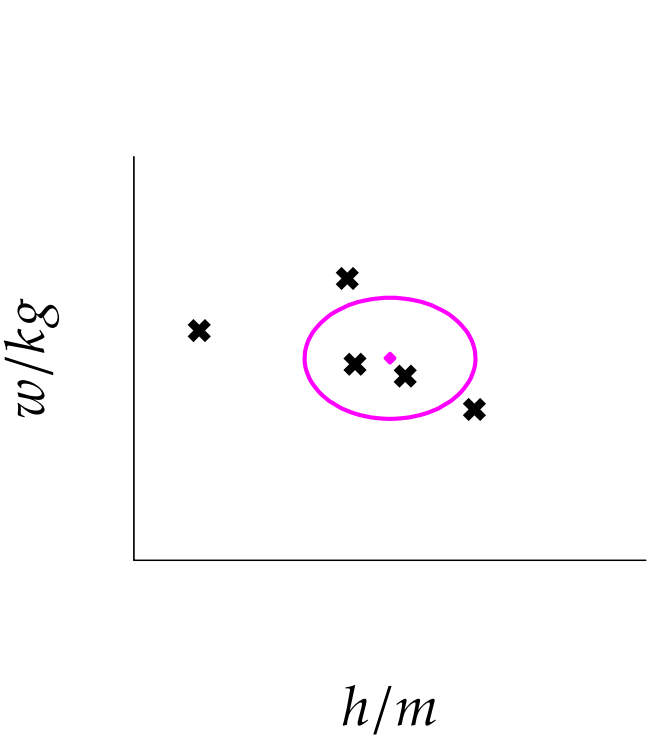


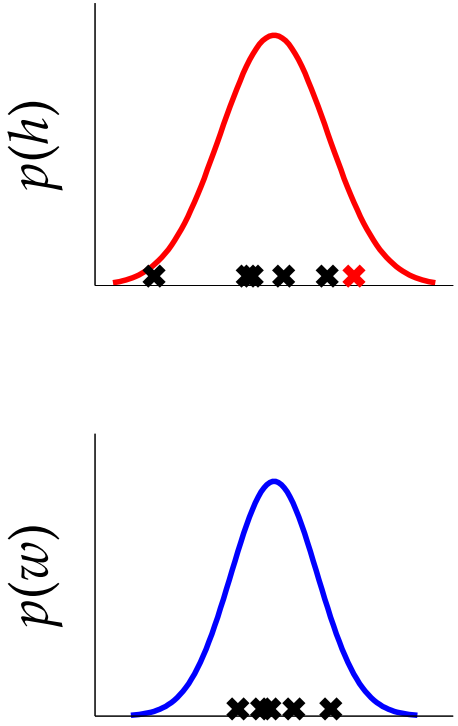
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



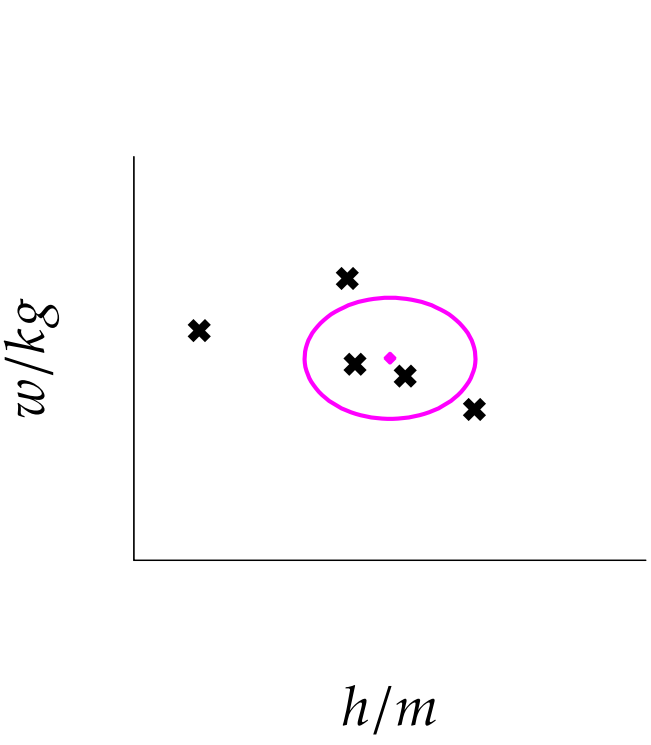


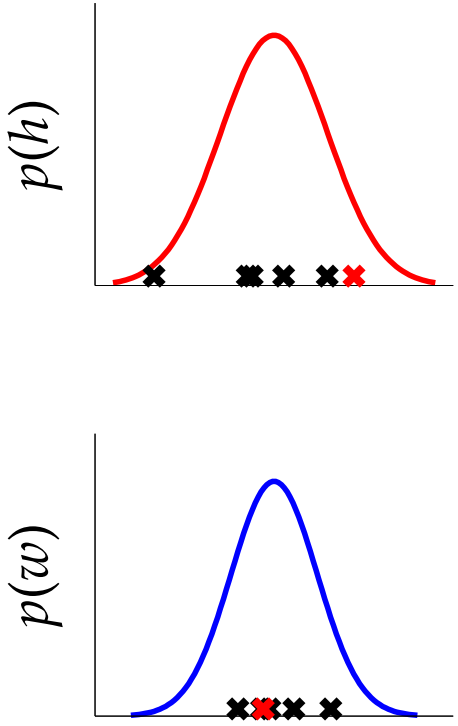
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



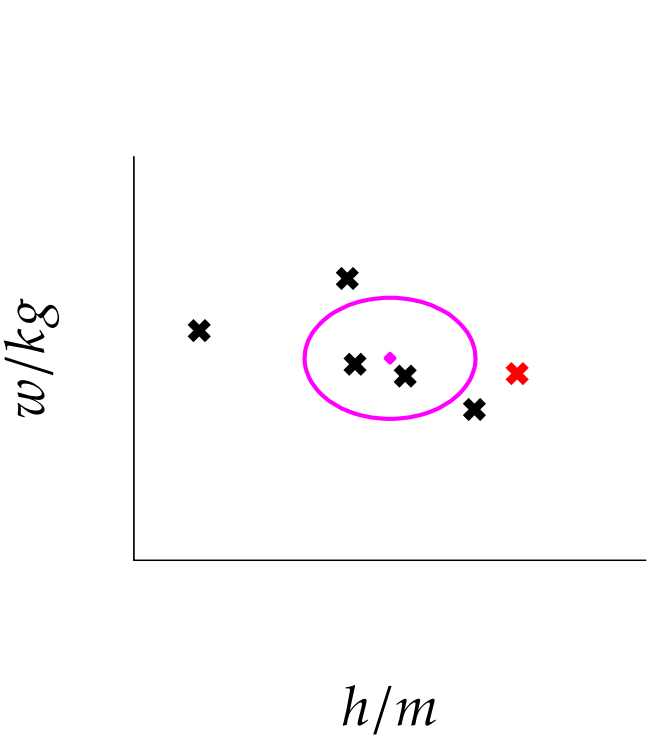


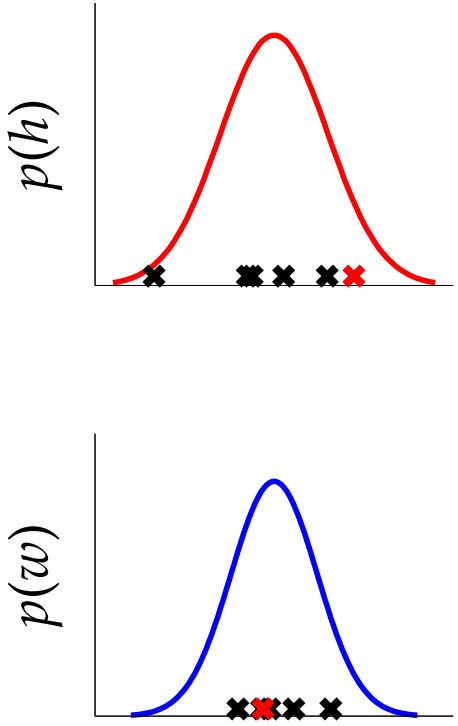
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



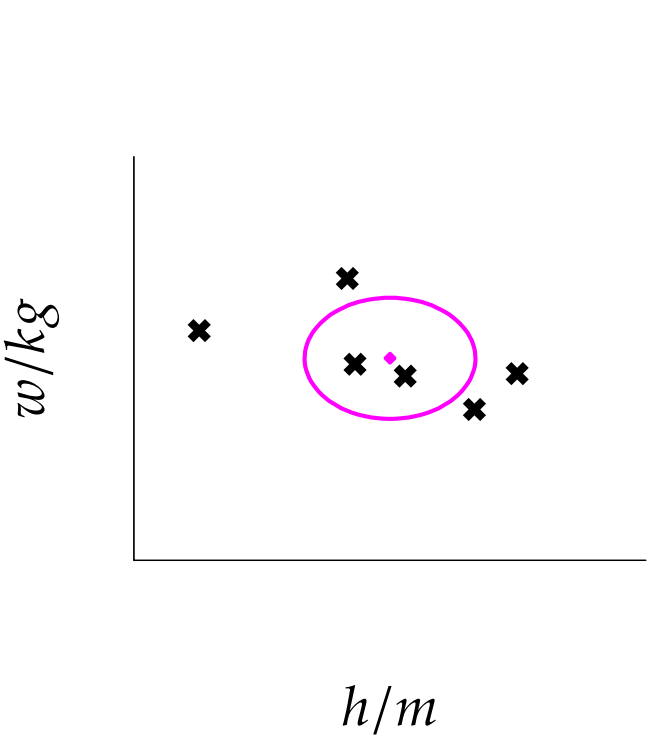


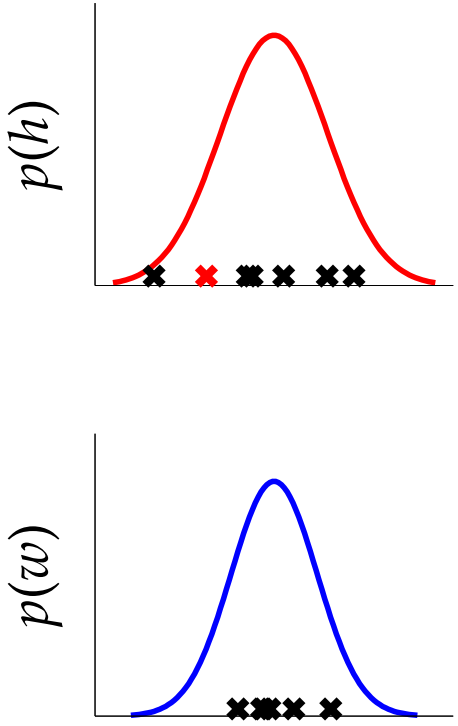
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



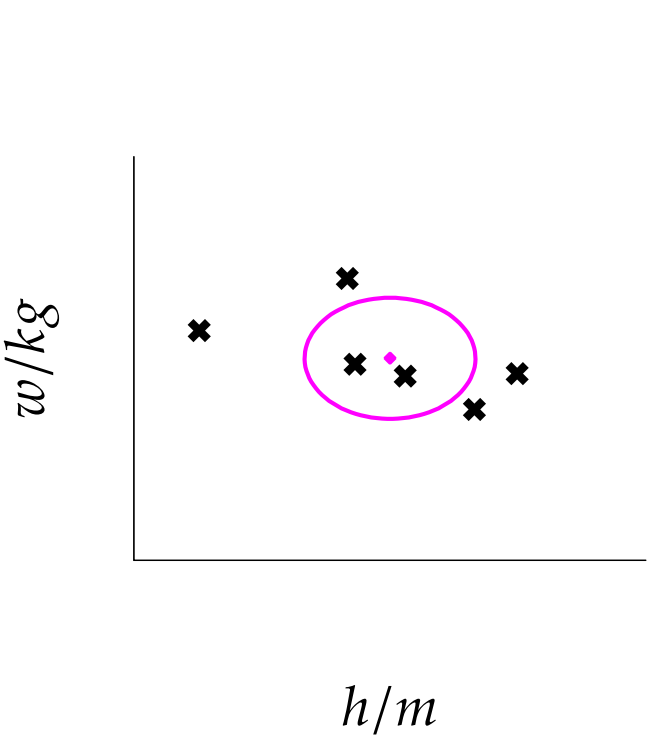


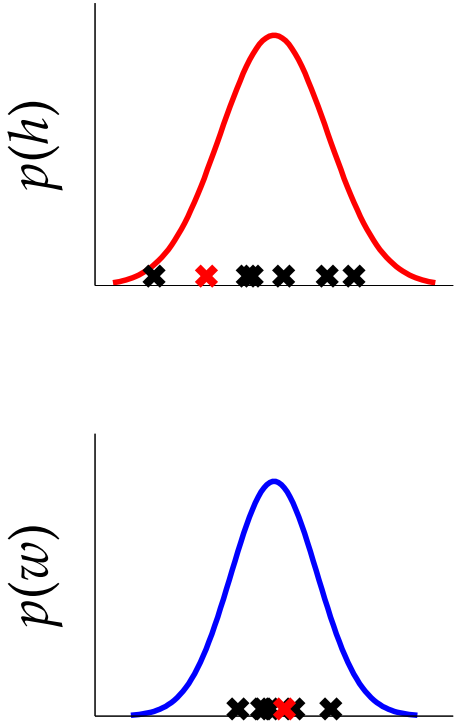
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



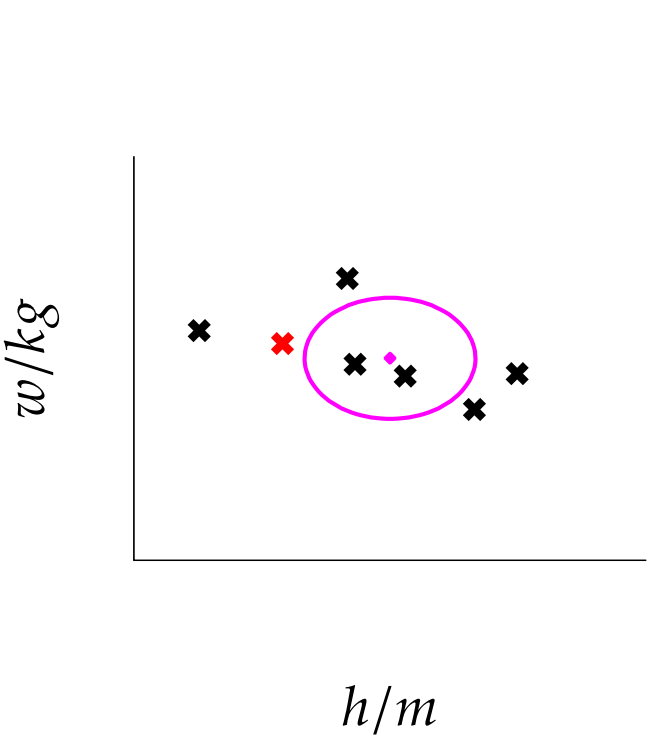


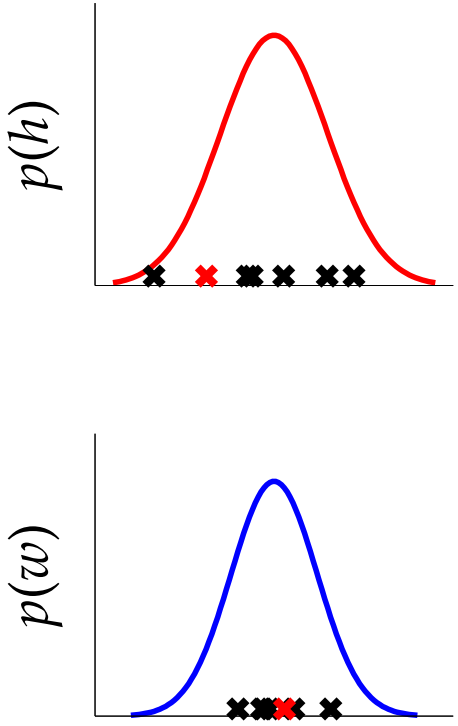
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



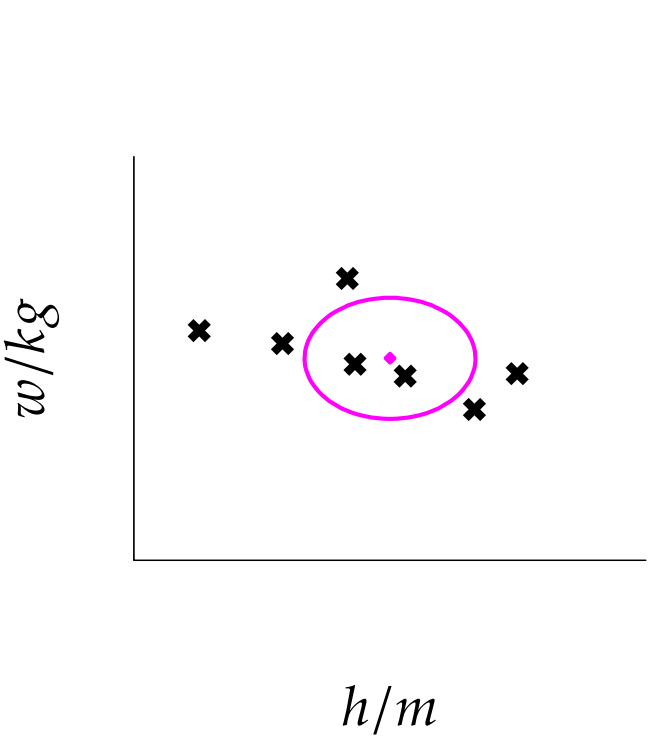


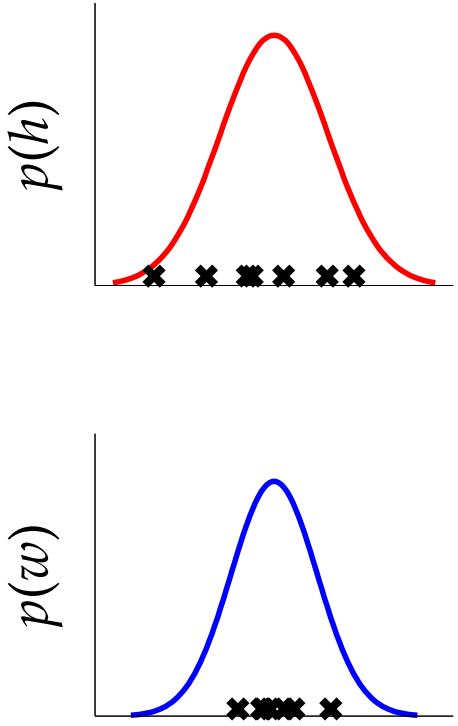
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution





Samples of height and weight

Independence Assumption

- This assumes height and weight are independent.

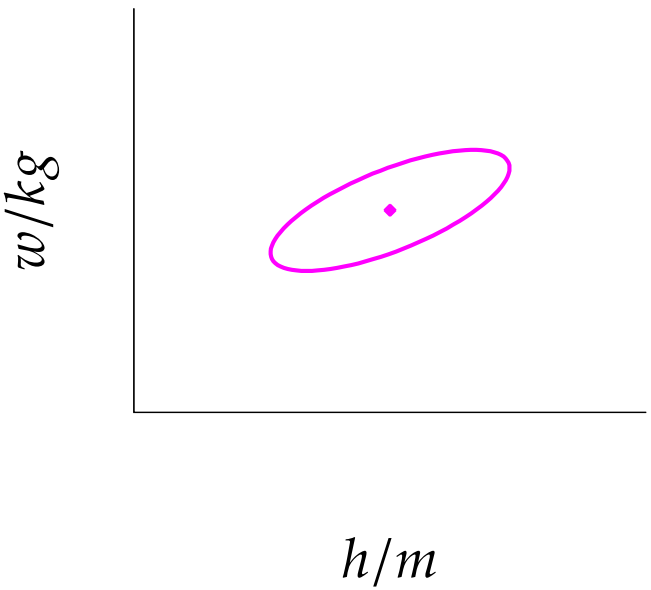
$$p(h, w) = p(h)p(w)$$

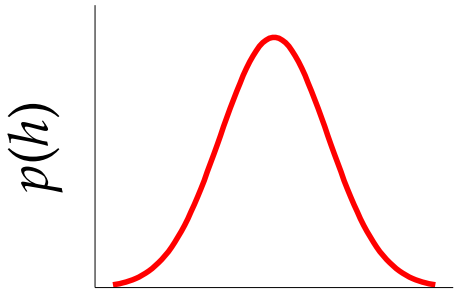
- In reality they are dependent (body mass index) $= \frac{w}{h^2}$.

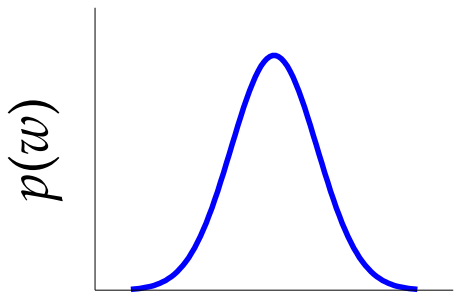
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



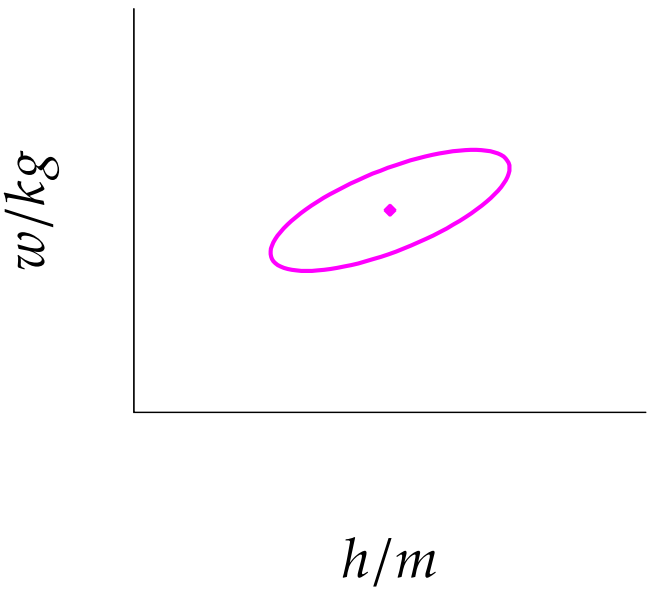


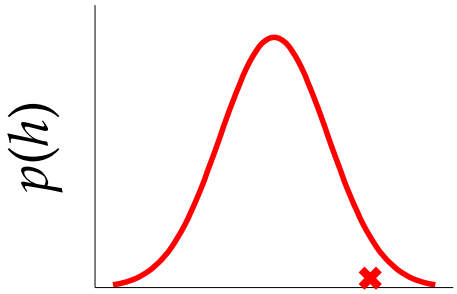


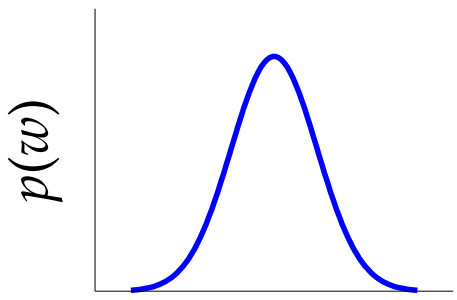
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



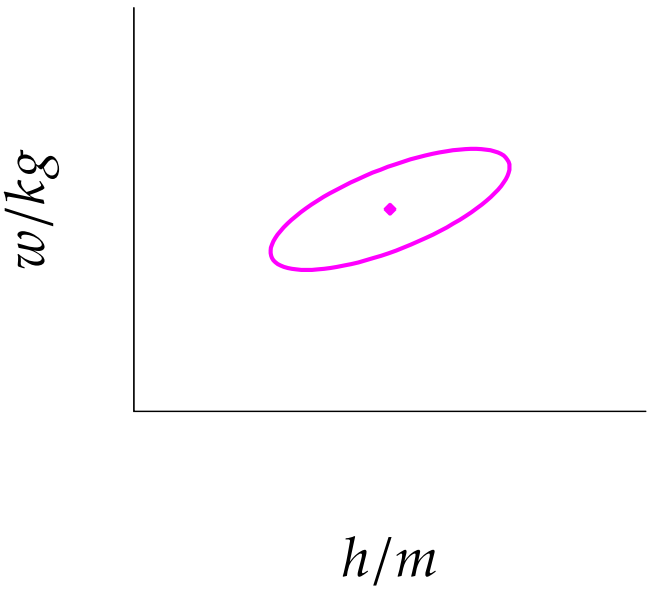


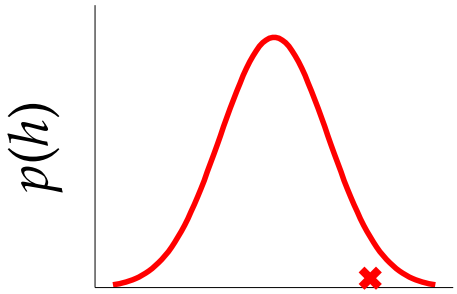


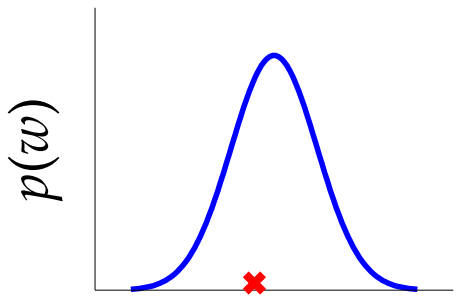
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



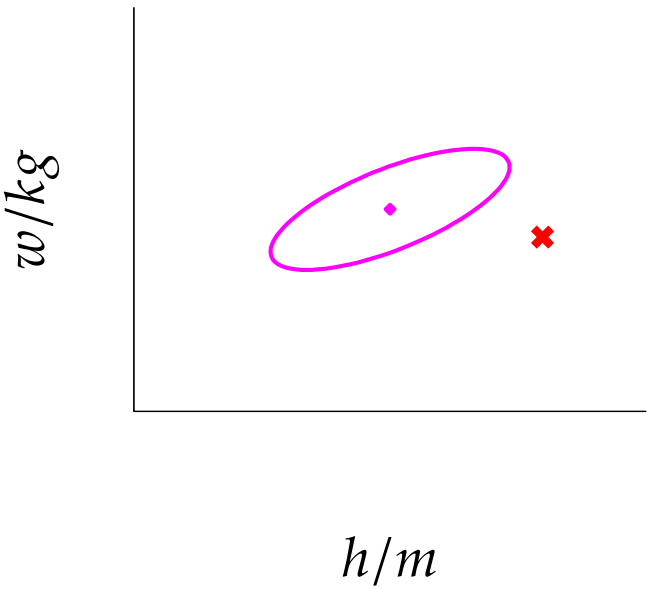


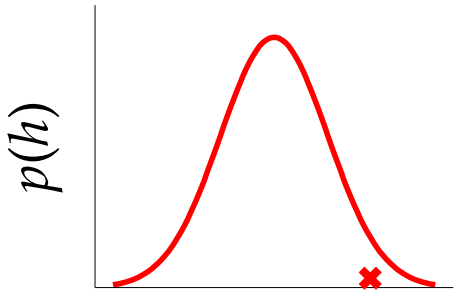


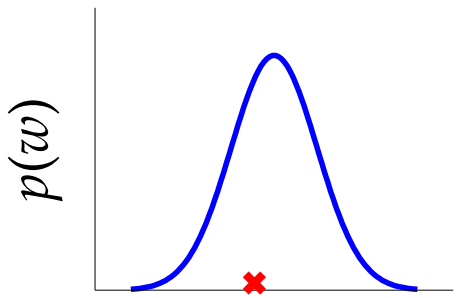
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



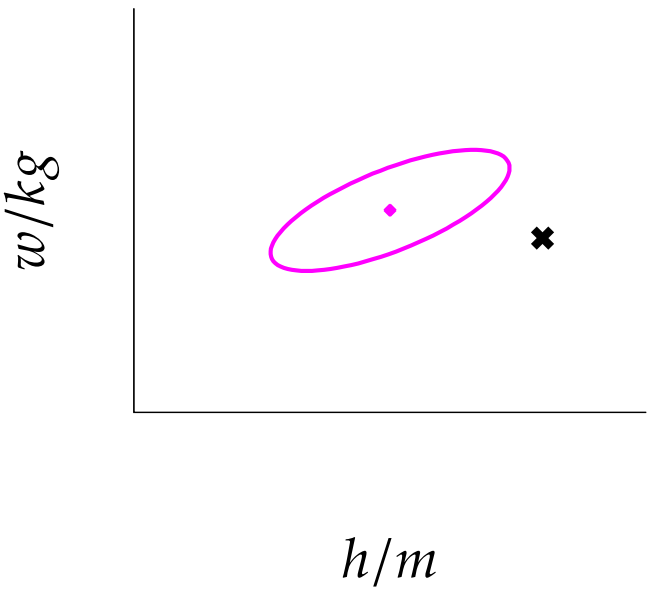


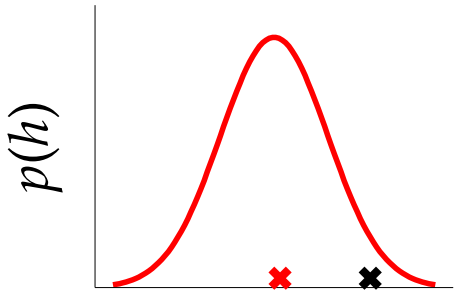


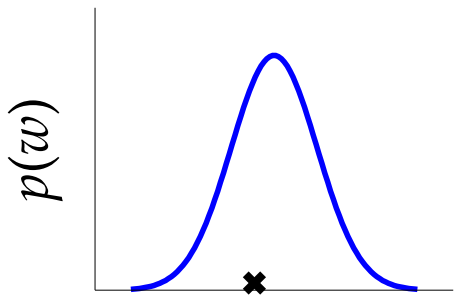
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



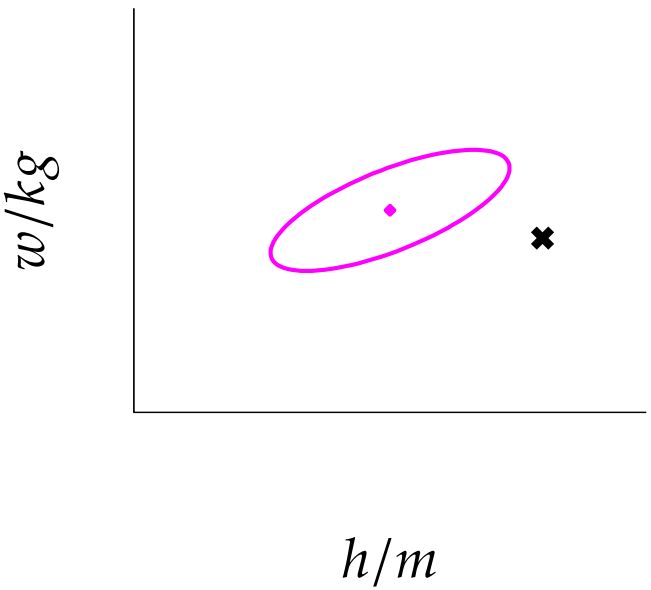


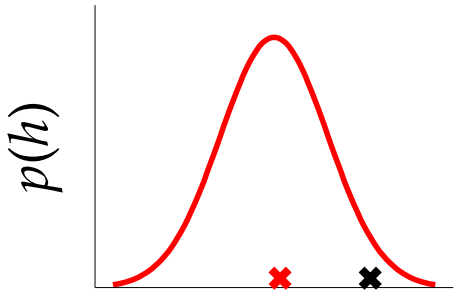


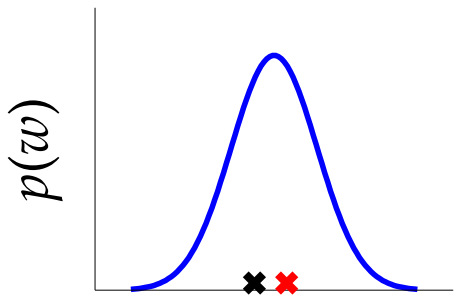
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



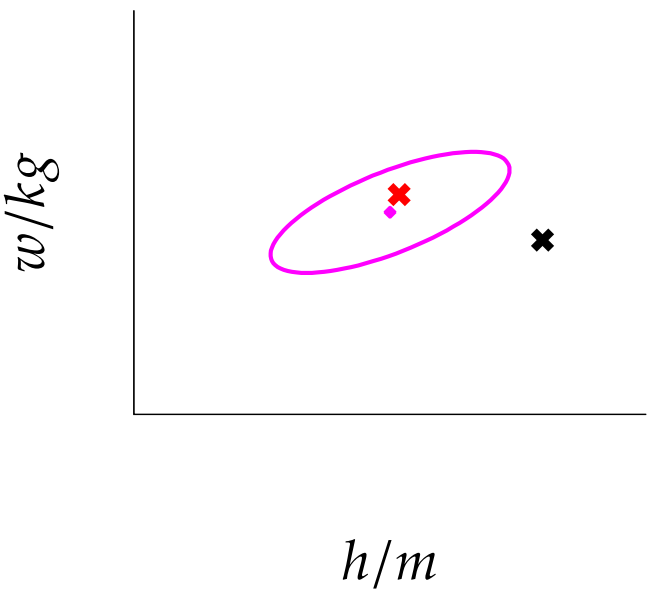


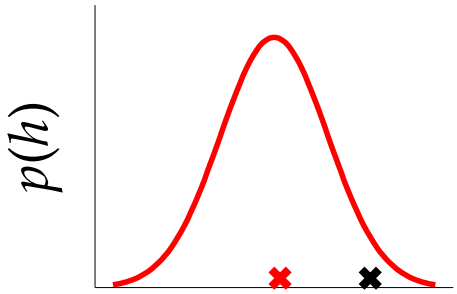


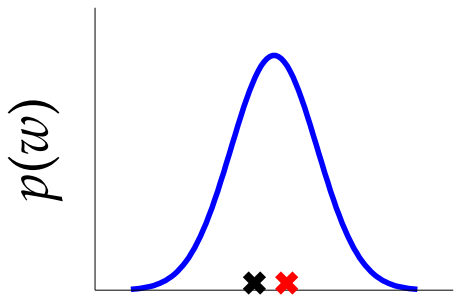
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



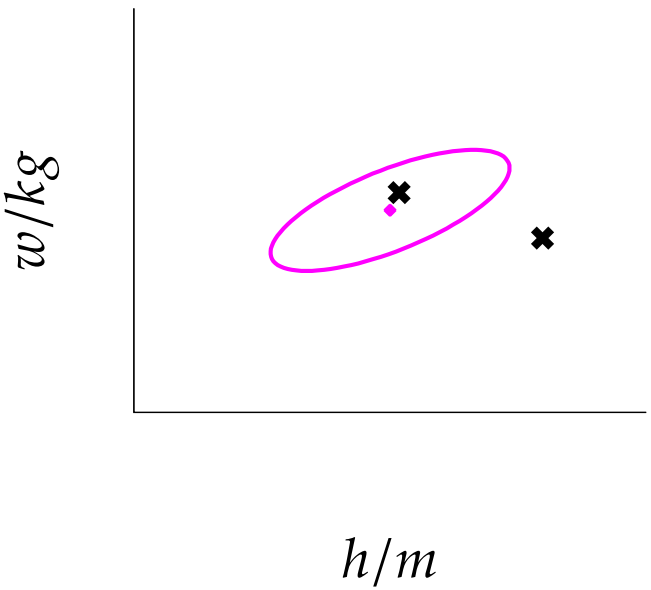


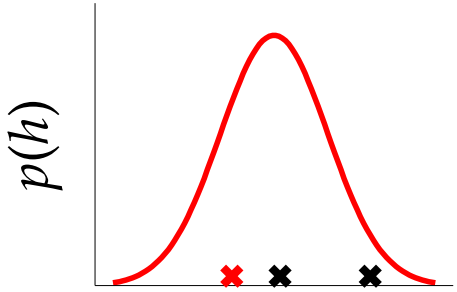


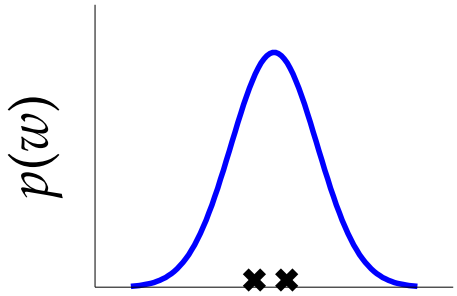
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



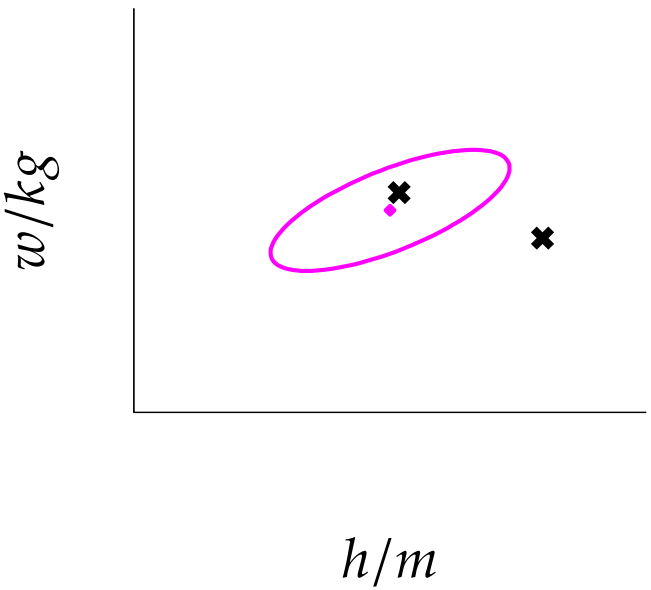


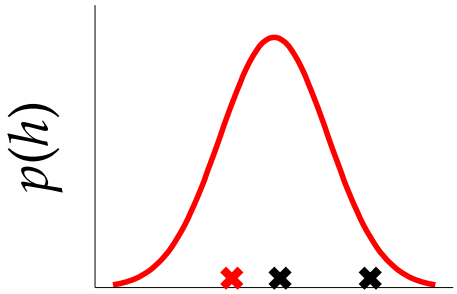


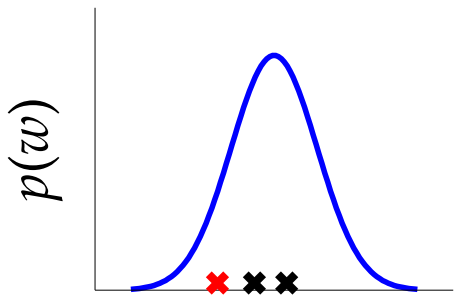
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



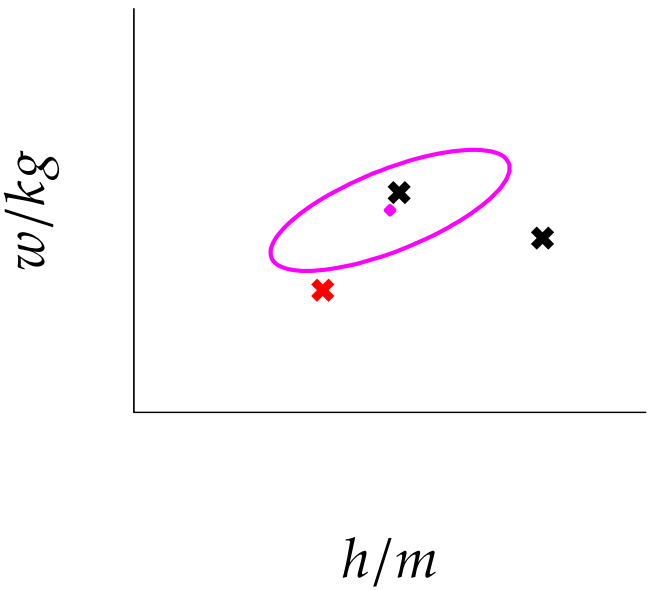


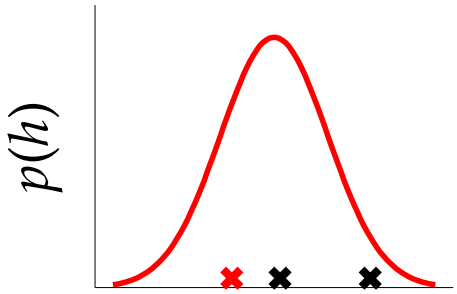


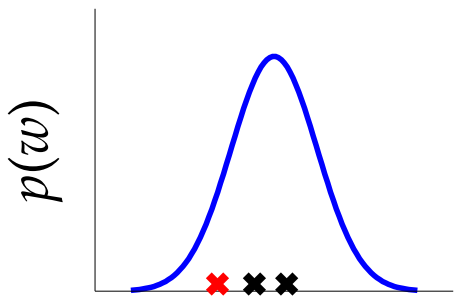
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



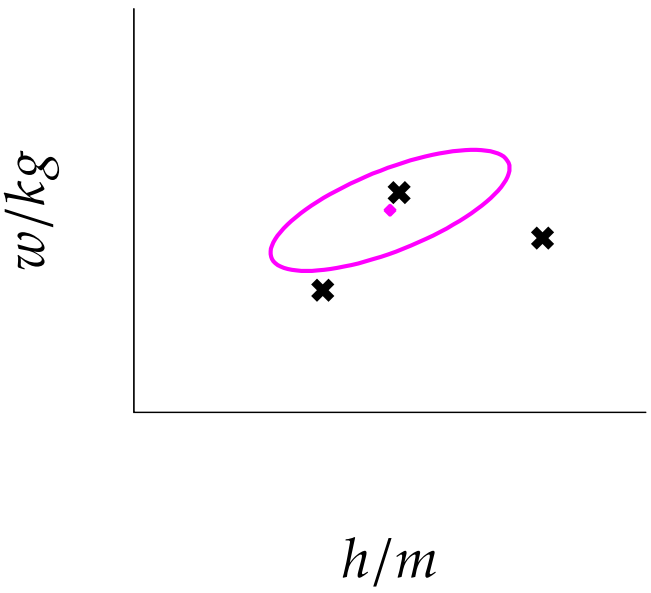


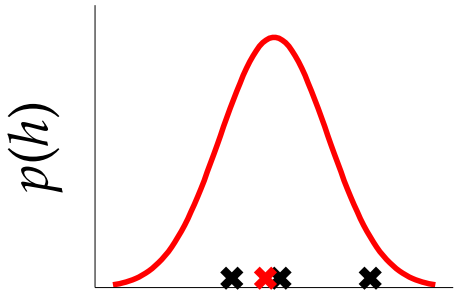


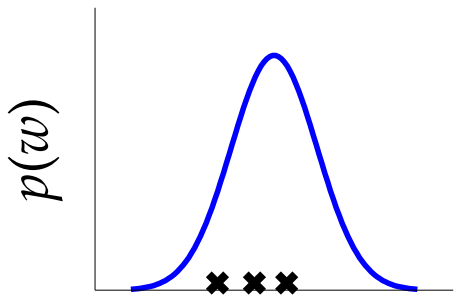
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



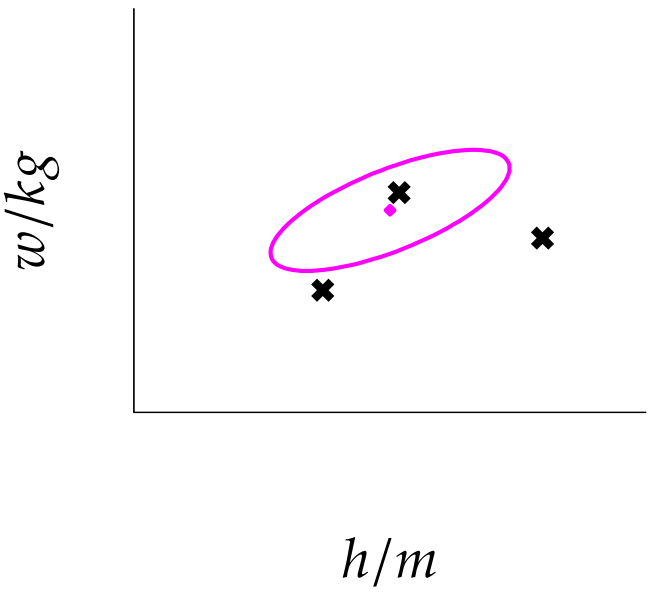


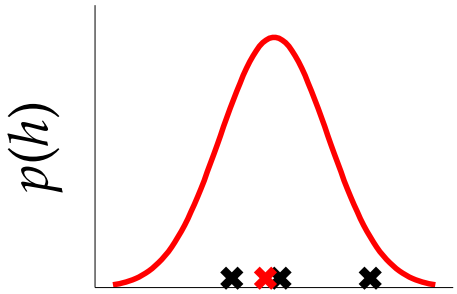


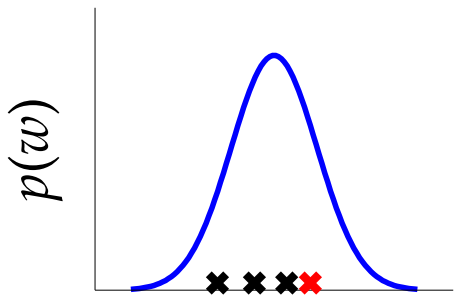
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



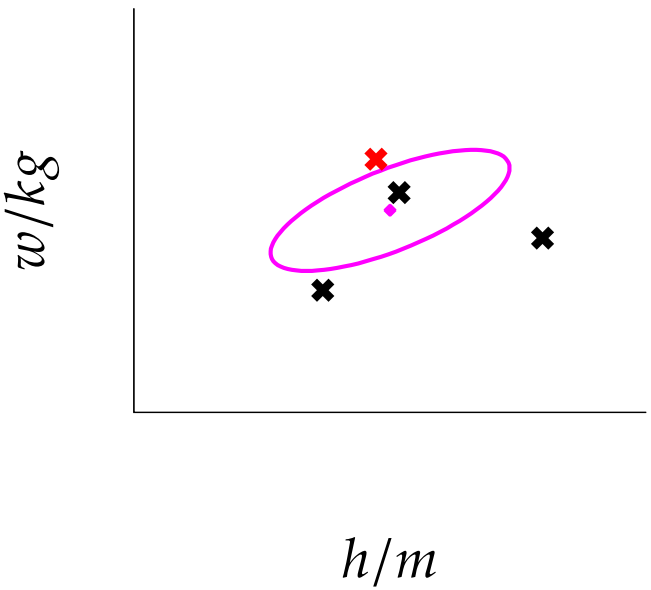


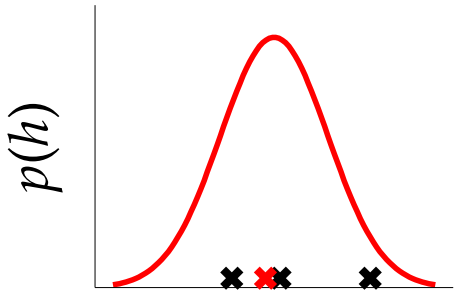


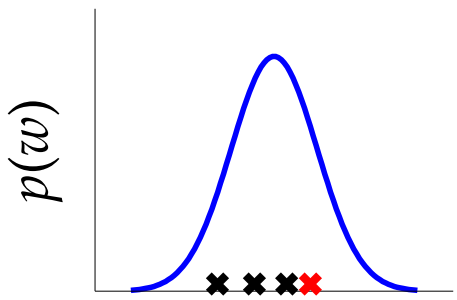
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



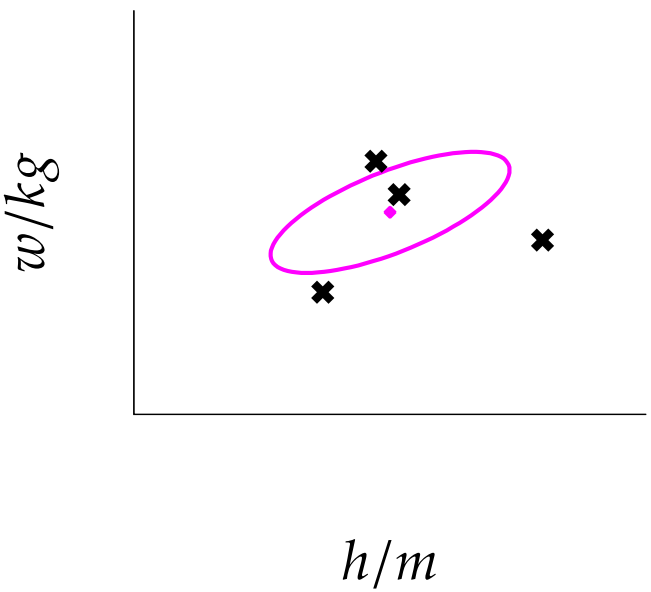


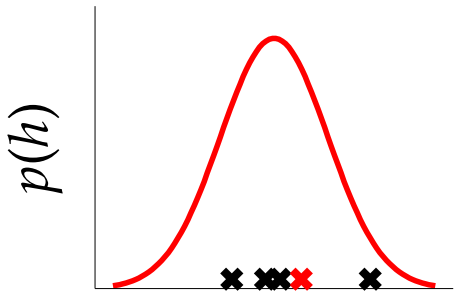


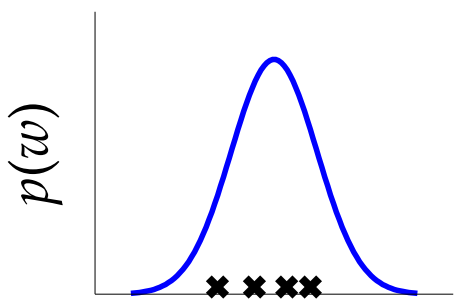
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



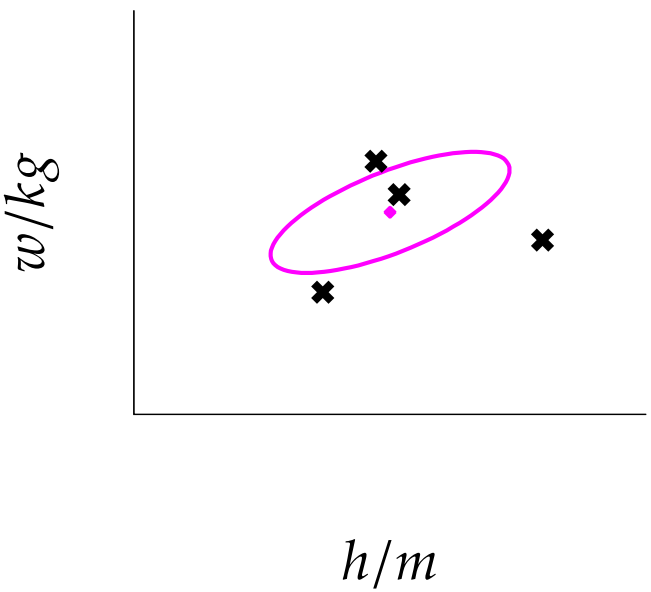


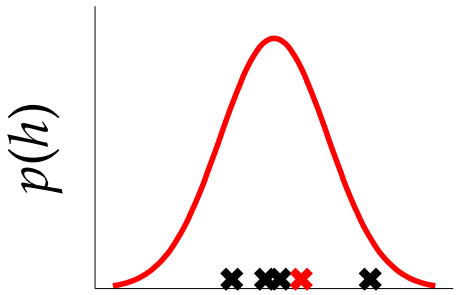


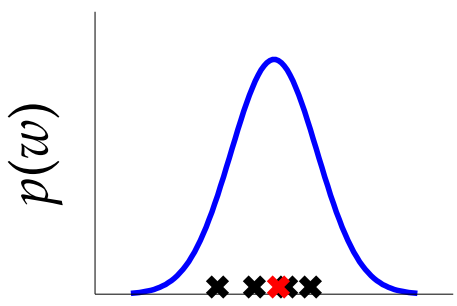
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



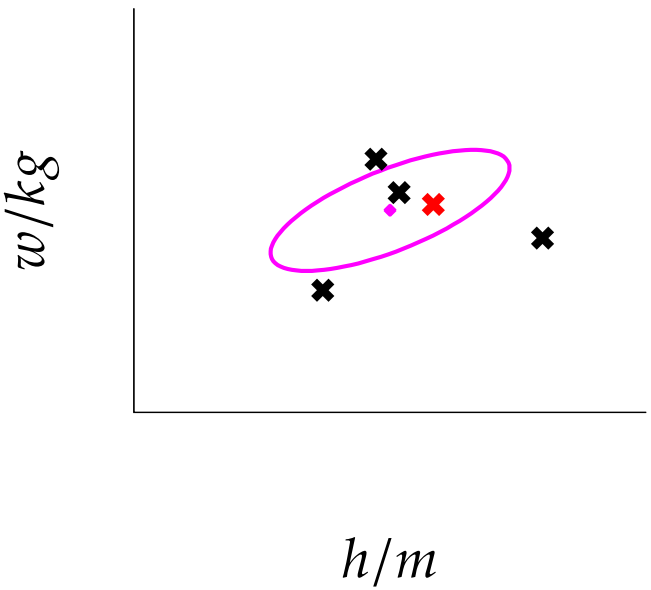


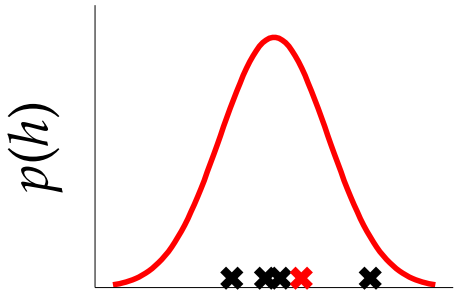


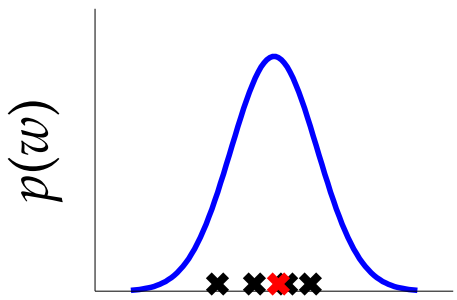
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



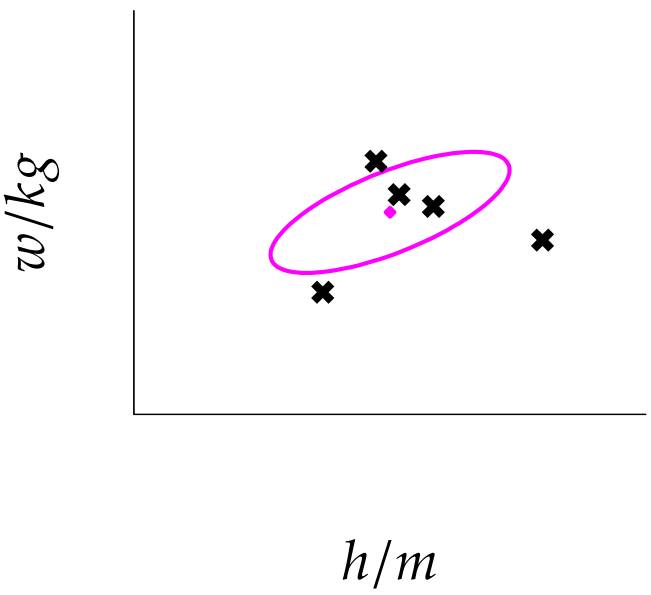


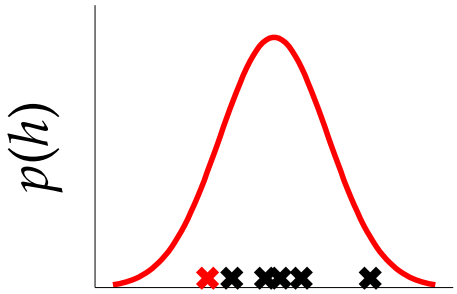


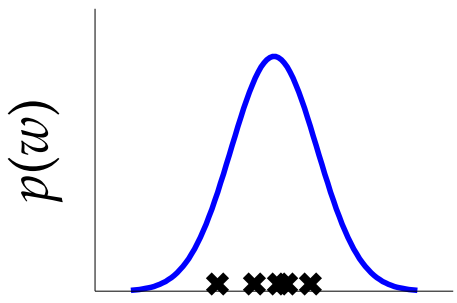
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



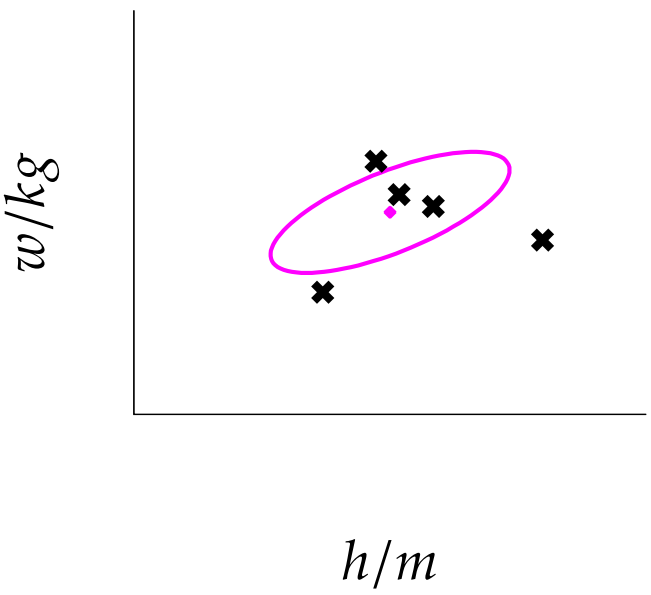


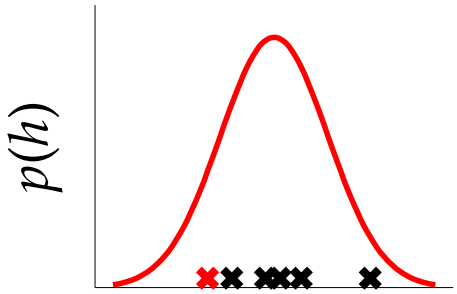


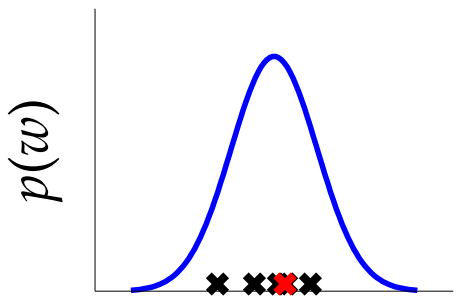
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



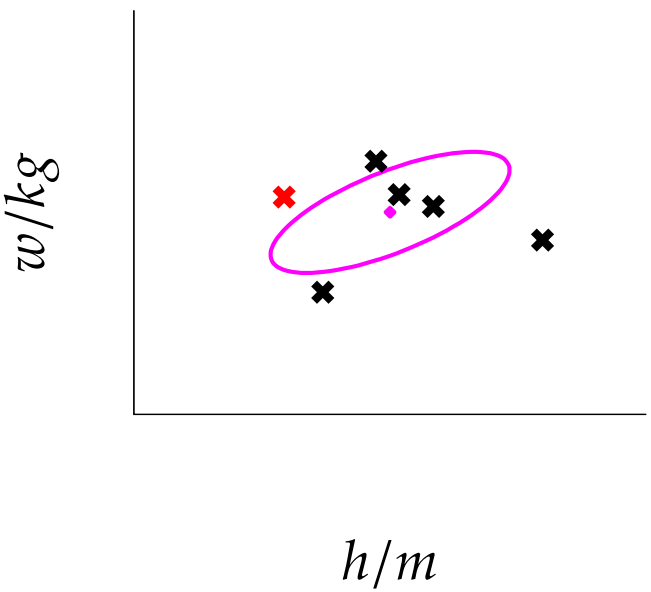


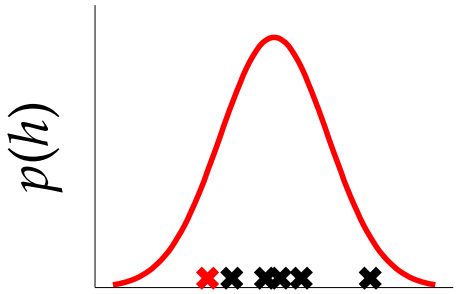


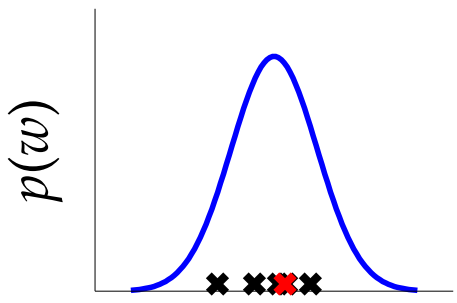
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



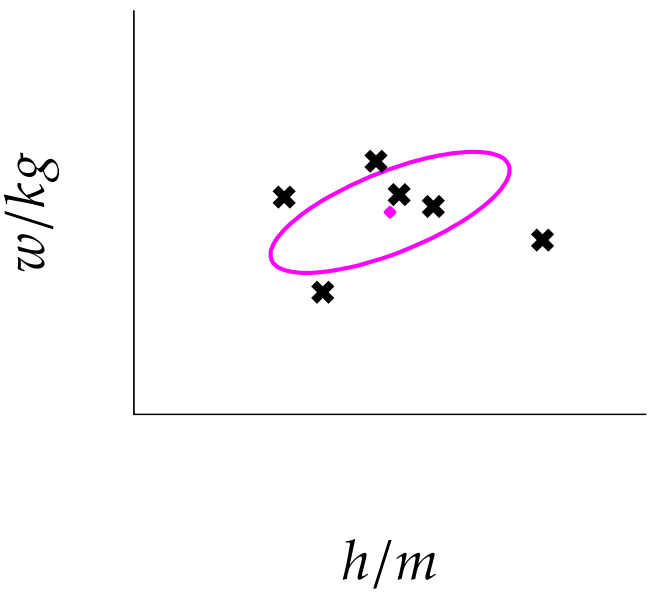


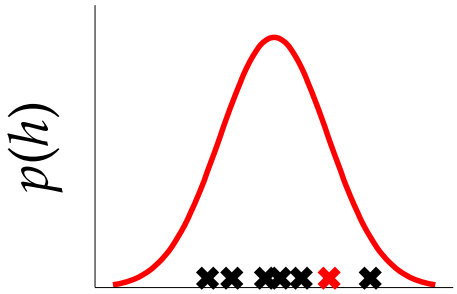


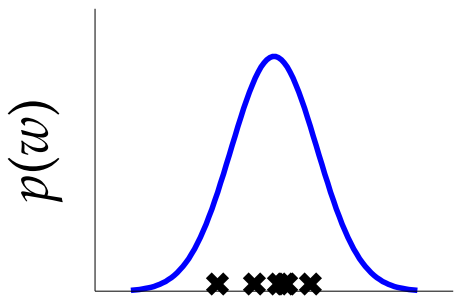
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



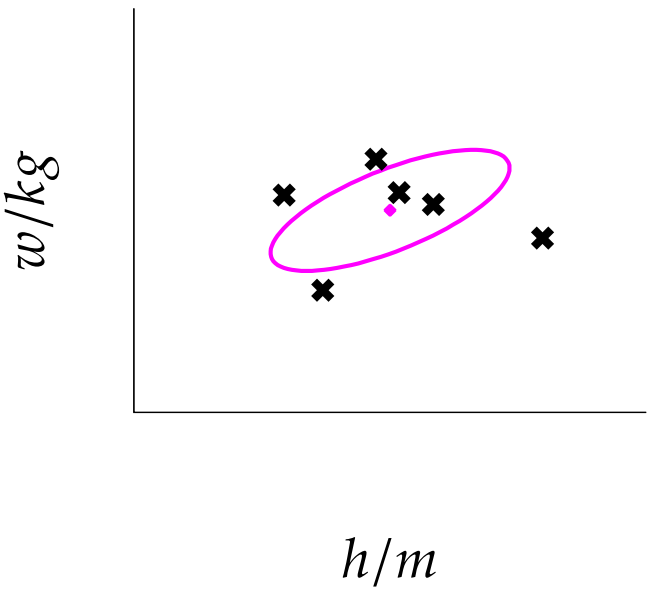


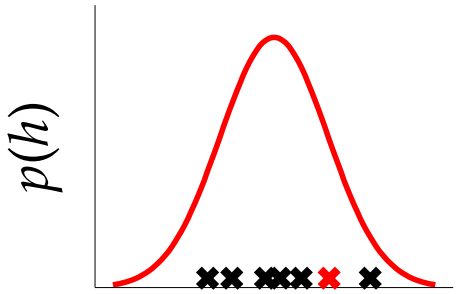


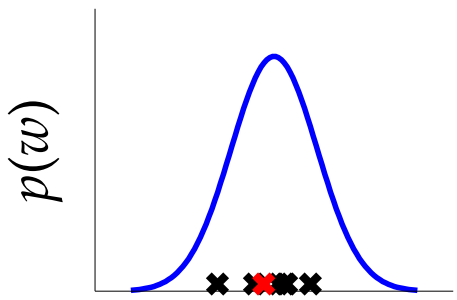
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



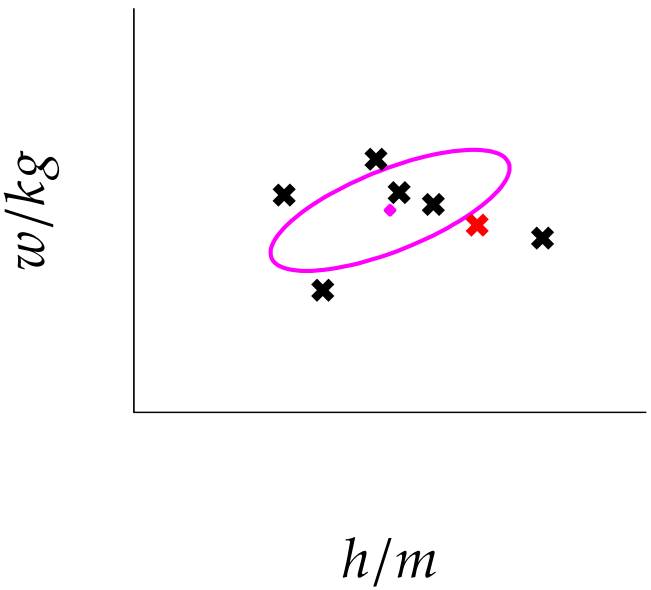


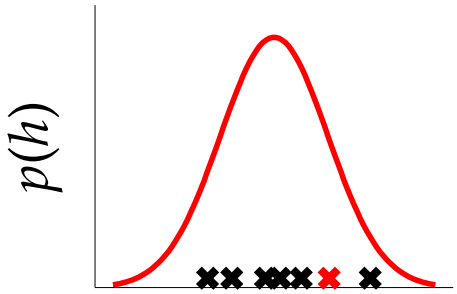


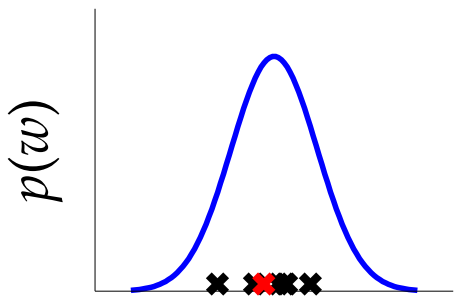
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



