Regression and Probability

Neil D. Lawrence

GPRS
19th–22nd January 2015

Outline

Regression

Bayesian Perspective

Gaussian Processes

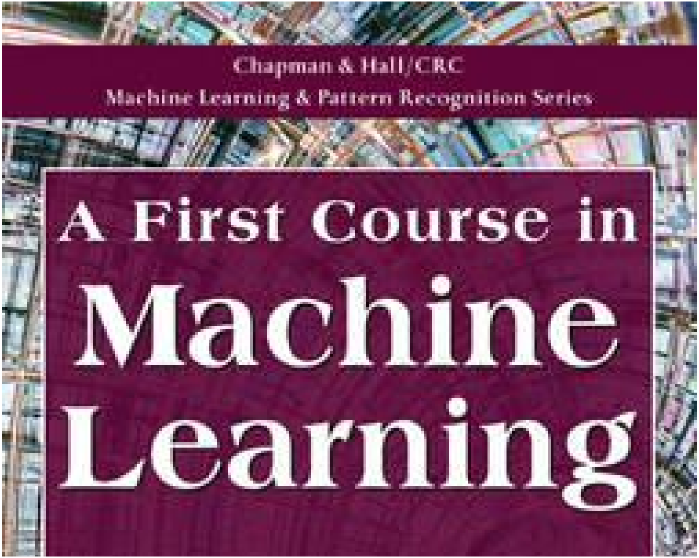
Multiple Output Processes

Latent Force Models

Approximations

Dimensionality Reduction

Rogers and Girolami



Bishop


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP

e

Regression Examples

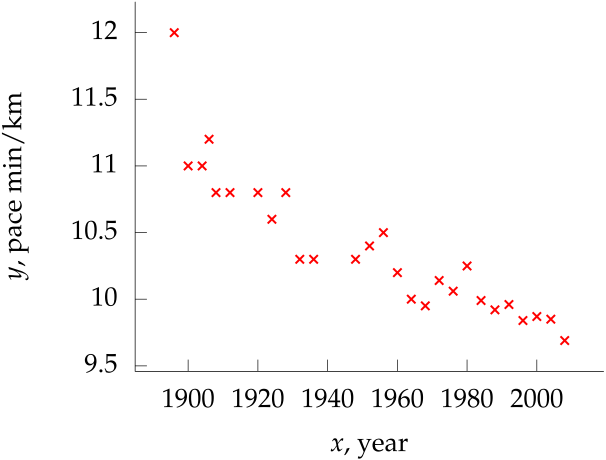
- Predict a real value, y_i given some inputs $\mathbf{x}_i$.
- Predict quality of meat given spectral measurements (Tecator data).
- Radiocarbon dating, the C14 calibration curve: predict age given quantity of C14 isotope.
- Predict quality of different Go or Backgammon moves given expert rated training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

Olympic 100m Data

- Gold medal times for Olympic 100 m runners since 1896.

Image from Wikimedia Commons
`http://bit.ly/191adDC`

Olympic 100m Data



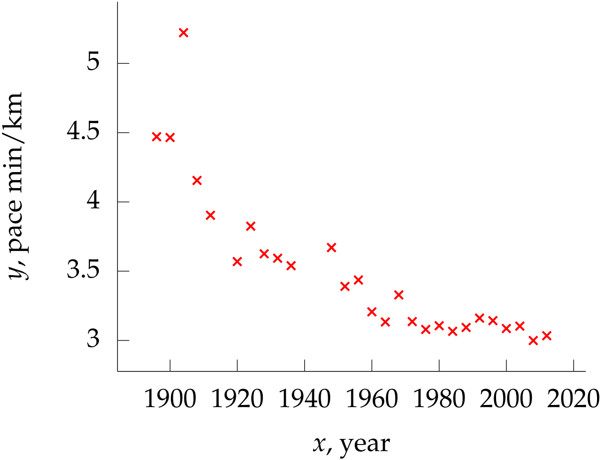
Olympic 100 m Data.

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

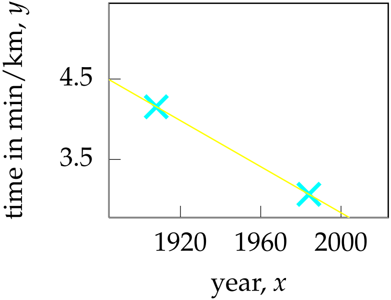
$$y = mx + c$$

- 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.

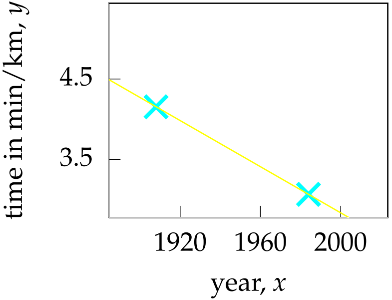
$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$



Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

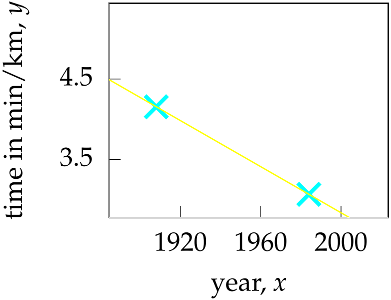
$$y_1 - y_2 = m(x_1 - x_2)$$



Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

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

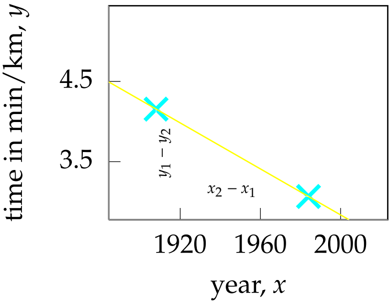


Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

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

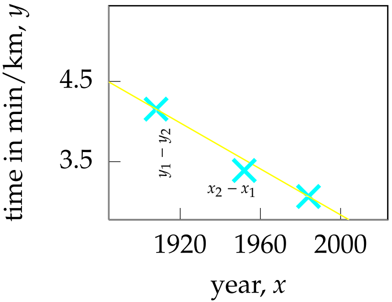
$$c = y_1 - mx_1$$



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:

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

- Additional observation leads to *overdetermined* system.

$$y_3 = mx_3 + c$$

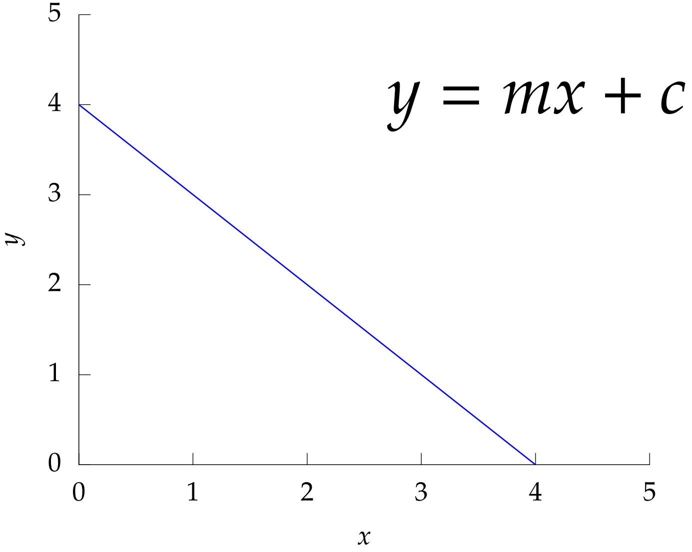
- This problem is solved through a noise model $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

$$y_1 = mx_1 + c + \epsilon_1$$
$$y_2 = mx_2 + c + \epsilon_2$$
$$y_3 = mx_3 + c + \epsilon_3$$

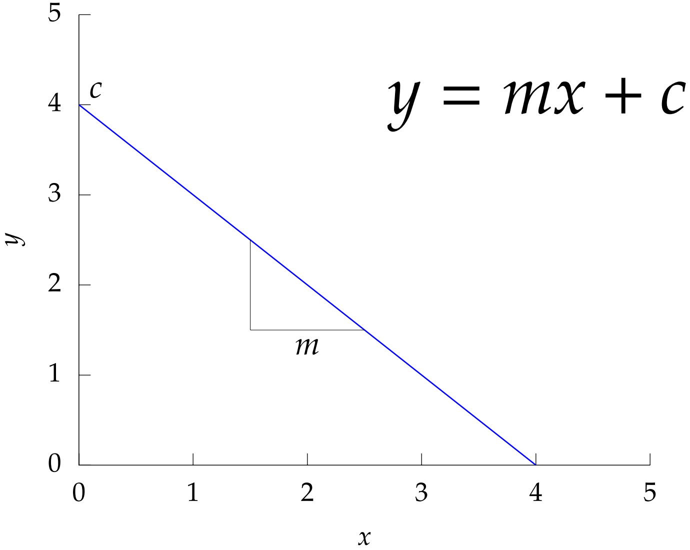
Noise Models

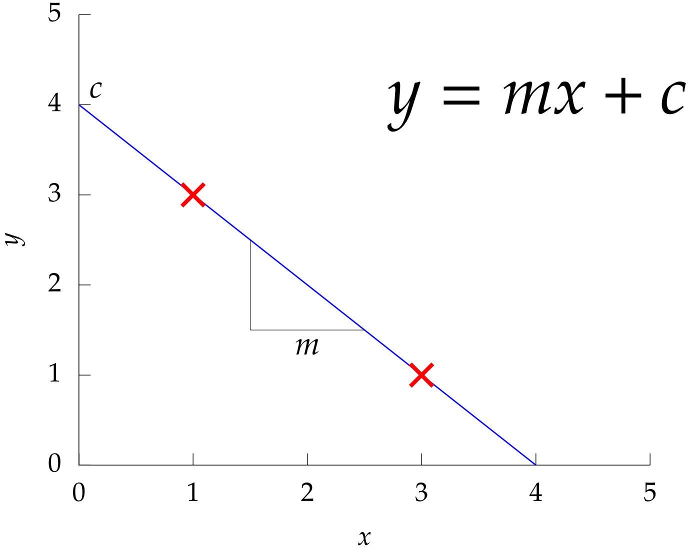
- We aren't modeling entire system.
- Noise model gives mismatch between model and data.
- Gaussian model justified by appeal to central limit theorem.
- Other models also possible (Student-t for heavy tails).
- Maximum likelihood with Gaussian noise leads to *least squares*.

$$y = mx + c$$

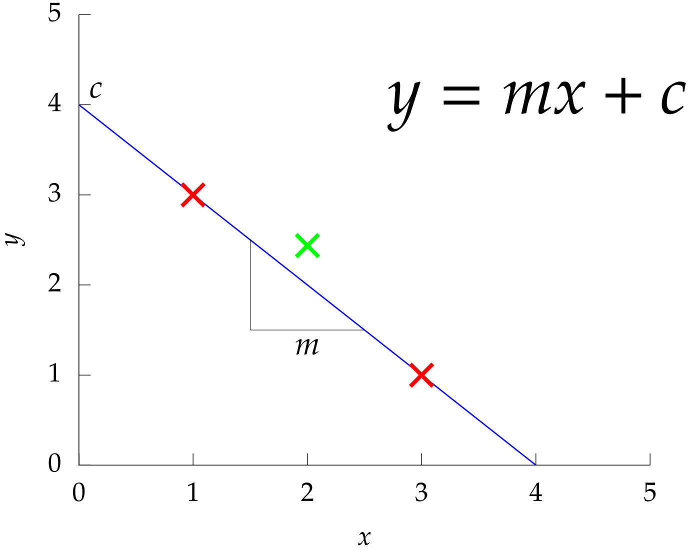


$y = mx + c$
y
x
0
1
2
3
4
5
0
1
2
3
4
5

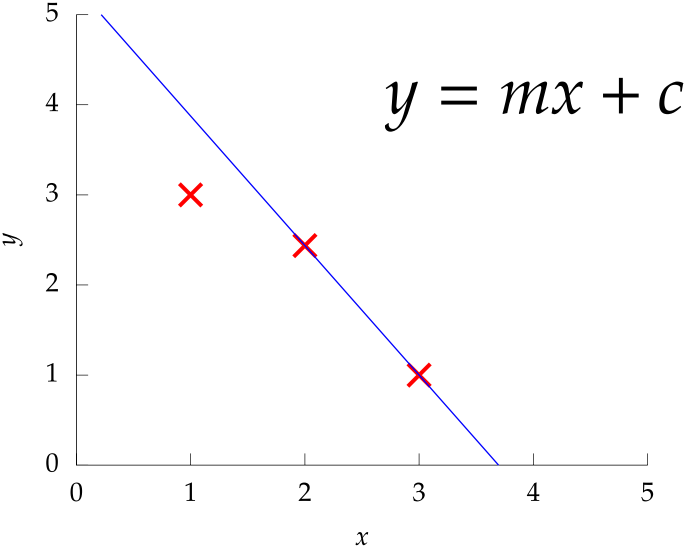

$y = mx + c$
c
m
y
x
0
1
2
3
4
5



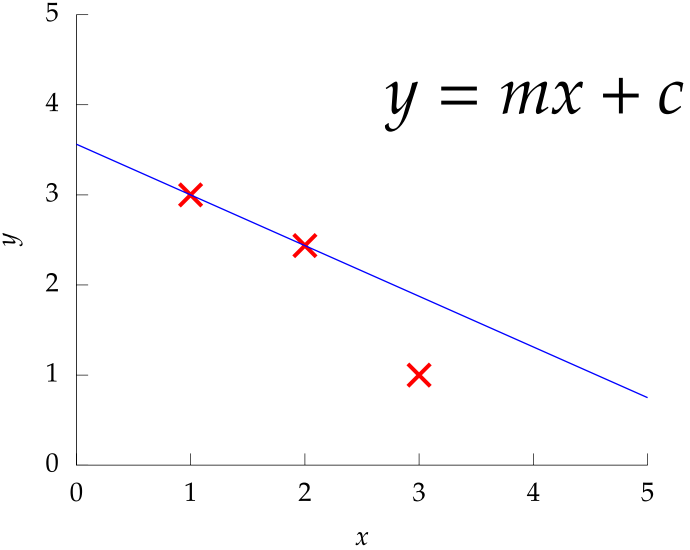
$y = mx + c$
c
m
y
x
0
1
2
3
4
5



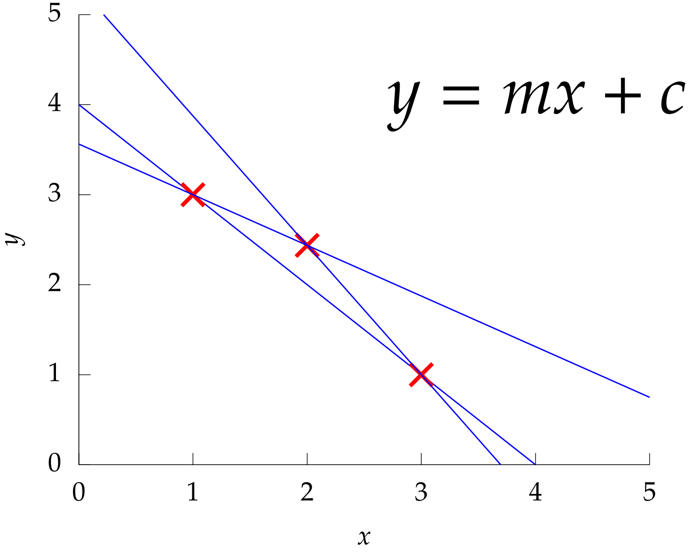
$y = mx + c$
c
m
y
x
0
1
2
3
4
5
0
1
2
3
4
5



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



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



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

$y = mx + c$

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

$$3 = m + c$$

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

$$1 = 3m + c$$

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

$$2.5 = 2m + c$$

riens. L'opinion contraire est une illusion de l'esprit qui, perdant de vue les raisons fugitives du choix de la volonté dans les choses indifférentes, se persuade qu'elle s'est déterminée d'elle-même et sans motifs.

Nous devons donc envisager l'état présent de l'univers, comme l'effet de son état antérieur, et comme la cause de celui qui va suivre. Une intelligence qui, pour un instant donné, connaîtrait toutes les forces dont la nature est animée, et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule les mouvemens des plus grands corps de l'univers et ceux du plus léger atome : rien ne serait incertain pour elle, et l'avenir comme le passé, serait présent à ses yeux. L'esprit humain offre, dans la perfection qu'il a su donner à l'Astronomie, une faible esquisse de cette intelligence. Ses découvertes en Mécanique et en Géométrie, jointes à celle de la pesanteur universelle, l'ont mis à portée de comprendre dans les mêmes expressions analytiques, les états passés et futurs du système du monde. En appliquant la même méthode à quelques autres objets de ses connaissances, il est parvenu à ramener à des lois générales, les phénomènes observés, et à prévoir ceux que des circonstances données doivent faire éclore. Tous ces efforts dans la recherche de la vérité, tendent à le rapprocher sans cesse de l'intelligence que nous venons de concevoir, mais dont il restera toujours infiniment éloigné. Cette tendance propre à l'espèce humaine, est ce qui la rend supérieure aux animaux; et ses progrès en ce genre, distinguent les nations et les siècles, et font leur véritable gloire.

Rappelons-nous qu'autrefois, et à une époque qui other, we say that its choice is an effect without a cause. It is then, says Leibnitz, the blind chance of the Epicureans. The contrary opinion is an illusion of the mind, which, losing sight of the evasive reasons of the choice of the will in indifferent things, believes that choice is determined of itself and without motives.

We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it—an intelligence sufficiently vast to submit these data to analysis—it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes. The human mind offers, in the perfection which it has been able to give to astronomy, a feeble idea of this intelligence. Its discoveries in mechanics and geometry, added to that of universal gravity, have enabled it to comprehend in the same analytical expressions the past and future states of the system of the world. Applying the same method to some other objects of its knowledge, it has succeeded in referring to general laws observed phenomena and in foreseeing those which given circumstances ought to produce. All these efforts in the search for truth tend to lead it back continually to the vast intelligence which we have just mentioned, but from which it will always remain infinitely removed. This tendency, peculiar to the human race, is that which renders it superior to animals; and their progress height: "The day will come when, by study pursued through several ages, the things now concealed will appear with evidence; and posterity will be astonished that truths so clear had escaped us." Clairaut then undertook to submit to analysis the perturbations which the comet had experienced by the action of the two great planets, Jupiter and Saturn; after immense calculations he fixed its next passage at the perihelion toward the beginning of April, 1759, which was actually verified by observation. The regularity which astronomy shows us in the movements of the comets doubtless exists also in all phenomena.

The curve described by a simple molecule of air or vapor is regulated in a manner just as certain as the planetary orbits; the only difference between them is that which comes from our ignorance.

Probability is relative, in part to this ignorance, in part to our knowledge. We know that of three or a greater number of events a single one ought to occur; but nothing induces us to believe that one of them will occur rather than the others. In this state of indecision it is impossible for us to announce their occurrence with certainty. It is, however, probable that one of these events, chosen at will, will not occur because we see several cases equally possible which exclude its occurrence, while only a single one favors it.

The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of point 1: $x = 1$, $y = 3$

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

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

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

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

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

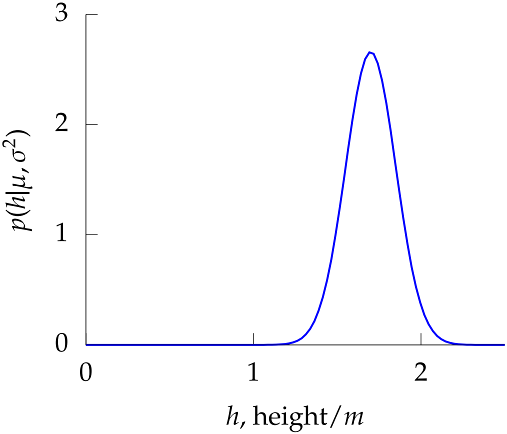
The Gaussian Density

- Perhaps the most common probability density.

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

- 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.

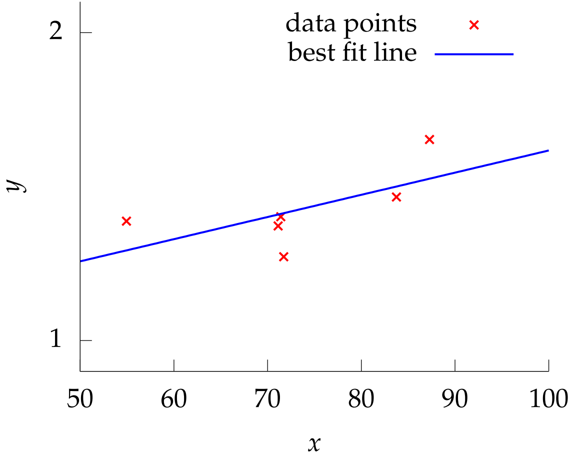
Height as a Function of Weight

- 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.

Mathematical Interpretation

- What is the mathematical interpretation?
  - There is a cost function.
  - It expresses mismatch between your prediction and reality.

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

  - This is known as the sum of squares error.

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

$$\frac{\mathrm{d}E(m)}{\mathrm{d}m} = -2\sum_{i=1}^{n} x_i \left(y_i - mx_i - c\right)$$

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

$$0 = -2\sum_{i=1}^{n} x_i \left(y_i - m x_i - c\right)$$

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

$$0 = -2\sum_{i=1}^{n} x_i y_i + 2\sum_{i=1}^{n} m x_i^2 + 2\sum_{i=1}^{n} c x_i$$

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

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

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

$$\frac{\mathrm{d}E(c)}{\mathrm{d}c} = -2\sum_{i=1}^{n}(y_i - mx_i - c)$$

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

$$0 = -2 \sum_{i=1}^{n} (y_i - mx_i - c)$$

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

$$0 = -2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n} mx_i + 2nc$$

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

$$c = \frac{\sum_{i=1}^{n} (y_i - mx_i)}{n}$$

Fixed Point Updates

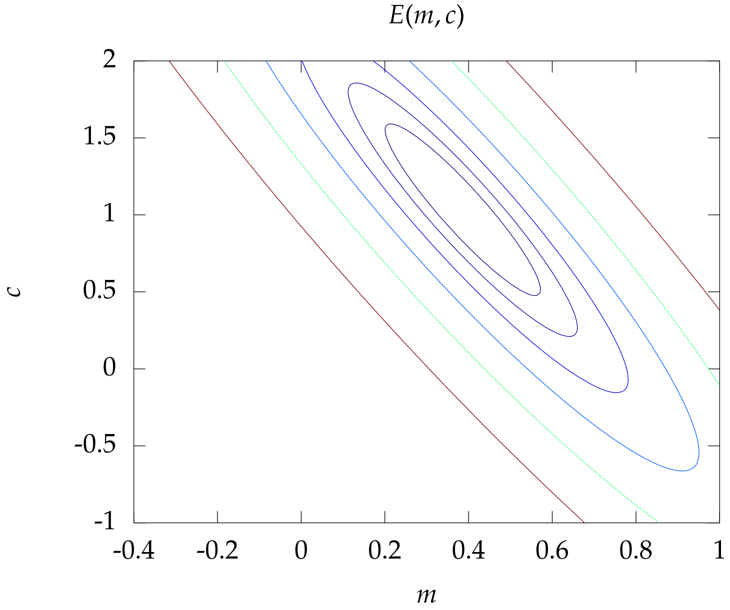
Worked example.

$$c^* = \frac{\sum_{i=1}^{n} (y_i - m^* x_i)}{n},$$

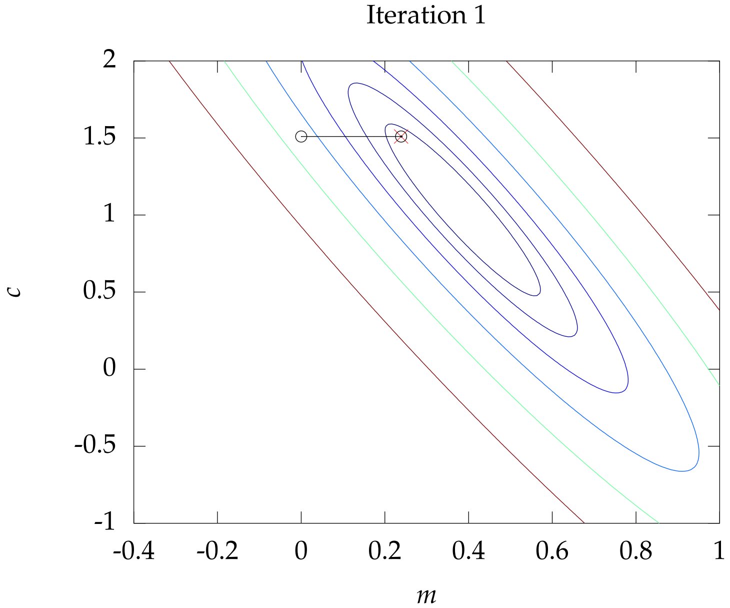
$$m^* = \frac{\sum_{i=1}^{n} x_i (y_i - c^*)}{\sum_{i=1}^{n} x_i^2},$$