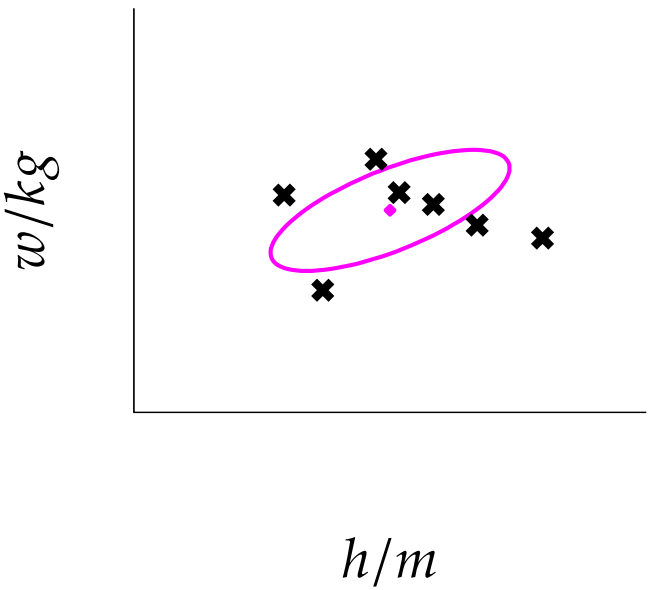


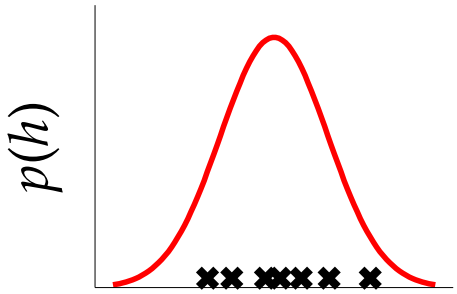


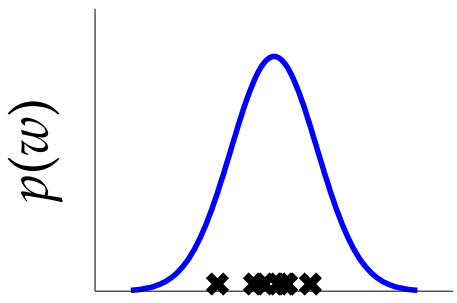
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution







Independent Gaussians

$$p(w, h) = p(w)p(h)$$

Independent Gaussians

$$p(w,h) = \frac{1}{\sqrt{2\pi\sigma_1^2}\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2}\left(\frac{(w-\mu_1)^2}{\sigma_1^2} + \frac{(h-\mu_2)^2}{\sigma_2^2}\right)\right)$$

Independent Gaussians

$$p(w,h) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} w \\ h \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)^\top \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \left(\begin{bmatrix} w \\ h \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)\right)$$

Independent Gaussians

$$p(\mathbf{y}) = \frac{1}{2\pi\,|\mathbf{D}|}\exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\mathbf{D}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix $\mathbf{R}$.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{D}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\mathbf{D}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix $\mathbf{R}$.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{D}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{R}^\top \mathbf{y} - \mathbf{R}^\top \boldsymbol{\mu})^\top \mathbf{D}^{-1}(\mathbf{R}^\top \mathbf{y} - \mathbf{R}^\top \boldsymbol{\mu})\right)$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix $\mathbf{R}$.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{D}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\mathbf{R}\mathbf{D}^{-1}\mathbf{R}^{\top}(\mathbf{y}-\boldsymbol{\mu})\right)$$

this gives a covariance matrix:

$$\mathbf{C}^{-1} = \mathbf{R}\mathbf{D}^{-1}\mathbf{R}^{\top}$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix $\mathbf{R}$.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{C}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\mathbf{C}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$$

this gives a covariance matrix:

$$\mathbf{C} = \mathbf{R}\mathbf{D}\mathbf{R}^{\top}$$

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

2. Scaling a Gaussian leads to a Gaussian.

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Multivariate Consequence

- If

$$\mathbf{x} \sim \mathcal{N}\left(\mu, \Sigma\right)$$

Multivariate Consequence

- If

$$\mathbf{x} \sim \mathcal{N}\left(\mu, \mathbf{\Sigma}\right)$$

- And

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

Multivariate Consequence

- If
$$\mathbf{x} \sim \mathcal{N}\left(\mu, \mathbf{\Sigma}\right)$$
- And
$$\mathbf{y} = \mathbf{W}\mathbf{x}$$
- Then
$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{W}\mu, \mathbf{W}\mathbf{\Sigma}\mathbf{W}^\top\right)$$

Multivariate Regression Likelihood

- Noise corrupted data point

$$y_i = \mathbf{w}^\top \mathbf{x}_{i,:} + \epsilon_i$$

Multivariate Regression Likelihood

- Noise corrupted data point

$$y_i = \mathbf{w}^\top \mathbf{x}_{i,:} + \epsilon_i$$

- Multivariate regression likelihood:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{w}^\top \mathbf{x}_{i,:}\right)^2\right)$$

Multivariate Regression Likelihood

- Noise corrupted data point

$$y_i = \mathbf{w}^\top \mathbf{x}_{i,:} + \epsilon_i$$

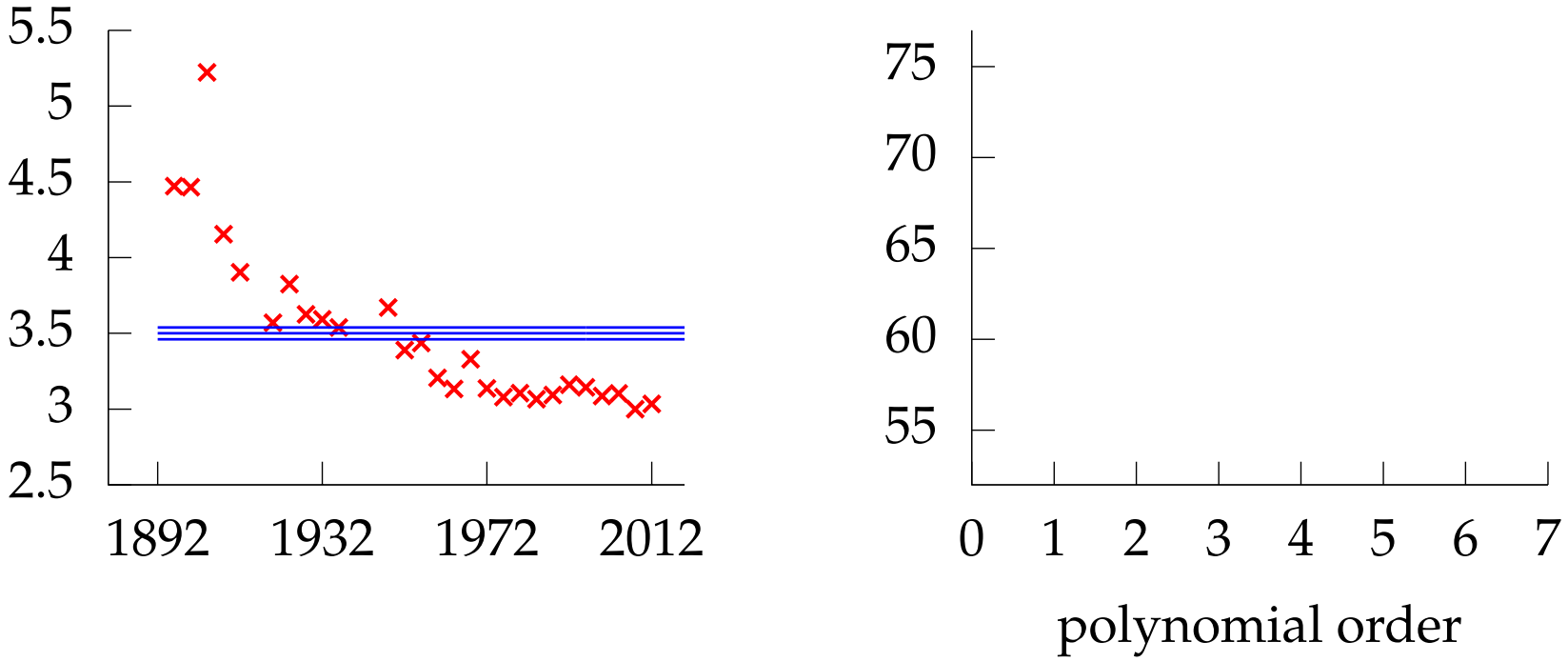
- Multivariate regression likelihood:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{w}^\top \mathbf{x}_{i,:}\right)^2\right)$$

- Now use a multivariate Gaussian prior:

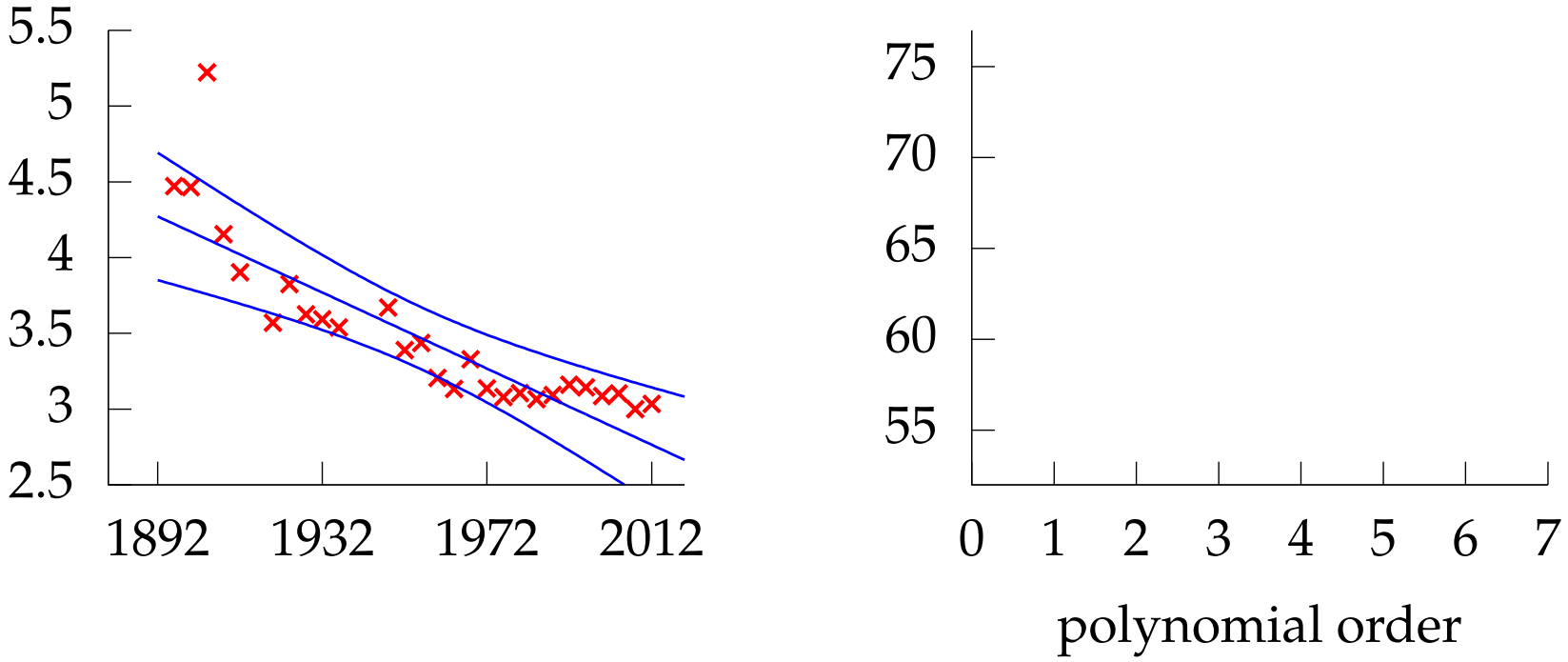
$$p(\mathbf{w}) = \frac{1}{(2\pi\alpha)^{\frac{p}{2}}} \exp\left(-\frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w}\right)$$

Polynomial Fits to Olympics Data



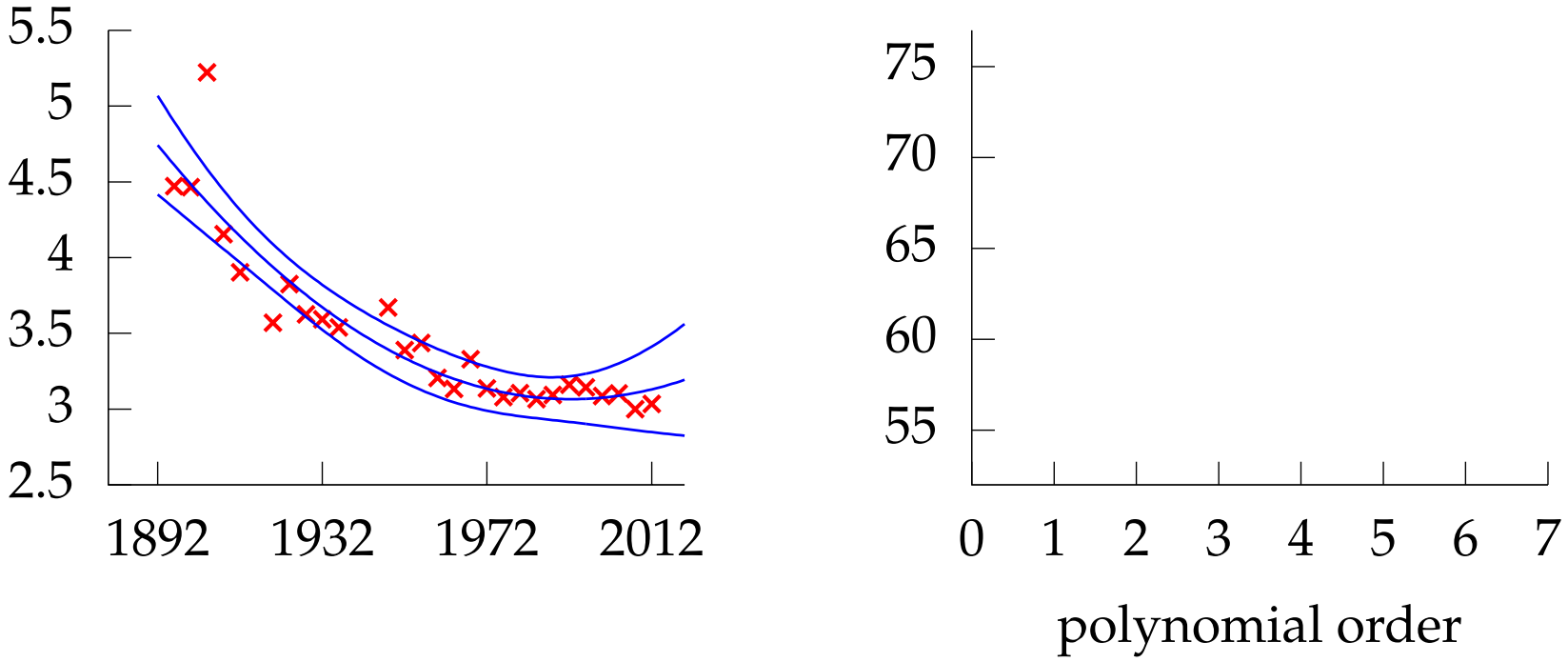
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 0, model error 29.757, $\sigma^2 = 0.286$, $\sigma = 0.535$.

Polynomial Fits to Olympics Data



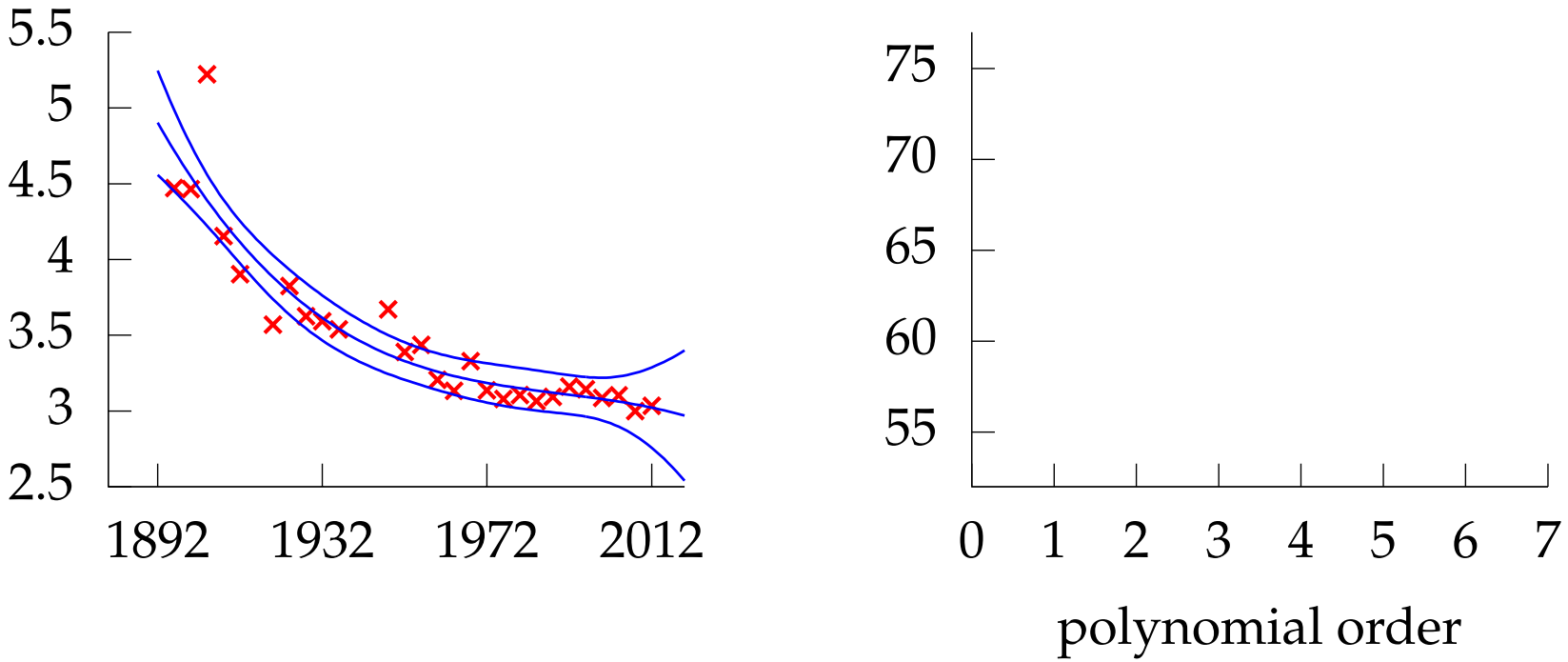
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 1, model error 14.942, $\sigma^2 = 0.0749$, $\sigma = 0.274$.

Polynomial Fits to Olympics Data



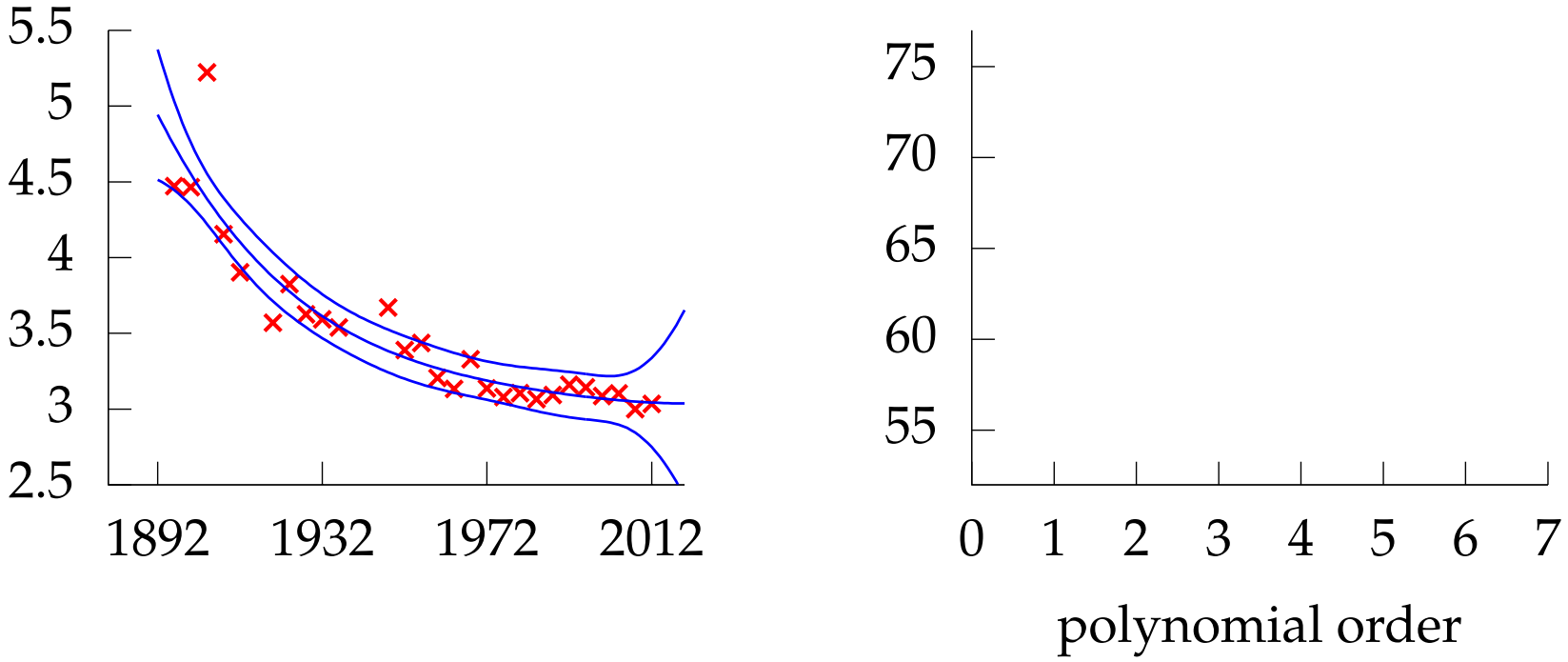
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 2, model error 9.7206, $\sigma^2 = 0.0427$, $\sigma = 0.207$.

Polynomial Fits to Olympics Data



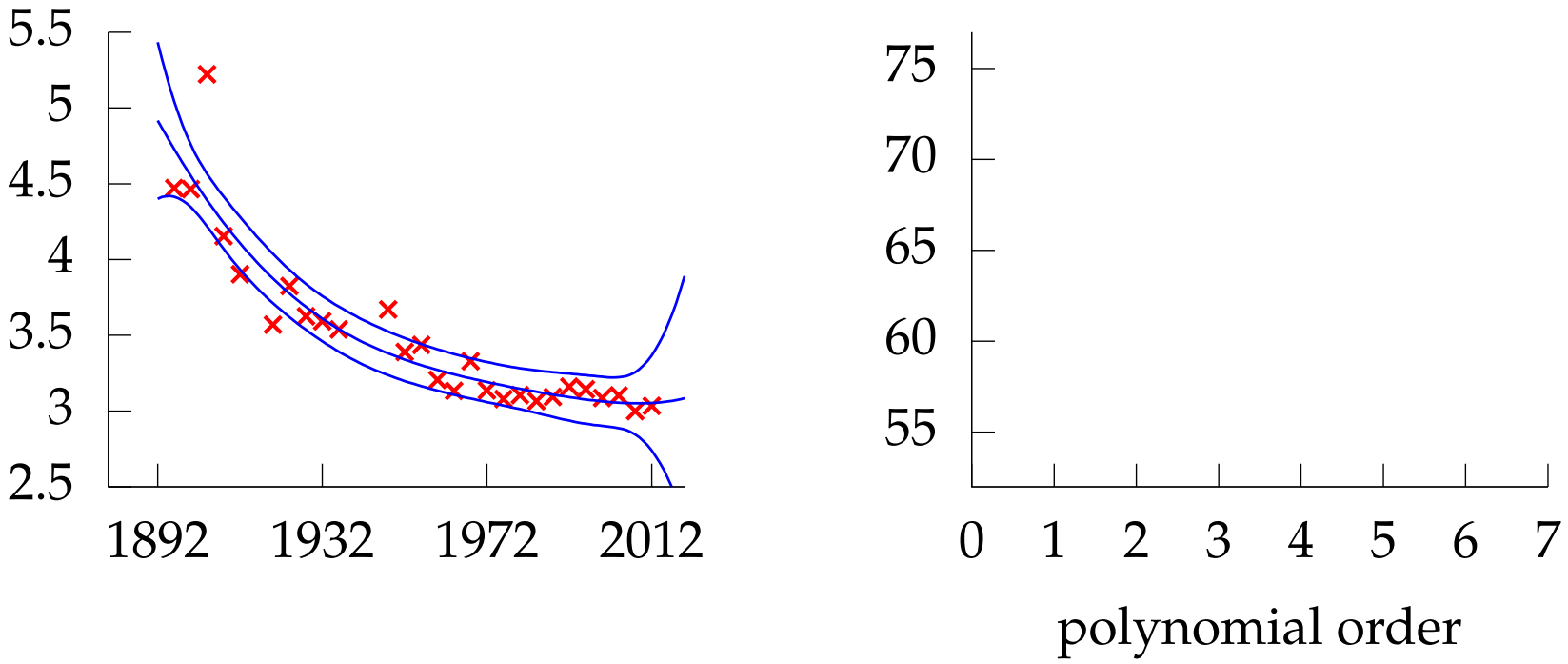
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 3, model error 10.416, $\sigma^2 = 0.0402$, $\sigma = 0.200$.

Polynomial Fits to Olympics Data



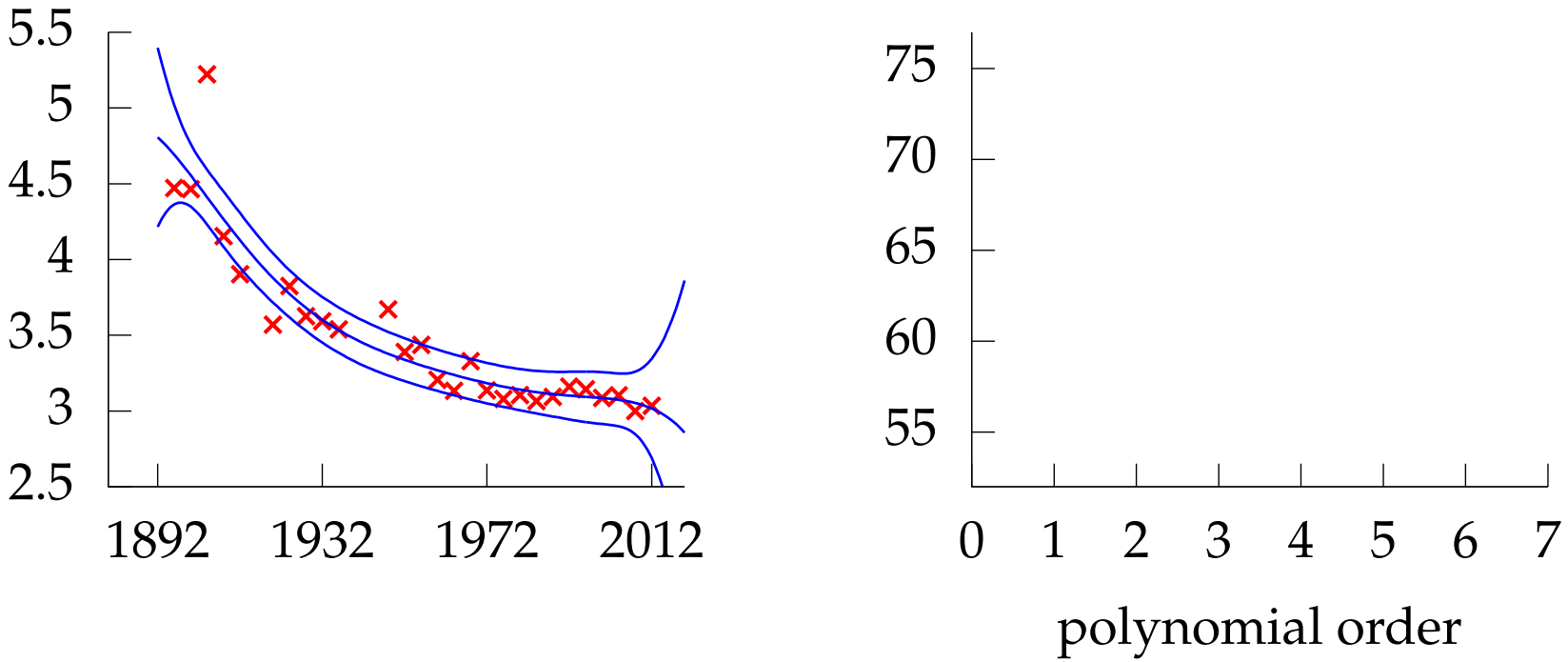
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 4, model error 11.34, $\sigma^2 = 0.0401$, $\sigma = 0.200$.

Polynomial Fits to Olympics Data



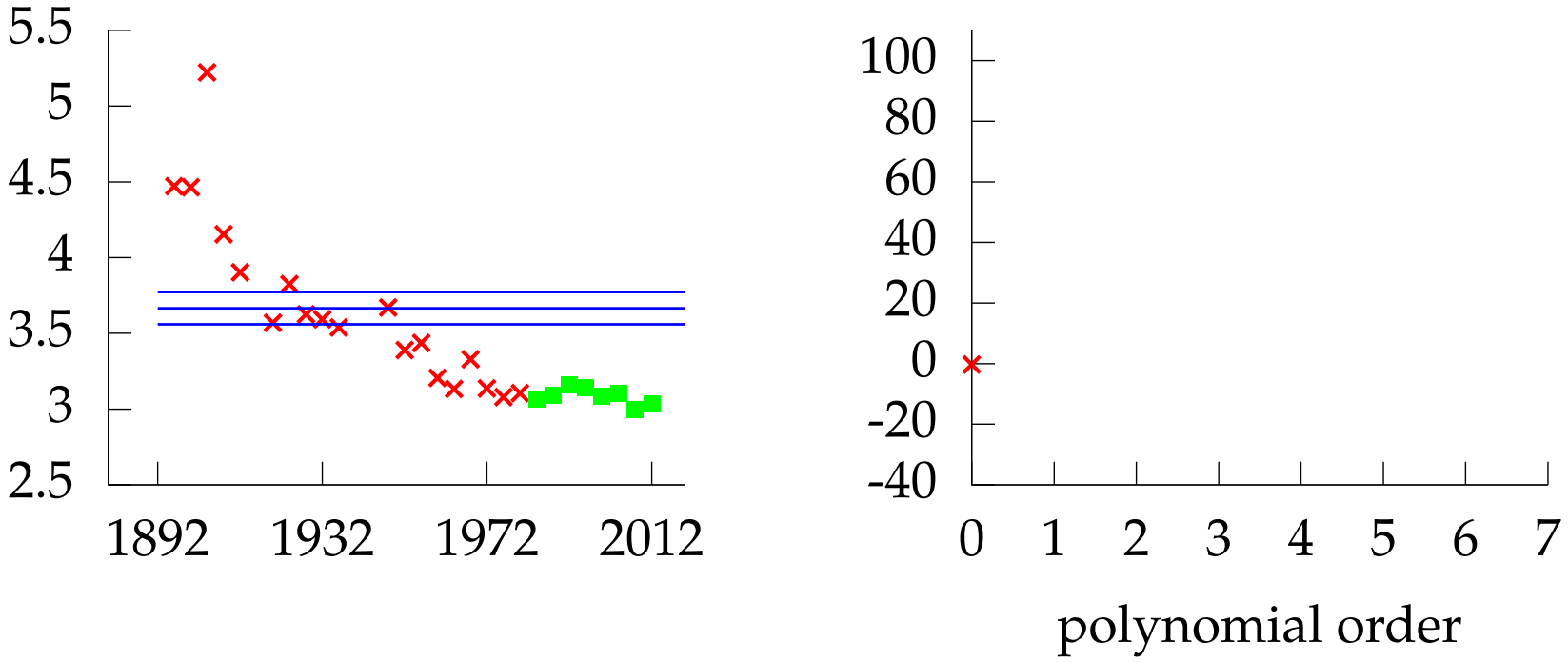
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 5, model error 11.986, $\sigma^2 = 0.0399$, $\sigma = 0.200$.

Polynomial Fits to Olympics Data



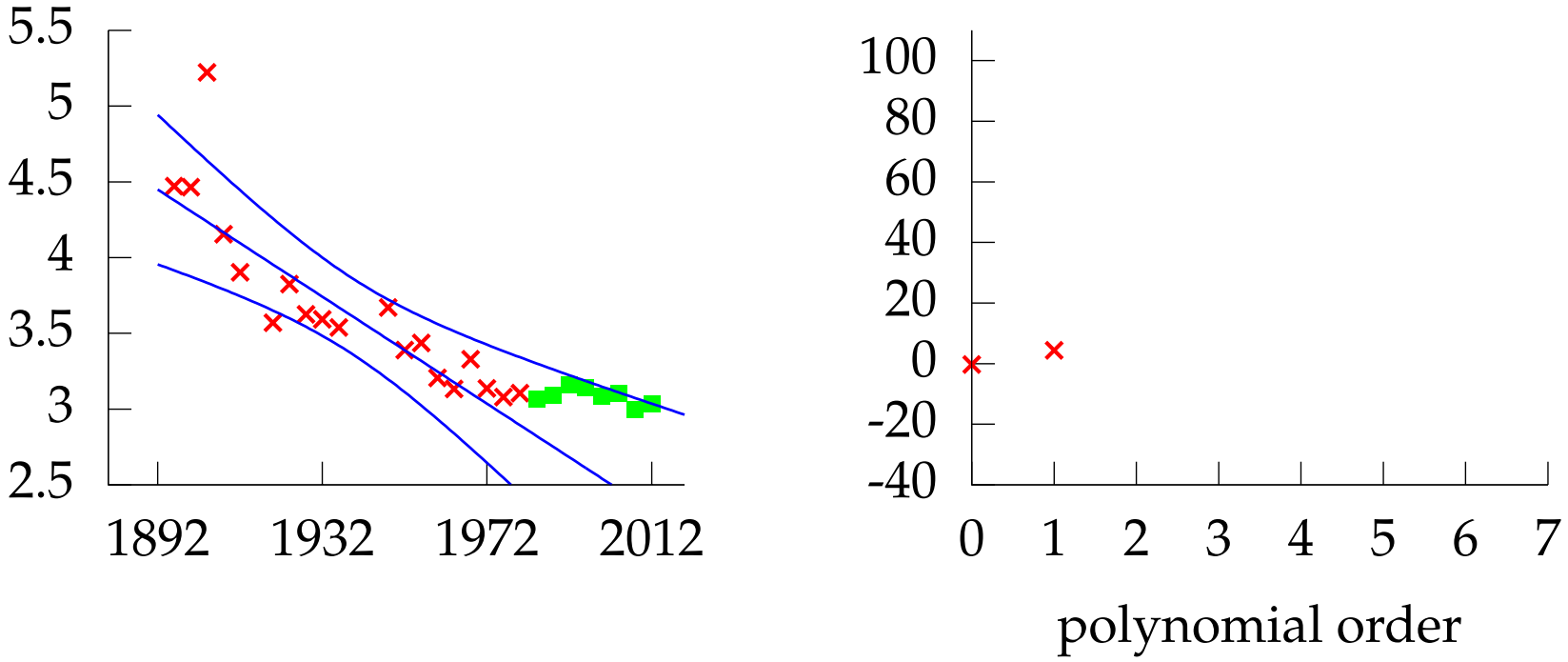
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 6, model error 12.369, $\sigma^2 = 0.0384$, $\sigma = 0.196$.

Validation Set



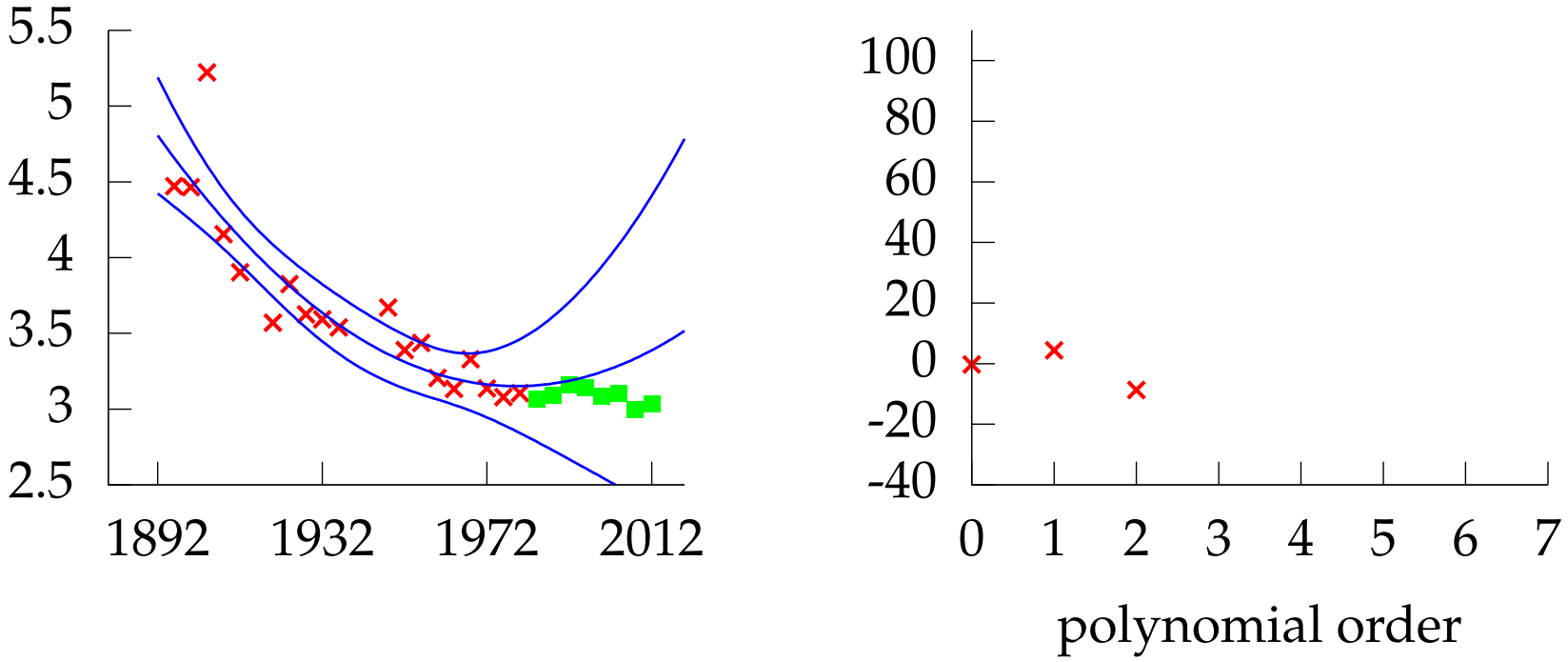
Left: fit to data, *Right*: model error. Polynomial order 0, training error 29.757, validation error -0.29243, $\sigma^2 = 0.302$, $\sigma = 0.550$.

Validation Set



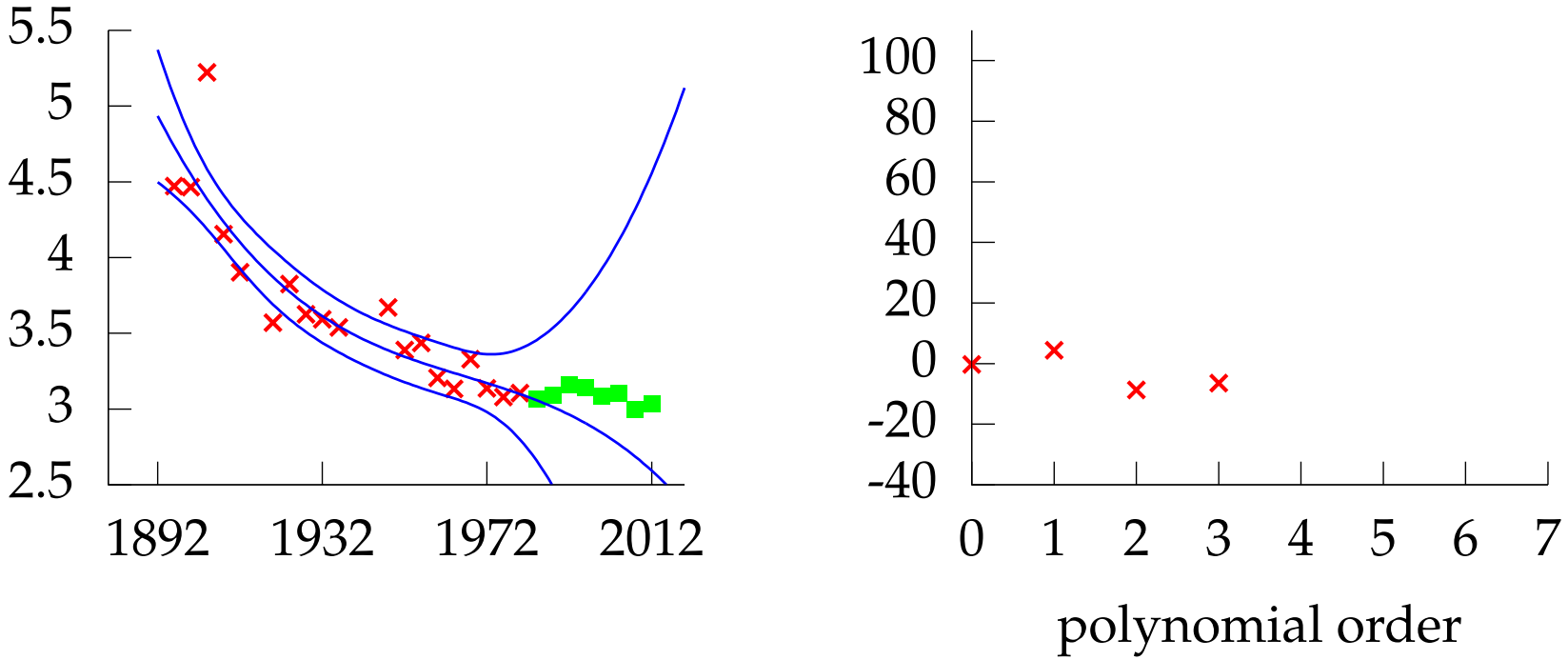
Left: fit to data, *Right*: model error. Polynomial order 1, training error 14.942, validation error 4.4027, $\sigma^2 = 0.0762$, $\sigma = 0.276$.

Validation Set



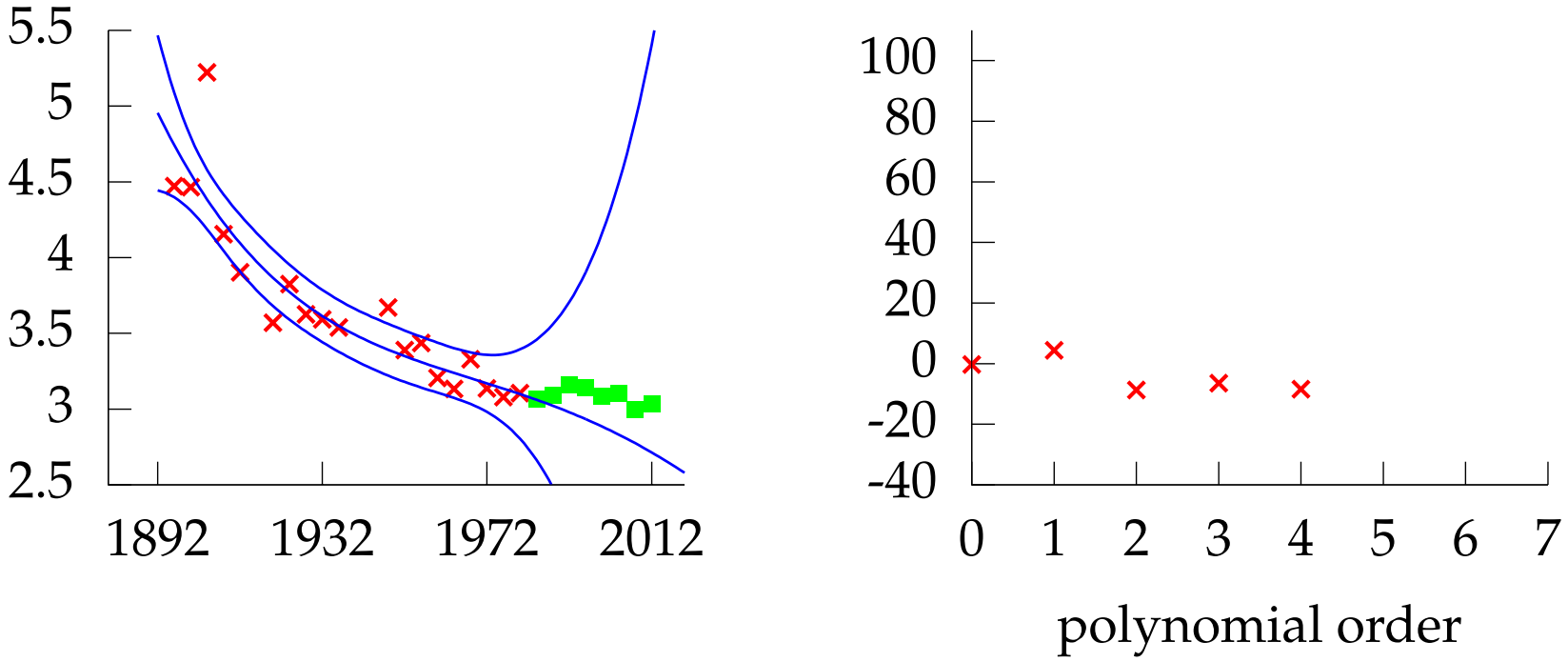
Left: fit to data, *Right*: model error. Polynomial order 2, training error 9.7206, validation error -8.6623, $\sigma^2 = 0.0580$, $\sigma = 0.241$.

Validation Set



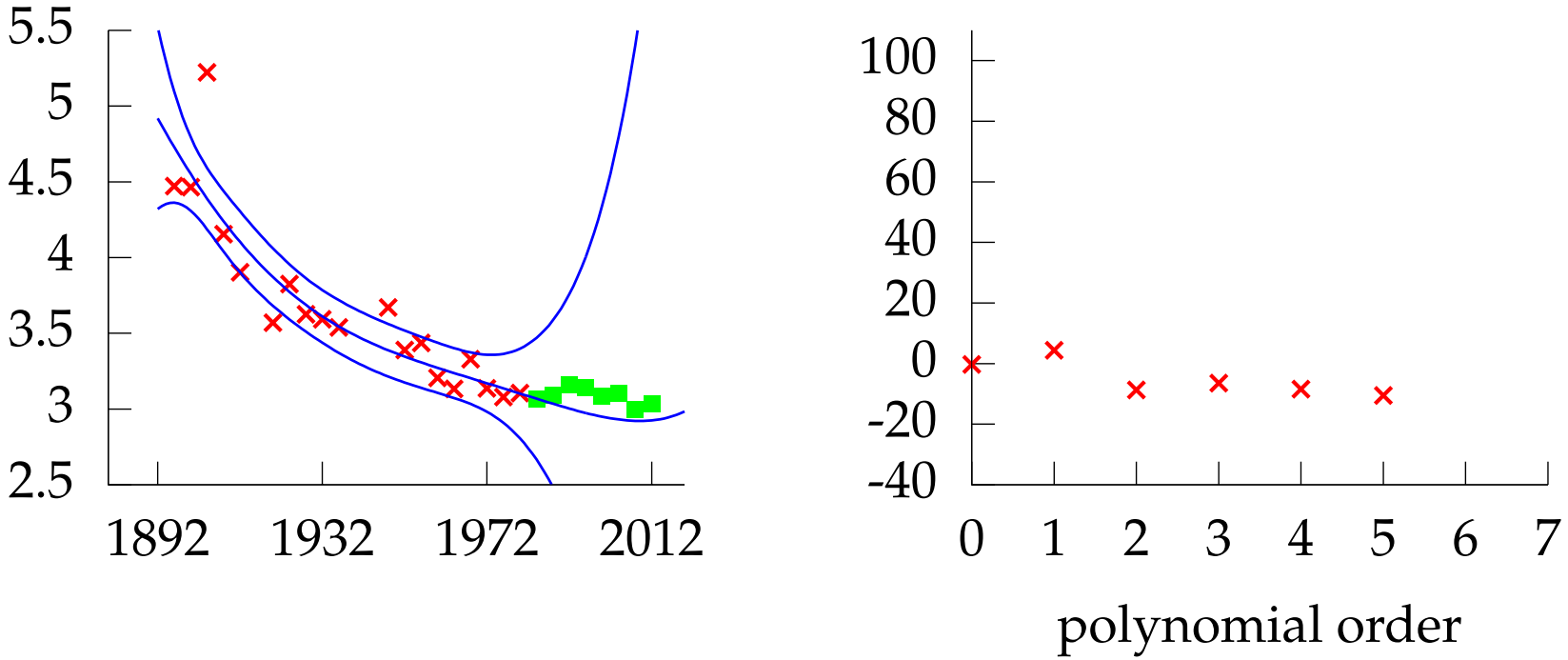
Left: fit to data, *Right*: model error. Polynomial order 3, training error 10.416, validation error -6.4726, $\sigma^2 = 0.0555$, $\sigma = 0.236$.

Validation Set



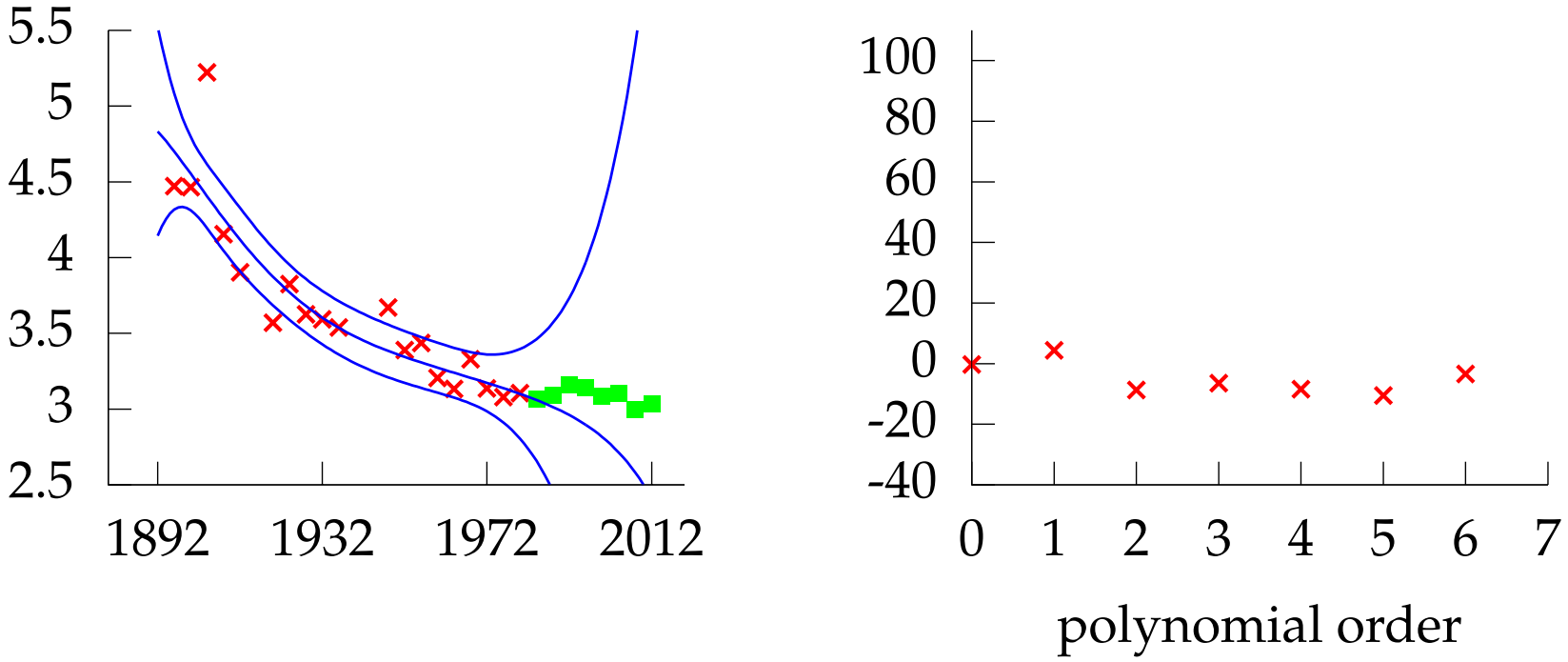
Left: fit to data, *Right*: model error. Polynomial order 4, training error 11.34, validation error -8.431, $\sigma^2 = 0.0555$, $\sigma = 0.236$.

Validation Set



Left: fit to data, *Right*: model error. Polynomial order 5, training error 11.986, validation error -10.483, $\sigma^2 = 0.0551$, $\sigma = 0.235$.

Validation Set



Left: fit to data, *Right*: model error. Polynomial order 6, training error 12.369, validation error -3.3823, $\sigma^2 = 0.0537$, $\sigma = 0.232$.

Example: GWAS Studies

- Try predicting phenotype ($\mathbf{Y}$) from a set of known mutations ($\mathbf{S}$):

$$\mathbf{y}_{i,:} = \mathbf{V}\mathbf{s}_{i,:} + \boldsymbol{\epsilon}_{i,:}$$

- Problem: observations are corrupted by environmental disturbances:

$$\mathbf{y}_{i,:} = \mathbf{V}\mathbf{s}_{i,:} + \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:}$$

  Here $\mathbf{x}_{i,:}$ is a vector of unobserved environmental factors (Parts et al., 2011).

- Our contribution: marginalize both $\mathbf{V}$ and $\mathbf{W}$.

Linear Dimensionality Reduction

Linear Latent Variable Model

- Represent data, $\mathbf{Y}$, with a lower dimensional set of latent variables $\mathbf{X}$.
- Assume a linear relationship of the form

$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:},$$

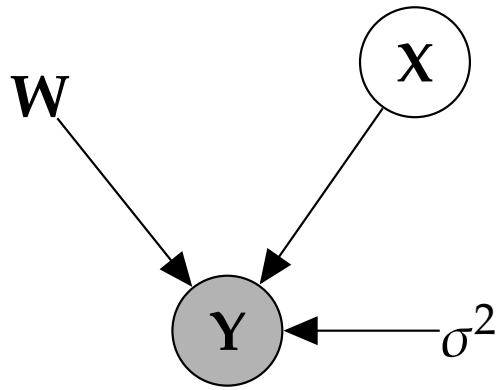
where

$$\boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right).$$

Linear Latent Variable Model

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.

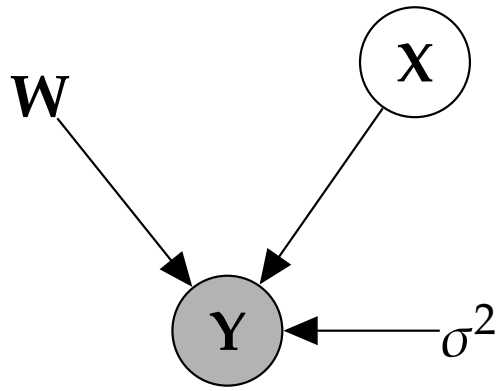


$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right)=\prod_{i=1}^{n}\mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:},\sigma^{2}\mathbf{I}\right)$$

Linear Latent Variable Model

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Standard** Latent variable approach:

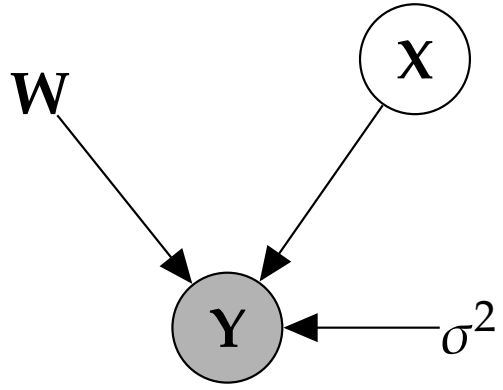


$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right)=\prod_{i=1}^{n}\mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:},\sigma^{2}\mathbf{I}\right)$$

Linear Latent Variable Model

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Standard** Latent variable approach:
  - Define Gaussian prior over *latent space*, $\mathbf{X}$.



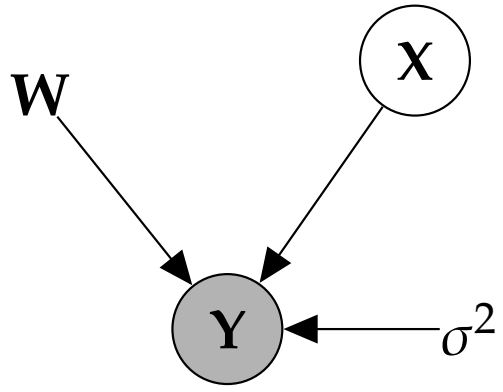
$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:}, \sigma^2\mathbf{I}\right)$$

$$p(\mathbf{X}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{x}_{i,:}|\mathbf{0}, \mathbf{I}\right)$$

Linear Latent Variable Model

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Standard** Latent variable approach:
  - Define Gaussian prior over *latent space*, $\mathbf{X}$.
  - Integrate out *latent variables*.



$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:}, \sigma^2\mathbf{I}\right)$$

$$p(\mathbf{X}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{x}_{i,:}|\mathbf{0}, \mathbf{I}\right)$$

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{0}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}\right)$$

Computation of the Marginal Likelihood

$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:}, \quad \mathbf{x}_{i,:} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$$

Computation of the Marginal Likelihood

$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:}, \quad \mathbf{x}_{i,:} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$$

$$\mathbf{W}\mathbf{x}_{i,:} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}\mathbf{W}^\top),$$

Computation of the Marginal Likelihood

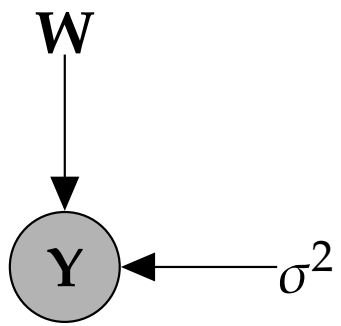
$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:}, \quad \mathbf{x}_{i,:} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$$

$$\mathbf{W}\mathbf{x}_{i,:} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}\mathbf{W}^\top),$$

$$\mathbf{W}\mathbf{x}_{i,:} + \epsilon_{i,:} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)



$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{0}, \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{0}, \mathbf{C}\right), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}$$

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{0}, \mathbf{C}\right), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}$$

$$\log p\left(\mathbf{Y}|\mathbf{W}\right) = -\frac{n}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{Y}^{\top}\mathbf{Y}\right) + \text{const.}$$

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{0}, \mathbf{C}\right), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}$$

$$\log p\left(\mathbf{Y}|\mathbf{W}\right) = -\frac{n}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{Y}^{\top}\mathbf{Y}\right) + \text{const.}$$

If $\mathbf{U}_q$ are first q principal eigenvectors of $n^{-1}\mathbf{Y}^{\top}\mathbf{Y}$ and the corresponding eigenvalues are $\mathbf{\Lambda}_q$,

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{0}, \mathbf{C}\right), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}$$

$$\log p\left(\mathbf{Y}|\mathbf{W}\right) = -\frac{n}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{Y}^{\top}\mathbf{Y}\right) + \text{const.}$$

If $\mathbf{U}_q$ are first q principal eigenvectors of $n^{-1}\mathbf{Y}^{\top}\mathbf{Y}$ and the corresponding eigenvalues are $\mathbf{\Lambda}_q$,

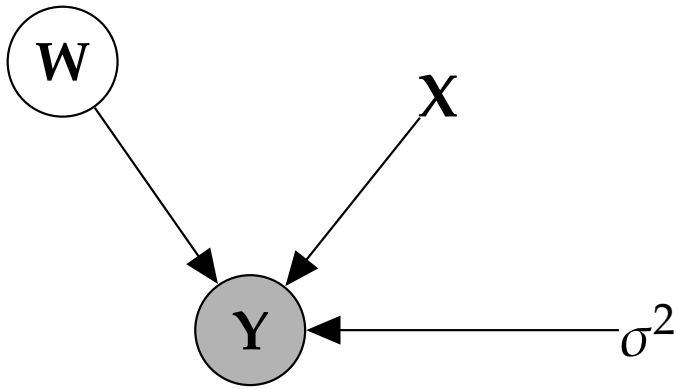
$$\mathbf{W} = \mathbf{U}_q\mathbf{L}\mathbf{R}^{\top}, \quad \mathbf{L} = \left(\mathbf{\Lambda}_q - \sigma^2\mathbf{I}\right)^{\frac{1}{2}}$$

where $\mathbf{R}$ is an arbitrary rotation matrix.

Linear Latent Variable Model III

Dual Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.

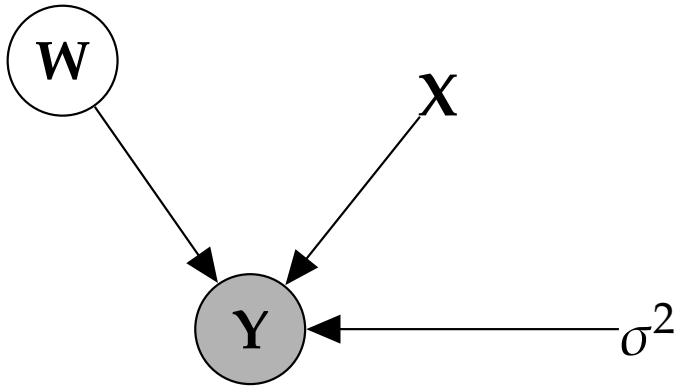


$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right)=\prod_{i=1}^{n}\mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:},\sigma^{2}\mathbf{I}\right)$$

Linear Latent Variable Model III

Dual Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Novel** Latent variable approach:

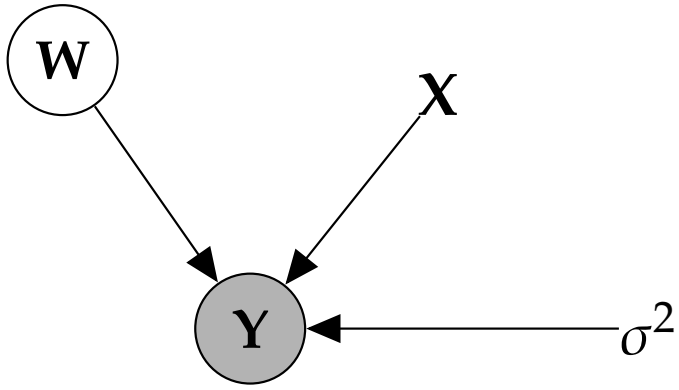


$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right)=\prod_{i=1}^{n}\mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:},\sigma^{2}\mathbf{I}\right)$$

Linear Latent Variable Model III

Dual Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Novel** Latent variable approach:
  - Define Gaussian prior over *parameters*, **W**.



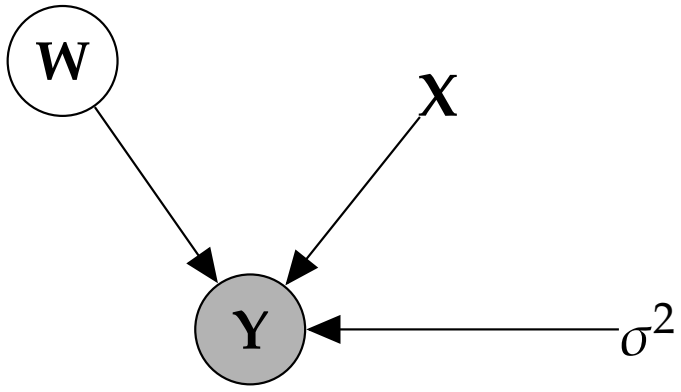
$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:}, \sigma^2\mathbf{I}\right)$$

$$p(\mathbf{W}) = \prod_{i=1}^{p} \mathcal{N}\left(\mathbf{w}_{i,:}|\mathbf{0}, \mathbf{I}\right)$$

Linear Latent Variable Model III

Dual Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Novel** Latent variable approach:
  - Define Gaussian prior over *parameters*, $\mathbf{W}$.
  - Integrate out *parameters*.



$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:}, \sigma^2\mathbf{I}\right)$$

$$p(\mathbf{W}) = \prod_{i=1}^{p} \mathcal{N}\left(\mathbf{w}_{i,:}|\mathbf{0}, \mathbf{I}\right)$$

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{p} \mathcal{N}\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}\right)$$

Computation of the Marginal Likelihood

$$\mathbf{y}_{:,j} = \mathbf{X}\mathbf{w}_{:,j} + \boldsymbol{\epsilon}_{:,j}, \quad \mathbf{w}_{:,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}\right)$$

Computation of the Marginal Likelihood

$$\mathbf{y}_{:,j} = \mathbf{X}\mathbf{w}_{:,j} + \boldsymbol{\epsilon}_{:,j}, \quad \mathbf{w}_{:,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}\right)$$

$$\mathbf{X}\mathbf{w}_{:,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{X}\mathbf{X}^\top),$$

Computation of the Marginal Likelihood

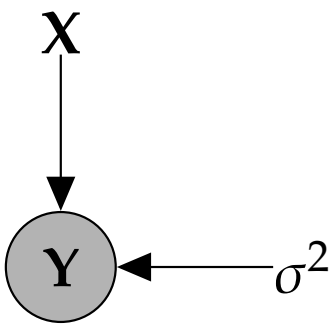
$$\mathbf{y}_{:,j} = \mathbf{X}\mathbf{w}_{:,j} + \boldsymbol{\epsilon}_{:,j}, \quad \mathbf{w}_{:,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}\right)$$

$$\mathbf{X}\mathbf{w}_{:,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{X}\mathbf{X}^\top),$$

$$\mathbf{X}\mathbf{w}_{:,j} + \epsilon_{:,j} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{X}\mathbf{X}^\top + \sigma^2 \mathbf{I}\right)$$

Linear Latent Variable Model IV

Dual Probabilistic PCA Max. Likelihood Soln (Lawrence, 2004, 2005)



$$p\left(\mathbf{Y}|\mathbf{X}\right) = \prod_{j=1}^{p} \mathcal{N}\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model IV

Dual PPCA Max. Likelihood Soln (Lawrence, 2004, 2005)

$$p\left(\mathbf{Y}|\mathbf{X}\right) = \prod_{j=1}^{p} \mathcal{N}\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K}\right), \quad \mathbf{K} = \mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I}$$

Linear Latent Variable Model IV

PPCA Max. Likelihood Soln (Tipping and Bishop, 1999)

$$p\left(\mathbf{Y}|\mathbf{X}\right) = \prod_{j=1}^{p} \mathcal{N}\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K}\right), \quad \mathbf{K} = \mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I}$$

$$\log p\left(\mathbf{Y}|\mathbf{X}\right) = -\frac{p}{2}\log|\mathbf{K}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^{\top}\right) + \text{const.}$$

Linear Latent Variable Model IV

PPCA Max. Likelihood Soln

$$p\left(\mathbf{Y}|\mathbf{X}\right) = \prod_{j=1}^{p} \mathcal{N}\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K}\right), \quad \mathbf{K} = \mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I}$$

$$\log p\left(\mathbf{Y}|\mathbf{X}\right) = -\frac{p}{2}\log|\mathbf{K}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^{\top}\right) + \text{const.}$$

If $\mathbf{U}'_q$ are first q principal eigenvectors of $p^{-1}\mathbf{Y}\mathbf{Y}^{\top}$ and the corresponding eigenvalues are $\mathbf{\Lambda}_q$,

Linear Latent Variable Model IV

PPCA Max. Likelihood Soln

$$p\left(\mathbf{Y}|\mathbf{X}\right) = \prod_{j=1}^{p} \mathcal{N}\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K}\right), \quad \mathbf{K} = \mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I}$$

$$\log p\left(\mathbf{Y}|\mathbf{X}\right) = -\frac{p}{2}\log|\mathbf{K}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^{\top}\right) + \text{const.}$$

If $\mathbf{U}'_q$ are first q principal eigenvectors of $p^{-1}\mathbf{Y}\mathbf{Y}^{\top}$ and the corresponding eigenvalues are $\mathbf{\Lambda}_q$,

$$\mathbf{X} = \mathbf{U}'_q\mathbf{L}\mathbf{R}^{\top}, \quad \mathbf{L} = \left(\mathbf{\Lambda}_q - \sigma^2\mathbf{I}\right)^{\frac{1}{2}}$$

where $\mathbf{R}$ is an arbitrary rotation matrix.

Linear Latent Variable Model IV

Dual PPCA Max. Likelihood Soln (Lawrence, 2004, 2005)

$$p\left(\mathbf{Y}|\mathbf{X}\right) = \prod_{j=1}^{p} \mathcal{N}\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K}\right), \quad \mathbf{K} = \mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I}$$

$$\log p\left(\mathbf{Y}|\mathbf{X}\right) = -\frac{p}{2}\log|\mathbf{K}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^{\top}\right) + \text{const.}$$

If $\mathbf{U}'_q$ are first q principal eigenvectors of $p^{-1}\mathbf{Y}\mathbf{Y}^{\top}$ and the corresponding eigenvalues are $\mathbf{\Lambda}_q$,

$$\mathbf{X} = \mathbf{U}'_q\mathbf{L}\mathbf{R}^{\top}, \quad \mathbf{L} = \left(\mathbf{\Lambda}_q - \sigma^2\mathbf{I}\right)^{\frac{1}{2}}$$

where $\mathbf{R}$ is an arbitrary rotation matrix.

Linear Latent Variable Model IV

PPCA Max. Likelihood Soln (Tipping and Bishop, 1999)

$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{0}, \mathbf{C}\right), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}$$

$$\log p\left(\mathbf{Y}|\mathbf{W}\right) = -\frac{n}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{Y}^{\top}\mathbf{Y}\right) + \text{const.}$$

If $\mathbf{U}_q$ are first q principal eigenvectors of $n^{-1}\mathbf{Y}^{\top}\mathbf{Y}$ and the corresponding eigenvalues are $\mathbf{\Lambda}_q$,

$$\mathbf{W} = \mathbf{U}_q\mathbf{L}\mathbf{R}^{\top}, \quad \mathbf{L} = \left(\mathbf{\Lambda}_q - \sigma^2\mathbf{I}\right)^{\frac{1}{2}}$$

where $\mathbf{R}$ is an arbitrary rotation matrix.

Equivalence of Formulations

The Eigenvalue Problems are equivalent

- Solution for Probabilistic PCA (solves for the mapping)

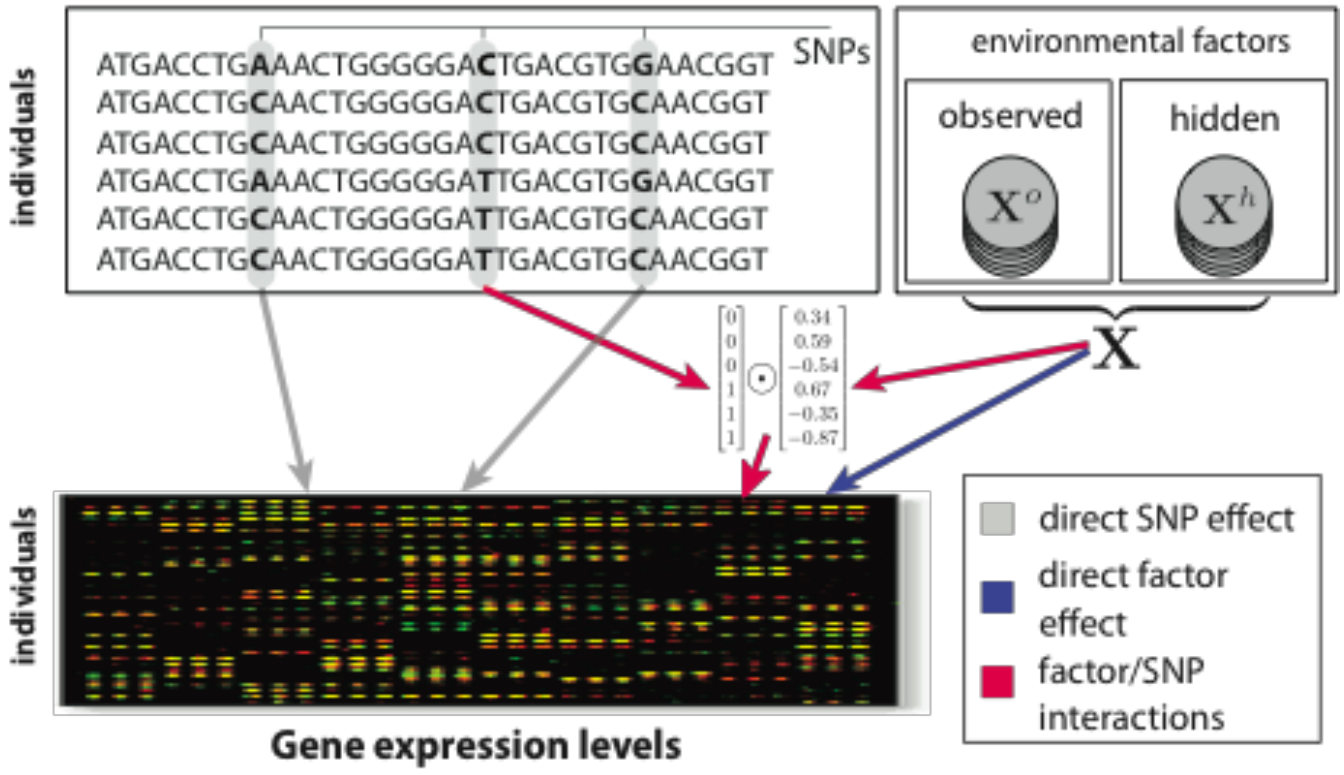
$$\mathbf{Y}^\top \mathbf{Y}\mathbf{U}_q = \mathbf{U}_q \mathbf{\Lambda}_q \qquad \mathbf{W} = \mathbf{U}_q \mathbf{L}\mathbf{R}^\top$$

- Solution for Dual Probabilistic PCA (solves for the latent positions)

$$\mathbf{Y}\mathbf{Y}^\top \mathbf{U}'_q = \mathbf{U}'_q \mathbf{\Lambda}_q \qquad \mathbf{X} = \mathbf{U}'_q \mathbf{L}\mathbf{R}^\top$$

- Equivalence is from

$$\mathbf{U}_q = \mathbf{Y}^\top \mathbf{U}'_q \mathbf{\Lambda}_q^{-\frac{1}{2}}$$


individuals
ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
SNPs
environmental factors
observed
hidden
$\mathbf{X}^o$
$\mathbf{X}^h$
$\mathbf{X}$
individuals
Gene expression levels
direct SNP effect
direct factor effect
factor/SNP interactions

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Joint Modelling of Confounding Factors and Prominent Genetic Regulators Provides Increased Accuracy in Genetical Genomics Studies

Nicoló Fusi[1☯]*, Oliver Stegle[2☯]*, Neil D. Lawrence[1]*

1 Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield, United Kingdom, **2** Machine Learning and Computational Biology Research Group, Max Planck Institute for Developmental Biology, Tübingen, Germany

BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 11 2013, pages 1382–1389
doi:10.1093/bioinformatics/btt148

Gene expression

Advance Access publication April 4, 2013

Detecting regulatory gene–environment interactions with unmeasured environmental factors

Nicoló Fusi[1,*], Christoph Lippert[2], Karsten Borgwardt[3,4], Neil D. Lawrence[1] and Oliver Stegle[3,5,*]

[1]Department of Computer Science, University of Sheffield, Sheffield S10 2HQ, UK, [2]Microsoft Research, Los Angeles, CA 90024, USA, [3]Machine Learning and Computational Biology Research Group, Max Planck Institutes, 72076 Tübingen, Germany, [4]Zentrum für Bioinformatik, Eberhard Karls Universität, 72074 Tübingen, Germany and [5]EMBL-European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK

Back to the full model

$$\mathbf{y}_g \sim \mathcal{N}\Big(\underbrace{\mu_g \mathbf{1}}_{\text{mean}}, \underbrace{\mathbf{K_S}}_{\substack{\text{SNP} \\ \text{effects}}} + \underbrace{\mathbf{K_X}}_{\substack{\text{direct factor} \\ \text{effects}}} + \underbrace{\mathbf{K_I}}_{\substack{\text{SNP-factor} \\ \text{interactions}}} + \underbrace{\sigma_p^2 \mathbf{K_P}}_{\substack{\text{population} \\ \text{structure}}} + \underbrace{\sigma_e^2 \mathbf{I}}_{\text{noise}} \Big)$$

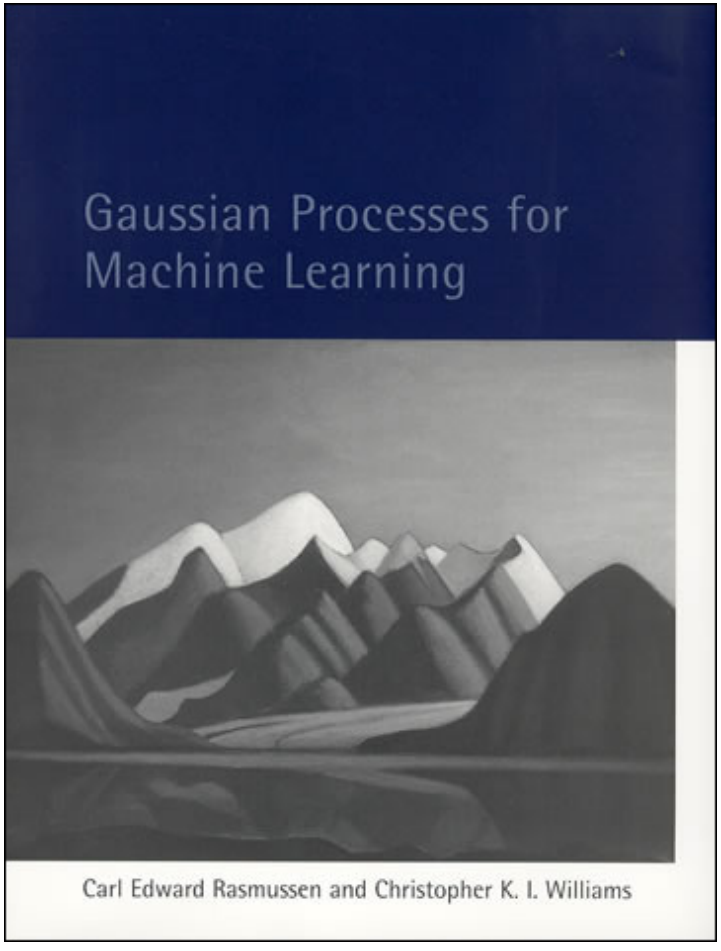
$$p(\mathbf{Y} \,|\, \mathbf{S}, \mathbf{X}, \mathbf{\Theta}_{\mathrm{K}}, \mathcal{I}, \mathcal{S}) = \prod_{g=1}^{G} \mathcal{N}(\mathbf{y}_g \,|\, \mathbf{0}, \mathbf{\Sigma}),$$

$$\mathbf{\Sigma} = \underbrace{\sum_{\forall k \in \mathcal{S}} \beta_k^2 \mathbf{s}_k \mathbf{s}_k^\top}_{\mathbf{K_S}} + \underbrace{\sum_{q=1}^{Q} \alpha_q^2 \mathbf{x}_q \mathbf{x}_q^\top}_{\mathbf{K_X}} + \underbrace{\sum_{\forall (k,q) \in \mathcal{I}} \gamma_{k,q}^2 (\mathbf{s}_k \odot \mathbf{x}_q)(\mathbf{s}_k \odot \mathbf{x}_q)^{\mathrm{T}}}_{\mathbf{K_I}} + \sigma_p^2 \mathbf{K_P} + \sigma_e^2 \mathbf{I}$$

Reading

- Section 2.3 of Bishop up to top of pg 85 (multivariate Gaussians).
- Section 3.3 of Bishop up to 159 (pg 152–159).
- The LIMMI paper (Fusi et al., 2013).
- The PANAMA paper (Fusi et al., 2012).

Book



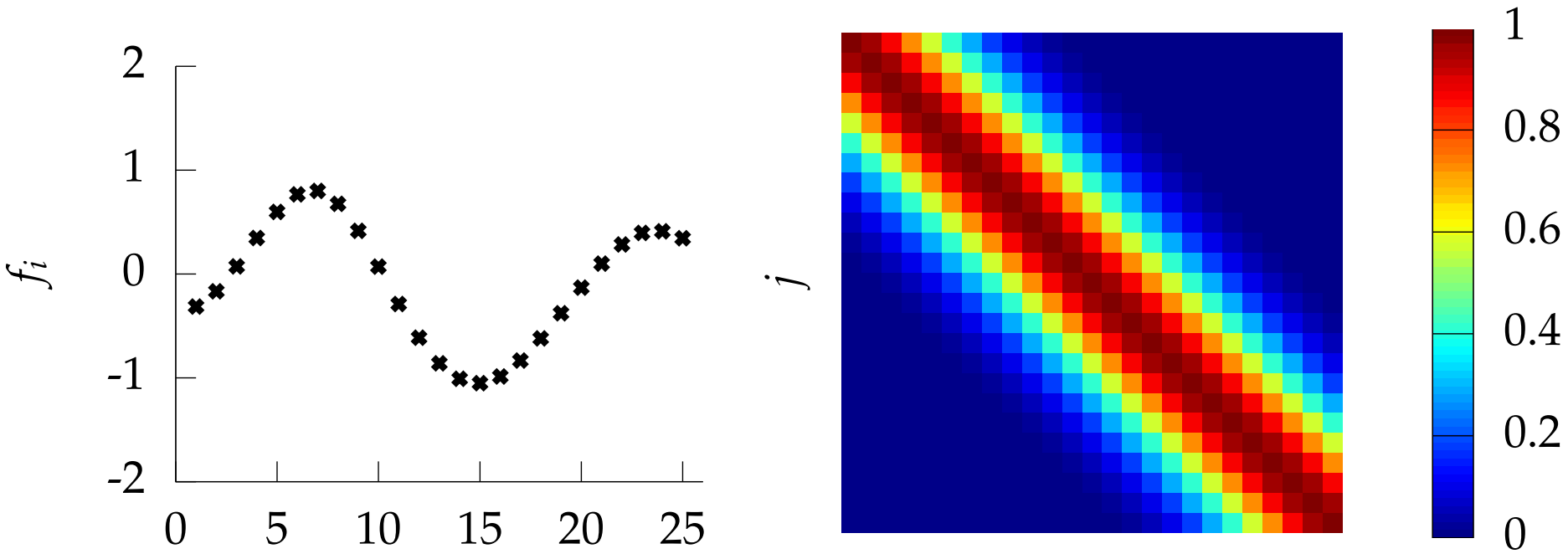
Rasmussen and Williams (2006)

Sampling a Function

Multi-variate Gaussians

- We will consider a Gaussian with a particular structure of covariance matrix.
- Generate a single sample from this 25 dimensional Gaussian distribution, $\mathbf{f} = [f_1, f_2 \dots f_{25}]$.
- We will plot these points against their index.

Gaussian Distribution Sample

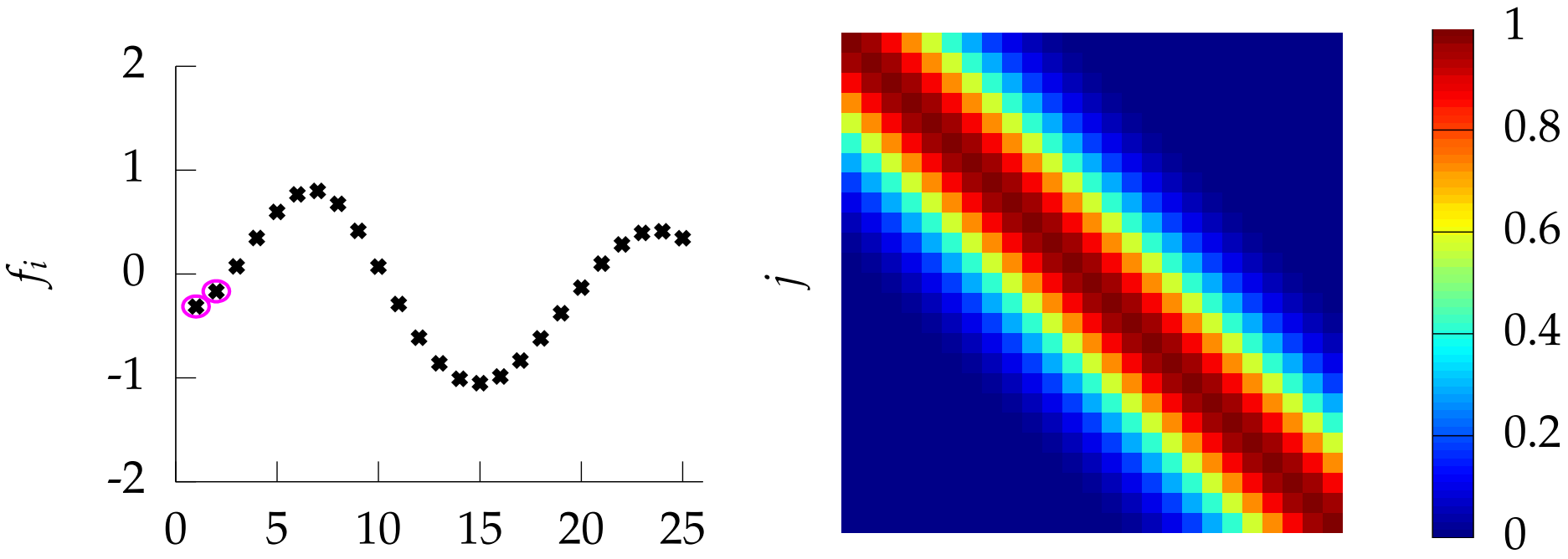


(a) A 25 dimensional correlated random variable (values ploted against index)

(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample

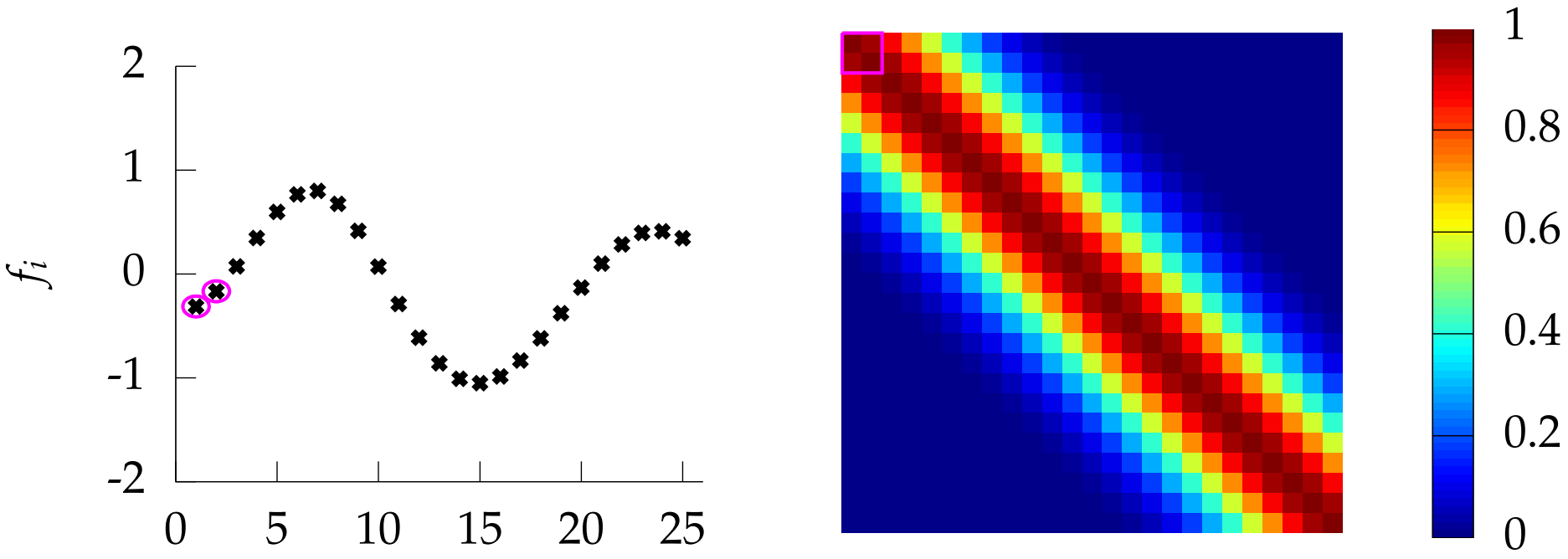


(a) A 25 dimensional correlated random variable (values ploted against index)

(b) colormap *i*showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample

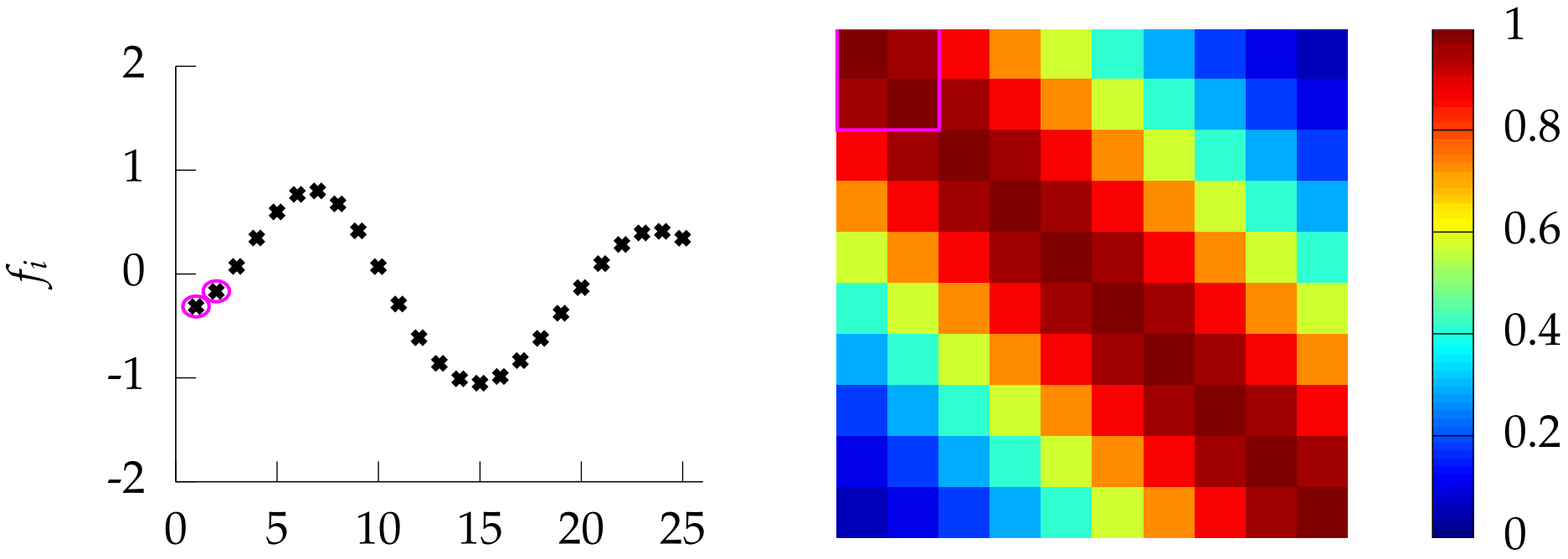


(a) A 25 dimensional correlated random variable (values ploted against index)

(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample

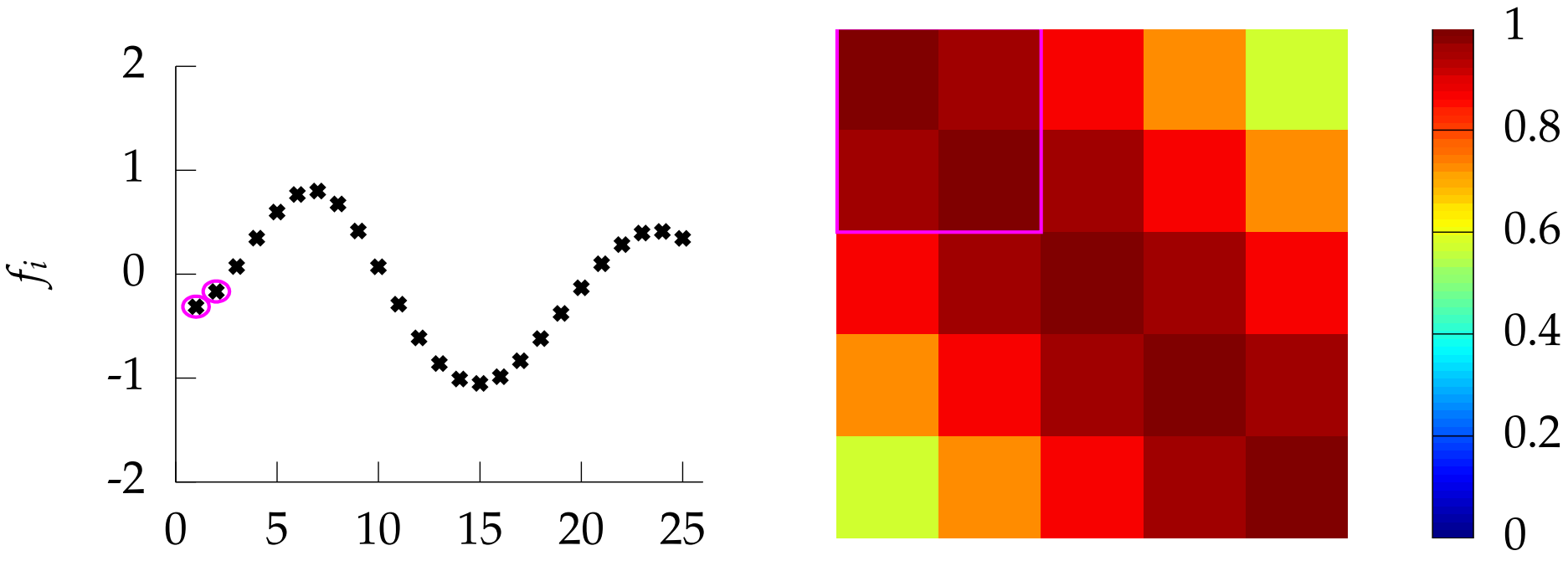


(a) A 25 dimensional correlated random variable (values ploted against index)

(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample

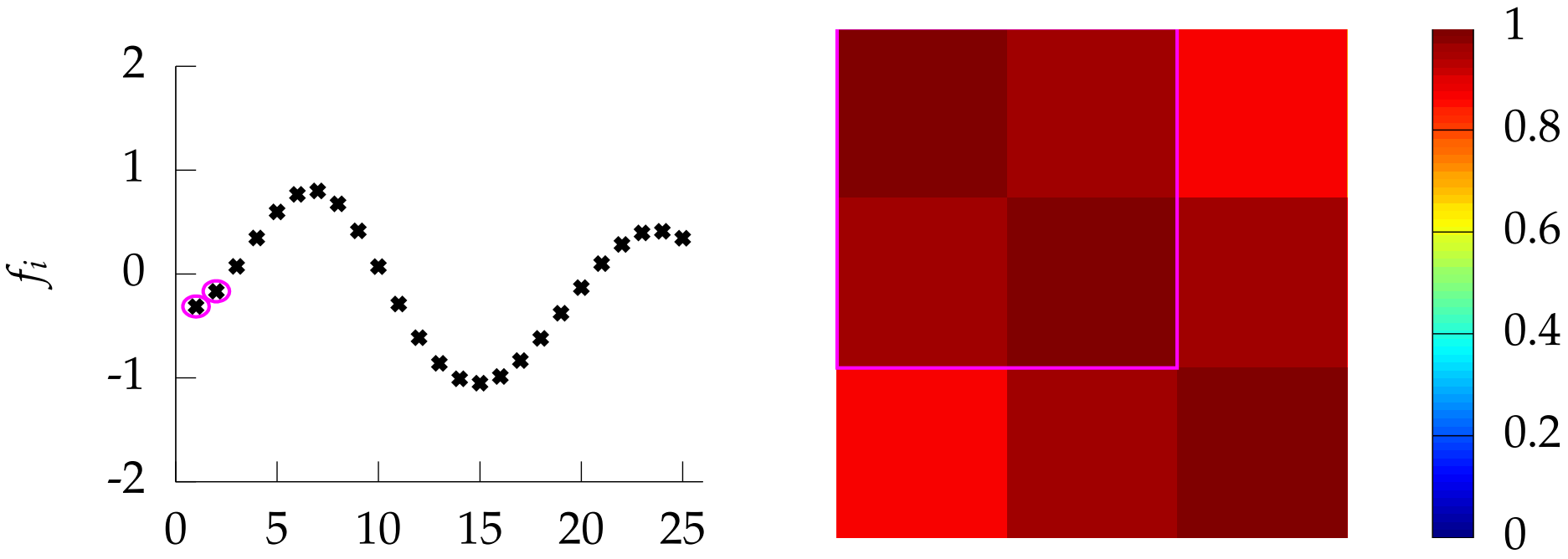


(a) A 25 dimensional correlated random variable (values ploted against index)

(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample

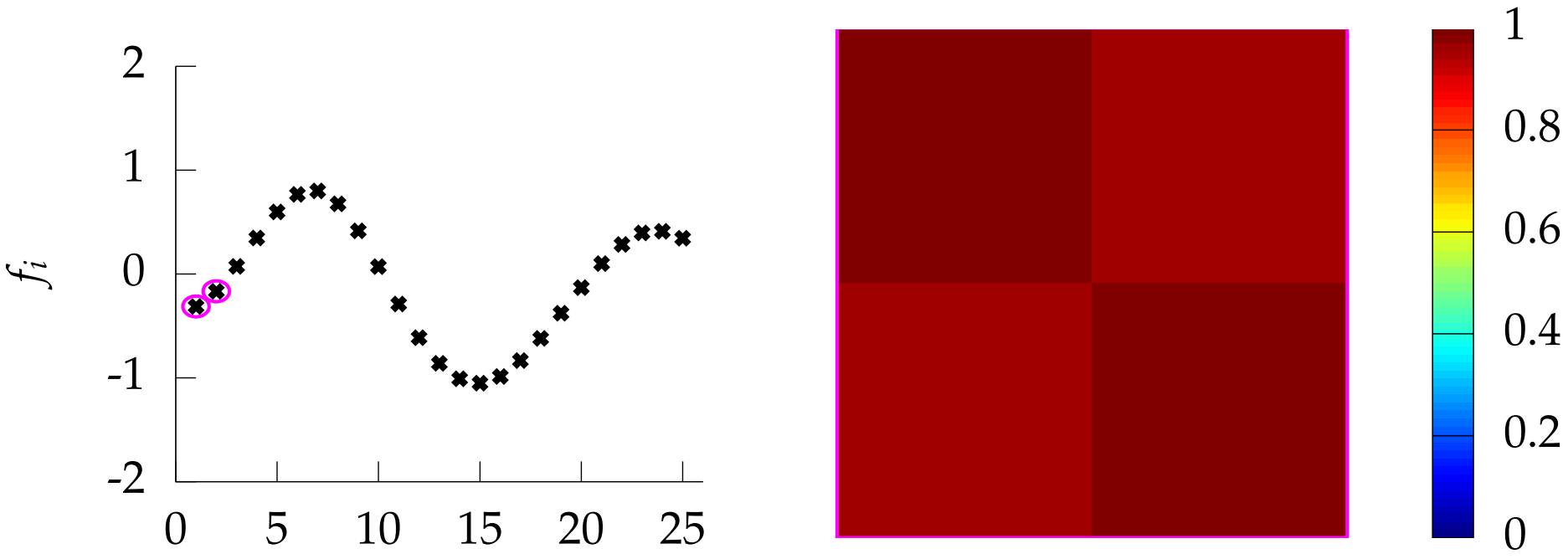


(a) A 25 dimensional correlated random variable (values ploted against index)

(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample

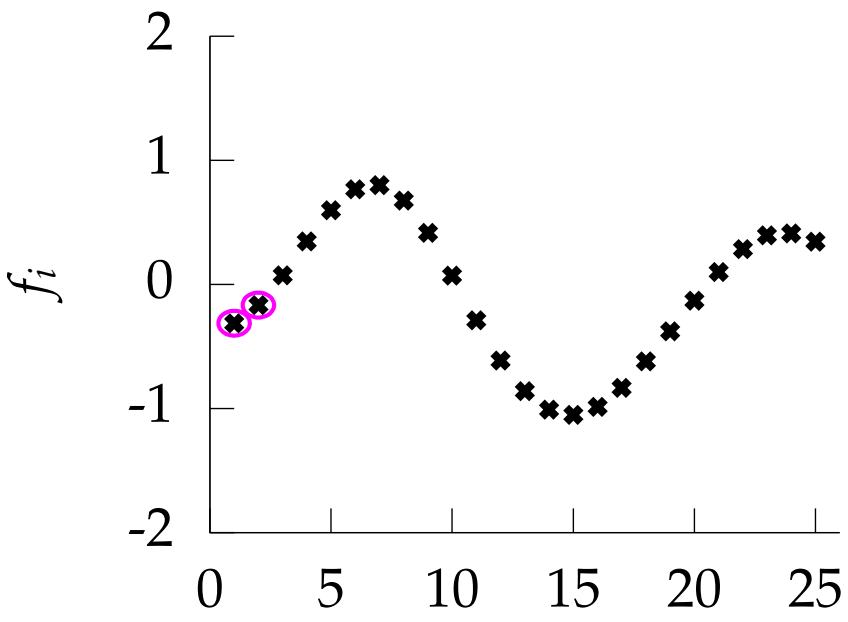


(a) A 25 dimensional correlated random variable (values ploted against index)

(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



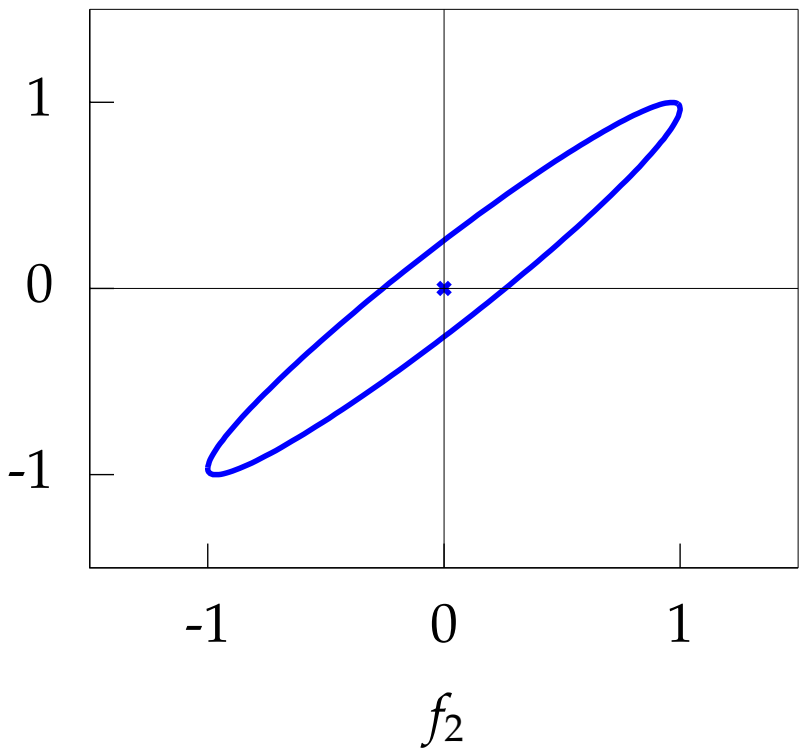
(a) A 25 dimensional correlated random variable (values ploted against index)

$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

(b) correlation between f_1 and f_2.

Figure: A sample from a 25 dimensional Gaussian distribution.

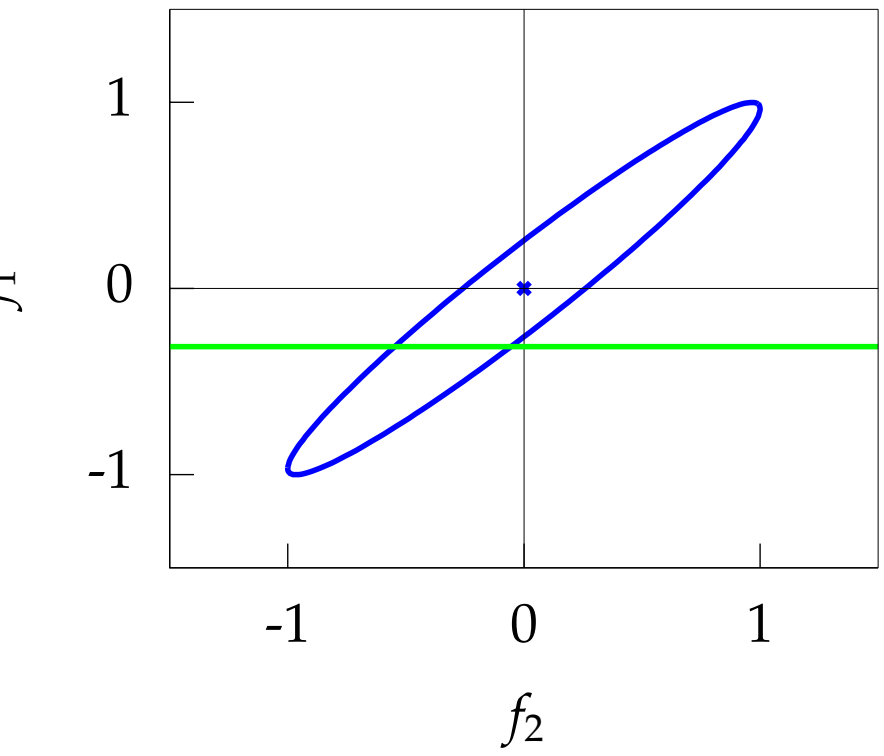
Prediction of f_2 from f_1



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.

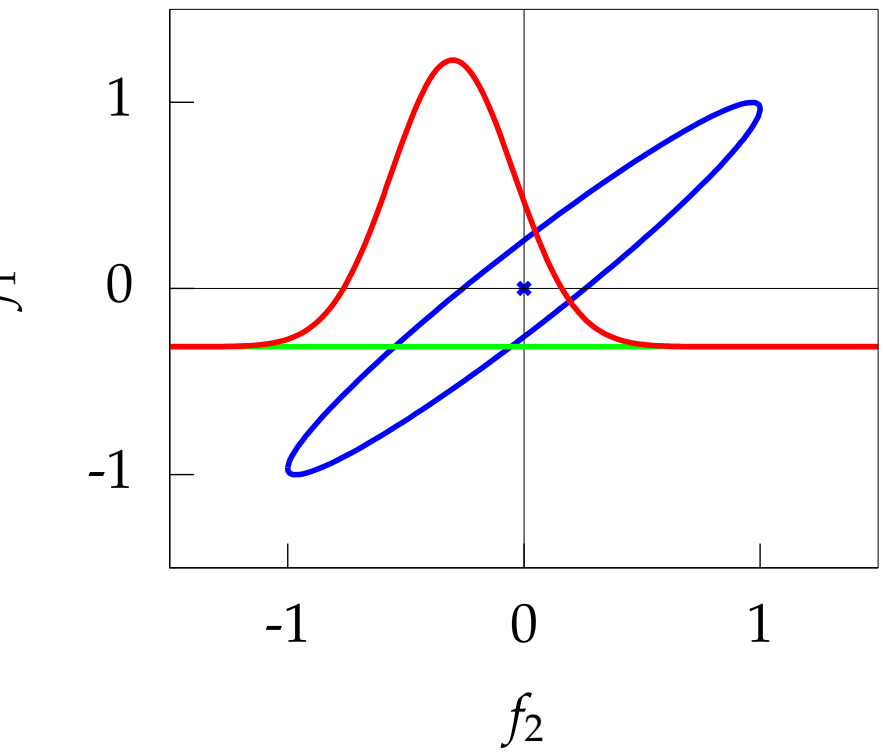
Prediction of f_2 from f_1



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.
- We observe that $f_1 = -0.313$.

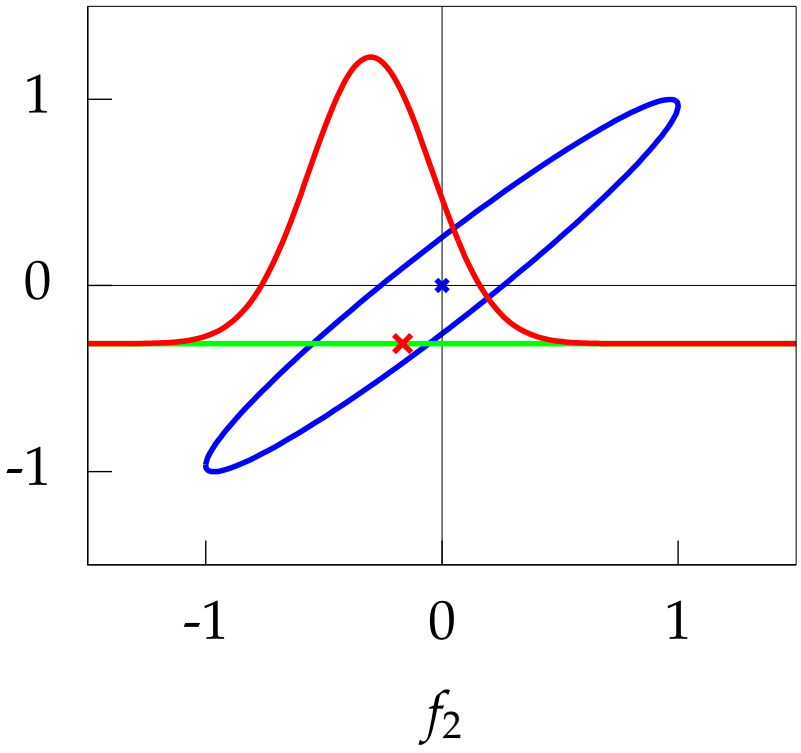
Prediction of f_2 from f_1



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_2|f_1 = -0.313)$.

Prediction of f_2 from f_1



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_2|f_1 = -0.313)$.

Prediction with Correlated Gaussians

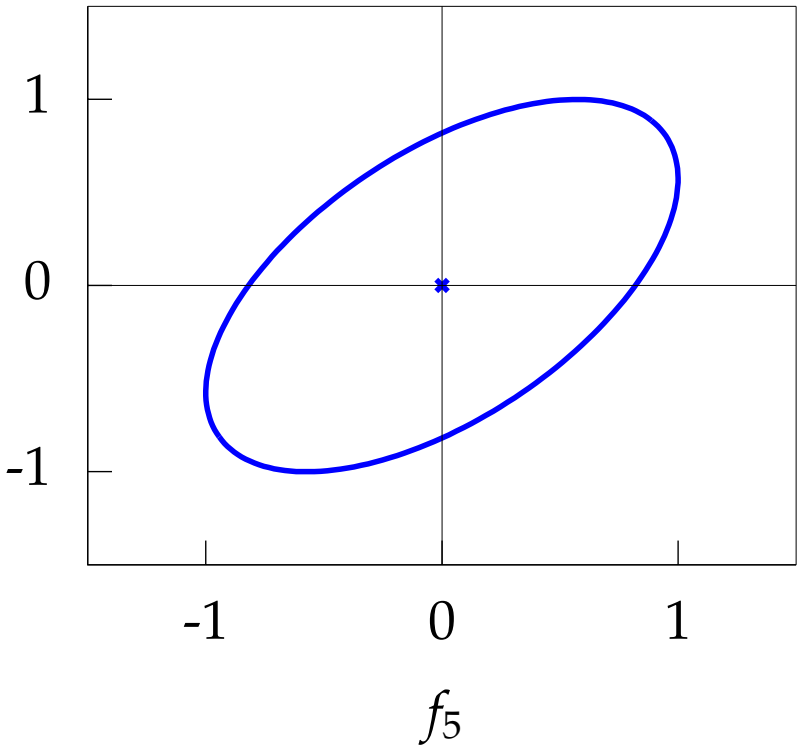
- Prediction of f_2 from f_1 requires *conditional density*.
- Conditional density is *also* Gaussian.

$$p(f_2|f_1) = \mathcal{N}\left(f_2|\frac{k_{1,2}}{k_{1,1}}f_1, k_{2,2} - \frac{k_{1,2}^2}{k_{1,1}}\right)$$

where covariance of joint density is given by

$$\mathbf{K} = \begin{bmatrix} k_{1,1} & k_{1,2} \\ k_{2,1} & k_{2,2} \end{bmatrix}$$

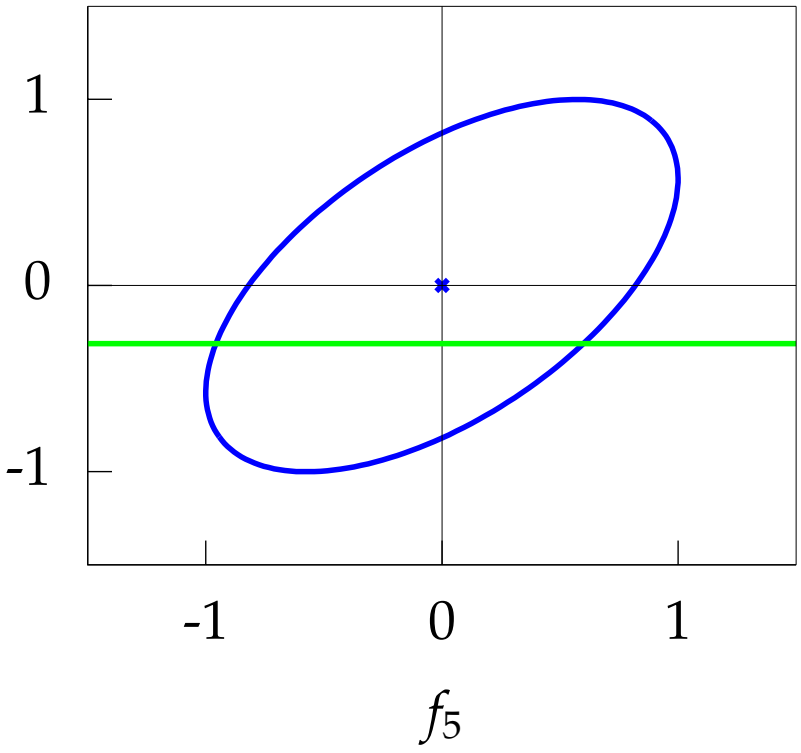
Prediction of f_5 from f_1



$$\begin{bmatrix} 1 & 0.57375 \\ 0.57375 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_5)$.

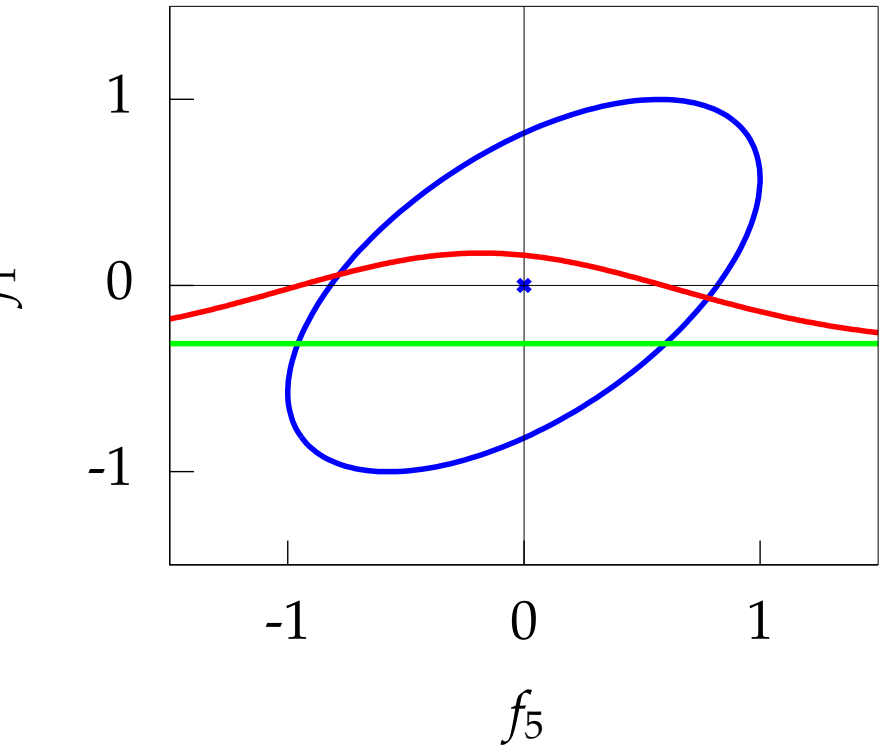
Prediction of f_5 from f_1



$$\begin{bmatrix} 1 & 0.57375 \\ 0.57375 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_5)$.
- We observe that $f_1 = -0.313$.

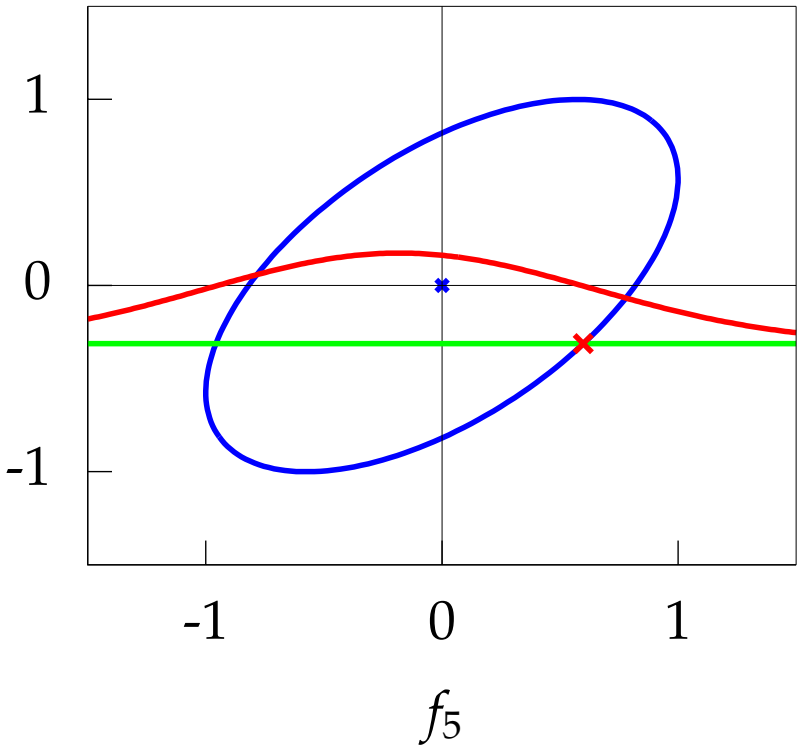
Prediction of f_5 from f_1



$$\begin{bmatrix} 1 & 0.57375 \\ 0.57375 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_5)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_5|f_1 = -0.313)$.

Prediction of f_5 from f_1



$$\begin{bmatrix} 1 & 0.57375 \\ 0.57375 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_5)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_5|f_1 = -0.313)$.

Prediction with Correlated Gaussians

- Prediction of $\mathbf{f}_*$ from $\mathbf{f}$ requires multivariate *conditional density*.
- Multivariate conditional density is *also* Gaussian.

$$p(\mathbf{f}_*|\mathbf{f}) = \mathcal{N}\left(\mathbf{f}_*|\mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{f}, \mathbf{K}_{*,*} - \mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{K}_{\mathbf{f},*}\right)$$

- Here covariance of joint density is given by

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{\mathbf{f},\mathbf{f}} & \mathbf{K}_{*,\mathbf{f}} \\ \mathbf{K}_{\mathbf{f},*} & \mathbf{K}_{*,*} \end{bmatrix}$$

Prediction with Correlated Gaussians

- Prediction of $\mathbf{f}_*$ from $\mathbf{f}$ requires multivariate *conditional density*.
- Multivariate conditional density is *also* Gaussian.

$$p(\mathbf{f}_*|\mathbf{f}) = \mathcal{N}\left(\mathbf{f}_*|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

$$\boldsymbol{\mu} = \mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{f}$$

$$\boldsymbol{\Sigma} = \mathbf{K}_{*,*} - \mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{K}_{\mathbf{f},*}$$

- Here covariance of joint density is given by

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{\mathbf{f},\mathbf{f}} & \mathbf{K}_{*,\mathbf{f}} \\ \mathbf{K}_{\mathbf{f},*} & \mathbf{K}_{*,*} \end{bmatrix}$$

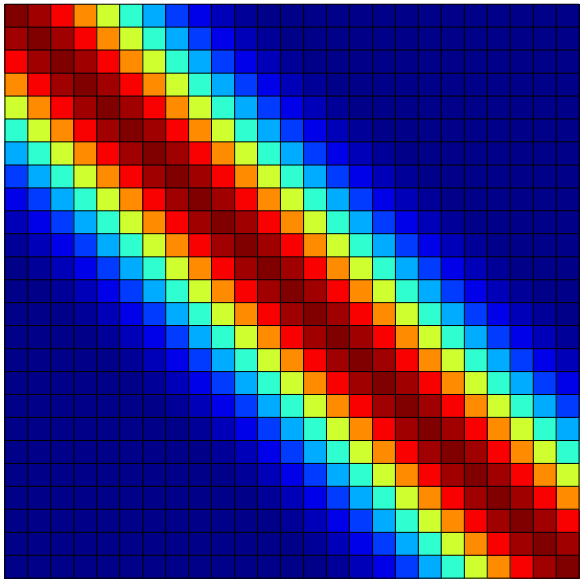
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.
- For the example above it was based on Euclidean distance.
- The covariance function is also know as a kernel.



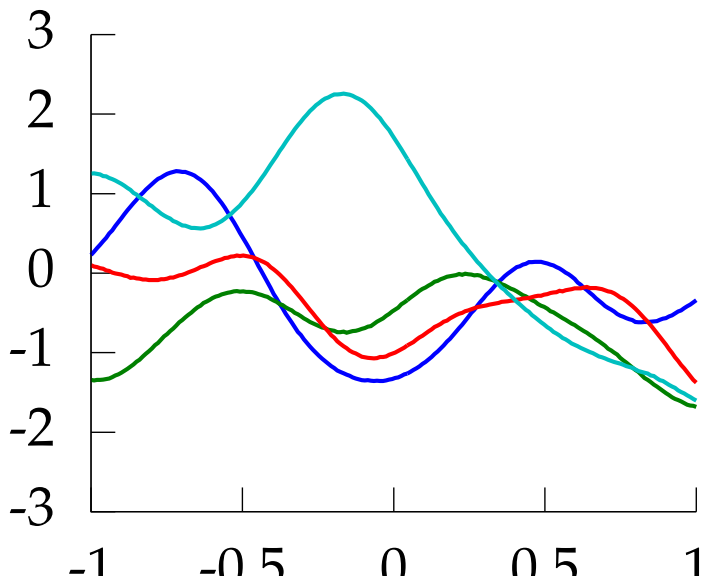
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.
- For the example above it was based on Euclidean distance.
- The covariance function is also know as a kernel.



Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_1 = -3.0, x_1 = -3.0$$

$$k_{1,1} = 1.00 \times \exp\left(-\frac{(-3.0 - -3.0)^2}{2 \times 2.00^2}\right)$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_1 = -3.0,\ x_1 = -3.0$$

$$k_{1,1} = 1.00 \times \exp\left(-\frac{(-3.0 - -3.0)^2}{2 \times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & & \\ & & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.20, x_1 = -3.0$$

$$k_{2,1} = 1.00 \times \exp\left(-\frac{(1.20 - -3.0)^2}{2 \times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & & \\ & & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_1 = -3.0$$

$$k_{2,1} = 1.00 \times \exp\left(-\frac{(1.20 - -3.0)^2}{2 \times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & & \\ 0.110 & & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_1 = -3.0$$

$$k_{2,1} = 1.00 \times \exp\left(-\frac{(1.20 - -3.0)^2}{2 \times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 \\ 0.110 & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_2 = 1.20$$

$$k_{2,2} = 1.00 \times \exp\left(-\frac{(1.20-1.20)^2}{2\times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 \\ 0.110 & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_2 = 1.20$$

$$k_{2,2} = 1.00 \times \exp\left(-\frac{(1.20-1.20)^2}{2\times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 \\ 0.110 & 1.00 \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40, x_1 = -3.0$$

$$\begin{bmatrix} 1.00 & 0.110 \\ 0.110 & 1.00 \end{bmatrix}$$

$$k_{3,1} = 1.00 \times \exp\left(-\frac{(1.40 - -3.0)^2}{2 \times 2.00^2}\right)$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\|x_i - x_j\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_1 = -3.0$$

$$\begin{bmatrix} 1.00 & 0.110 \\ 0.110 & 1.00 \\ 0.0889 & \end{bmatrix}$$

$$k_{3,1} = 1.00 \times \exp\left(-\frac{(1.40 - -3.0)^2}{2 \times 2.00^2}\right)$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_1 = -3.0$$

$$\begin{bmatrix} 1.00 & 0.110 & 0.0889 \\ 0.110 & 1.00 & \\ 0.0889 & & \end{bmatrix}$$

$$k_{3,1} = 1.00 \times \exp\left(-\frac{(1.40 - -3.0)^2}{2 \times 2.00^2}\right)$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_2 = 1.20$$

$$k_{3,2} = 1.00 \times \exp\left(-\frac{(1.40-1.20)^2}{2\times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 & 0.0889 \\ 0.110 & 1.00 & \\ 0.0889 & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_2 = 1.20$$

$$k_{3,2} = 1.00 \times \exp\left(-\frac{(1.40-1.20)^2}{2\times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 & 0.0889 \\ 0.110 & 1.00 & \\ 0.0889 & 0.995 & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_2 = 1.20$$

$$k_{3,2} = 1.00 \times \exp\left(-\frac{(1.40-1.20)^2}{2\times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 & 0.0889 \\ 0.110 & 1.00 & 0.995 \\ 0.0889 & 0.995 & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_3 = 1.40$$

$$k_{3,3} = 1.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2\times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 & 0.0889 \\ 0.110 & 1.00 & 0.995 \\ 0.0889 & 0.995 & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_3 = 1.40$$

$$k_{3,3} = 1.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2\times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 & 0.0889 \\ 0.110 & 1.00 & 0.995 \\ 0.0889 & 0.995 & 1.00 \end{bmatrix}$$

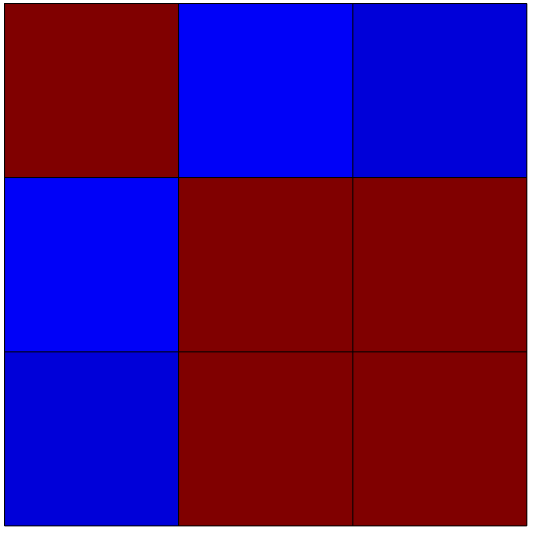
$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$x_3 = 1.40, x_3 = 1.40$



$$k_{3,3} = 1.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2\times 2.00^2}\right)$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_1 = -3, x_1 = -3$$

$$k_{1,1} = 1.0 \times \exp\left(-\frac{(-3--3)^2}{2\times 2.0^2}\right)$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_1 = -3, x_1 = -3$$

$$k_{1,1} = 1.0 \times \exp\left(-\frac{(-3--3)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & & \\ & & \\ & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.2, x_1 = -3$$

$$k_{2,1} = 1.0 \times \exp\left(-\frac{(1.2 - -3)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & & \\ & & \\ & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.2, x_1 = -3$$

$$k_{2,1} = 1.0 \times \exp\left(-\frac{(1.2 - -3)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & & \\ 0.11 & & \\ & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.2, x_1 = -3$$

$$k_{2,1} = 1.0 \times \exp\left(-\frac{(1.2 - -3)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 \\ 0.11 & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.2, x_2 = 1.2$$

$$\begin{bmatrix} 1.0 & 0.11 \\ 0.11 & \end{bmatrix}$$

$$k_{2,2} = 1.0 \times \exp\left(-\frac{(1.2-1.2)^2}{2\times 2.0^2}\right)$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.2,\ x_2 = 1.2$$

$$k_{2,2} = 1.0 \times \exp\left(-\frac{(1.2-1.2)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 \\ 0.11 & 1.0 \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.4, x_1 = -3$$

$$k_{3,1} = 1.0 \times \exp\left(-\frac{(1.4 - -3)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 \\ 0.11 & 1.0 \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.4,\ x_1 = -3$$

$$k_{3,1} = 1.0 \times \exp\left(-\frac{(1.4 - -3)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & \\ 0.11 & 1.0 & \\ \boxed{0.089} & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.4, x_1 = -3$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 \\ 0.11 & 1.0 & \\ 0.089 & & \end{bmatrix}$$

$$k_{3,1} = 1.0 \times \exp\left(-\frac{(1.4 - -3)^2}{2 \times 2.0^2}\right)$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.4,\ x_2 = 1.2$$

$$k_{3,2} = 1.0 \times \exp\left(-\frac{(1.4-1.2)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 \\ 0.11 & 1.0 & \\ 0.089 & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.4,\ x_2 = 1.2$$

$$k_{3,2} = 1.0 \times \exp\left(-\frac{(1.4-1.2)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 \\ 0.11 & 1.0 & \\ 0.089 & 1.0 & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.4,\ x_2 = 1.2$$

$$k_{3,2} = 1.0 \times \exp\left(-\frac{(1.4-1.2)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 \\ 0.11 & 1.0 & 1.0 \\ 0.089 & 1.0 & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.4,\ x_3 = 1.4$$

$$k_{3,3} = 1.0 \times \exp\left(-\frac{(1.4-1.4)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 \\ 0.11 & 1.0 & 1.0 \\ 0.089 & 1.0 & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.4,\ x_3 = 1.4$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 \\ 0.11 & 1.0 & 1.0 \\ 0.089 & 1.0 & 1.0 \end{bmatrix}$$

$$k_{3,3} = 1.0 \times \exp\left(-\frac{(1.4-1.4)^2}{2\times 2.0^2}\right)$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_4 = 2.0, x_1 = -3$$

$$k_{4,1} = 1.0 \times \exp\left(-\frac{(2.0 - -3)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 \\ 0.11 & 1.0 & 1.0 \\ 0.089 & 1.0 & 1.0 \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_1 = -3$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 \\ 0.11 & 1.0 & 1.0 \\ 0.089 & 1.0 & 1.0 \\ 0.044 & & \end{bmatrix}$$

$$k_{4,1} = 1.0 \times \exp\left(-\frac{(2.0 - -3)^2}{2 \times 2.0^2}\right)$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_1 = -3$$

$$k_{4,1} = 1.0 \times \exp\left(-\frac{(2.0 - -3)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & \\ 0.089 & 1.0 & 1.0 & \\ 0.044 & & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_2 = 1.2$$

$$k_{4,2} = 1.0 \times \exp\left(-\frac{(2.0-1.2)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & \\ 0.089 & 1.0 & 1.0 & \\ 0.044 & & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_2 = 1.2$$

$$k_{4,2} = 1.0 \times \exp\left(-\frac{(2.0-1.2)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & \\ 0.089 & 1.0 & 1.0 & \\ 0.044 & 0.92 & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_2 = 1.2$$

$$k_{4,2} = 1.0 \times \exp\left(-\frac{(2.0-1.2)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & 0.92 \\ 0.089 & 1.0 & 1.0 & \\ 0.044 & 0.92 & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_3 = 1.4$$

$$k_{4,3} = 1.0 \times \exp\left(-\frac{(2.0-1.4)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & 0.92 \\ 0.089 & 1.0 & 1.0 & \\ 0.044 & 0.92 & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_3 = 1.4$$

$$k_{4,3} = 1.0 \times \exp\left(-\frac{(2.0-1.4)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & 0.92 \\ 0.089 & 1.0 & 1.0 & \\ 0.044 & 0.92 & 0.96 & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_3 = 1.4$$

$$k_{4,3} = 1.0 \times \exp\left(-\frac{(2.0-1.4)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & 0.92 \\ 0.089 & 1.0 & 1.0 & 0.96 \\ 0.044 & 0.92 & 0.96 & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_4 = 2.0, x_4 = 2.0$$

$$k_{4,4} = 1.0 \times \exp\left(-\frac{(2.0-2.0)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & 0.92 \\ 0.089 & 1.0 & 1.0 & 0.96 \\ 0.044 & 0.92 & 0.96 & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_4 = 2.0$$

$$k_{4,4} = 1.0 \times \exp\left(-\frac{(2.0-2.0)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & 0.92 \\ 0.089 & 1.0 & 1.0 & 0.96 \\ 0.044 & 0.92 & 0.96 & 1.0 \end{bmatrix}$$

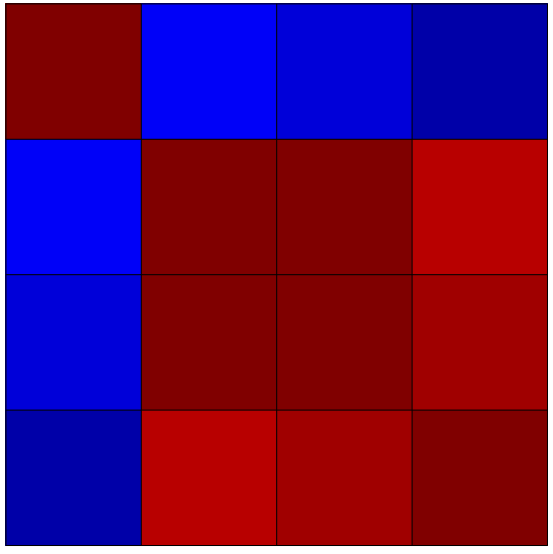
$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$x_4 = 2.0, x_4 = 2.0$



$$k_{4,4} = 1.0 \times \exp\left(-\frac{(2.0-2.0)^2}{2\times 2.0^2}\right)$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_1 = -3.0, x_1 = -3.0$$

$$k_{1,1} = 4.00 \times \exp\left(-\frac{(-3.0 - -3.0)^2}{2 \times 5.00^2}\right)$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$\begin{bmatrix} 4.00 & & \\ & & \\ & & \end{bmatrix}$$

$$x_1 = -3.0, x_1 = -3.0$$

$$k_{1,1} = 4.00 \times \exp\left(-\frac{(-3.0 - -3.0)^2}{2 \times 5.00^2}\right)$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_1 = -3.0$$

$$k_{2,1} = 4.00 \times \exp\left(-\frac{(1.20 - -3.0)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & & \\ & & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_1 = -3.0$$

$$k_{2,1} = 4.00 \times \exp\left(-\frac{(1.20 - -3.0)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & & \\ & 2.81 & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_1 = -3.0$$

$$k_{2,1} = 4.00 \times \exp\left(-\frac{(1.20 - -3.0)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 \\ 2.81 & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_2 = 1.20$$

$$k_{2,2} = 4.00 \times \exp\left(-\frac{(1.20-1.20)^2}{2\times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 \\ 2.81 & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_2 = 1.20$$

$$k_{2,2} = 4.00 \times \exp\left(-\frac{(1.20-1.20)^2}{2\times 5.00^2}\right)$$

$$\begin{bmatrix} & & \\ & 4.00 & 2.81 \\ & 2.81 & 4.00 \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40, x_1 = -3.0$$

$$\begin{bmatrix} 4.00 & 2.81 \\ 2.81 & 4.00 \end{bmatrix}$$

$$k_{3,1} = 4.00 \times \exp\left(-\frac{(1.40 - -3.0)^2}{2 \times 5.00^2}\right)$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_1 = -3.0$$

$$\begin{bmatrix} 4.00 & 2.81 \\ 2.81 & 4.00 \end{bmatrix}$$

$$k_{3,1} = 4.00 \times \exp\left(-\frac{(1.40 - -3.0)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} 2.72 \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_1 = -3.0$$

$$\begin{bmatrix} 4.00 & 2.81 & 2.72 \\ 2.81 & 4.00 & \\ 2.72 & & \end{bmatrix}$$

$$k_{3,1} = 4.00 \times \exp\left(-\frac{(1.40 - -3.0)^2}{2 \times 5.00^2}\right)$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_2 = 1.20$$

$$k_{3,2} = 4.00 \times \exp\left(-\frac{(1.40-1.20)^2}{2\times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 & 2.72 \\ 2.81 & 4.00 & \\ 2.72 & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_2 = 1.20$$

$$\begin{bmatrix} 4.00 & 2.81 & 2.72 \\ 2.81 & 4.00 & \\ 2.72 & 4.00 & \end{bmatrix}$$

$$k_{3,2} = 4.00 \times \exp\left(-\frac{(1.40-1.20)^2}{2\times 5.00^2}\right)$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_2 = 1.20$$

$$k_{3,2} = 4.00 \times \exp\left(-\frac{(1.40-1.20)^2}{2\times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 & 2.72 \\ 2.81 & 4.00 & 4.00 \\ 2.72 & 4.00 & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40, x_3 = 1.40$$

$$k_{3,3} = 4.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2\times 5.00^2}\right)$$

$$\begin{bmatrix} & & \\ 4.00 & 2.81 & 2.72 \\ & & \\ 2.81 & 4.00 & 4.00 \\ & & \\ 2.72 & 4.00 & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40, x_3 = 1.40$$

$$k_{3,3} = 4.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2\times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 & 2.72 \\ 2.81 & 4.00 & 4.00 \\ 2.72 & 4.00 & 4.00 \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

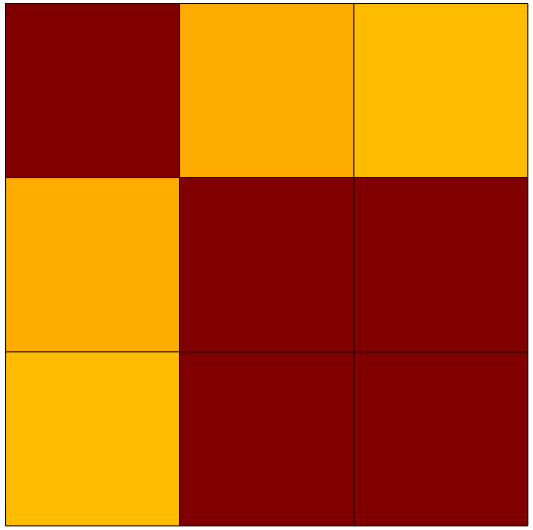
Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_3 = 1.40$$

$$k_{3,3} = 4.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2\times 5.00^2}\right)$$



$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Basis Function Form

Radial basis functions commonly have the form

$$\phi_k(\mathbf{x}_i) = \exp\left(-\frac{\left|\mathbf{x}_i - \boldsymbol{\mu}_k\right|^2}{2\ell^2}\right).$$

- Basis function maps data into a "feature space" in which a linear sum is a non linear function.