$$\sigma^{2*} = \frac{\sum_{i=1}^{n} (y_i - m^* x_i - c^*)^2}{n}$$

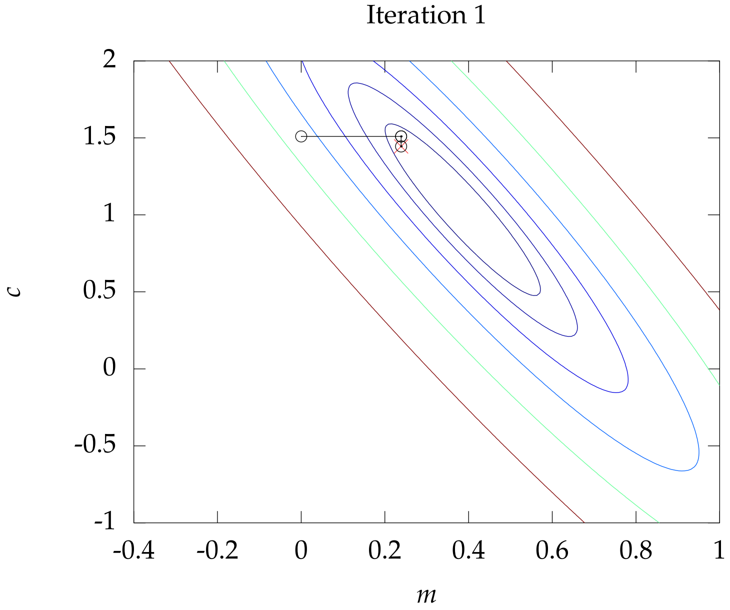
Coordinate Descent



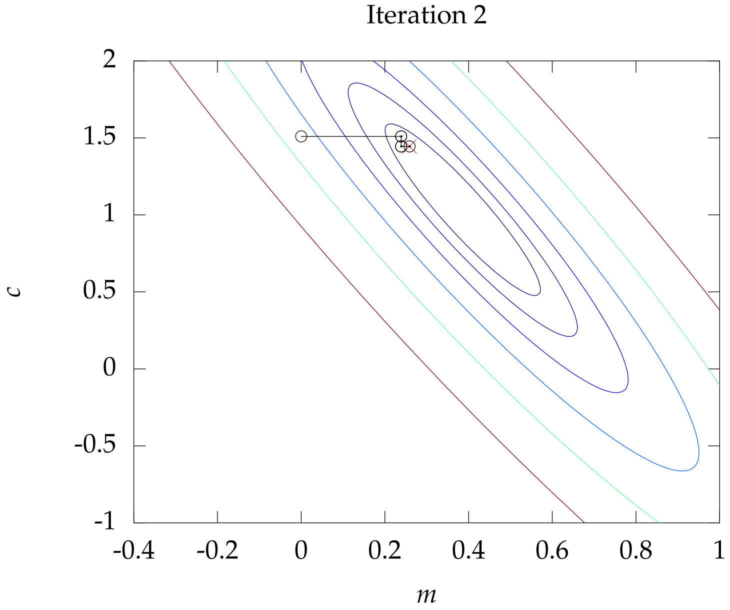
Coordinate Descent



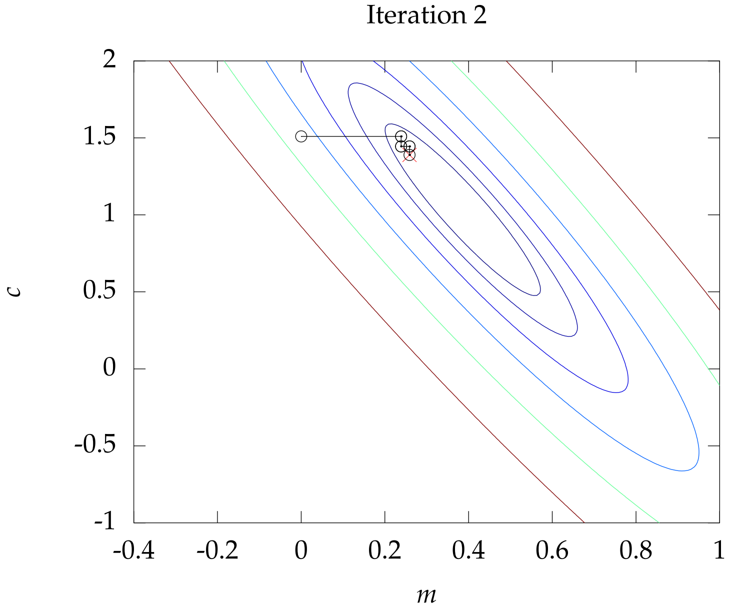
Coordinate Descent



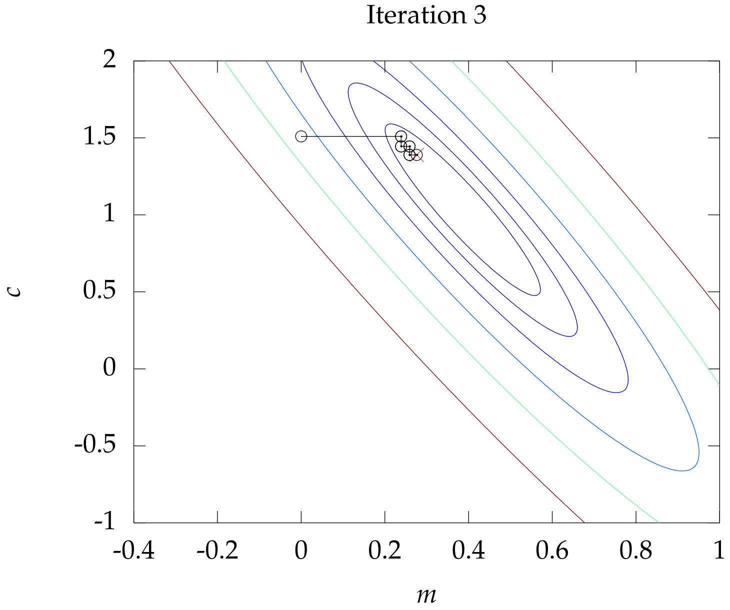
Coordinate Descent



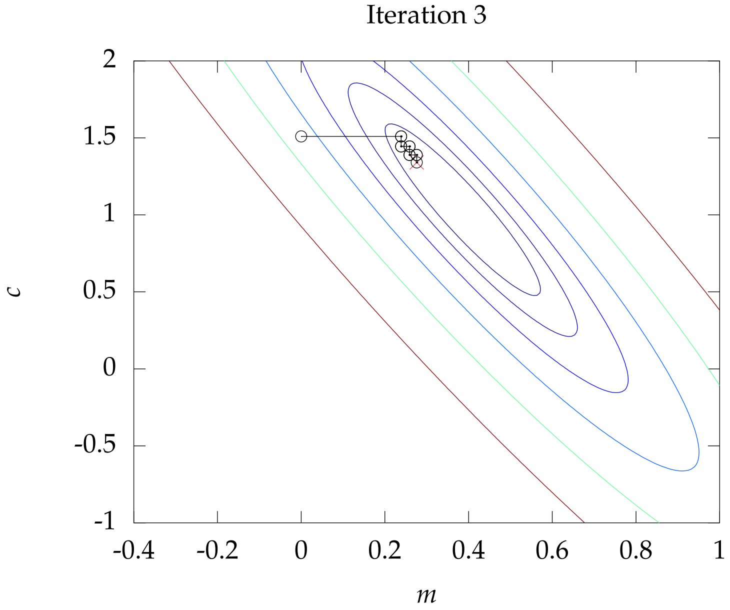
Coordinate Descent



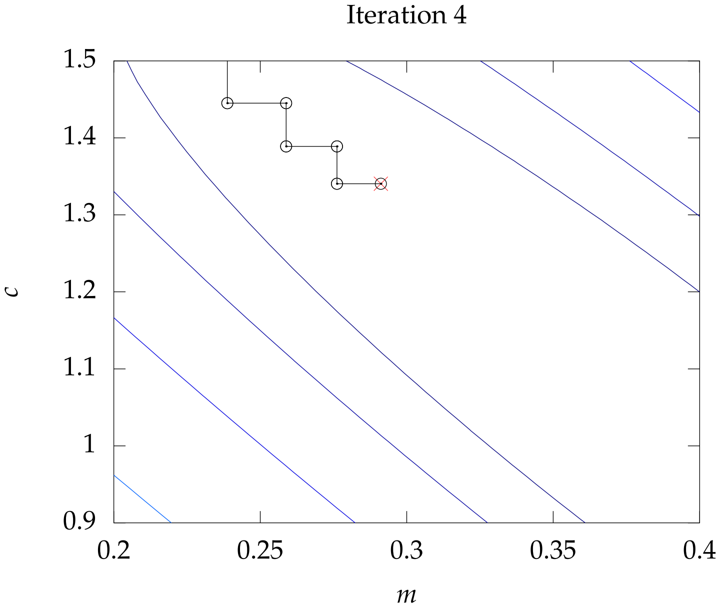
Coordinate Descent



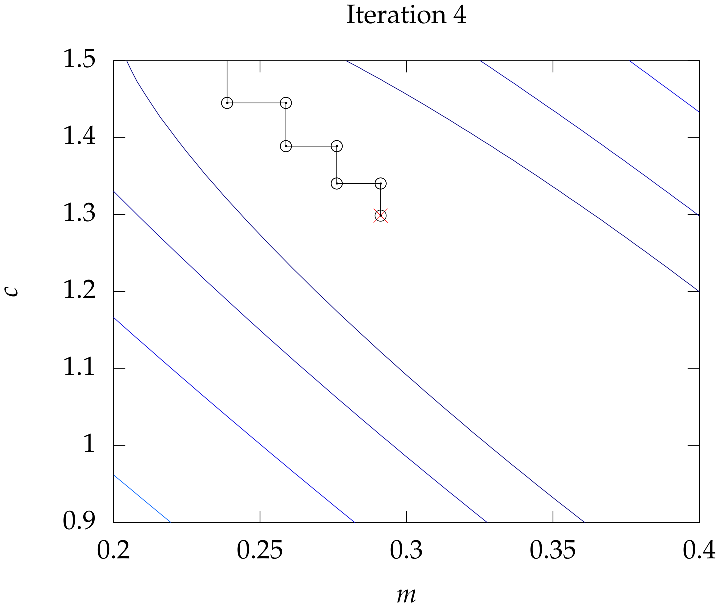
Coordinate Descent



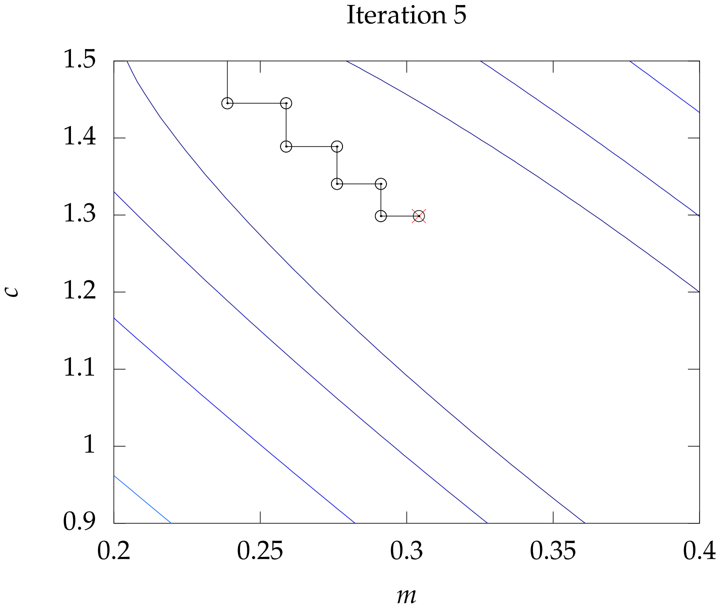
Coordinate Descent



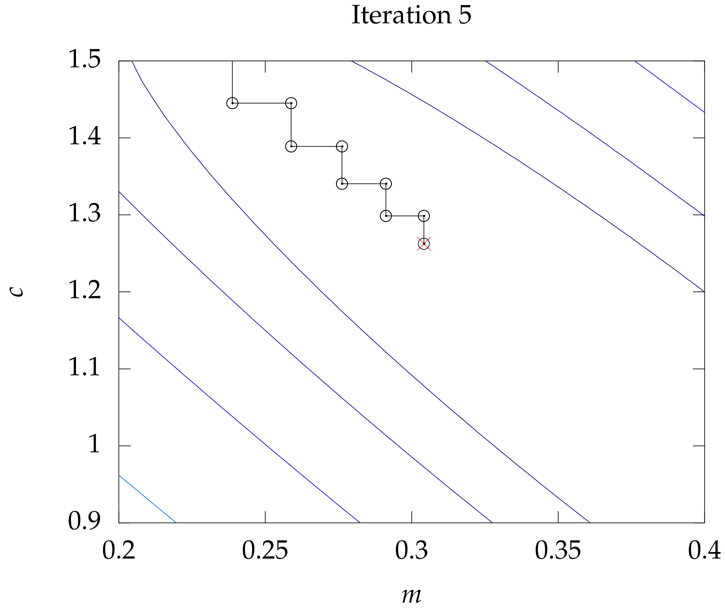
Coordinate Descent



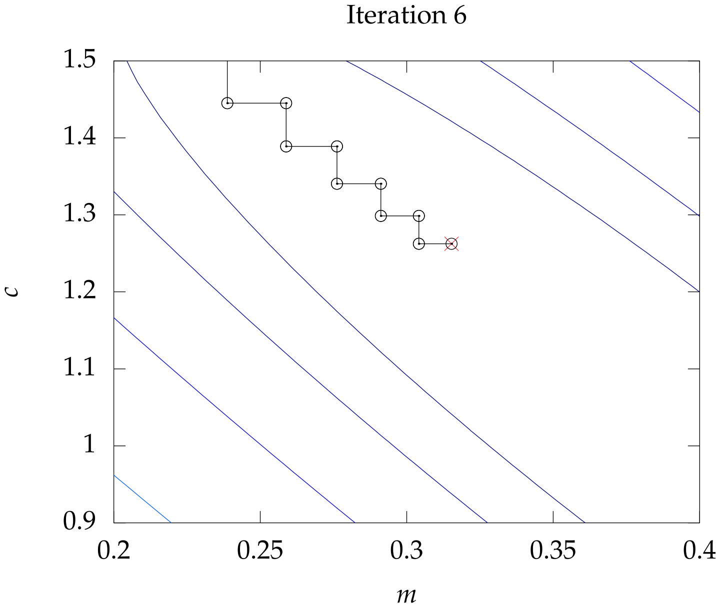
Coordinate Descent



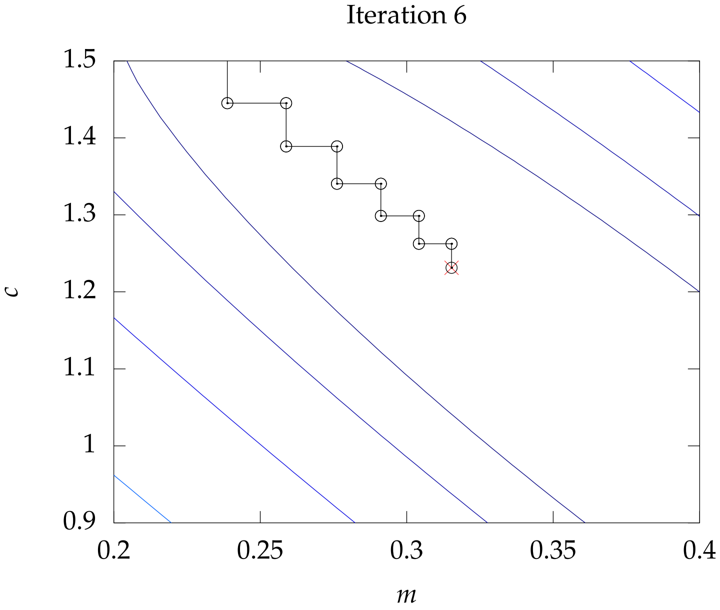
Coordinate Descent



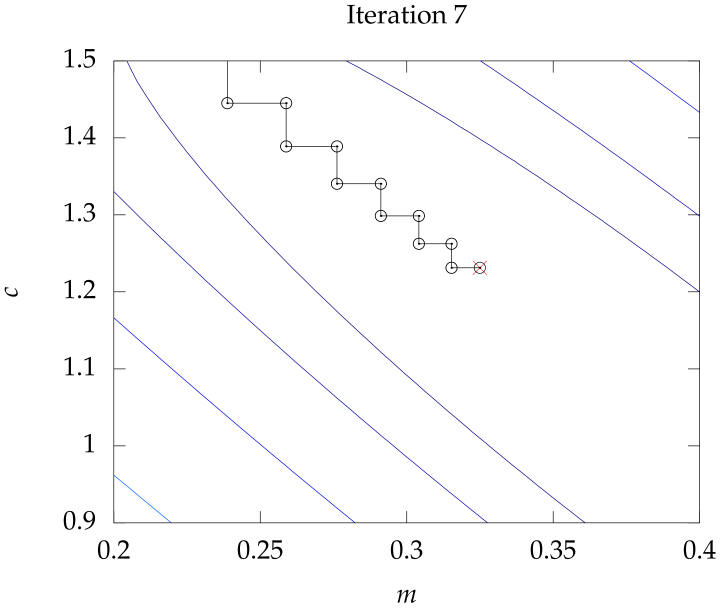
Coordinate Descent



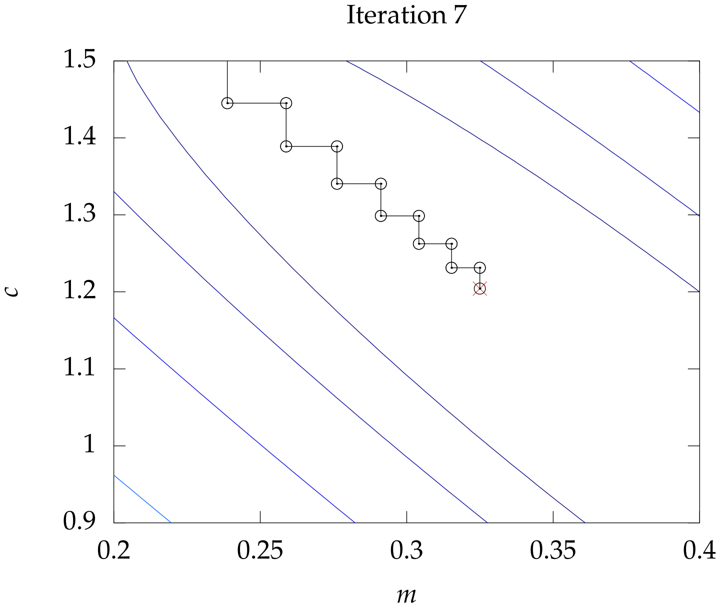
Coordinate Descent



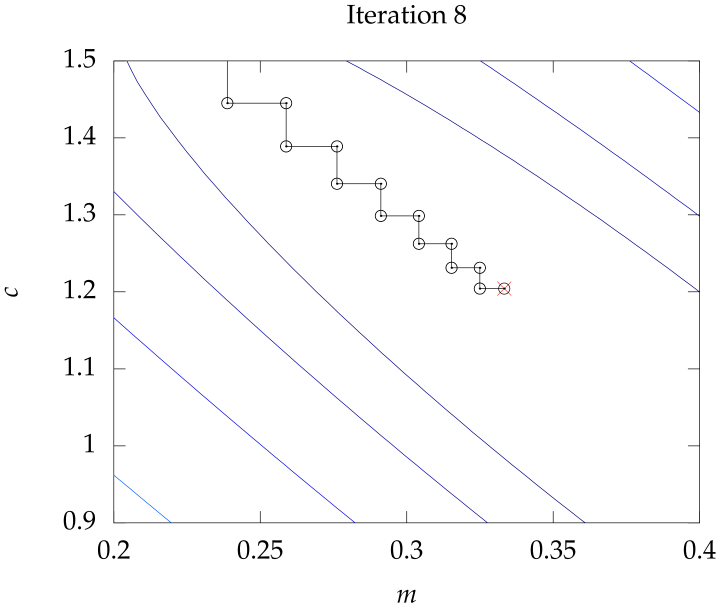
Coordinate Descent



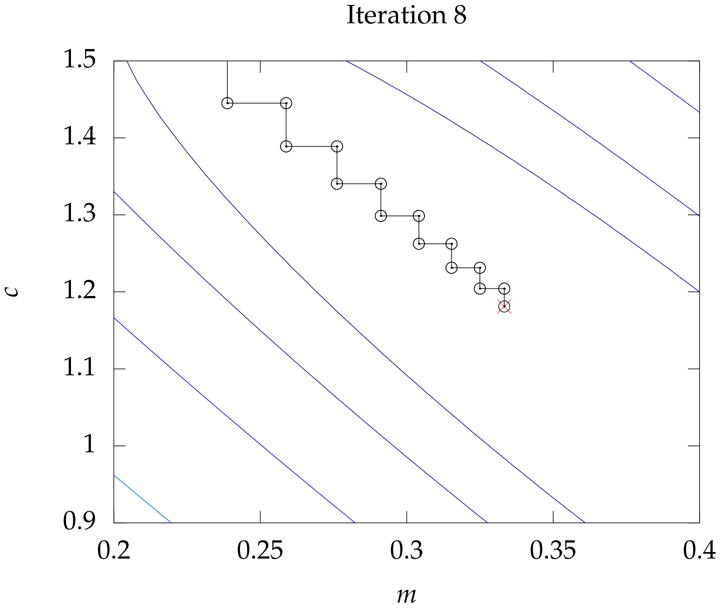
Coordinate Descent



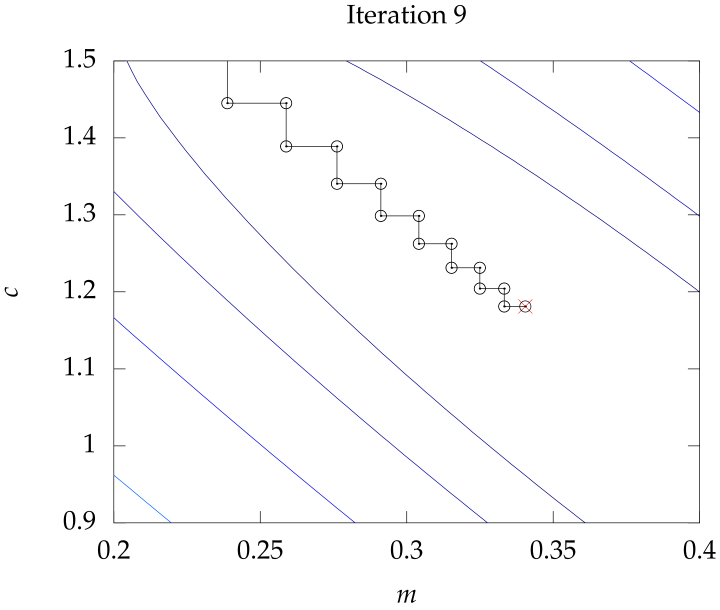
Coordinate Descent



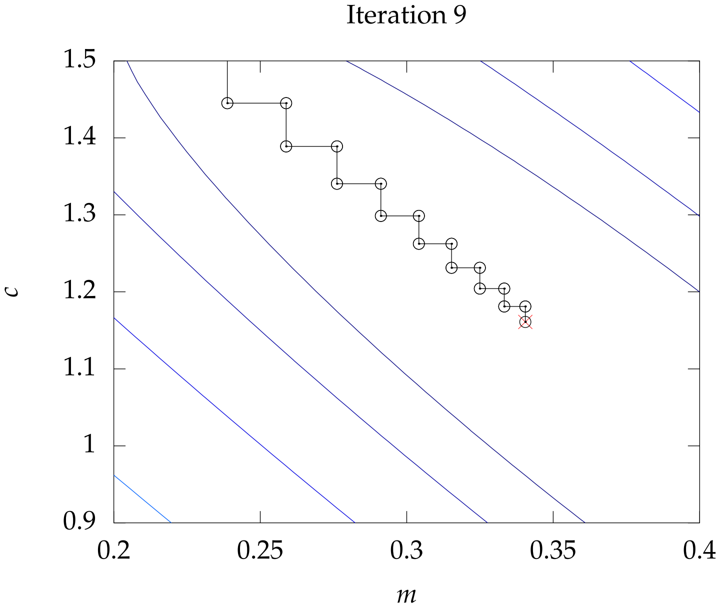
Coordinate Descent



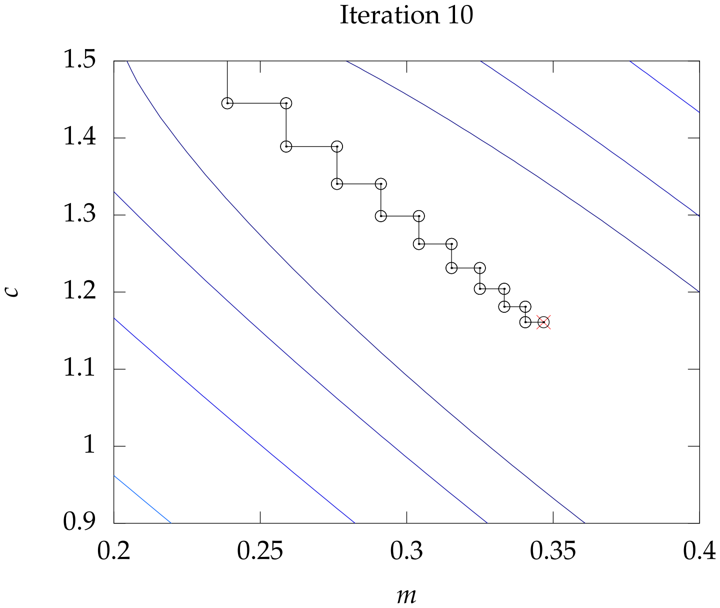
Coordinate Descent



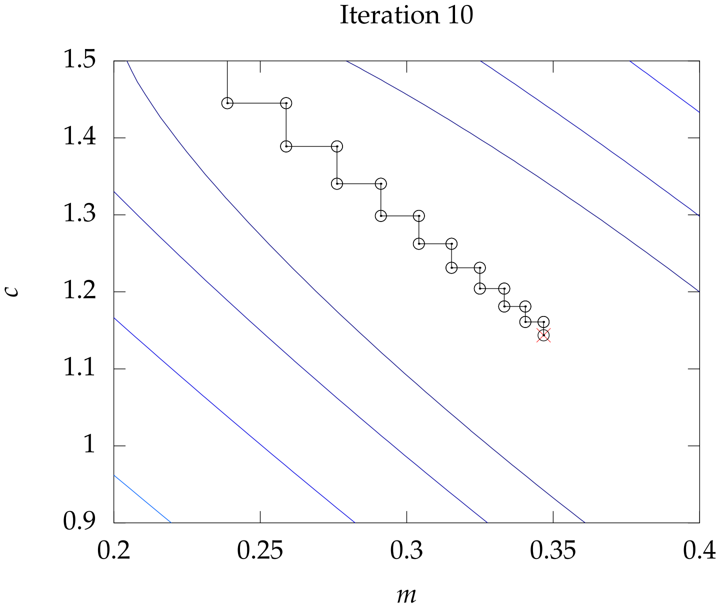
Coordinate Descent



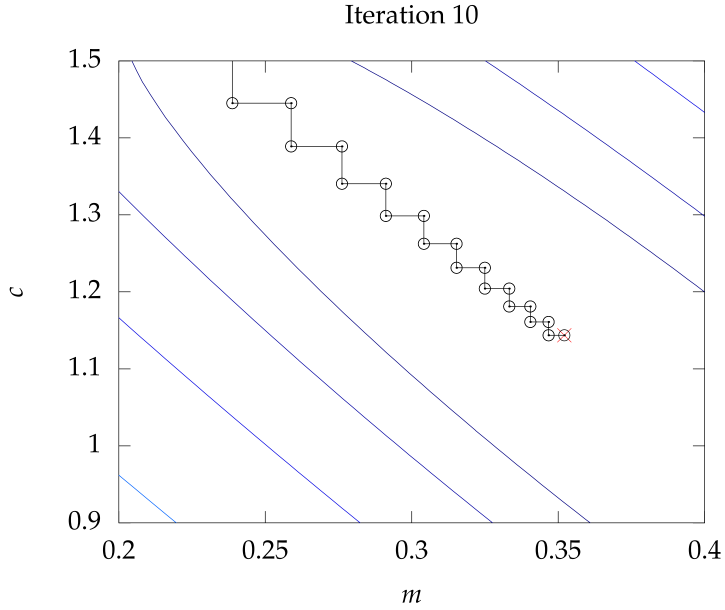
Coordinate Descent



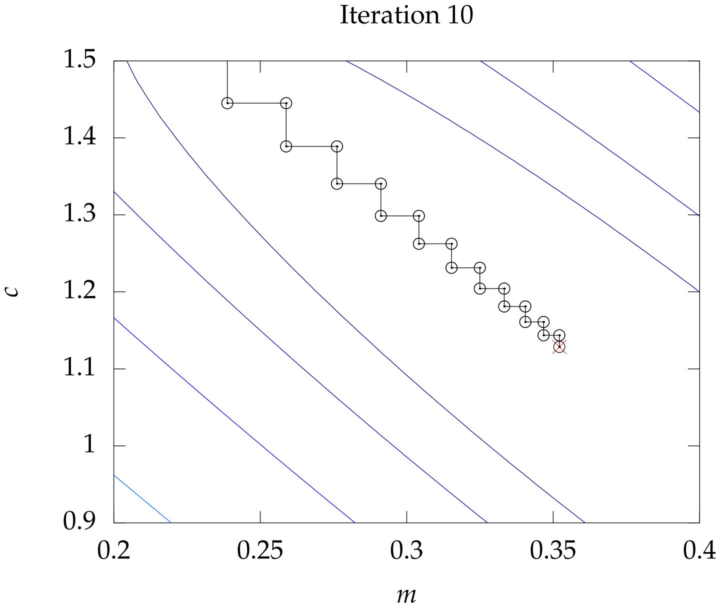
Coordinate Descent



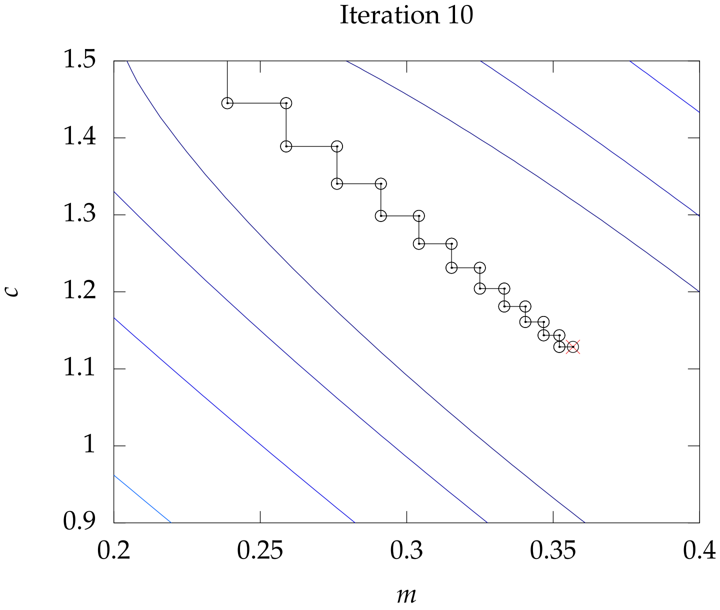
Coordinate Descent



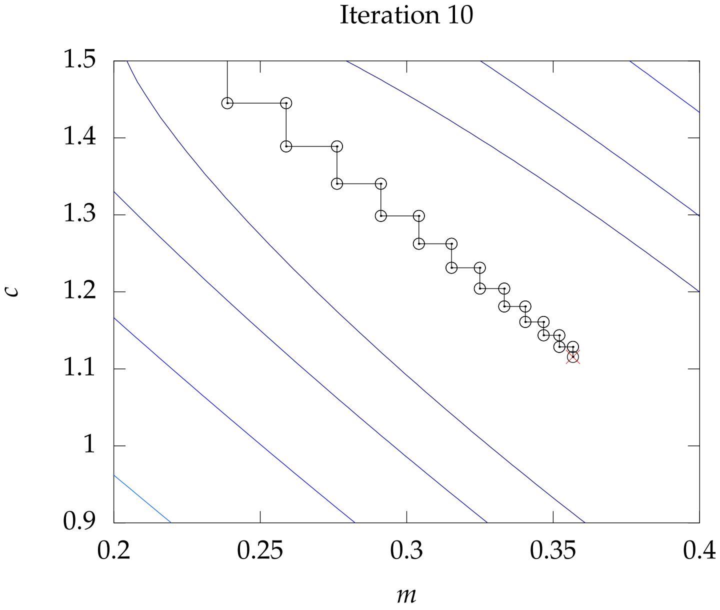
Coordinate Descent



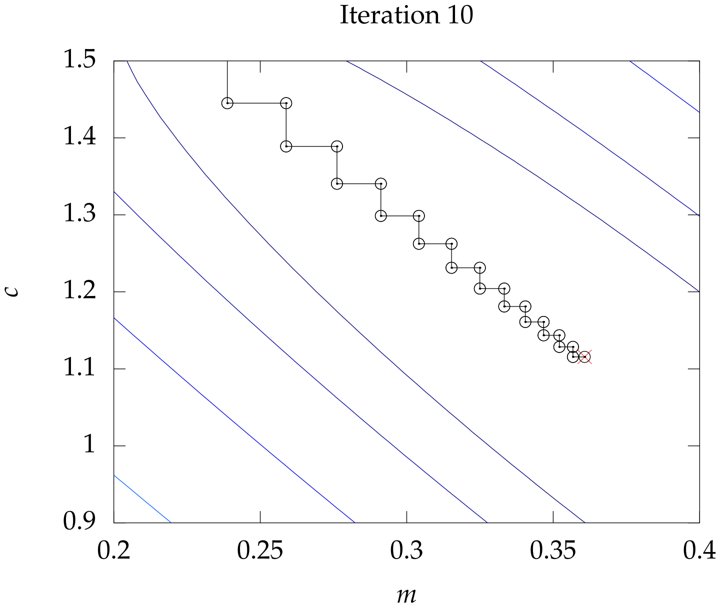
Coordinate Descent



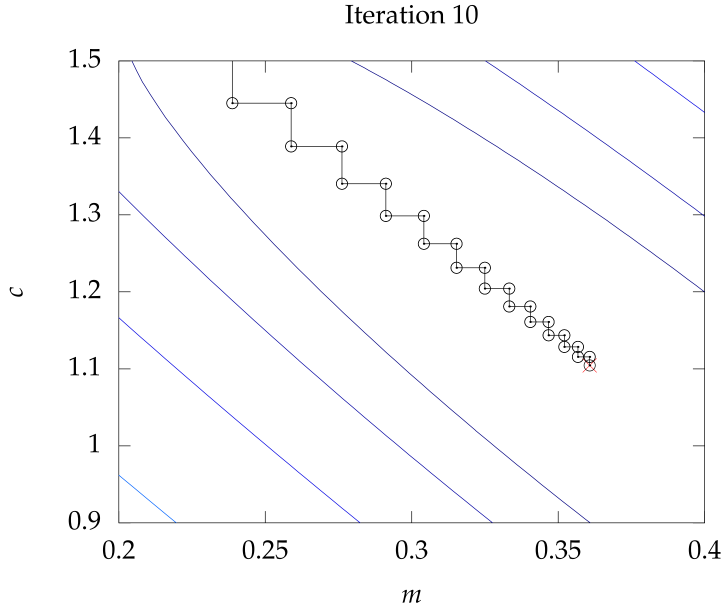
Coordinate Descent



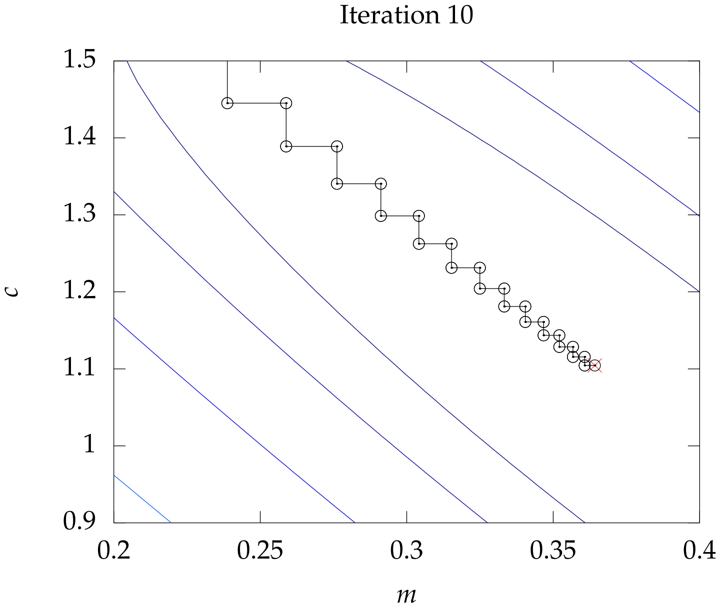
Coordinate Descent



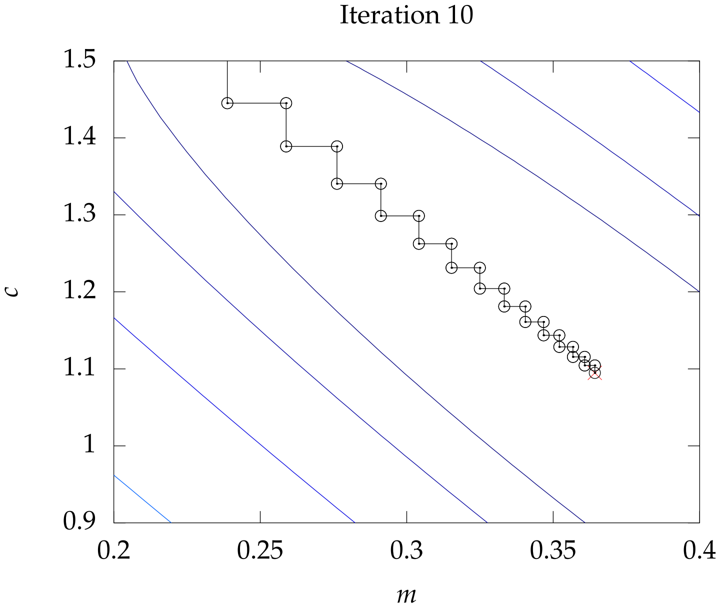
Coordinate Descent



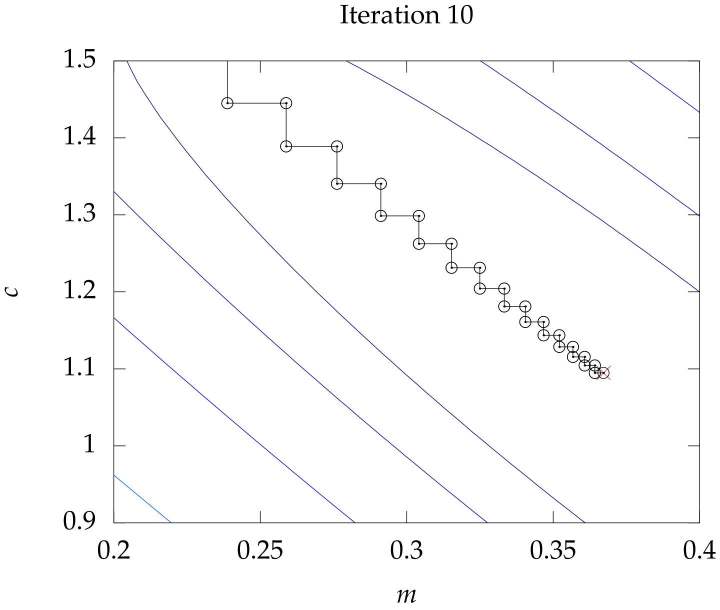
Coordinate Descent



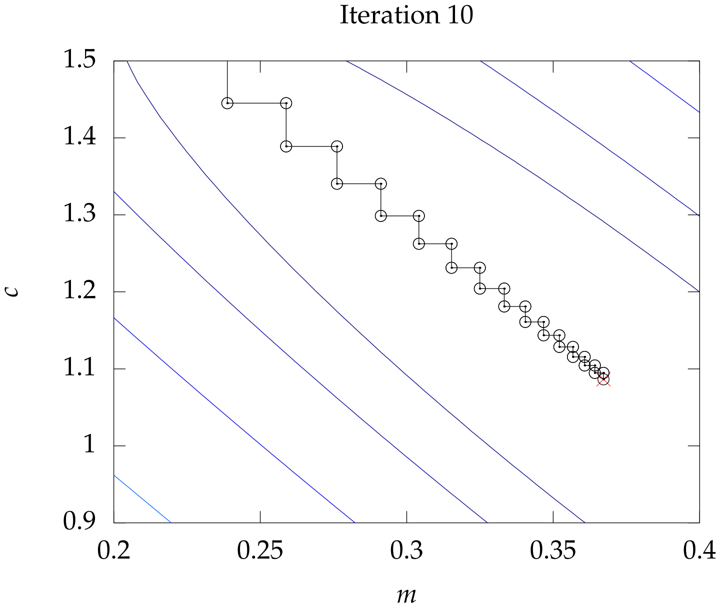
Coordinate Descent



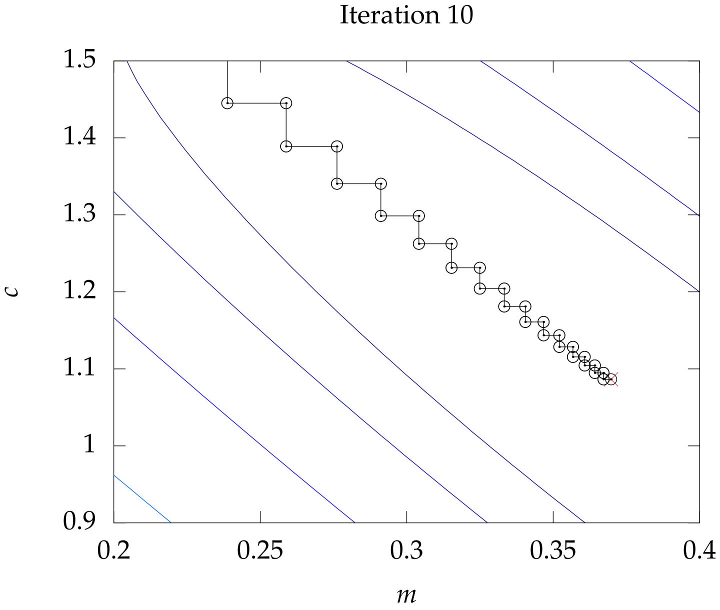
Coordinate Descent



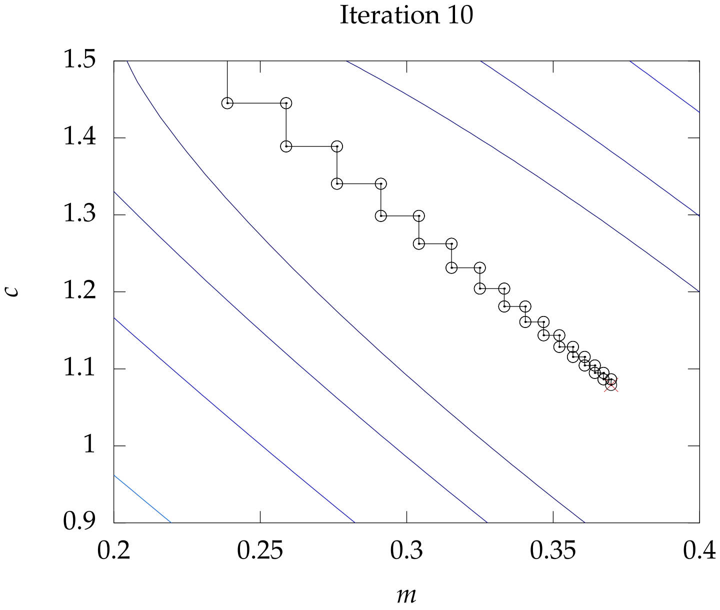
Coordinate Descent



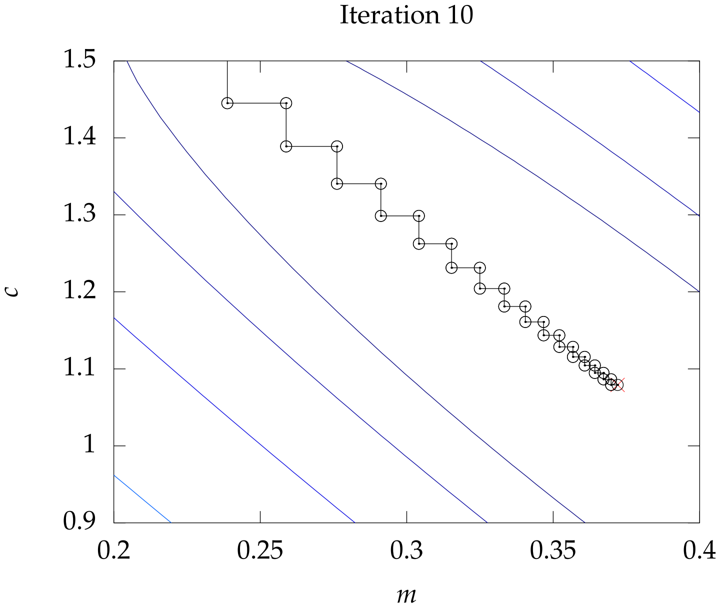
Coordinate Descent



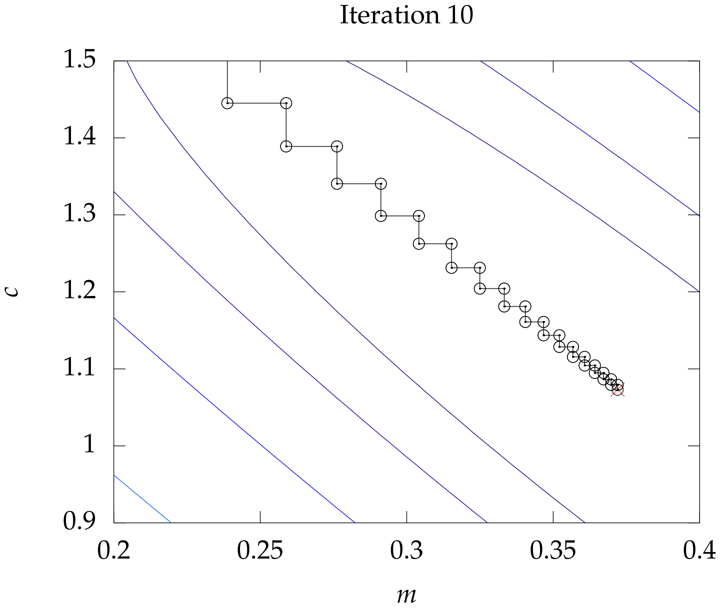
Coordinate Descent



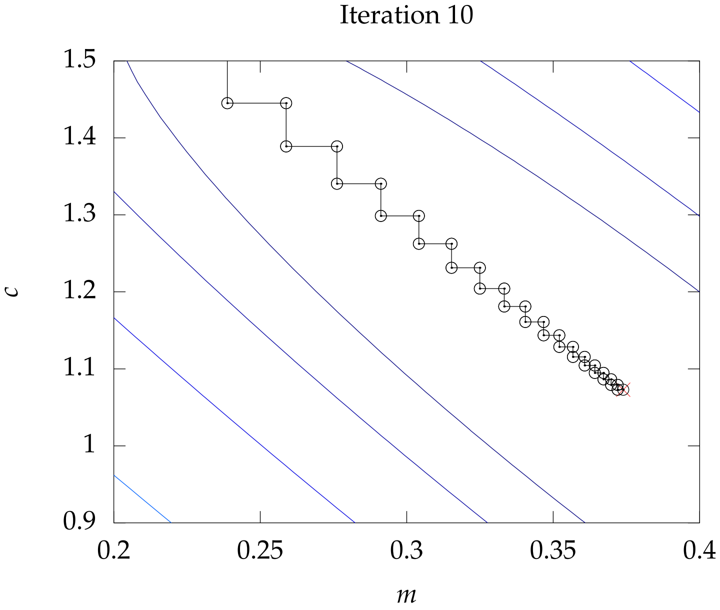
Coordinate Descent



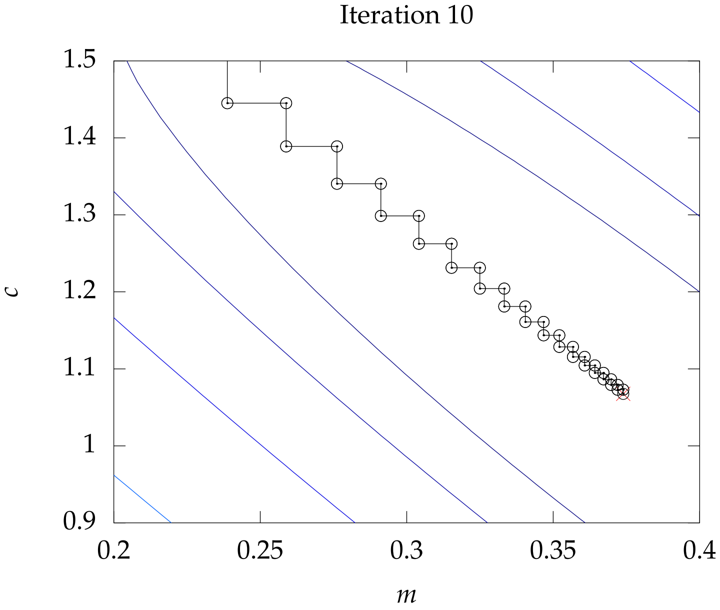
Coordinate Descent



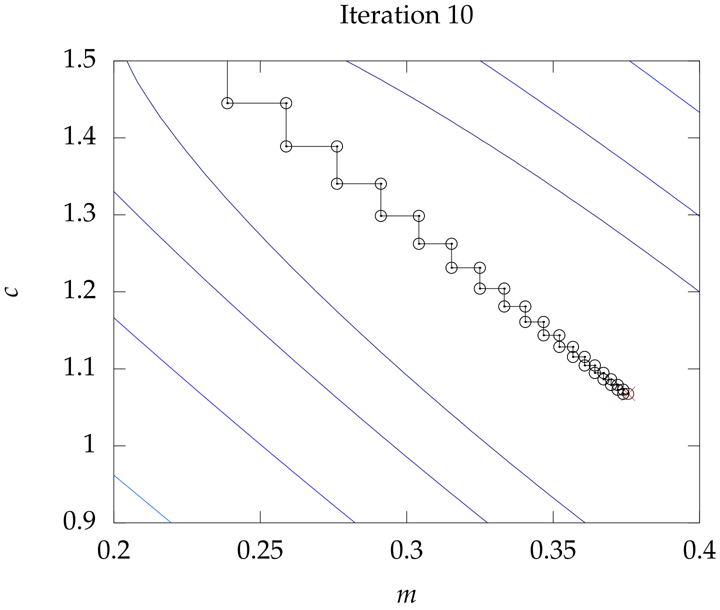
Coordinate Descent



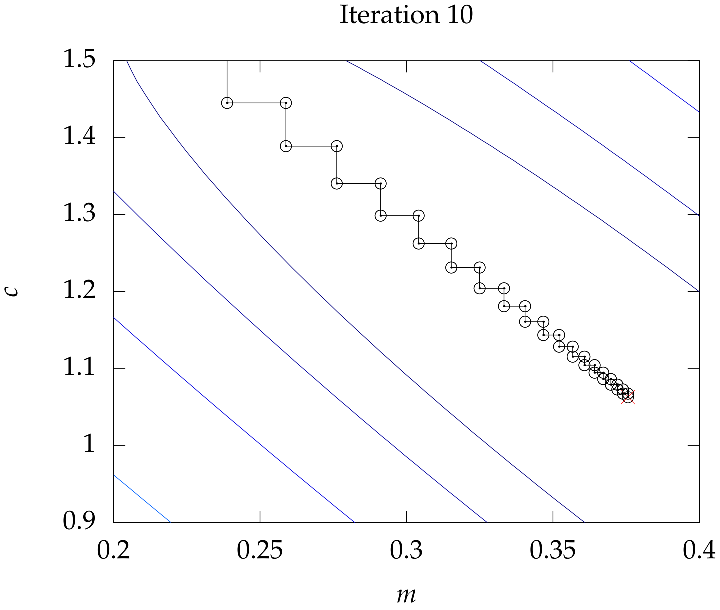
Coordinate Descent



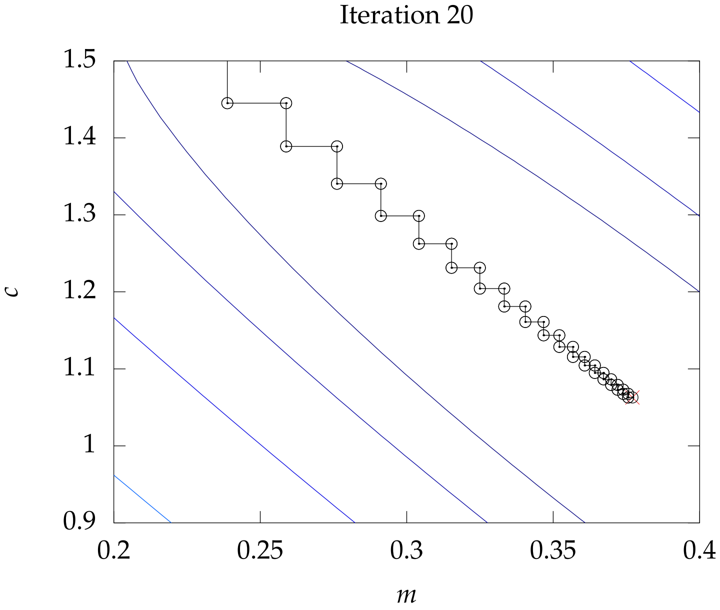
Coordinate Descent



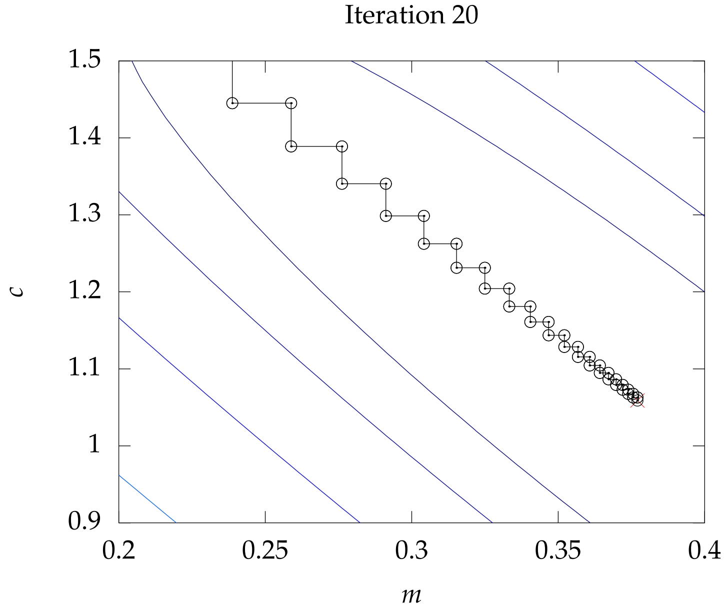
Coordinate Descent



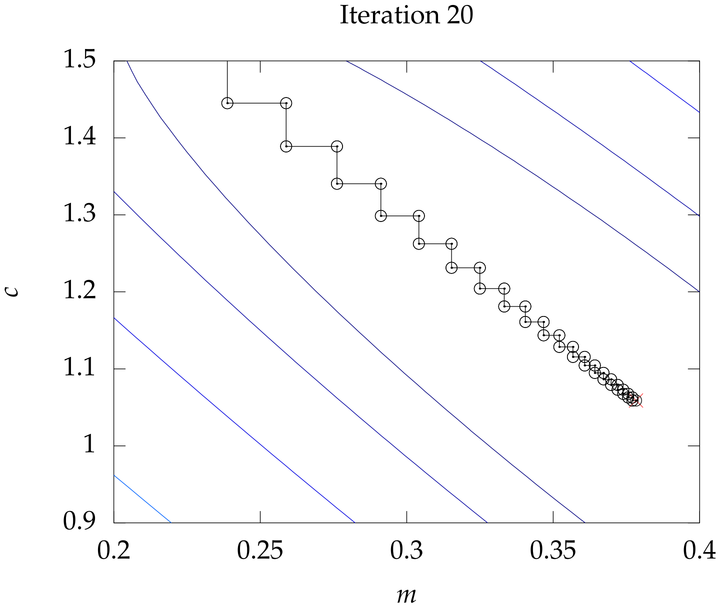
Coordinate Descent



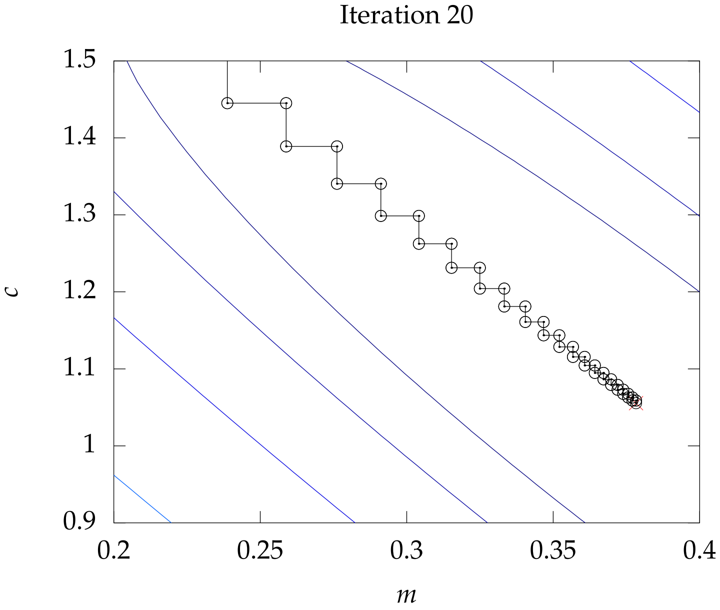
Coordinate Descent



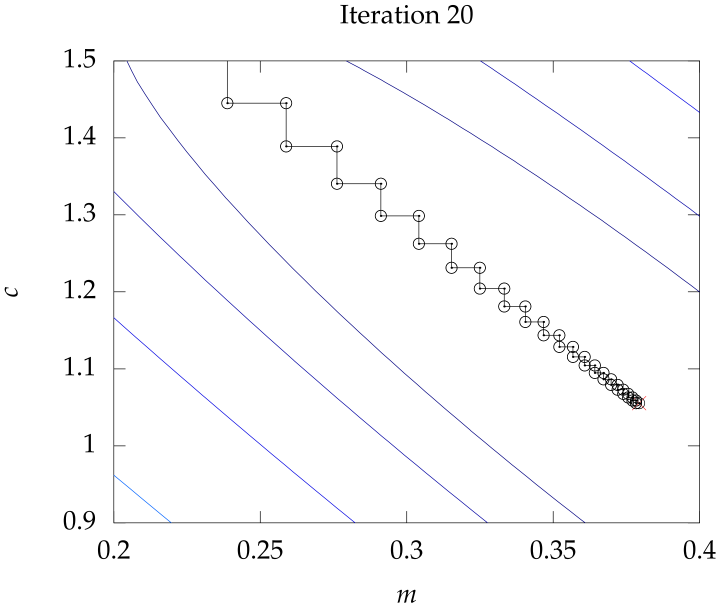
Coordinate Descent



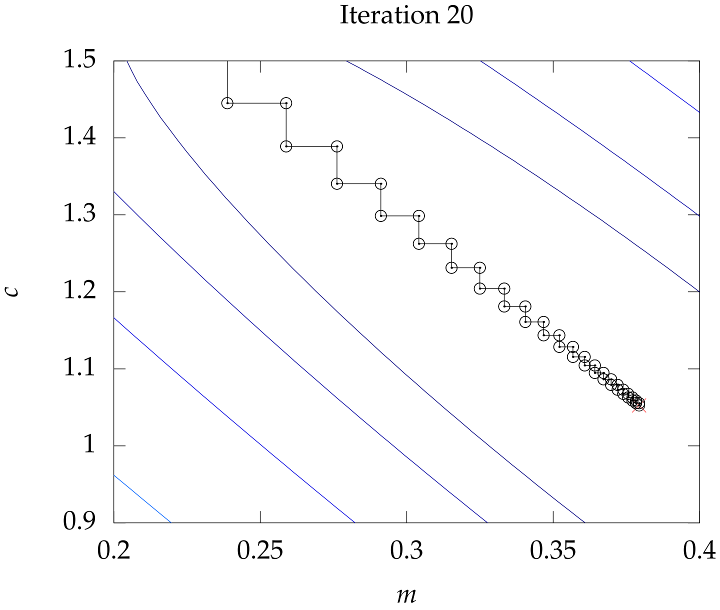
Coordinate Descent



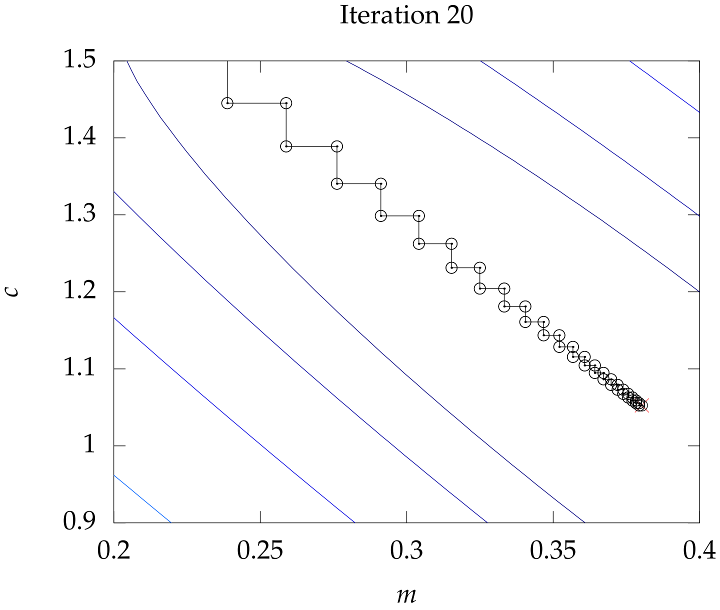
Coordinate Descent



Coordinate Descent

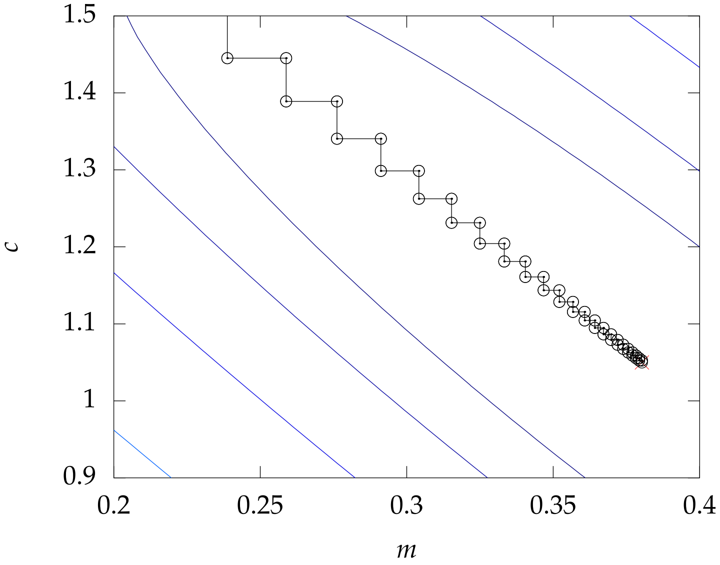


Coordinate Descent

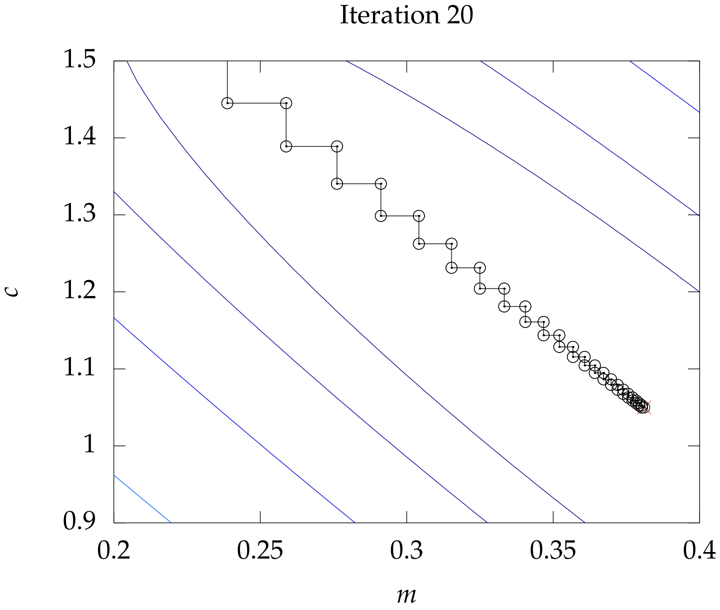


Coordinate Descent

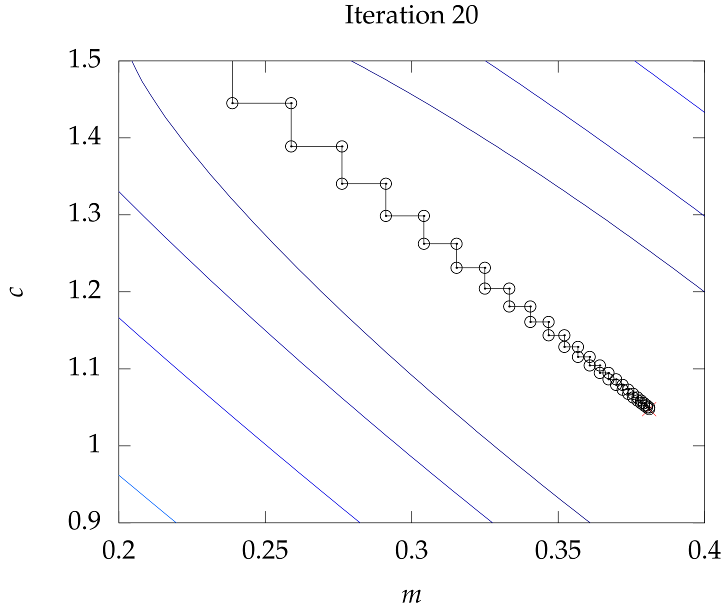
Iteration 20



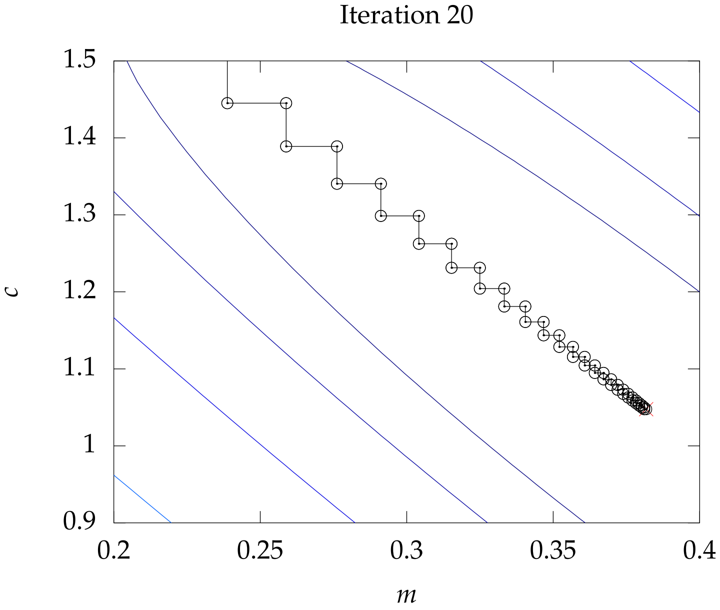
Coordinate Descent



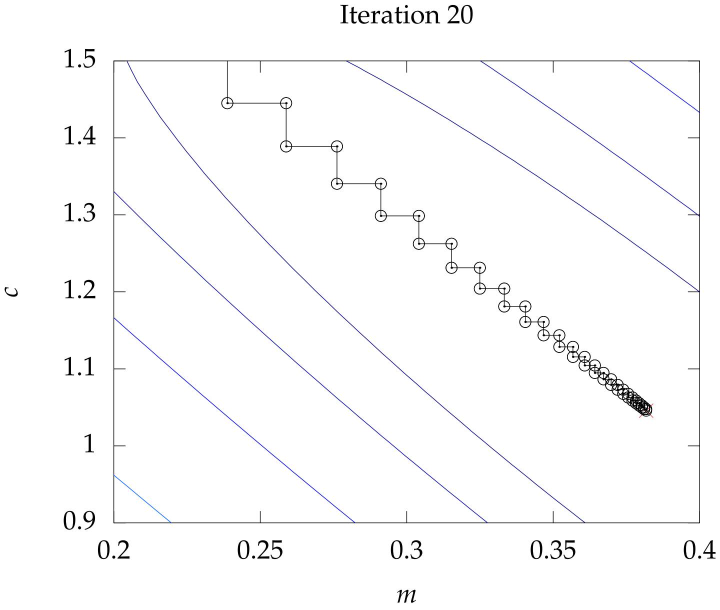
Coordinate Descent



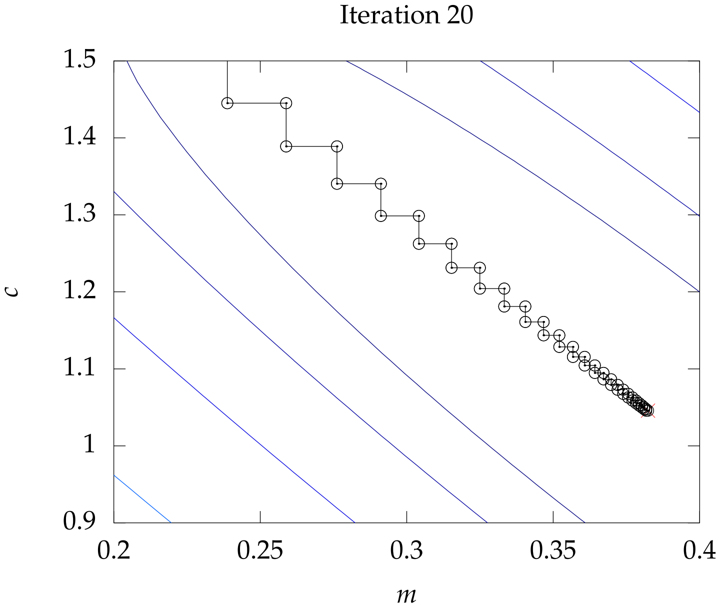
Coordinate Descent



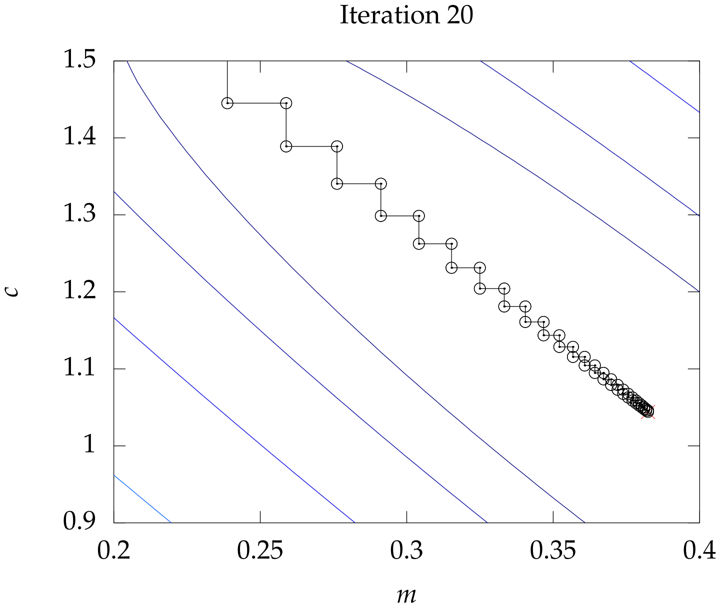
Coordinate Descent



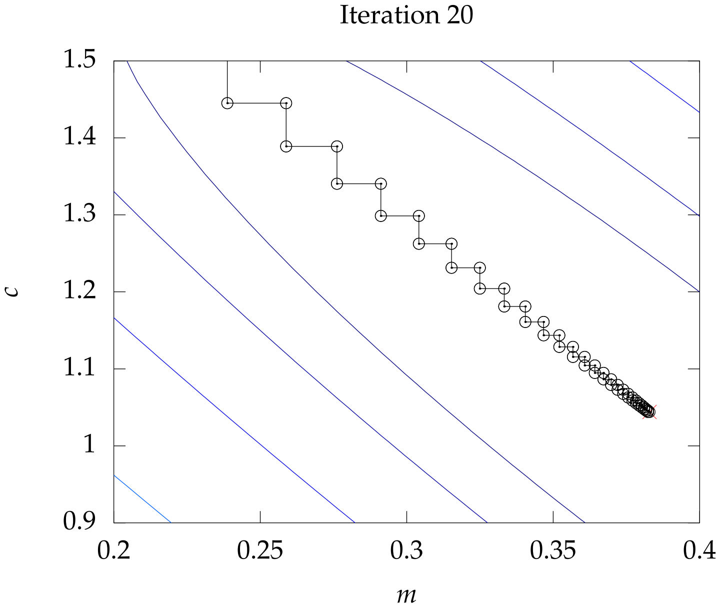
Coordinate Descent



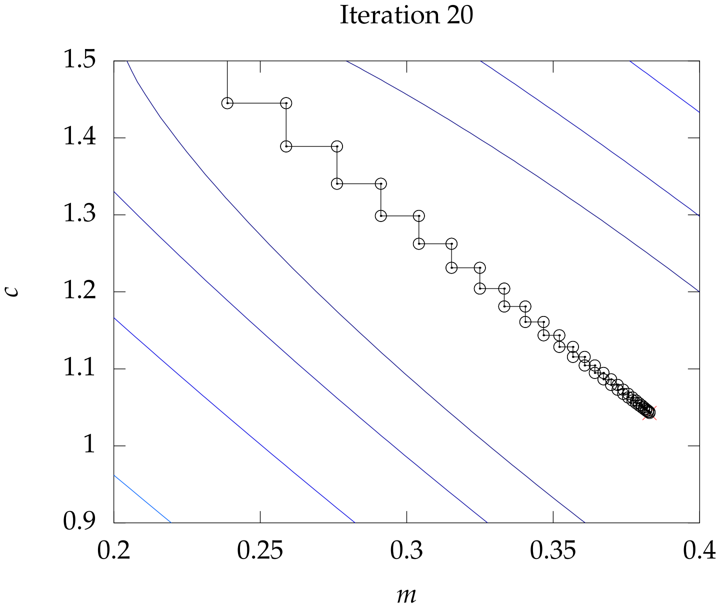
Coordinate Descent



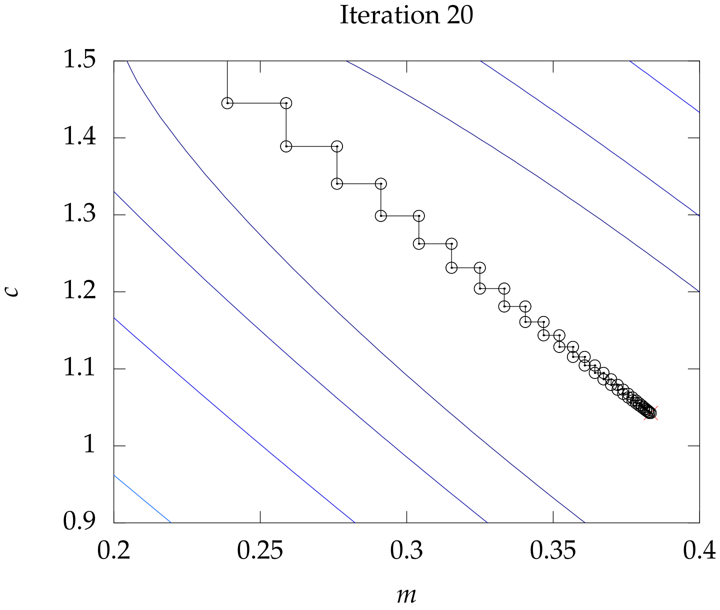
Coordinate Descent



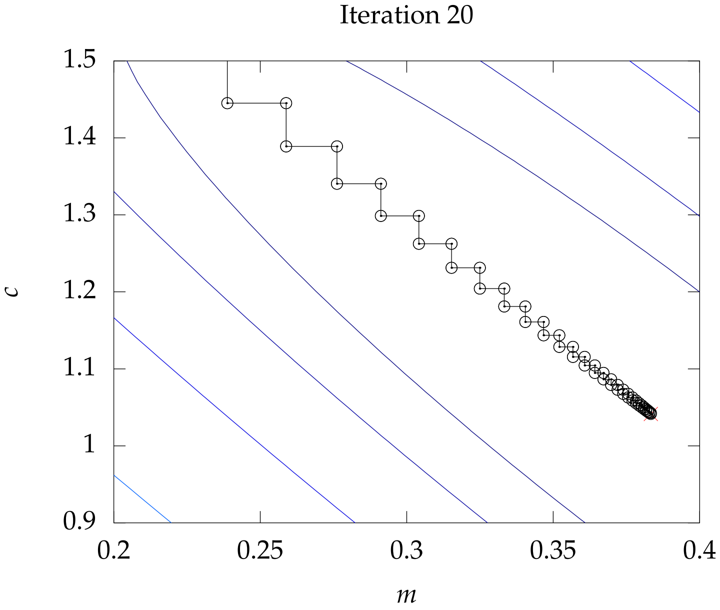
Coordinate Descent



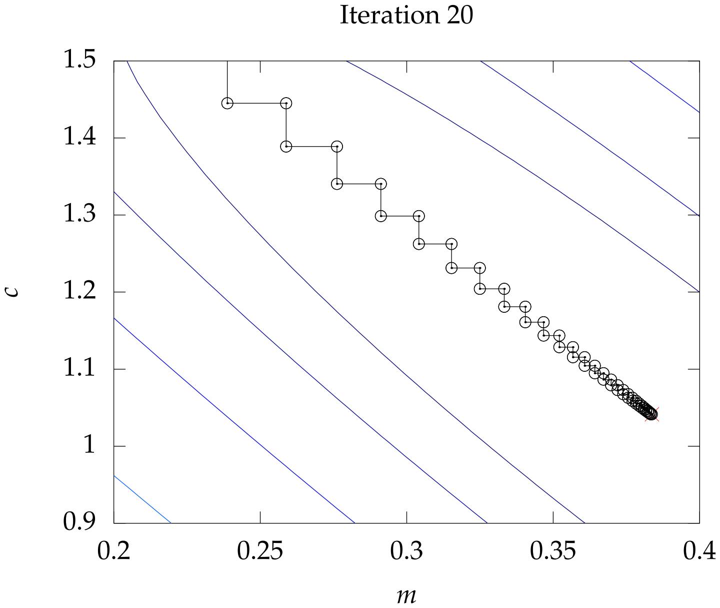
Coordinate Descent



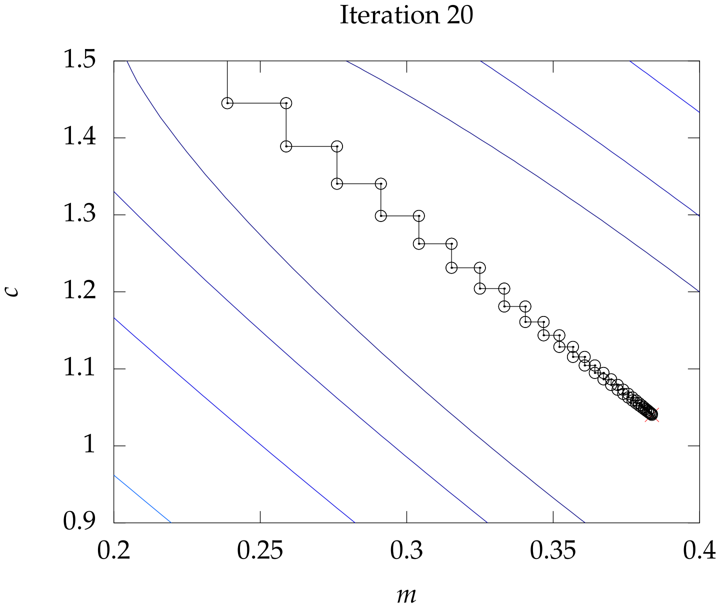
Coordinate Descent



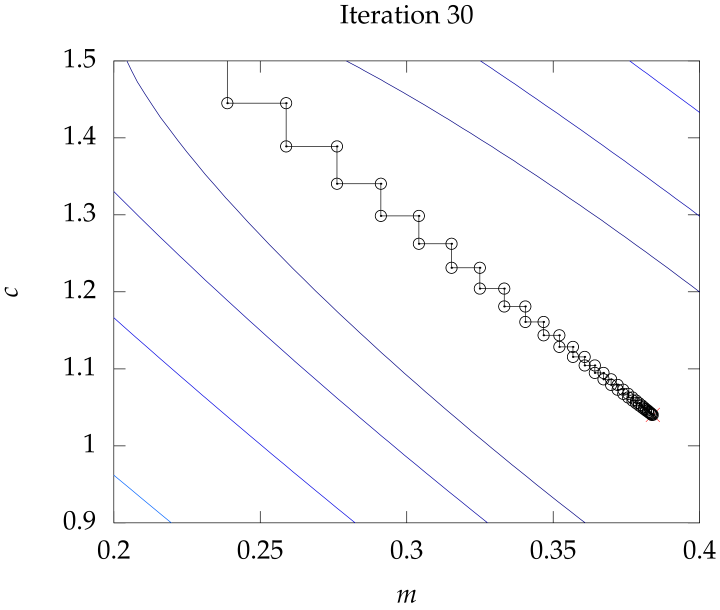
Coordinate Descent



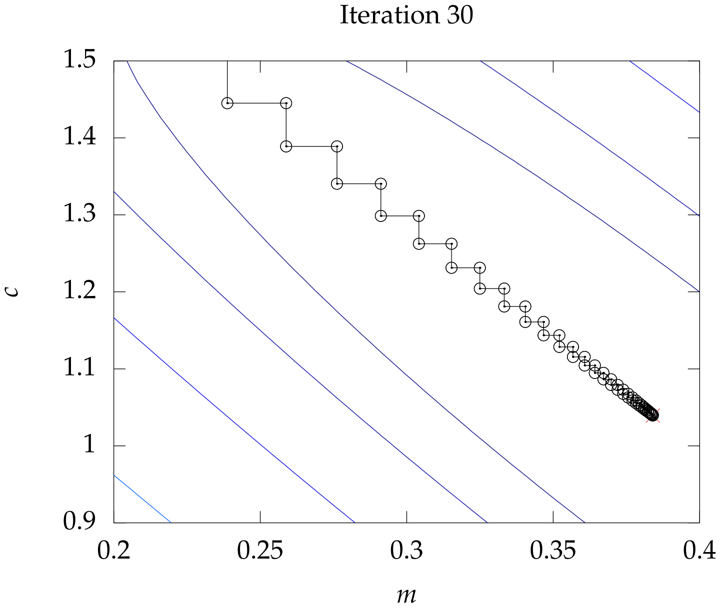
Coordinate Descent



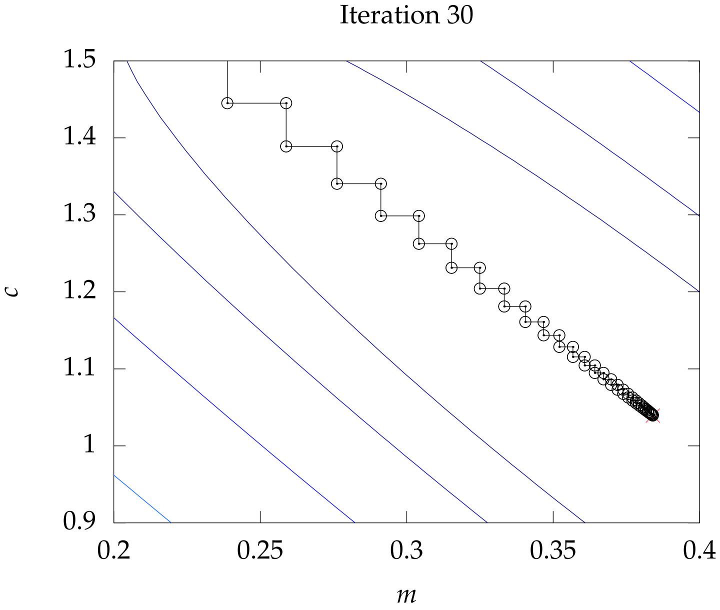
Coordinate Descent



Coordinate Descent



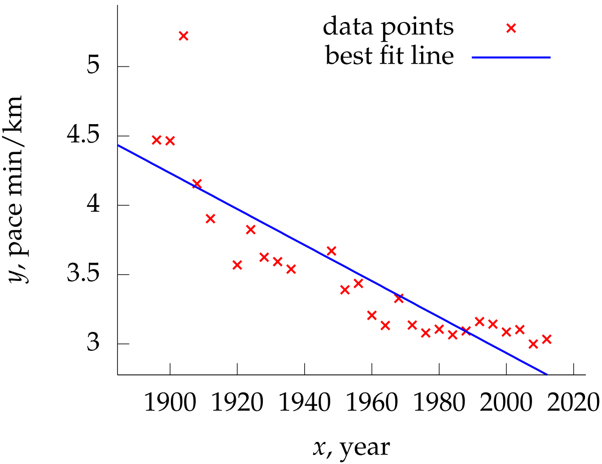
Coordinate Descent



Important Concepts Not Covered

- Optimization methods.
  - Second order methods, conjugate gradient, quasi-Newton and Newton.
  - Effective heuristics such as momentum.
- Local vs global solutions.

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.

- 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:

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

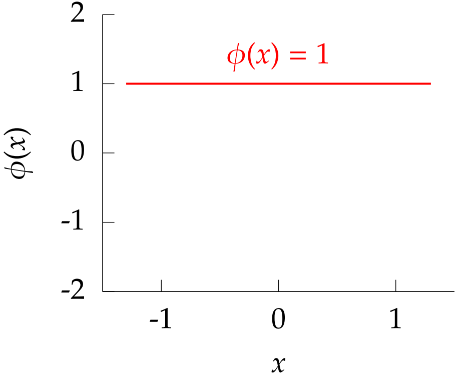


Figure : A quadratic basis.

Quadratic Basis

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

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

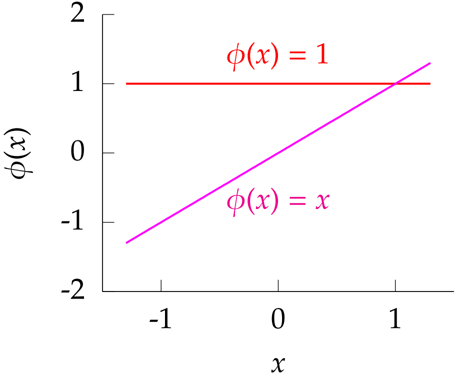


Figure : A quadratic basis.

Quadratic Basis

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

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

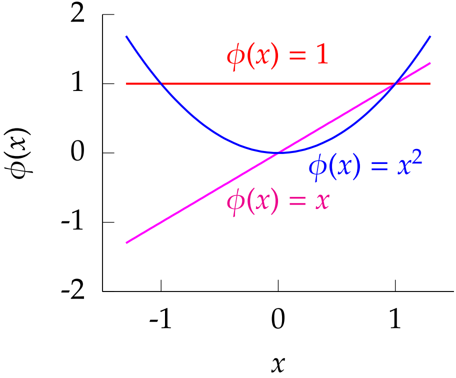


Figure : A quadratic basis.

Functions Derived from Quadratic Basis

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

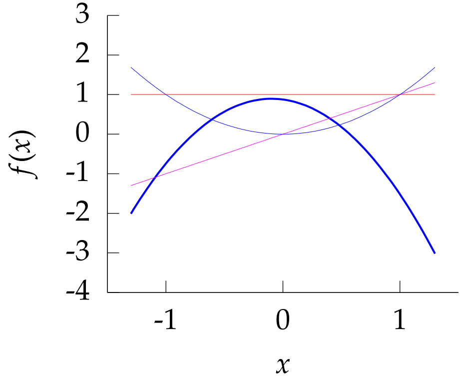


Figure : Function from quadratic basis with weights $w_1 = 0.87466$, $w_2 = -0.38835$, $w_3 = -2.0058$.

Functions Derived from Quadratic Basis

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

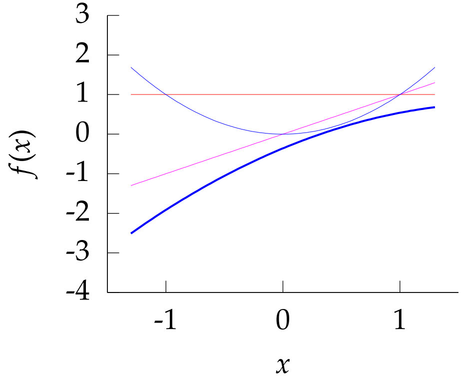


Figure : Function from quadratic basis with weights $w_1 = -0.35908$, $w_2 = 1.2274$, $w_3 = -0.32825$.

Functions Derived from Quadratic Basis

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

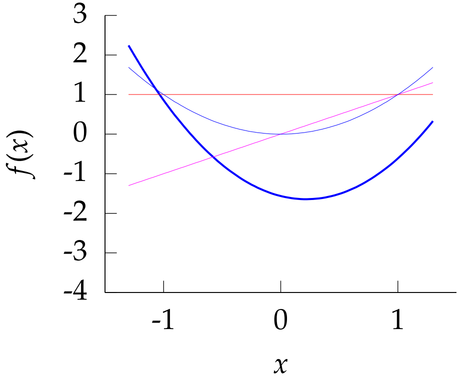


Figure : Function from quadratic basis with weights $w_1 = -1.5638$, $w_2 = -0.73577$, $w_3 = 1.6861$.