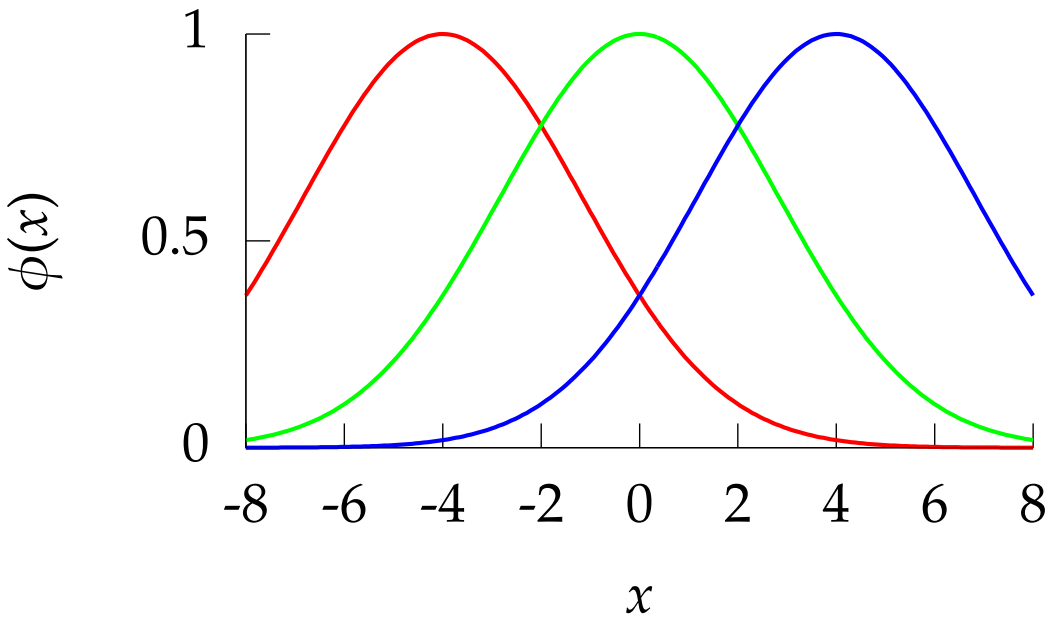


Figure: A set of radial basis functions with width $\ell = 2$ and location parameters $\boldsymbol{\mu} = [-4 \;\; 0 \;\; 4]^\top$.

Basis Function Representations

- Represent a function by a linear sum over a basis,

$$f(\mathbf{x}_{i,:}; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(\mathbf{x}_{i,:}), \tag{2}$$

- Here: m basis functions and $\phi_k(\cdot)$ is kth basis function and

$$\mathbf{w} = [w_1, \ldots, w_m]^\top .$$

- For standard linear model: $\phi_k(\mathbf{x}_{i,:}) = x_{i,k}$.

Random Functions

Functions derived using:

$$f(x) = \sum_{k=1}^{m} w_k \phi_k(x),$$

where $\mathbf{W}$ is sampled from a Gaussian density,

$$w_k \sim \mathcal{N}(0, \alpha).$$

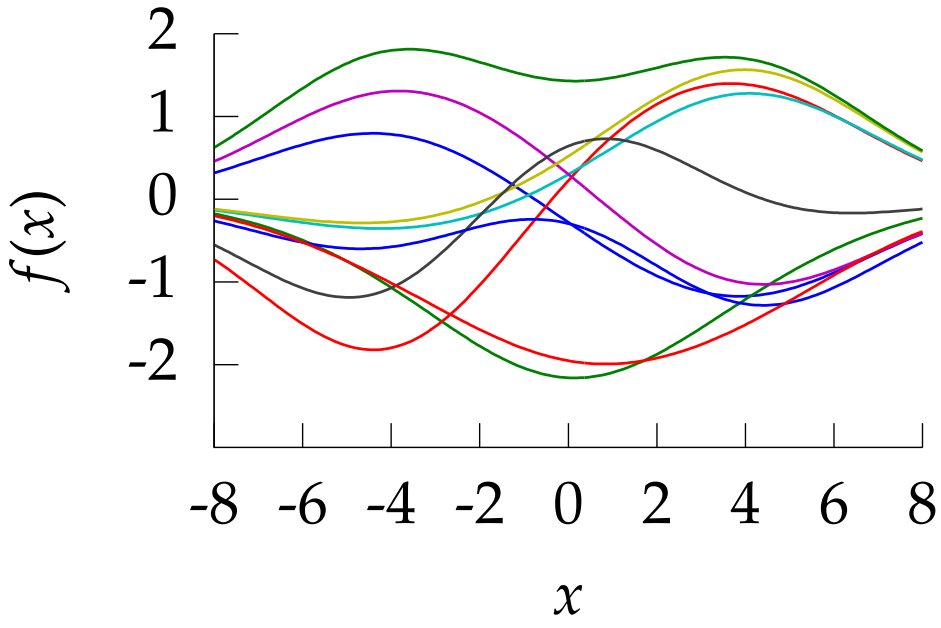


Figure: Functions sampled using the basis set from figure 8. Each line is a separate sample, generated by a weighted sum of the basis set. The weights, $\mathbf{w}$ are sampled from a Gaussian density with variance $\alpha = 1$.

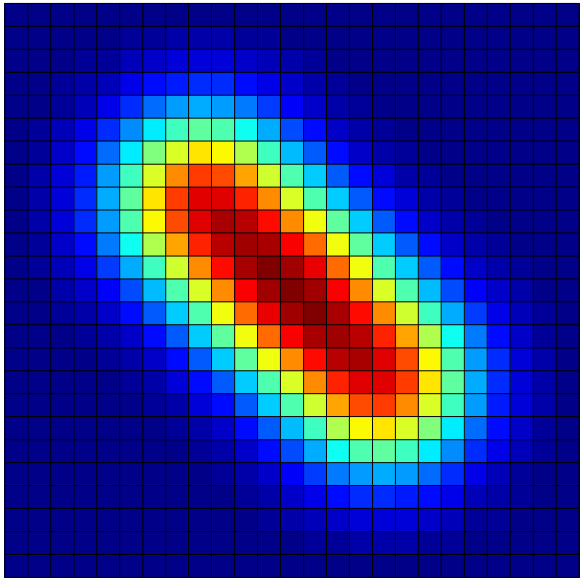
Covariance Functions

RBF Basis Functions

$$k(\mathbf{x}, \mathbf{x}') = \alpha \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{x}')$$

$$\phi_i(x) = \exp\left(-\frac{\left\|x - \mu_i\right\|_2^2}{\ell^2}\right)$$

$$\boldsymbol{\mu} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$



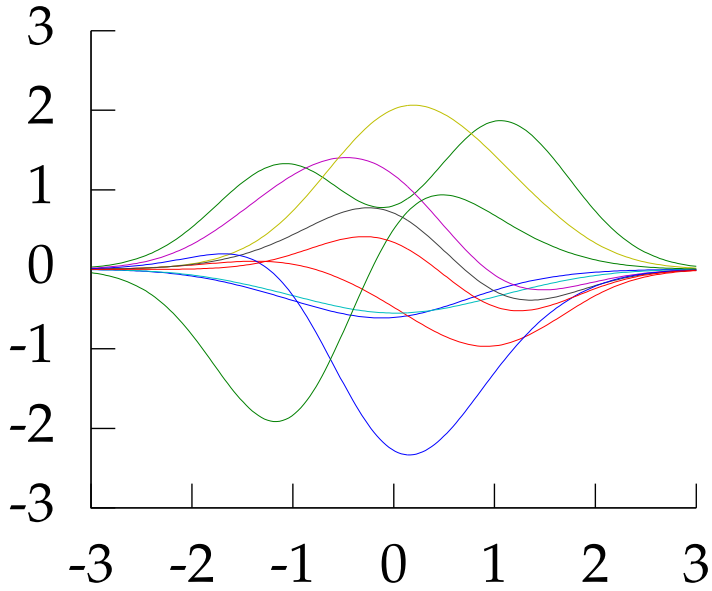
Covariance Functions

RBF Basis Functions

$$k(\mathbf{x}, \mathbf{x}') = \alpha \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{x}')$$

$$\phi_i(x) = \exp\left(-\frac{\left\|x - \mu_i\right\|_2^2}{\ell^2}\right)$$

$$\mu = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$



Direct Construction of Covariance Matrix

- Use matrix notation to write function,

$$f(\mathbf{x}_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(\mathbf{x}_i)$$

Direct Construction of Covariance Matrix

- Use matrix notation to write function,

$$f(\mathbf{x}_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(\mathbf{x}_i)$$

computed at training data gives a vector

$$\mathbf{f} = \mathbf{\Phi w}.$$

Direct Construction of Covariance Matrix

- Use matrix notation to write function,

$$f(\mathbf{x}_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(\mathbf{x}_i)$$

computed at training data gives a vector

$$\mathbf{f} = \mathbf{\Phi w}.$$

$\mathbf{w}$ and $\mathbf{f}$ are only related by an *inner product*.

Direct Construction of Covariance Matrix

- Use matrix notation to write function,

$$f(\mathbf{x}_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(\mathbf{x}_i)$$

computed at training data gives a vector

$$\mathbf{f} = \mathbf{\Phi w}.$$

$\mathbf{w}$ and $\mathbf{f}$ are only related by an *inner product*.

$\mathbf{\Phi} \in \Re^{n \times p}$ is a *design matrix*

Direct Construction of Covariance Matrix

- Use matrix notation to write function,

$$f(\mathbf{x}_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(\mathbf{x}_i)$$

computed at training data gives a vector

$$\mathbf{f} = \mathbf{\Phi w}.$$

$\mathbf{w}$ and $\mathbf{f}$ are only related by an *inner product*.

$\mathbf{\Phi} \in \Re^{n \times p}$ is a *design matrix*

$\mathbf{\Phi}$ is fixed and non-stochastic for a given training set.

Direct Construction of Covariance Matrix

- Use matrix notation to write function,

$$f(\mathbf{x}_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(\mathbf{x}_i)$$

computed at training data gives a vector

$$\mathbf{f} = \mathbf{\Phi}\mathbf{w}.$$

$\mathbf{w}$ and $\mathbf{f}$ are only related by an *inner product*.

$\mathbf{\Phi} \in \Re^{n \times p}$ is a *design matrix*

$\mathbf{\Phi}$ is fixed and non-stochastic for a given training set.

$\mathbf{f}$ is Gaussian distributed.

Expectations

- We have

$$\langle \mathbf{f} \rangle = \mathbf{\Phi} \langle \mathbf{w} \rangle .$$

We use $\langle \cdot \rangle$ to denote expectations under prior distributions.

Expectations

- We have
$$\langle \mathbf{f} \rangle = \mathbf{\Phi} \langle \mathbf{w} \rangle .$$
- Prior mean of $\mathbf{w}$ was zero giving
$$\langle \mathbf{f} \rangle = \mathbf{0}.$$

We use $\langle \cdot \rangle$ to denote expectations under prior distributions.

Expectations

- We have
$$\langle \mathbf{f} \rangle = \mathbf{\Phi} \langle \mathbf{w} \rangle .$$
- Prior mean of $\mathbf{w}$ was zero giving
$$\langle \mathbf{f} \rangle = \mathbf{0}.$$
- Prior covariance of $\mathbf{f}$ is
$$\mathbf{K} = \left\langle \mathbf{f}\mathbf{f}^\top \right\rangle - \langle \mathbf{f} \rangle \langle \mathbf{f} \rangle^\top$$

We use $\langle \cdot \rangle$ to denote expectations under prior distributions.

Expectations

- We have
$$\langle \mathbf{f} \rangle = \mathbf{\Phi} \langle \mathbf{w} \rangle .$$
- Prior mean of $\mathbf{w}$ was zero giving
$$\langle \mathbf{f} \rangle = \mathbf{0}.$$
- Prior covariance of $\mathbf{f}$ is
$$\mathbf{K} = \left\langle \mathbf{f}\mathbf{f}^\top \right\rangle - \langle \mathbf{f} \rangle \langle \mathbf{f} \rangle^\top$$
$$\left\langle \mathbf{f}\mathbf{f}^\top \right\rangle = \mathbf{\Phi} \left\langle \mathbf{w}\mathbf{w}^\top \right\rangle \mathbf{\Phi}^\top,$$
giving
$$\mathbf{K} = \gamma' \mathbf{\Phi}\mathbf{\Phi}^\top.$$

We use $\langle \cdot \rangle$ to denote expectations under prior distributions.

Back to the full model

$$\mathbf{y}_g \sim \mathcal{N}\Big(\underbrace{\mu_g \mathbf{1}}_{\text{mean}}, \underbrace{\mathbf{K_S}}_{\substack{\text{SNP} \\ \text{effects}}} + \underbrace{\mathbf{K_X}}_{\substack{\text{direct factor} \\ \text{effects}}} + \underbrace{\mathbf{K_I}}_{\substack{\text{SNP-factor} \\ \text{interactions}}} + \underbrace{\sigma_p^2 \mathbf{K_P}}_{\substack{\text{population} \\ \text{structure}}} + \underbrace{\sigma_e^2 \mathbf{I}}_{\text{noise}} \Big)$$

$$p(\mathbf{Y} \,|\, \mathbf{S}, \mathbf{X}, \mathbf{\Theta}_{\mathrm{K}}, \mathcal{I}, \mathcal{S}) = \prod_{g=1}^{G} \mathcal{N}(\mathbf{y}_g \,|\, \mathbf{0}, \mathbf{\Sigma}),$$

$$\mathbf{\Sigma} = \underbrace{\sum_{\forall k \in \mathcal{S}} \beta_k^2 \mathbf{s}_k \mathbf{s}_k^\top}_{\mathbf{K_S}} + \underbrace{\sum_{q=1}^{Q} \alpha_q^2 \mathbf{x}_q \mathbf{x}_q^\top}_{\mathbf{K_X}} + \underbrace{\sum_{\forall (k,q) \in \mathcal{I}} \gamma_{k,q}^2 (\mathbf{s}_k \odot \mathbf{x}_q)(\mathbf{s}_k \odot \mathbf{x}_q)^{\mathrm{T}}}_{\mathbf{K_I}} + \sigma_p^2 \mathbf{K_P} + \sigma_e^2 \mathbf{I}$$

Gaussian Process Interpolation

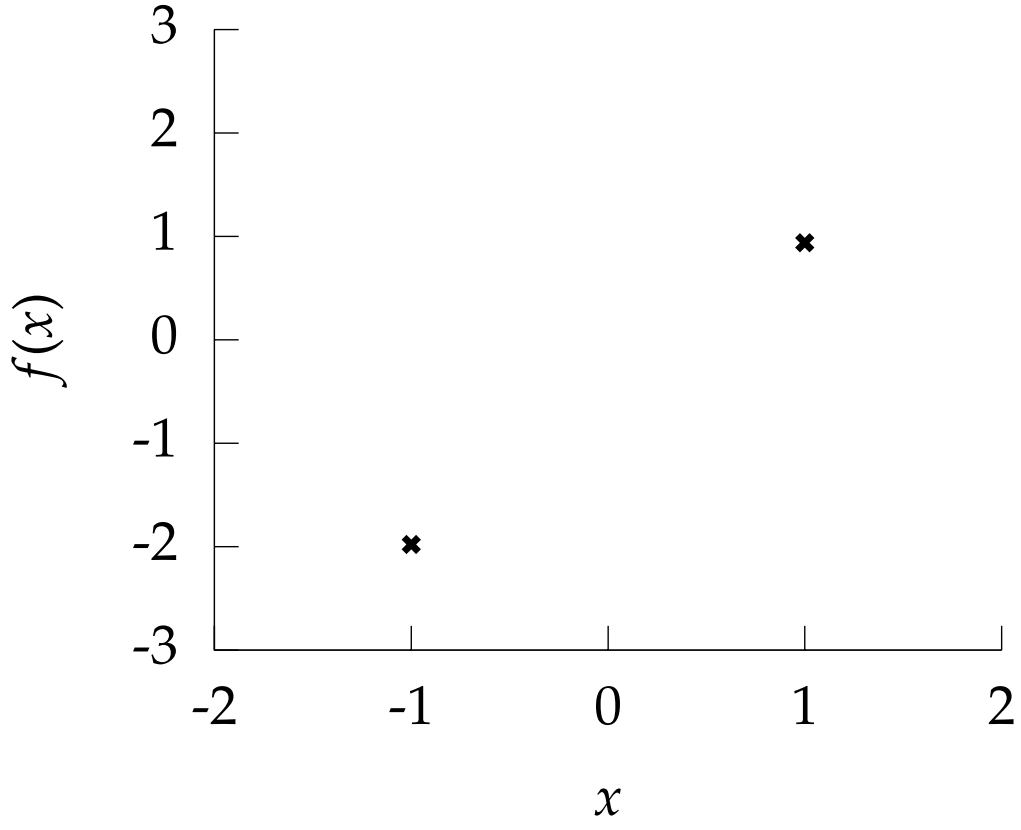


Figure: Real example: BACCO (see *e.g.* (Oakley and O'Hagan, 2002)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

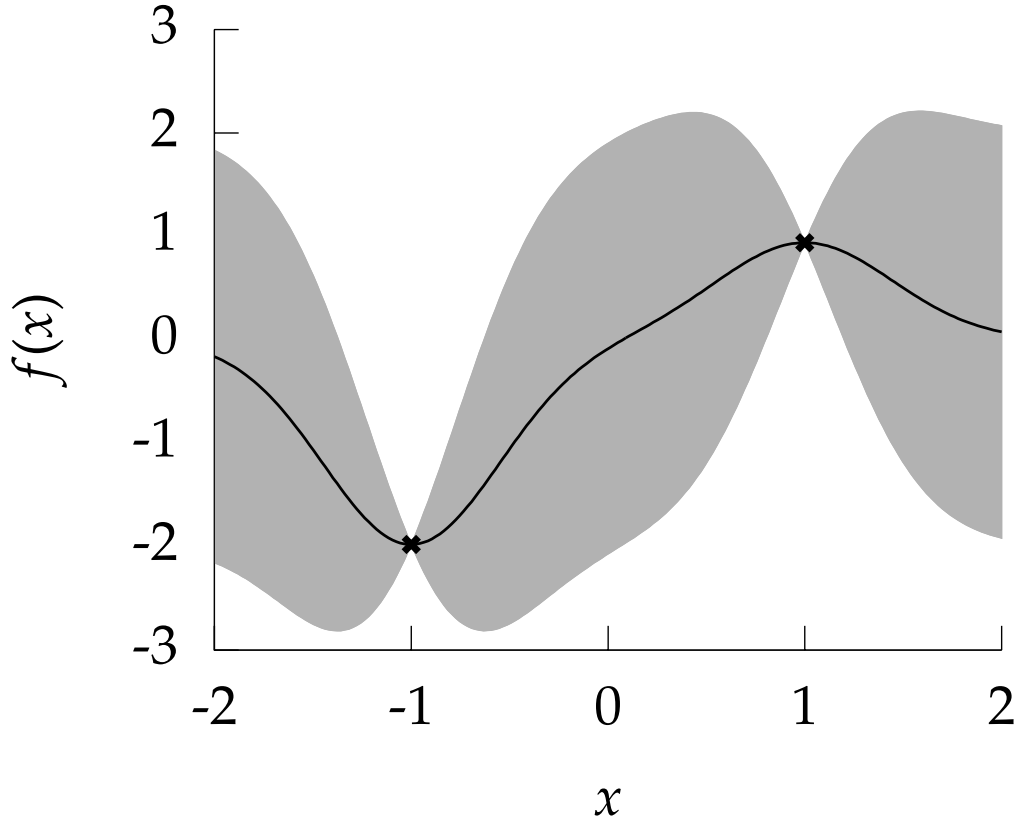


Figure: Real example: BACCO (see *e.g.* (Oakley and O'Hagan, 2002)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

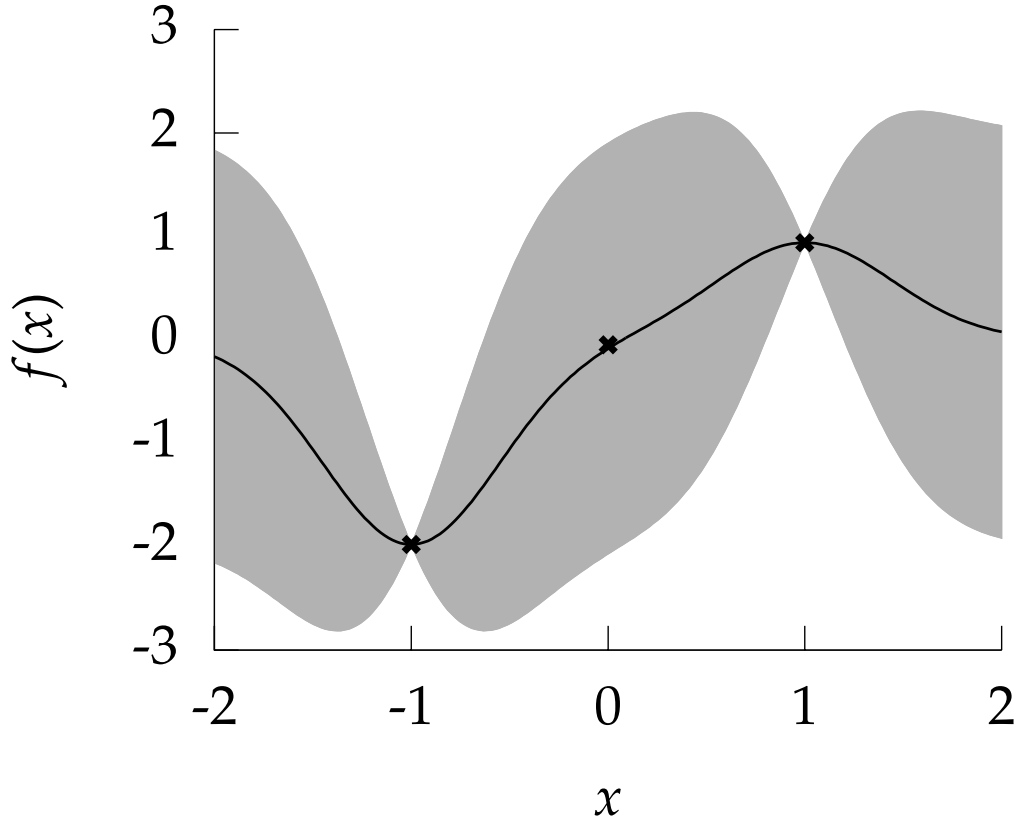


Figure: Real example: BACCO (see *e.g.* (Oakley and O'Hagan, 2002)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

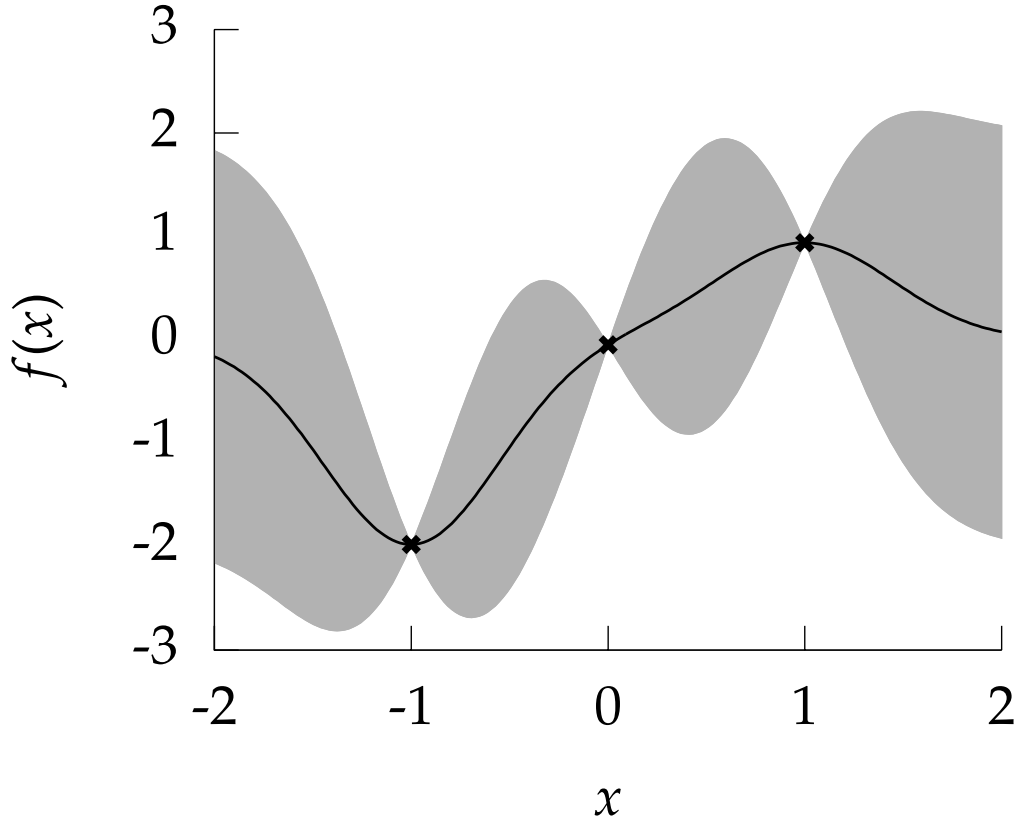


Figure: Real example: BACCO (see *e.g.* (Oakley and O'Hagan, 2002)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

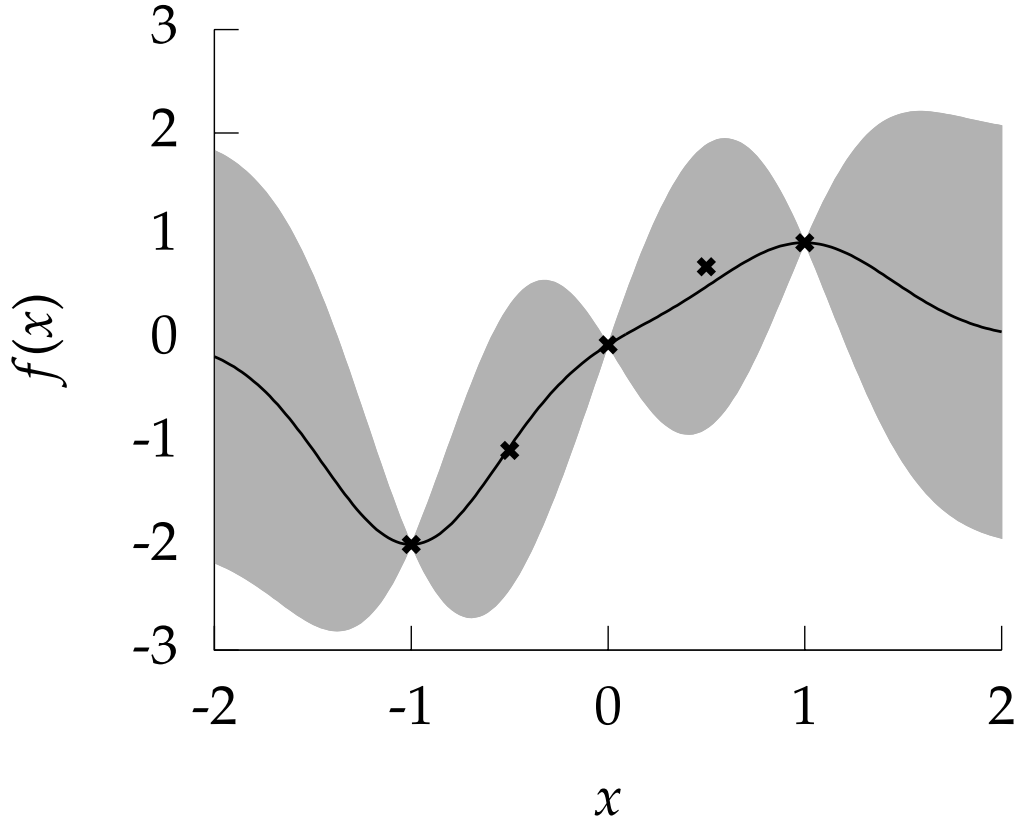


Figure: Real example: BACCO (see *e.g.* (Oakley and O'Hagan, 2002)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

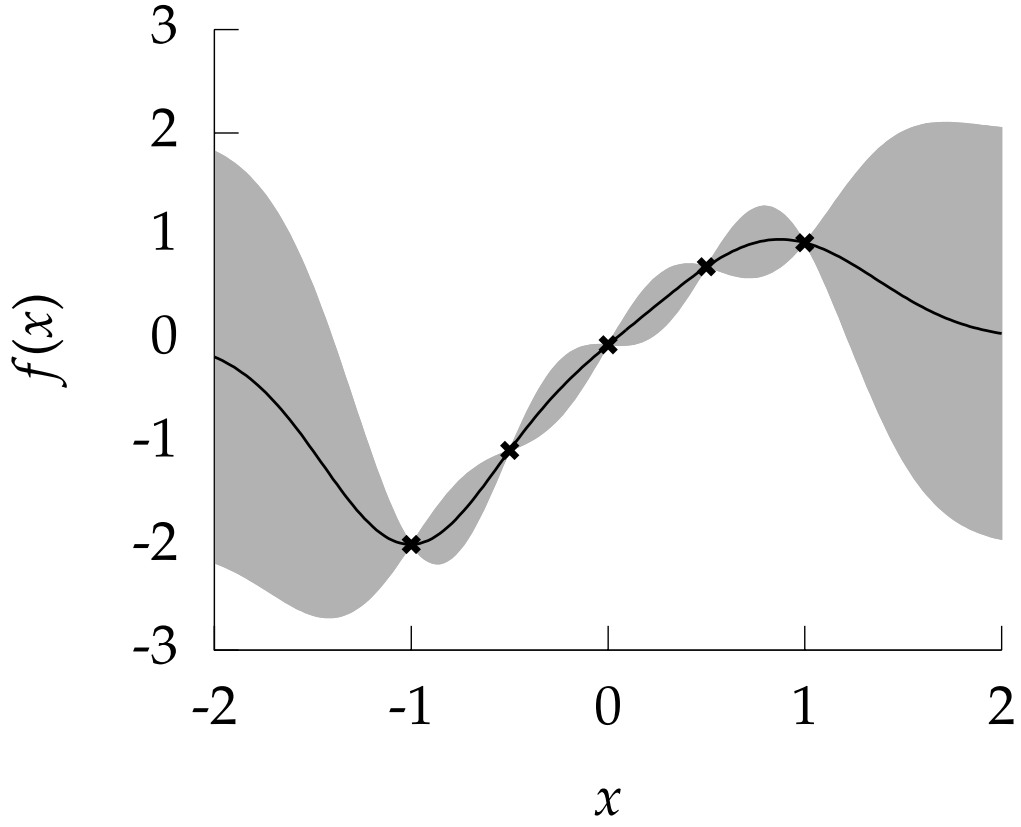


Figure: Real example: BACCO (see *e.g.* (Oakley and O'Hagan, 2002)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

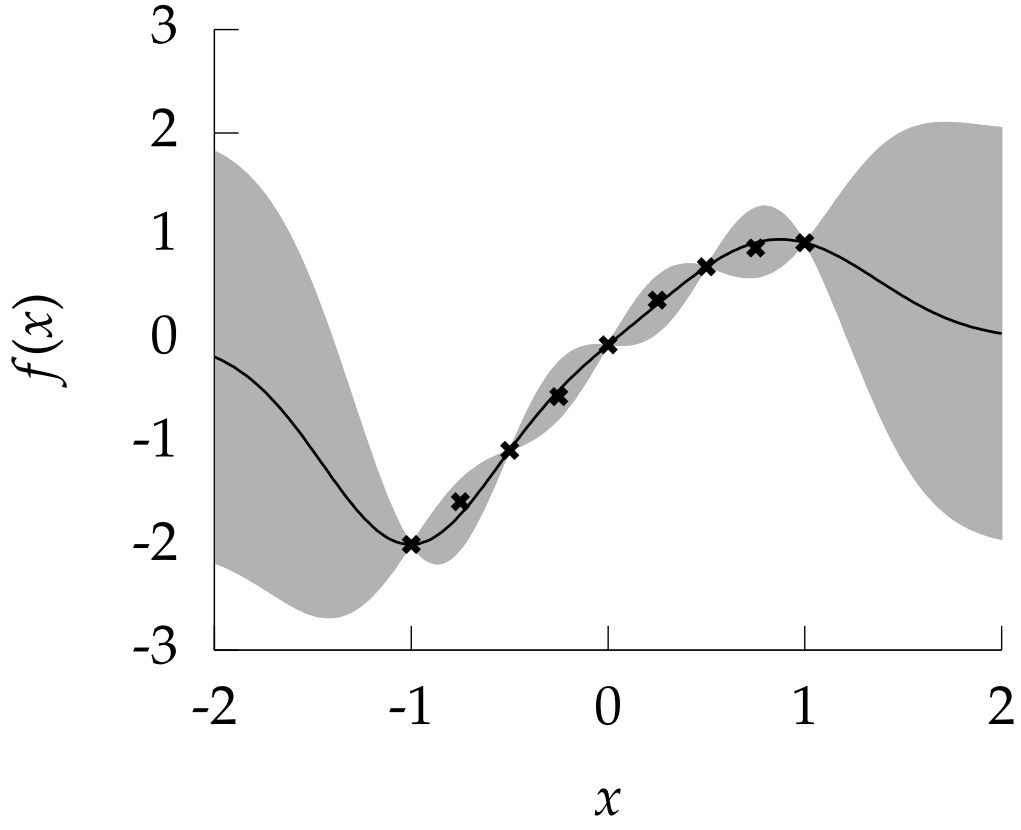


Figure: Real example: BACCO (see *e.g.* (Oakley and O'Hagan, 2002)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

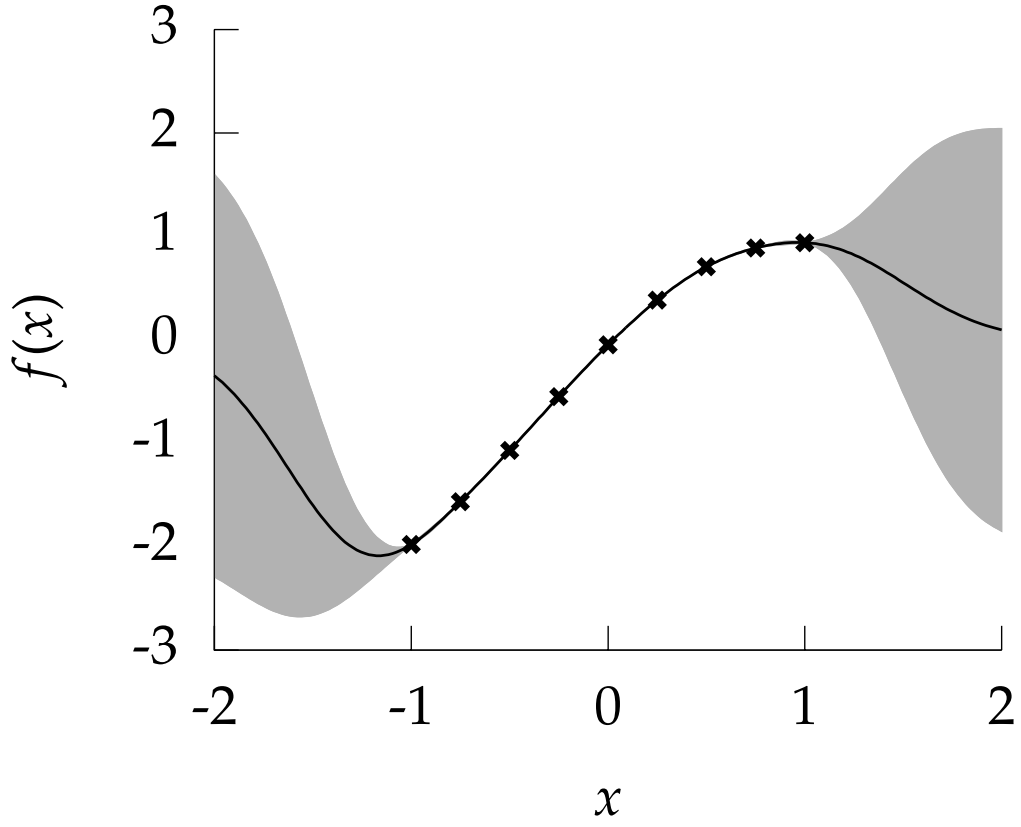


Figure: Real example: BACCO (see *e.g.* (Oakley and O'Hagan, 2002)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Regression

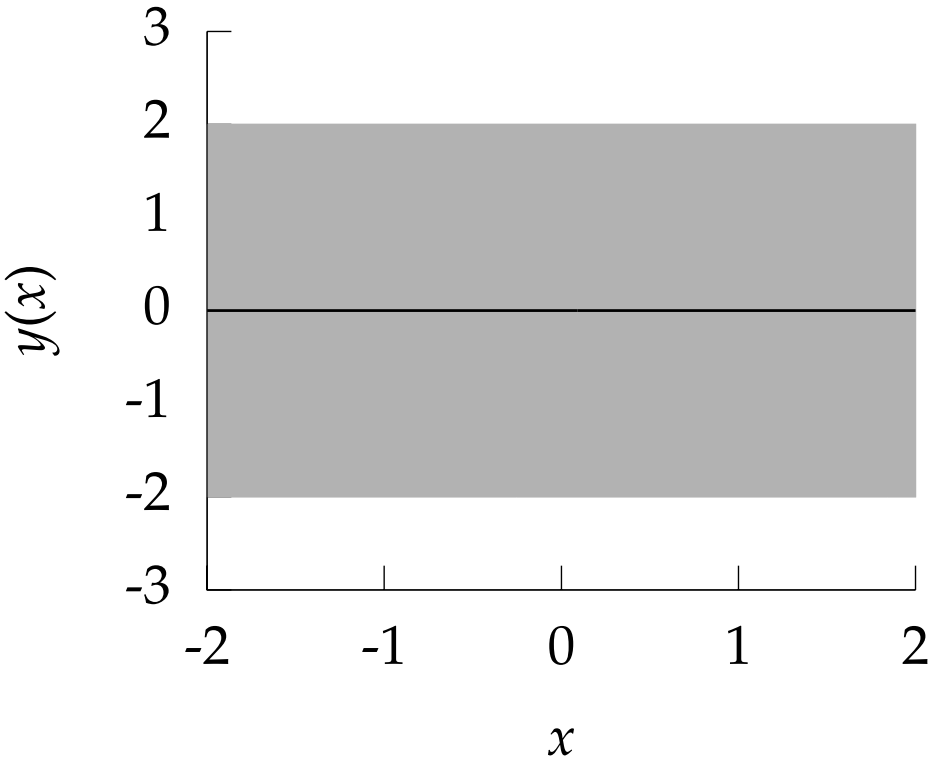


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

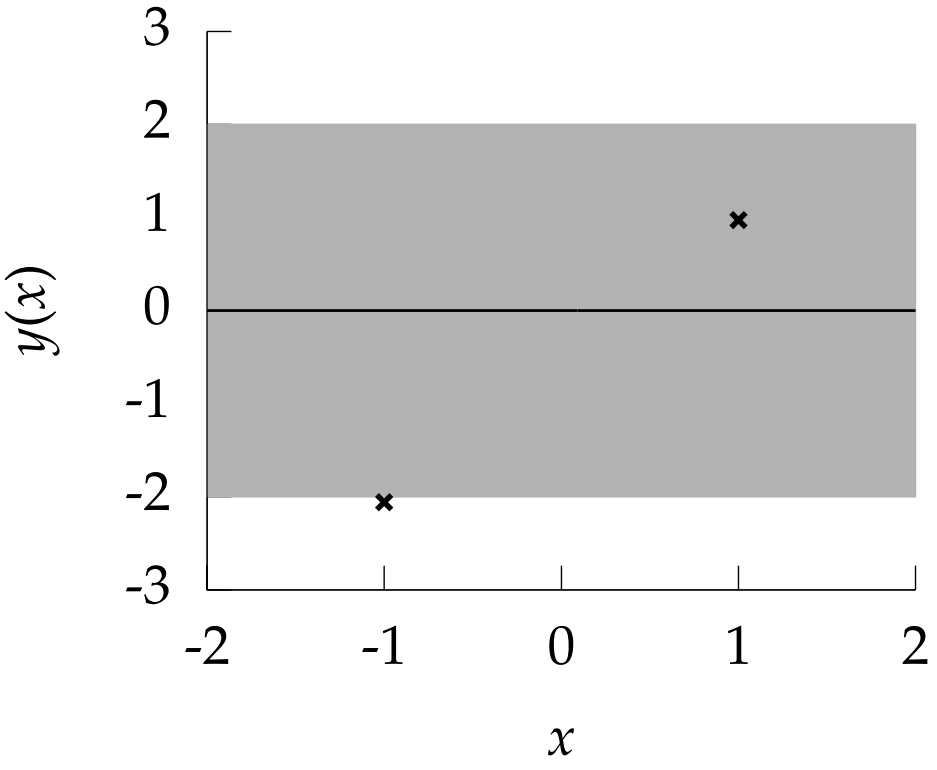


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

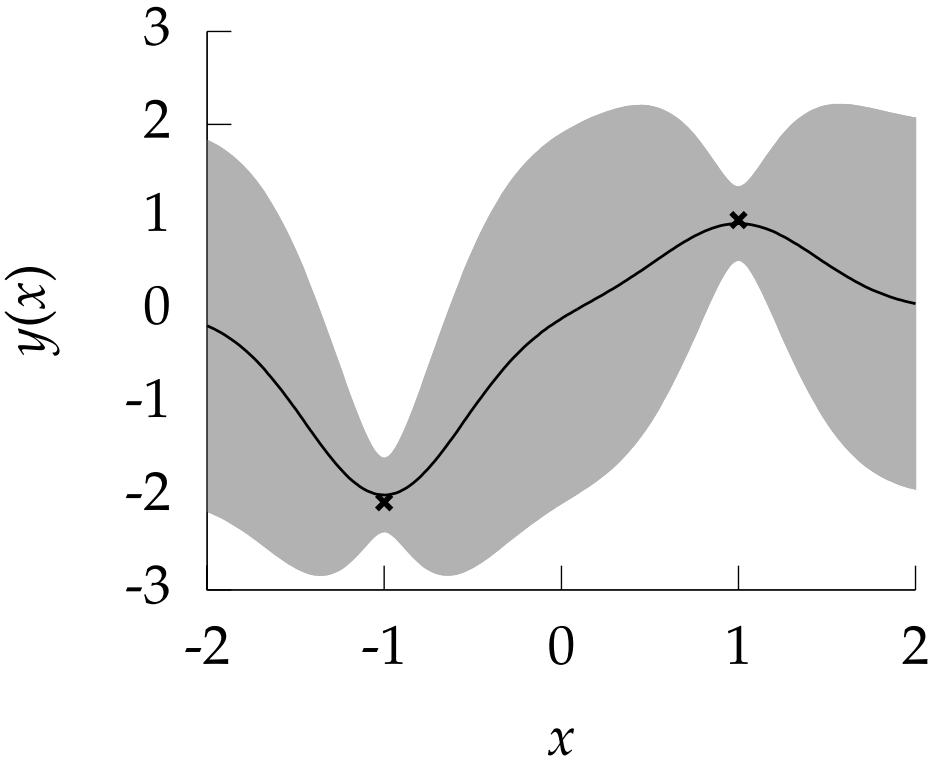


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

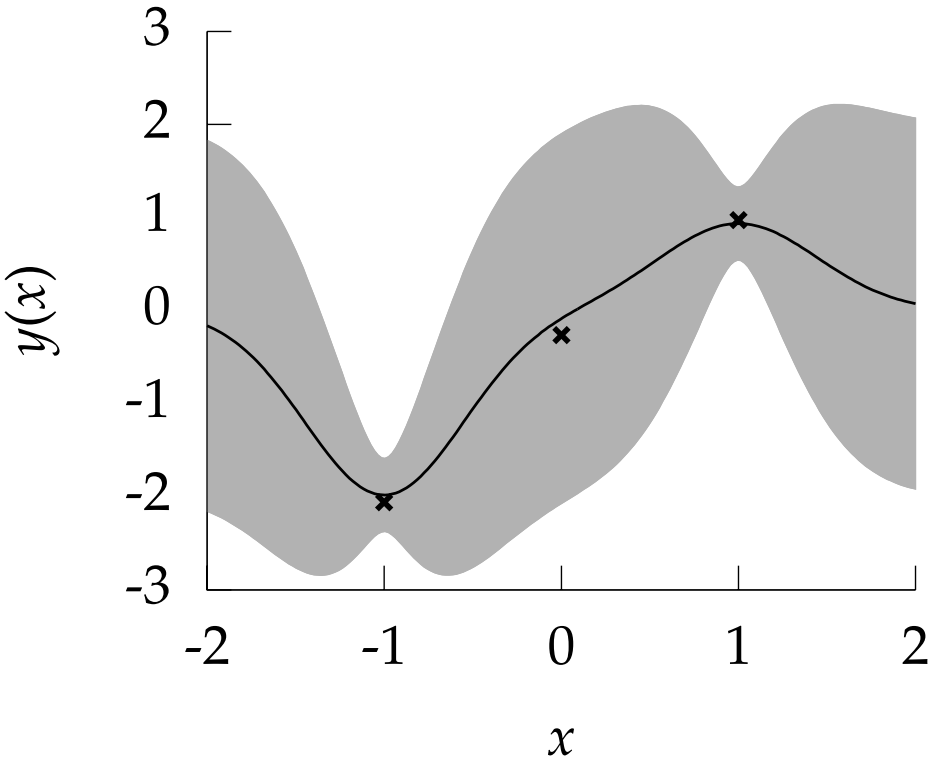


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

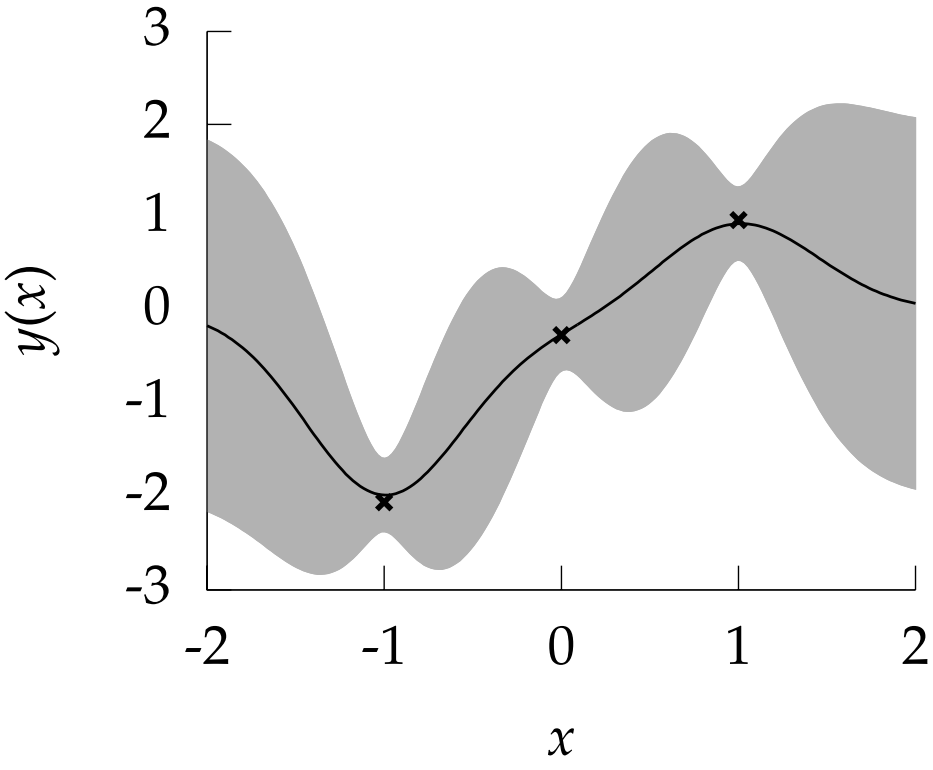


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

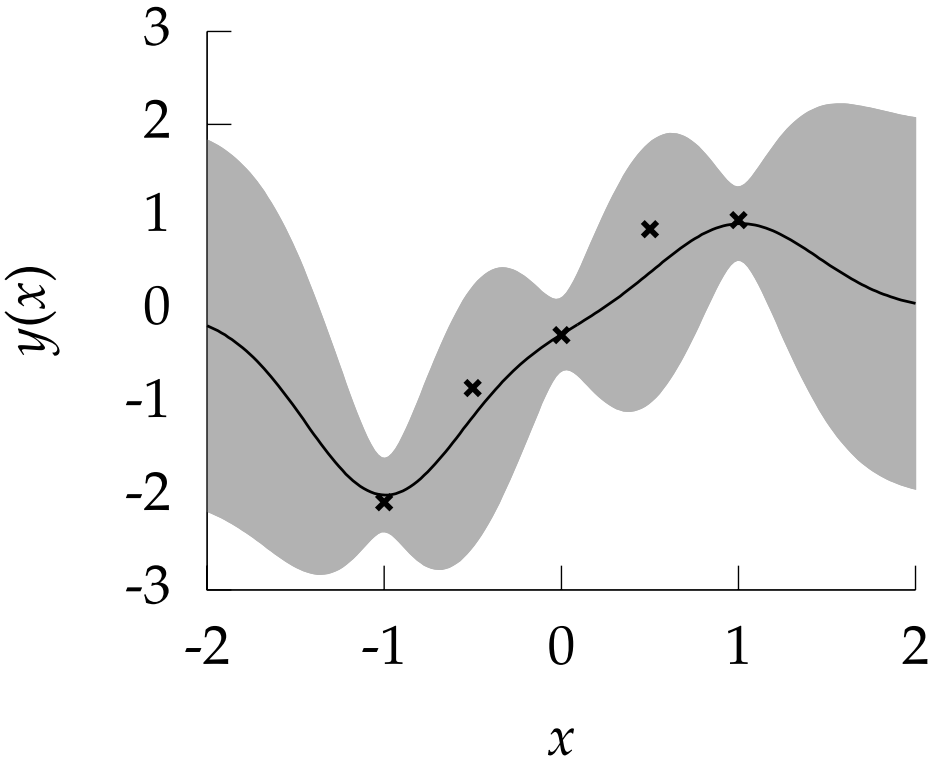


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

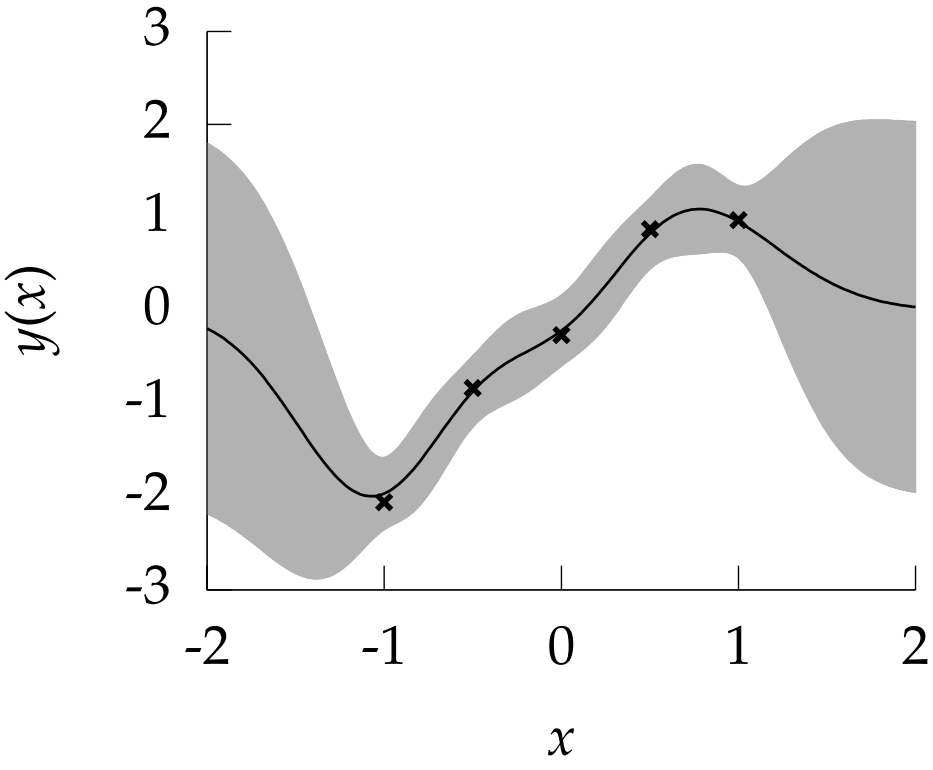


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

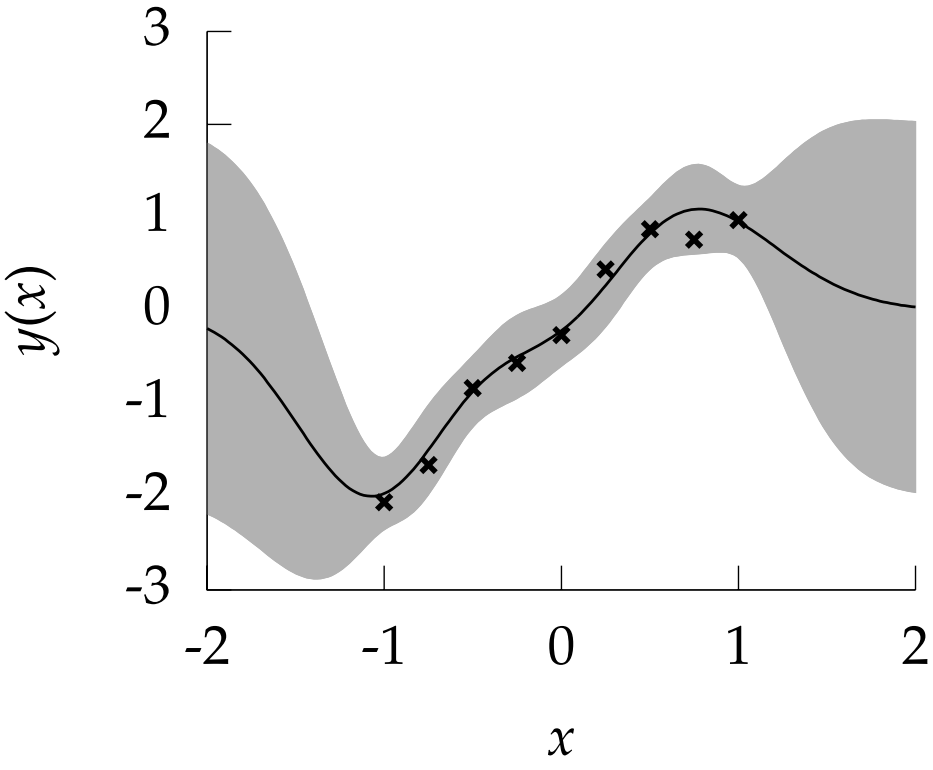


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

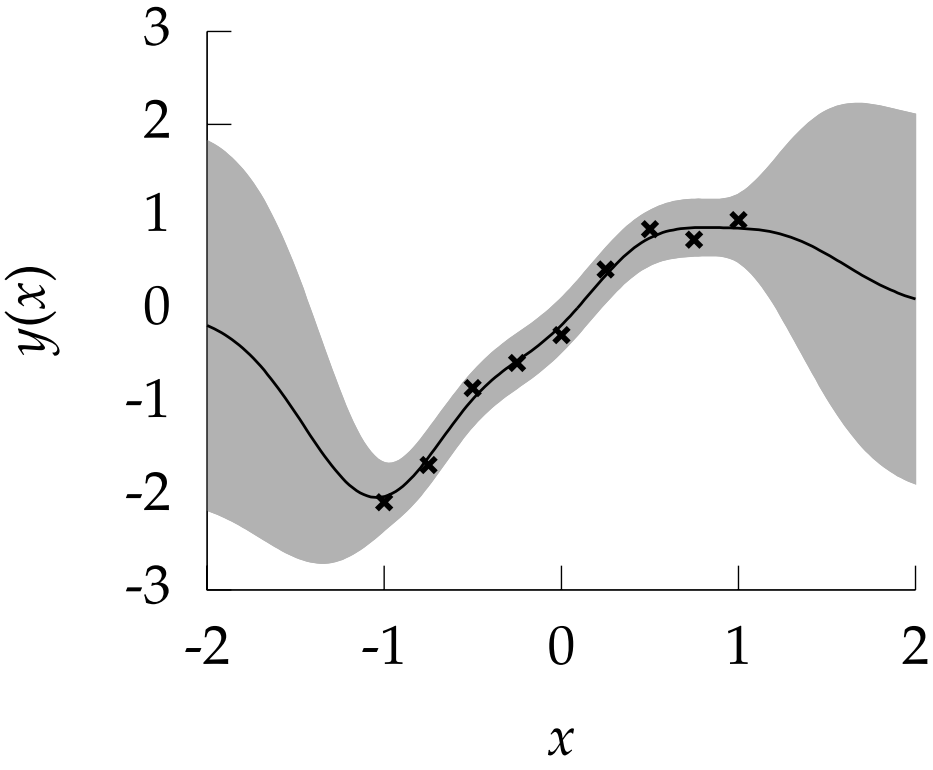


Figure: Examples include WiFi localization, C14 callibration curve.

Dealing with Non Gaussian Data

- Marginalization property of Gaussians very attractive.
- How to incorporate non-Gaussian data?
  - Data which isn't missing at random.
  - Binary data.
  - Ordinal categorical data.
  - Poisson counts.
  - Outliers.

Project Back into Gaussian

- Combine non-Gaussian likelihood with Gaussian prior.
- Either:
  - Project back to Gaussian posterior that is nearest in KL sense.
  - Expectation propagation.
- Or:
  - Fit a locally valid Gaussian approximation.
  - Laplace Approximation.

Ongoing work with Ricardo Andrade Pacheco (EP) and Alan Saul (Laplace) also James Hensman.

Gaussian Noise

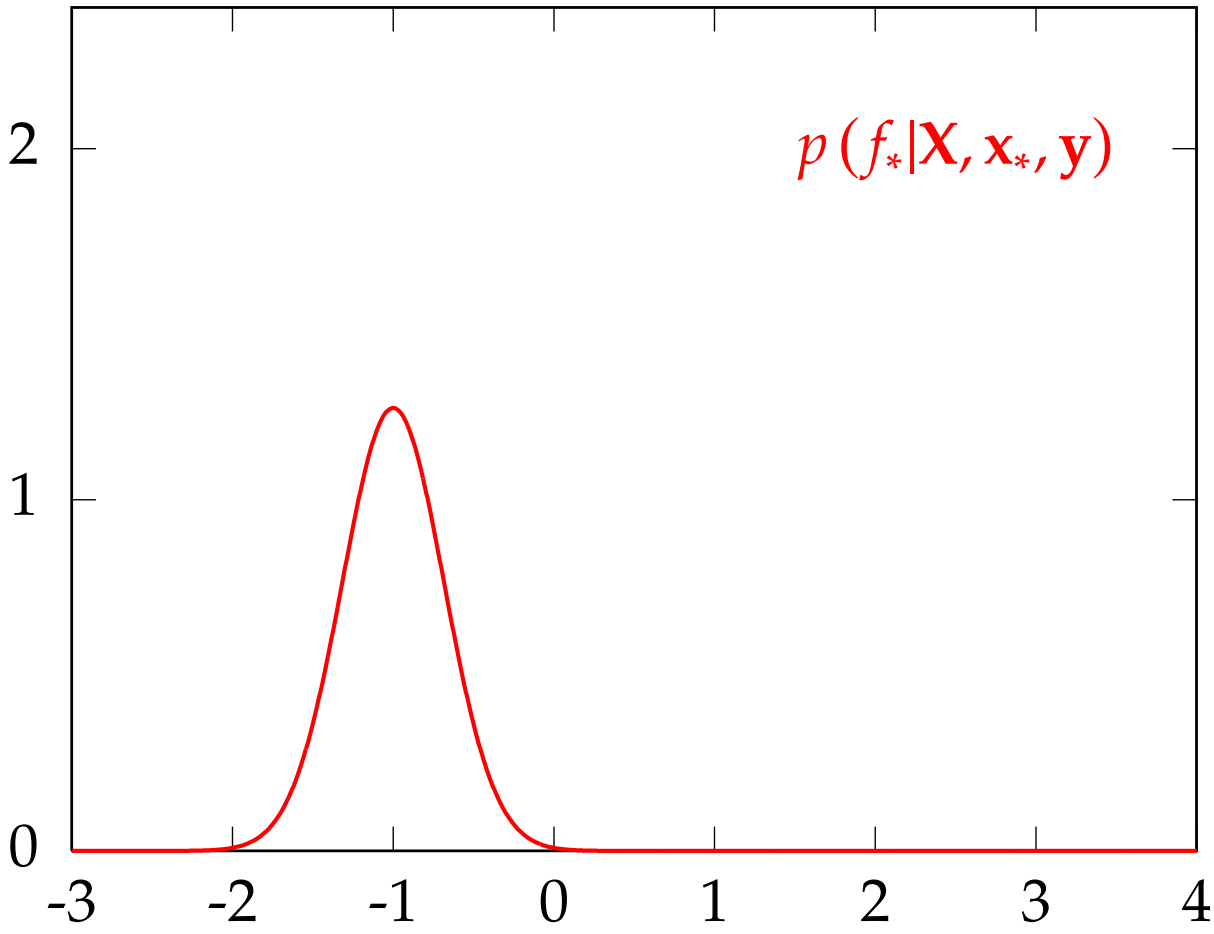


Figure: Inclusion of a data point with Gaussian noise.

Gaussian Noise

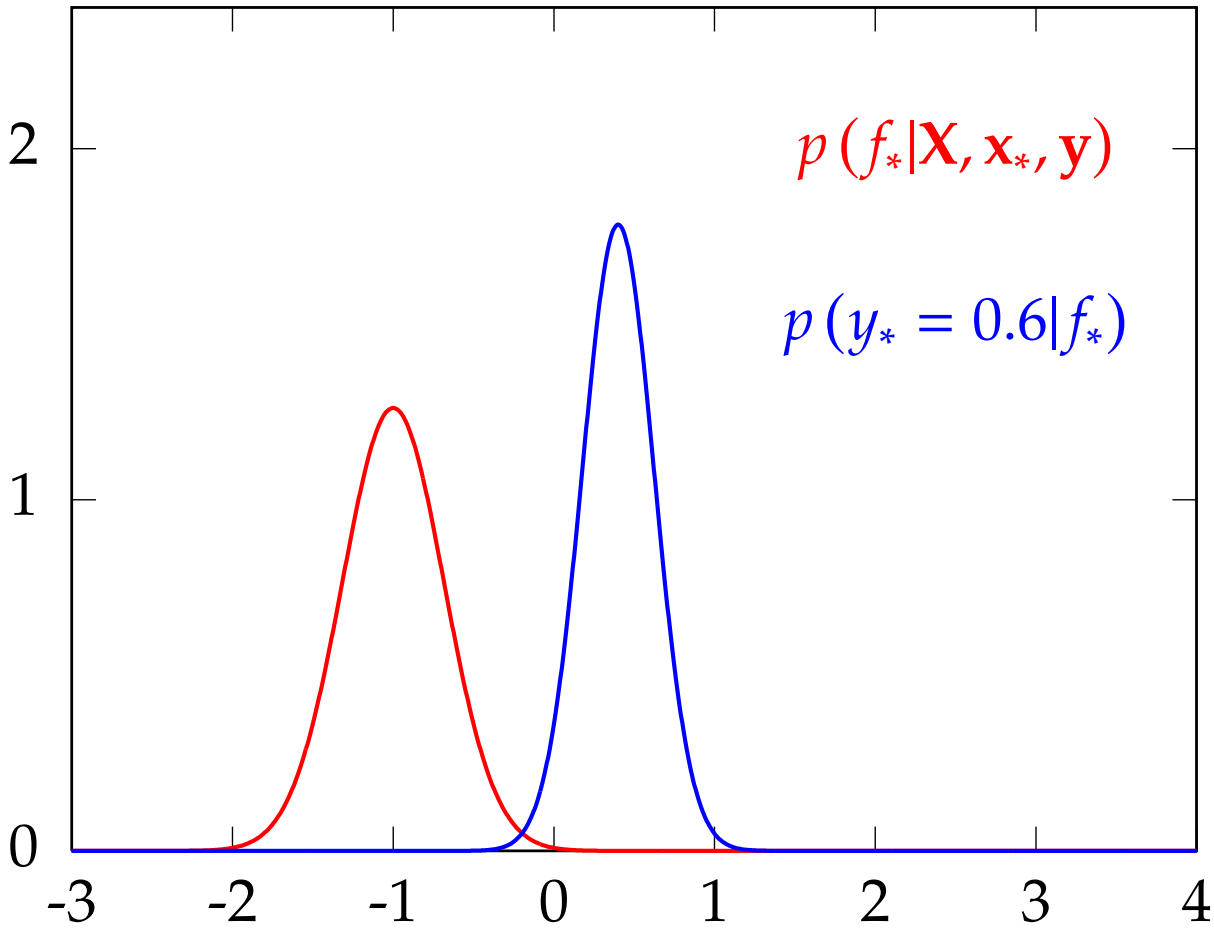


Figure: Inclusion of a data point with Gaussian noise.

Gaussian Noise

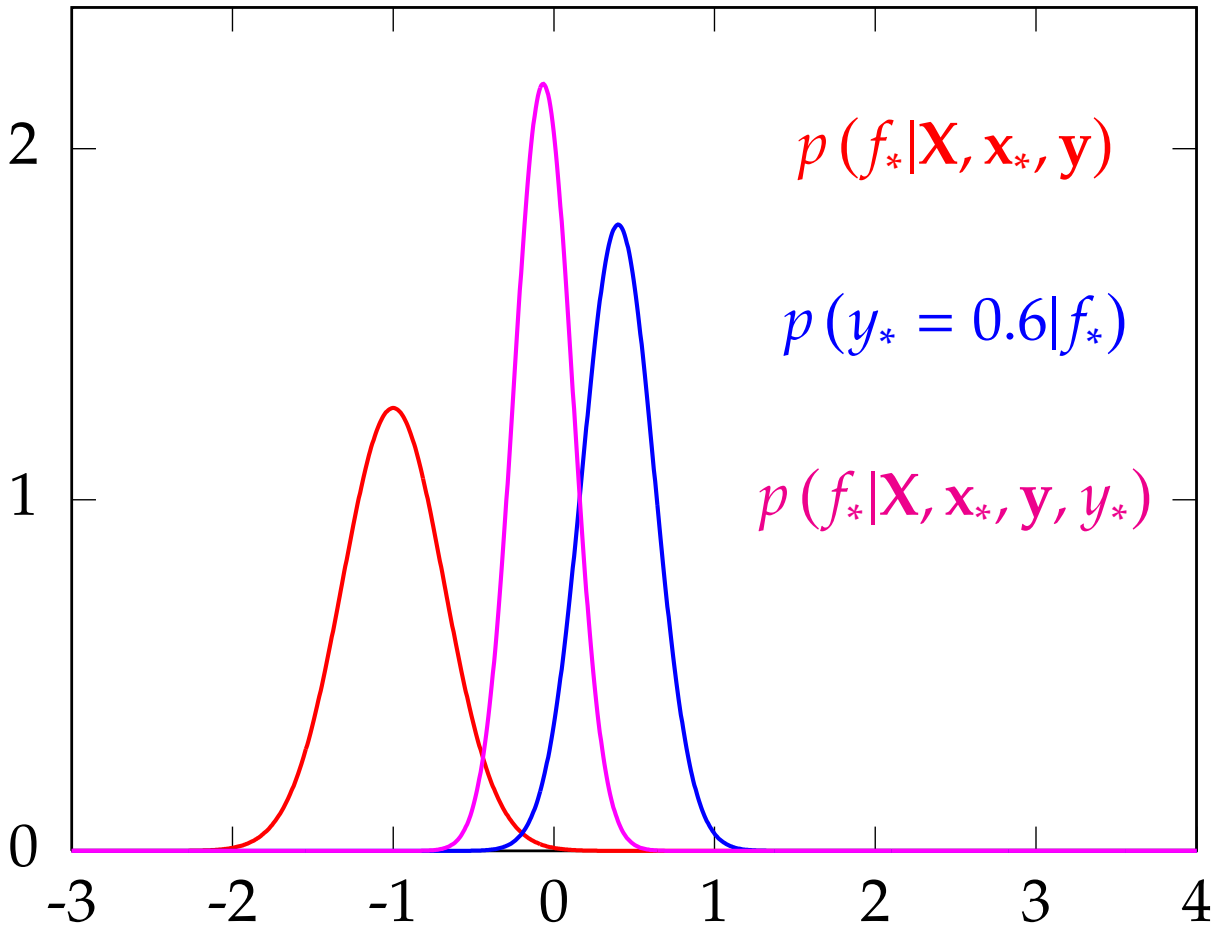


Figure: Inclusion of a data point with Gaussian noise.

Classification Noise Model

Probit Noise Model

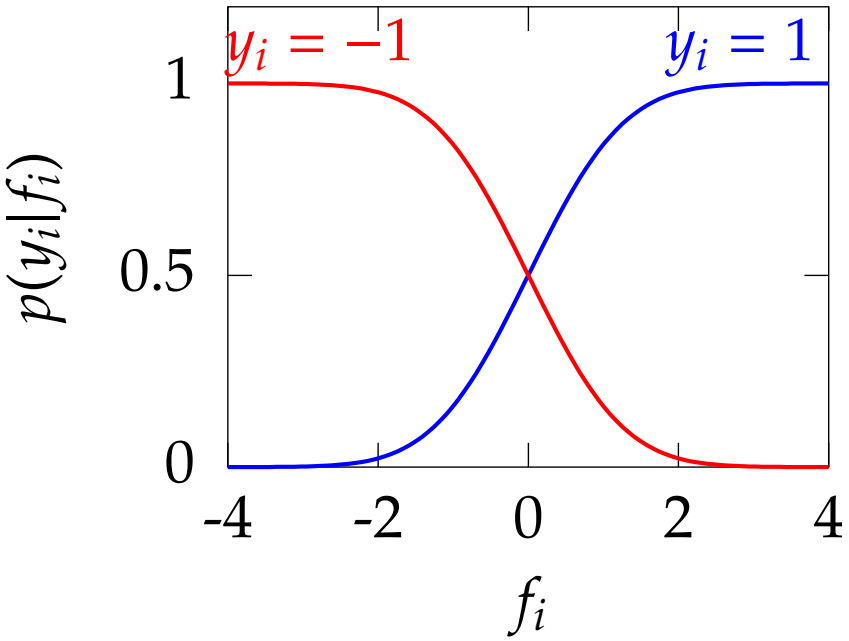


Figure: The probit model (classification). The plot shows $p(y_i|f_i)$ for different values of y_i. For $y_i = 1$ we have

$p(y_i|f_i) = \phi(f_i) = \int_{-\infty}^{f_i} \mathcal{N}(z|0,1)\,\mathrm{d}z.$

Classification

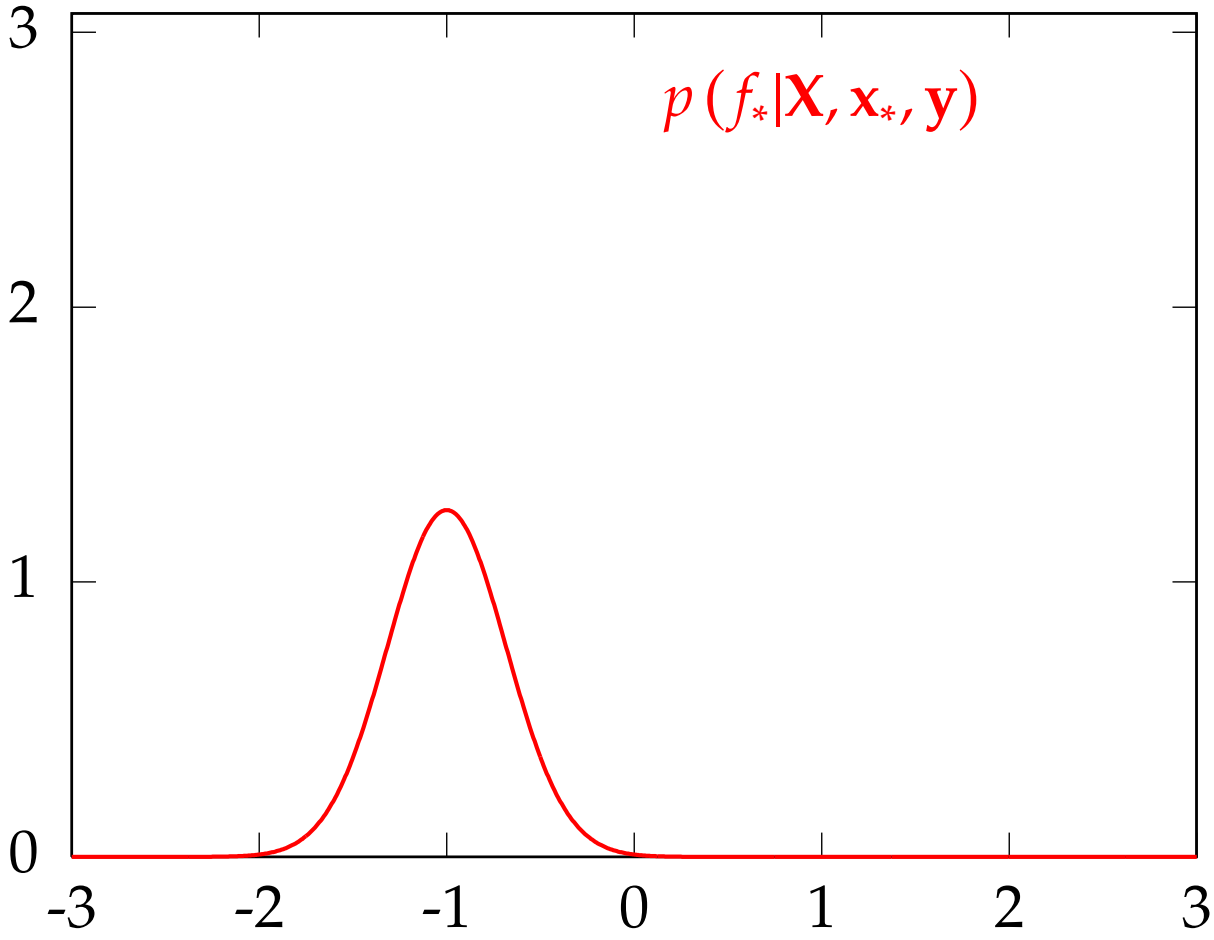


Figure: An EP style update with a classification noise model.

Classification

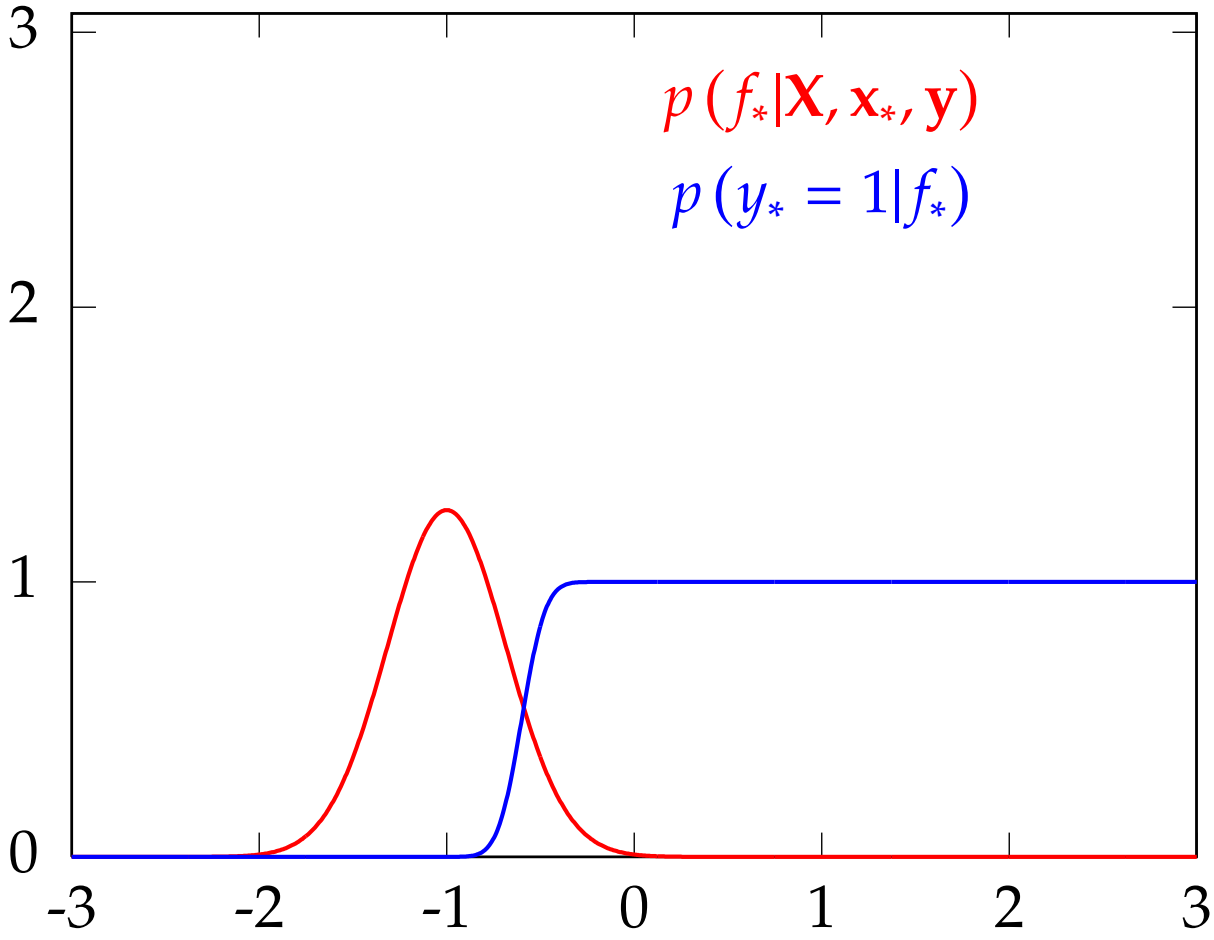


Figure: An EP style update with a classification noise model.

Classification

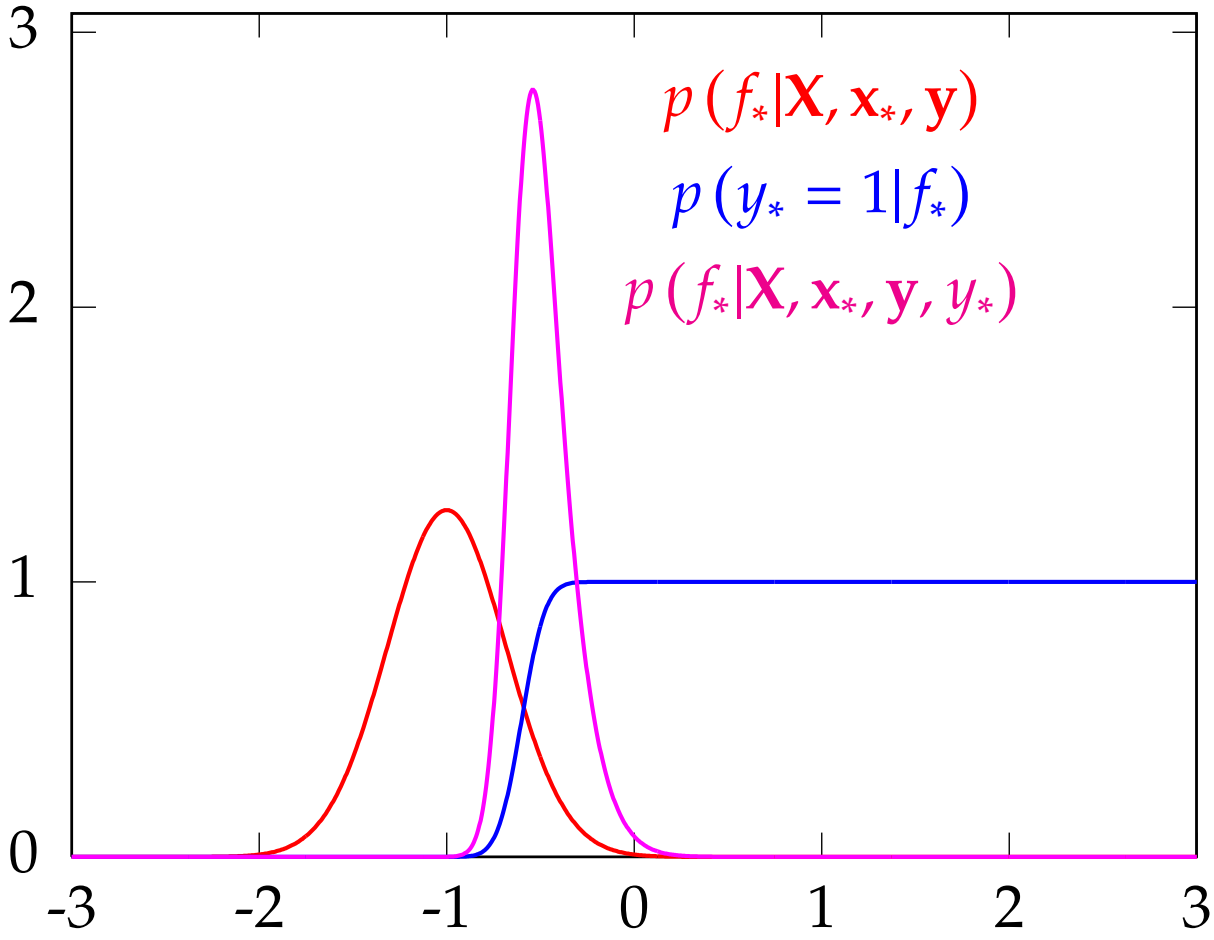


Figure: An EP style update with a classification noise model.

Classification

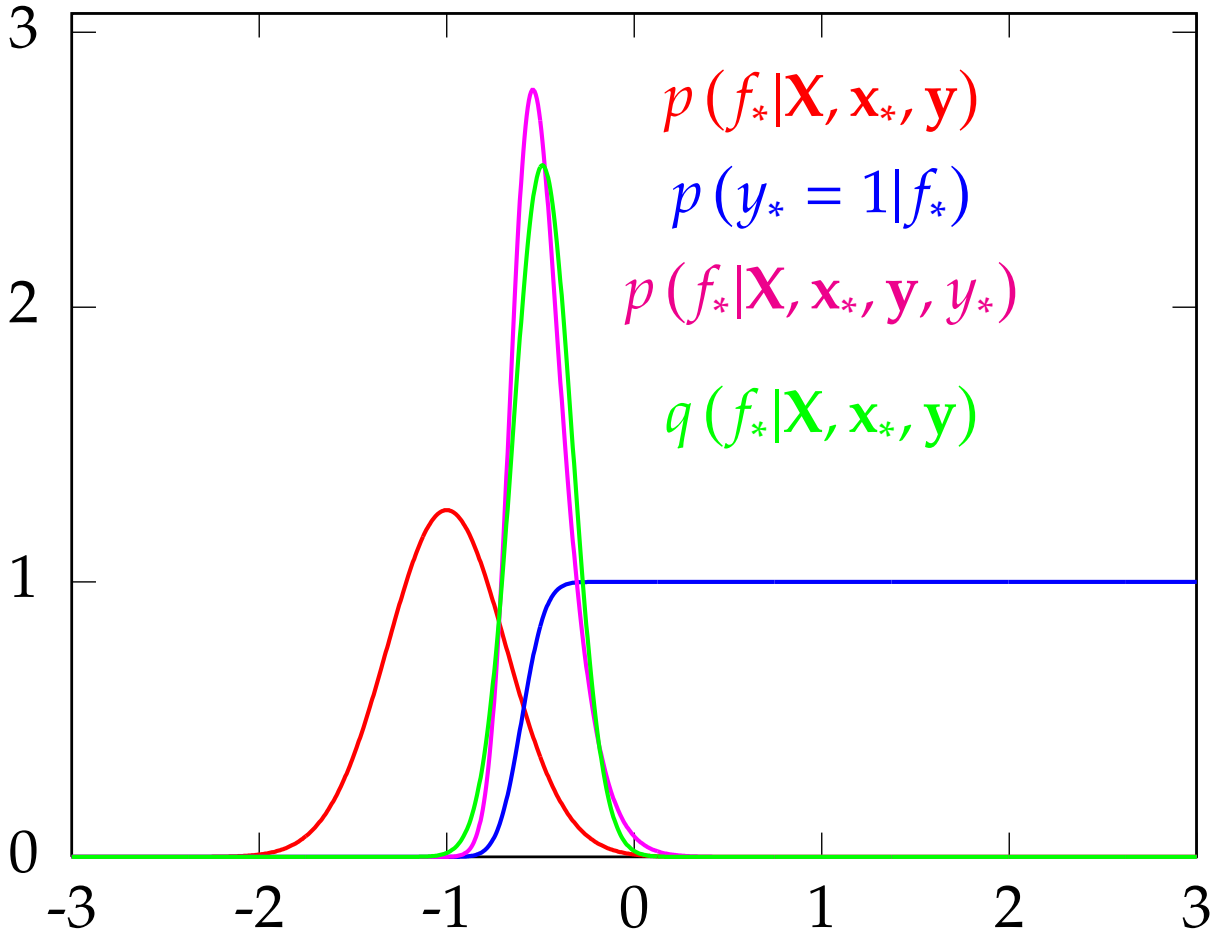


Figure: An EP style update with a classification noise model.

Ordinal Noise Model

Ordered Categories

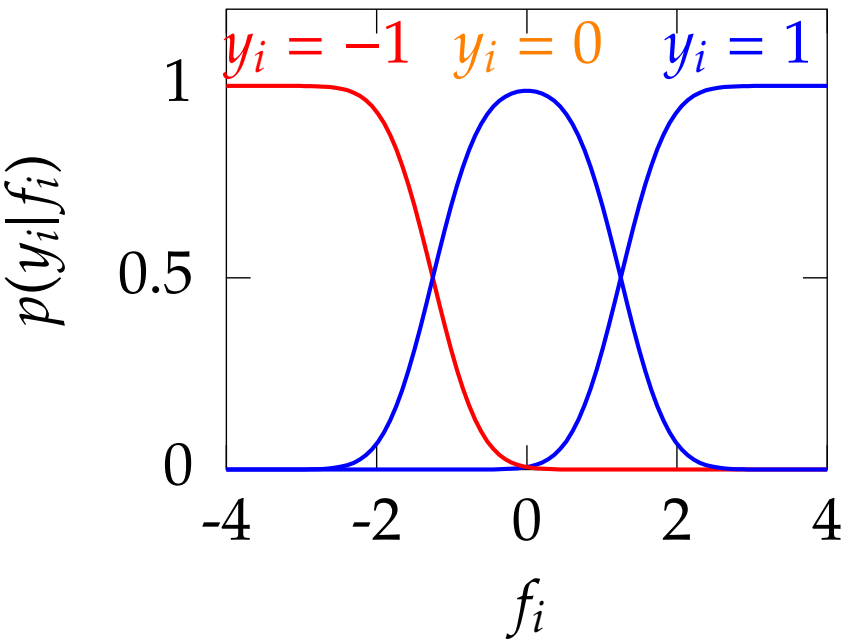


Figure: The ordered categorical noise model (ordinal regression). The plot shows $p(y_i|f_i)$ for different values of y_i. Here we have assumed three categories.

Ordinal Regression

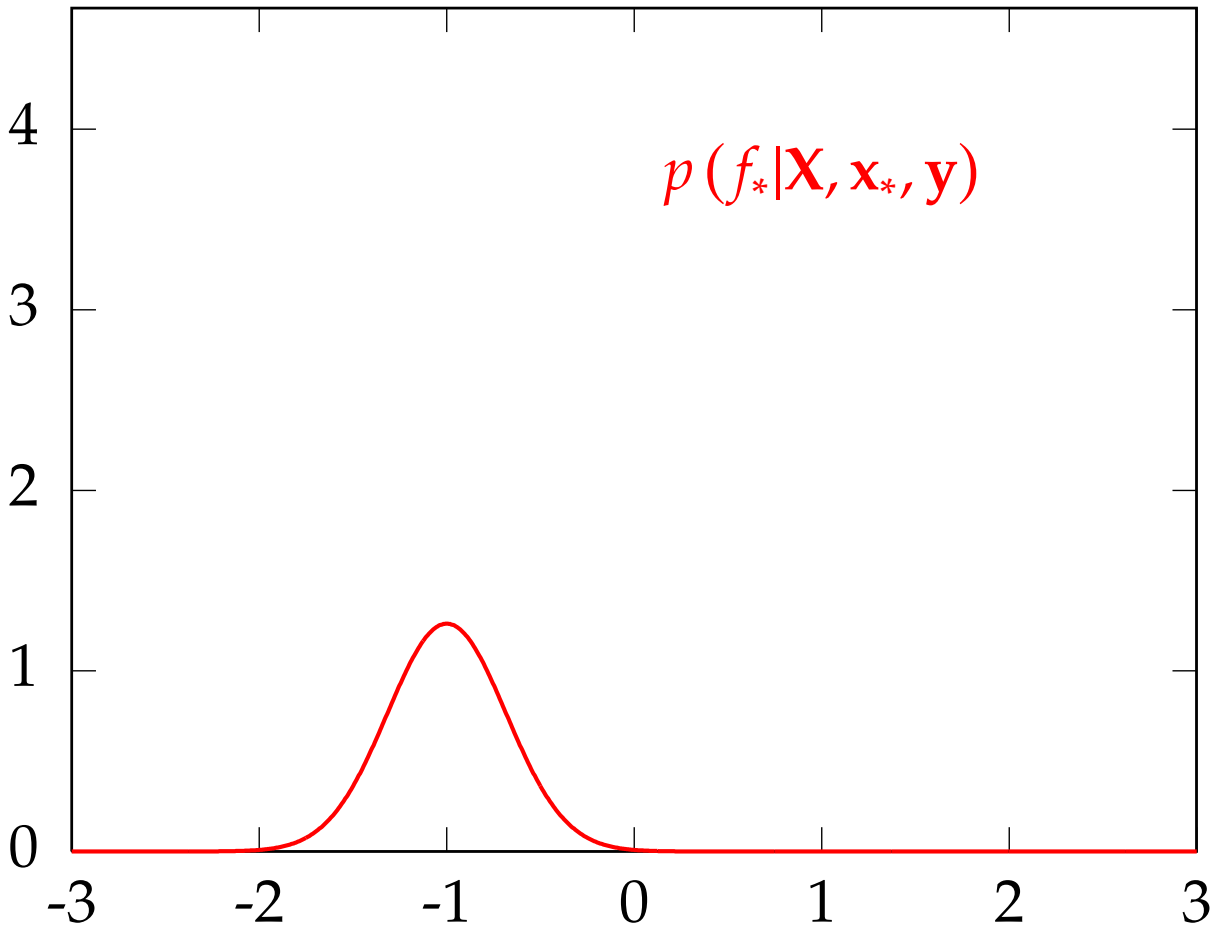


Figure: An EP style update with an ordered category noise model.

Return

Ordinal Regression

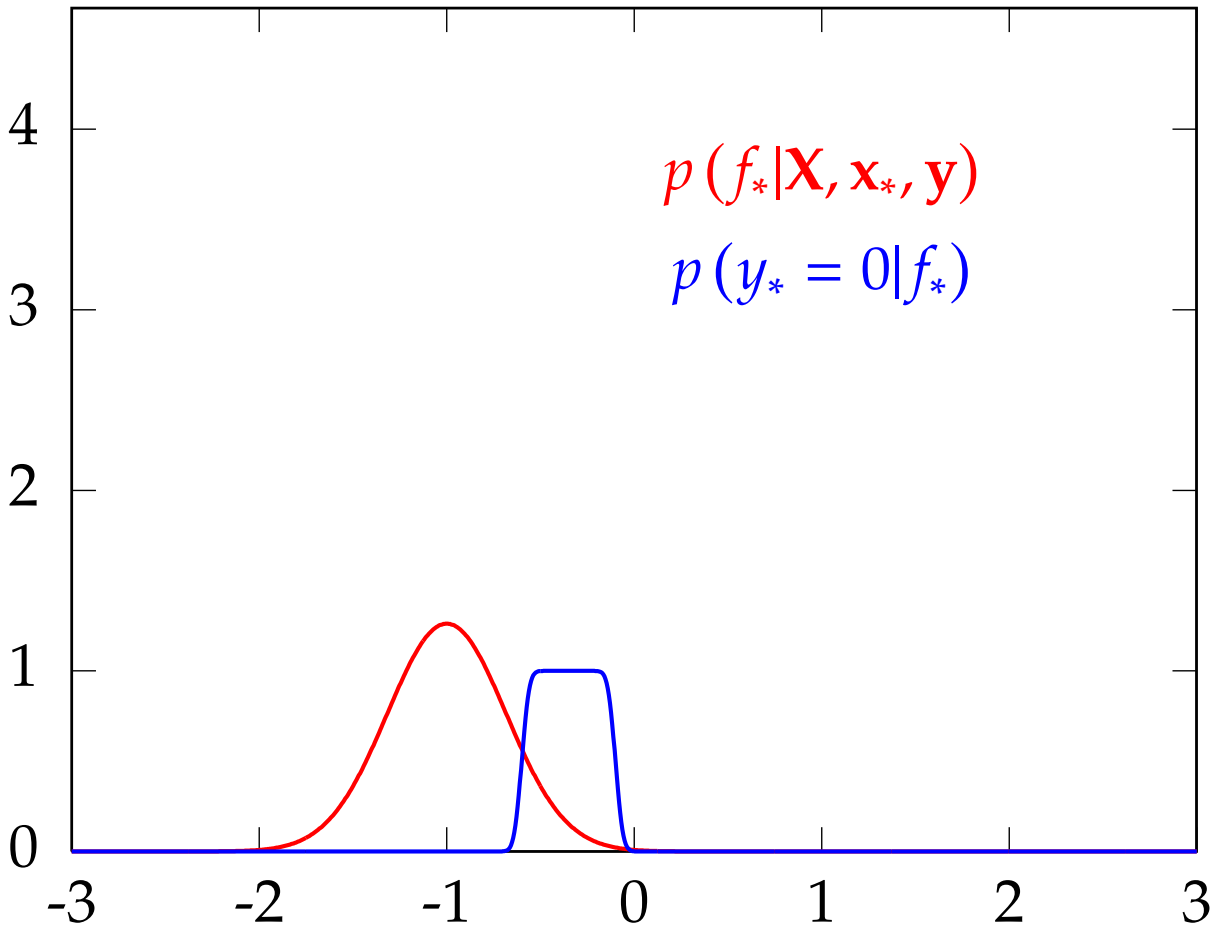


Figure: An EP style update with an ordered category noise model.

Return

Ordinal Regression

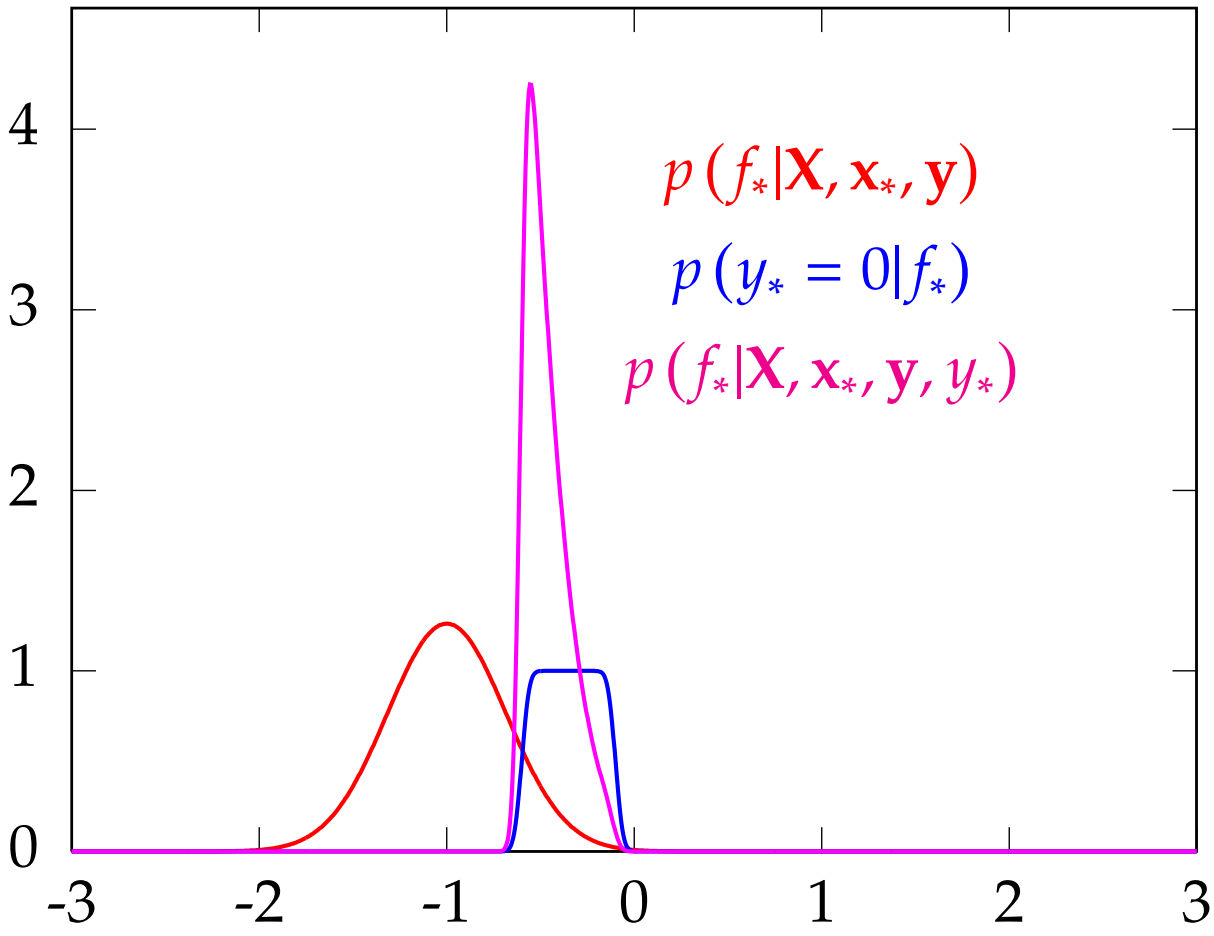


Figure: An EP style update with an ordered category noise model.

Return

Ordinal Regression

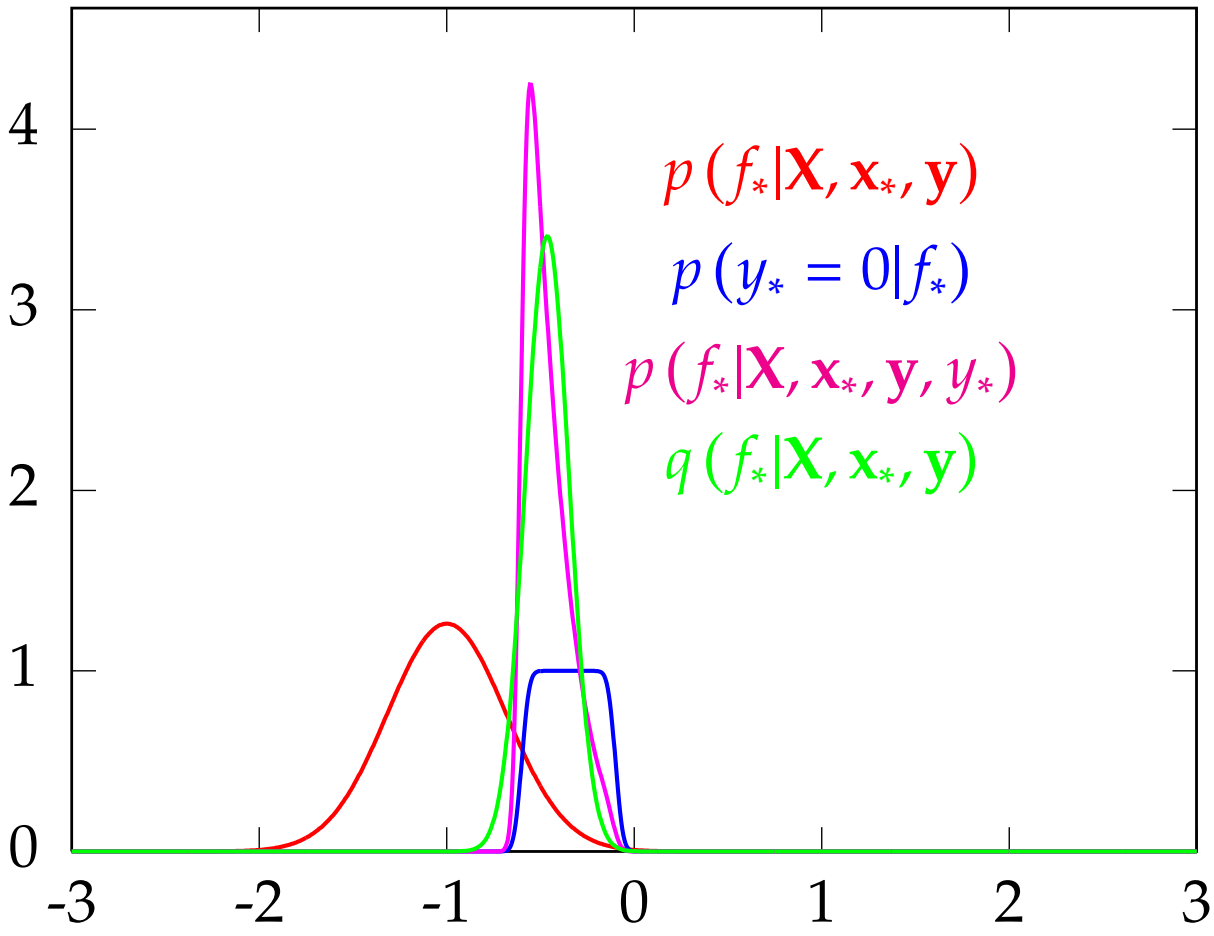


Figure: An EP style update with an ordered category noise model.

Return

Survival Models

Cox Gaussian Process Regression

$$h(t|\mathbf{x}) = \overbrace{h_0(t)}^{\exp(GP(t))} \exp\Big(\overbrace{\beta\mathbf{x}}^{GP(\mathbf{x}(t),t)} \Big)$$

- Apply these extremely flexible methods to Survival Analysis
- Alter assumptions of Cox Proportional Hazards Model to discover how significant they are, test whether we can increase our predictive power by:
  1. Breaking proportionality assumption
  2. Allowing for interactions between variables

Learning Covariance Parameters

Can we determine covariance parameters from the data?

$$\mathcal{N}\left(\mathbf{y}|\mathbf{0},\mathbf{K}\right) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{K}|}\exp\left(-\frac{\mathbf{y}^\top\mathbf{K}^{-1}\mathbf{y}}{2}\right)$$

The parameters are *inside* the covariance function (matrix).

$$k_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$$

Learning Covariance Parameters

Can we determine covariance parameters from the data?

$$\mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{K}|}\exp\left(-\frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}\right)$$

The parameters are *inside* the covariance function (matrix).

$$k_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$$

Learning Covariance Parameters

Can we determine covariance parameters from the data?

$$\log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}) = -\frac{1}{2}\log|\mathbf{K}| - \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2} - \frac{n}{2}\log 2\pi$$

The parameters are *inside* the covariance function (matrix).

$$k_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$$

Learning Covariance Parameters

Can we determine covariance parameters from the data?

$$E(\boldsymbol{\theta}) = \frac{1}{2}\log|\mathbf{K}| + \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

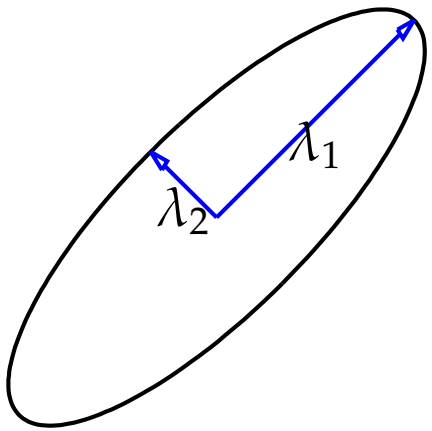
The parameters are *inside* the covariance function (matrix).

$$k_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$$

Eigendecomposition of Covariance

A useful decomposition for understanding the objective function.

$$\mathbf{K} = \mathbf{R}\mathbf{\Lambda}^2\mathbf{R}^\top$$

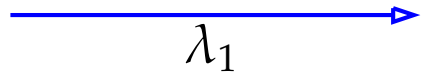


Diagonal of $\mathbf{\Lambda}$ represents distance along axes.
$\mathbf{R}$ gives a rotation of these axes.

where $\mathbf{\Lambda}$ is a *diagonal* matrix and $\mathbf{R}^\top\mathbf{R} = \mathbf{I}$.

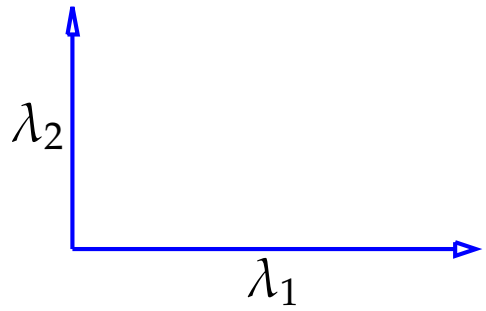
Capacity control: $\log|\mathbf{K}|$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$



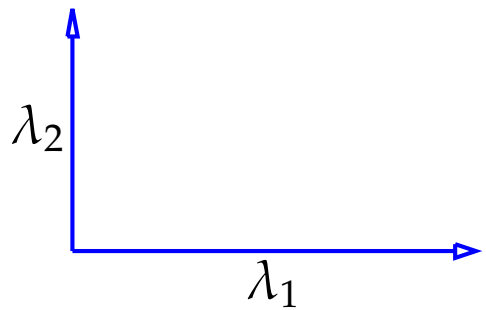
Capacity control: $\log|\mathbf{K}|$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$



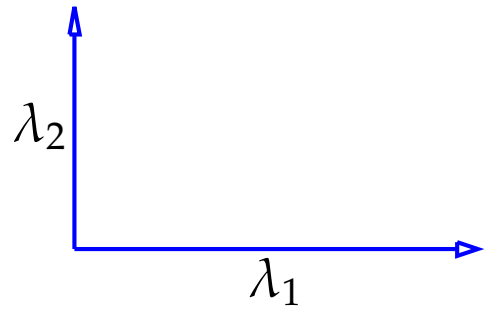
Capacity control: $\log|\mathbf{K}|$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$



Capacity control: $\log|\mathbf{K}|$

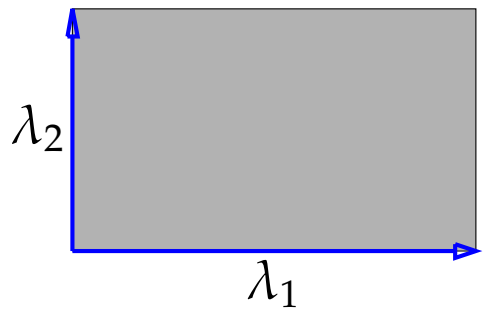
$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$



$$|\mathbf{\Lambda}| = \lambda_1 \lambda_2$$

Capacity control: $\log|\mathbf{K}|$

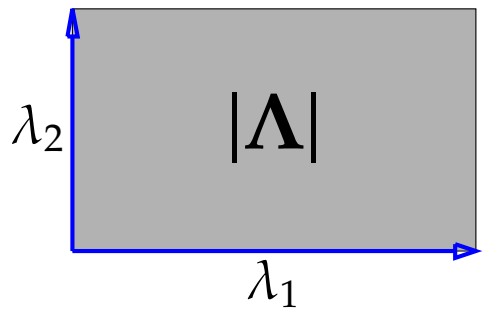
$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$



$$|\mathbf{\Lambda}| = \lambda_1 \lambda_2$$

Capacity control: $\log|\mathbf{K}|$

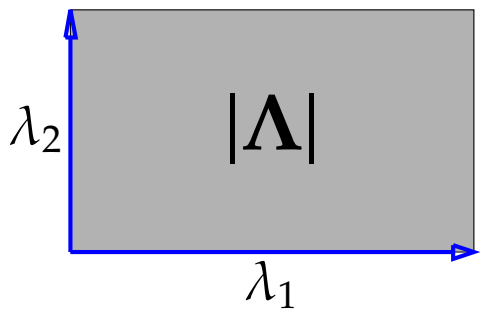
$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$



$$|\mathbf{\Lambda}| = \lambda_1 \lambda_2$$

Capacity control: $\log|\mathbf{K}|$

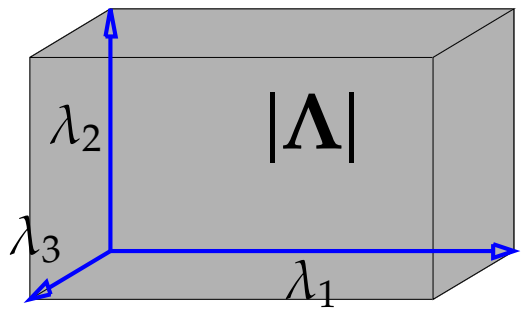
$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$



$$|\mathbf{\Lambda}| = \lambda_1 \lambda_2$$

Capacity control: $\log |\mathbf{K}|$

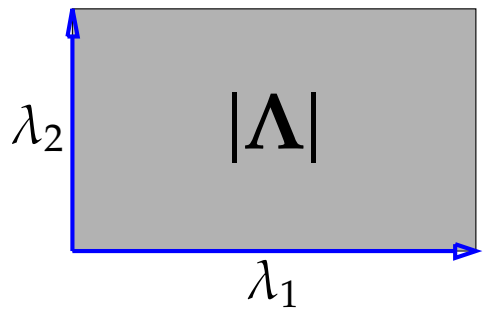
$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$



$$|\mathbf{\Lambda}| = \lambda_1 \lambda_2 \lambda_3$$

Capacity control: $\log|\mathbf{K}|$

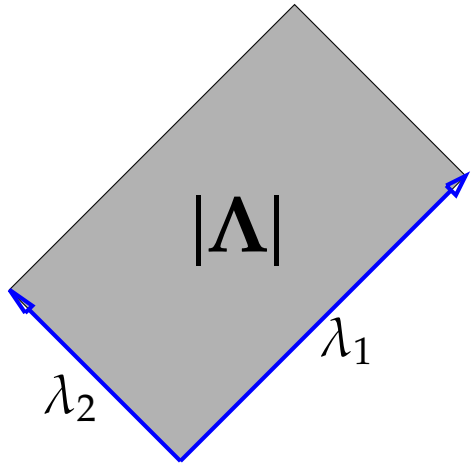
$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$



$$|\mathbf{\Lambda}| = \lambda_1 \lambda_2$$

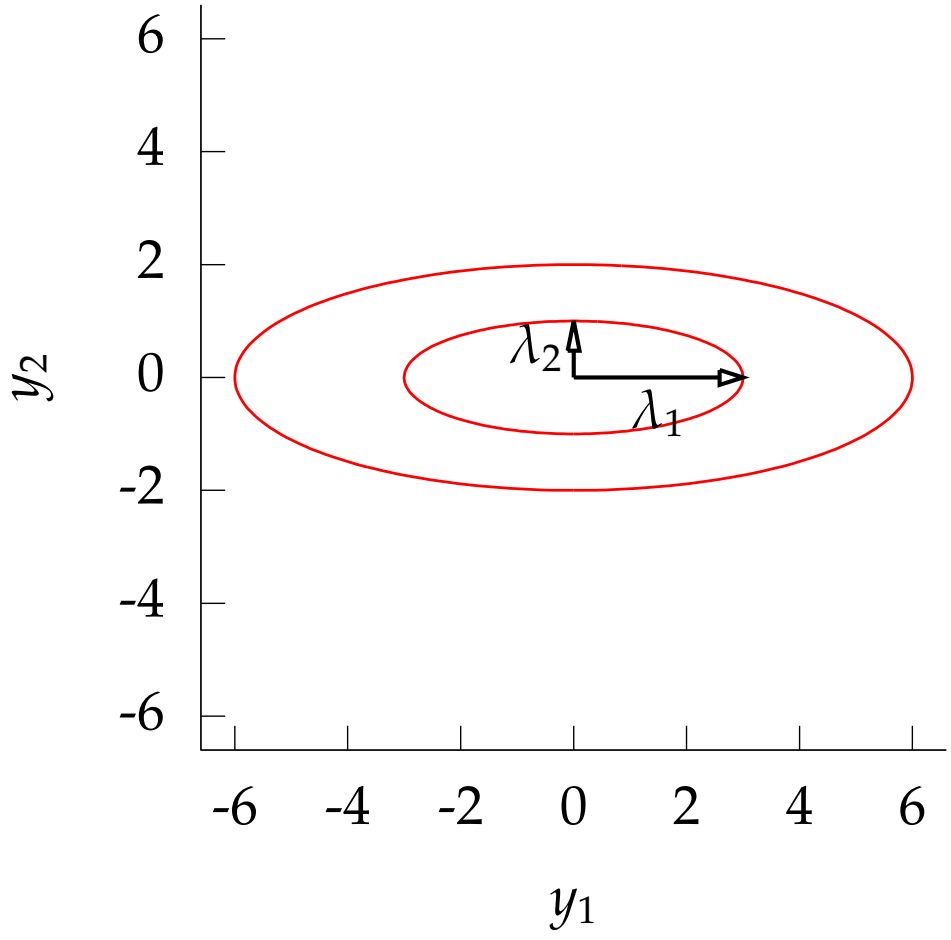
Capacity control: $\log |\mathbf{K}|$

$$\mathbf{R\Lambda} = \begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \end{bmatrix}$$

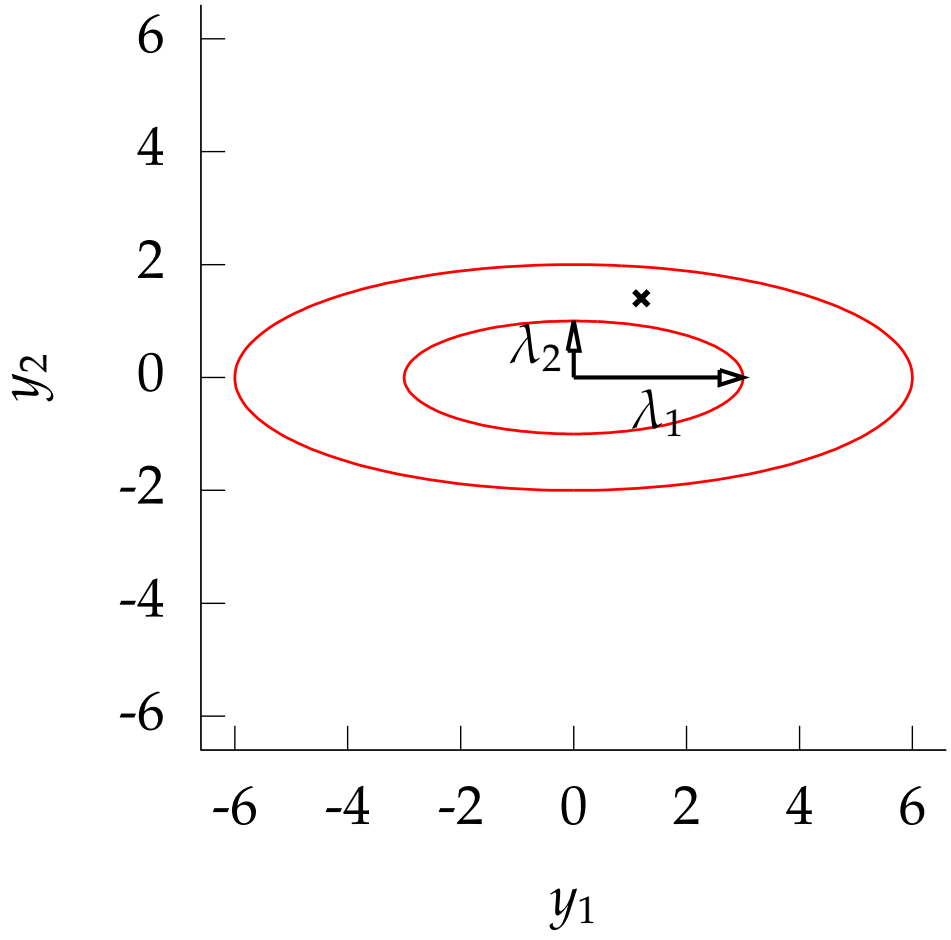


$$|\mathbf{R\Lambda}| = \lambda_1 \lambda_2$$

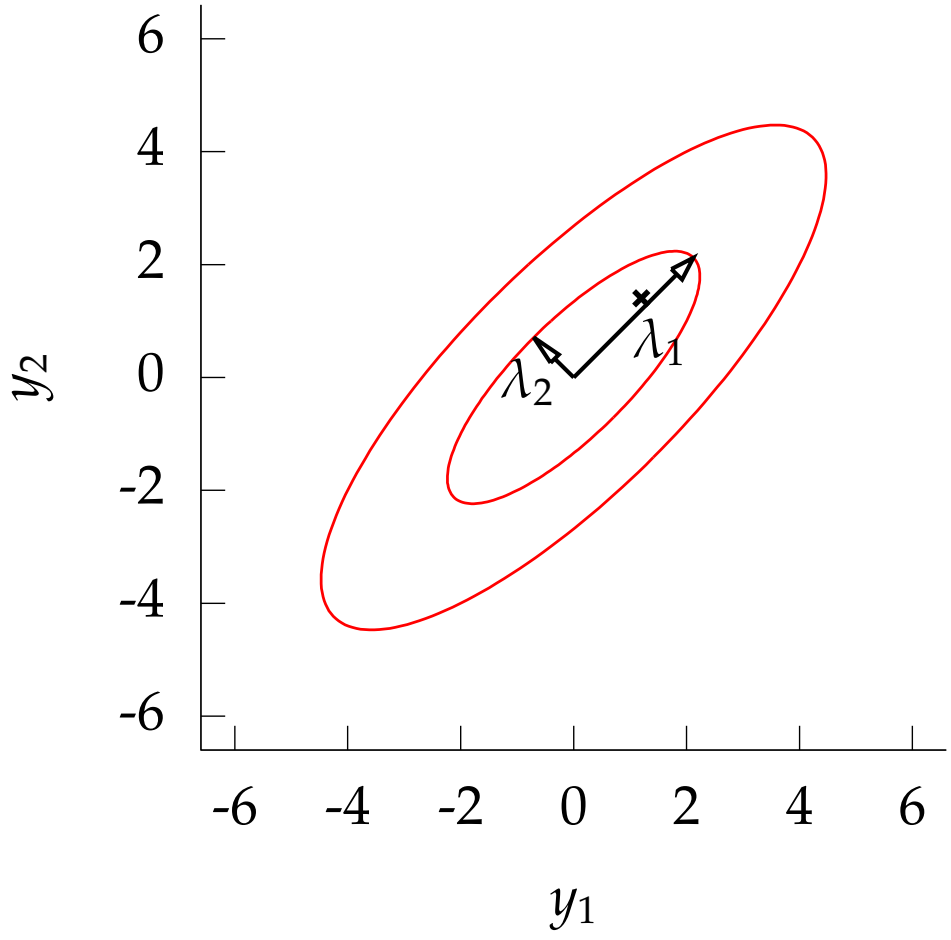
Data Fit: $\frac{\mathbf{y}^{-1}\mathbf{K}^{-1}\mathbf{y}}{2}$



Data Fit: $\frac{\mathbf{y}^{-1}\mathbf{K}^{-1}\mathbf{y}}{2}$

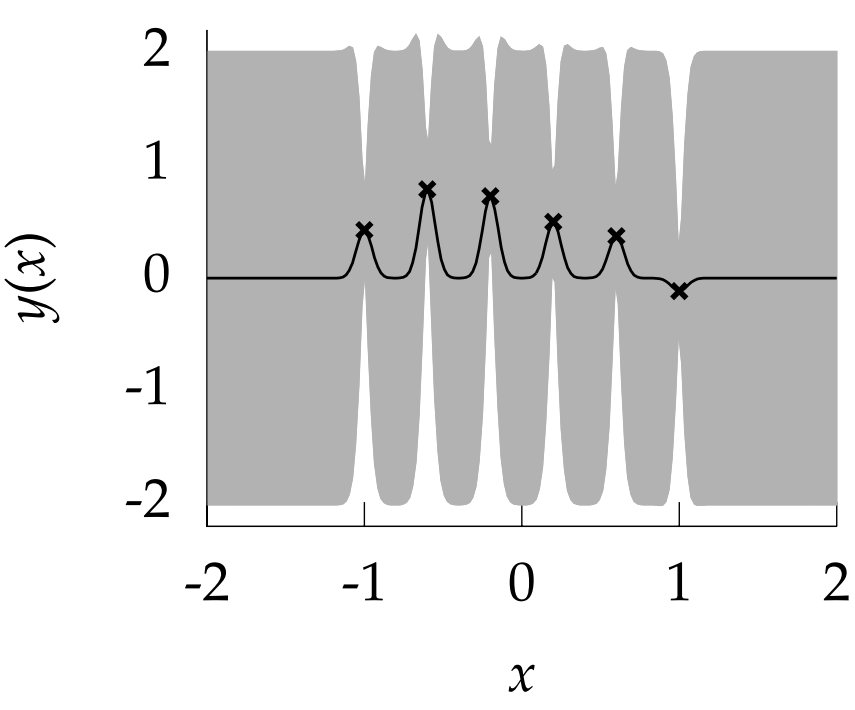


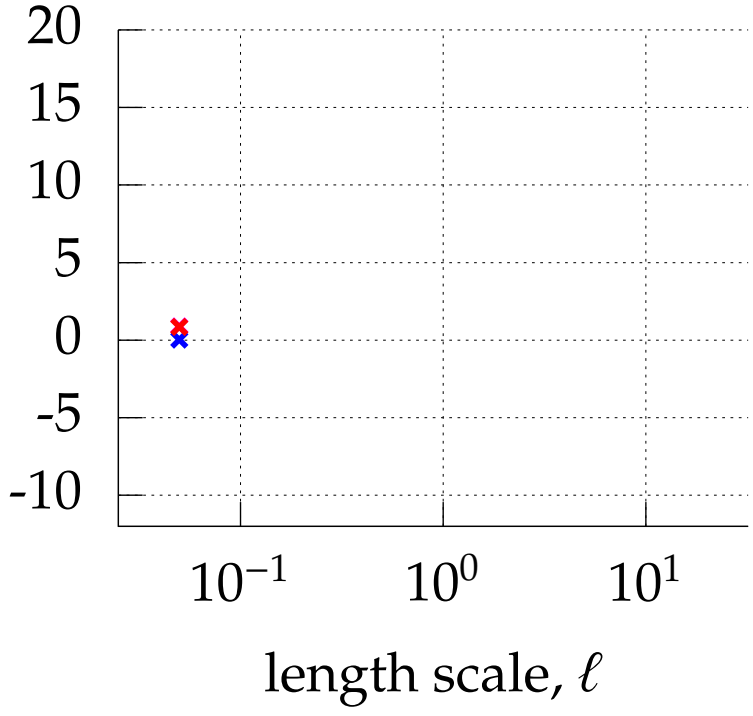
Data Fit: $\frac{\mathbf{y}^{-1}\mathbf{K}^{-1}\mathbf{y}}{2}$



Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

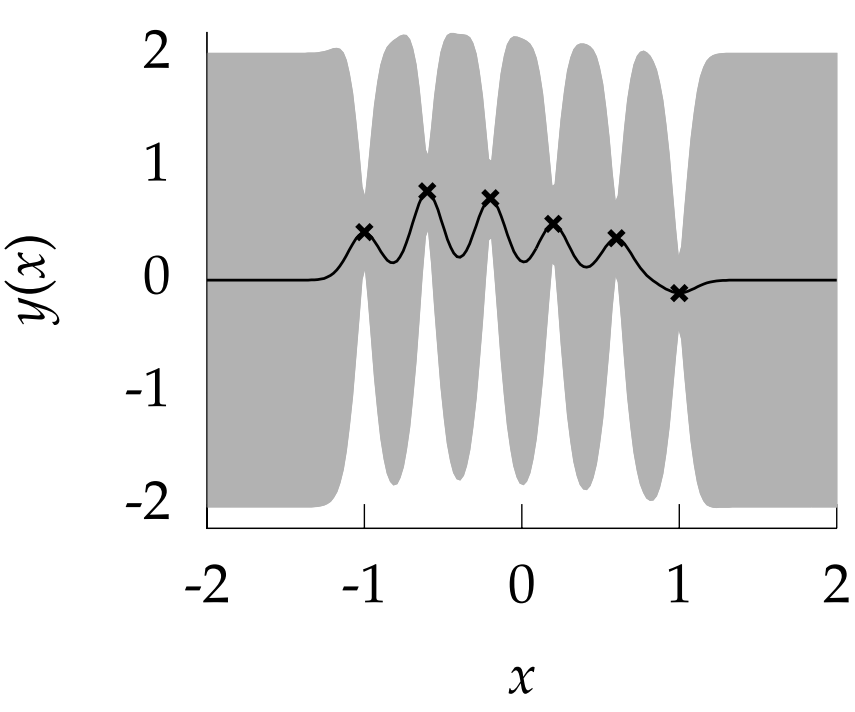


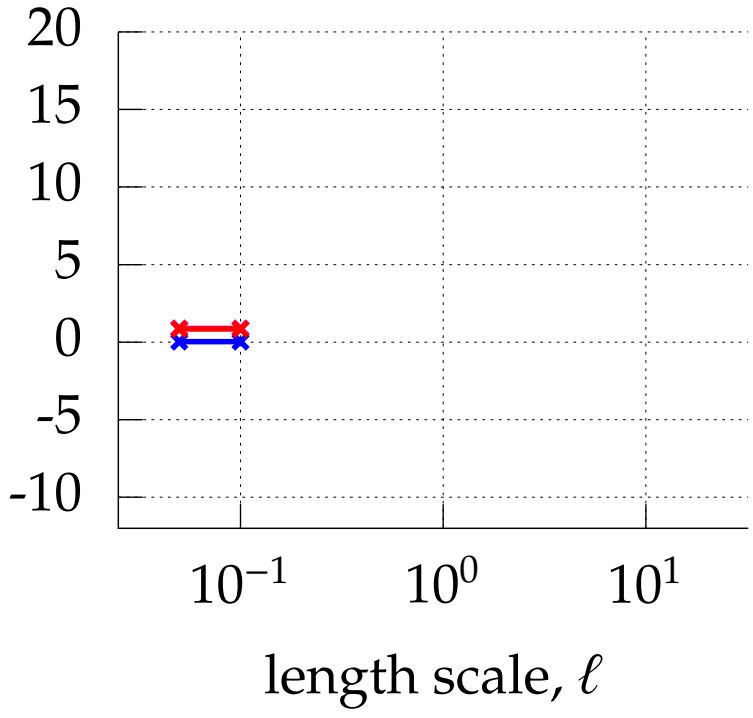


$$E(\boldsymbol{\theta}) = \frac{1}{2}\log|\mathbf{K}| + \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

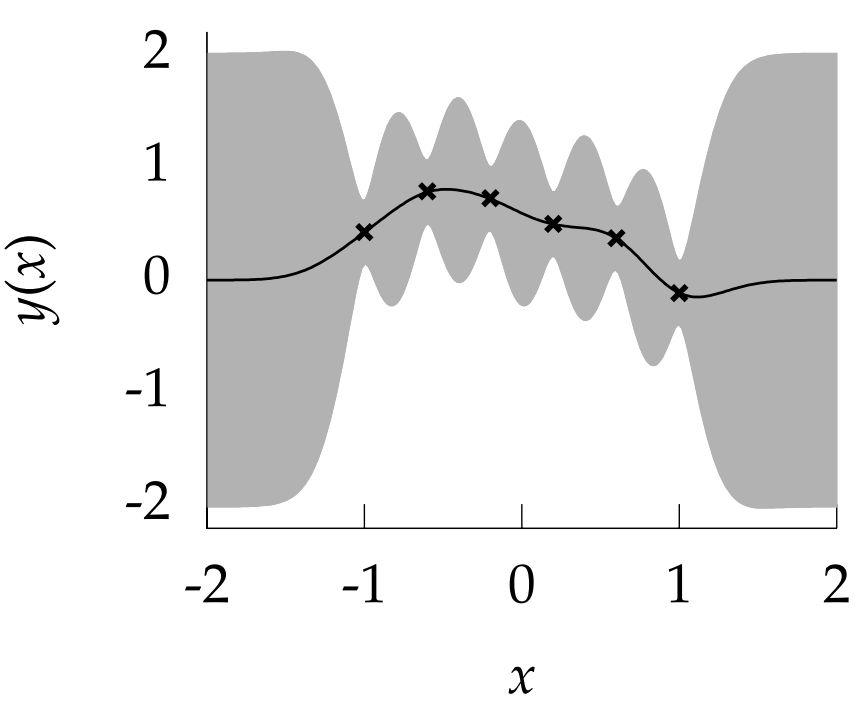


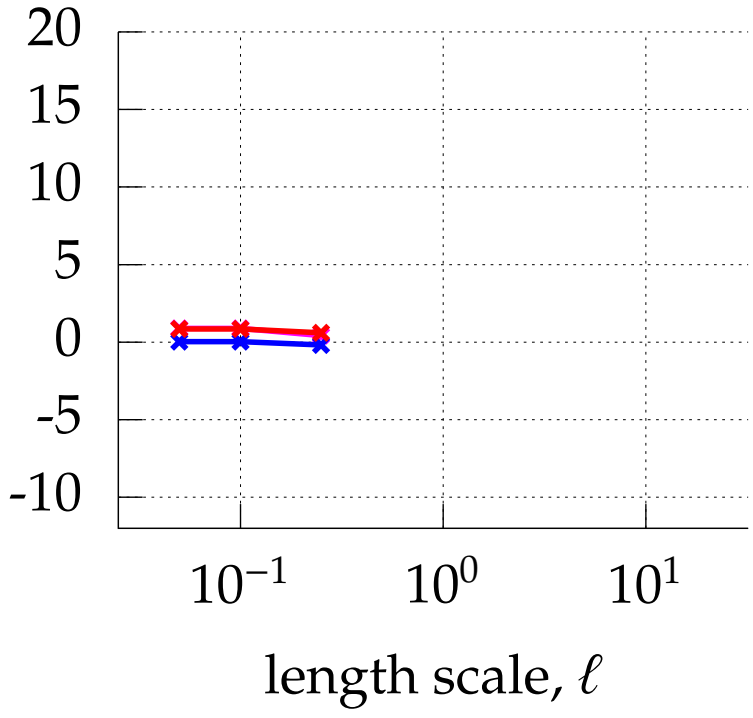


$$E(\boldsymbol{\theta}) = \frac{1}{2}\log|\mathbf{K}| + \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