Radial Basis Functions

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

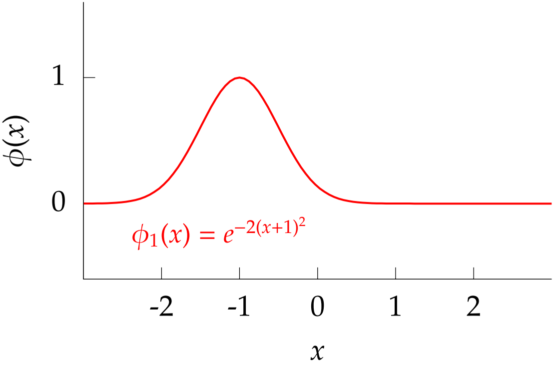


Figure : Radial basis functions.

Radial Basis Functions

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

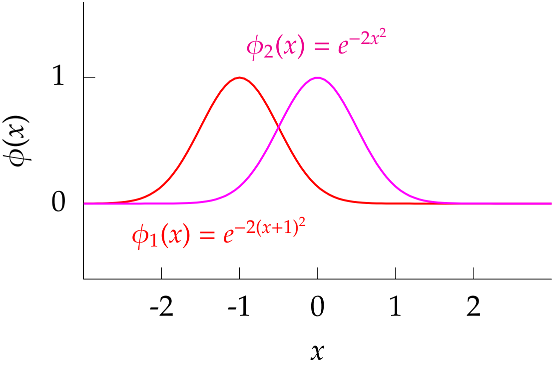


Figure : Radial basis functions.

Radial Basis Functions

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

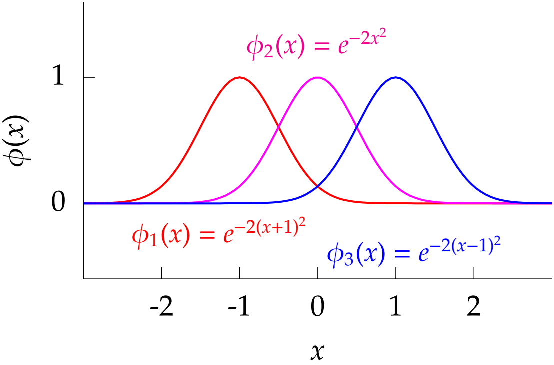


Figure : Radial basis functions.

Functions Derived from Radial Basis

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

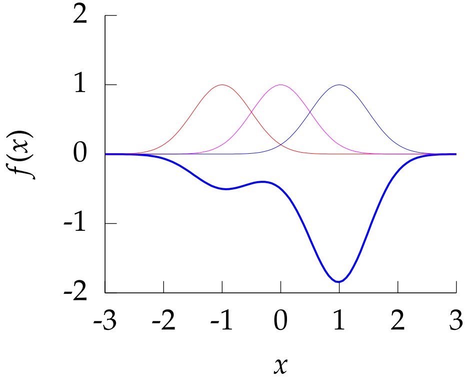


Figure : Function from radial basis with weights $w_1 = -0.47518$, $w_2 = -0.18924$, $w_3 = -1.8183$.

Functions Derived from Radial Basis

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

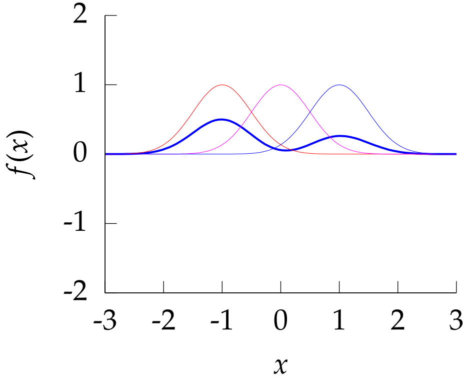


Figure : Function from radial basis with weights $w_1 = 0.50596$, $w_2 = -0.046315$, $w_3 = 0.26813$.

Functions Derived from Radial Basis

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

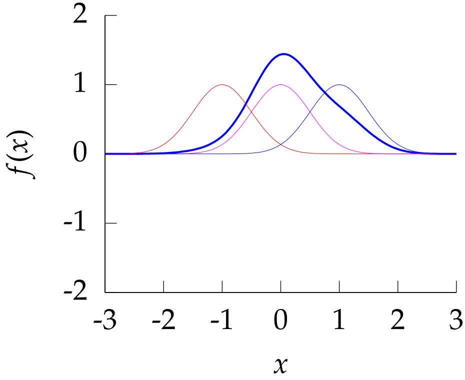


Figure : Function from radial basis with weights $w_1 = 0.07179$, $w_2 = 1.3591$, $w_3 = 0.50604$.

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.

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 **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 $\mathbf{w}^*$.

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

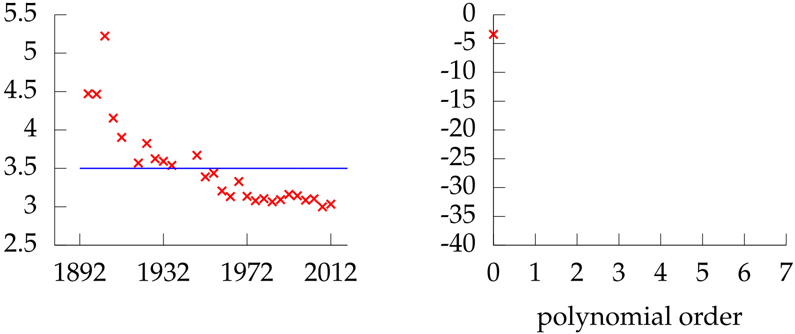
- The equation for ${\sigma^2}^*$ may also be found

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

Reading

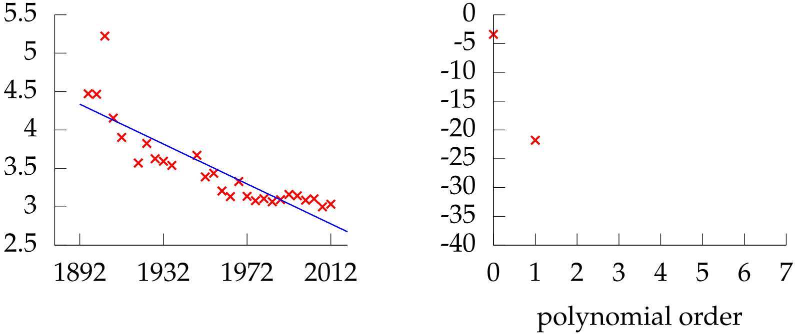
- Section 1.3 of Rogers and Girolami for Matrix & Vector Review.

Polynomial Fits to Olympics Data



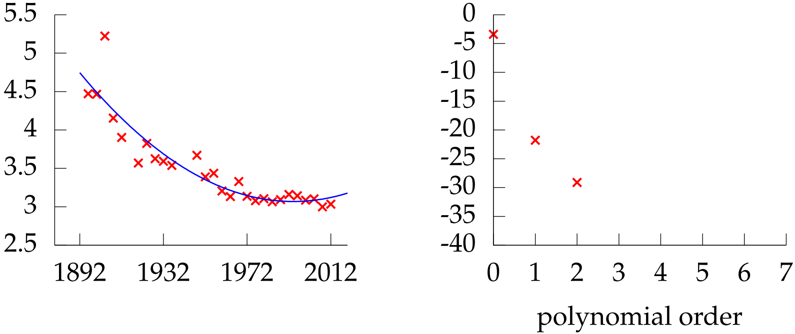
Left: fit to data, *Right*: model error. Polynomial order 0, model error -3.3989, $\sigma^2 = 0.286$, $\sigma = 0.535$.

Polynomial Fits to Olympics Data



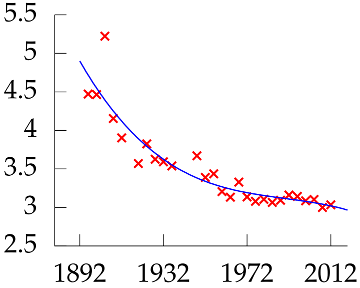
Left: fit to data, *Right*: model error. Polynomial order 1, model error -21.772, $\sigma^2 = 0.0733$, $\sigma = 0.271$.

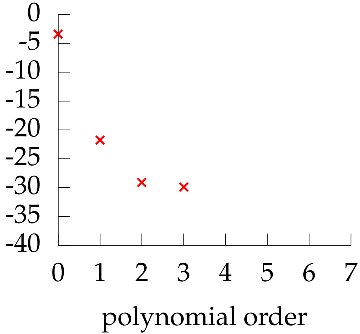
Polynomial Fits to Olympics Data



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

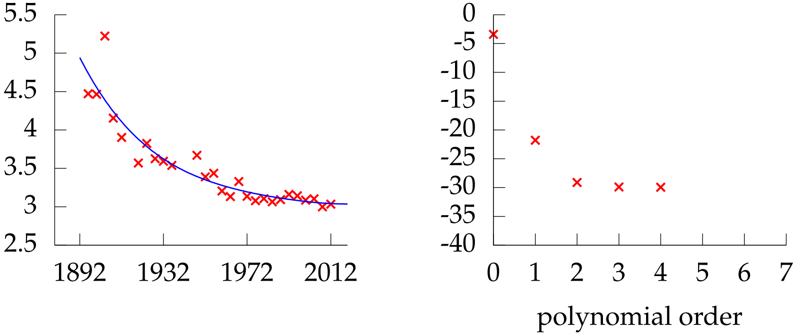
Polynomial Fits to Olympics Data





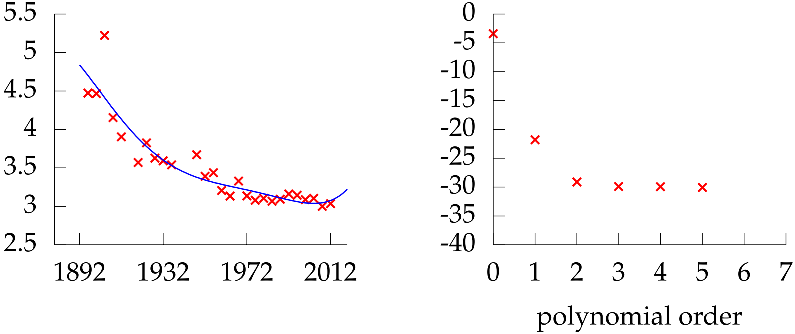
Left: fit to data, *Right*: model error. Polynomial order 3, model error -29.907, $\sigma^2 = 0.0401$, $\sigma = 0.200$.

Polynomial Fits to Olympics Data



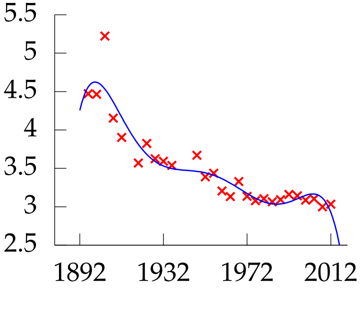
Left: fit to data, *Right*: model error. Polynomial order 4, model error -29.943, $\sigma^2 = 0.0400$, $\sigma = 0.200$.

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, $\sigma^2 = 0.0322$, $\sigma = 0.180$.

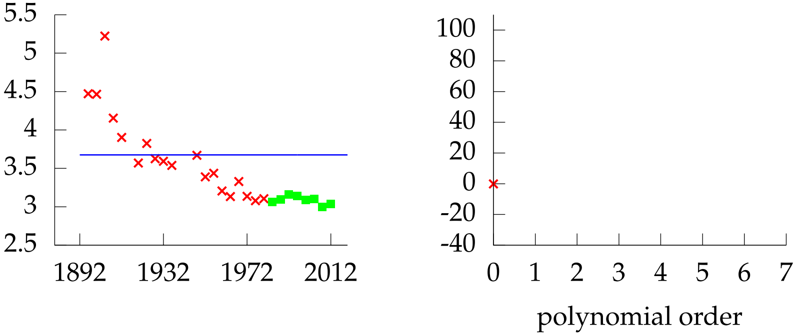
Overfitting

- 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

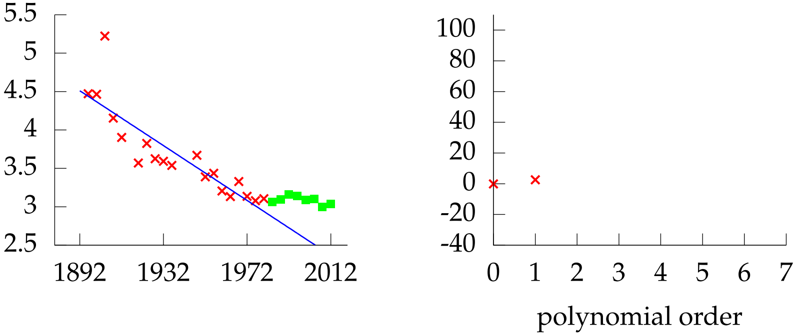
- We call the data used for fitting the model the ‘training set’.
- Data not used for training, but when the model is applied ‘in the field’ is called the ‘test data’.
- Challenge for generalization is to ensure a good performance on test data given only training data.

Validation Set



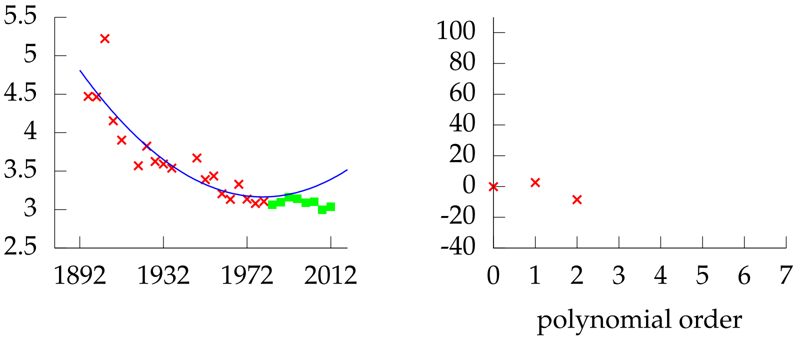
Left: fit to data, *Right*: model error. Polynomial order 0, training error -1.8774, validation error -0.13132, $\sigma^2 = 0.302$, $\sigma = 0.549$.

Validation Set



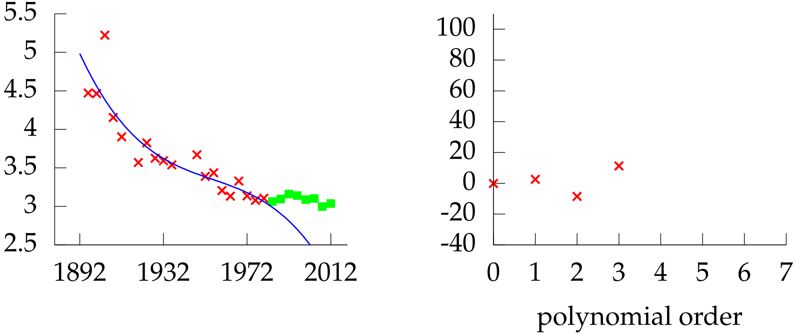
Left: fit to data, *Right*: model error. Polynomial order 1, training error -15.325, validation error 2.5863, $\sigma^2 = 0.0733$, $\sigma = 0.271$.

Validation Set



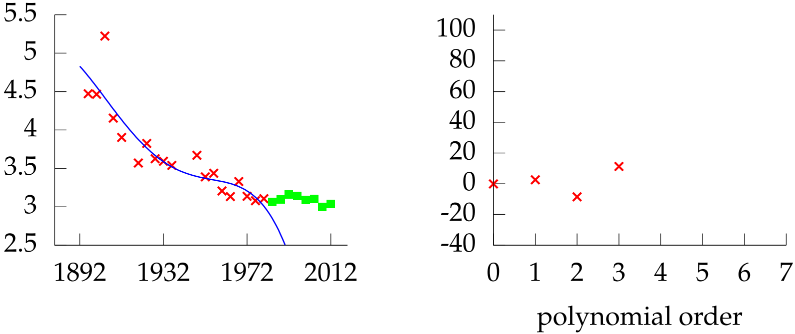
Left: fit to data, *Right*: model error. Polynomial order 2, training error -17.579, validation error -8.4831, $\sigma^2 = 0.0578$, $\sigma = 0.240$.

Validation Set



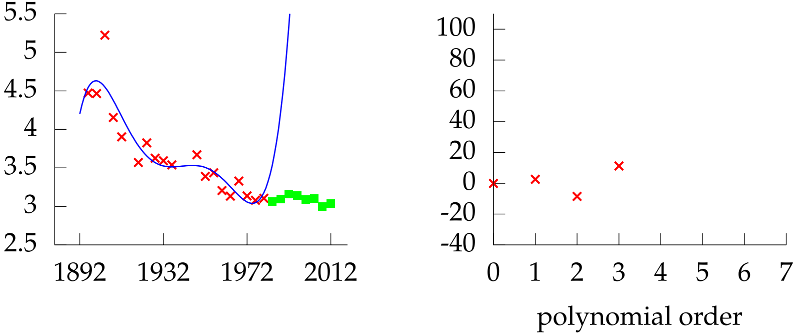
Left: fit to data, *Right*: model error. Polynomial order 3, training error -18.064, validation error 11.27, $\sigma^2 = 0.0549$, $\sigma = 0.234$.

Validation Set



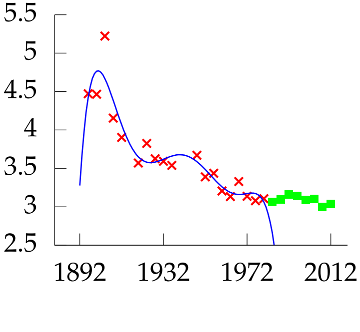
Left: fit to data, *Right*: model error. Polynomial order 4, training error -18.245, validation error 232.92, $\sigma^2 = 0.0539$, $\sigma = 0.232$.

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

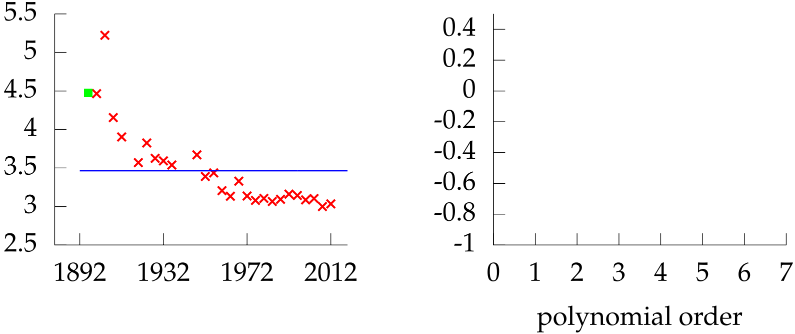


Left: fit to data, *Right*: model error. Polynomial order 6, training error -22.881, validation error 67775, σ^2 = 0.0331, σ = 0.182.

Leave One Out Error

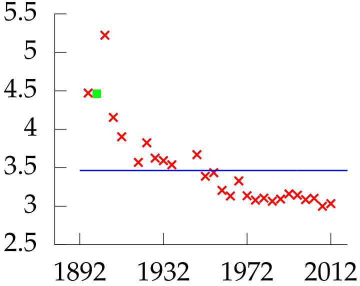
- Take training set and remove one point.
- Train on the remaining data.
- Compute the error on the point you removed (which wasn't in the training data).
- Do this for each point in the training set in turn.
- Average the resulting error. This is the leave one out error.

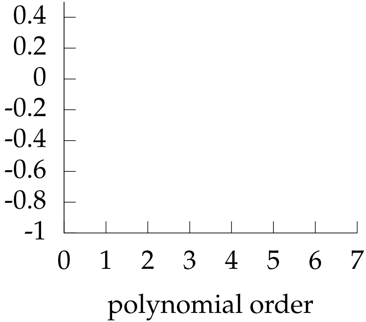
Leave One Out Error



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

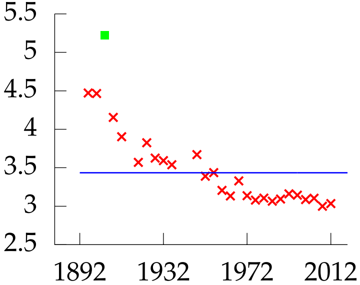
Leave One Out Error

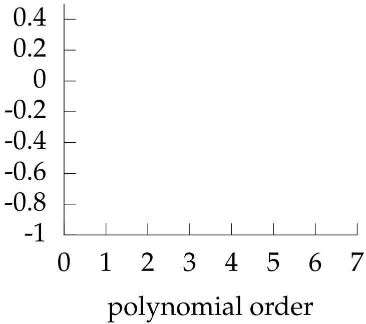




Polynomial order 0, training error -3.346, leave one out error 0.045811.

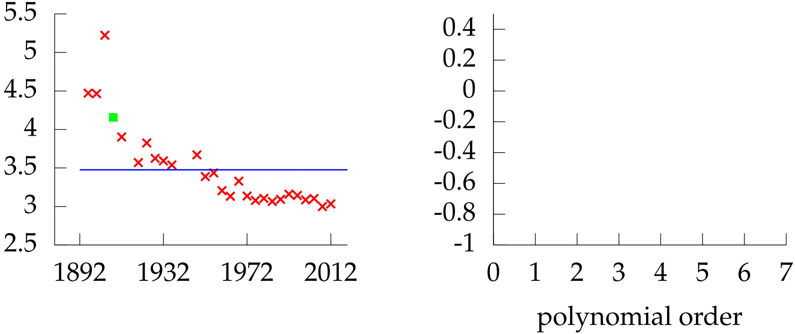
Leave One Out Error





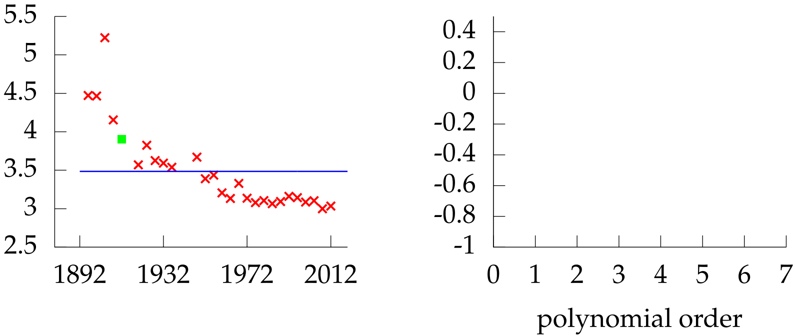
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



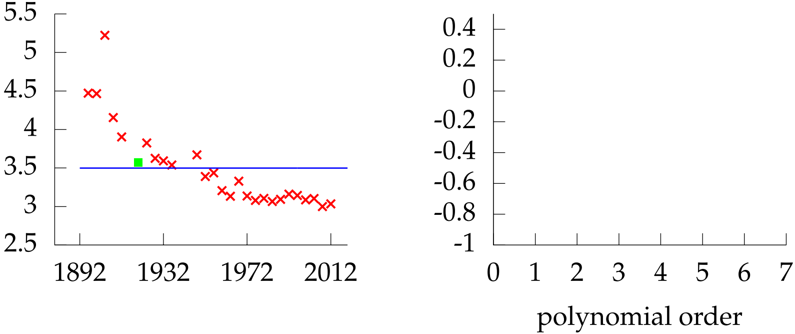
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



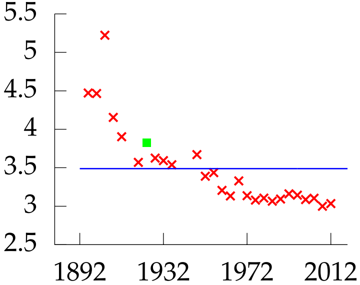
Polynomial order 0, training error -3.346, leave one out error 0.045811.

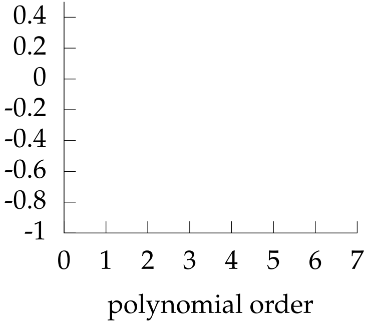
Leave One Out Error



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

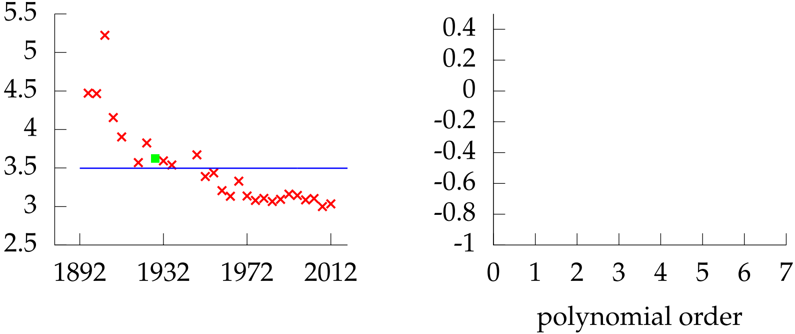
Leave One Out Error





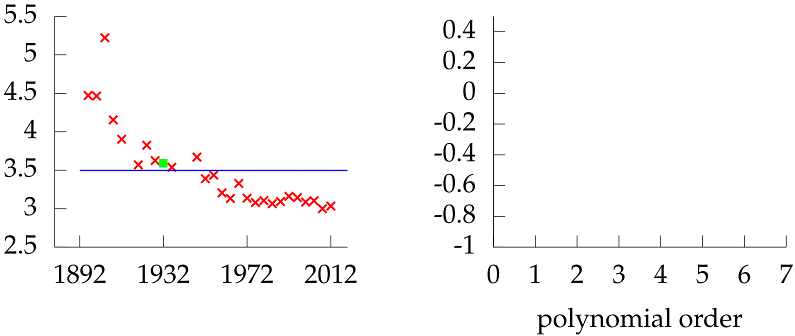
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



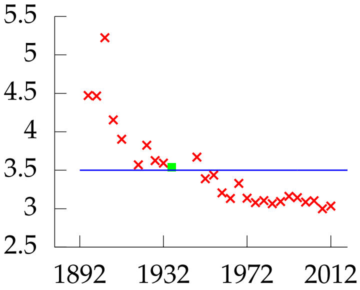
Polynomial order 0, training error -3.346, leave one out error 0.045811.

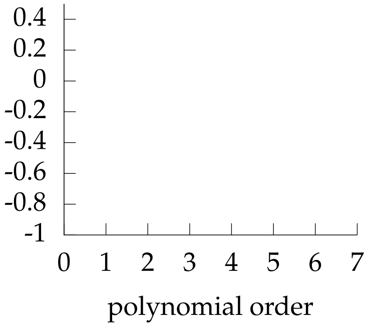
Leave One Out Error



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

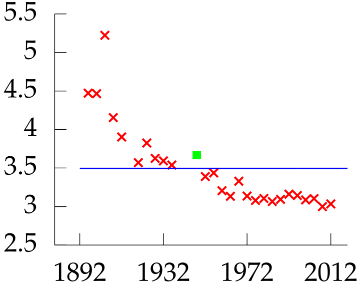
Leave One Out Error

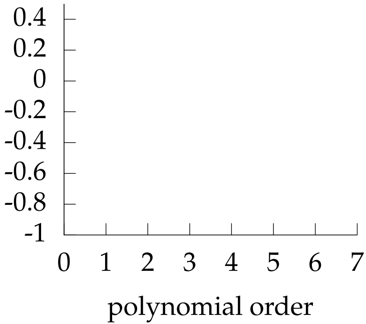




Polynomial order 0, training error -3.346, leave one out error 0.045811.

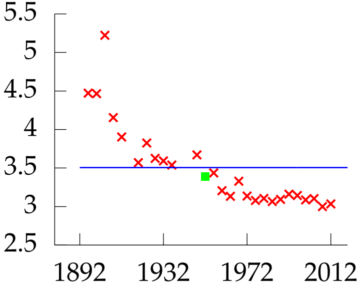
Leave One Out Error

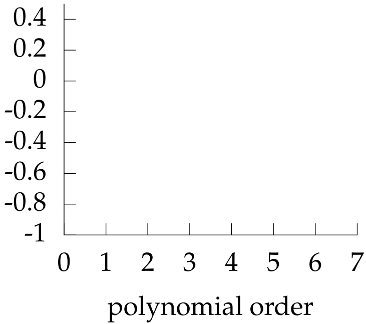




Polynomial order 0, training error -3.346, leave one out error 0.045811.

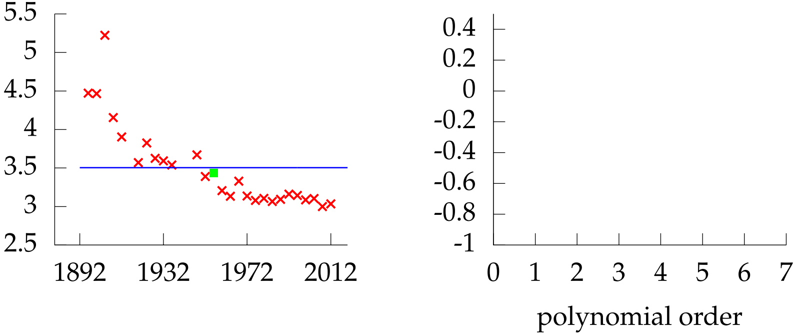
Leave One Out Error





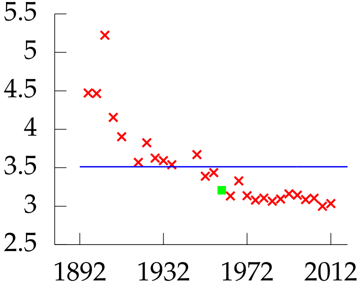
Polynomial order 0, training error -3.346, leave one out error 0.045811.

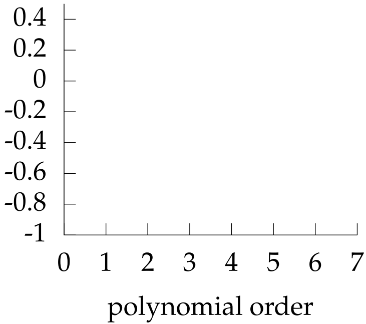
Leave One Out Error



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

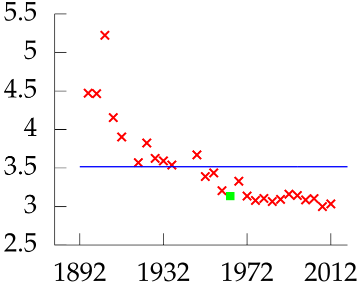
Leave One Out Error

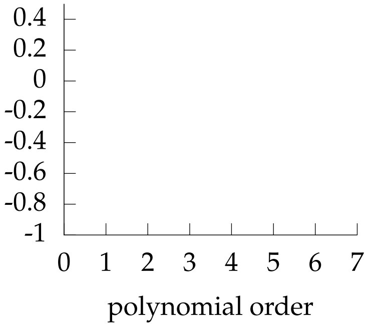




Polynomial order 0, training error -3.346, leave one out error 0.045811.

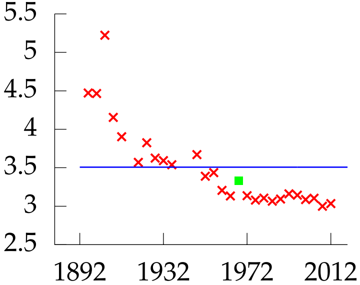
Leave One Out Error

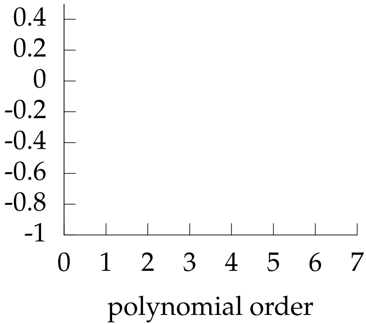




Polynomial order 0, training error -3.346, leave one out error 0.045811.

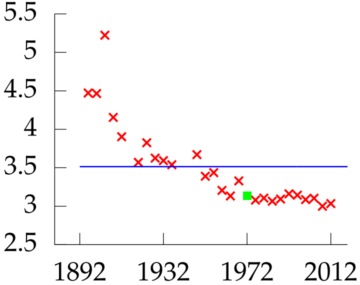
Leave One Out Error

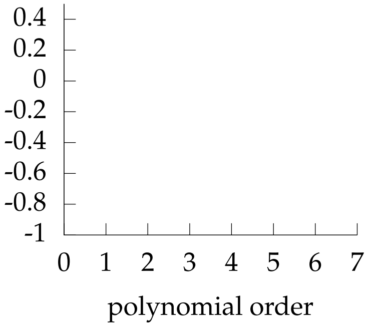




Polynomial order 0, training error -3.346, leave one out error 0.045811.

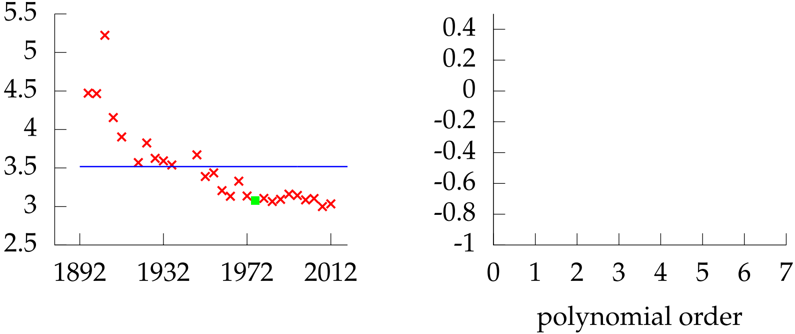
Leave One Out Error





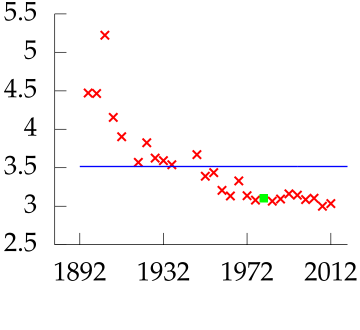
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



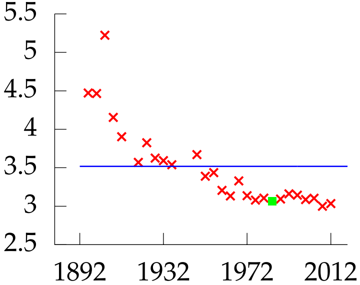
Polynomial order 0, training error -3.346, leave one out error 0.045811.

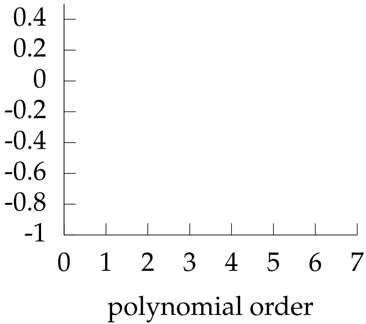
Leave One Out Error



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

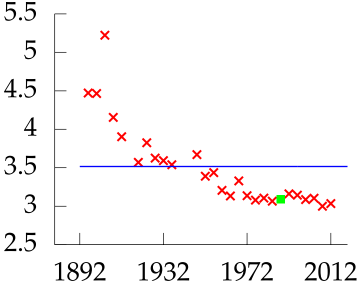
Leave One Out Error

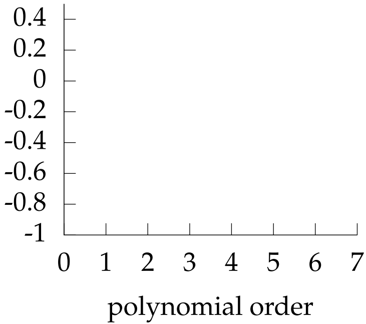




Polynomial order 0, training error -3.346, leave one out error 0.045811.

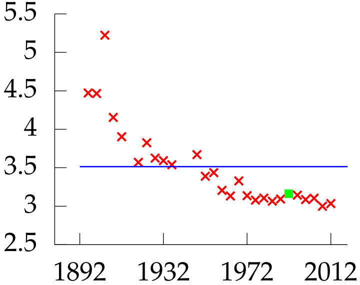
Leave One Out Error

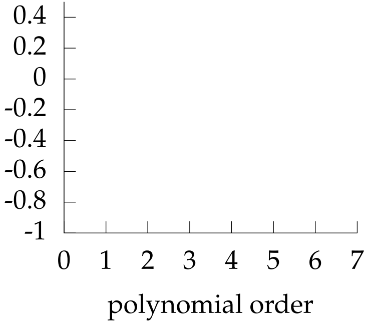




Polynomial order 0, training error -3.346, leave one out error 0.045811.

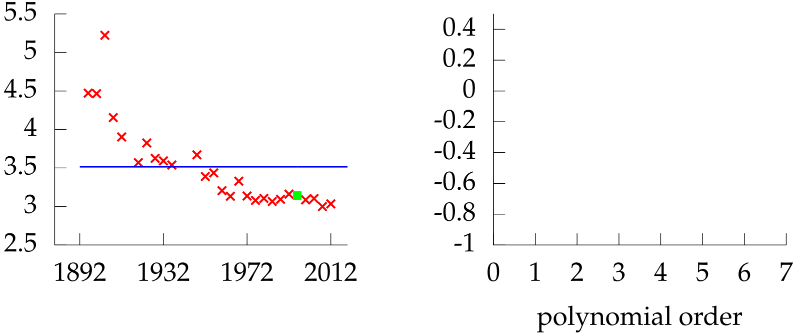
Leave One Out Error





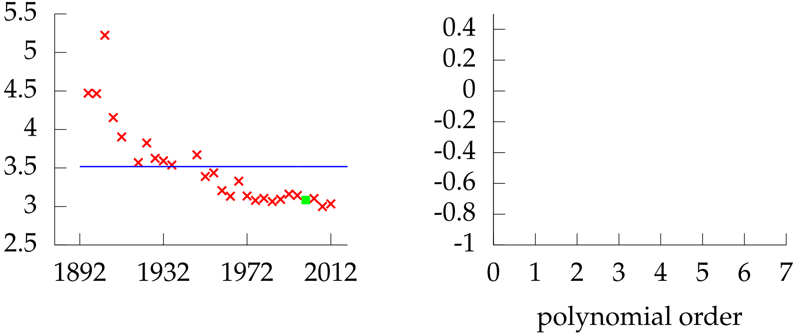
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



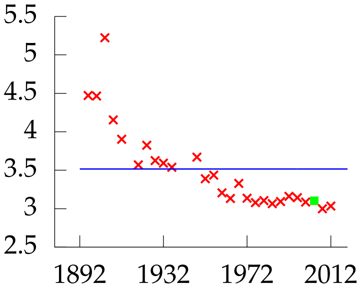
Polynomial order 0, training error -3.346, leave one out error 0.045811.

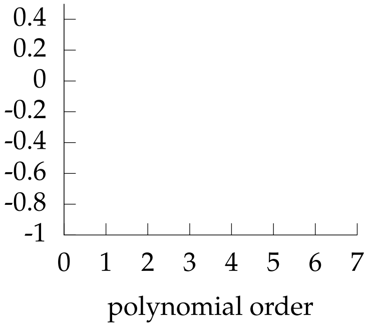
Leave One Out Error



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

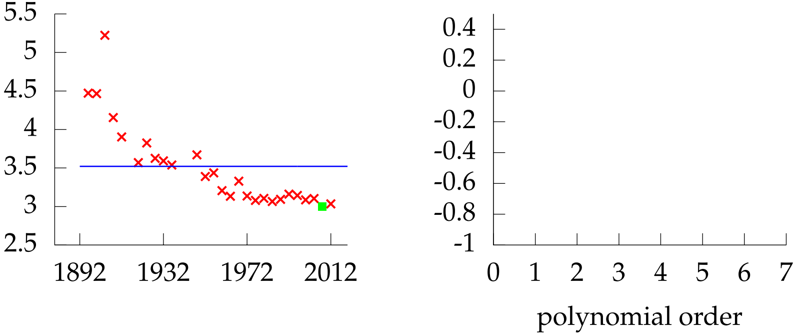
Leave One Out Error





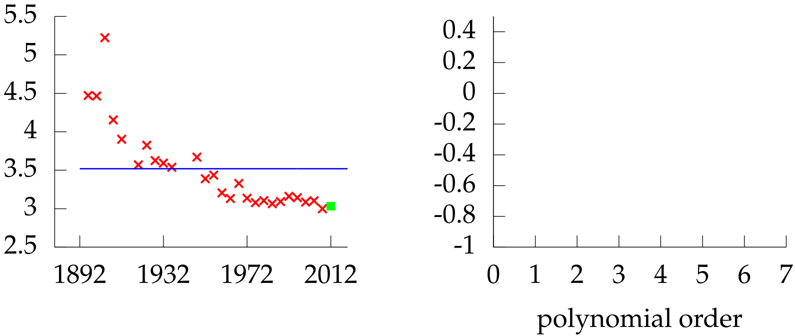
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



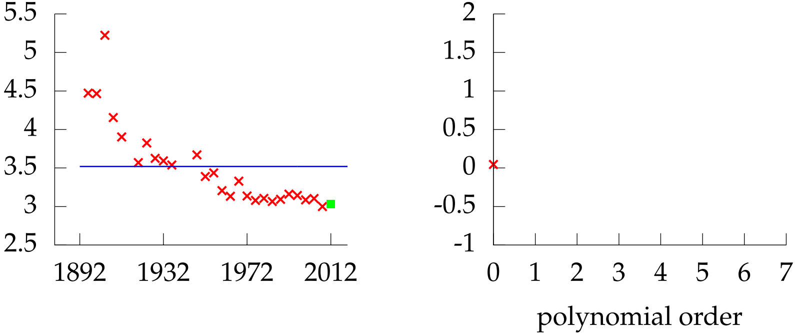
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



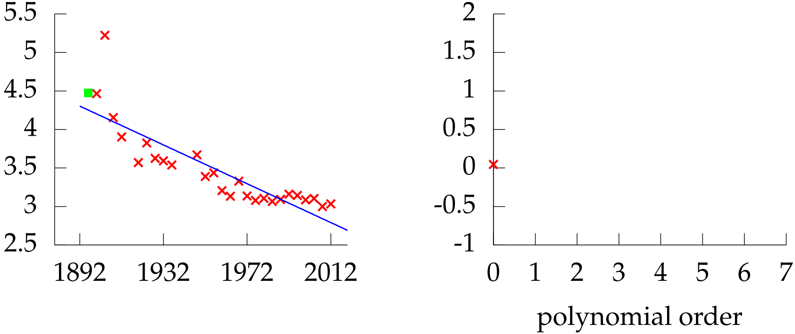
Polynomial order 0, training error -3.346, leave one out error 0.045811.

Leave One Out Error



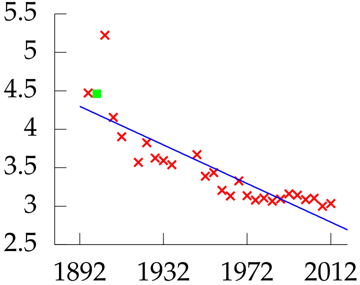
Polynomial order 0, training error -3.346, leave one out error 0.045811.

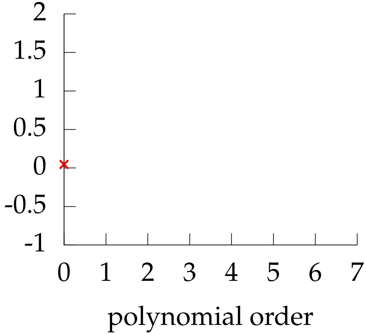
Leave One Out Error



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

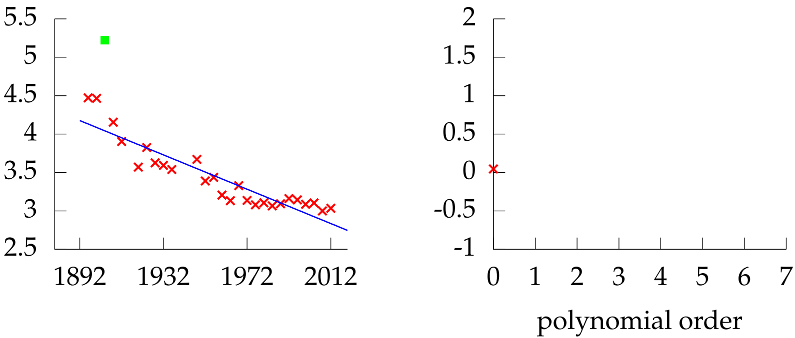
Leave One Out Error





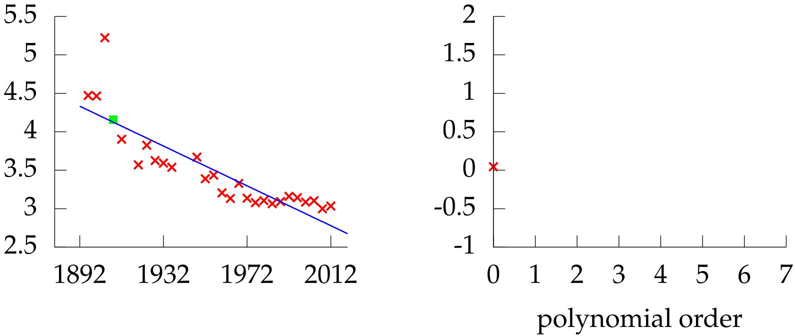
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



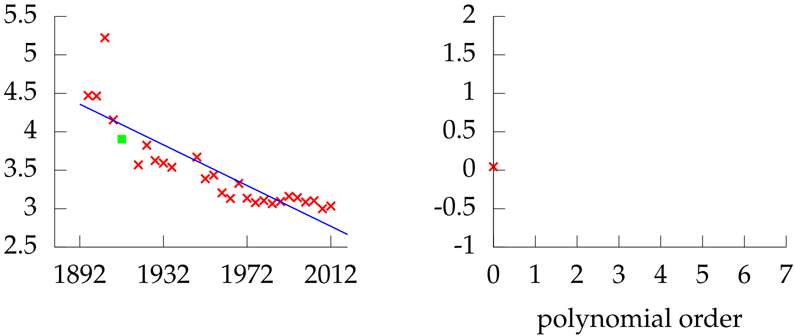
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



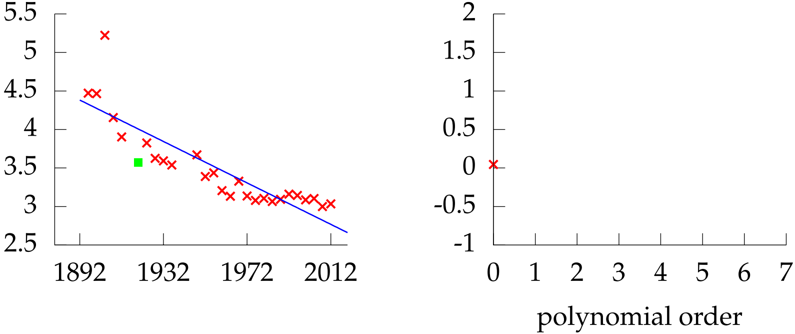
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



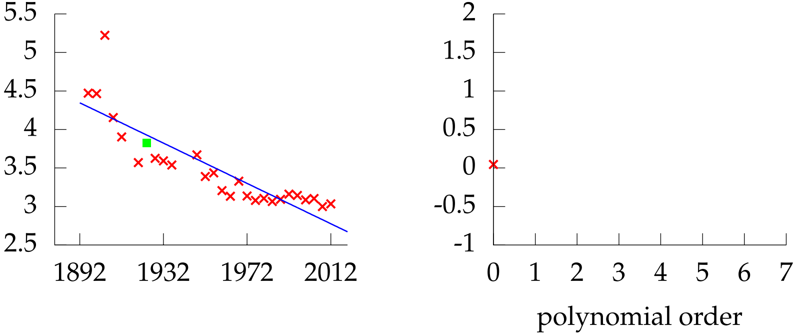
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



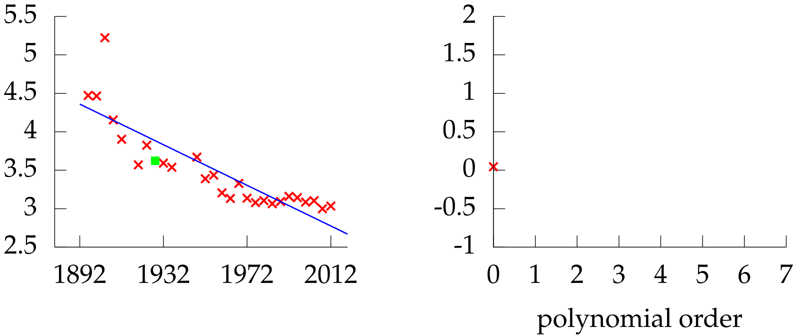
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



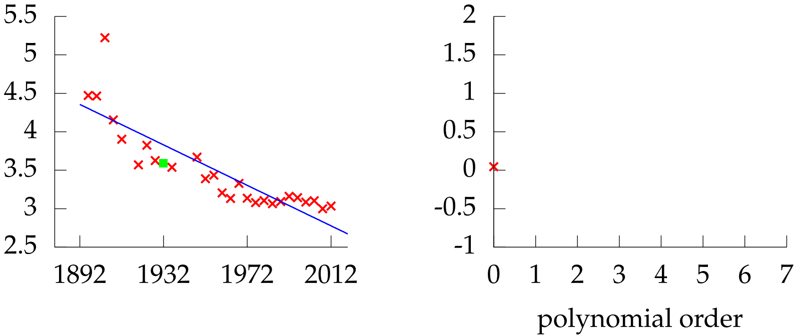
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



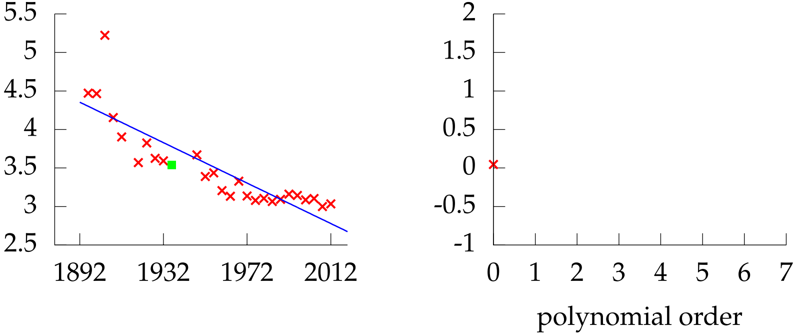
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



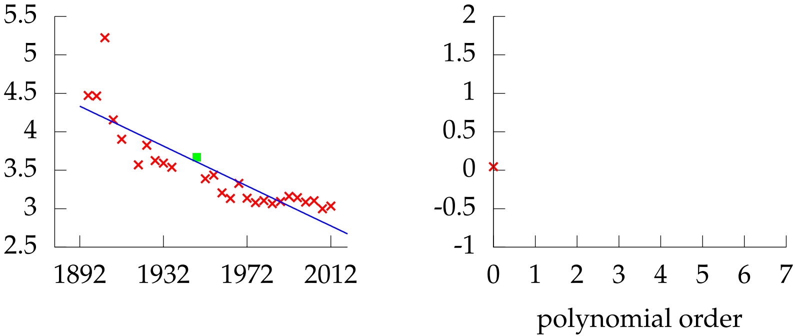
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



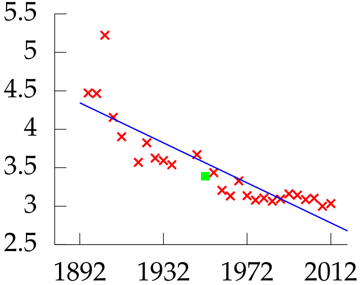
Polynomial order 1, training error -21.183, leave one out error -0.15413.

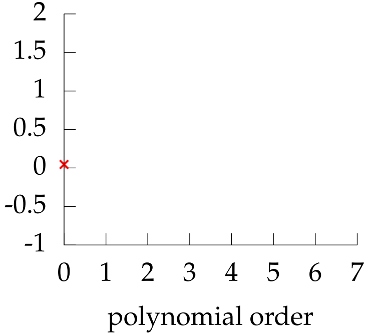
Leave One Out Error



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

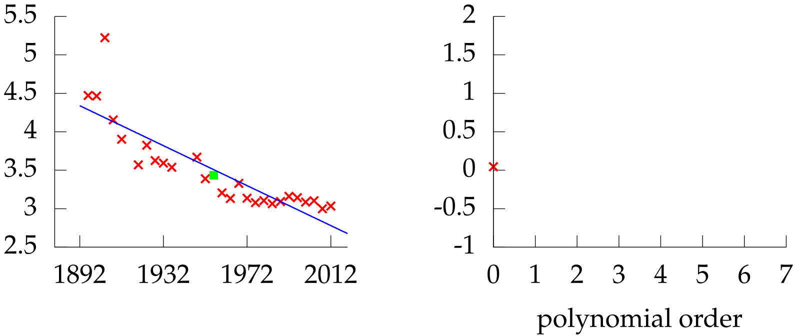
Leave One Out Error





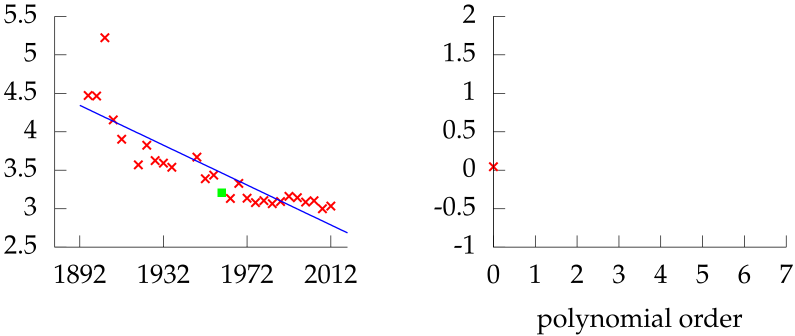
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



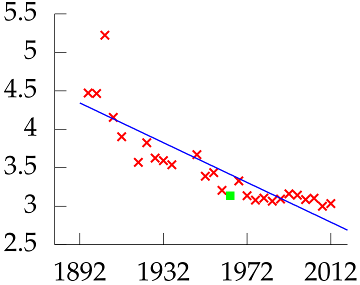
Polynomial order 1, training error -21.183, leave one out error -0.15413.

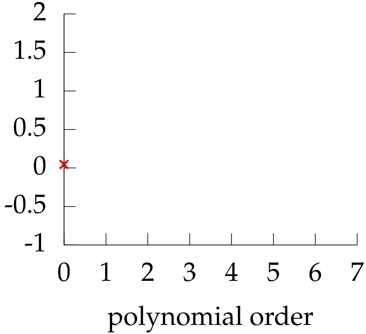
Leave One Out Error



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

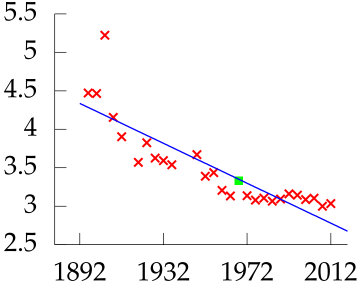
Leave One Out Error

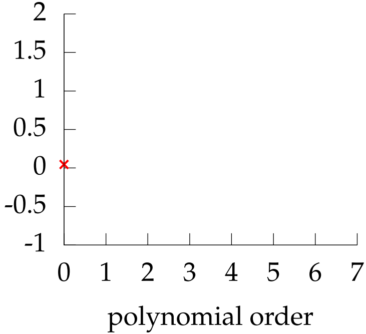




Polynomial order 1, training error -21.183, leave one out error -0.15413.

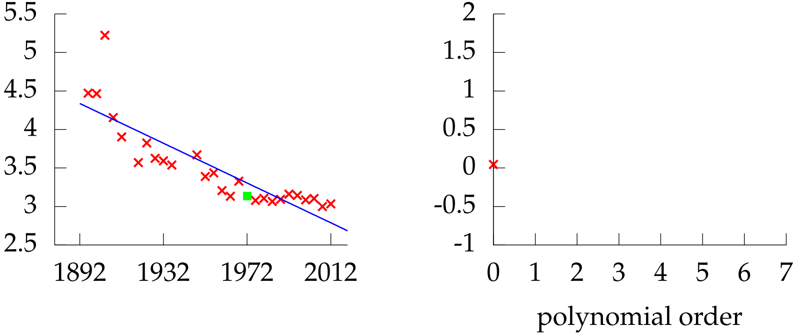
Leave One Out Error





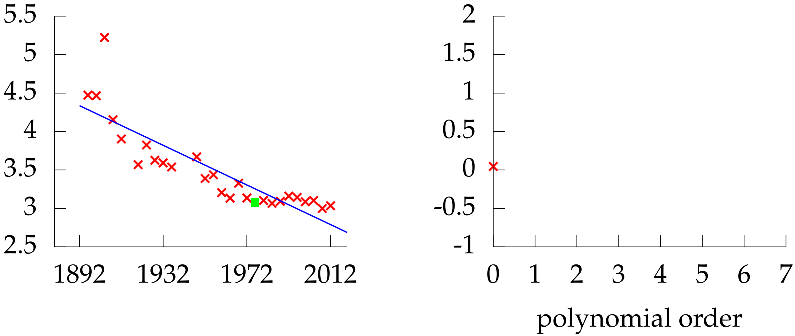
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



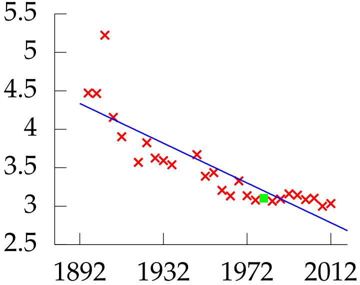
Polynomial order 1, training error -21.183, leave one out error -0.15413.

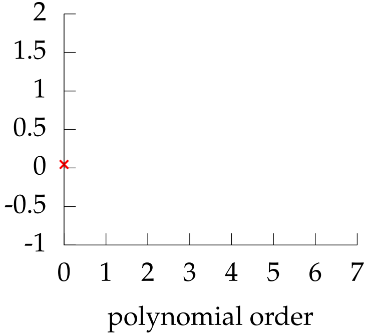
Leave One Out Error



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

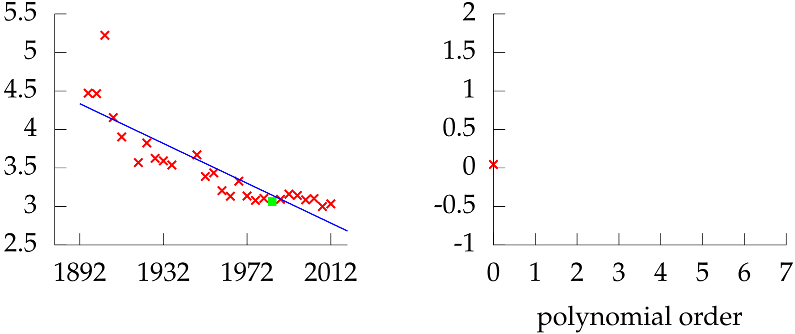
Leave One Out Error





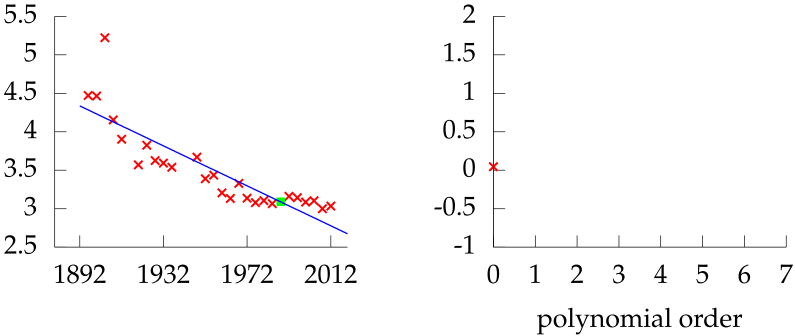
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



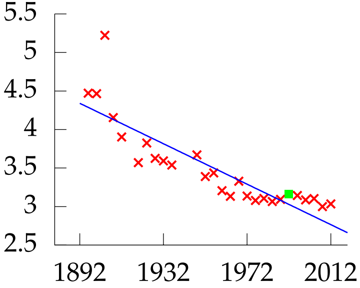
Polynomial order 1, training error -21.183, leave one out error -0.15413.

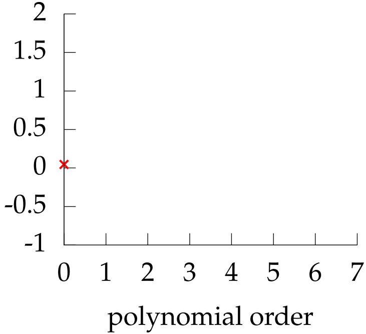
Leave One Out Error



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

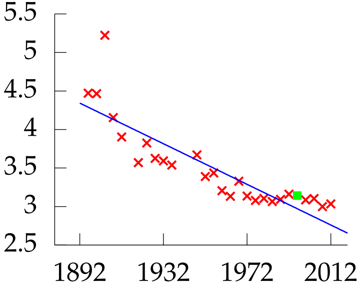
Leave One Out Error

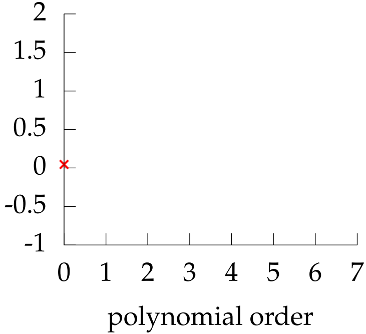




Polynomial order 1, training error -21.183, leave one out error -0.15413.

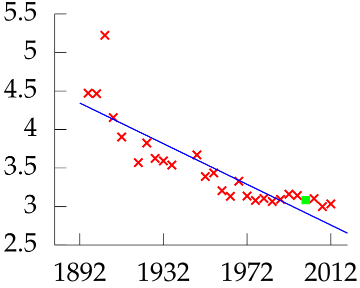
Leave One Out Error

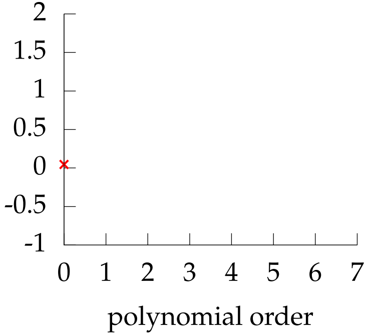




Polynomial order 1, training error -21.183, leave one out error -0.15413.

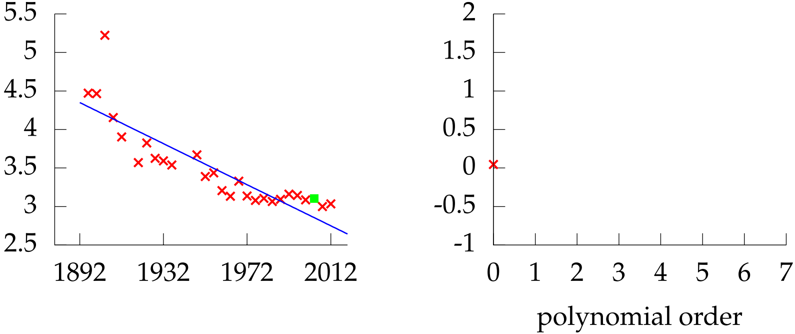
Leave One Out Error





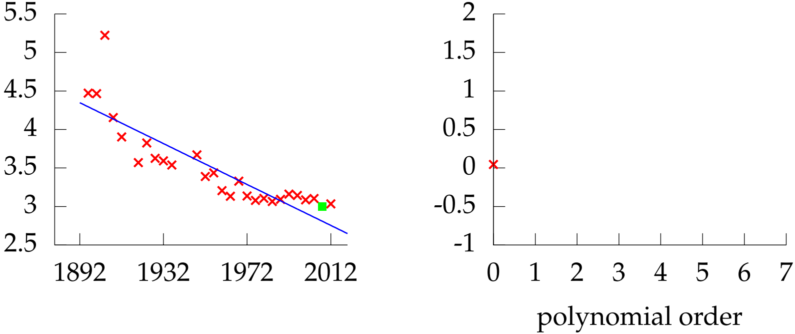
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



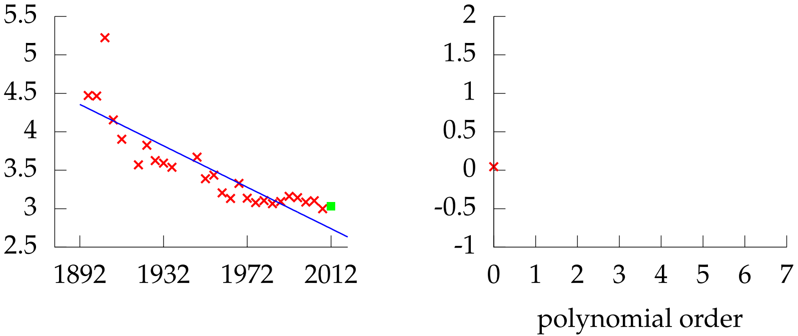
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



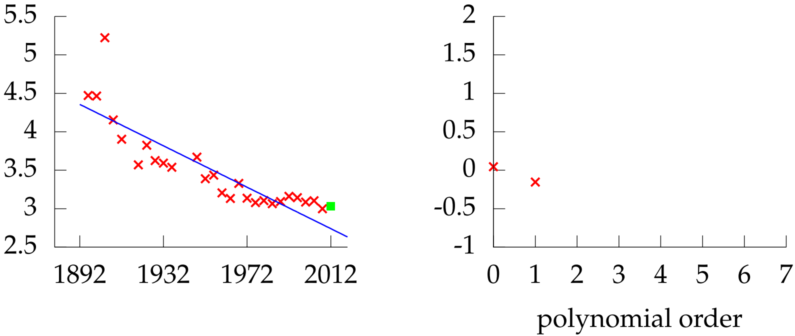
Polynomial order 1, training error -21.183, leave one out error -0.15413.

Leave One Out Error



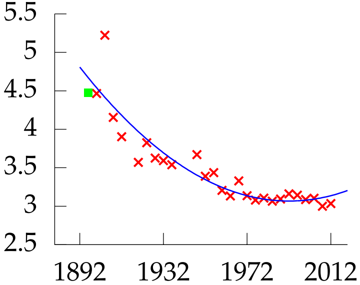
Polynomial order 1, training error -21.183, leave one out error -0.15413.