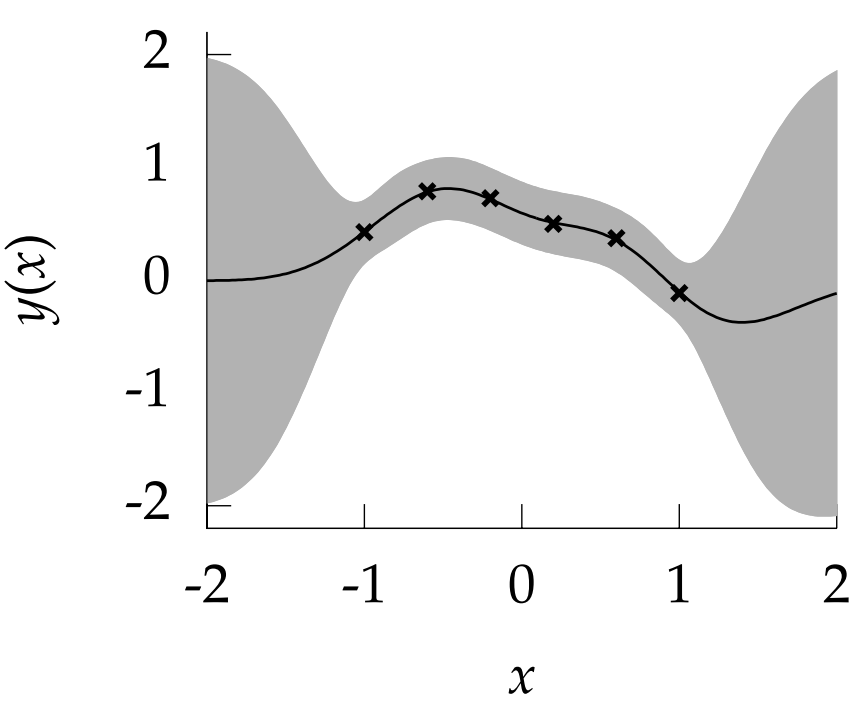


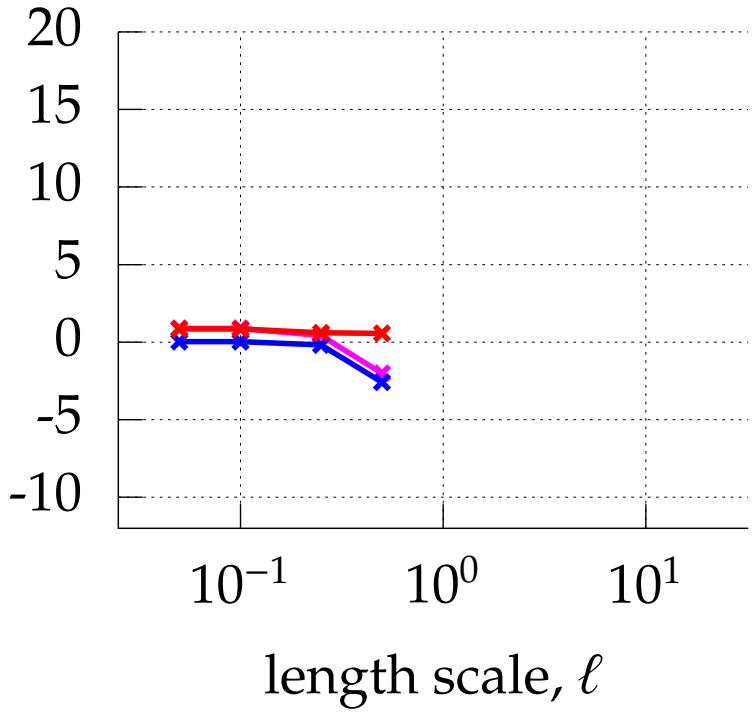


$$E(\boldsymbol{\theta}) = \frac{1}{2}\log|\mathbf{K}| + \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

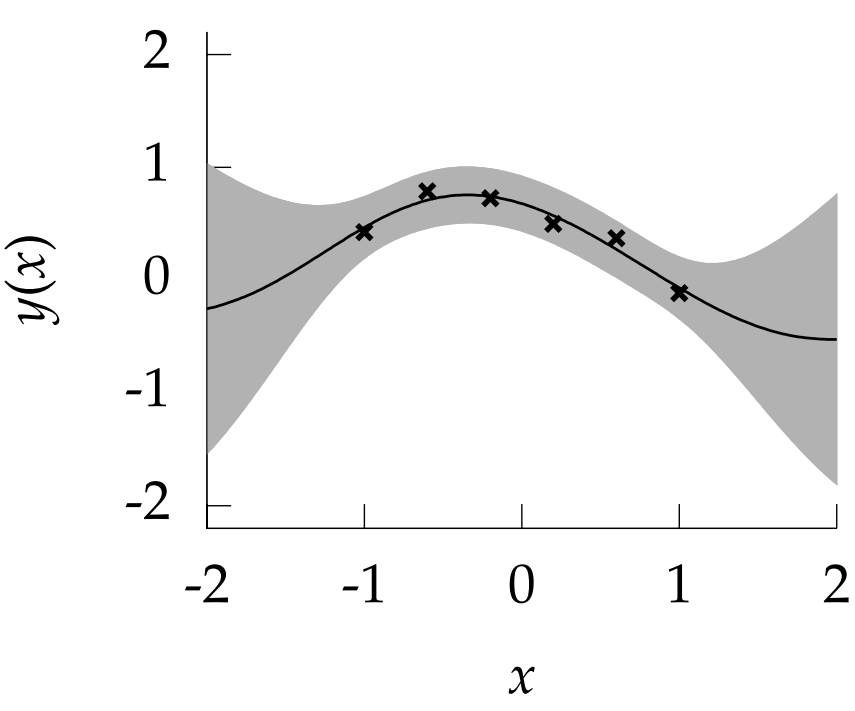


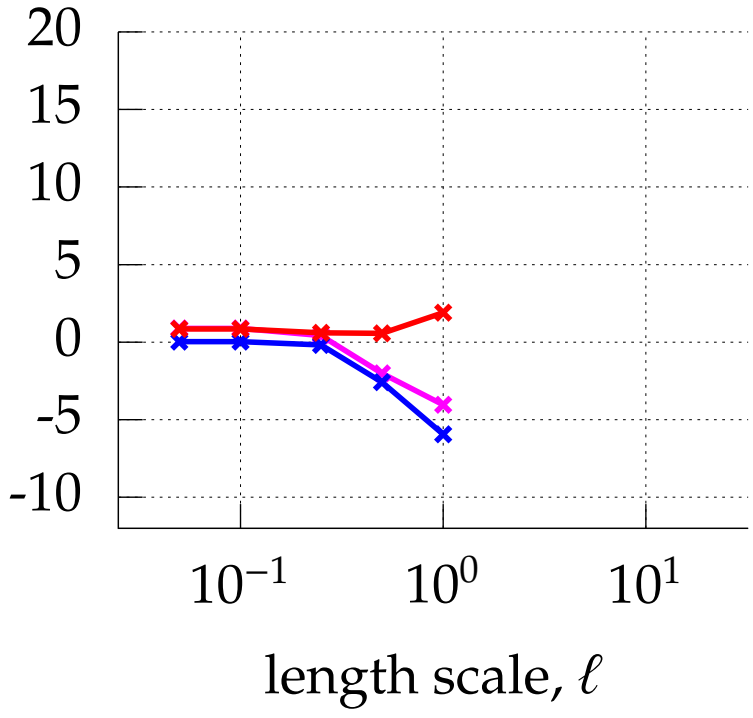


$$E(\boldsymbol{\theta}) = \frac{1}{2}\log|\mathbf{K}| + \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

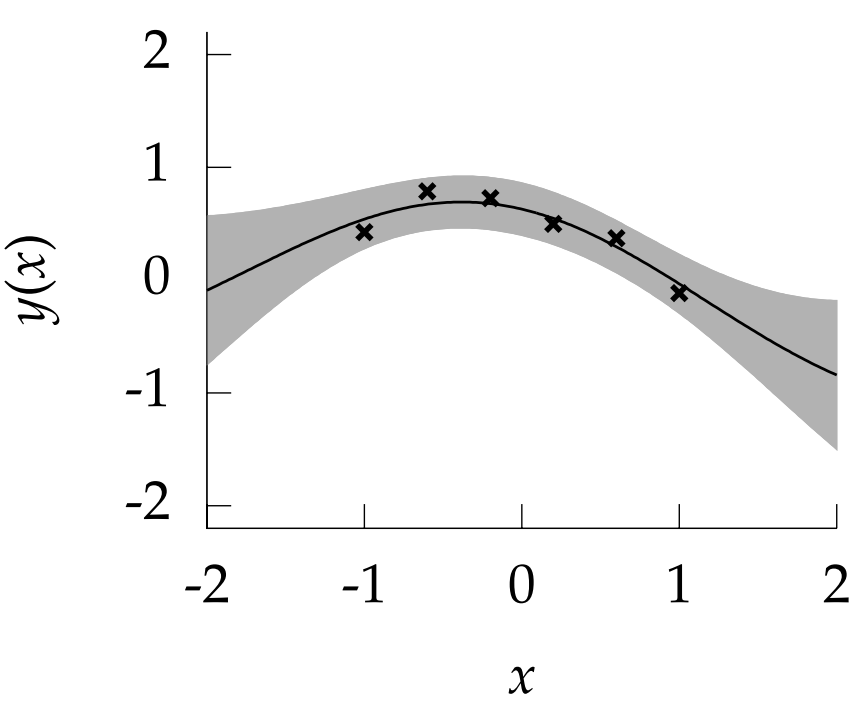


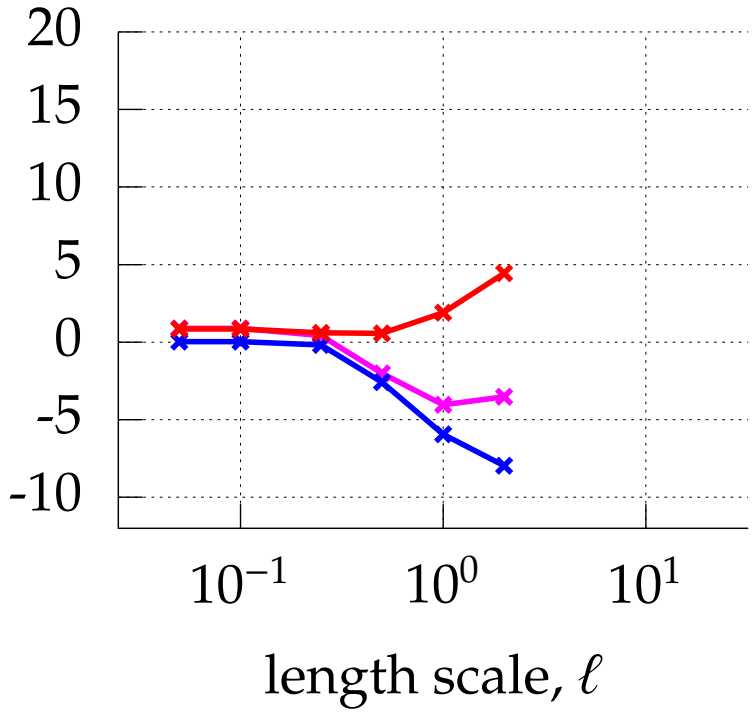


$$E(\boldsymbol{\theta}) = \frac{1}{2}\log|\mathbf{K}| + \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

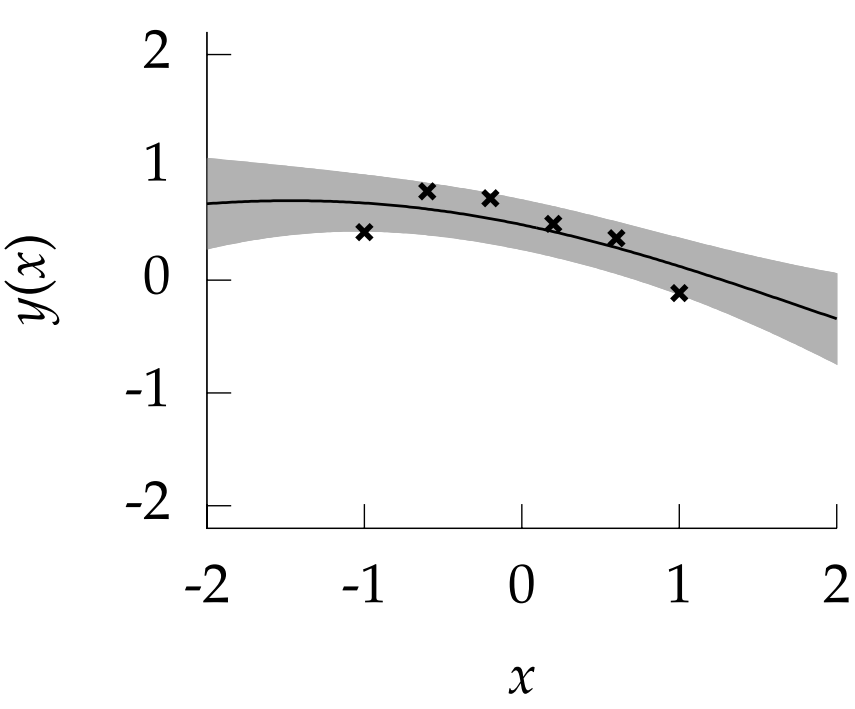


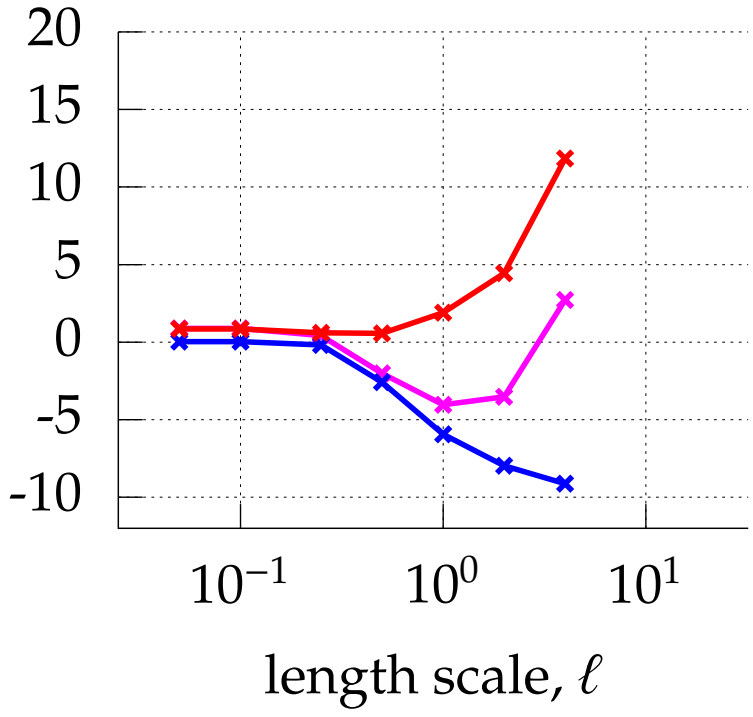


$$E(\boldsymbol{\theta}) = \frac{1}{2}\log|\mathbf{K}| + \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

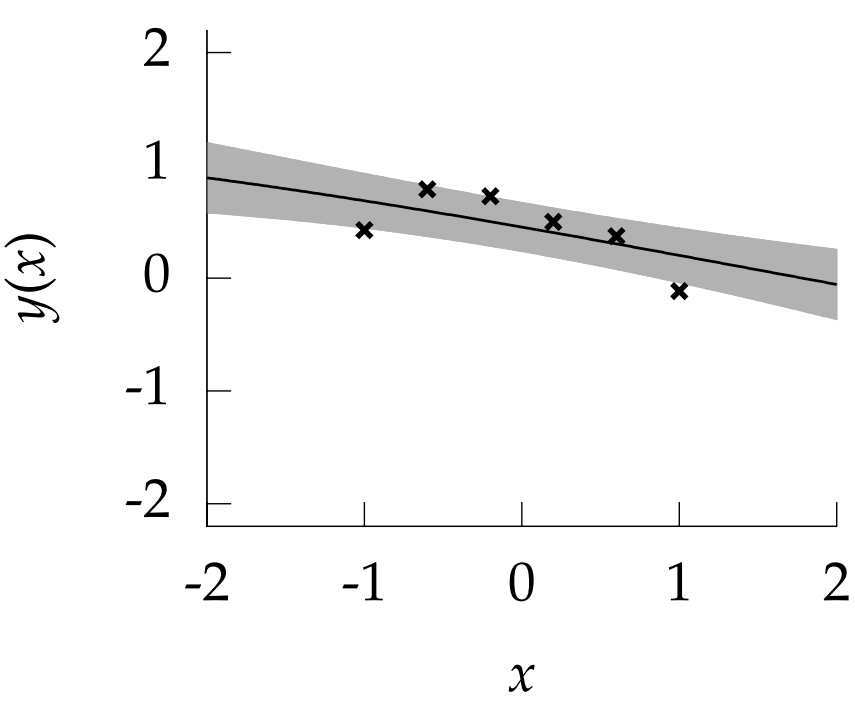


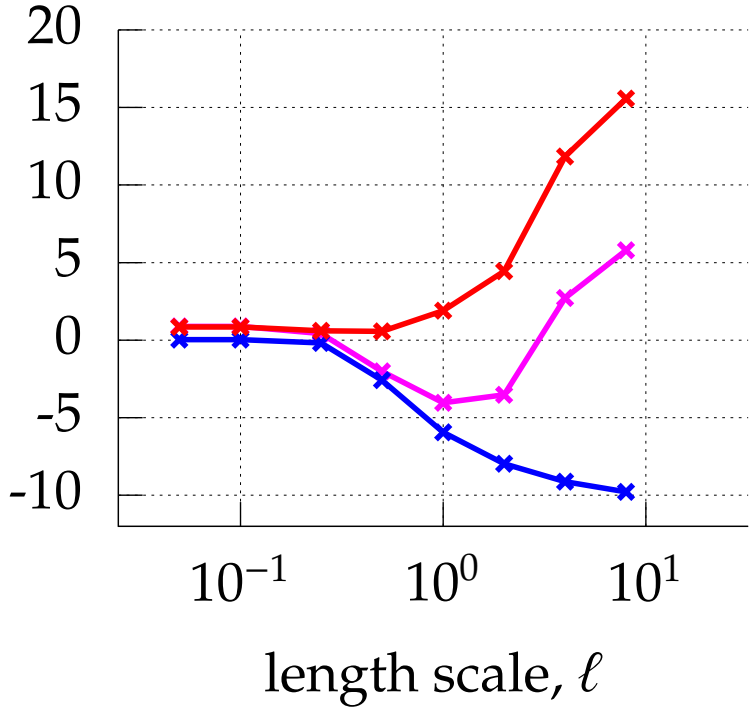


$$E(\boldsymbol{\theta}) = \frac{1}{2}\log|\mathbf{K}| + \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

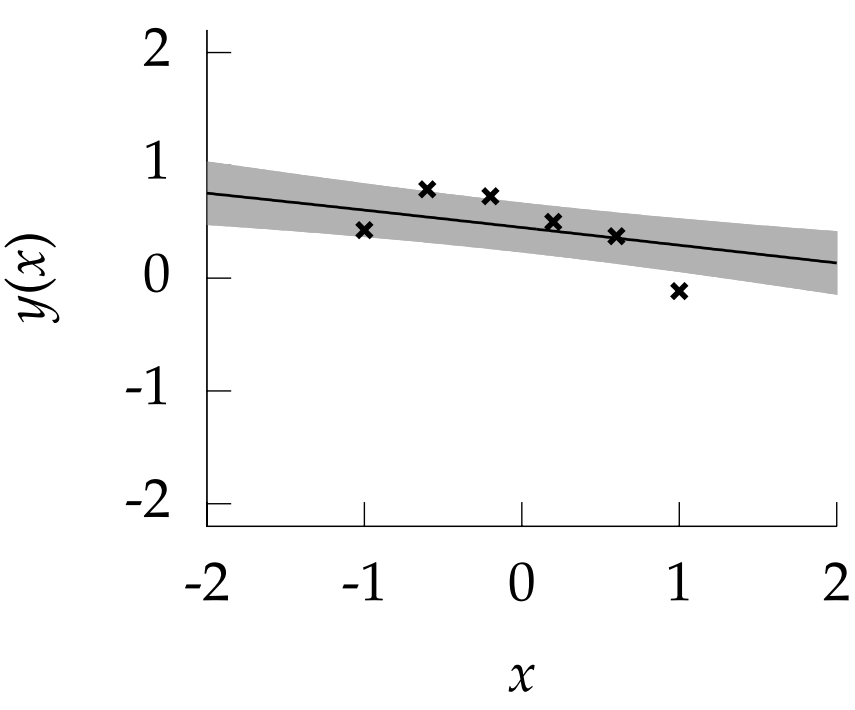


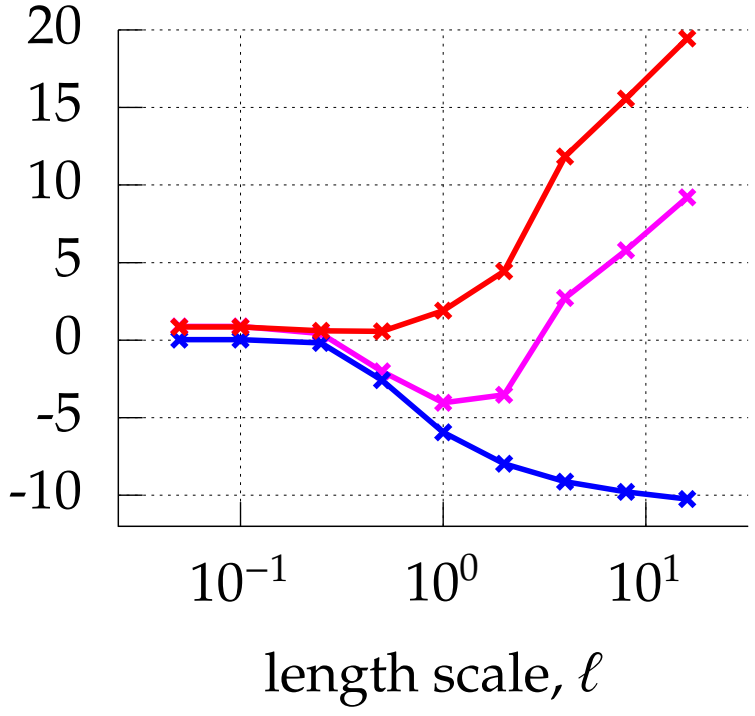


$$E(\boldsymbol{\theta}) = \frac{1}{2}\log|\mathbf{K}| + \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?





$$E(\boldsymbol{\theta}) = \frac{1}{2}\log|\mathbf{K}| + \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Gene Expression Example

- Given given expression levels in the form of a time series from Della Gatta et al. (2008).
- Want to detect if a gene is expressed or not, fit a GP to each gene (Kalaitzis and Lawrence, 2011).

Kalaitzis and Lawrence *BMC Bioinformatics* 2011, **12**:180
http://www.biomedcentral.com/1471-2105/12/180

BMC Bioinformatics

RESEARCH ARTICLE Open Access

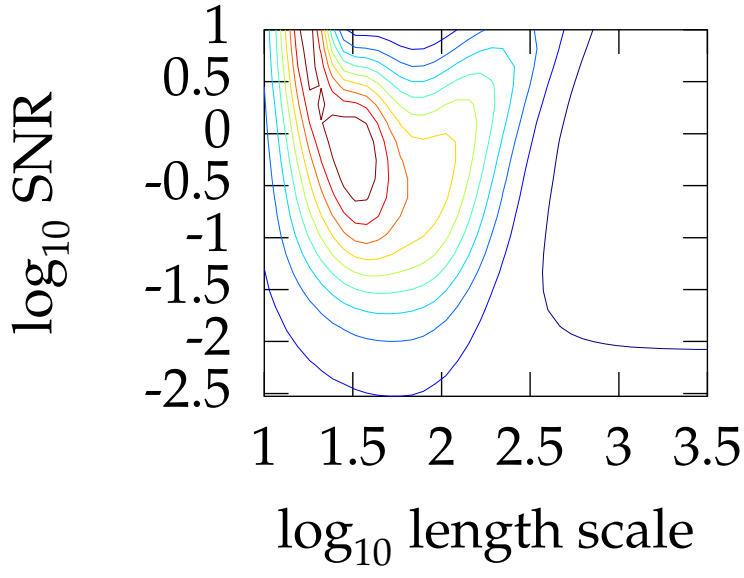
A Simple Approach to Ranking Differentially Expressed Gene Expression Time Courses through Gaussian Process Regression

Alfredo A Kalaitzis* and Neil D Lawrence*

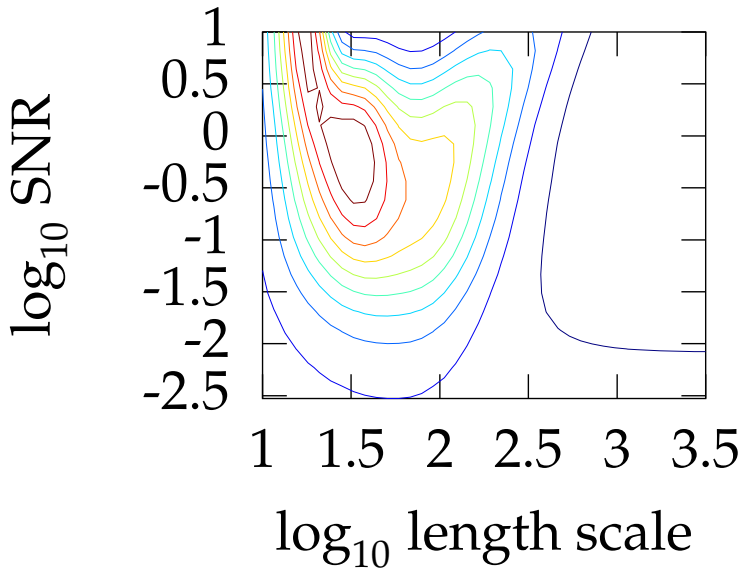
Abstract

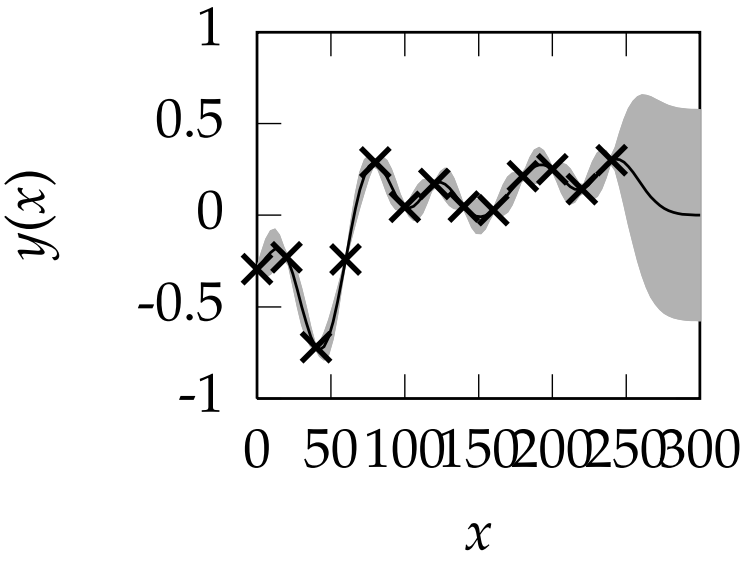
Background: The analysis of gene expression from time series underpins many biological studies. Two basic forms of analysis recur for data of this type: removing inactive (quiet) genes from the study and determining which genes are differentially expressed. Often these analysis stages are applied disregarding the fact that the data is drawn from a time series. In this paper we propose a simple model for accounting for the underlying temporal nature of the data based on a Gaussian process.

Results: We review Gaussian process (GP) regression for estimating the continuous trajectories underlying in gene expression time-series. We present a simple approach which can be used to filter quiet genes, or for the case of time series in the form of expression ratios, quantify differential expression. We assess via ROC curves the rankings produced by our regression framework and compare them to a recently proposed hierarchical Bayesian model for the analysis of gene expression time-series (BATS). We compare on both simulated and experimental data showing that the proposed approach considerably outperforms the current state of the art.

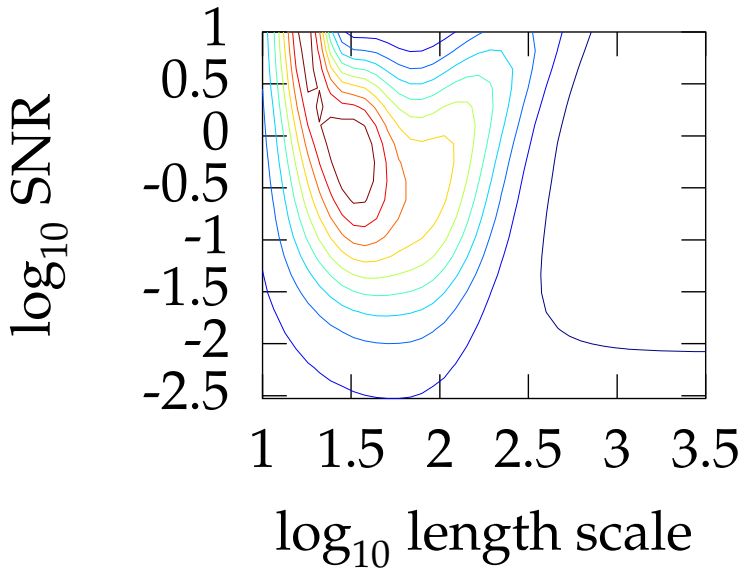


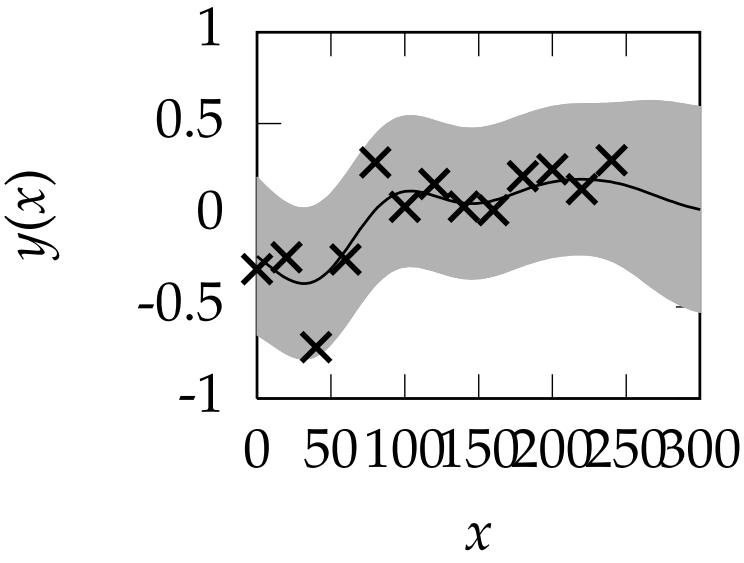
Contour plot of Gaussian process likelihood.



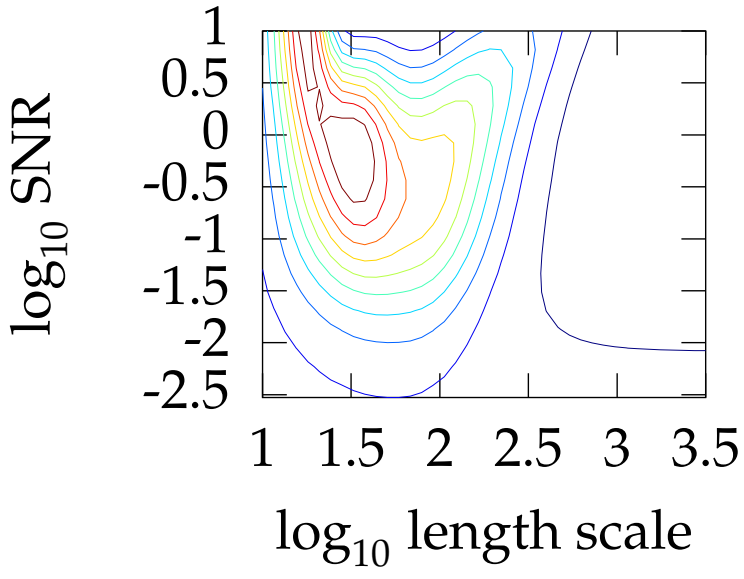


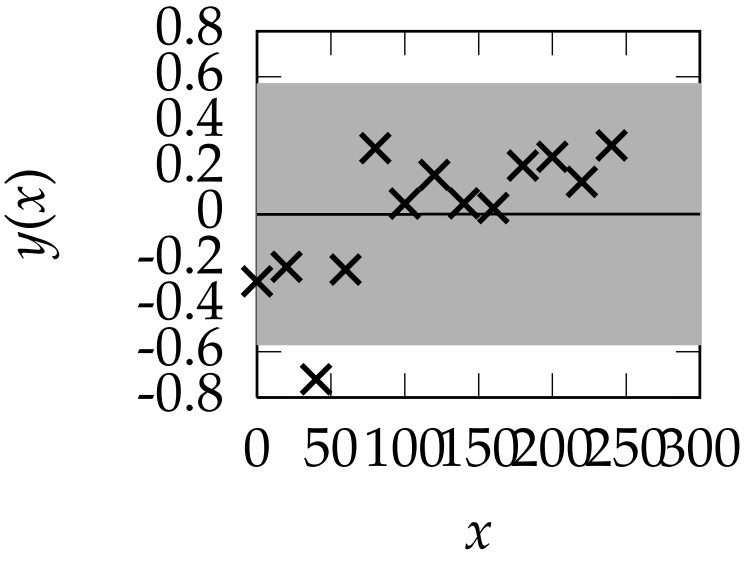
Optima: length scale of 1.2221 and $\log_{10}$ SNR of 1.9654 log likelihood is -0.22317.





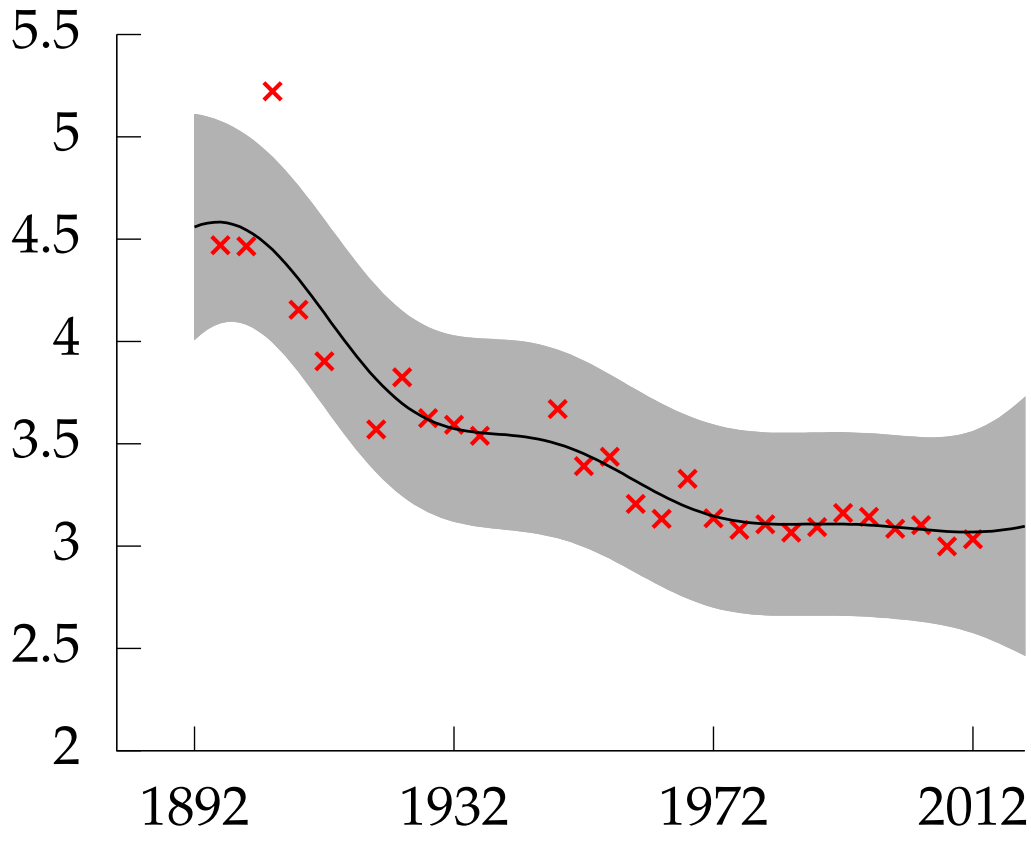
Optima: length scale of 1.5162 and $\log_{10}$ SNR of 0.21306 log likelihood is -0.23604.





Optima: length scale of 2.9886 and $\log_{10}$ SNR of -4.506 log likelihood is -2.1056.

Gaussian Process Fit to Olympic Marathon Data



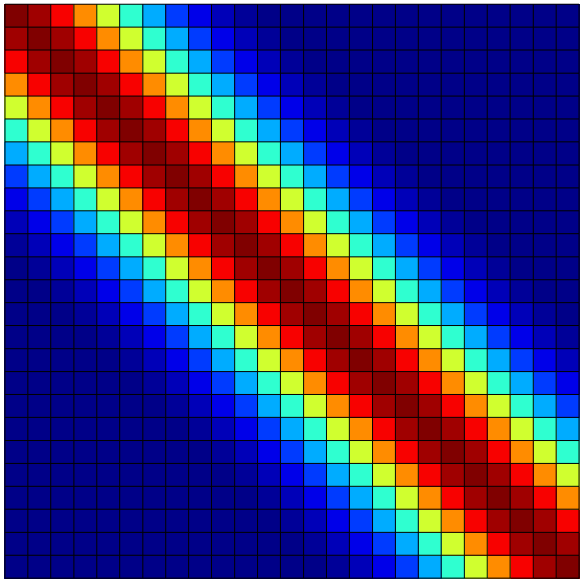
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.
- For the example above it was based on Euclidean distance.
- The covariance function is also know as a kernel.



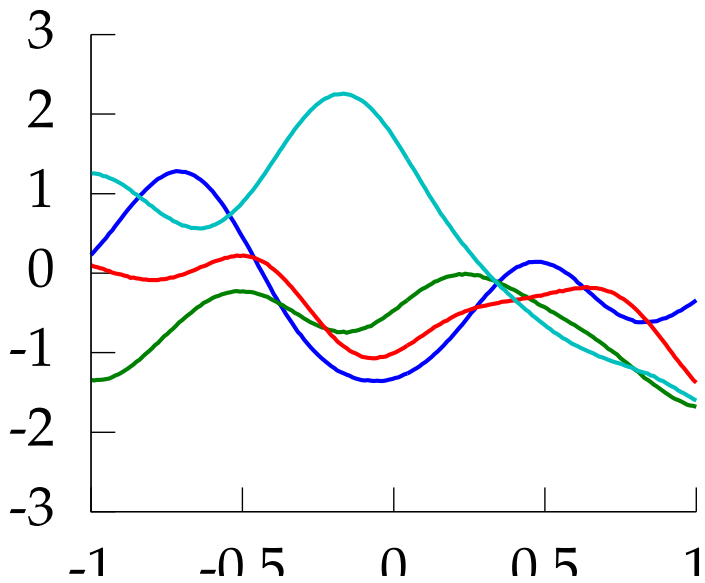
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.
- For the example above it was based on Euclidean distance.
- The covariance function is also know as a kernel.



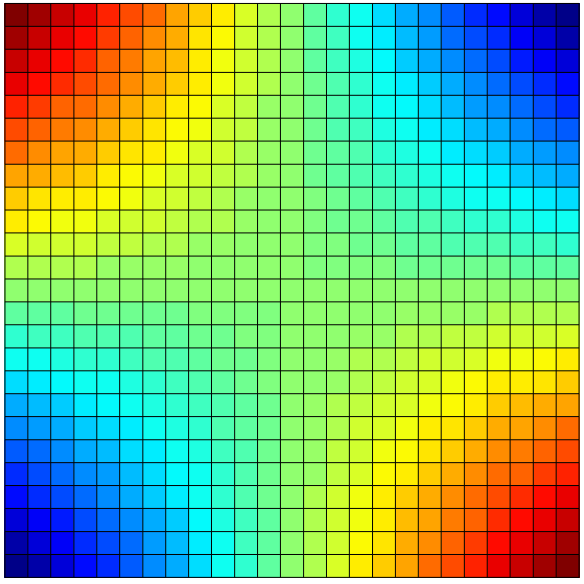
Covariance Functions

Linear Covariance Function

$$k(\mathbf{x}, \mathbf{x}') = \alpha \mathbf{x}^\top \mathbf{x}'$$

- Bayesian linear regression.

$$\alpha = 1$$



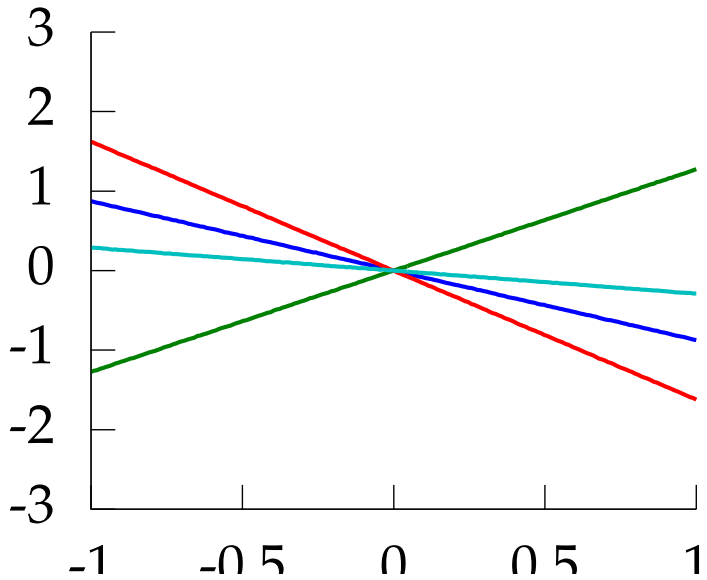
Covariance Functions

Linear Covariance Function

$$k(\mathbf{x}, \mathbf{x}') = \alpha \mathbf{x}^\top \mathbf{x}'$$

- Bayesian linear regression.

$$\alpha = 1$$



Covariance Functions

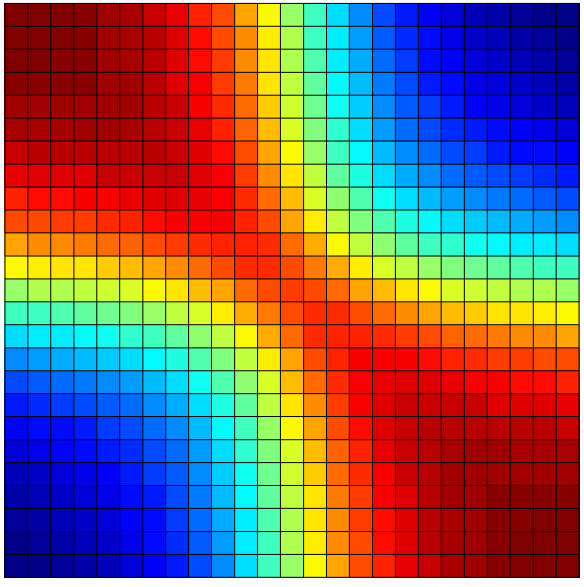
MLP Covariance Function

$$k(\mathbf{x}, \mathbf{x}') = \alpha \text{asin}\left(\frac{w\mathbf{x}^\top\mathbf{x}' + b}{\sqrt{w\mathbf{x}^\top\mathbf{x} + b + 1}\sqrt{w\mathbf{x}'^\top\mathbf{x}' + b + 1}}\right)$$

- Based on infinite neural network model.

$$w = 40$$

$$b = 4$$



Covariance Functions

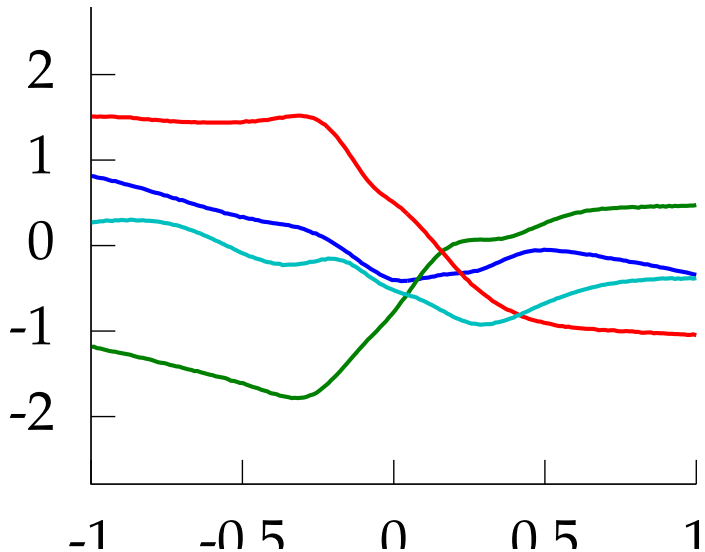
MLP Covariance Function

$$k(\mathbf{x}, \mathbf{x}') = \alpha \text{asin}\left(\frac{w\mathbf{x}^\top\mathbf{x}' + b}{\sqrt{w\mathbf{x}^\top\mathbf{x} + b + 1}\sqrt{w\mathbf{x}'^\top\mathbf{x}' + b + 1}}\right)$$

- Based on infinite neural network model.

$$w = 40$$

$$b = 4$$



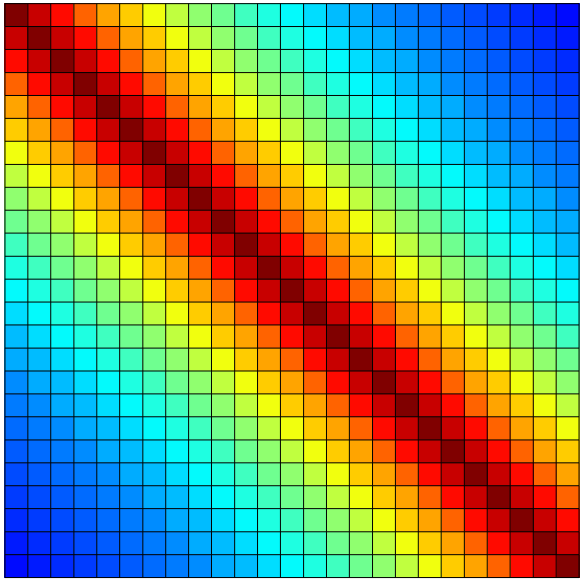
Covariance Functions

Where did this covariance matrix come from?

Ornstein-Uhlenbeck (stationary Gauss-Markov) covariance function

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp\left(-\frac{|\mathbf{x} - \mathbf{x}'|}{2\ell^2}\right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.



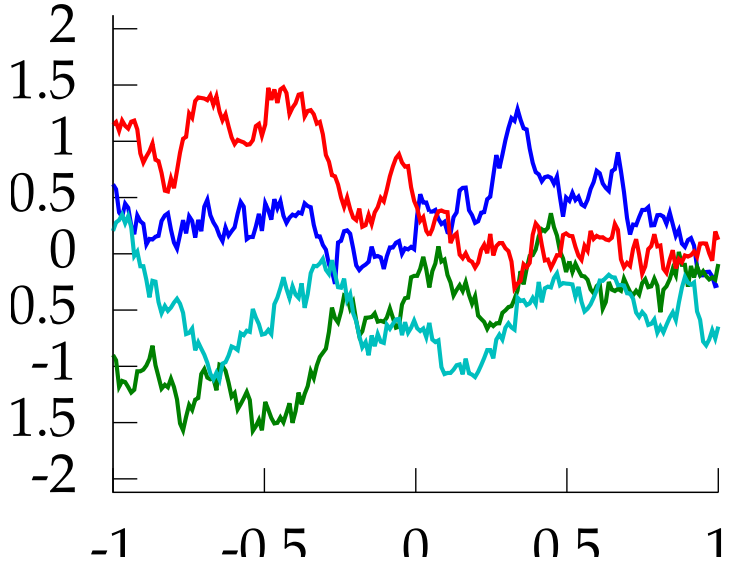
Covariance Functions

Where did this covariance matrix come from?

Ornstein-Uhlenbeck (stationary Gauss-Markov) covariance function

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp\left(-\frac{|\mathbf{x} - \mathbf{x}'|}{2\ell^2}\right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.



Outline

Simple Markov Chain

- Assume 1-d latent state, a vector over time, $\mathbf{x} = [x_1 \dots x_T]$.
- Markov property,

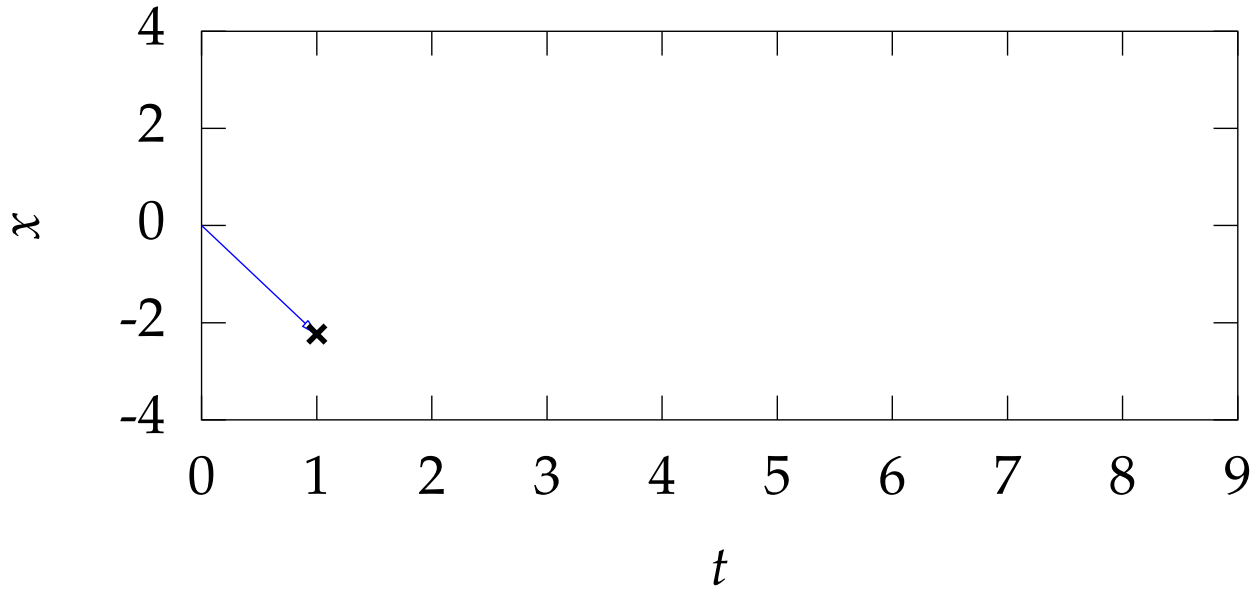
$$\begin{aligned} x_i =& x_{i-1} + \epsilon_i, \\ \epsilon_i \sim& \mathcal{N}(0, \alpha) \\ \implies x_i \sim& \mathcal{N}(x_{i-1}, \alpha) \end{aligned}$$

- Initial state,

$$x_0 \sim \mathcal{N}(0, \alpha_0)$$

- If $x_0 \sim \mathcal{N}(0, \alpha)$ we have a Markov chain for the latent states.
- Markov chain it is specified by an initial distribution (Gaussian) and a transition distribution (Gaussian).

Gauss Markov Chain

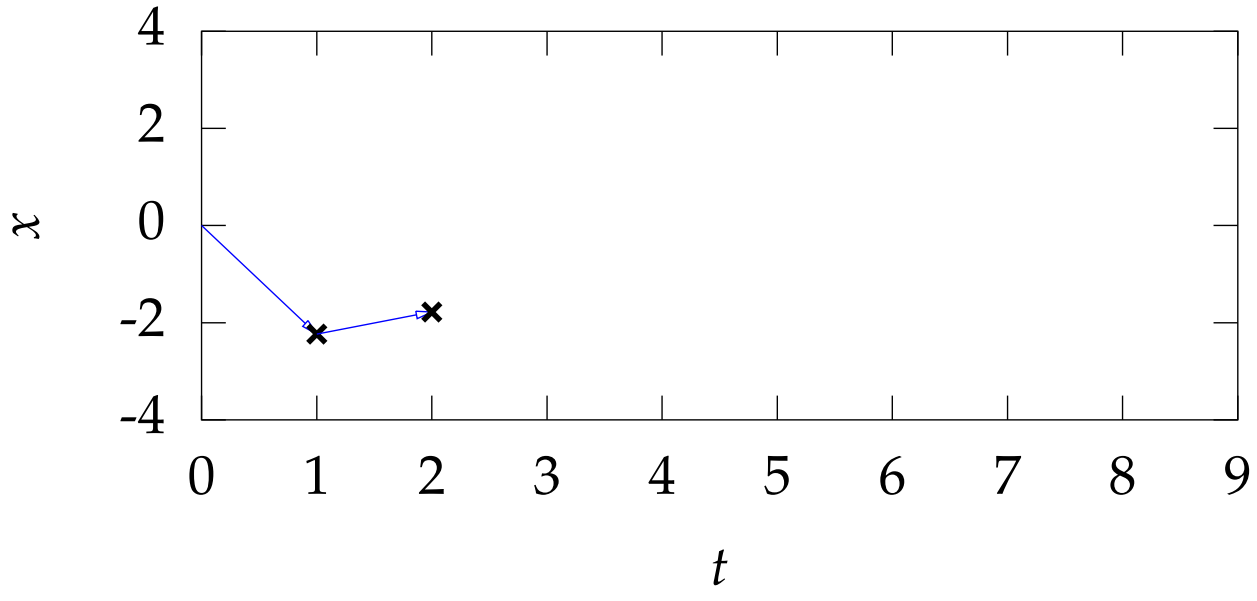


$$x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0,1)$$

$$x_0 = 0.000, \quad \epsilon_1 = -2.24$$

$$x_1 = 0.000 - 2.24 = -2.24$$

Gauss Markov Chain

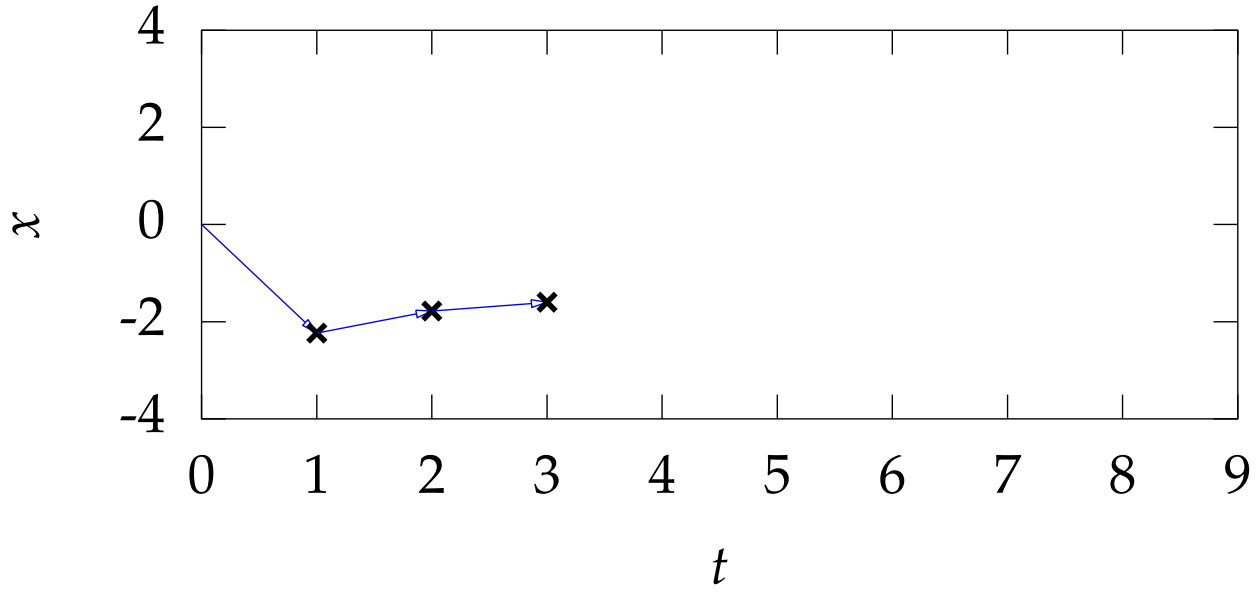


$$x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0,1)$$

$$x_1 = -2.24, \quad \epsilon_2 = 0.457$$

$$x_2 = -2.24 + 0.457 = -1.78$$

Gauss Markov Chain

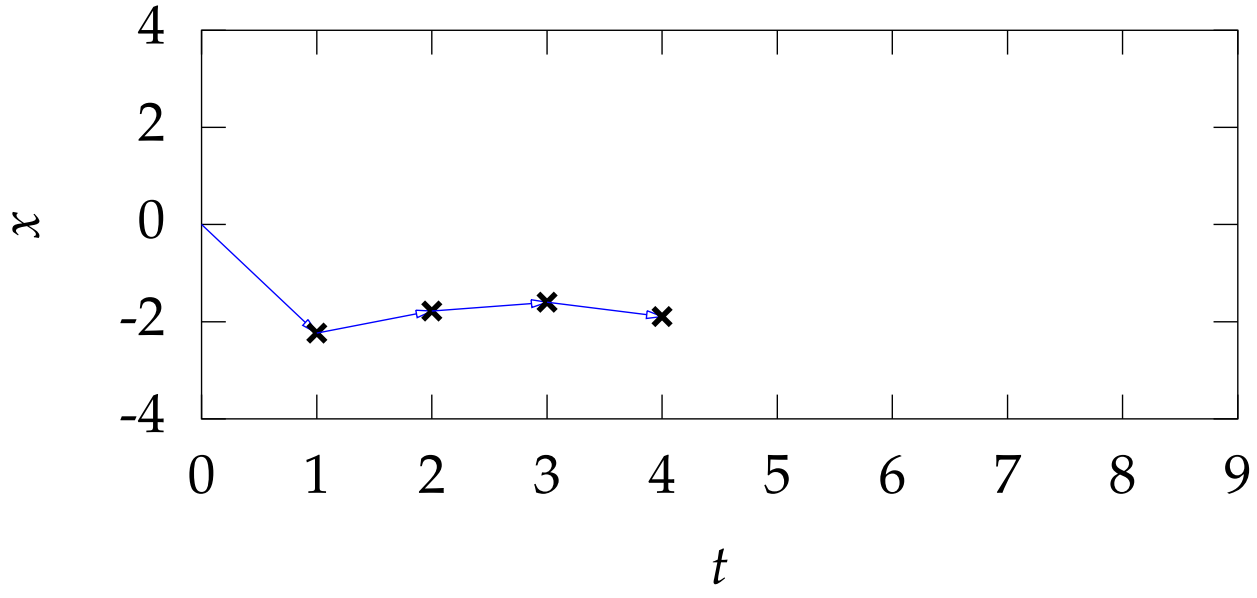


$$x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0,1)$$

$$x_2 = -1.78, \quad \epsilon_3 = 0.178$$

$$x_3 = -1.78 + 0.178 = -1.6$$

Gauss Markov Chain

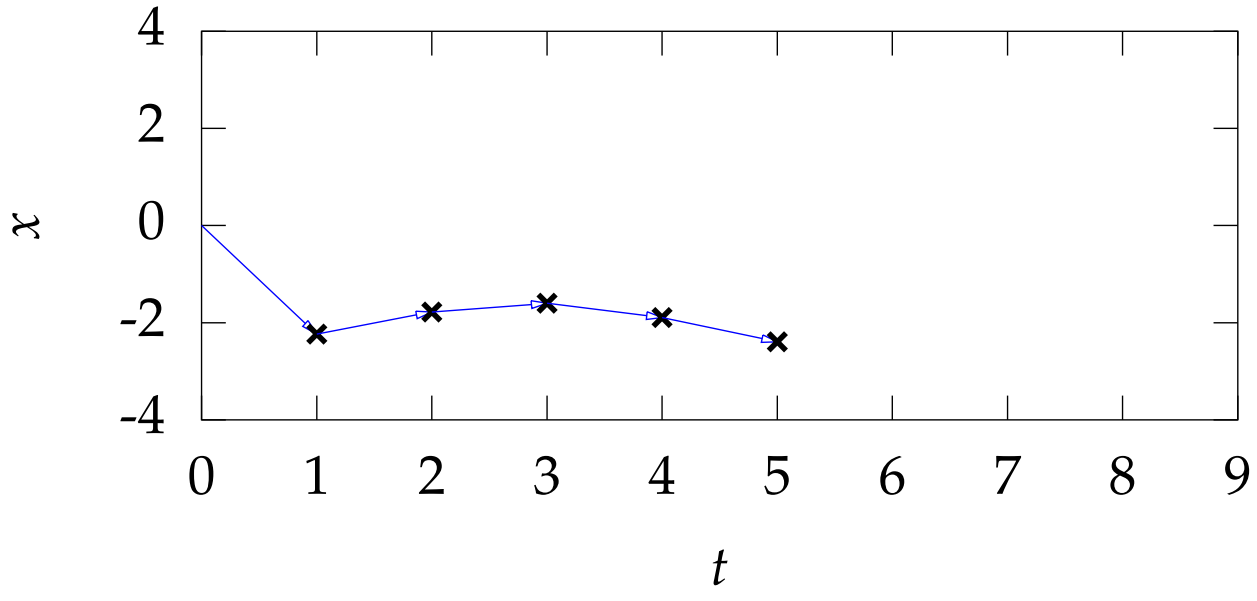


$$x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0,1)$$

$$x_3 = -1.6, \quad \epsilon_4 = -0.292$$

$$x_4 = -1.6 - 0.292 = -1.89$$

Gauss Markov Chain

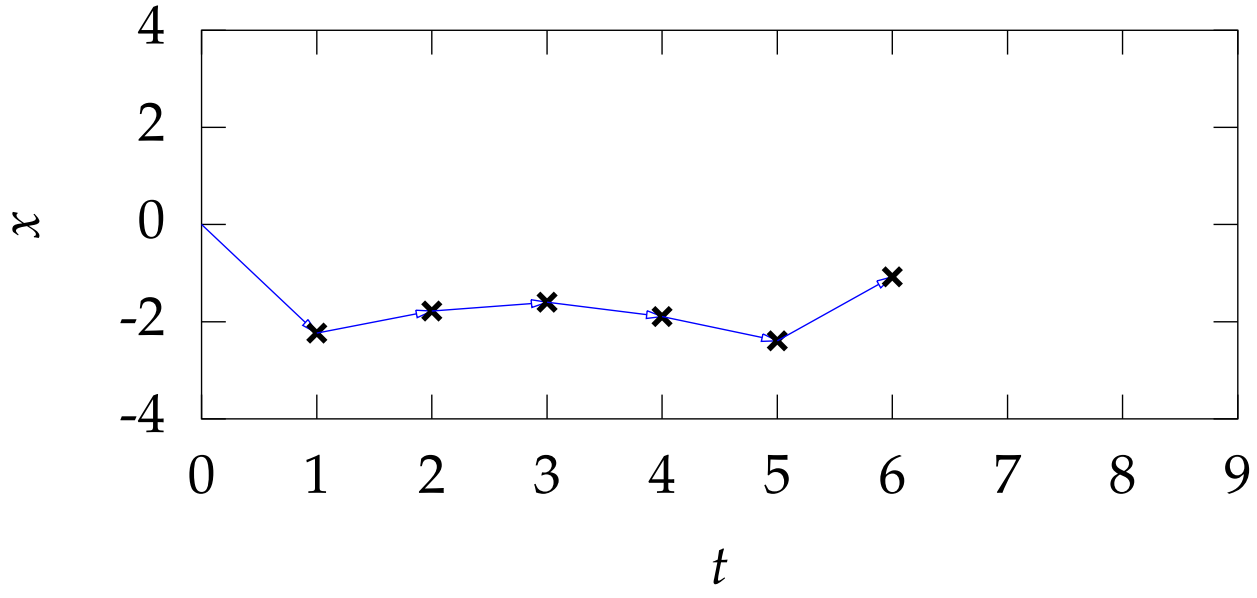


$$x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0,1)$$

$$x_4 = -1.89, \quad \epsilon_5 = -0.501$$

$$x_5 = -1.89 - 0.501 = -2.39$$

Gauss Markov Chain

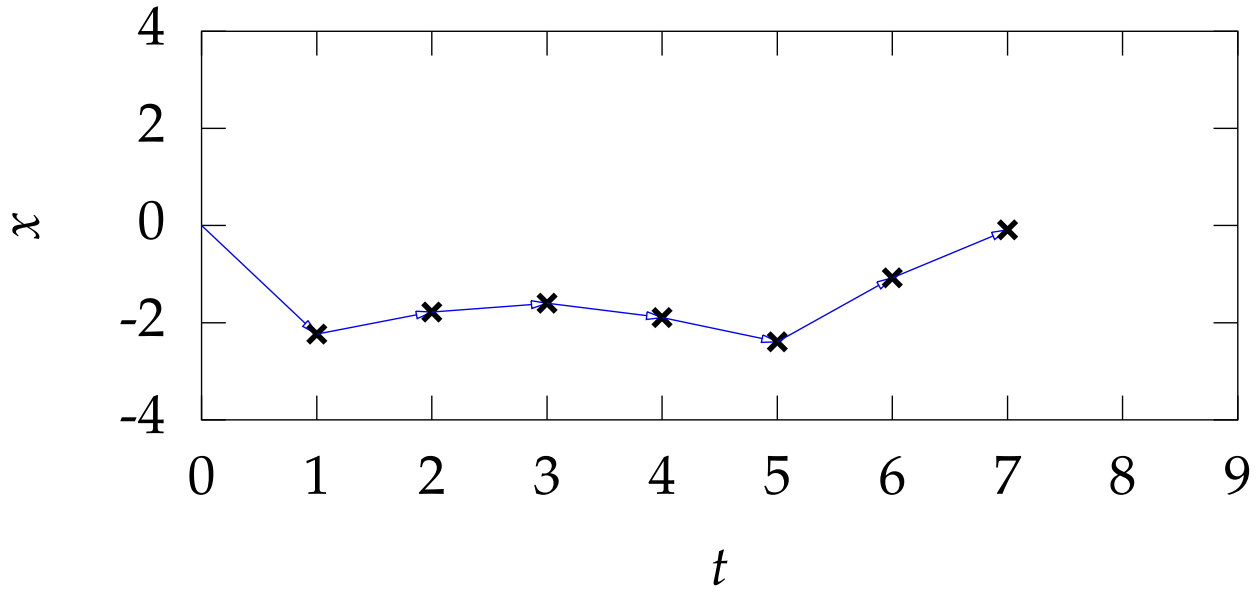


$$x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0,1)$$

$$x_5 = -2.39, \quad \epsilon_6 = 1.32$$

$$x_6 = -2.39 + 1.32 = -1.08$$

Gauss Markov Chain

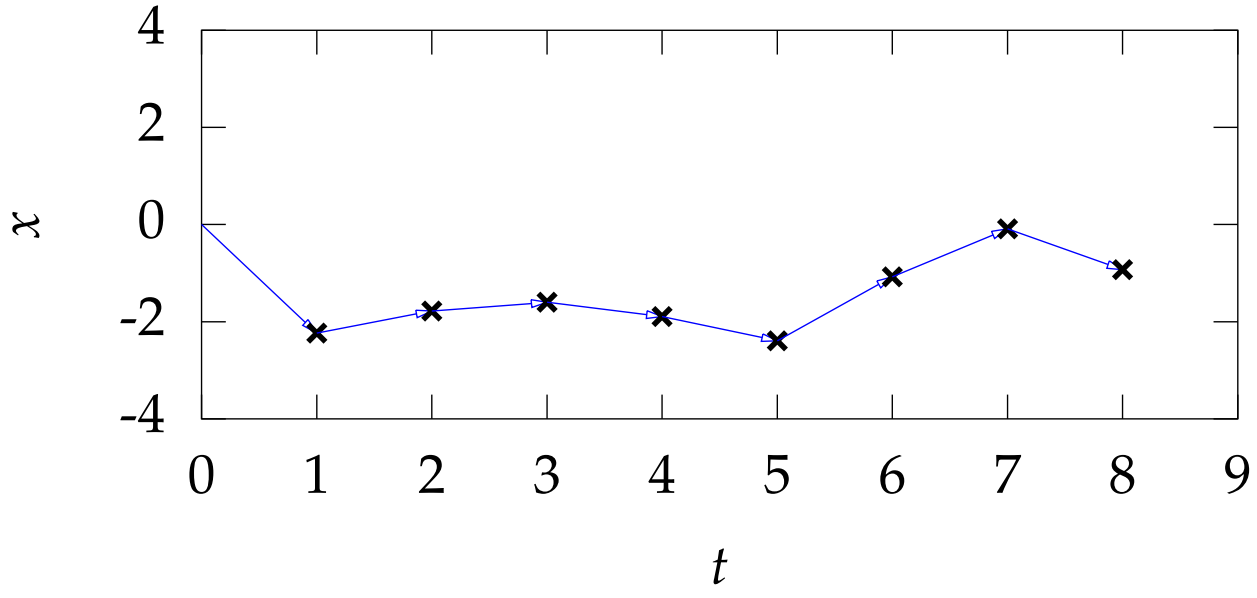


$$x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0,1)$$

$$x_6 = -1.08, \quad \epsilon_7 = 0.989$$

$$x_7 = -1.08 + 0.989 = -0.0881$$

Gauss Markov Chain

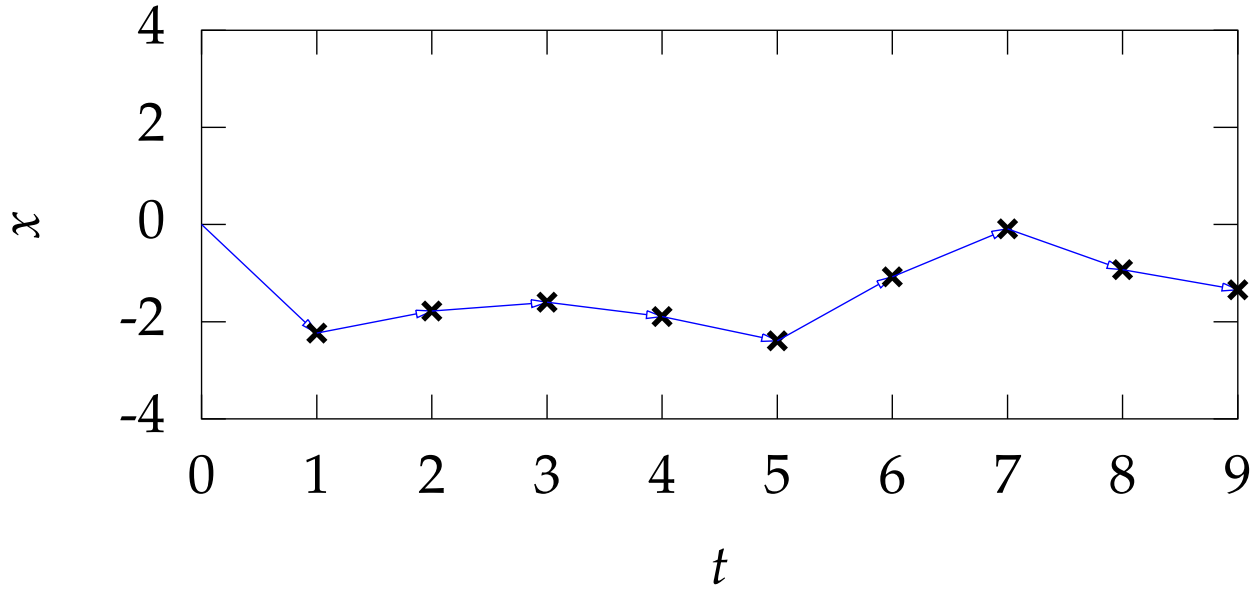


$$x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0,1)$$

$$x_7 = -0.0881, \quad \epsilon_8 = -0.842$$

$$x_8 = -0.0881 - 0.842 = -0.93$$

Gauss Markov Chain



$$x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0,1)$$

$$x_8 = -0.93, \quad \epsilon_9 = -0.41$$

$$x_9 = -0.93 - 0.410 = -1.34$$

Multivariate Gaussian Properties: Reminder

If

$$\mathbf{z} \sim \mathcal{N}\left(\mu, \mathbf{C}\right)$$

and

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \mathbf{b}$$

then

$$\mathbf{x} \sim \mathcal{N}\left(\mathbf{W}\mu + \mathbf{b}, \mathbf{W}\mathbf{C}\mathbf{W}^\top\right)$$

Multivariate Gaussian Properties: Reminder

Simplified: If

$$\mathbf{z} \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}\right)$$

and

$$\mathbf{x} = \mathbf{W}\mathbf{z}$$

then

$$\mathbf{x} \sim \mathcal{N}\left(0, \sigma^2 \mathbf{W}\mathbf{W}^\top\right)$$

Matrix Representation of Latent Variables

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix}$$

$$x_1 = \epsilon_1$$

Matrix Representation of Latent Variables

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix}$$

$$x_2 = \epsilon_1 + \epsilon_2$$

Matrix Representation of Latent Variables

$$\begin{bmatrix} x_1 \\ x_2 \\ \boxed{x_3} \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ \boxed{1 & 1 & 1 & 0 & 0} \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix}$$

$$x_3 = \epsilon_1 + \epsilon_2 + \epsilon_3$$

Matrix Representation of Latent Variables

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix}$$

$$x_4 = \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4$$

Matrix Representation of Latent Variables

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix}$$

$$x_5 = \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 + \epsilon_5$$

Matrix Representation of Latent Variables

$$\mathbf{x} = \mathbf{L}_1 \times \boldsymbol{\epsilon}$$

Multivariate Process

- Since $\mathbf{x}$ is linearly related to $\boldsymbol{\epsilon}$ we know $\mathbf{x}$ is a Gaussian process.
- Trick: we only need to compute the mean and covariance of $\mathbf{x}$ to determine that Gaussian.