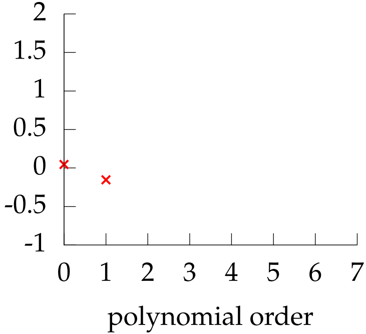
Leave One Out Error



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

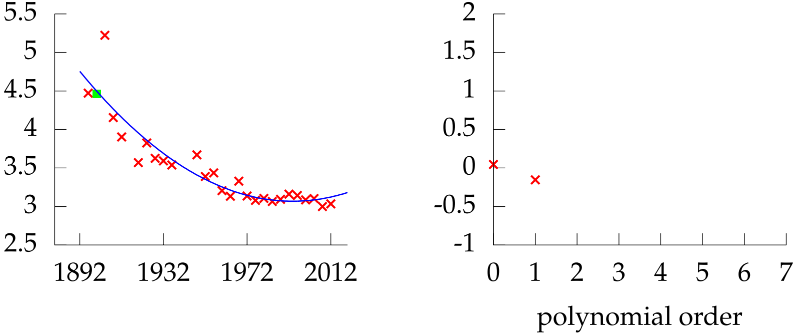
Leave One Out Error





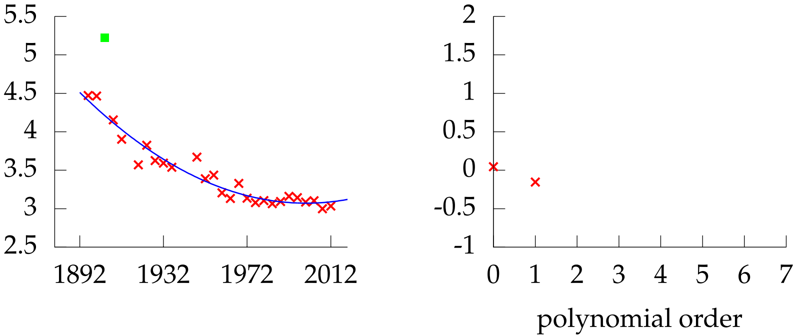
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



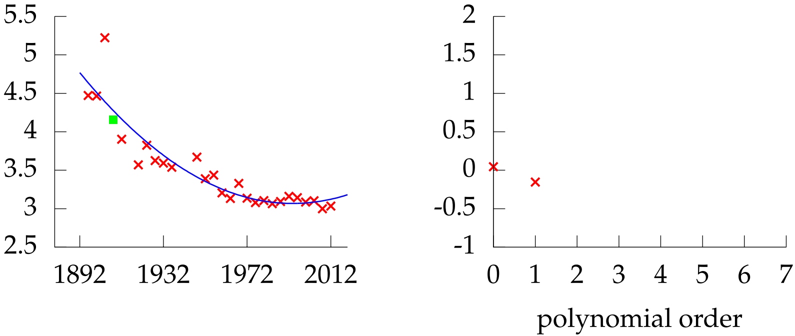
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



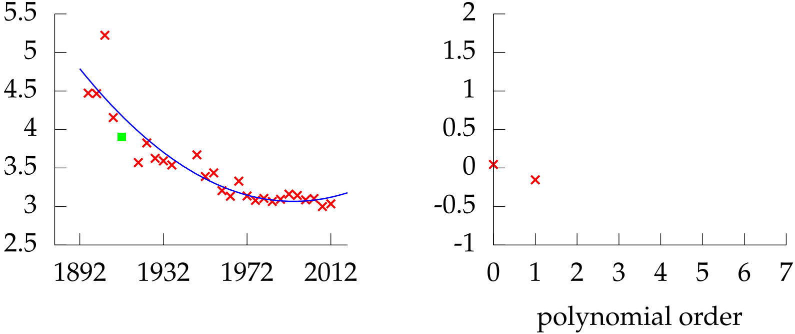
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



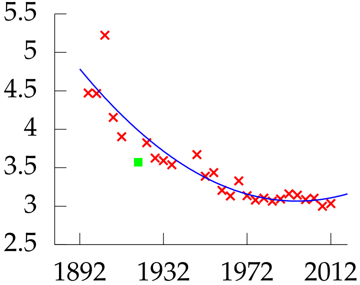
Polynomial order 2, training error -28.403, leave one out error 0.34669.

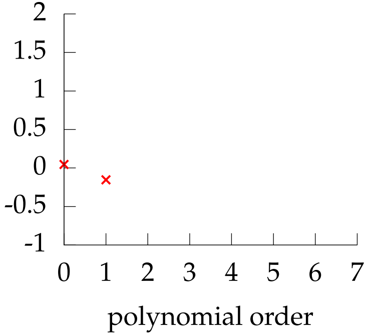
Leave One Out Error



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

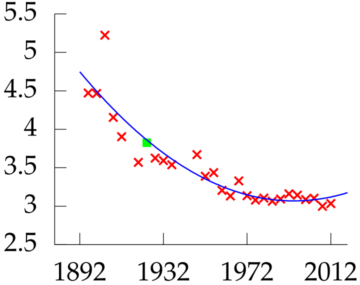
Leave One Out Error

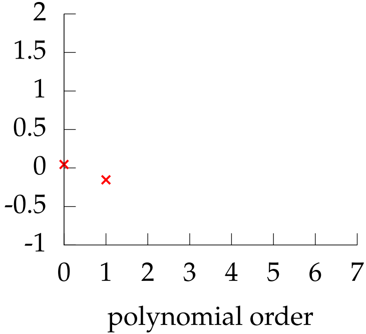




Polynomial order 2, training error -28.403, leave one out error 0.34669.

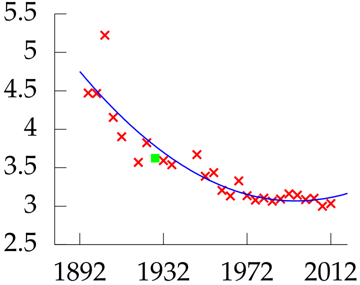
Leave One Out Error

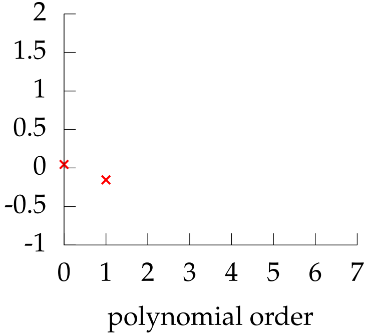




Polynomial order 2, training error -28.403, leave one out error 0.34669.

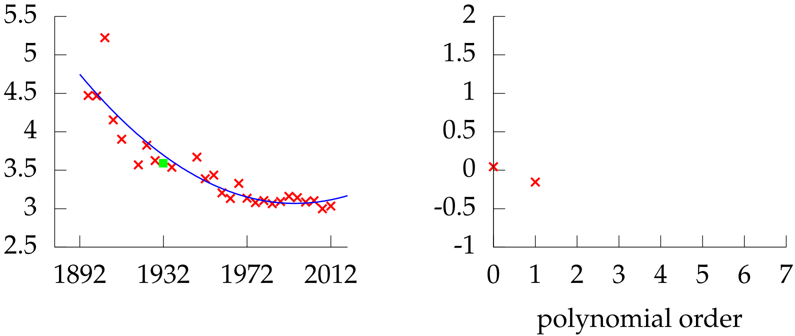
Leave One Out Error





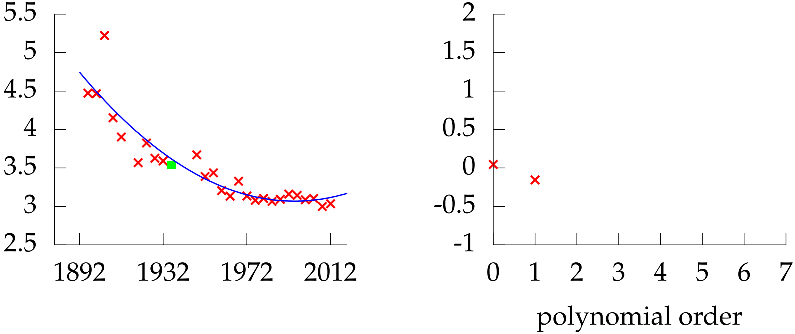
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



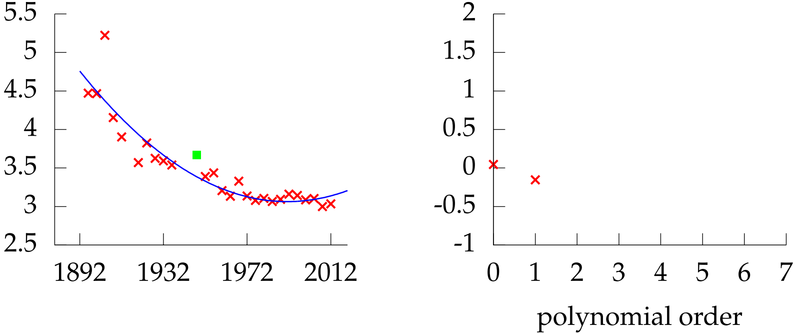
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



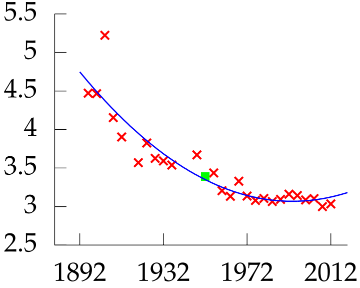
Polynomial order 2, training error -28.403, leave one out error 0.34669.

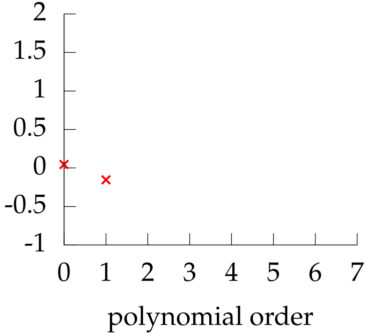
Leave One Out Error



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

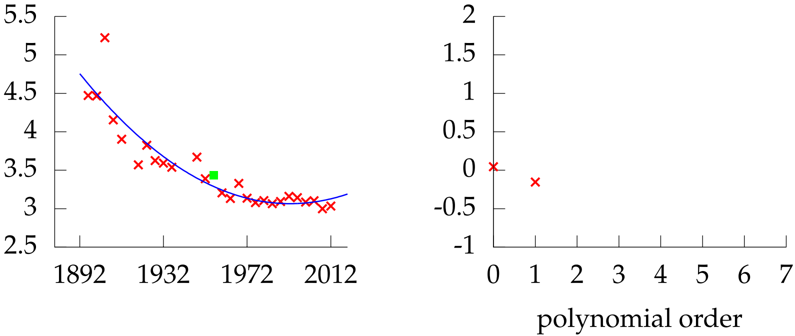
Leave One Out Error





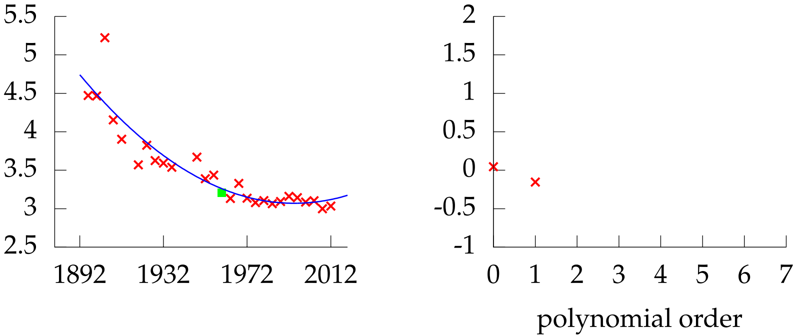
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



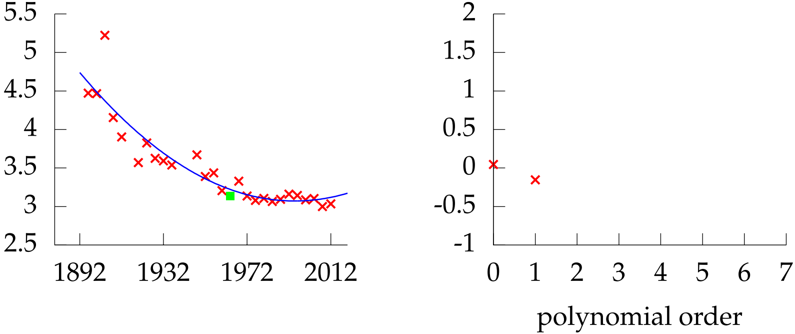
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



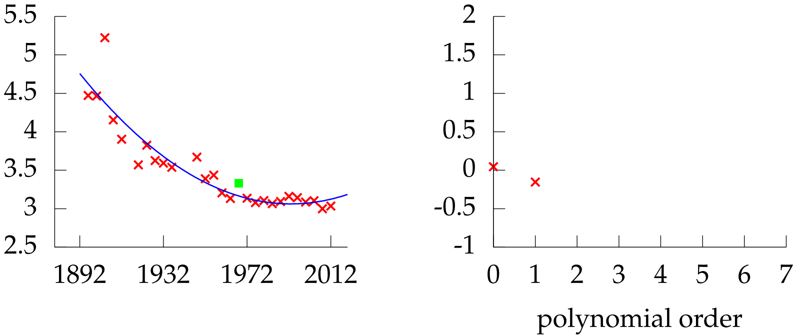
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



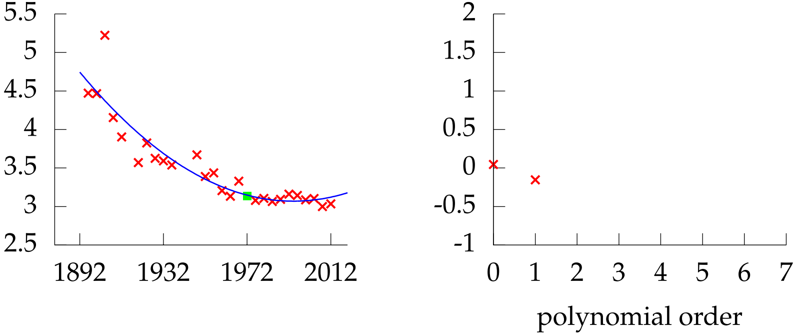
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



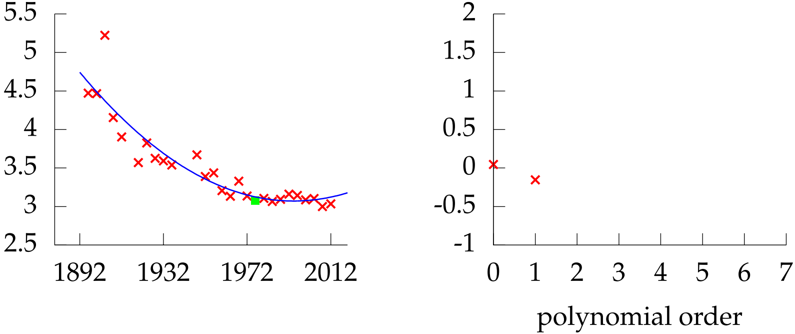
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



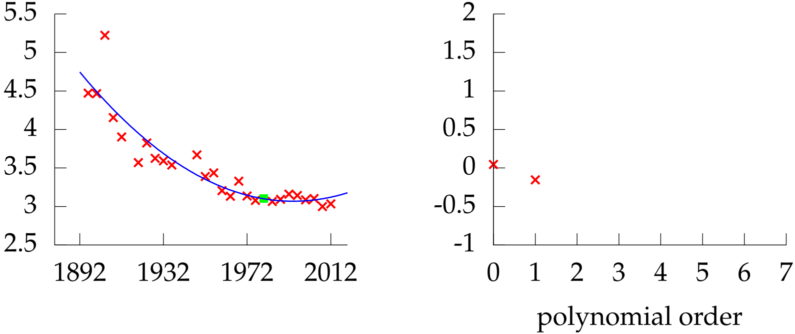
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



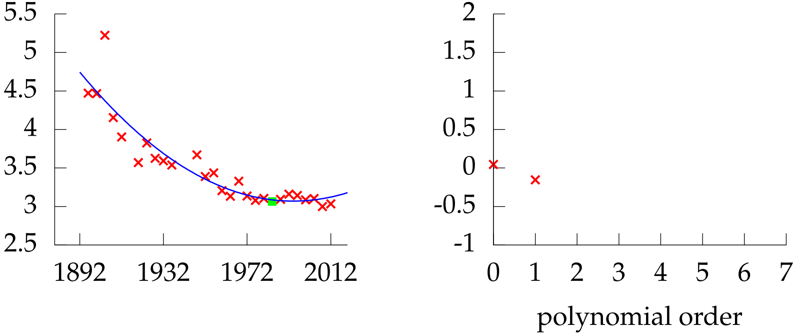
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



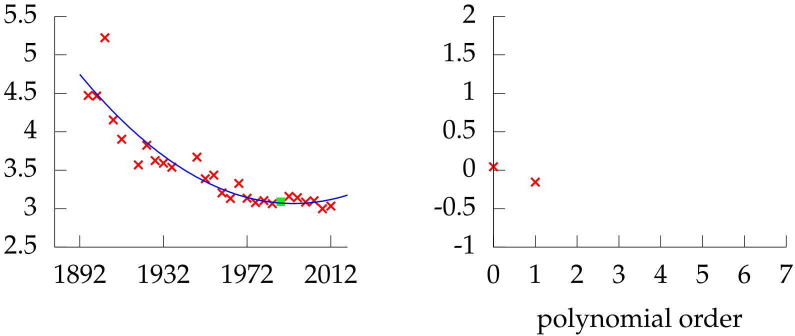
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



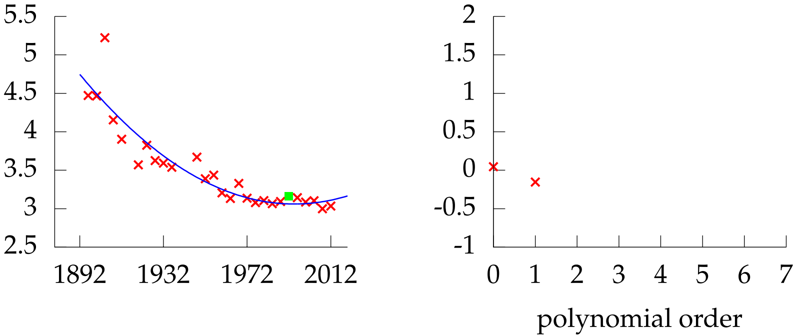
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



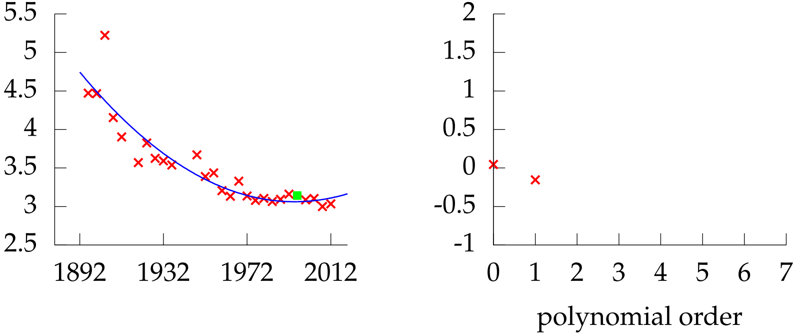
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



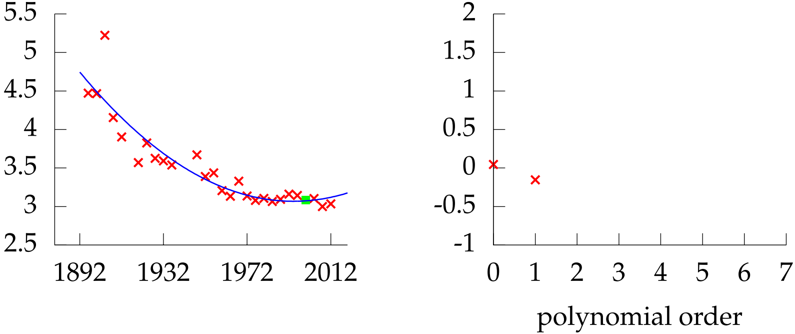
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



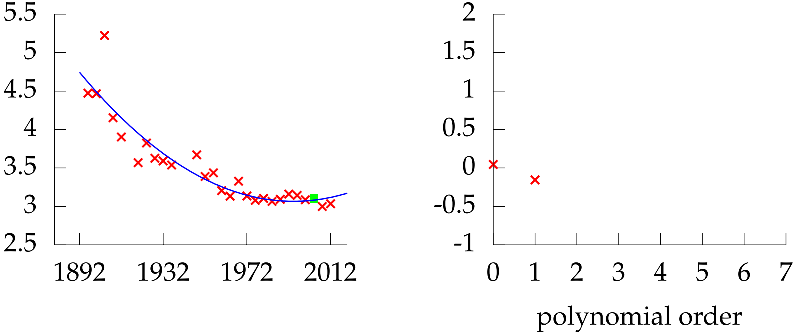
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



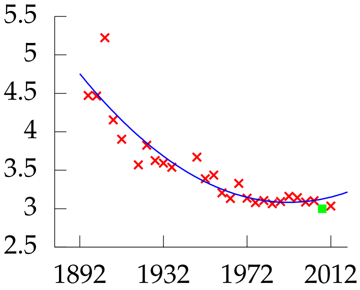
Polynomial order 2, training error -28.403, leave one out error 0.34669.

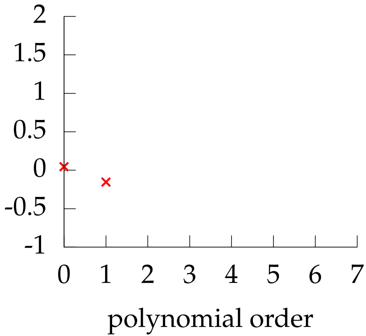
Leave One Out Error



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

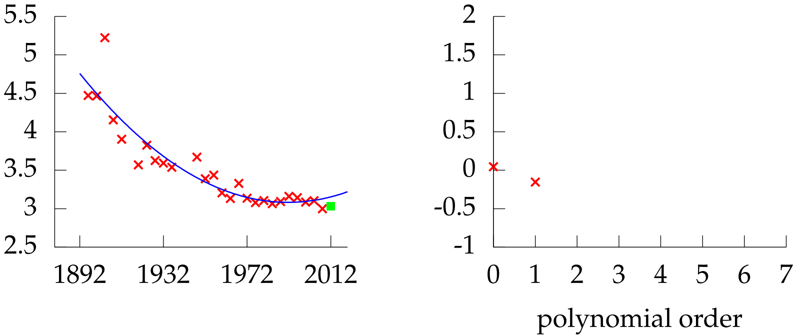
Leave One Out Error





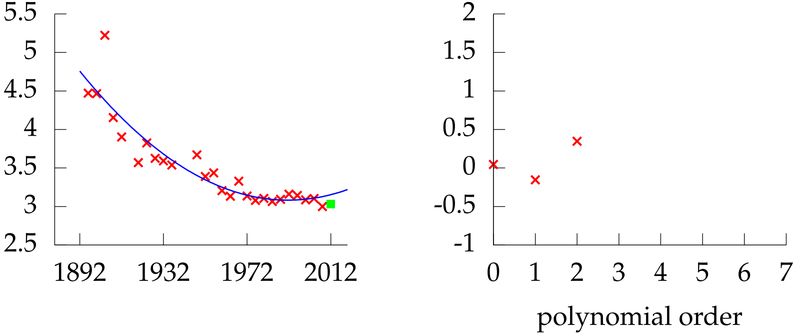
Polynomial order 2, training error -28.403, leave one out error 0.34669.

Leave One Out Error



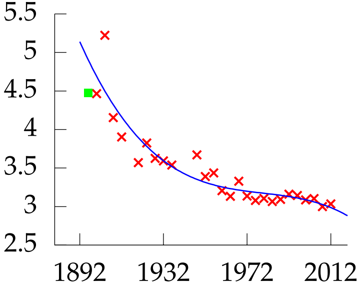
Polynomial order 2, training error -28.403, leave one out error 0.34669.

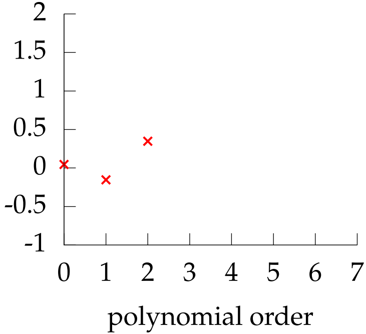
Leave One Out Error



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

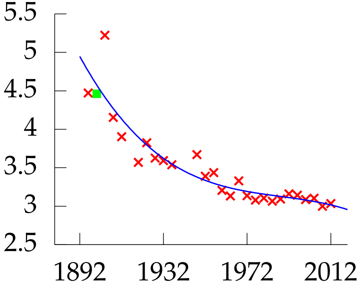
Leave One Out Error

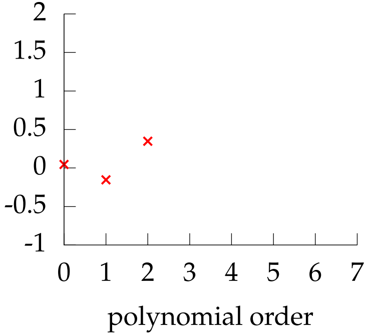




Polynomial order 3, training error -29.223, leave one out error 0.51621.

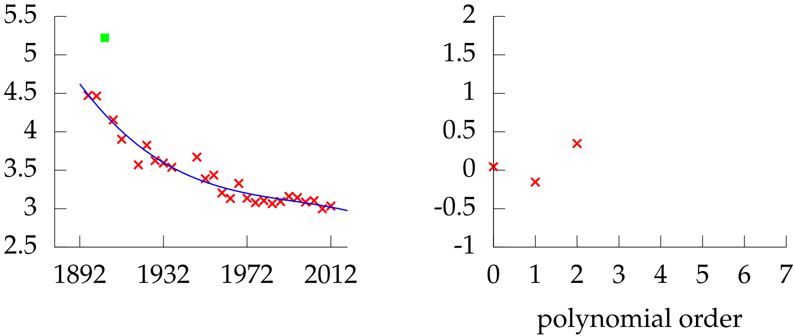
Leave One Out Error





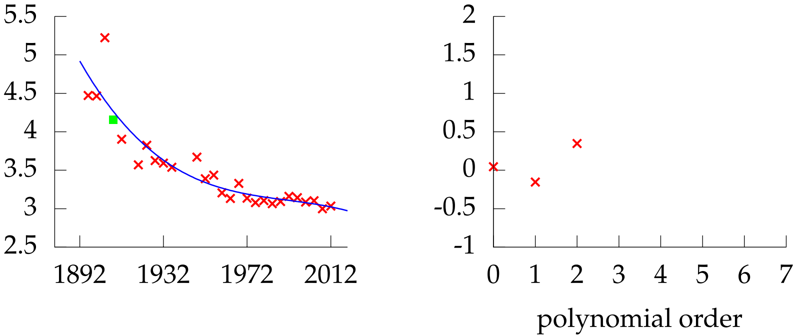
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



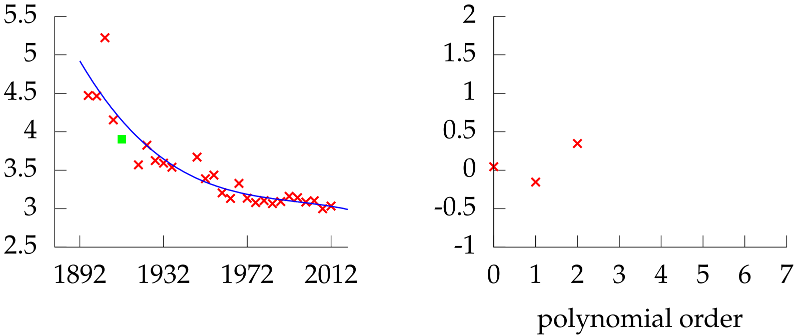
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



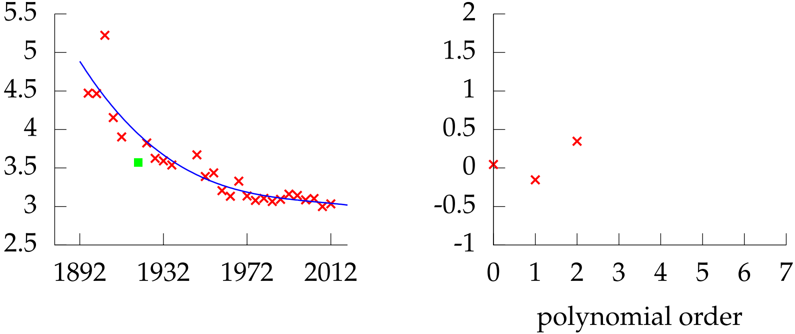
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



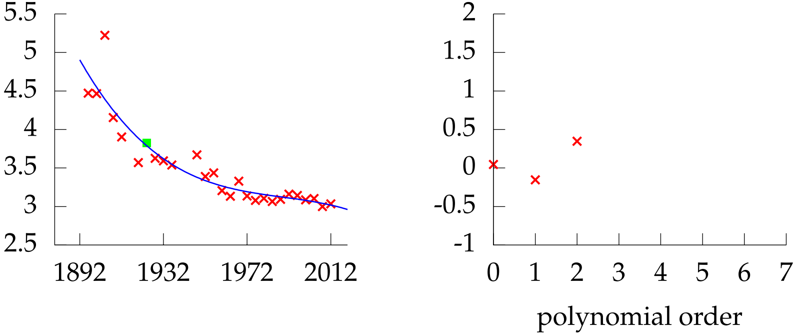
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



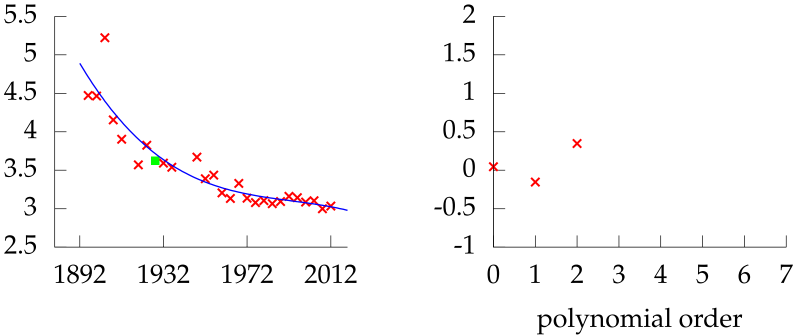
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



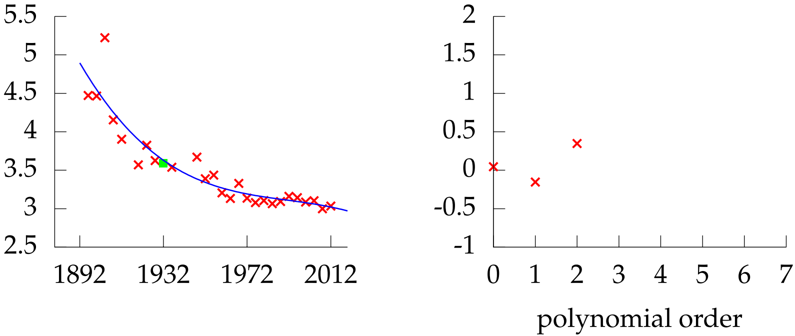
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



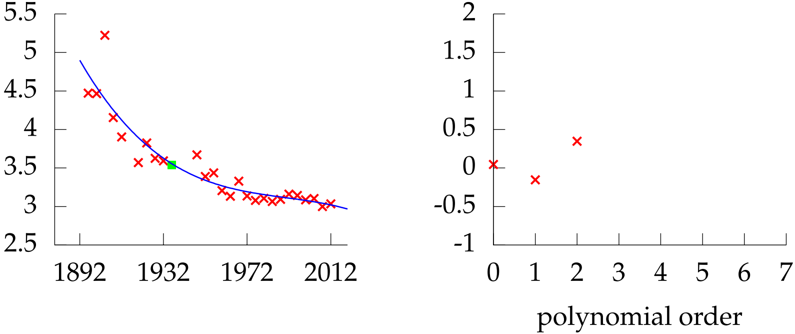
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



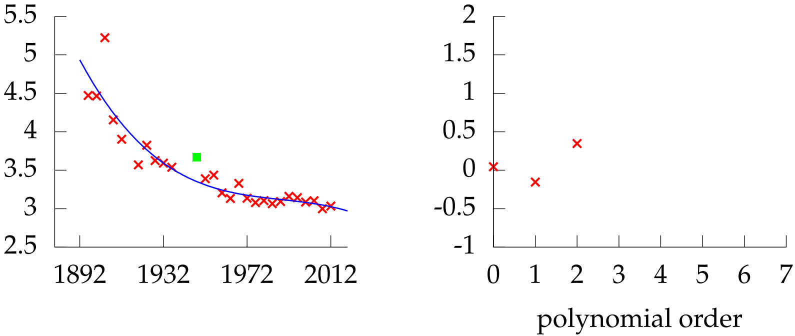
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



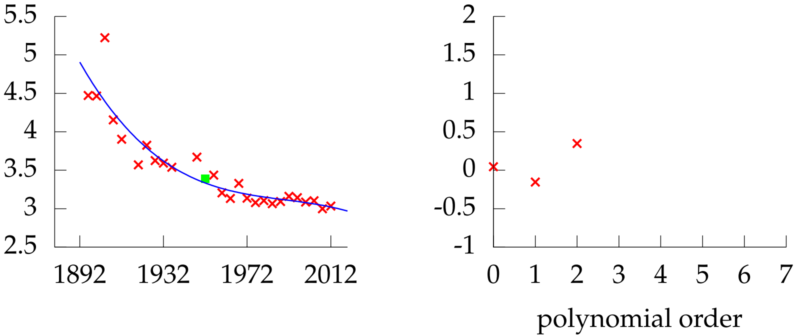
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



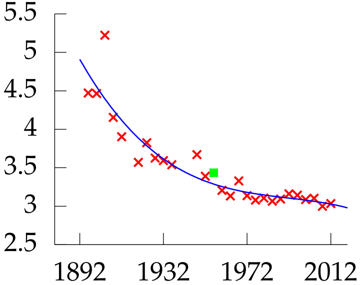
Polynomial order 3, training error -29.223, leave one out error 0.51621.

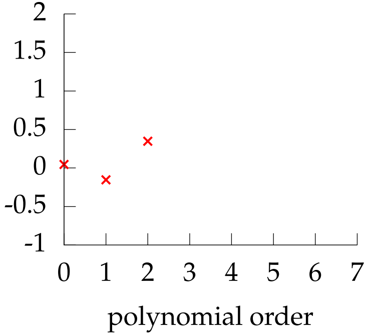
Leave One Out Error



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

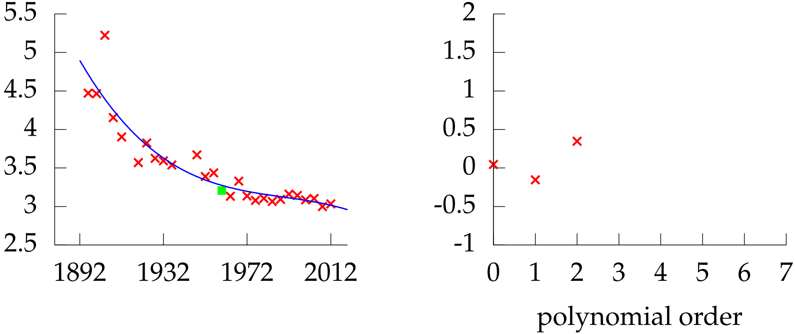
Leave One Out Error





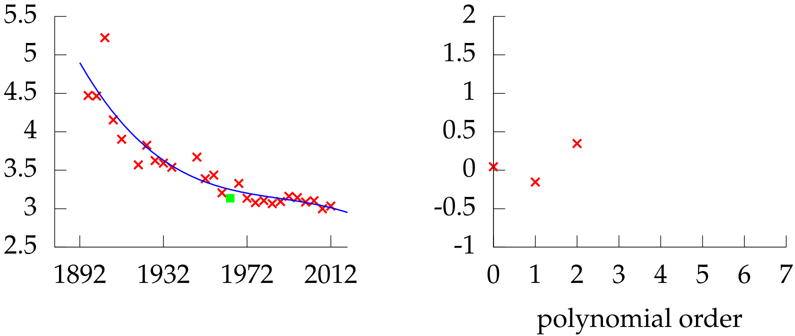
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



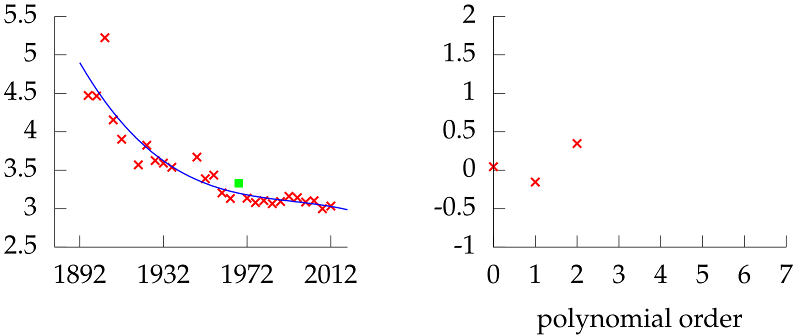
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



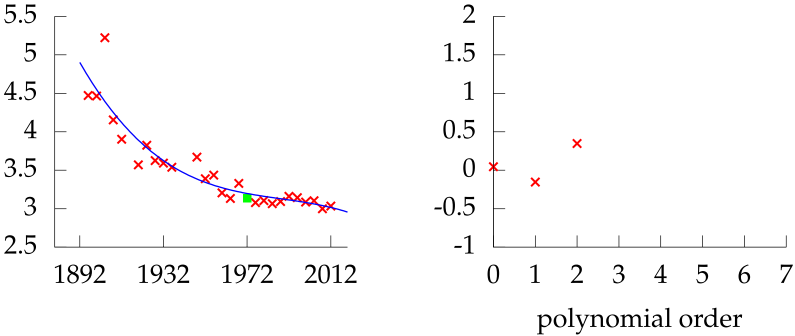
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



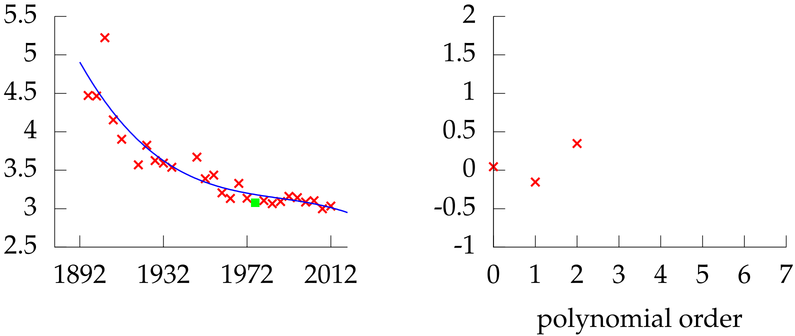
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



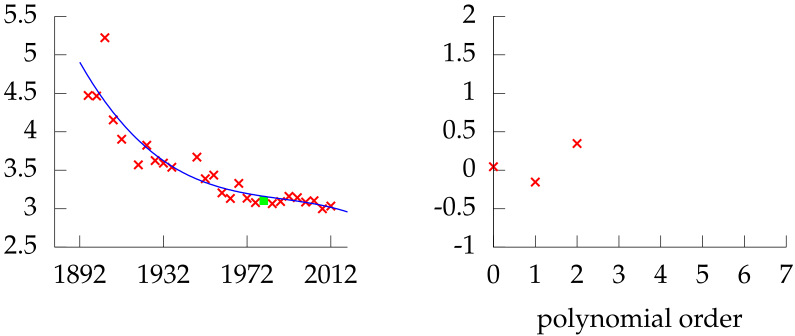
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



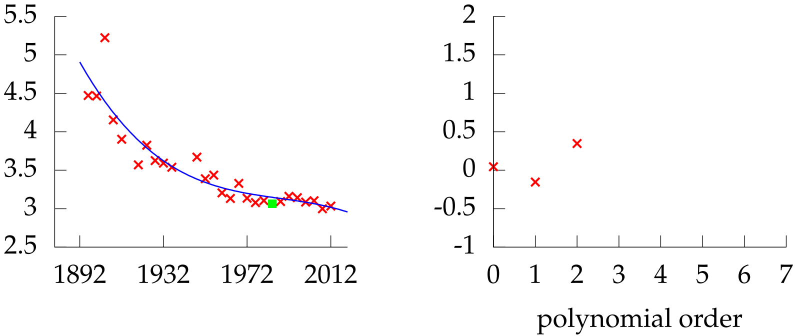
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



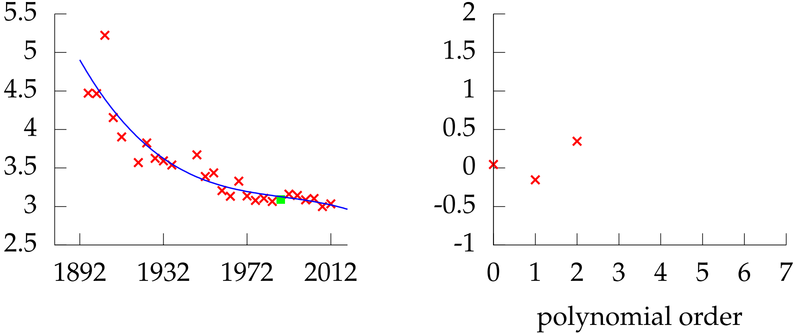
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



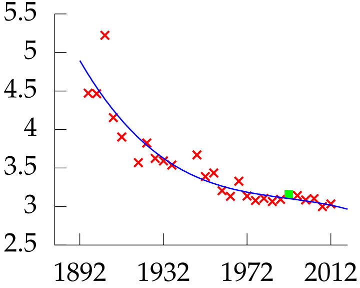
Polynomial order 3, training error -29.223, leave one out error 0.51621.

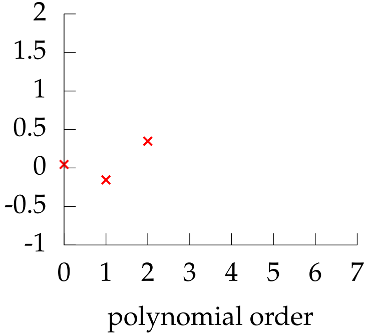
Leave One Out Error



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

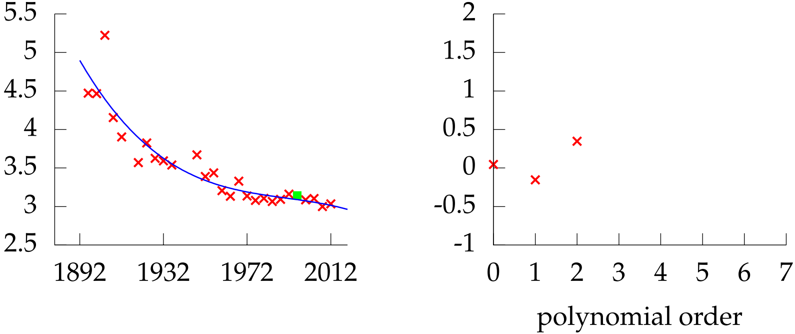
Leave One Out Error





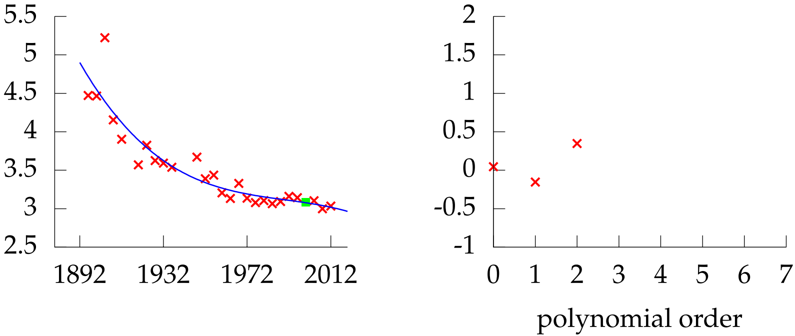
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



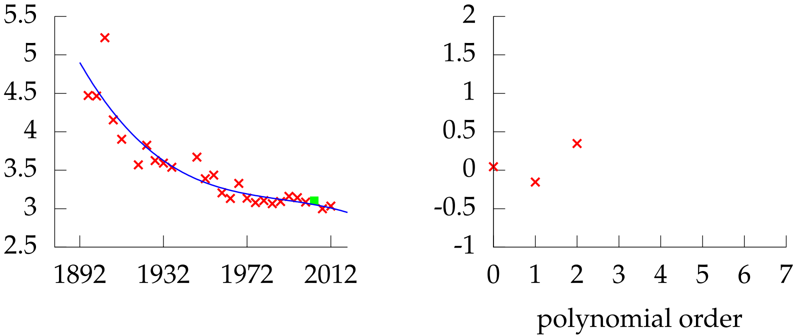
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



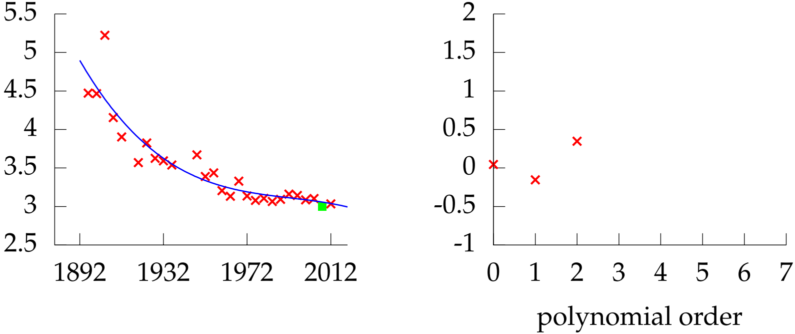
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



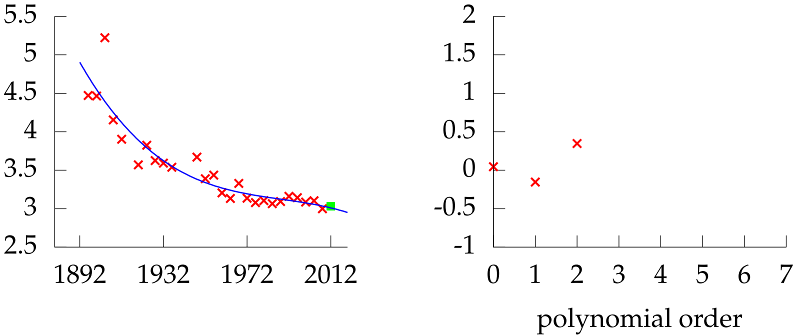
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



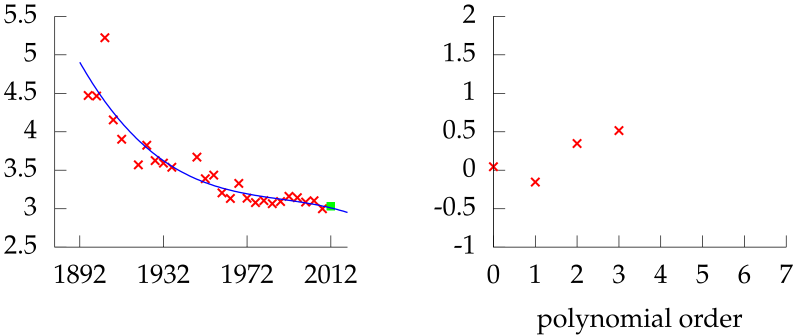
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



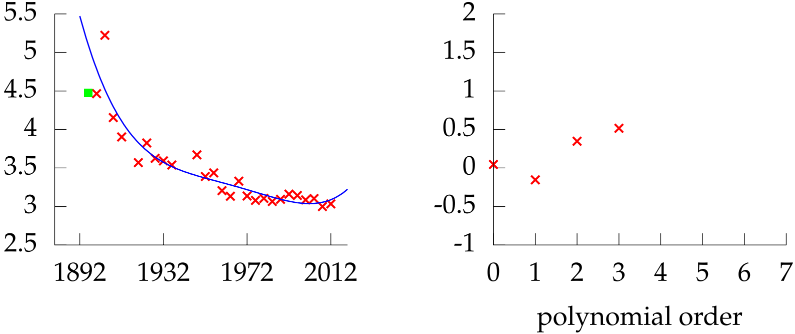
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



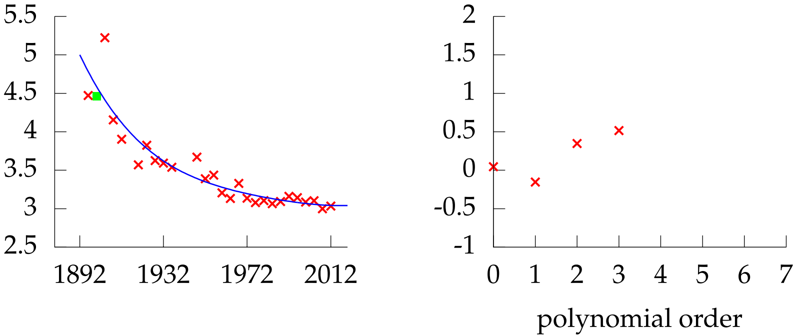
Polynomial order 3, training error -29.223, leave one out error 0.51621.