Latent Process Mean

$$\mathbf{x} = \mathbf{L}_1 \epsilon$$

Latent Process Mean

$$\langle \mathbf{x} \rangle = \langle \mathbf{L}_1 \boldsymbol{\epsilon} \rangle$$

Latent Process Mean

$$\langle \mathbf{x} \rangle = \mathbf{L}_1 \langle \epsilon \rangle$$

Latent Process Mean

$$\langle \mathbf{x} \rangle = \mathbf{L}_1 \langle \epsilon \rangle$$

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \alpha \mathbf{I})$$

Latent Process Mean

$$\langle \mathbf{x} \rangle = \mathbf{L}_1 \mathbf{0}$$

Latent Process Mean

$$\langle \mathbf{x} \rangle = \mathbf{0}$$

Latent Process Covariance

$$\mathbf{x}\mathbf{x}^\top = \mathbf{L}_1 \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mathbf{L}_1^\top$$

$$\mathbf{x}^\top = \boldsymbol{\epsilon}^\top \mathbf{L}^\top$$

Latent Process Covariance

$$\left\langle \mathbf{x}\mathbf{x}^{\top} \right\rangle = \left\langle \mathbf{L}_1 \boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\top} \mathbf{L}_1^{\top} \right\rangle$$

Latent Process Covariance

$$\left\langle \mathbf{x}\mathbf{x}^{\top} \right\rangle = \mathbf{L}_1 \left\langle \boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\top} \right\rangle \mathbf{L}_1^{\top}$$

Latent Process Covariance

$$\left\langle \mathbf{x}\mathbf{x}^{\top} \right\rangle = \mathbf{L}_1 \left\langle \boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\top} \right\rangle \mathbf{L}_1^{\top}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \alpha\mathbf{I}\right)$$

Latent Process Covariance

$$\left\langle \mathbf{x}\mathbf{x}^{\top} \right\rangle = \alpha \mathbf{L}_1 \mathbf{L}_1^{\top}$$

Latent Process

$$\mathbf{x} = \mathbf{L}_1 \epsilon$$

Latent Process

$$\mathbf{x} = \mathbf{L}_1 \boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \alpha \mathbf{I})$$

Latent Process

$$\mathbf{x} = \mathbf{L}_1 \epsilon$$

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \alpha \mathbf{I})$$

$$\Longrightarrow$$

Latent Process

$$\mathbf{x} = \mathbf{L}_1 \boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \alpha \mathbf{I})$$

$$\Longrightarrow$$

$$\mathbf{x} \sim \mathcal{N}\left(\mathbf{0}, \alpha \mathbf{L}_1 \mathbf{L}_1^\top\right)$$

Covariance for Latent Process II

- Make the variance dependent on time interval.
- Assume variance grows *linearly* with time.
- Justification: sum of two Gaussian distributed random variables is distributed as Gaussian with sum of variances.
- If variable's movement is additive over time (as described) variance scales linearly with time.

Covariance for Latent Process II

- Given
$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \alpha\mathbf{I}) \Longrightarrow \boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \alpha\mathbf{L_1}\mathbf{L_1}^\top\right).$$
Then
$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \Delta t\alpha\mathbf{I}) \Longrightarrow \boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \Delta t\alpha\mathbf{L_1}\mathbf{L_1}^\top\right).$$
where Δt is the time interval between observations.

Covariance for Latent Process II

$$\epsilon \sim \mathcal{N}\left(0, \alpha \Delta t \mathbf{I}\right), \quad \mathbf{x} \sim \mathcal{N}\left(0, \alpha \Delta t \mathbf{L_1} \mathbf{L_1}^\top\right)$$

Covariance for Latent Process II

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \alpha \Delta t \mathbf{I}\right), \quad \mathbf{x} \sim \mathcal{N}\left(0, \alpha \Delta t \mathbf{L_1} \mathbf{L_1}^\top\right)$$

$$\mathbf{K} = \alpha \Delta t \mathbf{L_1} \mathbf{L_1}^\top$$

Covariance for Latent Process II

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \alpha\Delta t \mathbf{I}\right), \quad \mathbf{x} \sim \mathcal{N}\left(0, \alpha\Delta t \mathbf{L_1}\mathbf{L_1}^\top\right)$$

$$\mathbf{K} = \alpha\Delta t \mathbf{L_1}\mathbf{L_1}^\top$$

$$k_{i,j} = \alpha\Delta t \mathbf{l}_{:,i}^\top \mathbf{l}_{:,j}$$

where $\mathbf{l}_{:,k}$ is a vector from the kth row of $\mathbf{L_1}$: the first k elements are one, the next $T - k$ are zero.

Covariance for Latent Process II

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \alpha\Delta t \mathbf{I}\right), \quad \mathbf{x} \sim \mathcal{N}\left(0, \alpha\Delta t \mathbf{L_1}\mathbf{L_1}^\top\right)$$

$$\mathbf{K} = \alpha\Delta t \mathbf{L_1}\mathbf{L_1}^\top$$

$$k_{i,j} = \alpha\Delta t \mathbf{l}_{:,i}^\top \mathbf{l}_{:,j}$$

where $\mathbf{l}_{:,k}$ is a vector from the kth row of $\mathbf{L_1}$: the first k elements are one, the next $T - k$ are zero.

$$k_{i,j} = \alpha\Delta t \min(i, j)$$

define $\Delta t i = t_i$ so

$$k_{i,j} = \alpha \min(t_i, t_j) = k(t_i, t_j)$$

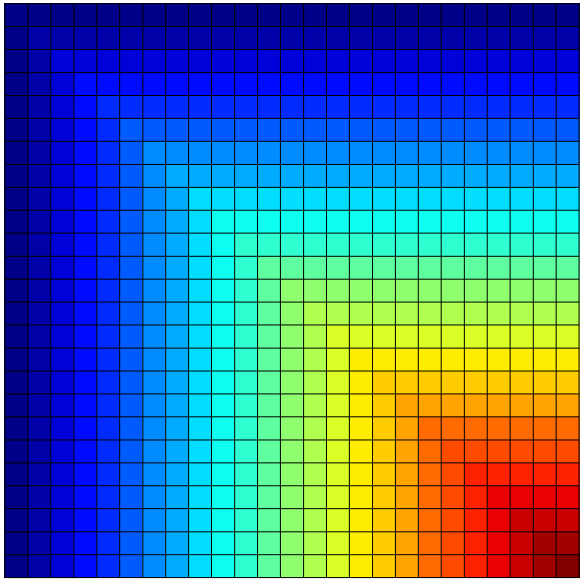
Covariance Functions

Where did this covariance matrix come from?

Markov Process

$$k(t, t') = \alpha \min(t, t')$$

- Covariance matrix is built using the *inputs* to the function t.



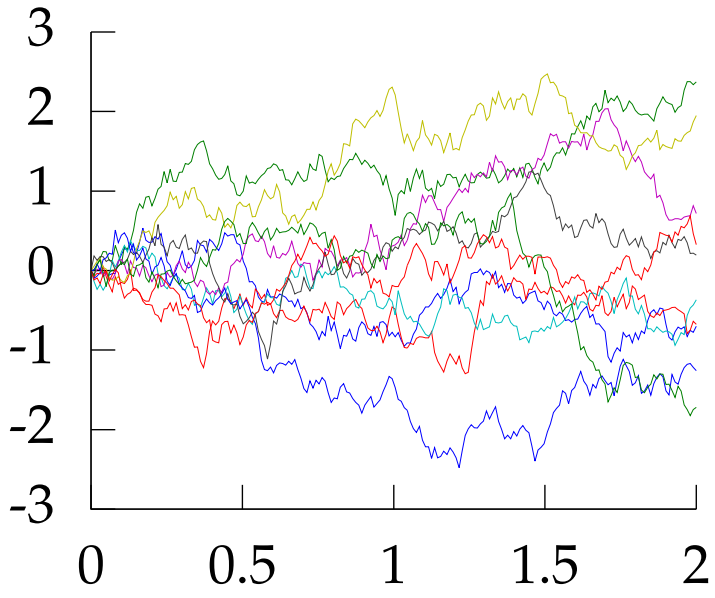
Covariance Functions

Where did this covariance matrix come from?

Markov Process

$$k(t, t') = \alpha \min(t, t')$$

- Covariance matrix is built using the *inputs* to the function t.



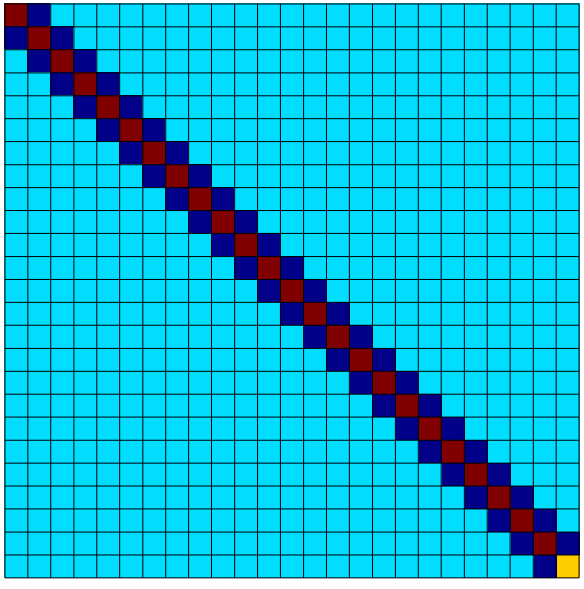
Covariance Functions

Where did this covariance matrix come from?

Markov Process

Visualization of inverse covariance (precision).

- Precision matrix is sparse: only neighbours in matrix are non-zero.
- This reflects *conditional* independencies in data.
- In this case *Markov* structure.



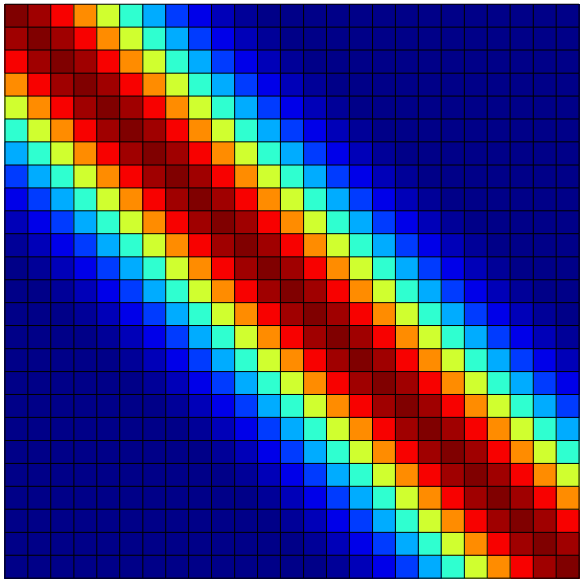
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.
- For the example above it was based on Euclidean distance.
- The covariance function is also know as a kernel.



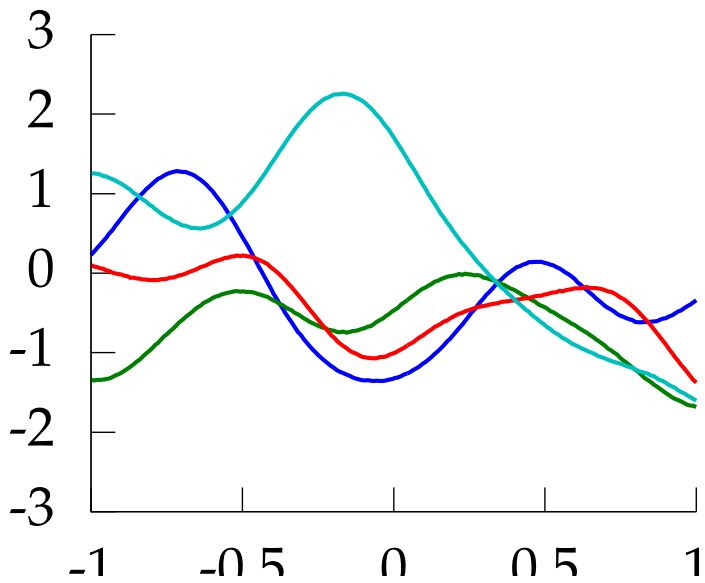
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.
- For the example above it was based on Euclidean distance.
- The covariance function is also know as a kernel.



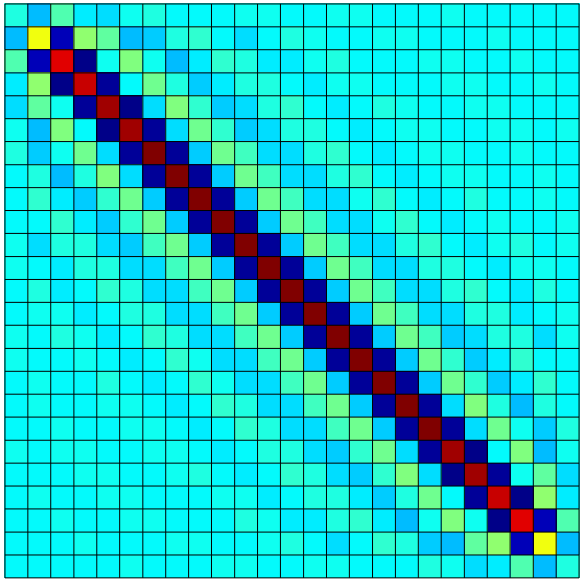
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic

Visualization of inverse covariance (precision).

- Precision matrix is not sparse.
- Each point is dependent on all the others.
- In this case non-Markovian.



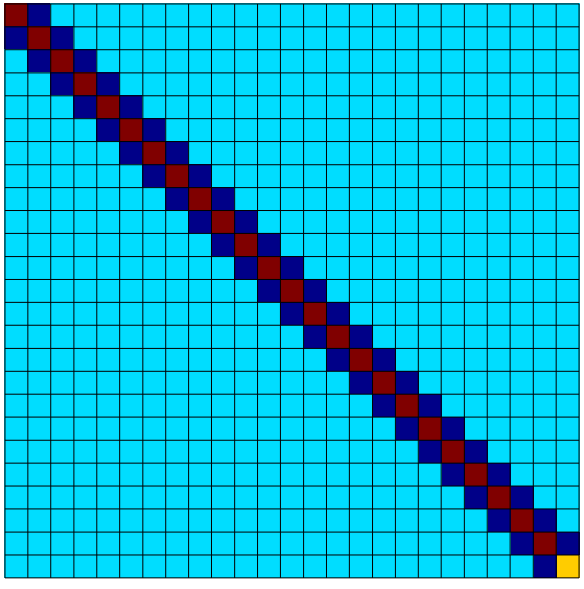
Covariance Functions

Where did this covariance matrix come from?

Markov Process

Visualization of inverse covariance (precision).

- Precision matrix is sparse: only neighbours in matrix are non-zero.
- This reflects *conditional* independencies in data.
- In this case *Markov* structure.



Simple Kalman Filter I

- We have state vector $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \dots \mathbf{x}_q \end{bmatrix} \in \mathbb{R}^{T \times q}$ and if each state evolves independently we have

$$p(\mathbf{X}) = \prod_{i=1}^{q} p(\mathbf{x}_{:,i})$$
$$p(\mathbf{x}_{:,i}) = \mathcal{N}\left(\mathbf{x}_{:,i} | \mathbf{0}, \mathbf{K}\right).$$

- We want to obtain outputs through:

$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:}$$

Stacking and Kronecker Products I

- Represent with a 'stacked' system:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I} \otimes \mathbf{K})$$

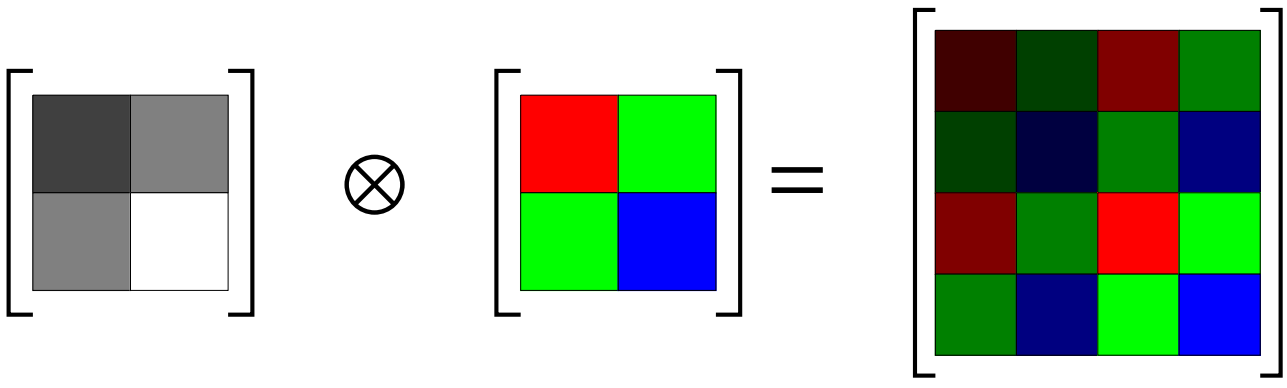
where the stacking is placing each column of $\mathbf{X}$ one on top of another as

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_{:,1} \\ \mathbf{x}_{:,2} \\ \vdots \\ \mathbf{x}_{:,q} \end{bmatrix}$$

Kronecker Product

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \otimes \mathbf{K} = \begin{bmatrix} a\mathbf{K} & b\mathbf{K} \\ c\mathbf{K} & d\mathbf{K} \end{bmatrix}$$

Kronecker Product



Stacking and Kronecker Products I

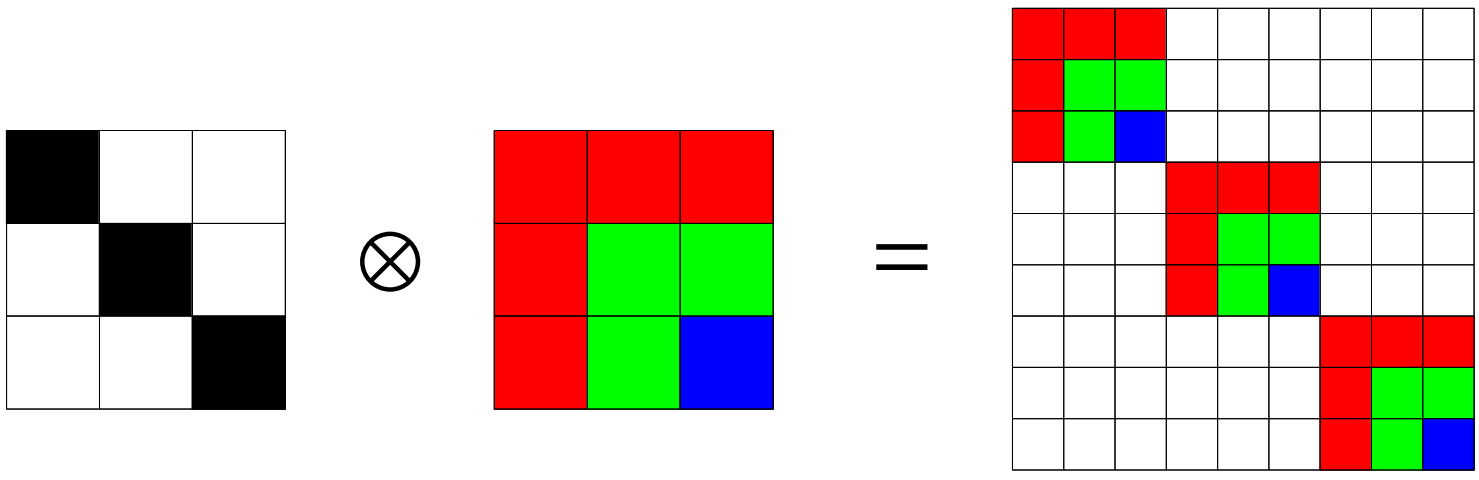
- Represent with a 'stacked' system:

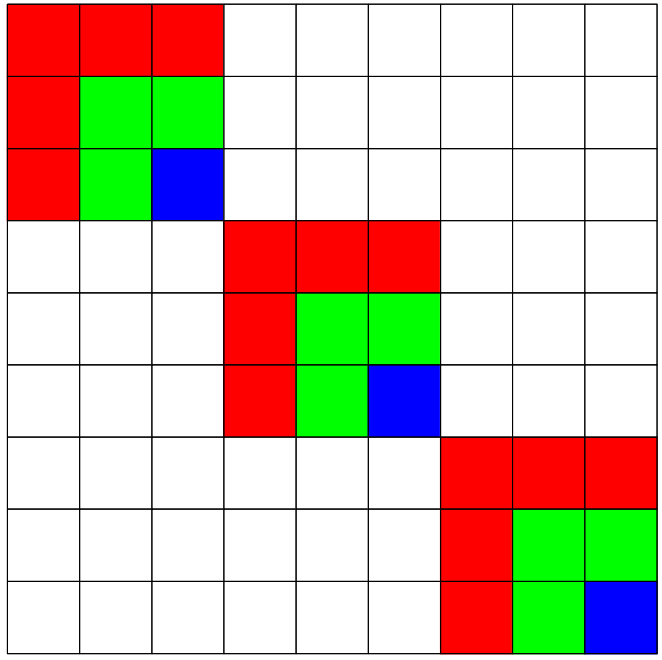
$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I} \otimes \mathbf{K})$$

where the stacking is placing each column of $\mathbf{X}$ one on top of another as

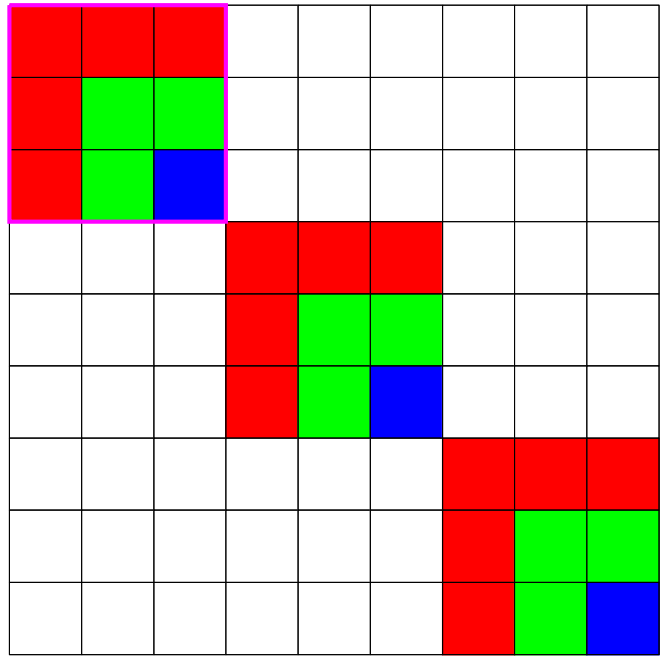
$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_{:,1} \\ \mathbf{x}_{:,2} \\ \vdots \\ \mathbf{x}_{:,q} \end{bmatrix}$$

Column Stacking

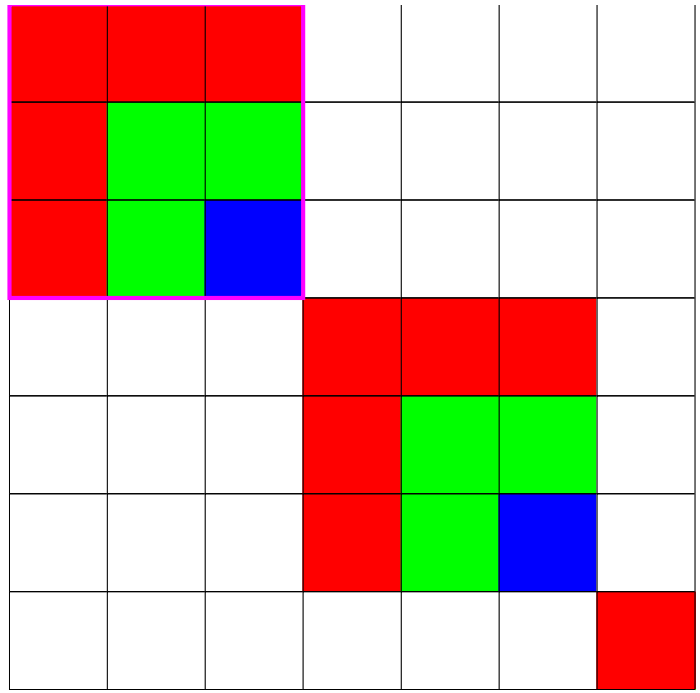




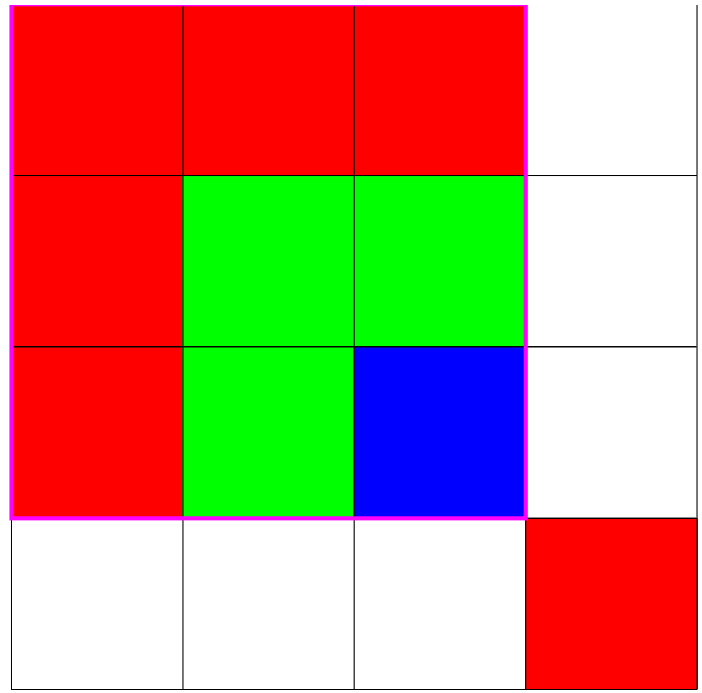
For this stacking the marginal distribution over *time* is given by the block diagonals.



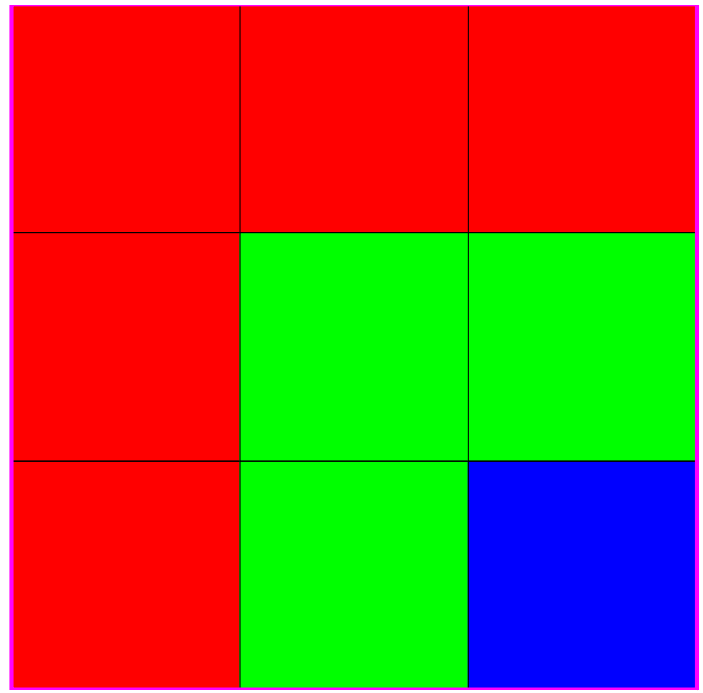
For this stacking the marginal distribution over *time* is given by the block diagonals.



For this stacking the marginal distribution over *time* is given by the block diagonals.



For this stacking the marginal distribution over *time* is given by the block diagonals.



For this stacking the marginal distribution over *time* is given by the block diagonals.

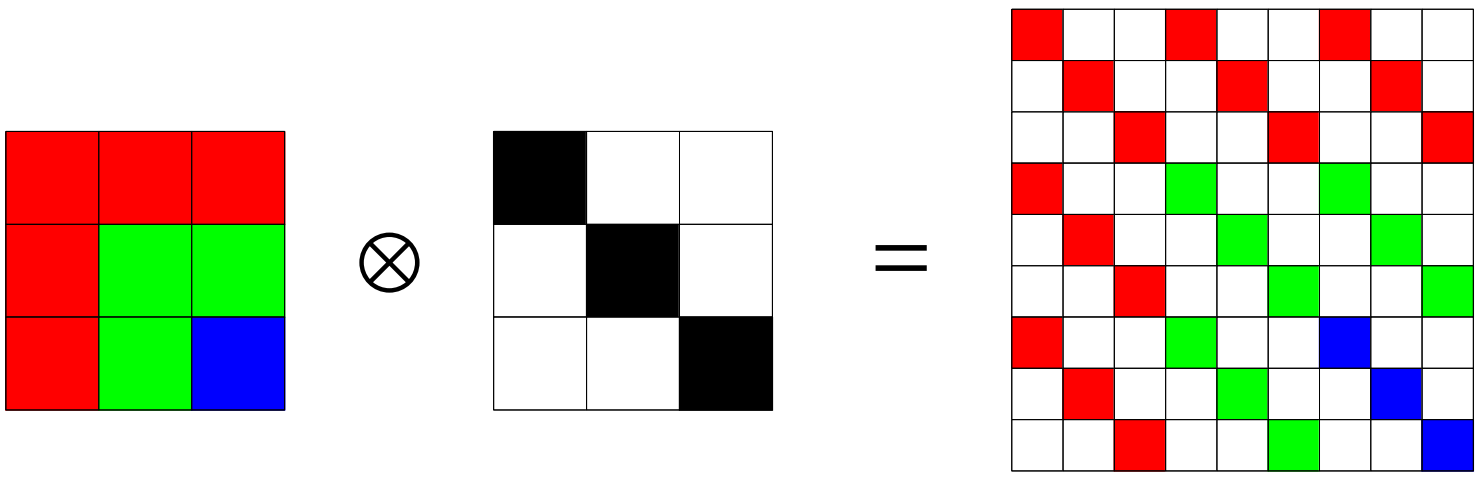
Two Ways of Stacking

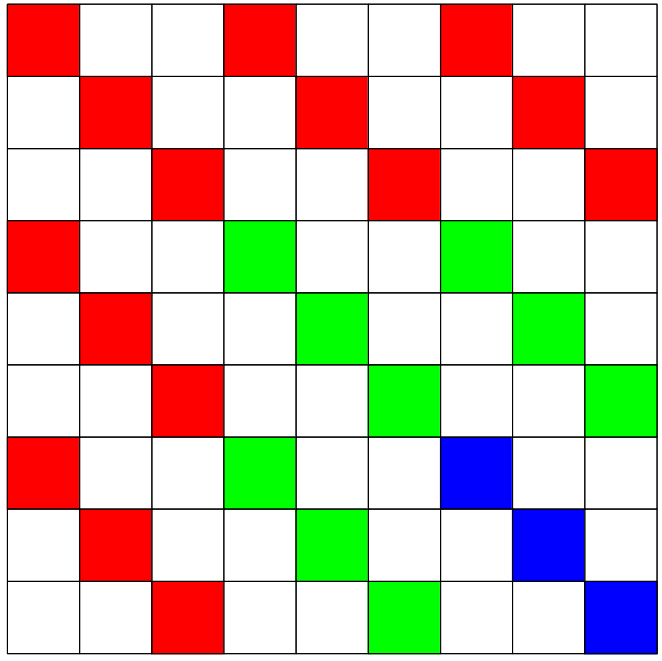
Can also stack each row of $\mathbf{X}$ to form column vector:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_{1,:} \\ \mathbf{x}_{2,:} \\ \vdots \\ \mathbf{x}_{T,:} \end{bmatrix}$$

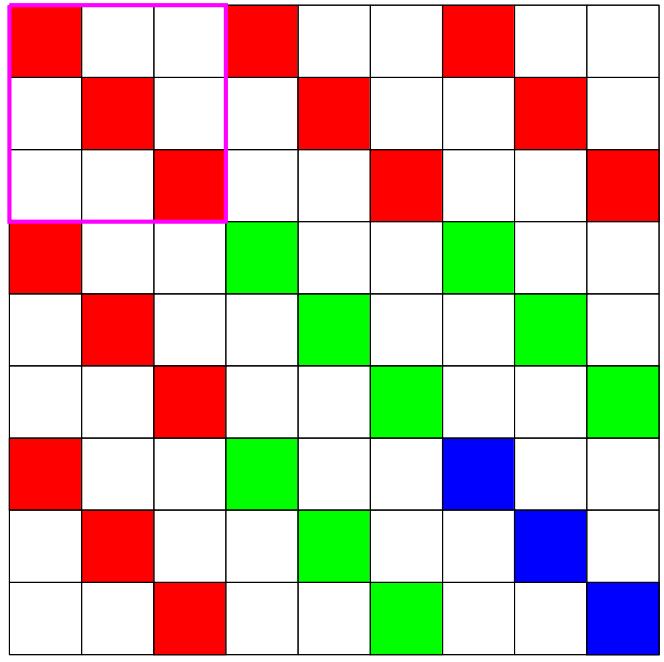
$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\mathbf{0}, \mathbf{K} \otimes \mathbf{I}\right)$$

Row Stacking

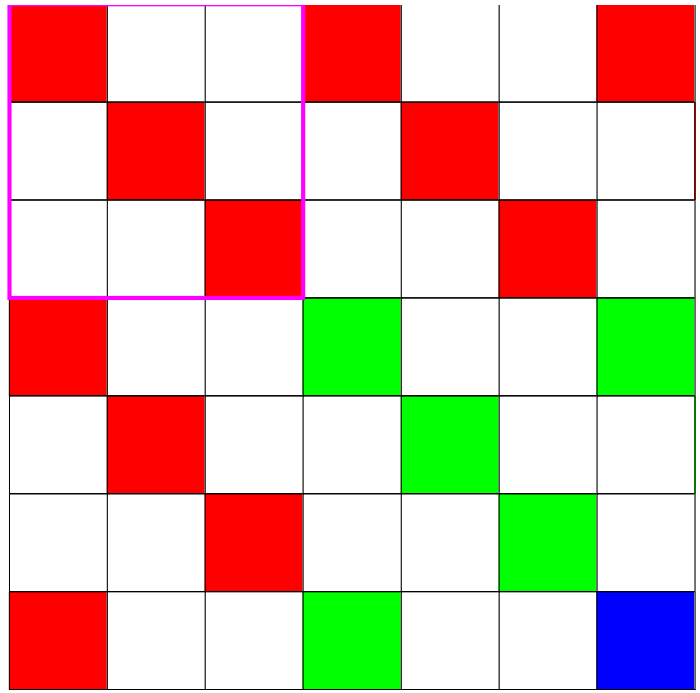




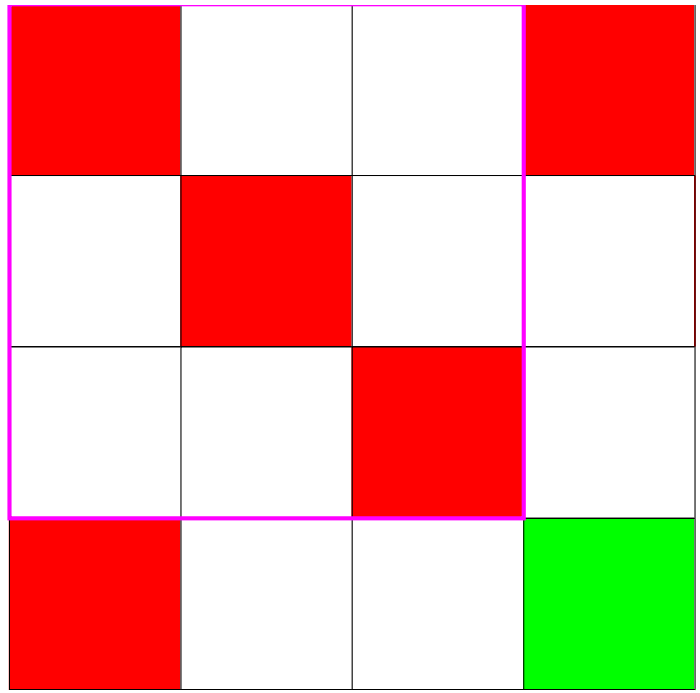
For this stacking the marginal distribution over the latent *dimensions* is given by the block diagonals.



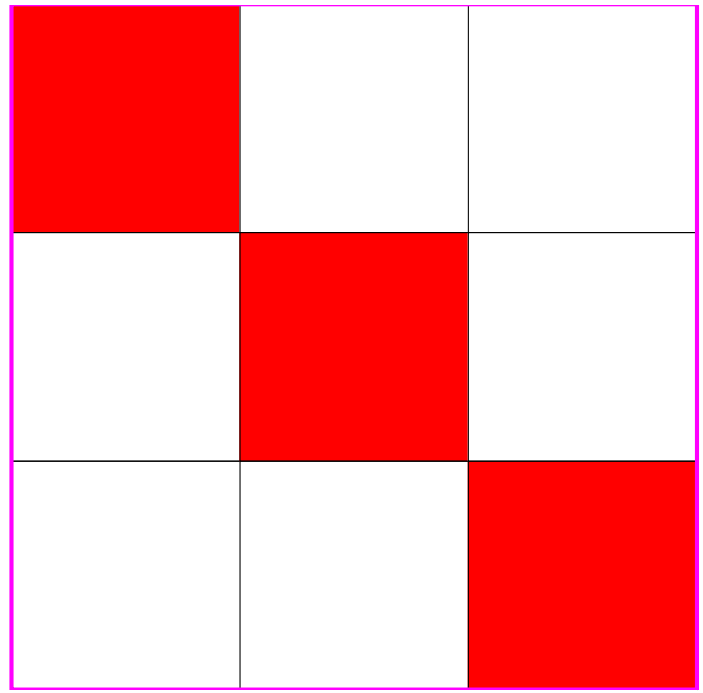
For this stacking the marginal distribution over the latent *dimensions* is given by the block diagonals.



For this stacking the marginal distribution over the latent *dimensions* is given by the block diagonals.



For this stacking the marginal distribution over the latent *dimensions* is given by the block diagonals.



For this stacking the marginal distribution over the latent *dimensions* is given by the block diagonals.

Observed Process

The observations are related to the latent points by a linear mapping matrix,

$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right)$$

Mapping from Latent Process to Observed

$$\begin{bmatrix} \mathbf{W} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{W} \end{bmatrix} \times \begin{bmatrix} \mathbf{x}_{1,:} \\ \mathbf{x}_{2,:} \\ \mathbf{x}_{3,:} \end{bmatrix} = \begin{bmatrix} \mathbf{W}\mathbf{x}_{1,:} \\ \mathbf{W}\mathbf{x}_{2,:} \\ \mathbf{W}\mathbf{x}_{3,:} \end{bmatrix}$$

Output Covariance

This leads to a covariance of the form

$$(\mathbf{I} \otimes \mathbf{W})(\mathbf{K} \otimes \mathbf{I})(\mathbf{I} \otimes \mathbf{W}^\top) + \mathbf{I}\sigma^2$$

Using $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$ This leads to

$$\mathbf{K} \otimes \mathbf{W}\mathbf{W}^\top + \mathbf{I}\sigma^2$$

or

$$\mathbf{y} \sim \mathcal{N}\left(0, \mathbf{W}\mathbf{W}^\top \otimes \mathbf{K} + \mathbf{I}\sigma^2\right)$$

Foundations and Trends® in
Machine Learning
Vol. 4, No. 3 (2011) 195–266

DOI: 10.1561/2200000036

now
the essence of knowledge

Kernels for Vector-Valued Functions: A Review

By Mauricio A. Álvarez,
Lorenzo Rosasco and Neil D. Lawrence

Kronecker Structure GPs

- This Kronecker structure leads to several published models.

$$(\mathbf{K}(\mathbf{x}, \mathbf{x}'))_{j,j'} = k(\mathbf{x}, \mathbf{x}')k_T(j, j'),$$

  where k has $\mathbf{x}$ and k_T has i as inputs.
- Can think of multiple output covariance functions as covariances with augmented input.
- Alongside $\mathbf{x}$ we also input the j associated with the *output* of interest.

Separable Covariance Functions

- Taking $\mathbf{B} = \mathbf{W}\mathbf{W}^\top$ we have a matrix expression across outputs.

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}')\mathbf{B},$$

  where $\mathbf{B}$ is a $p \times p$ symmetric and positive semi-definite matrix.
- $\mathbf{B}$ is called the *coregionalization* matrix.
- We call this class of covariance functions *separable* due to their product structure.

Sum of Separable Covariance Functions

- In the same spirit a more general class of kernels is given by

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{q} k_j(\mathbf{x}, \mathbf{x}')\mathbf{B}_j.$$

- This can also be written as

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \sum_{j=1}^{q} \mathbf{B}_j \otimes k_j(\mathbf{X}, \mathbf{X}),$$

- This is like several Kalman filter-type models added together, but each one with a different set of latent functions.
- We call this class of kernels sum of separable kernels (SoS kernels).

Geostatistics

- Use of GPs in Geostatistics is called kriging.
- These multi-output GPs pioneered in geostatistics: prediction over vector-valued output data is known as *cokriging*.
- The model in geostatistics is known as the *linear model of coregionalization* (LMC, Journel and Huijbregts (1978); Goovaerts (1997)).
- Most machine learning multitask models can be placed in the context of the LMC model.

Weighted sum of Latent Functions

- In the linear model of coregionalization (LMC) outputs are expressed as linear combinations of independent random functions.
- In the LMC, each component f_j is expressed as a linear sum

$$f_j(\mathbf{x}) = \sum_{j=1}^{q} w_{j,j} u_j(\mathbf{x}).$$

  where the latent functions are independent and have covariance functions $k_j(\mathbf{x}, \mathbf{x}')$.
- The processes $\{f_j(\mathbf{x})\}_{j=1}^{q}$ are independent for $q \neq j'$.

Kalman Filter Special Case

- The Kalman filter is an example of the LMC where $u_i(\mathbf{x}) \to x_i(t)$.
- I.e. we've moved form time input to a more general input space.
- In matrix notation:
  1. Kalman filter
  $$\mathbf{F} = \mathbf{WX}$$
  2. LMC
  $$\mathbf{F} = \mathbf{WU}$$

  where the rows of these matrices $\mathbf{F}$, $\mathbf{X}$, $\mathbf{U}$ each contain q samples from their corresponding functions at a different time (Kalman filter) or spatial location (LMC).

Intrinsic Coregionalization Model

- If one covariance used for latent functions (like in Kalman filter).
- This is called the intrinsic coregionalization model (ICM, Goovaerts (1997)).
- The kernel matrix corresponding to a dataset $\mathbf{X}$ takes the form

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \mathbf{B} \otimes k(\mathbf{X}, \mathbf{X}).$$

Autokrigeability

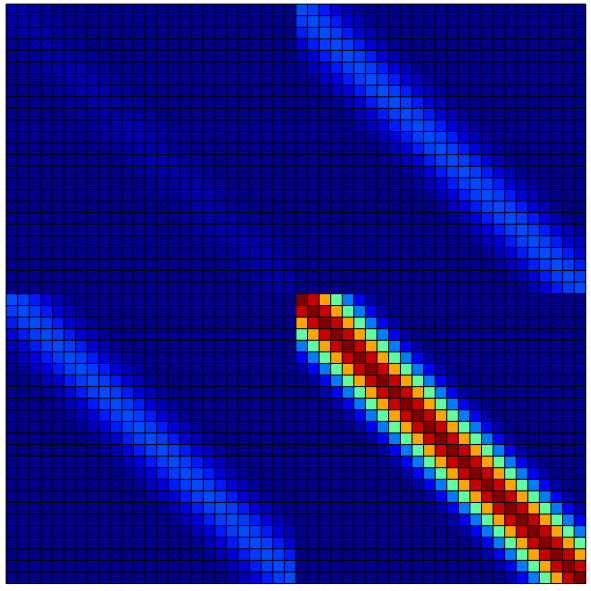
- If outputs are noise-free, maximum likelihood is equivalent to independent fits of $\mathbf{B}$ and $k(\mathbf{x}, \mathbf{x}')$ (Helterbrand and Cressie, 1994).
- In geostatistics this is known as autokrigeability (Wackernagel, 2003).
- In multitask learning its the cancellation of intertask transfer (Bonilla et al., 2008).

Intrinsic Coregionalization Model

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \mathbf{w}\mathbf{w}^\top \otimes k(\mathbf{X}, \mathbf{X}).$$

$$\mathbf{w} = \begin{bmatrix} 1 \\ 5 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 1 & 5 \\ 5 & 25 \end{bmatrix}$$

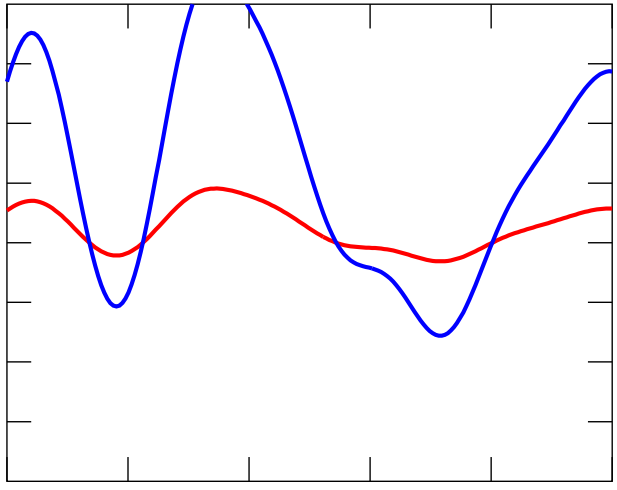


Intrinsic Coregionalization Model

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \mathbf{w}\mathbf{w}^\top \otimes k(\mathbf{X}, \mathbf{X}).$$

$$\mathbf{w} = \begin{bmatrix} 1 \\ 5 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 1 & 5 \\ 5 & 25 \end{bmatrix}$$

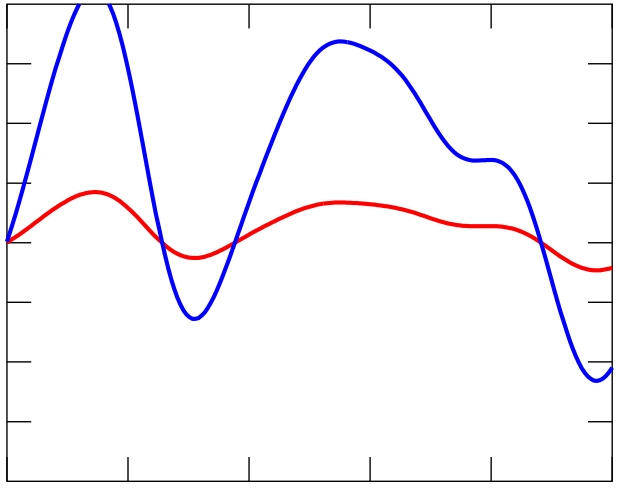


Intrinsic Coregionalization Model

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \mathbf{w}\mathbf{w}^\top \otimes k(\mathbf{X}, \mathbf{X}).$$

$$\mathbf{w} = \begin{bmatrix} 1 \\ 5 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 1 & 5 \\ 5 & 25 \end{bmatrix}$$

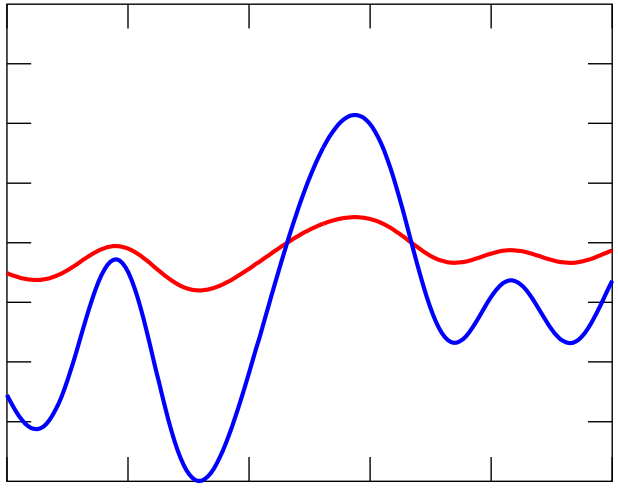


Intrinsic Coregionalization Model

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \mathbf{w}\mathbf{w}^\top \otimes k(\mathbf{X}, \mathbf{X}).$$

$$\mathbf{w} = \begin{bmatrix} 1 \\ 5 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 1 & 5 \\ 5 & 25 \end{bmatrix}$$

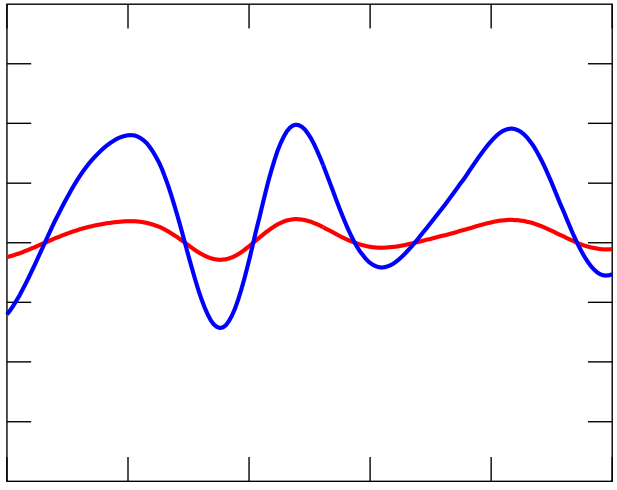


Intrinsic Coregionalization Model

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \mathbf{w}\mathbf{w}^\top \otimes k(\mathbf{X}, \mathbf{X}).$$

$$\mathbf{w} = \begin{bmatrix} 1 \\ 5 \end{bmatrix}$$

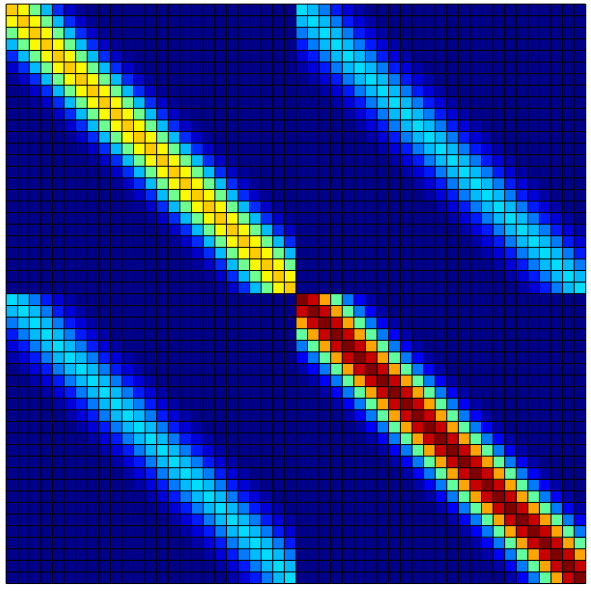
$$\mathbf{B} = \begin{bmatrix} 1 & 5 \\ 5 & 25 \end{bmatrix}$$



Intrinsic Coregionalization Model

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \mathbf{B} \otimes k(\mathbf{X}, \mathbf{X}).$$

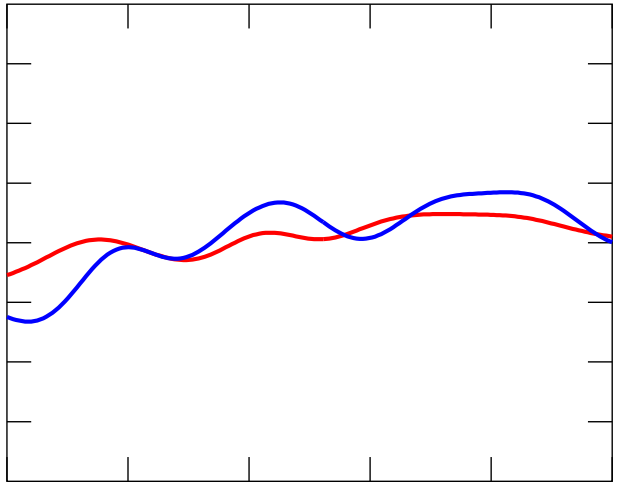
$$\mathbf{B} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.5 \end{bmatrix}$$



Intrinsic Coregionalization Model

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \mathbf{B} \otimes k(\mathbf{X}, \mathbf{X}).$$

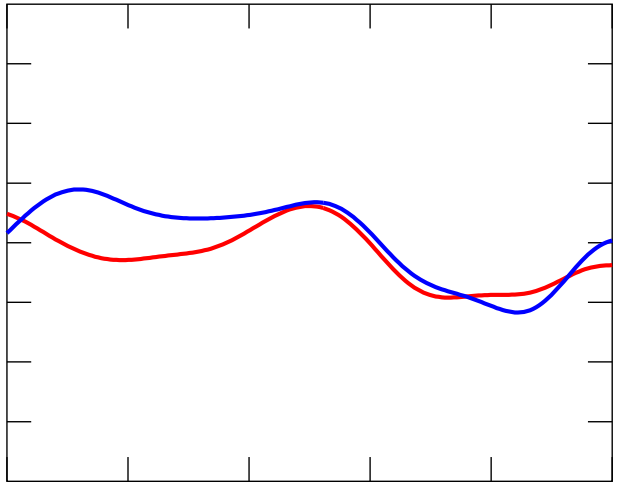
$$\mathbf{B} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.5 \end{bmatrix}$$



Intrinsic Coregionalization Model

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \mathbf{B} \otimes k(\mathbf{X}, \mathbf{X}).$$

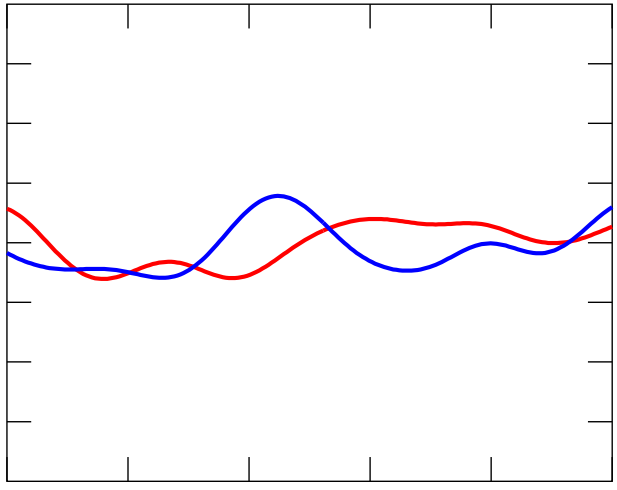
$$\mathbf{B} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.5 \end{bmatrix}$$



Intrinsic Coregionalization Model

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \mathbf{B} \otimes k(\mathbf{X}, \mathbf{X}).$$

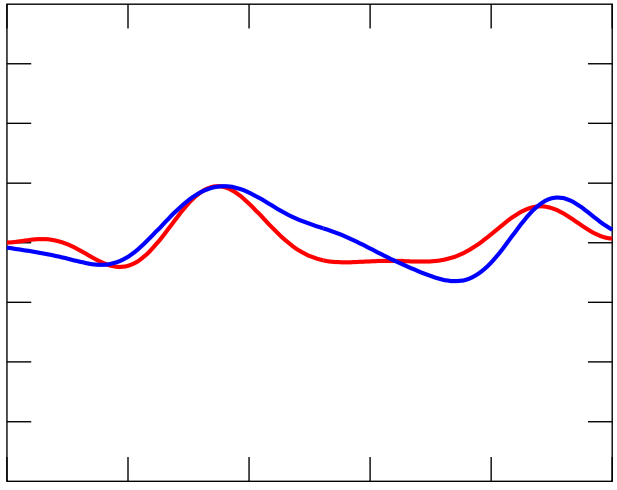
$$\mathbf{B} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.5 \end{bmatrix}$$



Intrinsic Coregionalization Model

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \mathbf{B} \otimes k(\mathbf{X}, \mathbf{X}).$$

$$\mathbf{B} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.5 \end{bmatrix}$$



LMC Samples

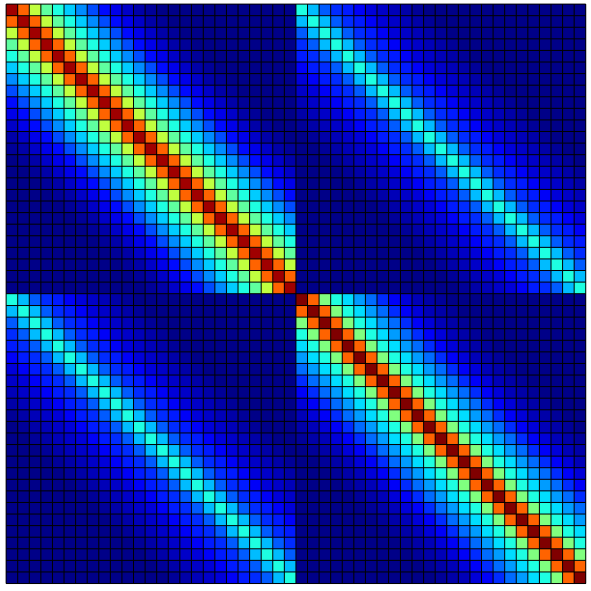
$$\mathbf{K}(\mathbf{X},\mathbf{X}) = \mathbf{B}_1 \otimes k_1(\mathbf{X},\mathbf{X}) + \mathbf{B}_2 \otimes k_2(\mathbf{X},\mathbf{X})$$

$$\mathbf{B}_1 = \begin{bmatrix} 1.4 & 0.5 \\ 0.5 & 1.2 \end{bmatrix}$$

$$\ell_1 = 1$$

$$\mathbf{B}_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.3 \end{bmatrix}$$

$$\ell_2 = 0.2$$



LMC Samples

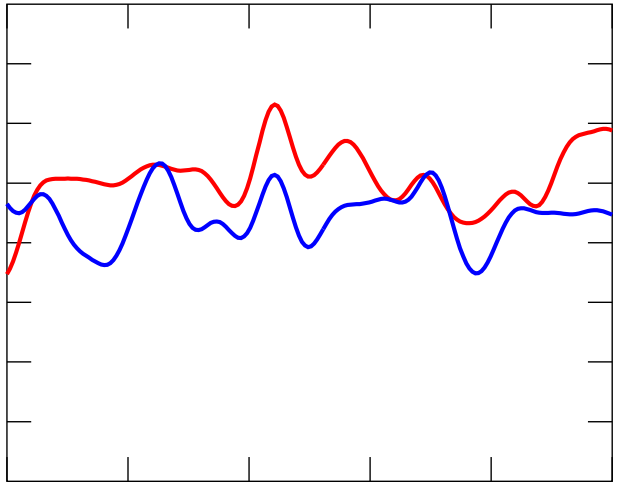
$$\mathbf{K}(\mathbf{X},\mathbf{X}) = \mathbf{B}_1 \otimes k_1(\mathbf{X},\mathbf{X}) + \mathbf{B}_2 \otimes k_2(\mathbf{X},\mathbf{X})$$

$$\mathbf{B}_1 = \begin{bmatrix} 1.4 & 0.5 \\ 0.5 & 1.2 \end{bmatrix}$$

$$\ell_1 = 1$$

$$\mathbf{B}_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.3 \end{bmatrix}$$

$$\ell_2 = 0.2$$



LMC Samples

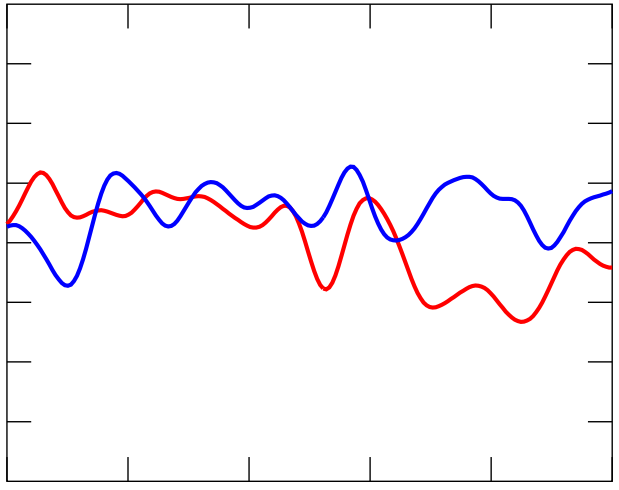
$$\mathbf{K}(\mathbf{X},\mathbf{X}) = \mathbf{B}_1 \otimes k_1(\mathbf{X},\mathbf{X}) + \mathbf{B}_2 \otimes k_2(\mathbf{X},\mathbf{X})$$

$$\mathbf{B}_1 = \begin{bmatrix} 1.4 & 0.5 \\ 0.5 & 1.2 \end{bmatrix}$$

$$\ell_1 = 1$$

$$\mathbf{B}_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.3 \end{bmatrix}$$

$$\ell_2 = 0.2$$



LMC Samples

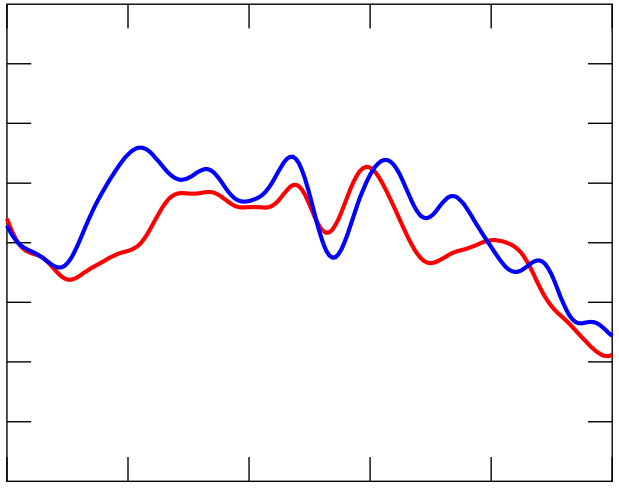
$$\mathbf{K}(\mathbf{X},\mathbf{X}) = \mathbf{B}_1 \otimes k_1(\mathbf{X},\mathbf{X}) + \mathbf{B}_2 \otimes k_2(\mathbf{X},\mathbf{X})$$

$$\mathbf{B}_1 = \begin{bmatrix} 1.4 & 0.5 \\ 0.5 & 1.2 \end{bmatrix}$$

$$\ell_1 = 1$$

$$\mathbf{B}_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.3 \end{bmatrix}$$

$$\ell_2 = 0.2$$



LMC Samples

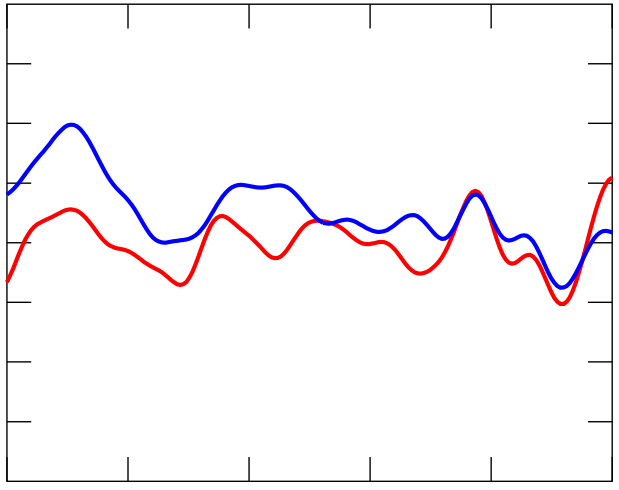
$$\mathbf{K}(\mathbf{X},\mathbf{X}) = \mathbf{B}_1 \otimes k_1(\mathbf{X},\mathbf{X}) + \mathbf{B}_2 \otimes k_2(\mathbf{X},\mathbf{X})$$

$$\mathbf{B}_1 = \begin{bmatrix} 1.4 & 0.5 \\ 0.5 & 1.2 \end{bmatrix}$$

$$\ell_1 = 1$$

$$\mathbf{B}_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.3 \end{bmatrix}$$

$$\ell_2 = 0.2$$



LMC in Machine Learning and Statistics

- Used in machine learning for GPs for multivariate regression and in statistics for computer emulation of expensive multivariate computer codes.
- Imposes the correlation of the outputs explicitly through the set of coregionalization matrices.
- Setting $\mathbf{B} = \mathbf{I}_p$ assumes outputs are conditionally independent given the parameters $\boldsymbol{\theta}$. (Minka and Picard, 1997; Lawrence and Platt, 2004; Yu et al., 2005).
- More recent approaches for multiple output modeling are different versions of the linear model of coregionalization.