Leave One Out Error



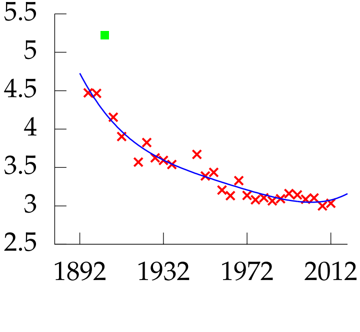
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



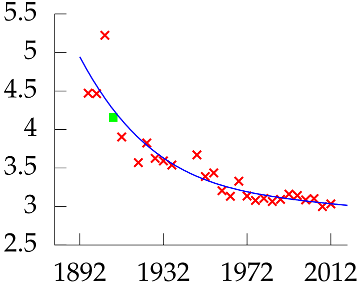
Polynomial order 4, training error -29.324, leave one out error 0.84844.

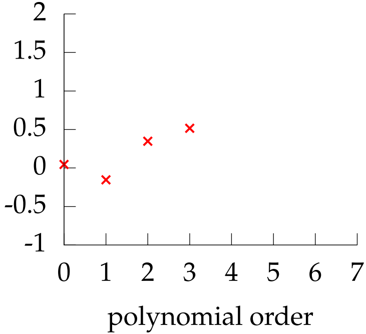
Leave One Out Error



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

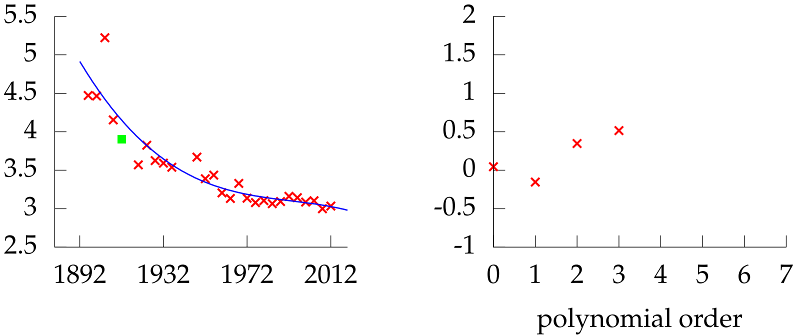
Leave One Out Error





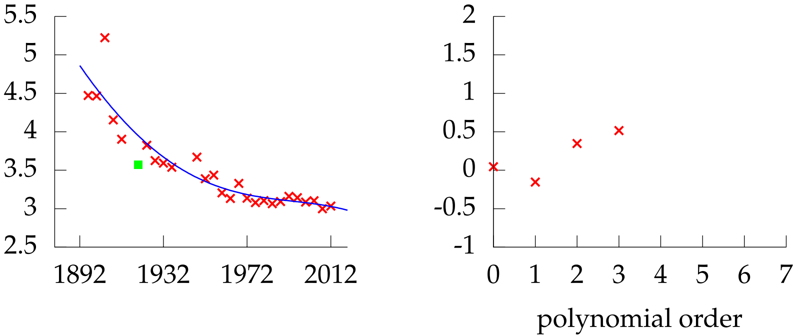
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



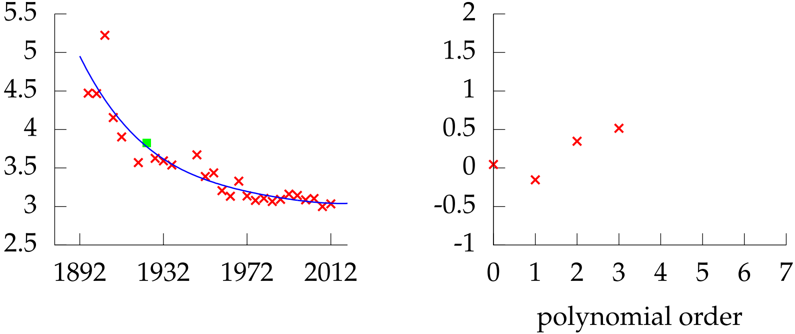
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



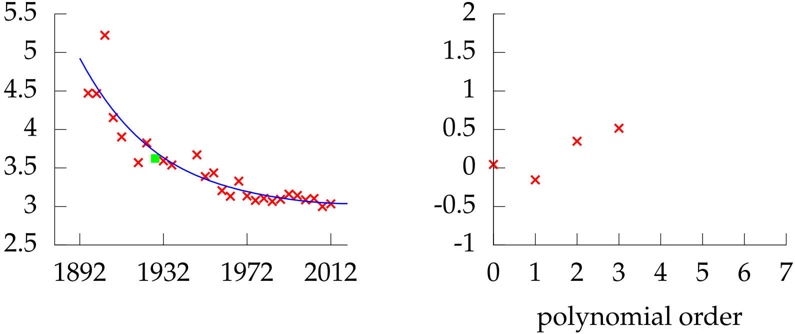
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



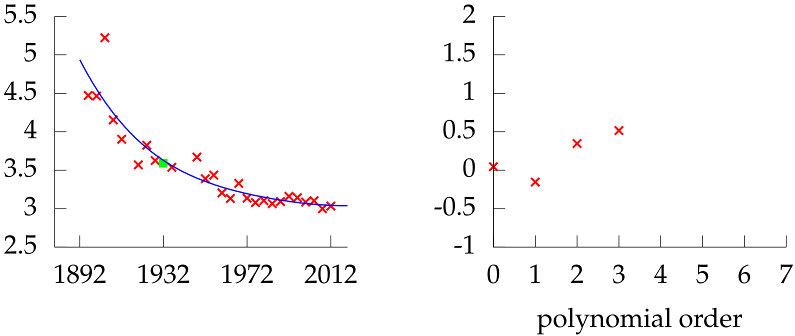
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



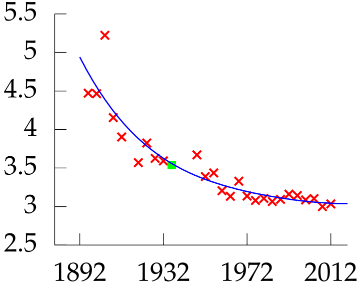
Polynomial order 4, training error -29.324, leave one out error 0.84844.

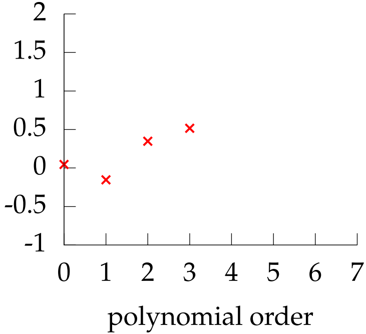
Leave One Out Error



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

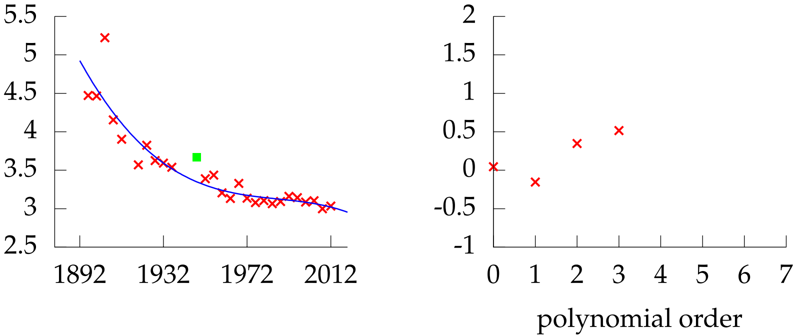
Leave One Out Error





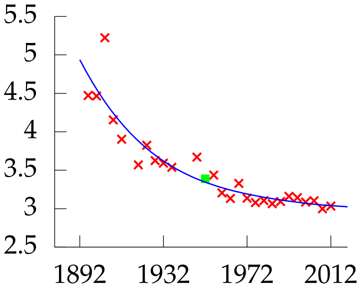
Polynomial order 4, training error -29.324, leave one out error 0.84844.

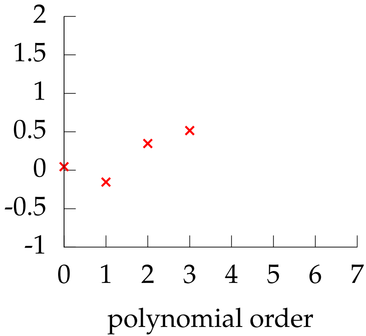
Leave One Out Error



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

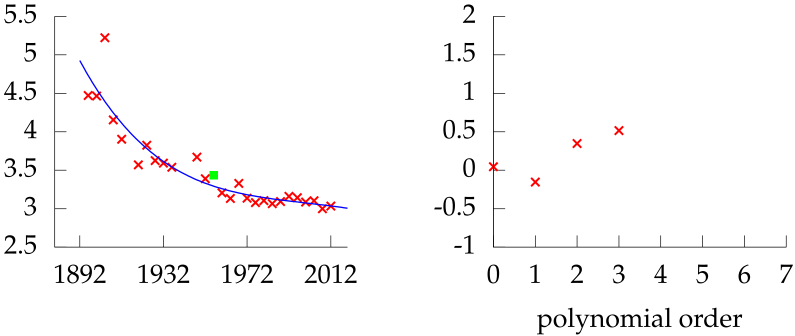
Leave One Out Error





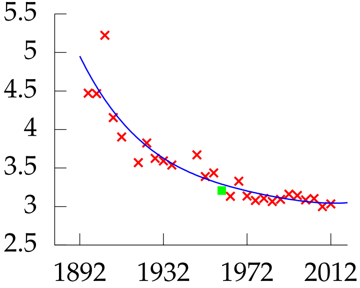
Polynomial order 4, training error -29.324, leave one out error 0.84844.

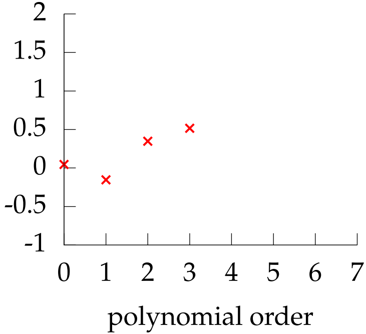
Leave One Out Error



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

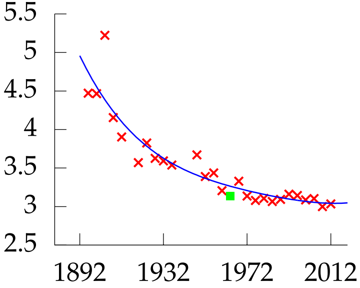
Leave One Out Error

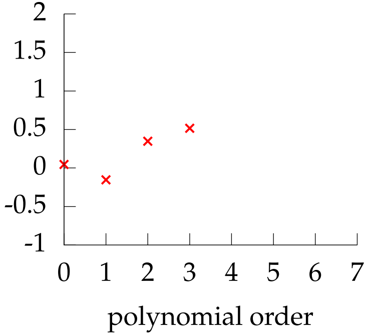




Polynomial order 4, training error -29.324, leave one out error 0.84844.

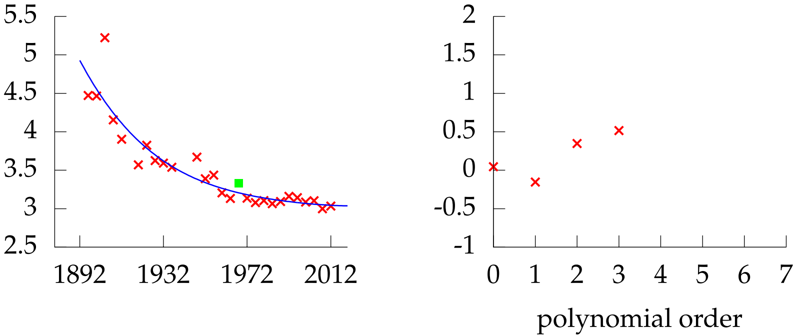
Leave One Out Error





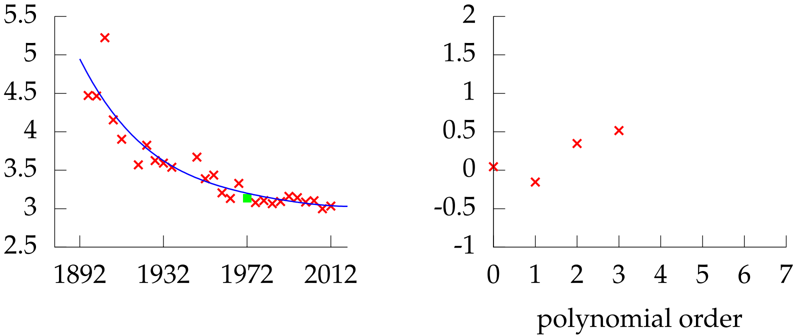
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



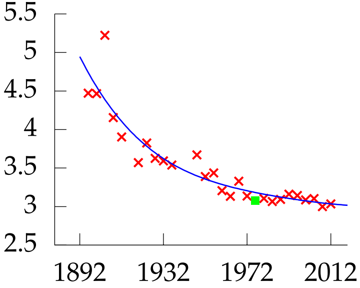
Polynomial order 4, training error -29.324, leave one out error 0.84844.

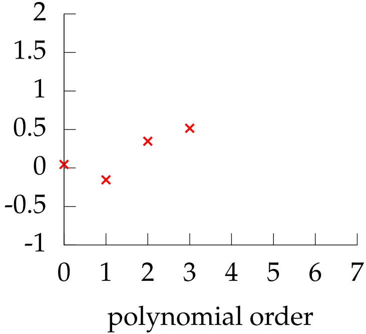
Leave One Out Error



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

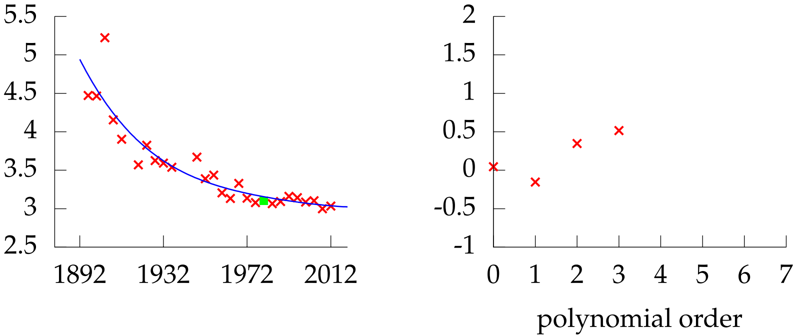
Leave One Out Error





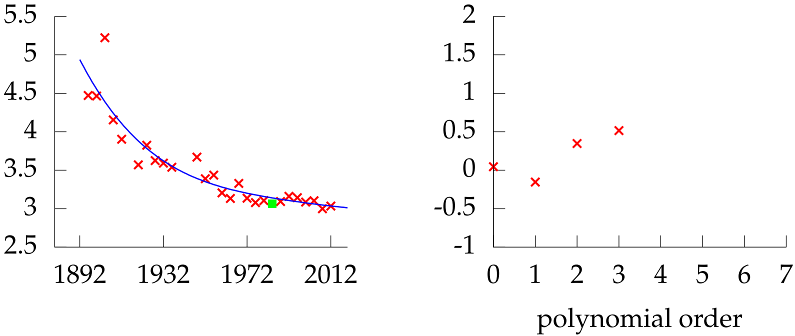
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



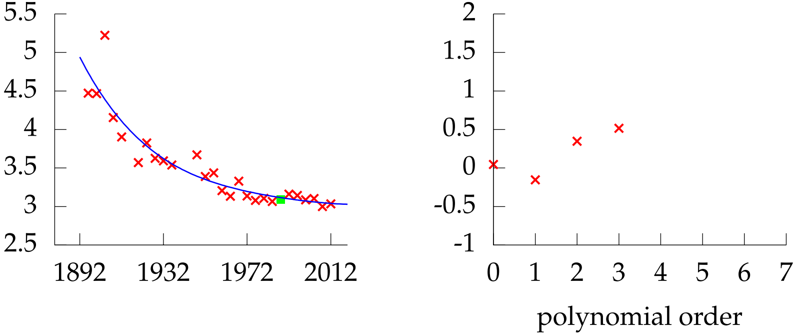
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



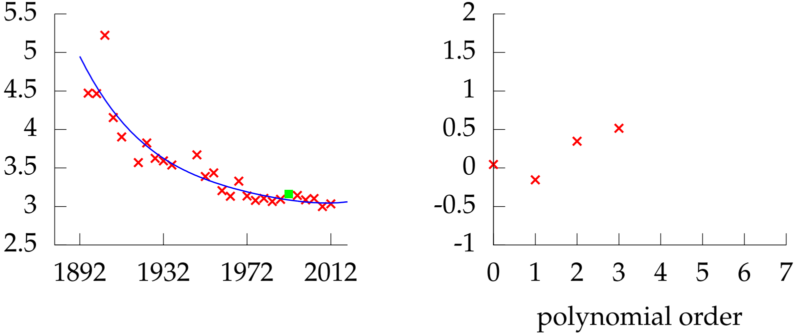
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



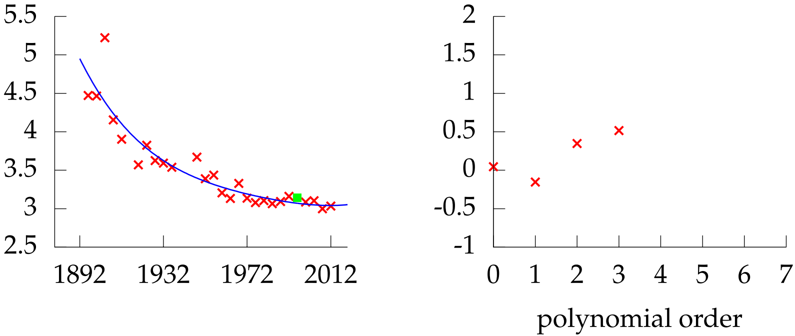
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



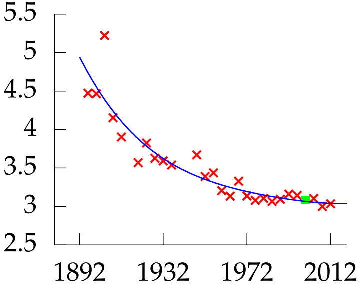
Polynomial order 4, training error -29.324, leave one out error 0.84844.

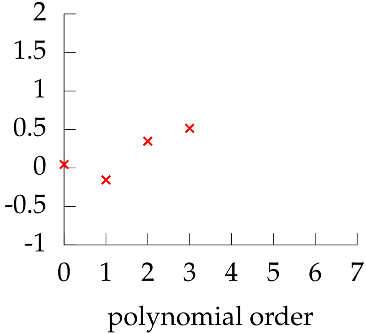
Leave One Out Error



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

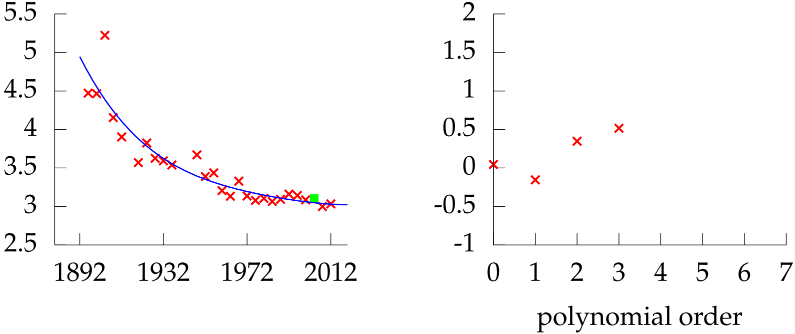
Leave One Out Error





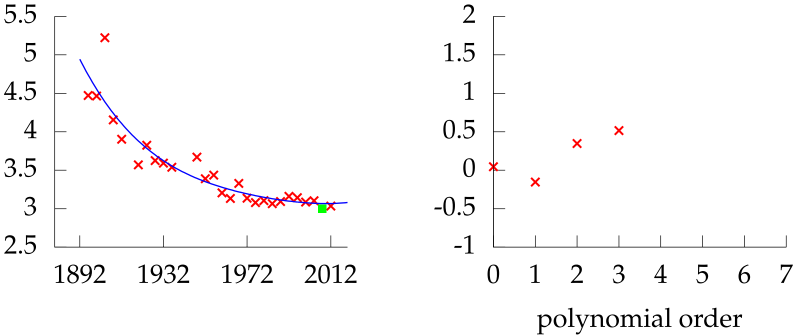
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



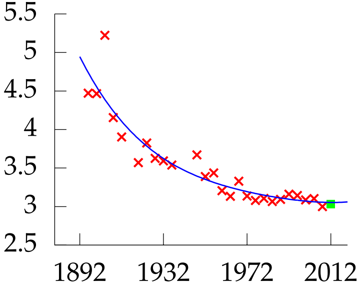
Polynomial order 4, training error -29.324, leave one out error 0.84844.

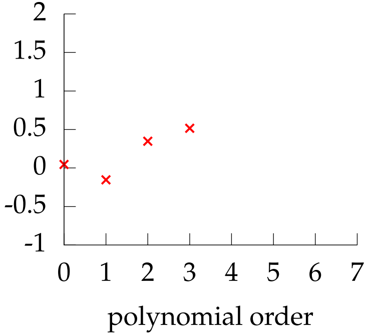
Leave One Out Error



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

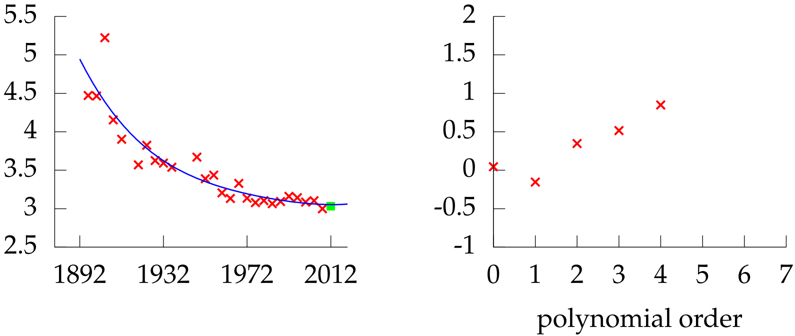
Leave One Out Error





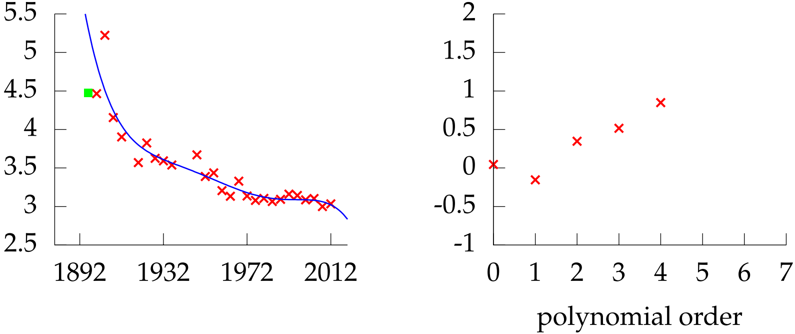
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



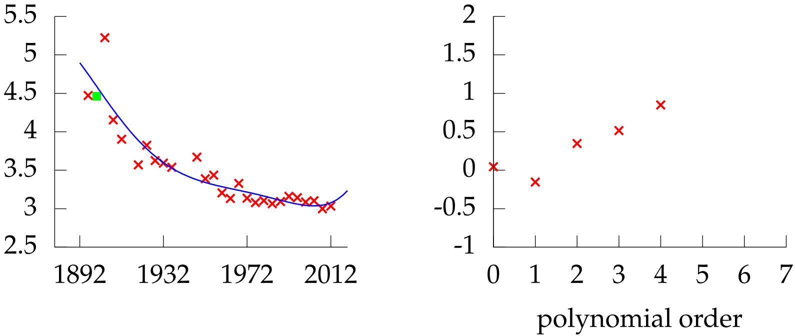
Polynomial order 4, training error -29.324, leave one out error 0.84844.

Leave One Out Error



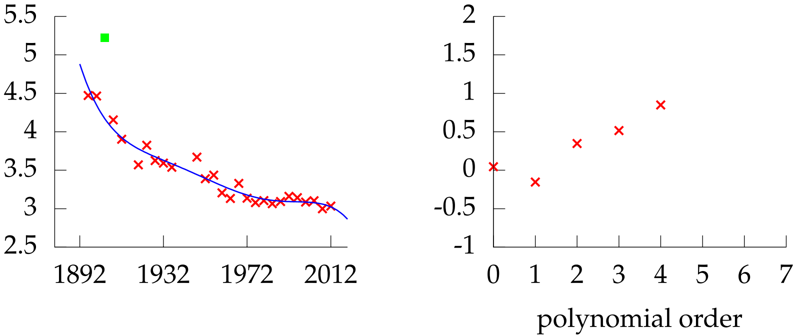
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



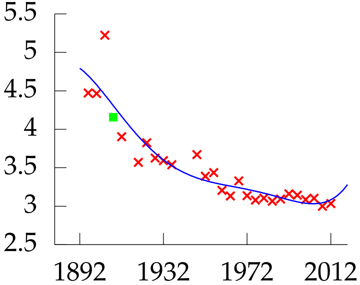
Polynomial order 5, training error -29.524, leave one out error 1.48.

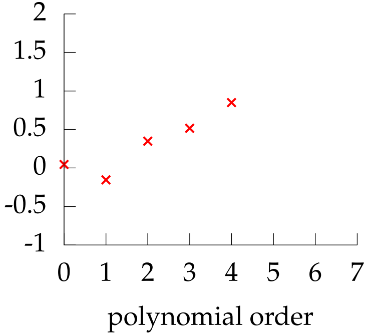
Leave One Out Error



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

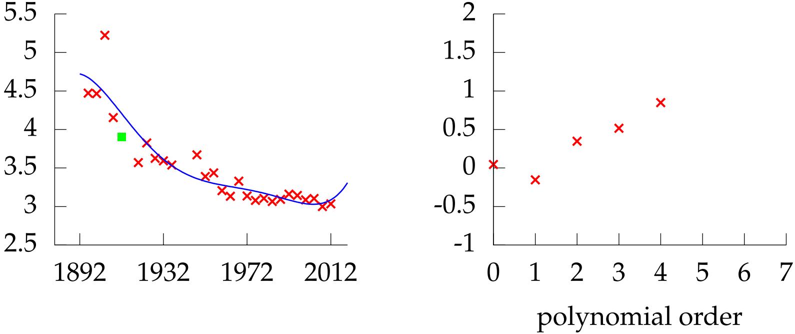
Leave One Out Error





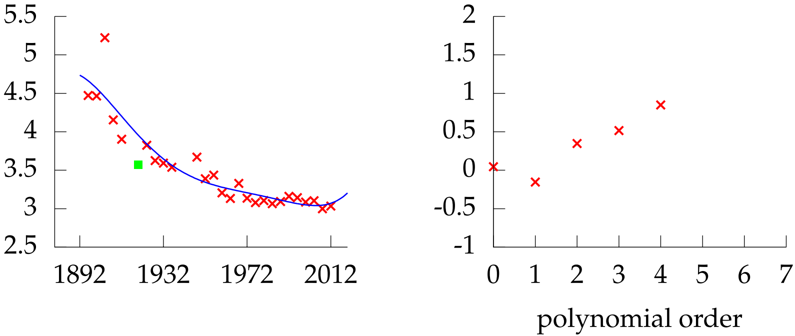
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



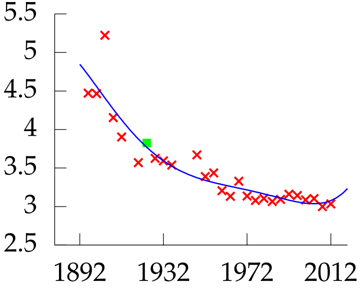
Polynomial order 5, training error -29.524, leave one out error 1.48.

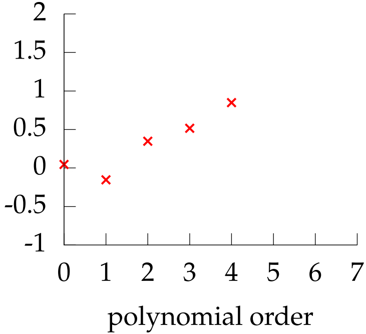
Leave One Out Error



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

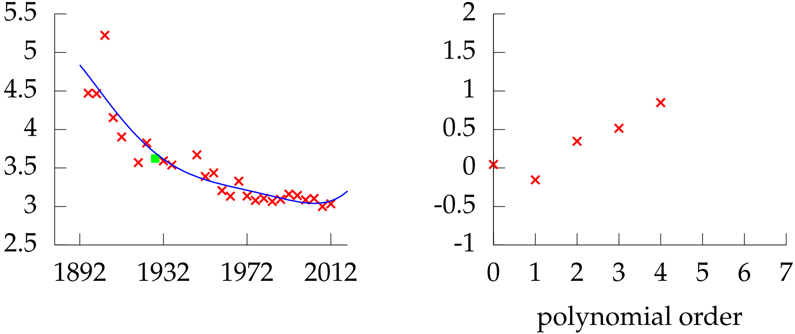
Leave One Out Error





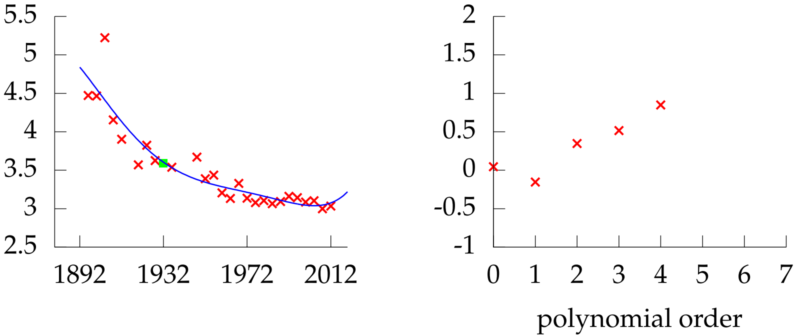
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



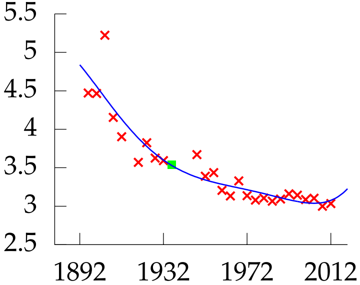
Polynomial order 5, training error -29.524, leave one out error 1.48.

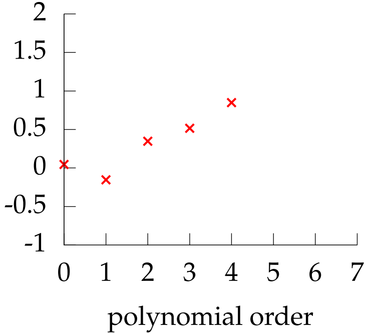
Leave One Out Error



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

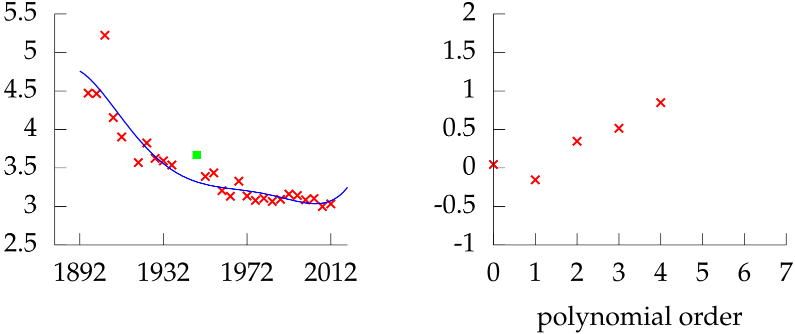
Leave One Out Error





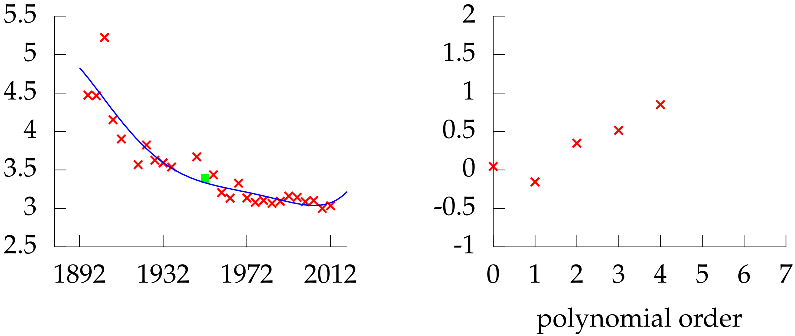
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



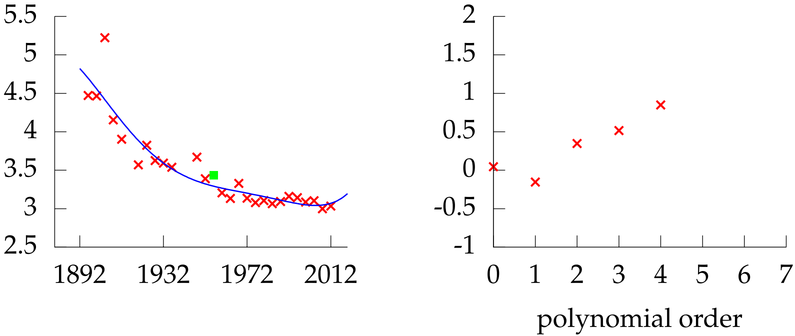
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



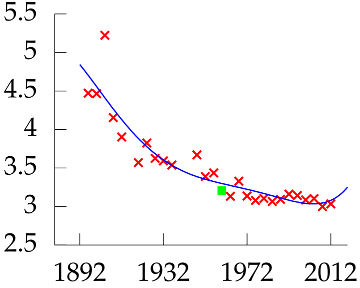
Polynomial order 5, training error -29.524, leave one out error 1.48.

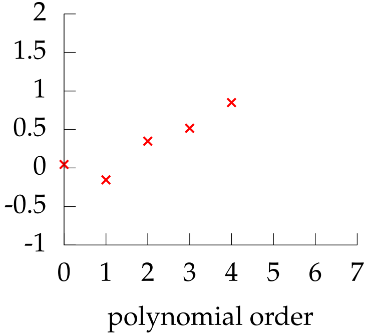
Leave One Out Error



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

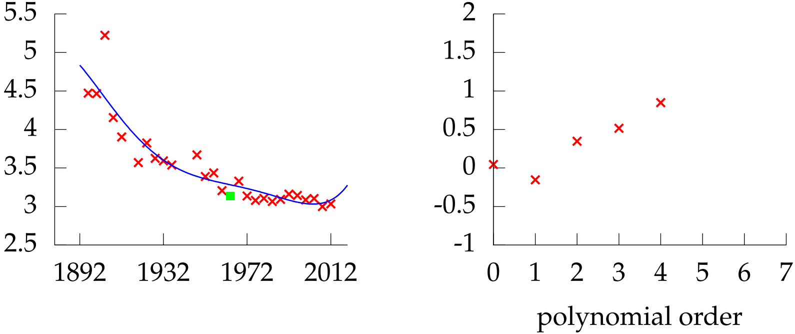
Leave One Out Error





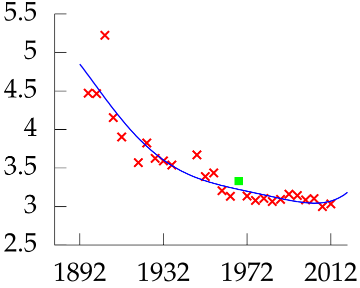
Polynomial order 5, training error -29.524, leave one out error 1.48.

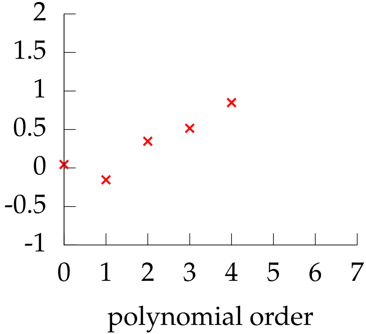
Leave One Out Error



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

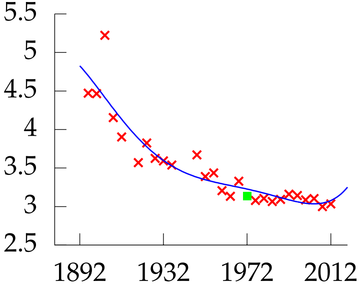
Leave One Out Error

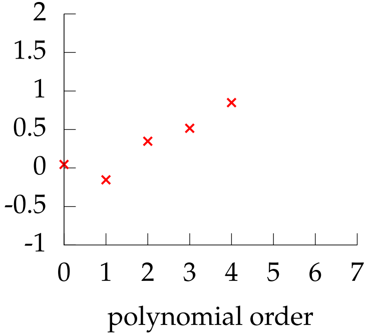




Polynomial order 5, training error -29.524, leave one out error 1.48.

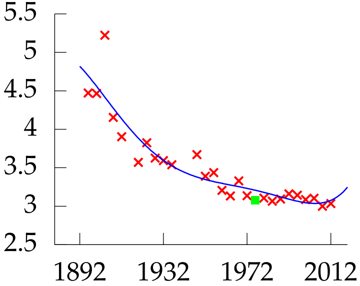
Leave One Out Error

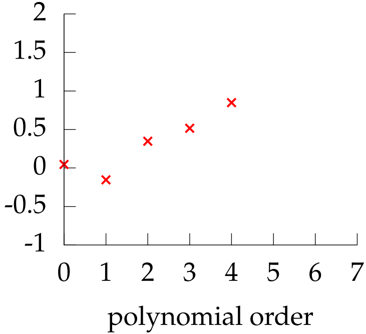




Polynomial order 5, training error -29.524, leave one out error 1.48.

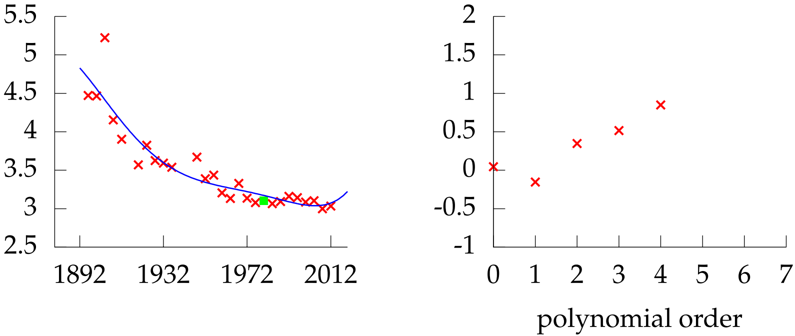
Leave One Out Error





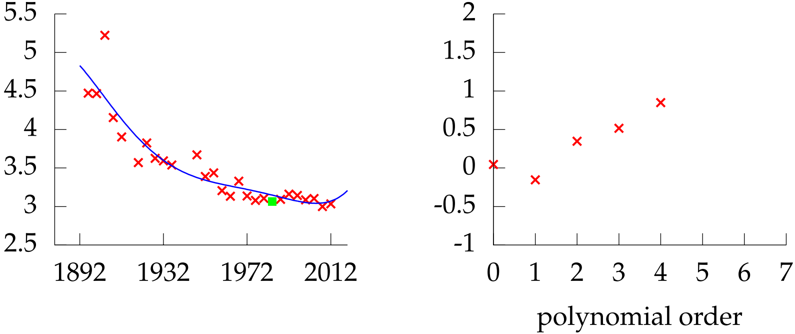
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



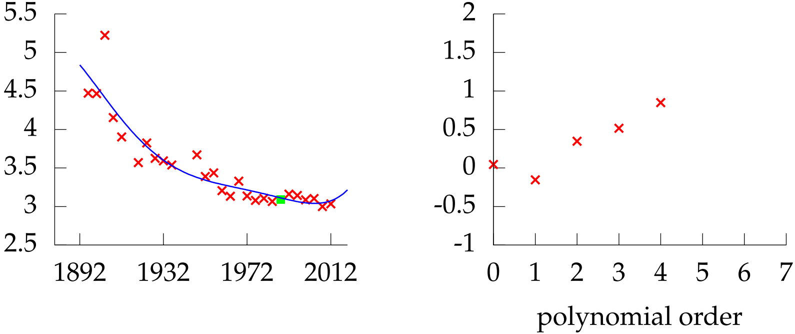
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



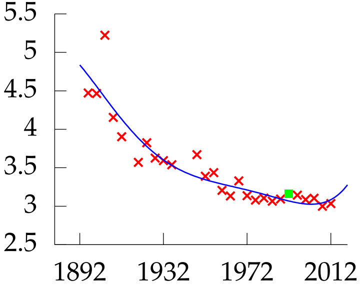
Polynomial order 5, training error -29.524, leave one out error 1.48.

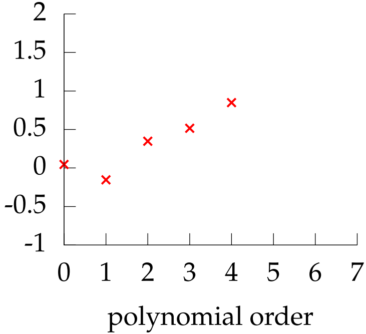
Leave One Out Error



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

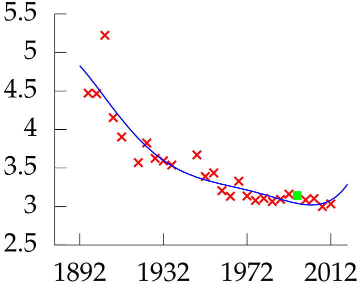
Leave One Out Error

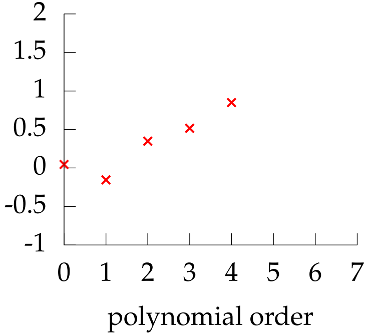




Polynomial order 5, training error -29.524, leave one out error 1.48.

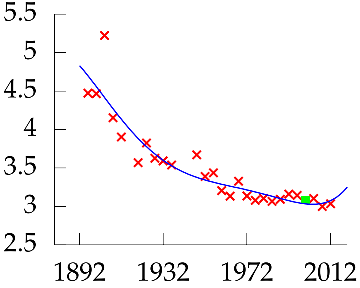
Leave One Out Error

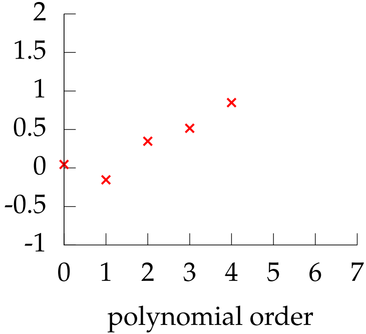




Polynomial order 5, training error -29.524, leave one out error 1.48.