Semiparametric Latent Factor Model

- Coregionalization matrices are rank 1 Teh et al. (2005). rewrite equation (??) as

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \sum_{j=1}^{q} \mathbf{w}_{:,j}\mathbf{w}_{:,j}^{\top} \otimes k_j(\mathbf{X}, \mathbf{X}).$$

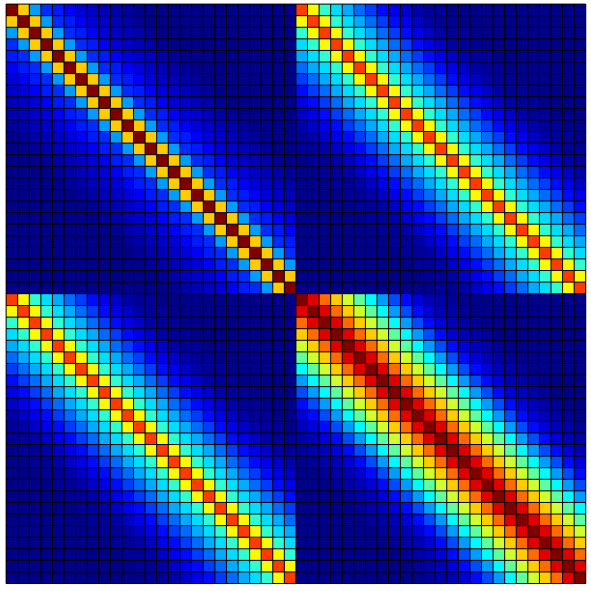
- Like the Kalman filter, but each latent function has a *different* covariance.
- Authors suggest using an exponentiated quadratic characteristic length-scale for each input dimension.

Semiparametric Latent Factor Model Samples

$$\mathbf{K}(\mathbf{X},\mathbf{X}) = \mathbf{w}_{:,1}\mathbf{w}_{:,1}^{\top} \otimes k_1(\mathbf{X},\mathbf{X}) + \mathbf{w}_{:,2}\mathbf{w}_{:,2}^{\top} \otimes k_2(\mathbf{X},\mathbf{X})$$

$$\mathbf{w}_1 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$$

$$\mathbf{w}_2 = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$

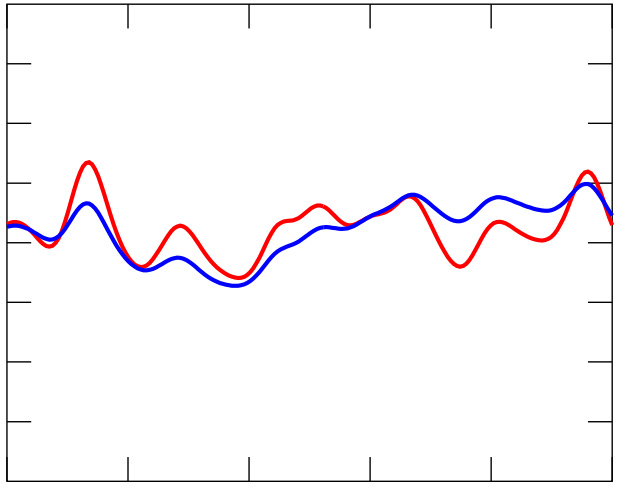


Semiparametric Latent Factor Model Samples

$$\mathbf{K}(\mathbf{X},\mathbf{X}) = \mathbf{w}_{:,1}\mathbf{w}_{:,1}^{\top} \otimes k_1(\mathbf{X},\mathbf{X}) + \mathbf{w}_{:,2}\mathbf{w}_{:,2}^{\top} \otimes k_2(\mathbf{X},\mathbf{X})$$

$$\mathbf{w}_1 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$$

$$\mathbf{w}_2 = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$

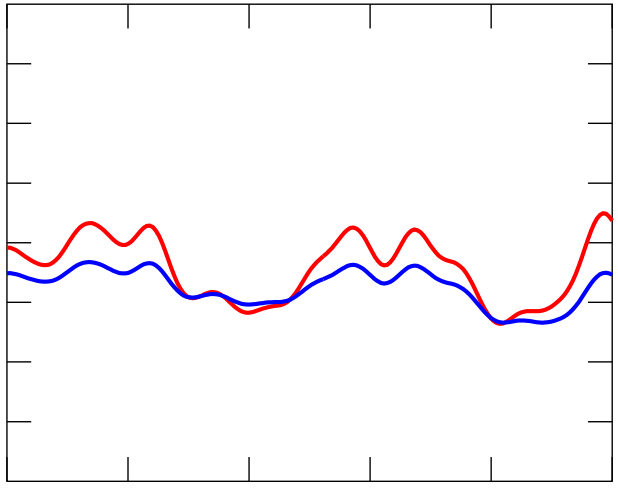


Semiparametric Latent Factor Model Samples

$$\mathbf{K}(\mathbf{X},\mathbf{X}) = \mathbf{w}_{:,1}\mathbf{w}_{:,1}^{\top} \otimes k_1(\mathbf{X},\mathbf{X}) + \mathbf{w}_{:,2}\mathbf{w}_{:,2}^{\top} \otimes k_2(\mathbf{X},\mathbf{X})$$

$$\mathbf{w}_1 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$$

$$\mathbf{w}_2 = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$

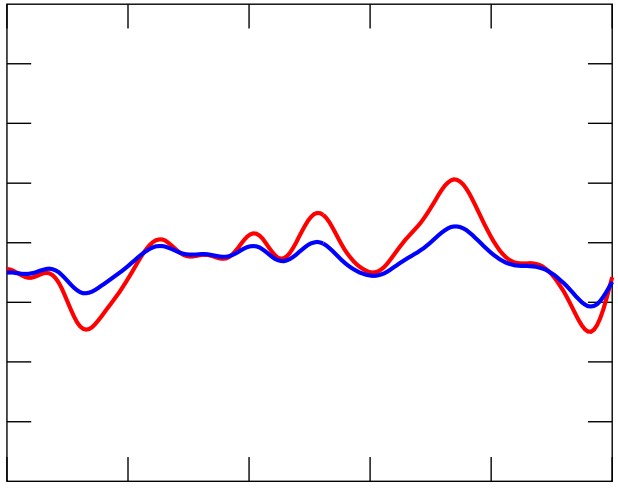


Semiparametric Latent Factor Model Samples

$$\mathbf{K}(\mathbf{X},\mathbf{X}) = \mathbf{w}_{:,1}\mathbf{w}_{:,1}^{\top} \otimes k_1(\mathbf{X},\mathbf{X}) + \mathbf{w}_{:,2}\mathbf{w}_{:,2}^{\top} \otimes k_2(\mathbf{X},\mathbf{X})$$

$$\mathbf{w}_1 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$$

$$\mathbf{w}_2 = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$

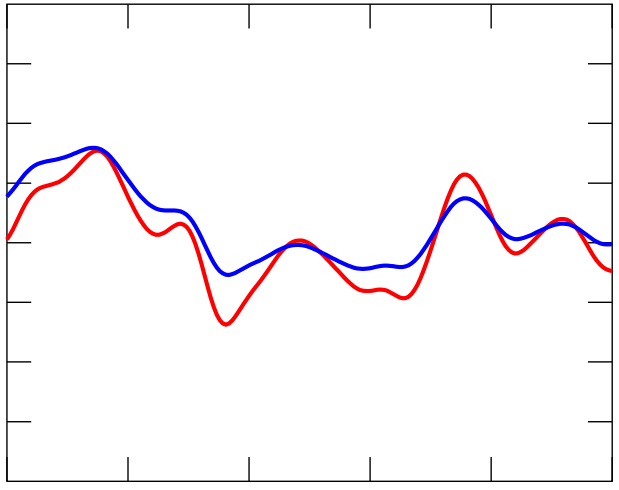


Semiparametric Latent Factor Model Samples

$$\mathbf{K}(\mathbf{X},\mathbf{X}) = \mathbf{w}_{:,1}\mathbf{w}_{:,1}^{\top} \otimes k_1(\mathbf{X},\mathbf{X}) + \mathbf{w}_{:,2}\mathbf{w}_{:,2}^{\top} \otimes k_2(\mathbf{X},\mathbf{X})$$

$$\mathbf{w}_1 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$$

$$\mathbf{w}_2 = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$



Gaussian processes for Multi-task, Multi-output and Multi-class

- Bonilla et al. (2008) suggest ICM for multitask learning.
- Use a PPCA form for $\mathbf{B}$: similar to our Kalman filter example.
- Refer to the autokrigeability effect as the cancellation of inter-task transfer.
- Also discuss the similarities between the multi-task GP and the ICM, and its relationship to the SLFM and the LMC.

Multitask Classification

- Mostly restricted to the case where the outputs are conditionally independent given the hyperparameters ϕ (Minka and Picard, 1997; Williams and Barber, 1998; Lawrence and Platt, 2004; Seeger and Jordan, 2004; Yu et al., 2005; Rasmussen and Williams, 2006).
- Intrinsic coregionalization model has been used in the multiclass scenario. Skolidis and Sanguinetti (2011) use the intrinsic coregionalization model for classification, by introducing a probit noise model as the likelihood.
- Posterior distribution is no longer analytically tractable: approximate inference is required.

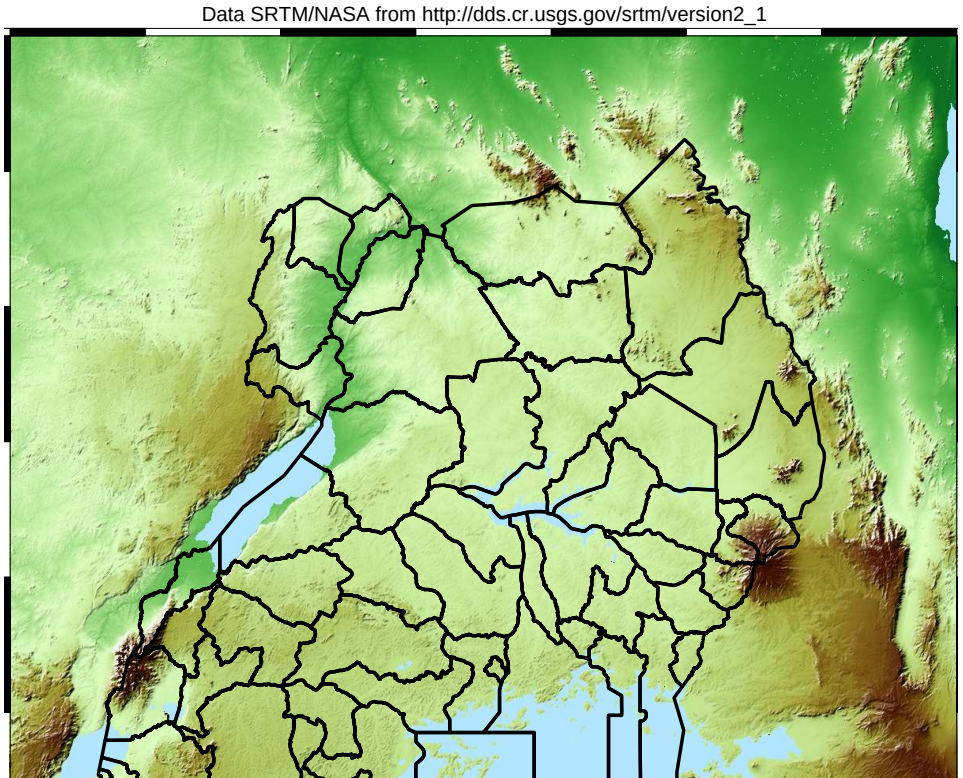
Computer Emulation

- A statistical model used as a surrogate for a computationally expensive computer model.
- Higdon et al. (2008) use the linear model of coregionalization to model images representing the evolution of the implosion of steel cylinders.
- In Conti and O'Hagan (2009) use the ICM to model a vegetation model: called the Sheffield Dynamic Global Vegetation Model (Woodward et al., 1998).

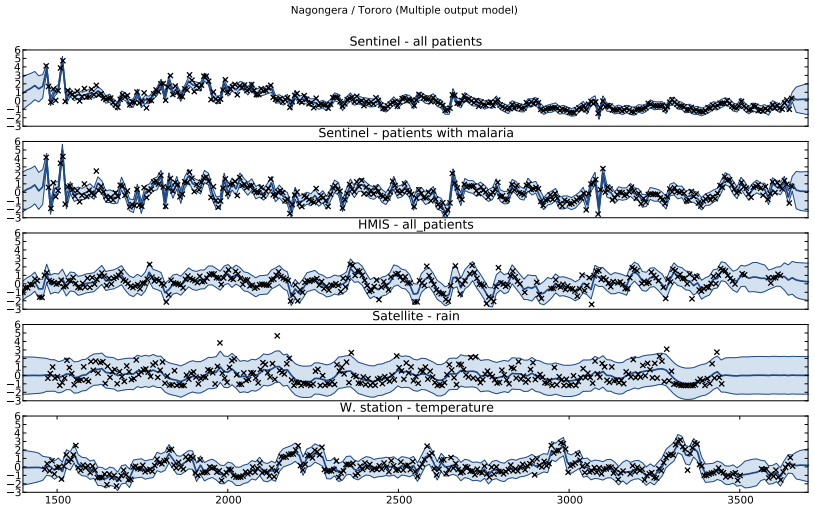
Example: Prediction of Malaria Incidence in Uganda

- Work with John Quinn and Martin Mubaganzi (Makerere University, Uganda)
- See http://cit.mak.ac.ug/cs/aigroup/.

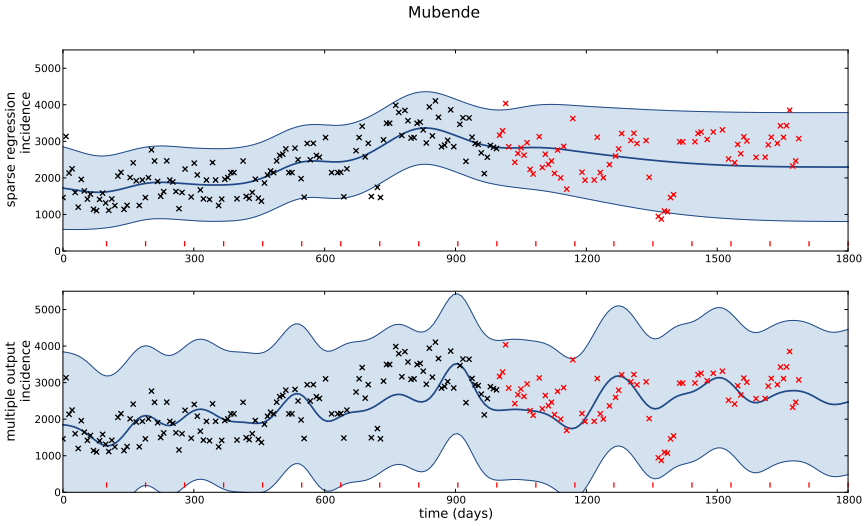
Malaria Prediction in Uganda



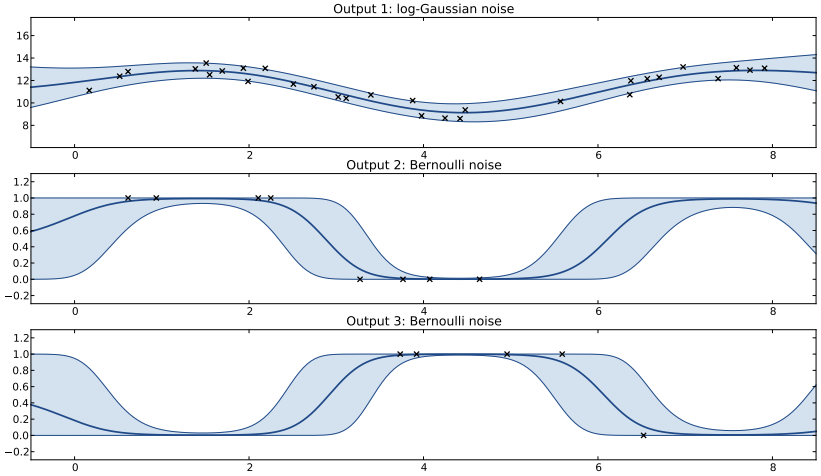
Malaria Prediction in Uganda



Malaria Prediction in Uganda



Mixed Noise Models



Scientists See Promise in Deep-Learning Programs

Hao Zhang/The New York Times

A voice recognition program translated a speech given by Richard F. Rashid, Microsoft's top scientist, into Mandarin Chinese.

By JOHN MARKOFF
Published: November 23, 2012

Using an artificial intelligence technique inspired by theories about how the brain recognizes patterns, technology companies are reporting startling gains in fields as diverse as computer vision, speech recognition and the identification of promising new molecules for designing drugs.

点击查看本文中文版.

The advances have led to widespread enthusiasm among researchers who design software to perform human

THE NEW YORKER

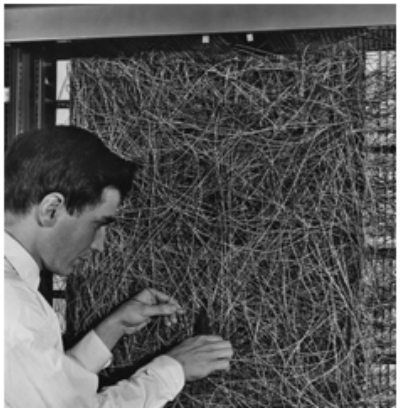
NEWS DESK
Reporting the latest on Washington and the world.

« How Susan Rice Sees the World | Main | Moral Machines »

NOVEMBER 25, 2012

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

POSTED BY GARY MARCUS



Can a new technique known as deep learning revolutionize artificial intelligence, as yesterday's front-page article at the New York *Times* suggests? There is good reason to be excited about deep learning, a sophisticated "machine learning" algorithm that far exceeds many of its predecessors in its abilities to recognize syllables and images. But there's also good reason to be skeptical. While the *Times* reports that "advances in an artificial intelligence technology that can recognize patterns

Google To Expand Knowledge Graph Through Hire Of Geoffrey Hinton

PREV STORY | NEXT STORY

Mar 14, 2013 • 8:23 am | (10)
by Barry Schwartz | Filed Under Google Search Engine

If I had to place one search priority above all else, I'd say right now, Google's most ambitious project is the knowledge graph. Yea, they are pushing Google+ big time, but the knowledge graph is a level above all of that technically.

Of course, Google has an outstanding team working on this project lead by one of the smartest people I've ever met Amit Singhal.

To take the knowledge graph to the next level, Google has hired/acquired Geoffrey Hinton and his team at DNNresearch. Geoffrey posted a note on his Google+ page about it:

> Last summer, I spent several months working with Google's Knowledge team in Mountain View, working with Jeff Dean and an incredible group of scientists and engineers who have a real shot at making spectacular progress in machine learning. Together with two of my recent graduate students, Ilya Sutskever and Alex Krizhevsky (who won the 2012 ImageNet competition), I am betting on Google's team to be the epicenter of future breakthroughs. That means we'll soon be joining Google to work with some of the smartest engineering minds to tackle some of the biggest challenges in computer science. I'll remain part-time at the University of Toronto, where I still have a lot of excellent graduate students, but at Google I will get to see what we can do with very large-scale computation.

I know we just scratched the surface of the knowledge graph and I am excited to see where it takes us in the future.

I am just glad I don't have to figure out how to get us there. I get to just sit and enjoy the ride.

Google Hires Brains that Helped Supercharge Machine Learning

BY ROBERT MCMILLAN 03.13.13 6:30 AM

Follow @bobmcmillan

FOLLOW WIRED ENTERPRISE

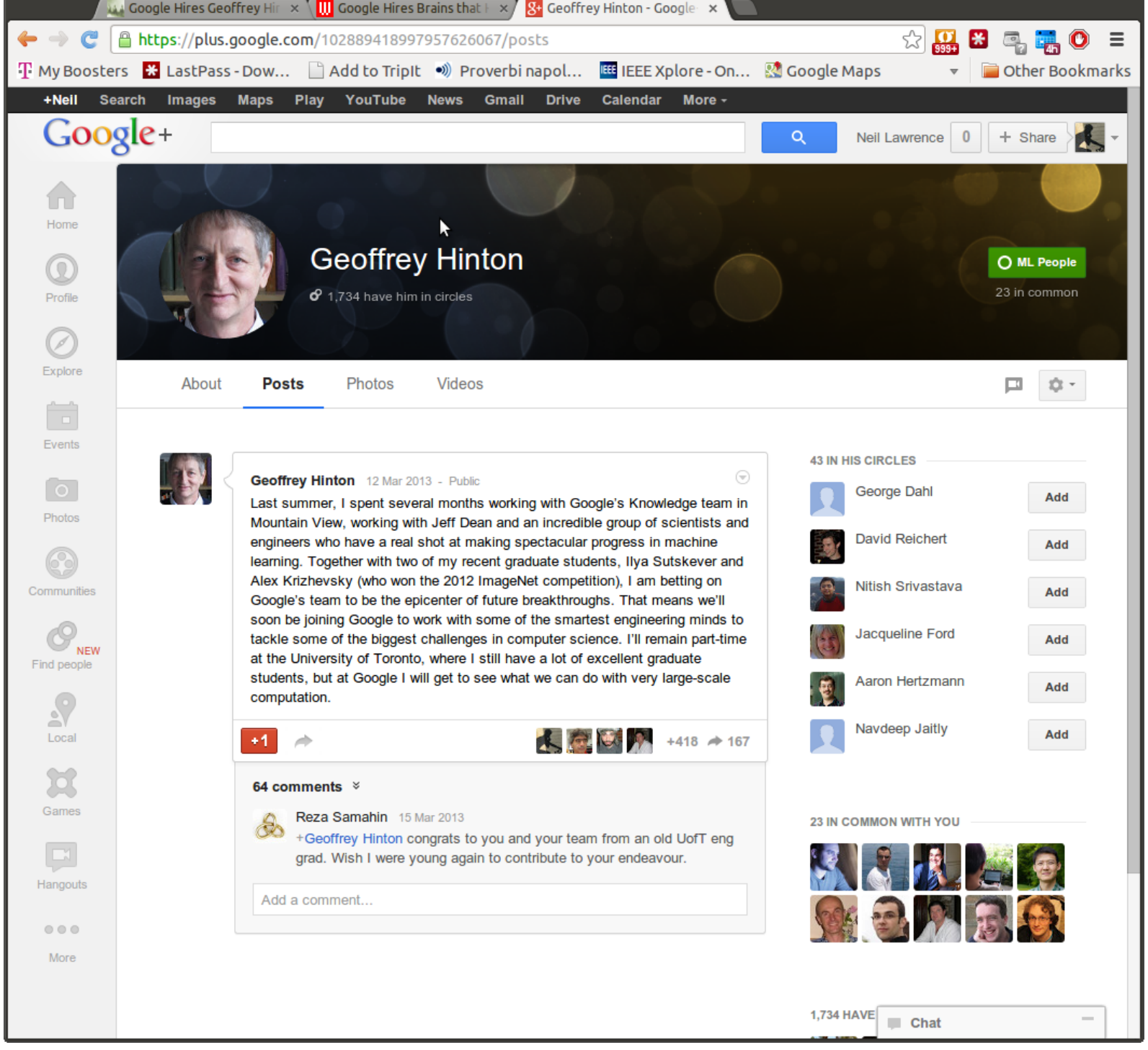
MOST RECENT WIRED POSTS


Jawbone's Up Fitness Band Is Now Android-Compatible


Review: Ecovacs Winbot, a Window-Cleaning Robot


Sherlock, Professor X and Margaery Tyrell Team for Neil Gaiman Radio Play


Video: Robo-Chopper Dives and Grabs Objects Like a Bird of Prey


Google Hires Geoffrey Hir
Google Hires Brains that
Geoffrey Hinton - Google
https://plus.google.com/102889418997957626067/posts
My Boosters
LastPass - Dow...
Add to TripIt
Proverbi napol...
IEEE Xplore - On...
Google Maps
Other Bookmarks
+Neil
Search
Images
Maps
Play
YouTube
News
Gmail
Drive
Calendar
More
Google+
Neil Lawrence
0
+ Share
Home
Profile
Explore
Events
Photos
Communities
NEW
Find people
Local
Games
Hangouts
More
Geoffrey Hinton
1,734 have him in circles
ML People
23 in common
About
Posts
Photos
Videos
Geoffrey Hinton 12 Mar 2013 - Public
Last summer, I spent several months working with Google's Knowledge team in Mountain View, working with Jeff Dean and an incredible group of scientists and engineers who have a real shot at making spectacular progress in machine learning. Together with two of my recent graduate students, Ilya Sutskever and Alex Krizhevsky (who won the 2012 ImageNet competition), I am betting on Google's team to be the epicenter of future breakthroughs. That means we'll soon be joining Google to work with some of the smartest engineering minds to tackle some of the biggest challenges in computer science. I'll remain part-time at the University of Toronto, where I still have a lot of excellent graduate students, but at Google I will get to see what we can do with very large-scale computation.
+1
+418
167
64 comments
Reza Samahin 15 Mar 2013
+Geoffrey Hinton congrats to you and your team from an old UofT eng grad. Wish I were young again to contribute to your endeavour.
Add a comment...
43 IN HIS CIRCLES
George Dahl
David Reichert
Nitish Srivastava
Jacqueline Ford
Aaron Hertzmann
Navdeep Jaitly
Add
23 IN COMMON WITH YOU
1,734 HAVE
Chat

direction for further research.

11.1. HAVE WE THROWN THE BABY OUT WITH THE BATH WATER?

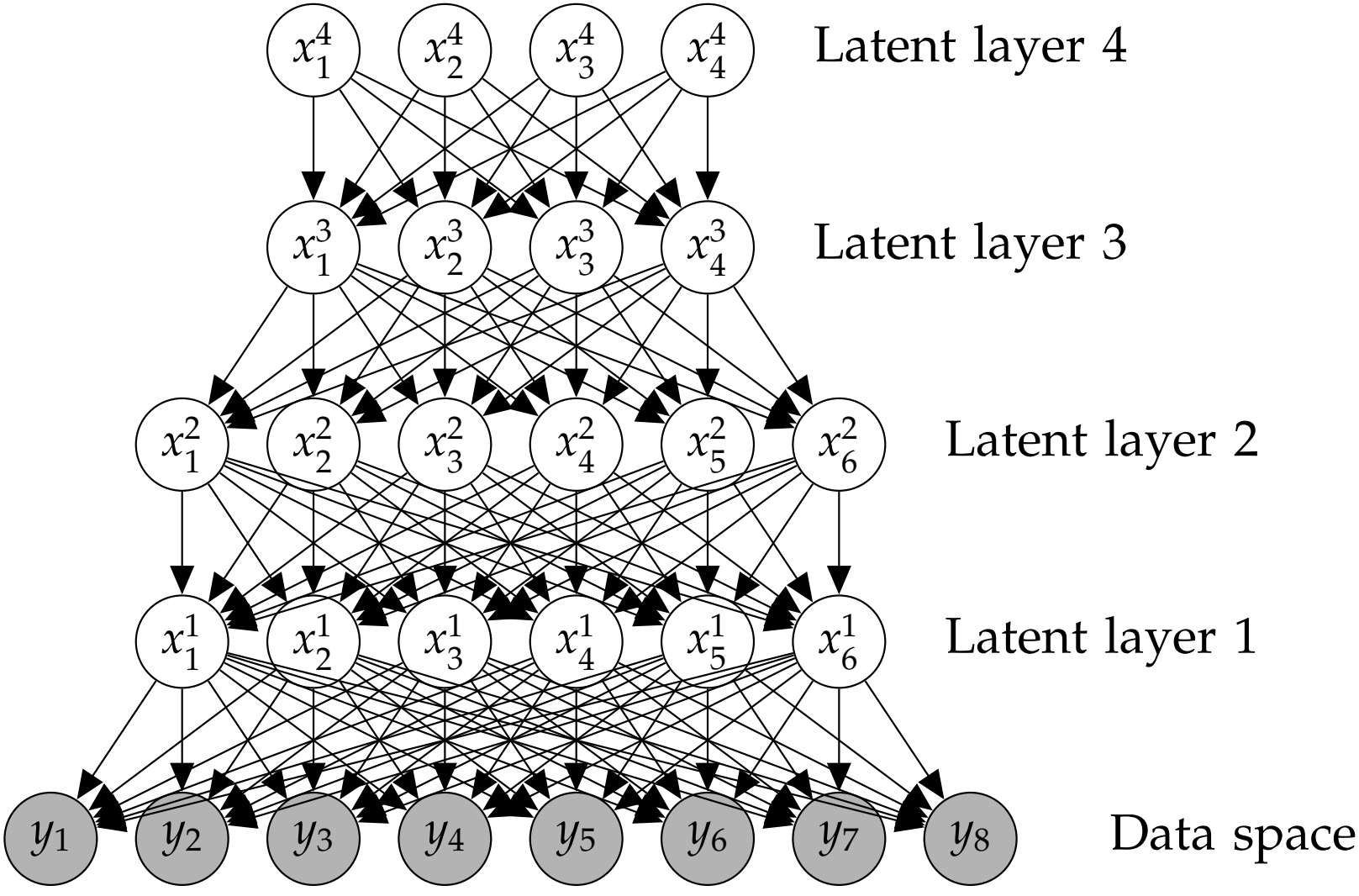
According to the hype of 1987, neural networks were meant to be intelligent models which discovered features and patterns in data. Gaussian processes in contrast are simply smoothing devices. How can Gaussian processes possibly replace neural networks? What is going on?

I think what the work of Williams and Rasmussen (1996) shows is that many real–world data modelling problems are perfectly well solved by sensible smoothing methods. The most interesting problems, the task of feature discovery for example, are not ones which Gaussian processes will solve. But maybe multilayer perceptrons can't solve them either. On the other hand, it may be that the limit of an infinite number of hidden units, to which Gaussian processes correspond, was a bad limit to take; maybe we should backtrack, or modify the prior on neural network parameters, so as to create new models more interesting than Gaussian processes. Evidence that this infinite limit has lost something compared with finite neural networks comes from the observation that in a finite neural network with more than one output, there are non–trivial correlations between the outputs (since they share inputs from common hidden units); but in the limit of an infinite number of hidden units, these correlations vanish. Radford Neal has suggested the use of non–Gaussian priors in networks with multiple hidden layers. Or perhaps a completely fresh start is needed, approaching the problem of machine learning from a paradigm different from the supervised feedforward mapping.

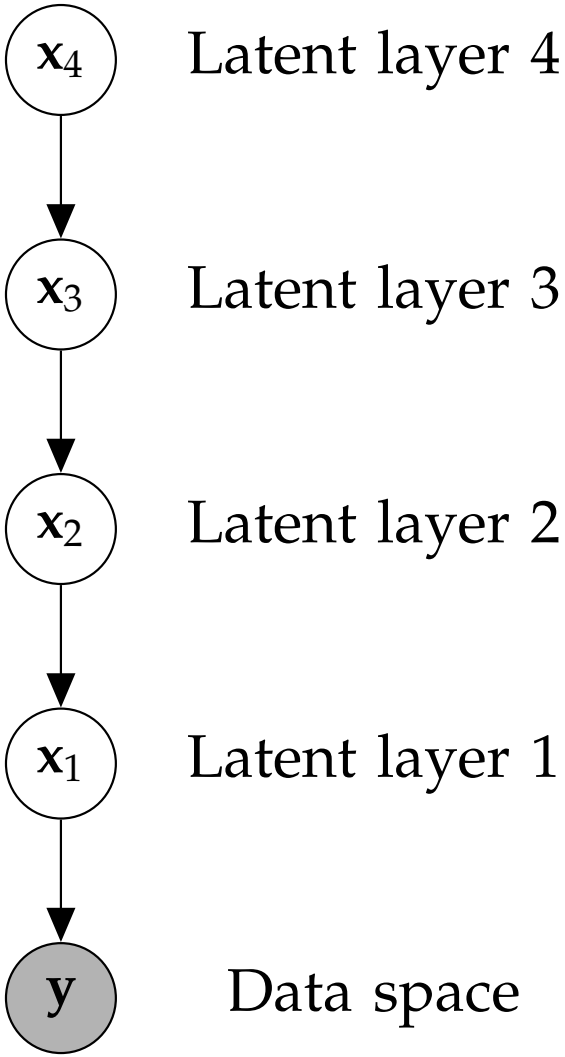
Structure of Priors

MacKay: NIPS Tutorial 1997 “Have we thrown out the baby with the bathwater?” (Published as MacKay, 1998) Also noted by (Wilson et al., 2012)

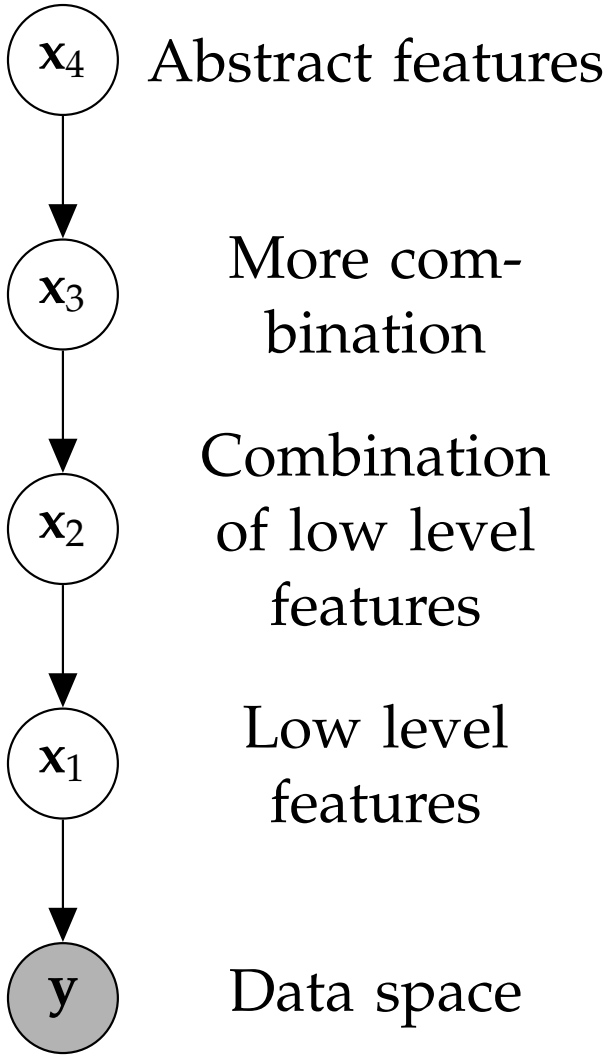
Deep Models


x_1^4 x_2^4 x_3^4 x_4^4
Latent layer 4
x_1^3 x_2^3 x_3^3 x_4^3
Latent layer 3
x_1^2 x_2^2 x_3^2 x_4^2 x_5^2 x_6^2
Latent layer 2
x_1^1 x_2^1 x_3^1 x_4^1 x_5^1 x_6^1
Latent layer 1
y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8
Data space

Deep Models


$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Deep Models

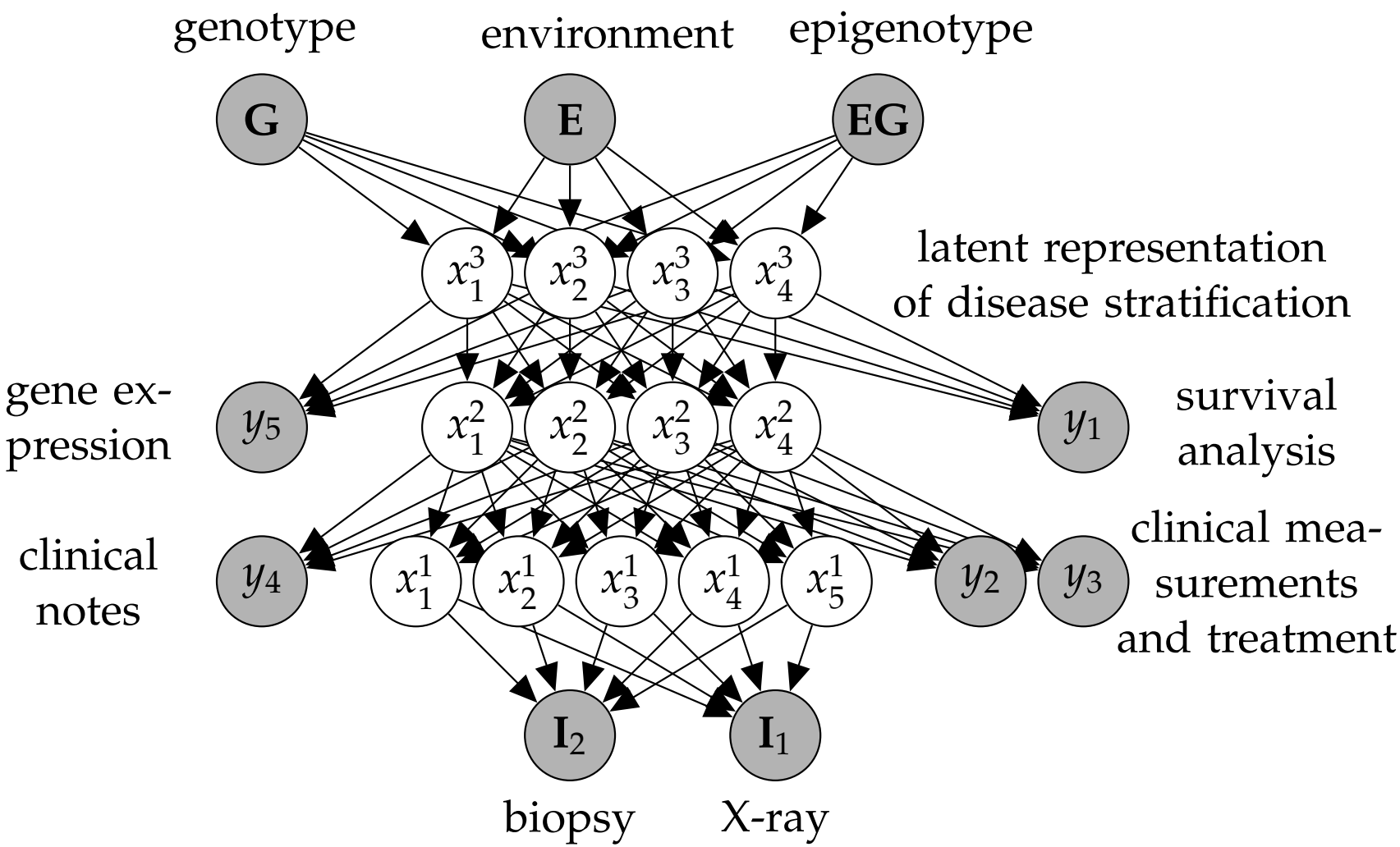

$\mathbf{x}_4$
Abstract features
$\mathbf{x}_3$
More combination
$\mathbf{x}_2$
Combination of low level features
$\mathbf{x}_1$
Low level features
$\mathbf{y}$
Data space

Deep Gaussian Processes

Damianou and Lawrence (2013)

- Deep architectures allow abstraction of features (Bengio, 2009; Hinton and Osindero, 2006; Salakhutdinov and Murray, 2008).
- We use variational approach to stack GP models.

Deep Health

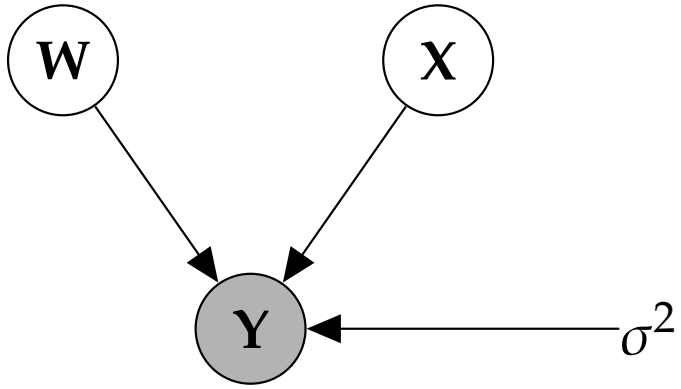

genotype
environment
epigenotype
G
E
EG
x_1^3 x_2^3 x_3^3 x_4^3
latent representation of disease stratification
gene expression
y_5
x_1^2 x_2^2 x_3^2 x_4^2
y_1
survival analysis
clinical notes
y_4
x_1^1 x_2^1 x_3^1 x_4^1 x_5^1
y_2
y_3
clinical measurements and treatment
I_2
I_1
biopsy
X-ray

Deep GPs

- Stacking PPCA still leads to a linear latent variable model.
- To stack latent variable models, need a non-linear model.
- The GP-LVM is a non-linear latent variable model.
- Stacking GP-LVM leads to hierarchical GP-LVM.

Bayesian GP-LVM

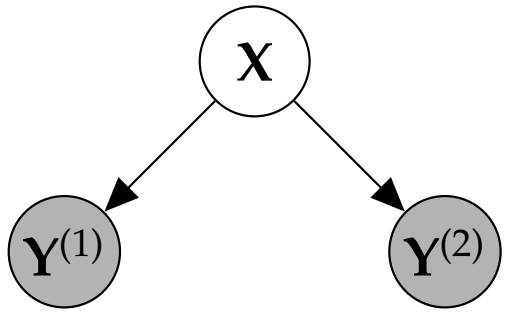
- Bayesian GP-LVM allows variational marginalization of $\mathbf{X}$ and $\mathbf{W}$.



- This leads to a Bayesian model where latent dimensionality can be learnt.

Modeling Multiple 'Views'

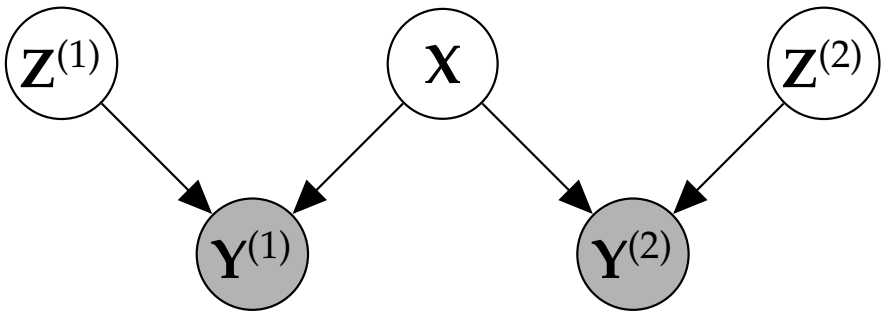
- Single space to model correlations between two different data sources, e.g., images & text, image & pose.
- Shared latent spaces: (Shon et al., 2006; Navaratnam et al., 2007; Ek et al., 2008b)



- Effective when the 'views' are correlated.
- But not all information is shared between both 'views'.
- PCA applied to concatenated data vs CCA applied to data.

Shared-Private Factorization

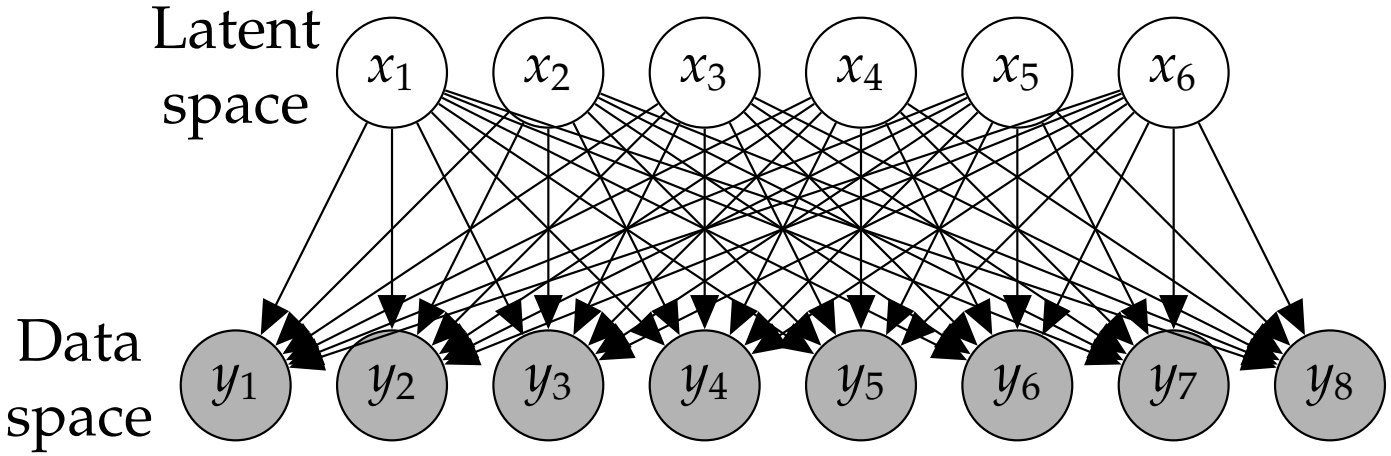
- In real scenarios, the 'views' are neither fully independent, nor fully correlated.
- Shared models
  - either allow information relevant to a single view to be mixed in the shared signal,
  - or are unable to model such private information.
- Solution: Model shared and private information (Virtanen et al., 2011; Ek et al., 2008a; Leen and Fyfe, 2006; Klami and Kaski, 2007, 2008; Tucker, 1958)



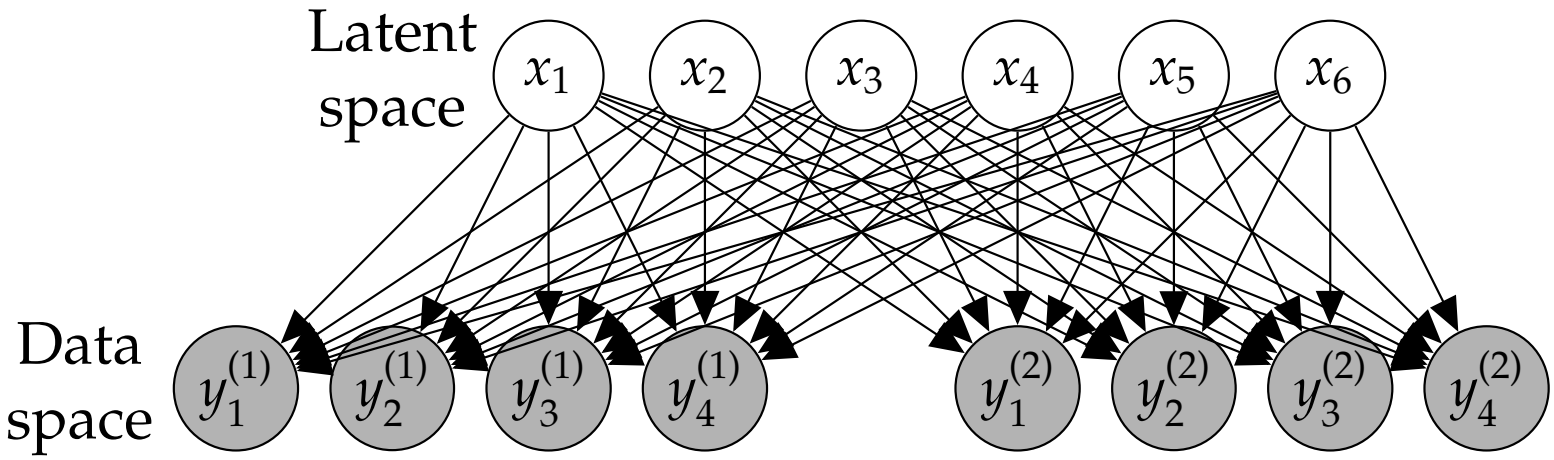
- Probabilistic CCA is case when dimensionality of $\mathbf{Z}$ matches $\mathbf{Y}^{(i)}$ (cf Inter Battery Factor Analysis (Tucker, 1958)).

Manifold Relevance Determination

Damianou et al. (2012)



Shared GP-LVM

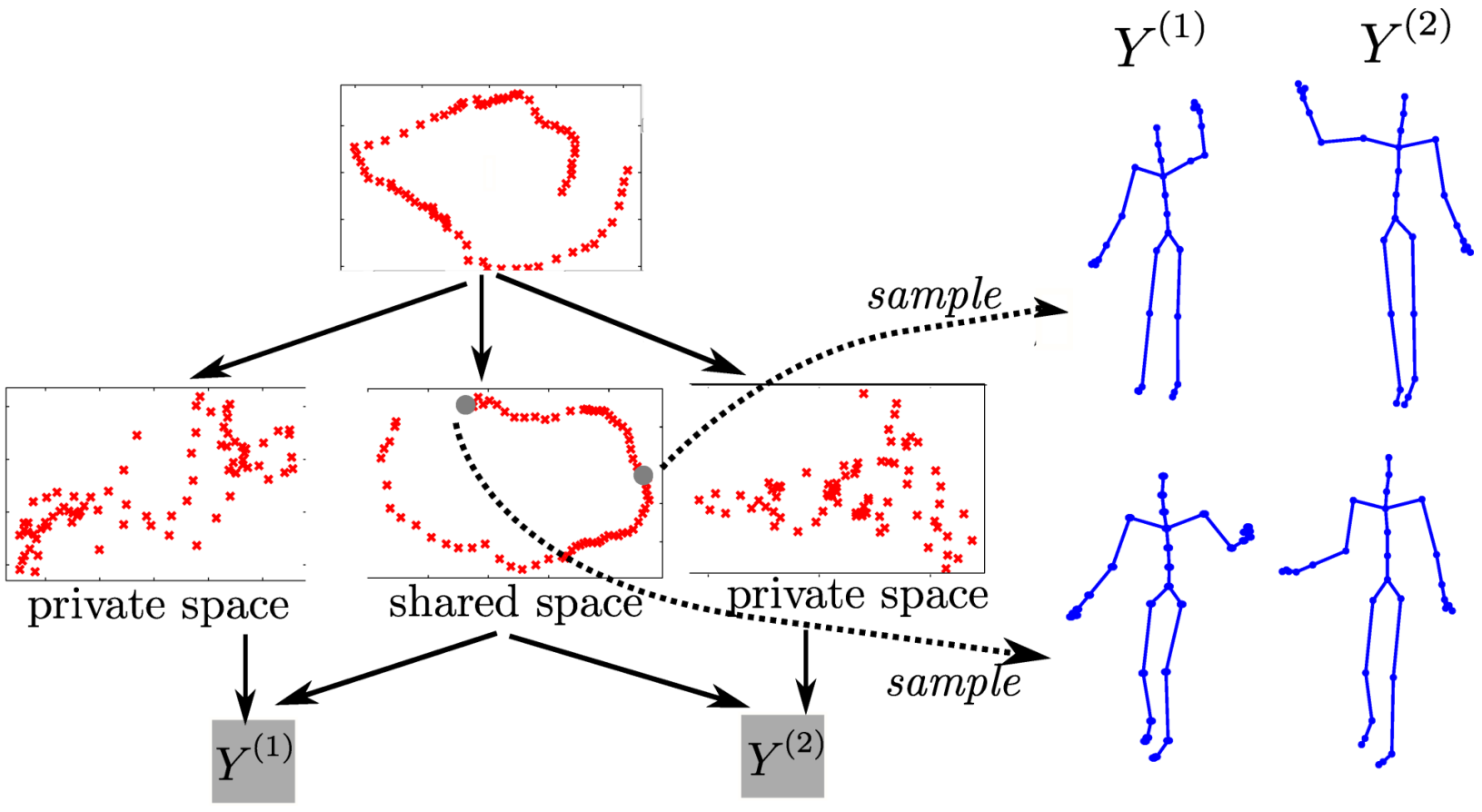


Separate ARD parameters for mappings to $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$.

Motion Capture

- Revisit 'high five' data.
- This time allow model to learn structure, rather than imposing it.

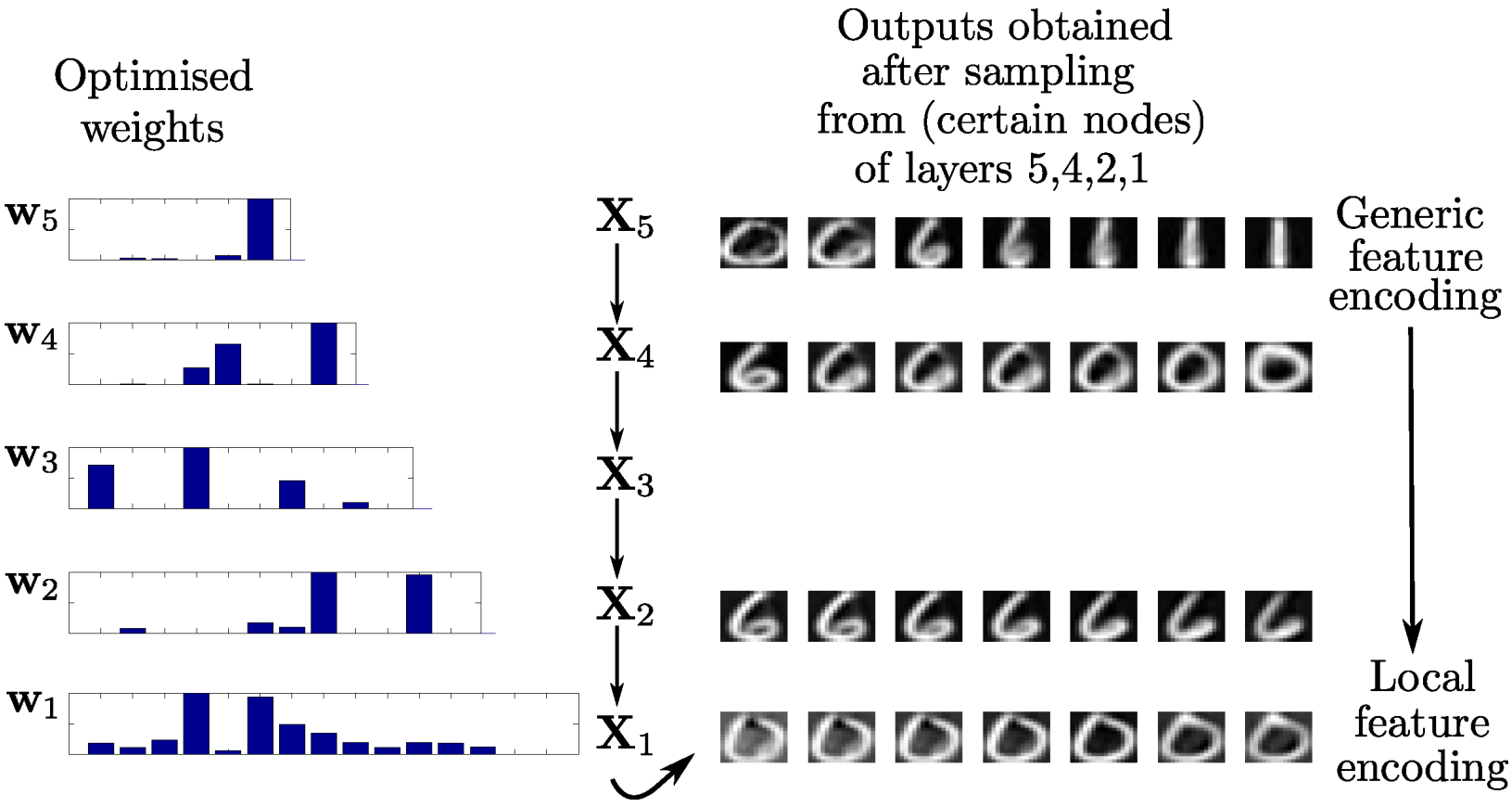
Deep hierarchies – motion capture



Digits Data Set

- Are deep hierarchies justified for small data sets?
- We can lower bound the evidence for different depths.
- For 150 6s, 0s and 1s from MNIST we found at least 5 layers are required.

Deep hierarchies – MNIST



Summary

- Gaussian models good for missing data.
- Disparate data types handled with EP and Laplace.
- Deep models allow complex abstract representation of data sets at higher levels.
- Current limitation is on data set size.
- Addressing this through work by James Hensman on Stochastic Variational Inference for GPs (recent UAI paper).
- Intention is to deploy these models for assimilating a wide range of data types in personalized health (text, survival times, images, genotype, phenotype).
- Requires population scale models with millions of features.

References I

Y. Bengio. Learning Deep Architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, Jan. 2009. ISSN 1935-8237. [DOI].

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006. [Google Books] .

E. V. Bonilla, K. M. Chai, and C. K. I. Williams. Multi-task Gaussian process prediction. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, Cambridge, MA, 2008. MIT Press.

S. Conti and A. O'Hagan. Bayesian emulation of complex multi-output and dynamic computer models. *Journal of Statistical Planning and Inference*, 140(3):640–651, 2009. [DOI].

A. Damianou, C. H. Ek, M. K. Titsias, and N. D. Lawrence. Manifold relevance determination. In Langford and Pineau (2012). [PDF].

A. Damianou and N. D. Lawrence. Deep Gaussian processes. In C. Carvalho and P. Ravikumar, editors, *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics*, volume 31, AZ, USA, 2013. JMLR W&CP 31. [PDF].

G. Della Gatta, M. Bansal, A. Ambesi-Impiombato, D. Antonini, C. Missero, and D. di Bernardo. Direct targets of the trp63 transcription factor revealed by a combination of gene expression profiling and reverse engineering. *Genome Research*, 18(6):939–948, Jun 2008. [URL]. [DOI].

C. H. Ek, J. Rihan, P. Torr, G. Rogez, and N. D. Lawrence. Ambiguity modeling in latent spaces. In A. Popescu-Belis and R. Stiefelhagen, editors, *Machine Learning for Multimodal Interaction (MLMI 2008)*, LNCS, pages 62–73. Springer-Verlag, 28–30 June 2008a. [PDF].

C. H. Ek, P. H. Torr, and N. D. Lawrence. Gaussian process latent variable models for human pose estimation. In A. Popescu-Belis, S. Renals, and H. Bourlard, editors, *Machine Learning for Multimodal Interaction (MLMI 2007)*, volume 4892 of *LNCS*, pages 132–143, Brno, Czech Republic, 2008b. Springer-Verlag. [PDF].

N. Fusi, C. Lippert, K. Borgwardt, N. D. Lawrence, and O. Stegle. Detecting regulatory gene-environment interactions with unmeasured environmental factors. *Bioinformatics*, 2013. [DOI].

N. Fusi, O. Stegle, and N. D. Lawrence. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Computat Biol*, 8:e1002330, 2012. [URL]. [PDF]. [DOI].

References II

P. Goovaerts. *Geostatistics For Natural Resources Evaluation*. Oxford University Press, 1997. [Google Books] .

J. D. Helterbrand and N. A. C. Cressie. Universal cokriging under intrinsic coregionalization. *Mathematical Geology*, 26(2):205–226, 1994.

D. M. Higdon, J. Gattiker, B. Williams, and M. Rightley. Computer model calibration using high dimensional output. *Journal of the American Statistical Association*, 103(482):570–583, 2008.

G. E. Hinton and S. Osindero. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:2006, 2006.

A. G. Journel and C. J. Huijbregts. *Mining Geostatistics*. Academic Press, London, 1978. [Google Books] .

A. A. Kalaitzis and N. D. Lawrence. A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinformatics*, 12(180), 2011. [DOI].

A. Klami and S. Kaski. Local dependent components analysis. In Z. Ghahramani, editor, *Proceedings of the International Conference in Machine Learning*, volume 24. Omnipress, 2007. [Google Books] .

A. Klami and S. Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72: 39–46, 2008.

J. Langford and J. Pineau, editors. *Proceedings of the International Conference in Machine Learning*, volume 29, San Francisco, CA, 2012. Morgan Kauffman.

P. S. Laplace. Mémoire sur la probabilité des causes par les évènemens. In *Mémoires de mathèmatique et de physique, presentés à lAcadémie Royale des Sciences, par divers savans, & lù dans ses assemblées 6*, pages 621–656, 1774. Translated in Stigler (1986).

N. D. Lawrence. Gaussian process models for visualisation of high dimensional data. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16, pages 329–336, Cambridge, MA, 2004. MIT Press.

N. D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 11 2005.

N. D. Lawrence and J. C. Platt. Learning to learn with the informative vector machine. In R. Greiner and D. Schuurmans, editors, *Proceedings of the International Conference in Machine Learning*, volume 21, pages 512–519. Omnipress, 2004. [PDF].

G. Leen and C. Fyfe. A Gaussian process latent variable model formulation of canonical correlation analysis. Bruges (Belgium), 26-28 April 2006 2006.

References III

D. J. C. MacKay. Introduction to Gaussian Processes. In C. M. Bishop, editor, *Neural Networks and Machine Learning*, volume 168 of *Series F: Computer and Systems Sciences*, pages 133–166. Springer-Verlag, Berlin, 1998.

T. P. Minka and R. W. Picard. Learning how to learn is learning with point sets. Available on-line., 1997. [URL]. Revised 1999, available at `http://www.stat.cmu.edu/~{}minka/`.

R. Navaratnam, A. Fitzgibbon, and R. Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society Press, 2007.

J. Oakley and A. O'Hagan. Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, 89(4):769–784, 2002.

L. Parts, O. Stegle, J. Winn, and R. Durbin. Joint genetic analysis of gene expression data with inferred cellular phenotypes. *PLoS Genet*, 7(1):e1001276, 2011. [URL]. [DOI].

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. [Google Books] .

S. Rogers and M. Girolami. *A First Course in Machine Learning*. CRC Press, 2011. [Google Books] .

R. Salakhutdinov and I. Murray. On the quantitative analysis of deep belief networks. In S. Roweis and A. McCallum, editors, *Proceedings of the International Conference in Machine Learning*, volume 25, pages 872–879. Omnipress, 2008.

M. Seeger and M. I. Jordan. Sparse Gaussian Process Classification With Multiple Classes. Technical Report 661, Department of Statistics, University of California at Berkeley,

A. P. Shon, K. Grochow, A. Hertzmann, and R. P. N. Rao. Learning shared latent structure for image synthesis and robotic imitation. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA, 2006. MIT Press.

G. Skolidis and G. Sanguinetti. Bayesian multitask classification with Gaussian process priors. *IEEE Transactions on Neural Networks*, 22(12):2011 – 2021, 2011.

S. M. Stigler. Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1:359–378, 1986.

Y. W. Teh, M. Seeger, and M. I. Jordan. Semiparametric latent factor models. In R. G. Cowell and Z. Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 333–340, Barbados, 6-8 January 2005. Society for Artificial Intelligence and Statistics.

References IV

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 6 (3):611–622, 1999. [PDF]. [DOI].

L. R. Tucker. An inter-battery method of factor analysis. *Psychometrika*, 23(2):111–136, 1958.

S. Virtanen, A. Klami, and S. Kaski. Bayesian CCA via group sparsity. In L. Getoor and T. Scheffer, editors, *Proceedings of the International Conference in Machine Learning*, volume 28, 2011.

H. Wackernagel. *Multivariate Geostatistics: An Introduction With Applications*. Springer-Verlag, 3rd edition, 2003. [Google Books] .

C. K. Williams and D. Barber. Bayesian Classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.

A. G. Wilson, D. A. Knowles, and Z. Ghahramani. Gaussian process regression networks. In Langford and Pineau (2012).

I. Woodward, M. R. Lomas, and R. A. Betts. Vegetation-climate feedbacks in a greenhouse world. *Philosophical Transactions: Biological Sciences*, 353(1365):29–39, 1998.

K. Yu, V. Tresp, and A. Schwaighofer. Learning Gaussian processes from multiple tasks. In *Proceedings of the 22nd International Conference on Machine Learning (ICML 2005)*, pages 1012–1019, 2005.