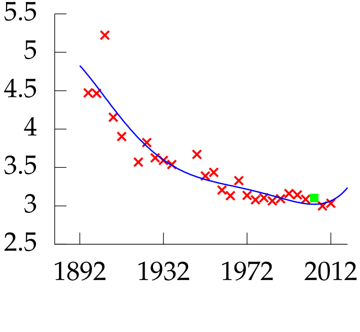
Leave One Out Error





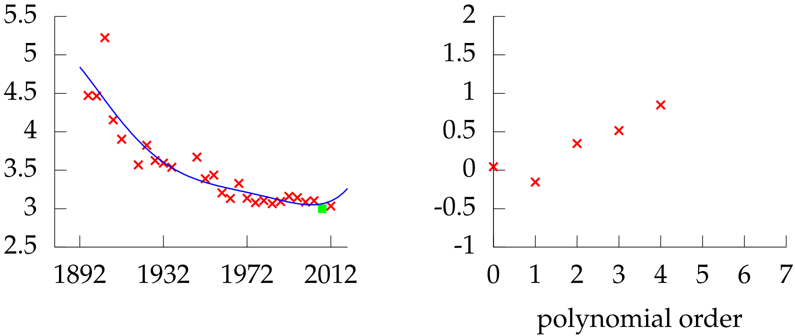
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



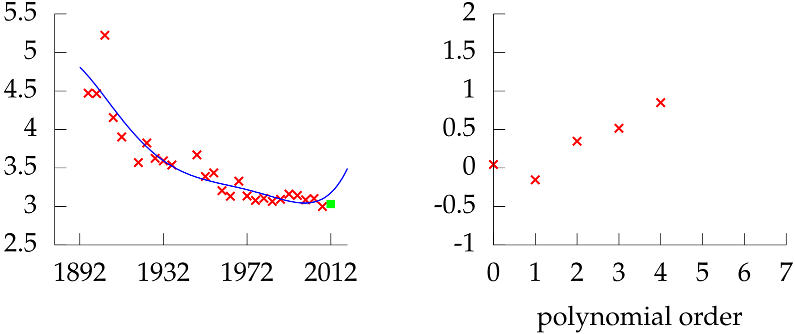
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



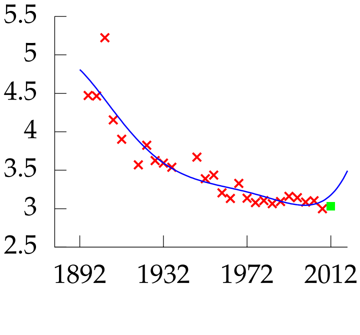
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



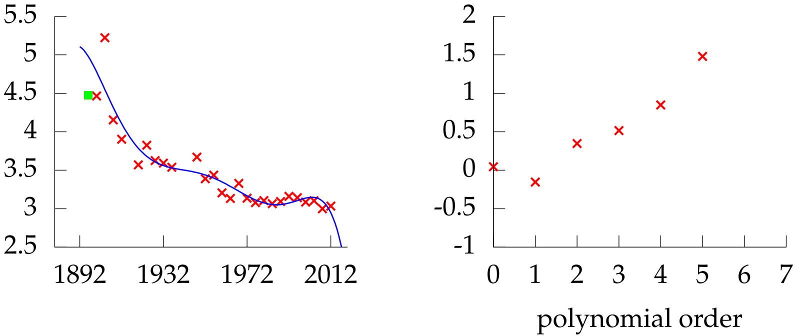
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



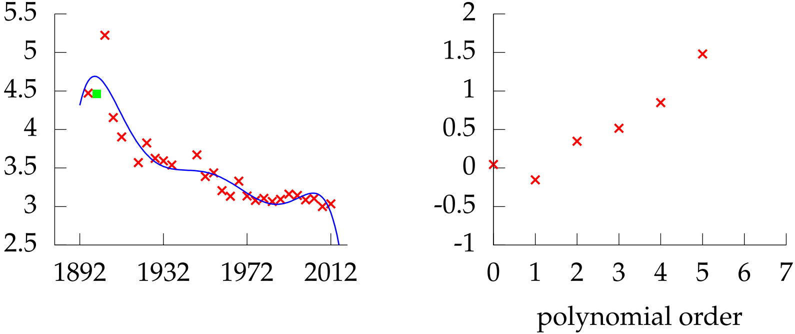
Polynomial order 5, training error -29.524, leave one out error 1.48.

Leave One Out Error



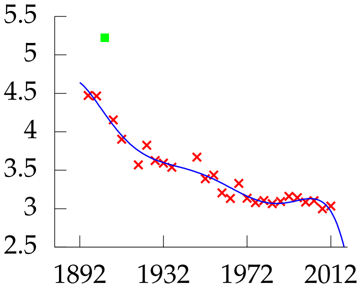
Polynomial order 6, training error -32.237, leave one out error 1.5047.

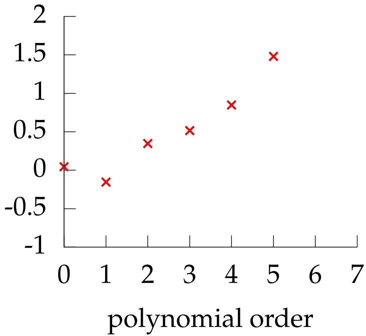
Leave One Out Error



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

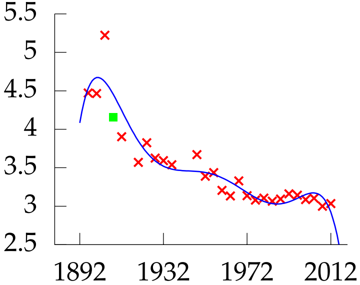
Leave One Out Error

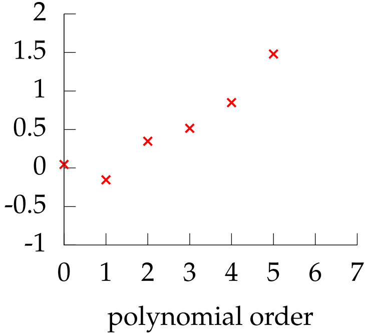




Polynomial order 6, training error -32.237, leave one out error 1.5047.

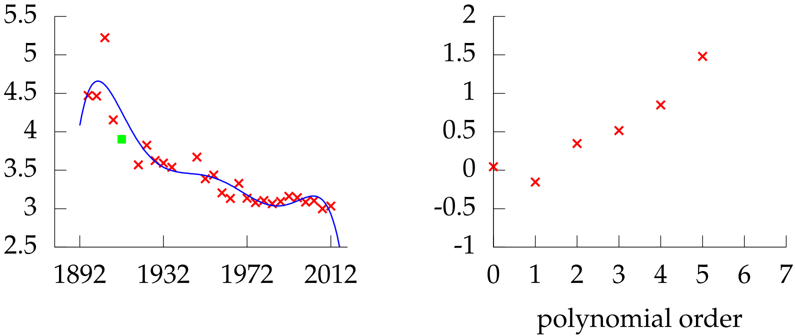
Leave One Out Error





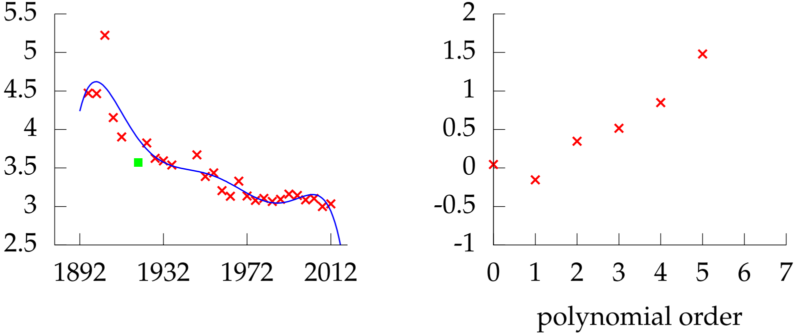
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



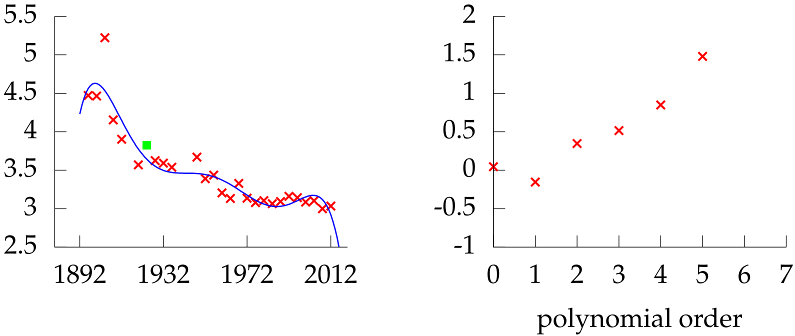
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



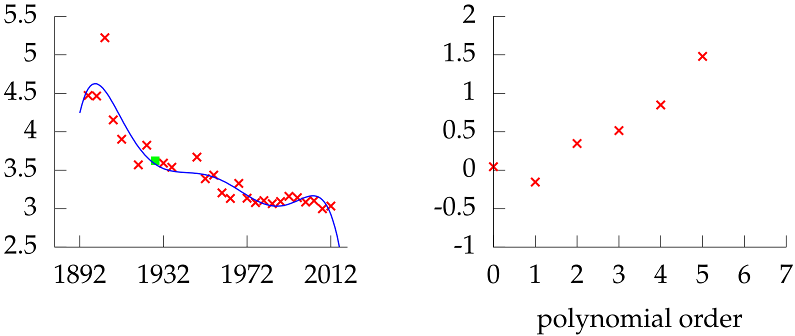
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



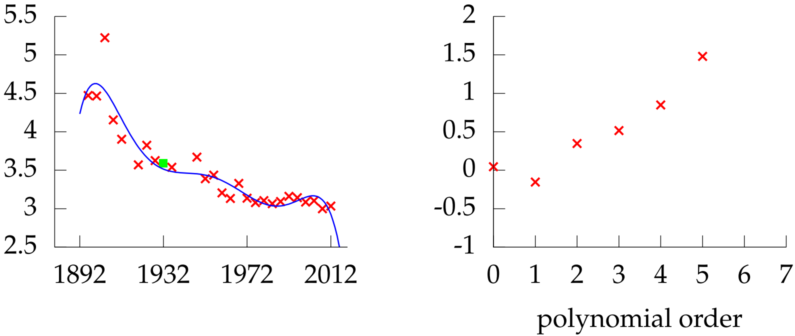
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



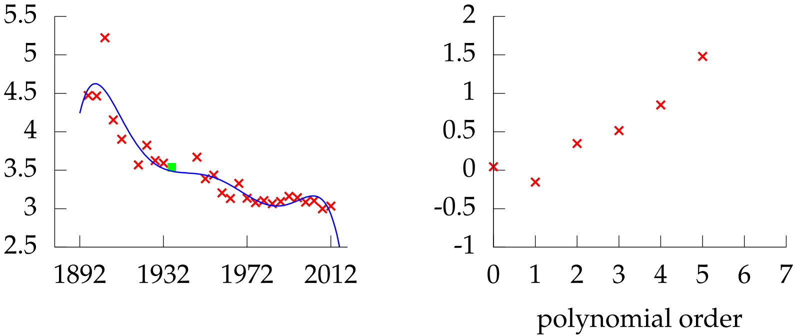
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



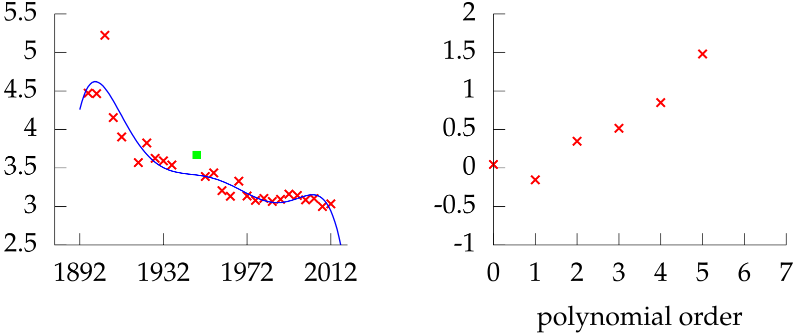
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



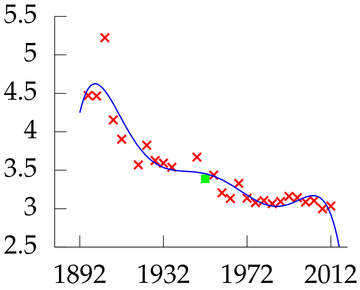
Polynomial order 6, training error -32.237, leave one out error 1.5047.

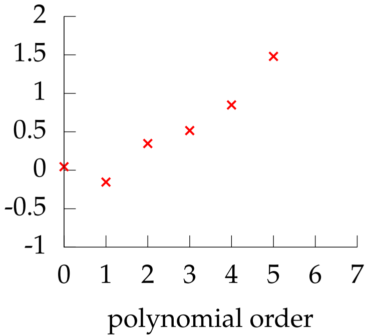
Leave One Out Error



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

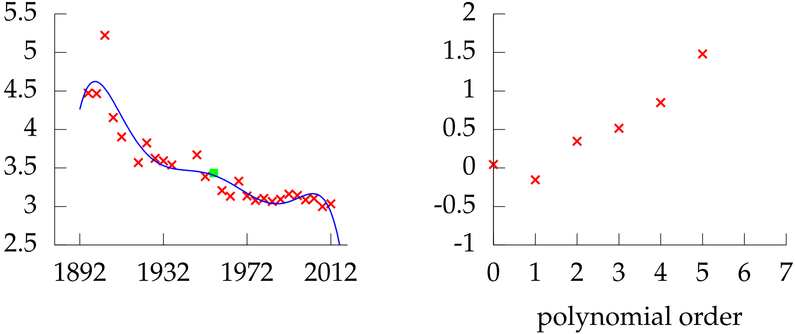
Leave One Out Error





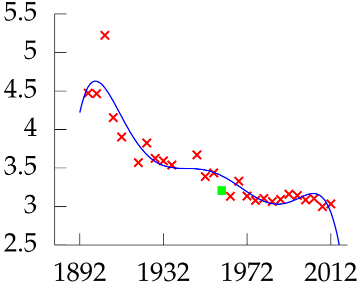
Polynomial order 6, training error -32.237, leave one out error 1.5047.

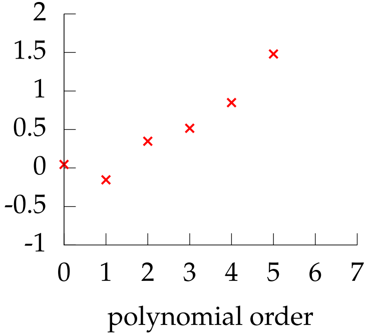
Leave One Out Error



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

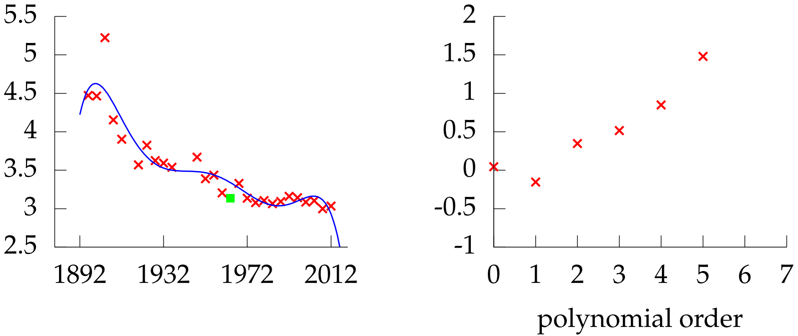
Leave One Out Error





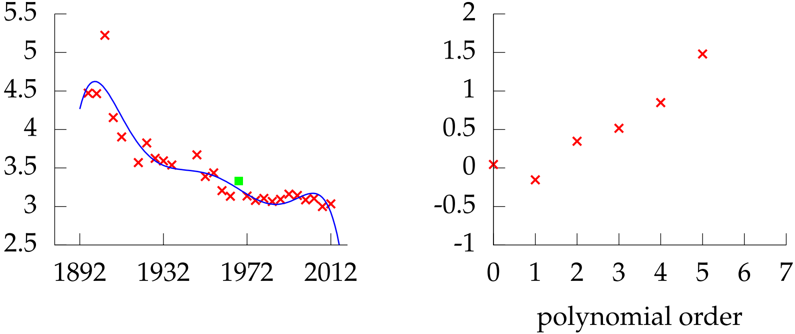
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



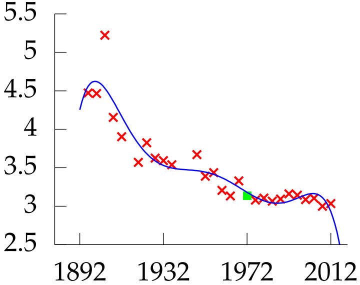
Polynomial order 6, training error -32.237, leave one out error 1.5047.

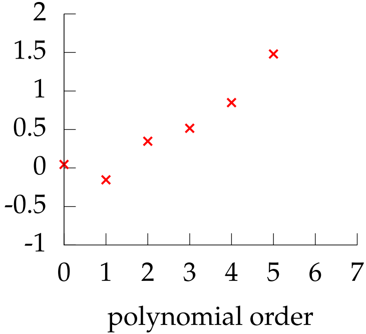
Leave One Out Error



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

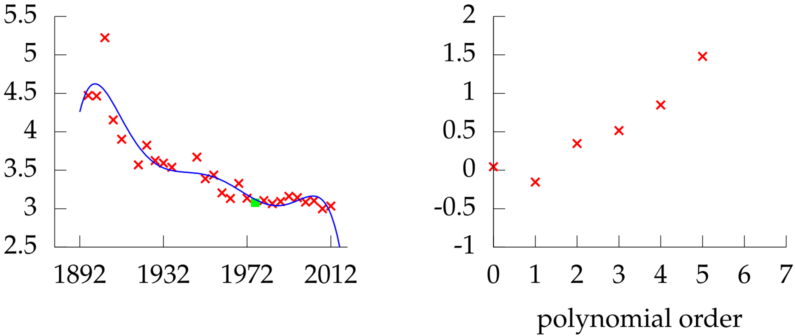
Leave One Out Error





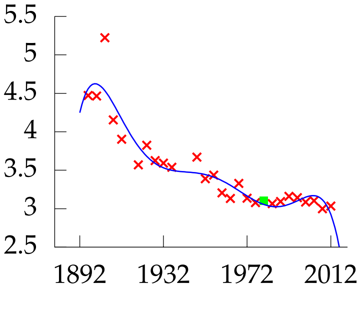
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



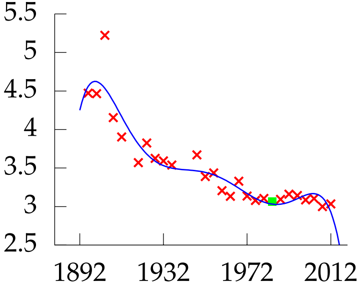
Polynomial order 6, training error -32.237, leave one out error 1.5047.

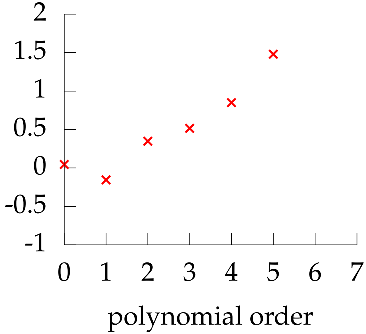
Leave One Out Error



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

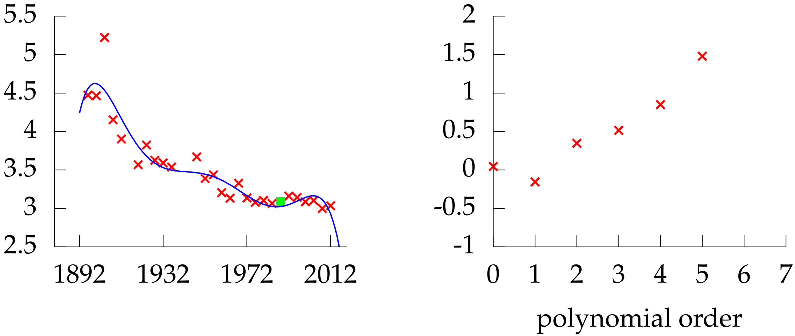
Leave One Out Error





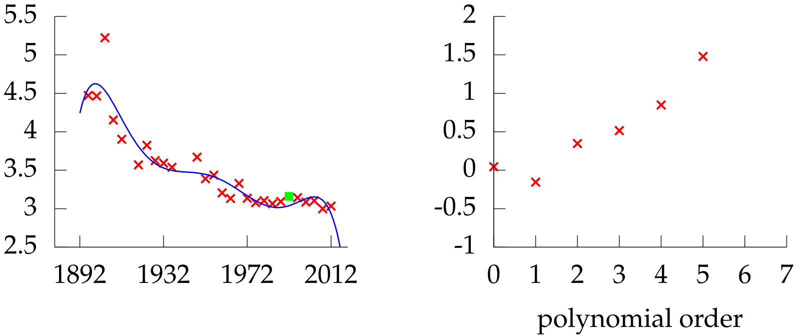
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



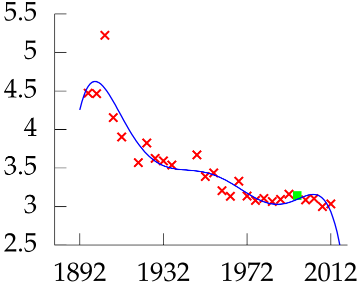
Polynomial order 6, training error -32.237, leave one out error 1.5047.

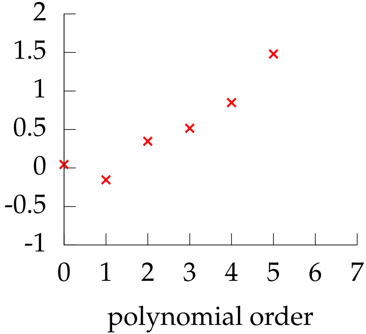
Leave One Out Error



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

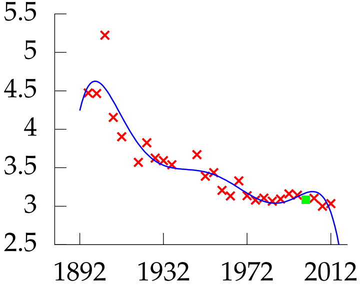
Leave One Out Error

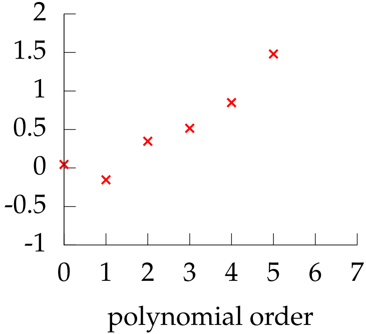




Polynomial order 6, training error -32.237, leave one out error 1.5047.

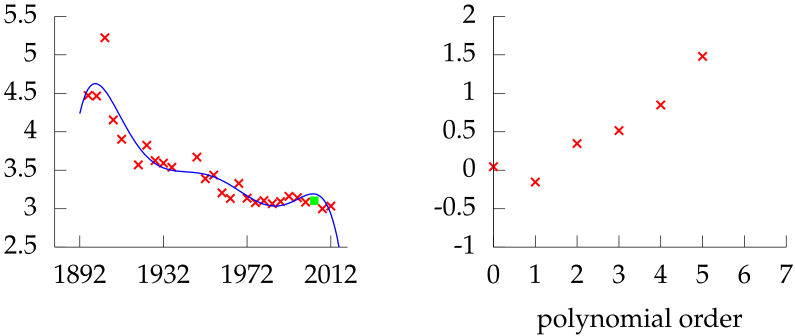
Leave One Out Error





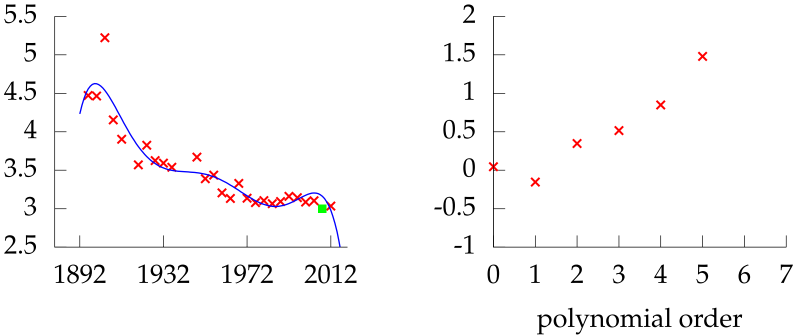
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



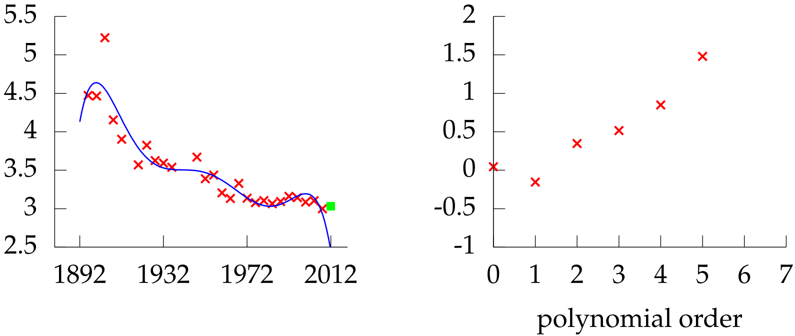
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



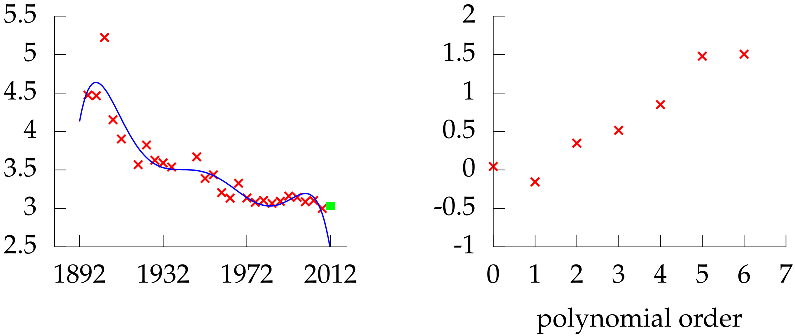
Polynomial order 6, training error -32.237, leave one out error 1.5047.

Leave One Out Error



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

Leave One Out Error

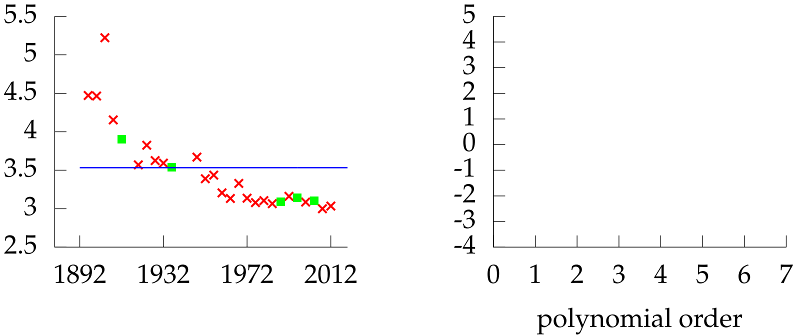


Polynomial order 6, training error -32.237, leave one out error 1.5047.

k Fold Cross Validation

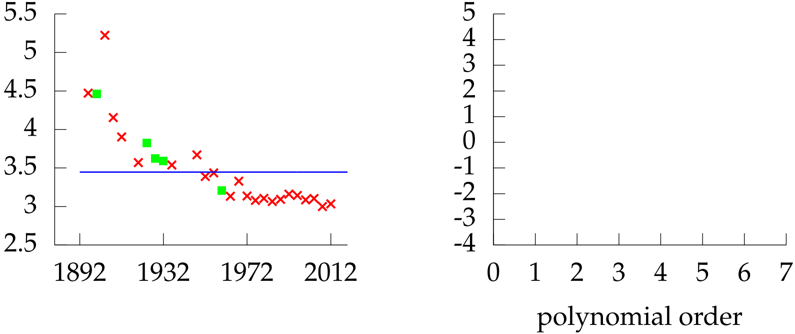
- Leave one out cross validation can be very time consuming!
- Need to train your algorithm n times.
- An alternative: k fold cross validation.

Cross Validation Error



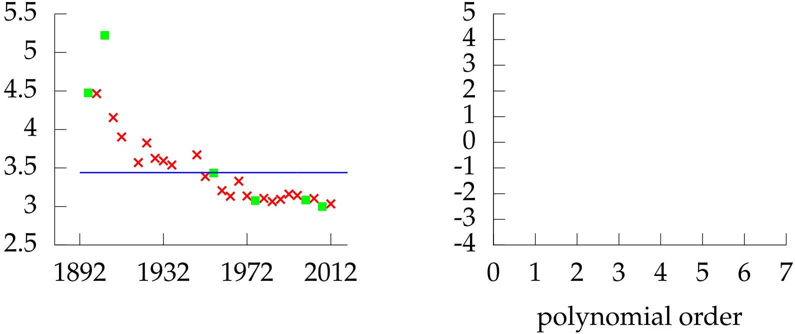
Polynomial order 0, training error -3.2644, leave one out error 0.045811.

Cross Validation Error



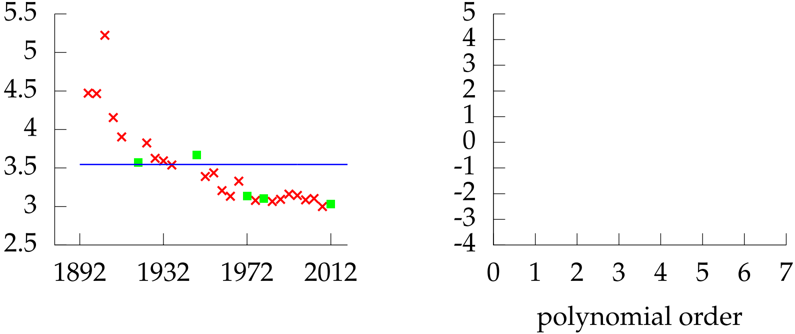
Polynomial order 0, training error -3.2644, leave one out error 0.045811.

Cross Validation Error



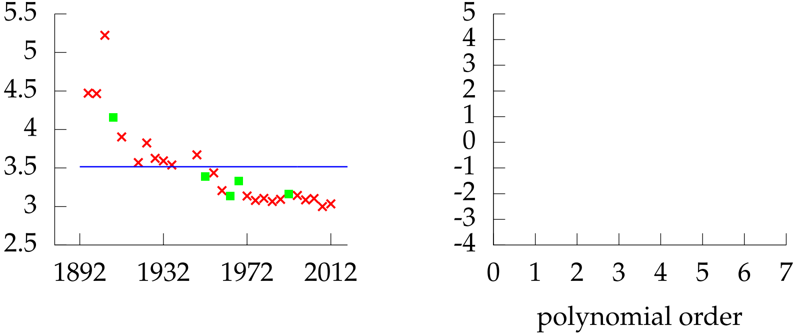
Polynomial order 0, training error -3.2644, leave one out error 0.045811.

Cross Validation Error



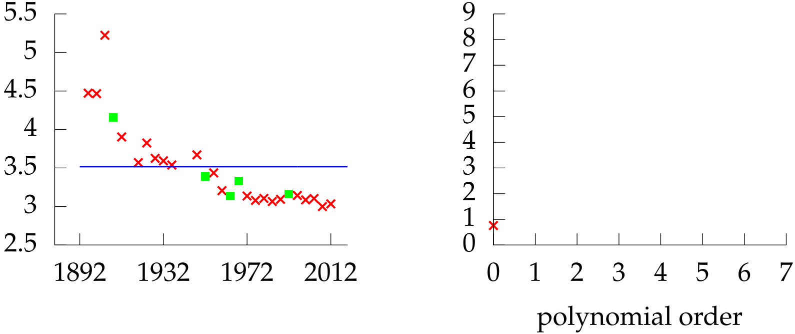
Polynomial order 0, training error -3.2644, leave one out error 0.045811.

Cross Validation Error



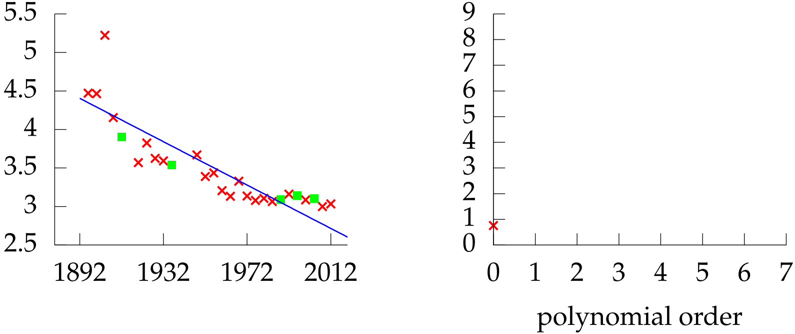
Polynomial order 0, training error -3.2644, leave one out error 0.045811.

Cross Validation Error



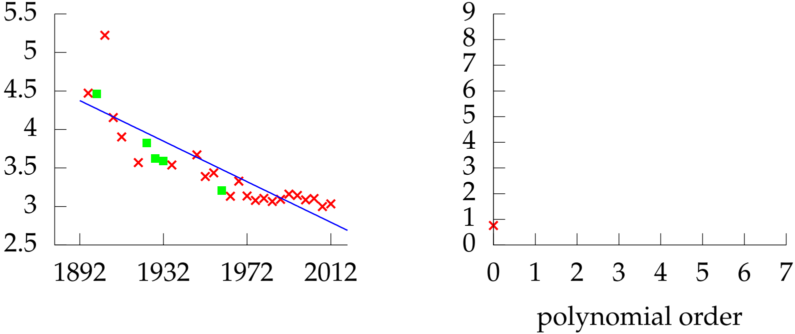
Polynomial order 0, training error -3.2644, leave one out error 0.045811.

Cross Validation Error



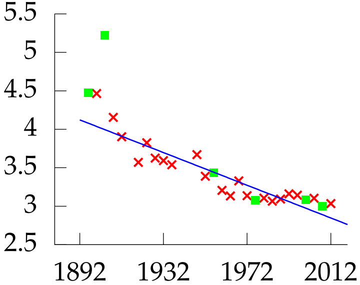
Polynomial order 1, training error -18.873, leave one out error -0.15413.

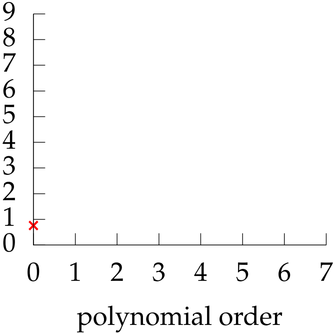
Cross Validation Error



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

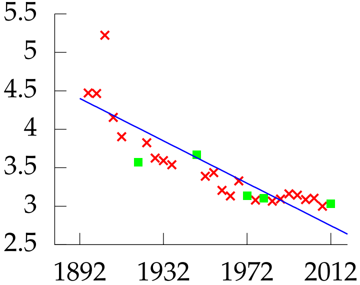
Cross Validation Error

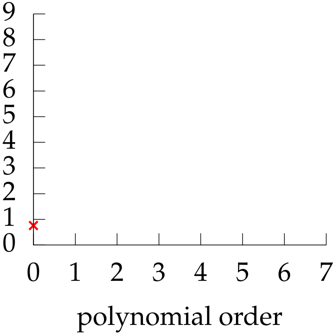




Polynomial order 1, training error -18.873, leave one out error -0.15413.

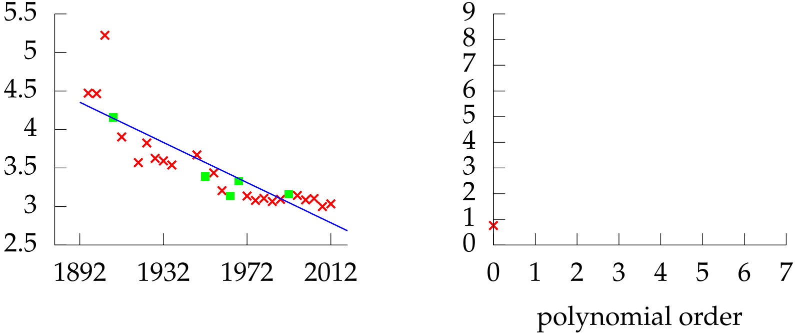
Cross Validation Error





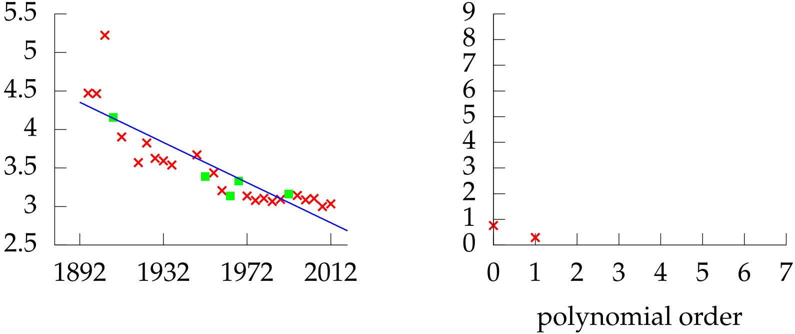
Polynomial order 1, training error -18.873, leave one out error -0.15413.

Cross Validation Error



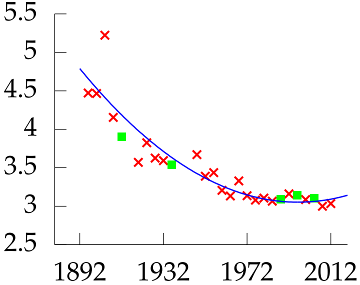
Polynomial order 1, training error -18.873, leave one out error -0.15413.

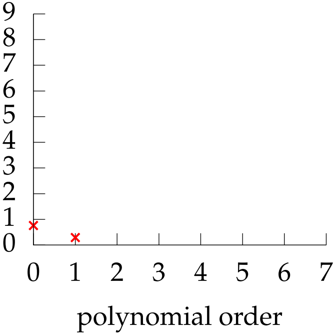
Cross Validation Error



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

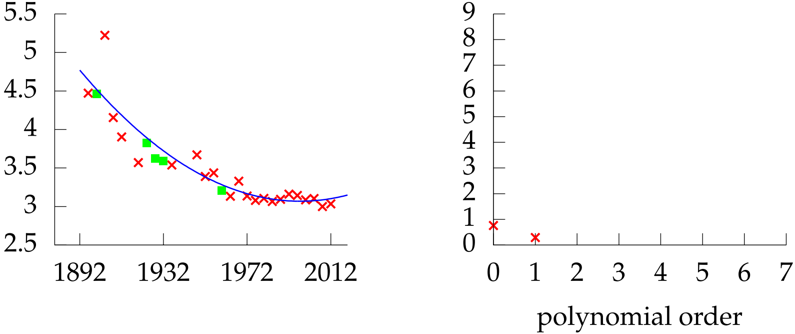
Cross Validation Error





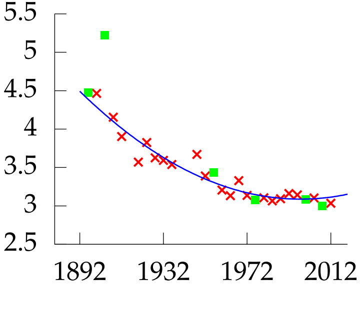
Polynomial order 2, training error -25.177, leave one out error 0.34669.

Cross Validation Error



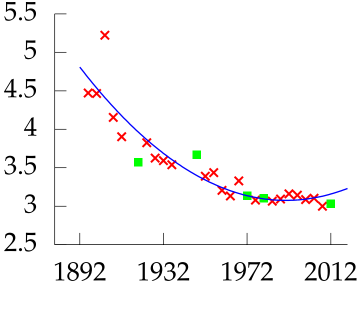
Polynomial order 2, training error -25.177, leave one out error 0.34669.

Cross Validation Error



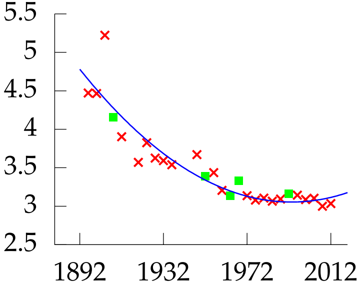
Polynomial order 2, training error -25.177, leave one out error 0.34669.

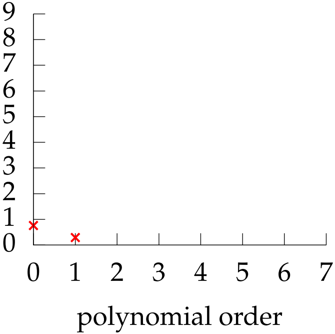
Cross Validation Error



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

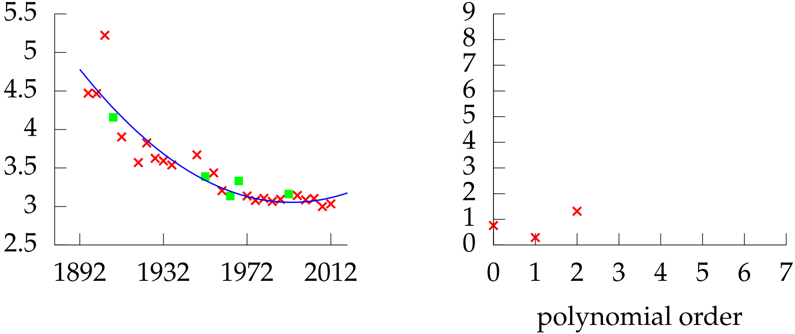
Cross Validation Error





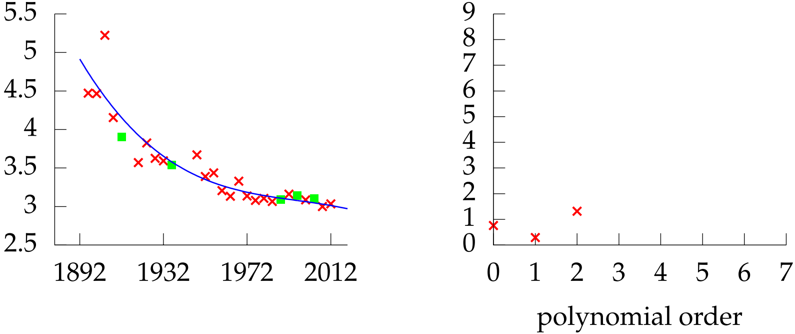
Polynomial order 2, training error -25.177, leave one out error 0.34669.

Cross Validation Error



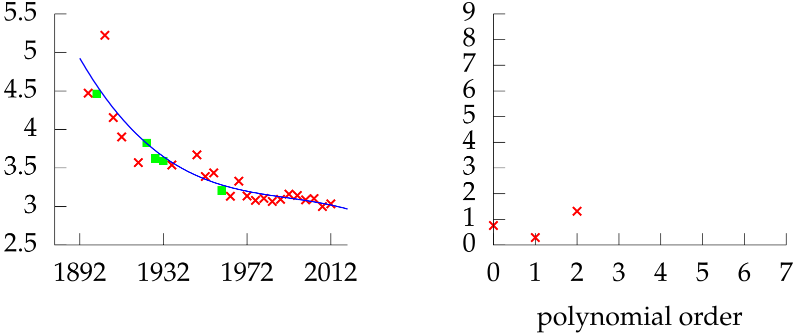
Polynomial order 2, training error -25.177, leave one out error 0.34669.

Cross Validation Error



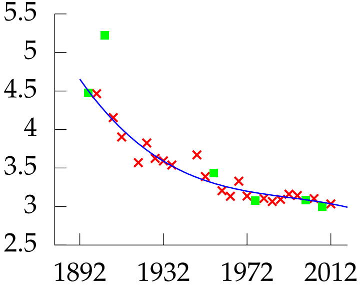
Polynomial order 3, training error -25.777, leave one out error 0.51621.

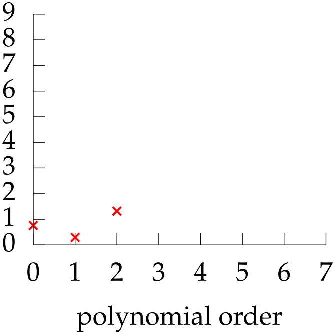
Cross Validation Error



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

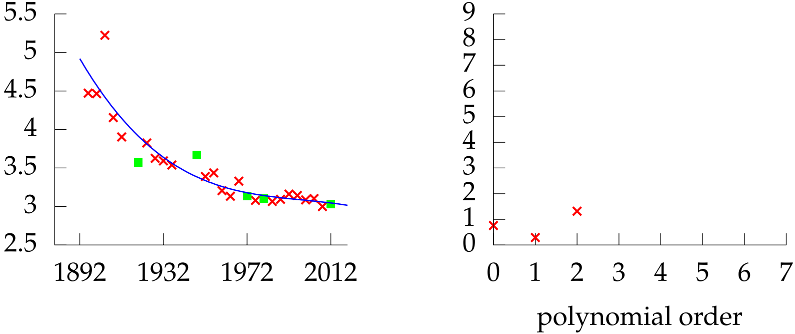
Cross Validation Error





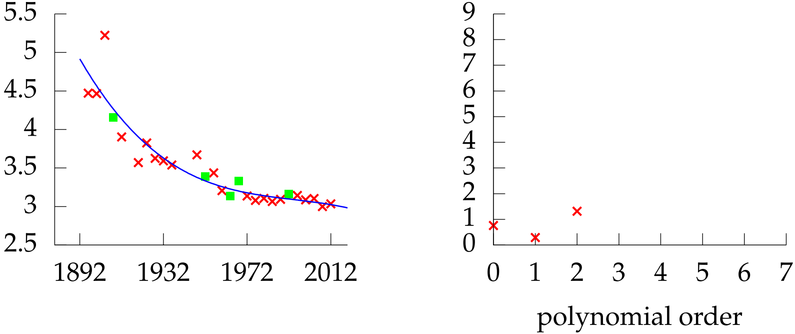
Polynomial order 3, training error -25.777, leave one out error 0.51621.

Cross Validation Error



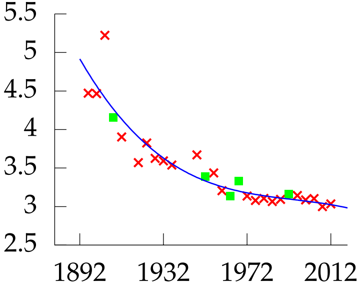
Polynomial order 3, training error -25.777, leave one out error 0.51621.

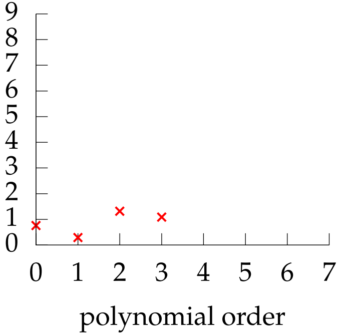
Cross Validation Error



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

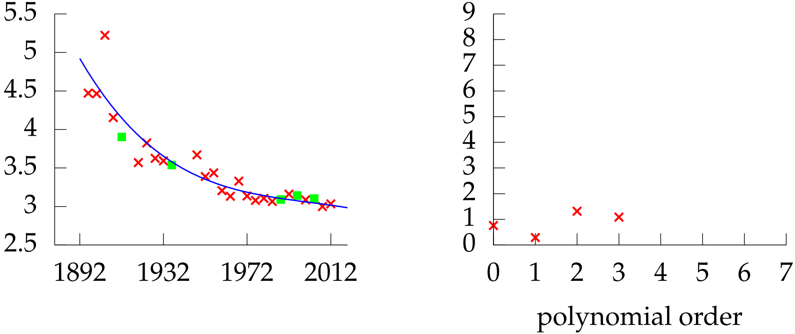
Cross Validation Error





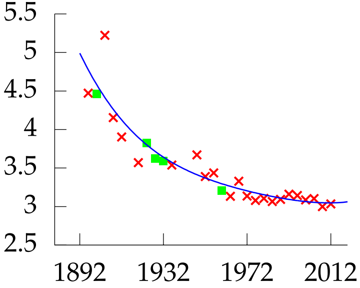
Polynomial order 3, training error -25.777, leave one out error 0.51621.

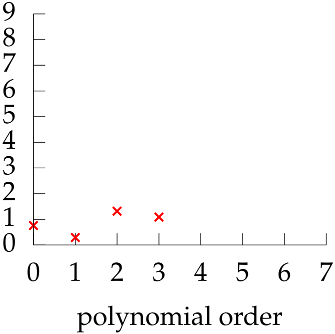
Cross Validation Error



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

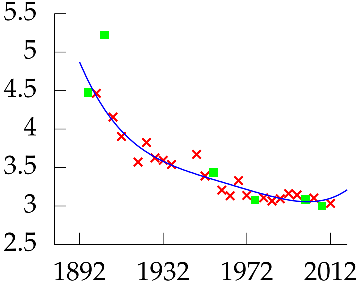
Cross Validation Error

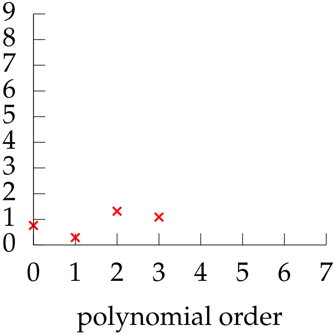




Polynomial order 4, training error -26.048, leave one out error 0.84844.

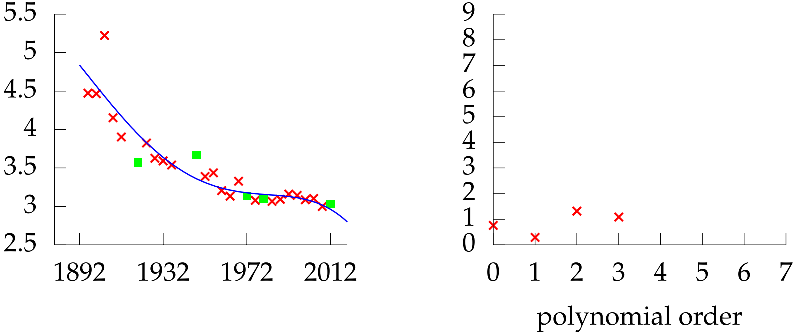
Cross Validation Error





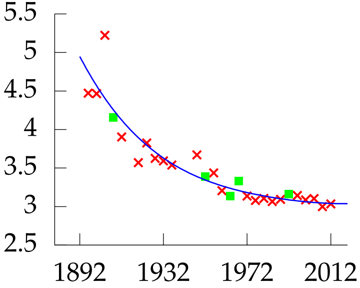
Polynomial order 4, training error -26.048, leave one out error 0.84844.

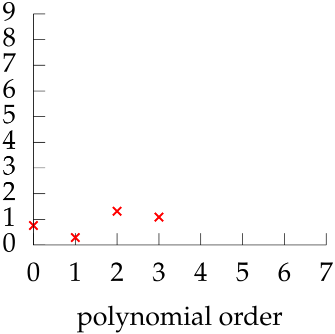
Cross Validation Error



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

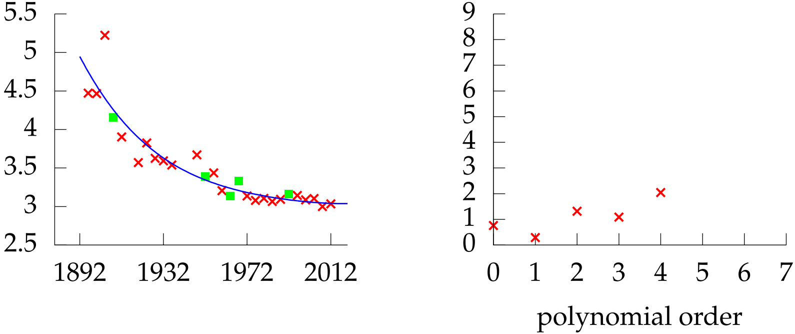
Cross Validation Error





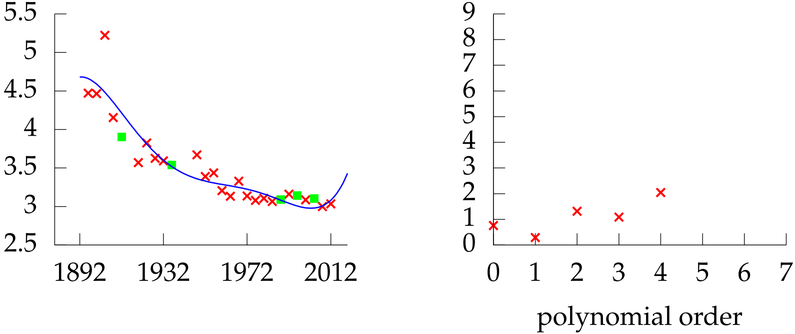
Polynomial order 4, training error -26.048, leave one out error 0.84844.

Cross Validation Error



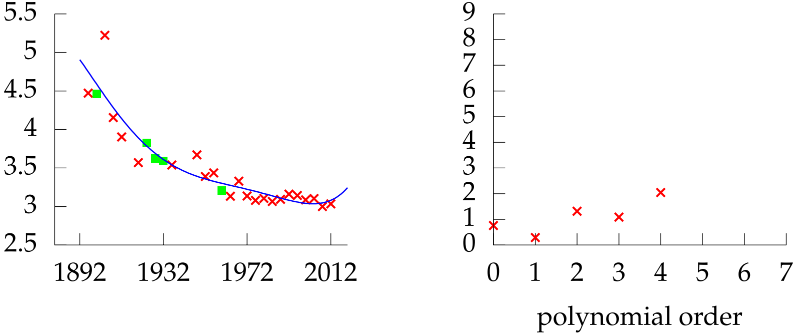
Polynomial order 4, training error -26.048, leave one out error 0.84844.

Cross Validation Error



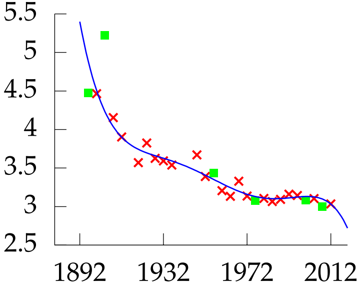
Polynomial order 5, training error -26.892, leave one out error 1.48.

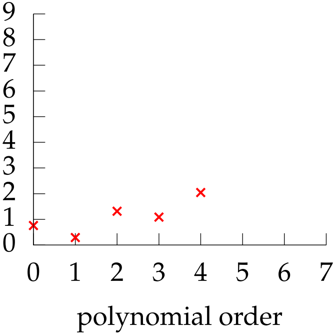
Cross Validation Error



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

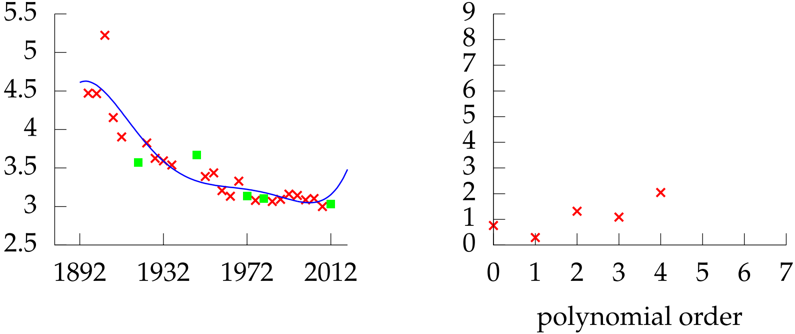
Cross Validation Error





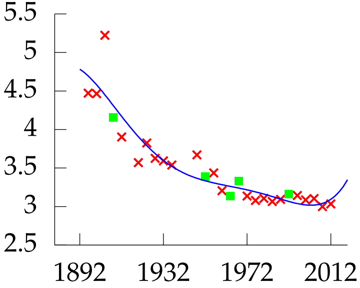
Polynomial order 5, training error -26.892, leave one out error 1.48.

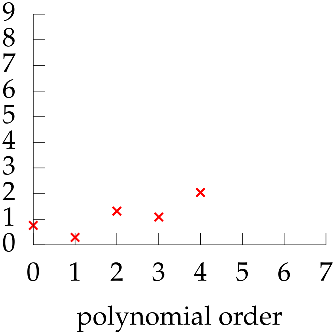
Cross Validation Error



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

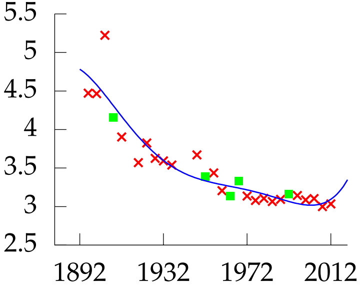
Cross Validation Error

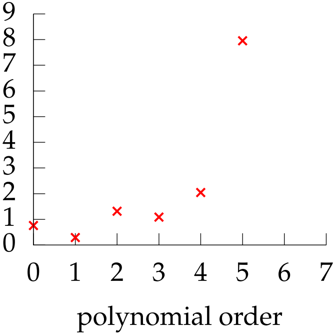




Polynomial order 5, training error -26.892, leave one out error 1.48.

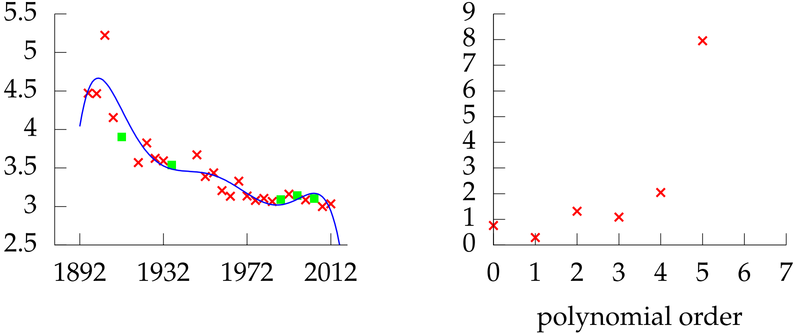
Cross Validation Error





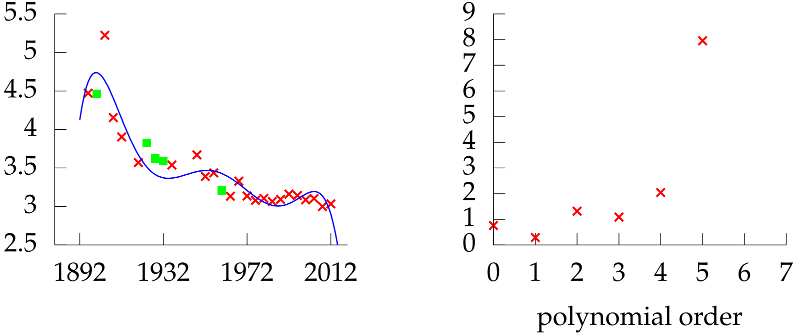
Polynomial order 5, training error -26.892, leave one out error 1.48.

Cross Validation Error



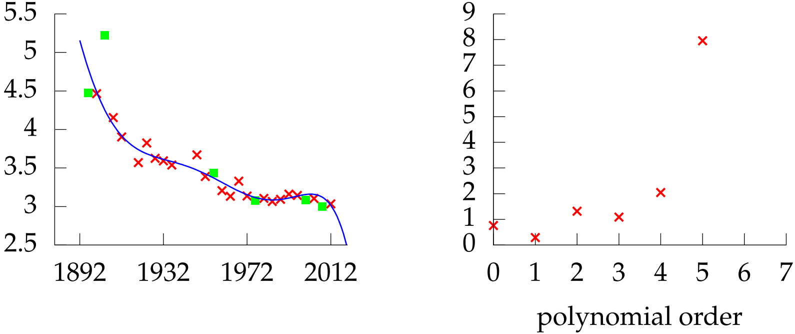
Polynomial order 6, training error -29.395, leave one out error 1.5047.

Cross Validation Error



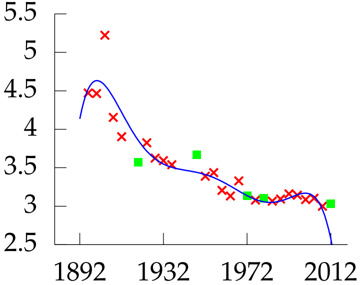
Polynomial order 6, training error -29.395, leave one out error 1.5047.

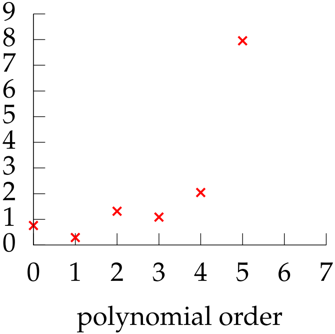
Cross Validation Error



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

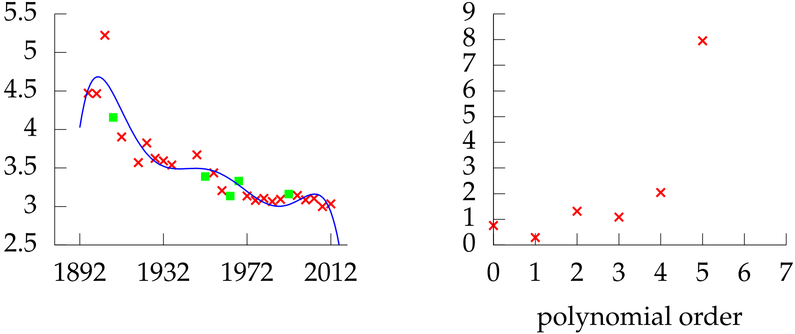
Cross Validation Error





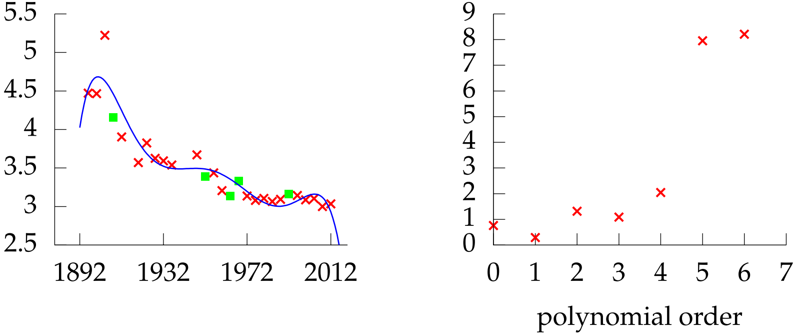
Polynomial order 6, training error -29.395, leave one out error 1.5047.

Cross Validation Error



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

Cross Validation Error

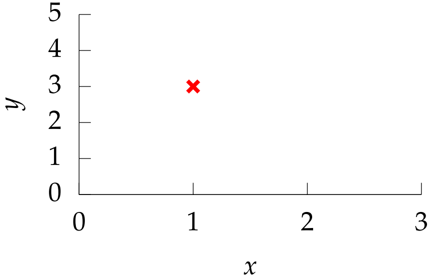


Polynomial order 6, training error -29.395, leave one out error 1.5047.

Underdetermined System

What about two unknowns and *one* observation?

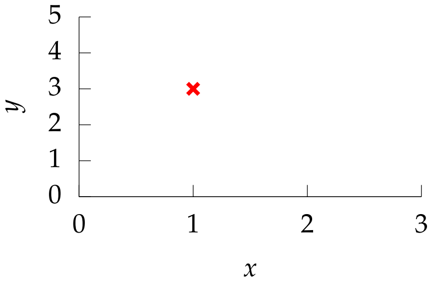
$$y_1 = mx_1 + c$$



Underdetermined System

Can compute m given c.

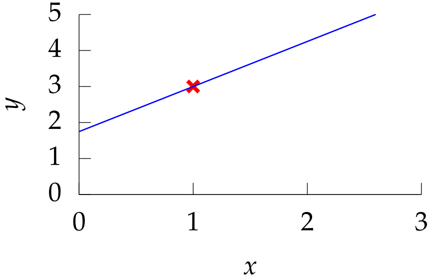
$$m = \frac{y_1 - c}{x}$$



Underdetermined System

Can compute m given c.

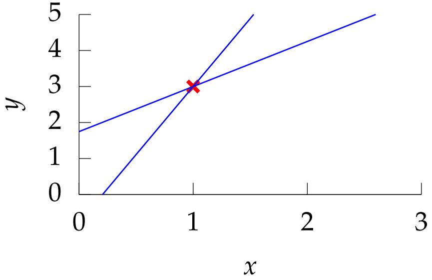
$$c = 1.75 \implies m = 1.25$$



Underdetermined System

Can compute m given c.

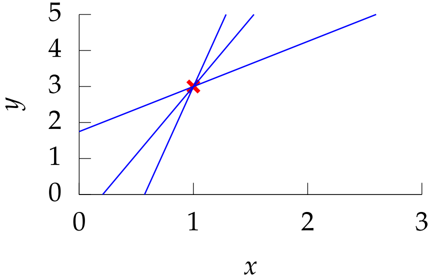
$$c = -0.777 \Longrightarrow m = 3.78$$



Underdetermined System

Can compute m given c.

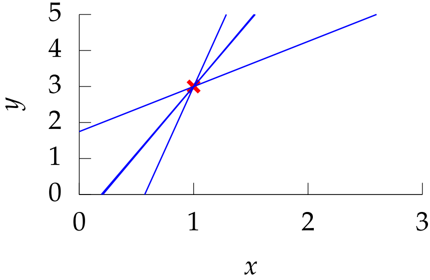
$$c = -4.01 \implies m = 7.01$$



Underdetermined System

Can compute m given c.

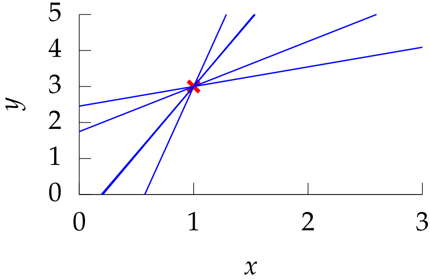
$$c = -0.718 \implies m = 3.72$$



Underdetermined System

Can compute m given c.

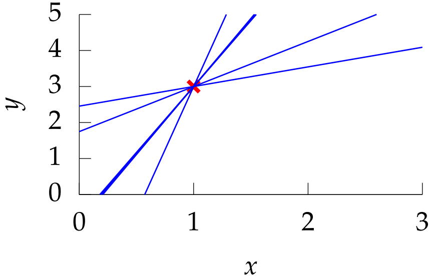
$$c = 2.45 \Longrightarrow m = 0.545$$



Underdetermined System

Can compute m given c.

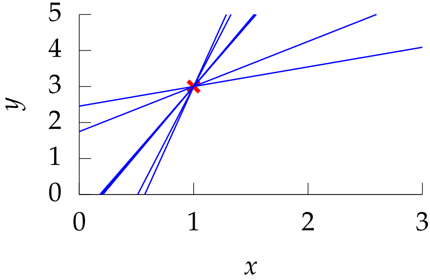
$$c = -0.657 \implies m = 3.66$$



Underdetermined System

Can compute m given c.

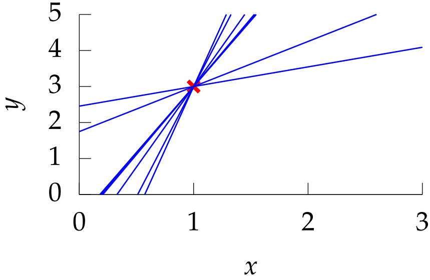
$$c = -3.13 \implies m = 6.13$$



Underdetermined System

Can compute m given c.

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

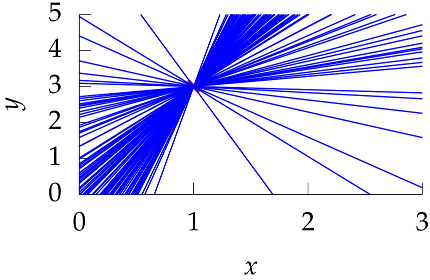


Underdetermined System

Can compute m given c. Assume

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

we find a distribution of solutions.



Bayesian Perspective

Neil D. Lawrence

GPRS
19th–22nd January 2015

Outline

- Regression
- **Bayesian Perspective**
- Gaussian Processes
- Multiple Output Processes
- Latent Force Models
- Approximations
- Dimensionality Reduction

Outline

Bayesian Approach

- Likelihood for the regression example has the form

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\mathbf{w}^\top \phi_i, \sigma^2\right).$$

- Suggestion was to maximize this likelihood with respect to $\mathbf{w}$.
- This can be done with gradient based optimization of the log likelihood.
- Alternative approach: integration across $\mathbf{w}$.
- Consider expected value of likelihood under a range of potential $\mathbf{w}$s.
- This is known as the *Bayesian* approach.

Note on the Term Bayesian

- We will use Bayes' rule to invert probabilities in the Bayesian approach.
  - Bayesian is not named after Bayes' rule (v. common confusion).
  - The term Bayesian refers to the treatment of the parameters as stochastic variables.
  - This approach was proposed by Laplace (1774) and Bayes (1763) independently.
  - For early statisticians this was very controversial (Fisher et al).

Bayesian Controversy

- Bayesian controversy relates to treating *epistemic* uncertainty as *aleatoric* uncertainty.
- Another analogy:
  - Before a football match the uncertainty about the result is *aleatoric*.
  - If I watch a recorded match *without* knowing the result the uncertainty is *epistemic*.

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

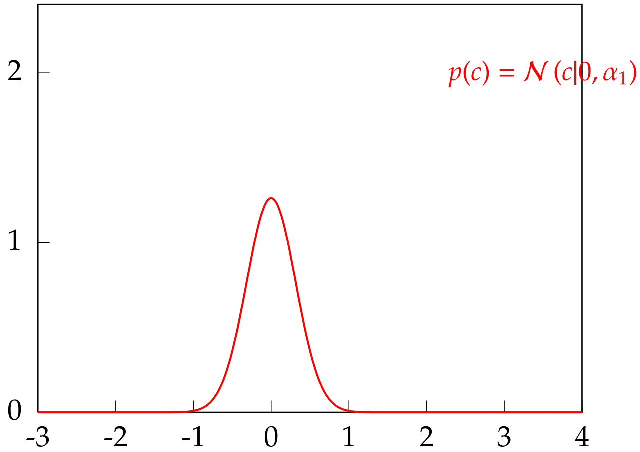


Figure : A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Bayes Update

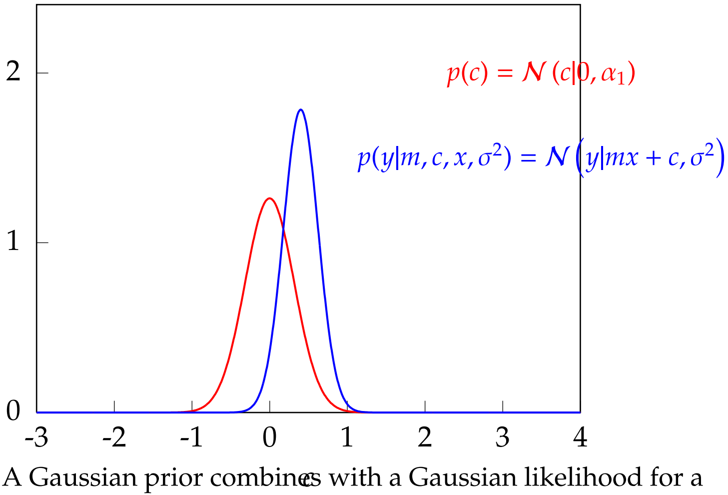


Figure : A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Bayes Update

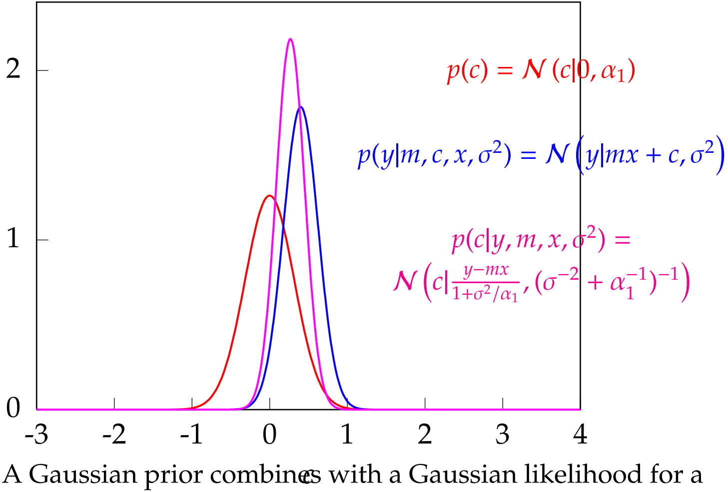


Figure : A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

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.
- We will use the multivariate Gaussian to put a prior *directly* on the function (a Gaussian process).

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.
- We will use the multivariate Gaussian to put a prior *directly* on the function (a Gaussian process).

Two Dimensional Gaussian

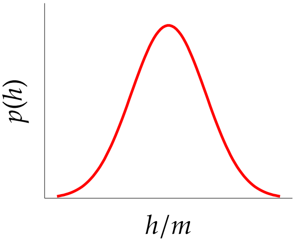
- Consider height, h/m and weight, w/kg.
- Could sample height from a distribution:

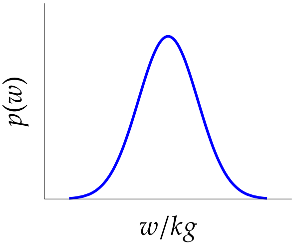
$$p(h) \sim \mathcal{N}(1.7, 0.0225)$$

- And similarly weight:

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

Height and Weight Models



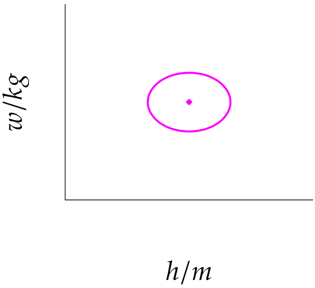


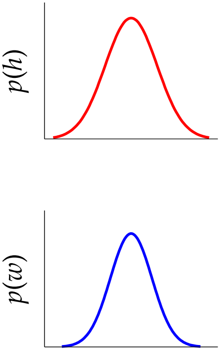
Gaussian distributions for height and weight.

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



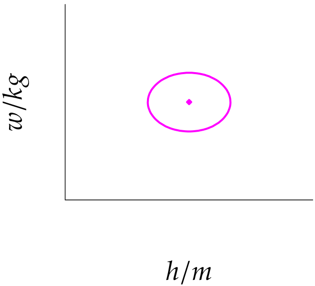


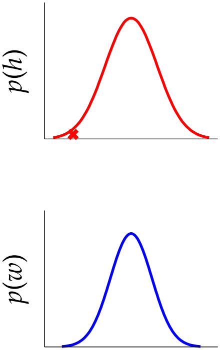
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



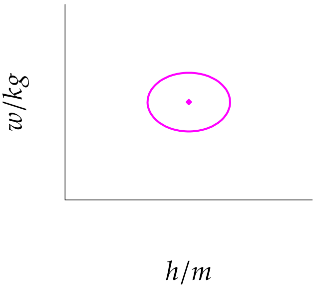


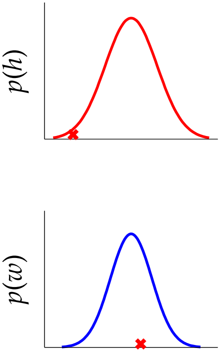
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

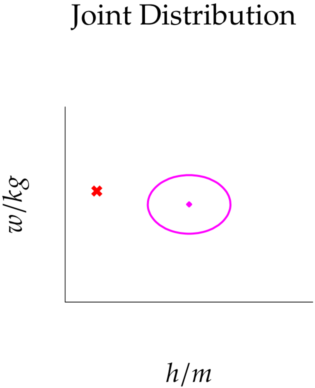
Joint Distribution

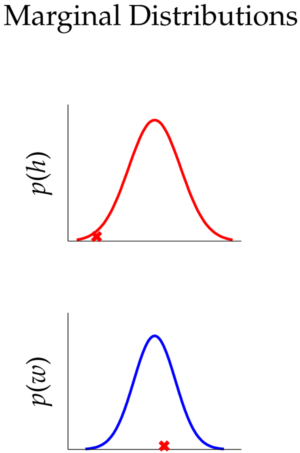




Samples of height and weight

Sampling Two Dimensional Variables



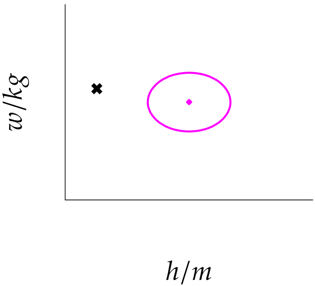


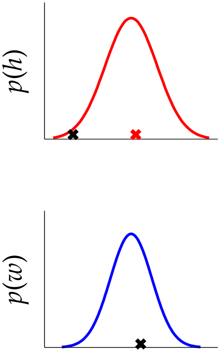
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



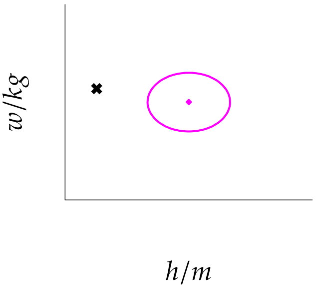


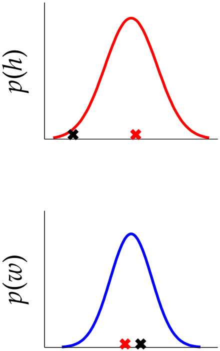
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



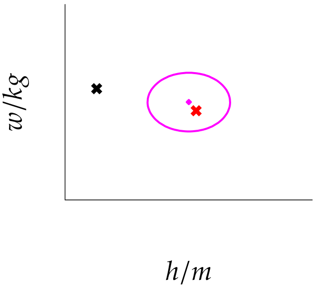


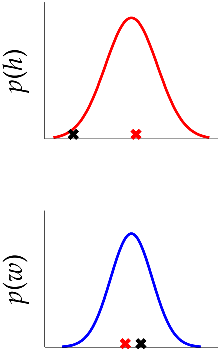
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution

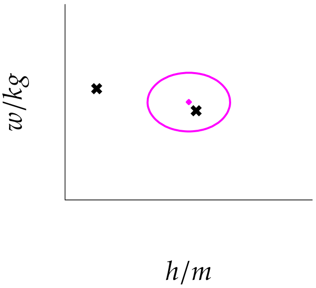




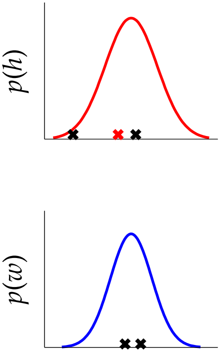
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



Marginal Distributions

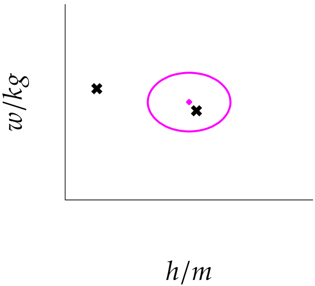


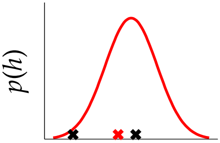
Samples of height and weight

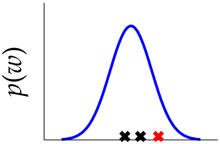
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution





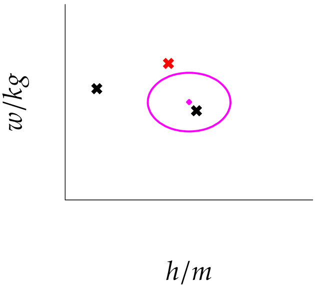


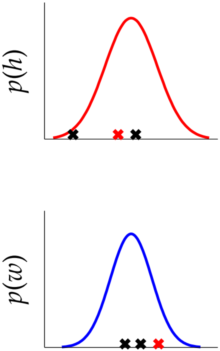
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



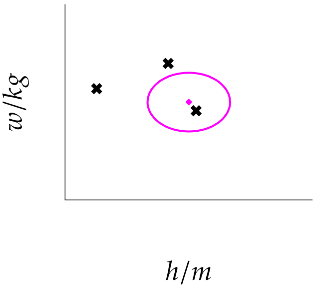


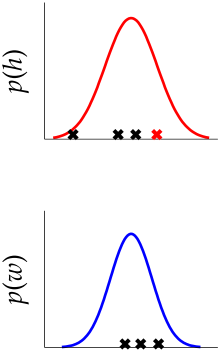
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

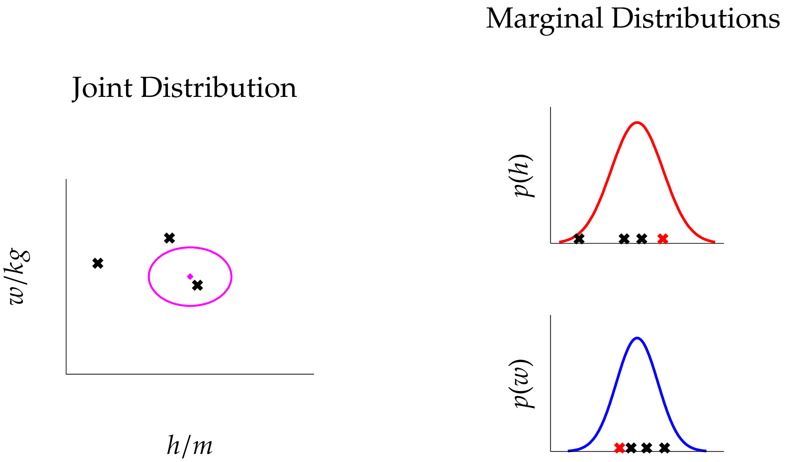
Joint Distribution





Samples of height and weight

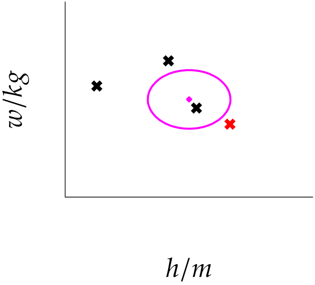
Sampling Two Dimensional Variables



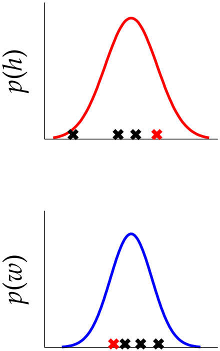
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



Marginal Distributions

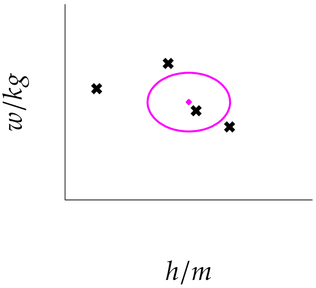


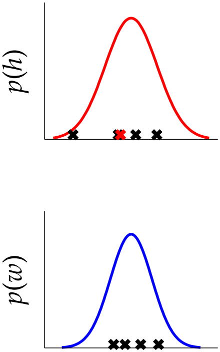
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

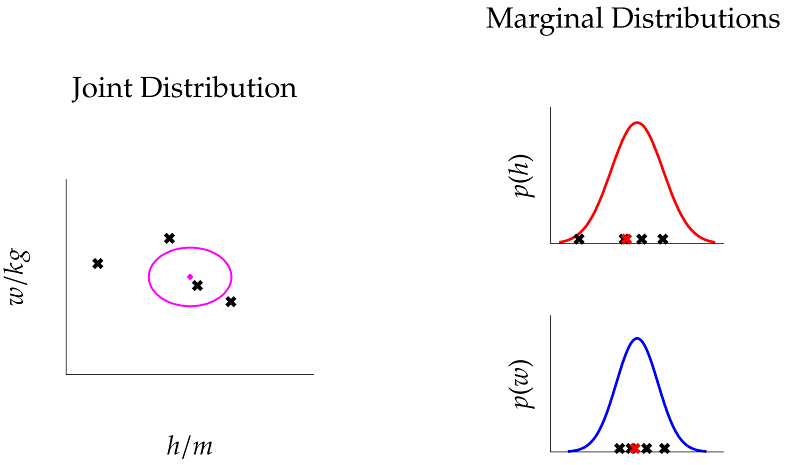
Joint Distribution





Samples of height and weight

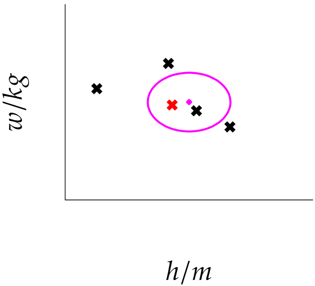
Sampling Two Dimensional Variables



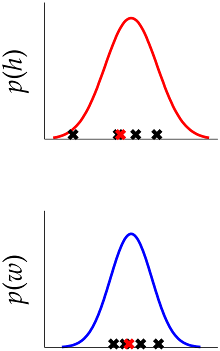
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



Marginal Distributions

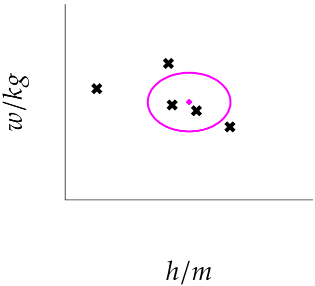


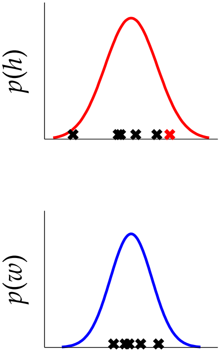
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

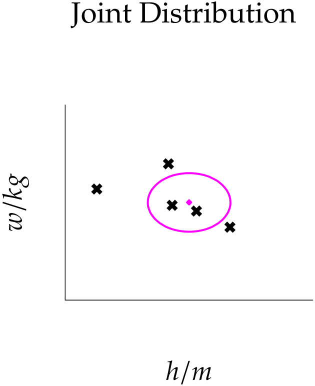
Joint Distribution

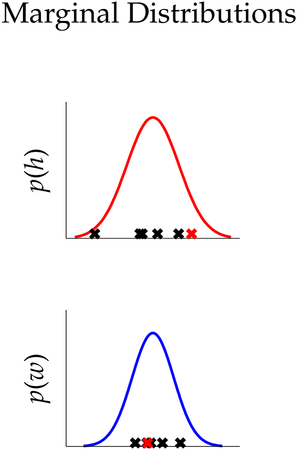




Samples of height and weight

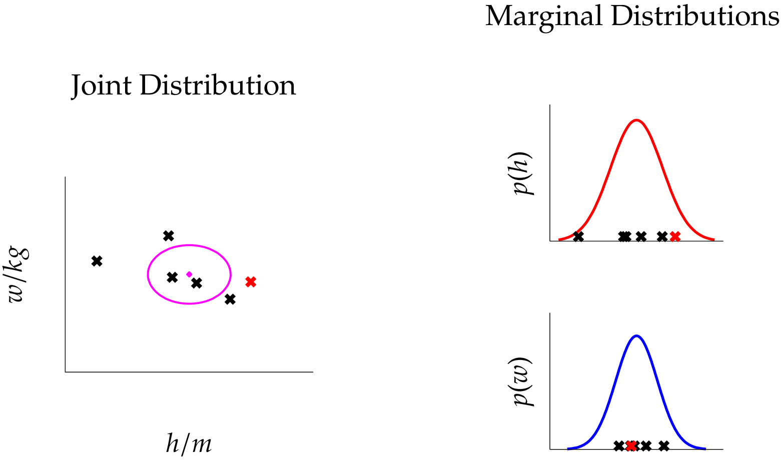
Sampling Two Dimensional Variables





Samples of height and weight

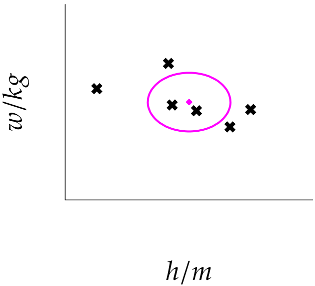
Sampling Two Dimensional Variables



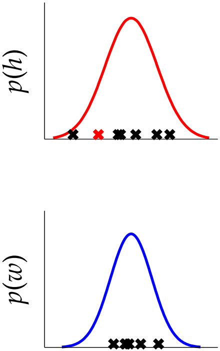
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



Marginal Distributions

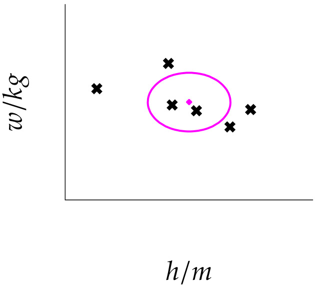


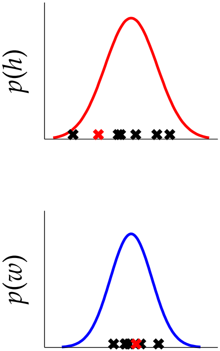
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



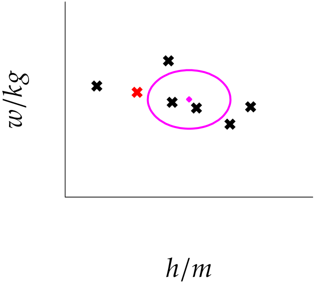


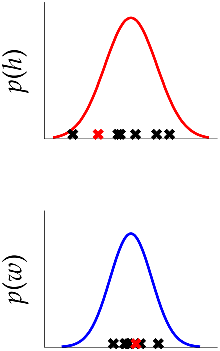
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution

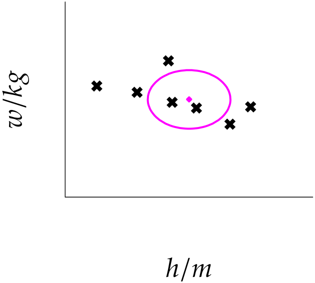




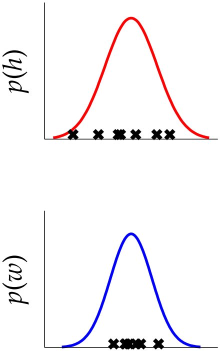
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



Marginal Distributions



Samples of height and weight

Independence Assumption

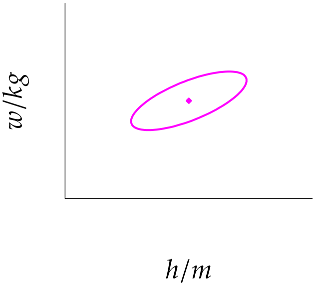
- This assumes height and weight are independent.

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

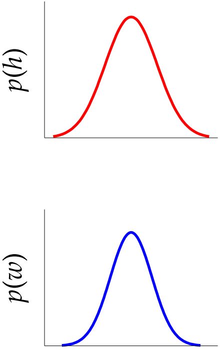
- In reality they are dependent (body mass index) $= \frac{w}{h^2}$.

Sampling Two Dimensional Variables

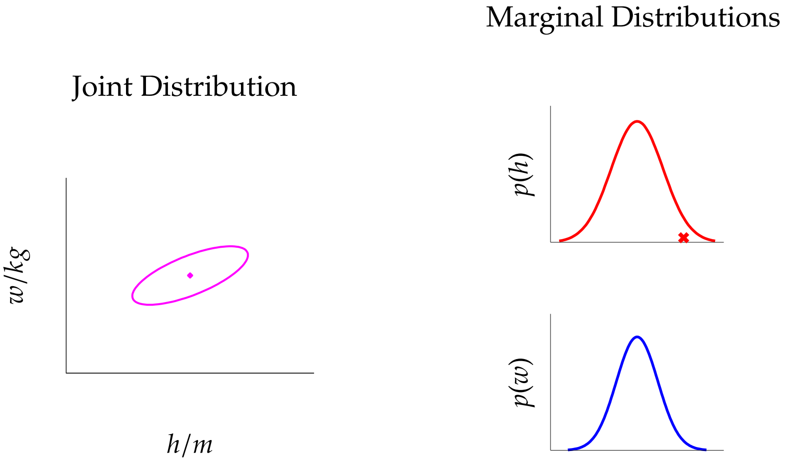
Joint Distribution



Marginal Distributions

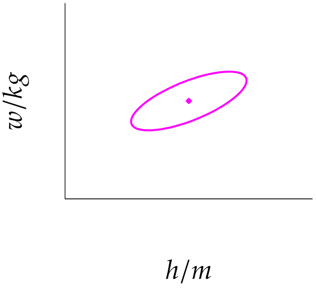


Sampling Two Dimensional Variables

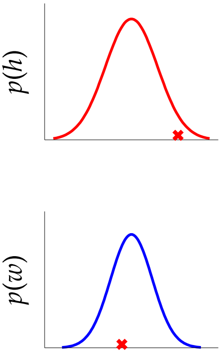


Sampling Two Dimensional Variables

Joint Distribution

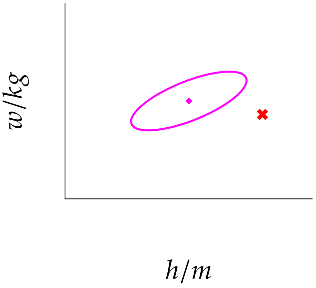


Marginal Distributions

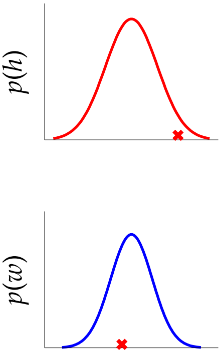


Sampling Two Dimensional Variables

Joint Distribution

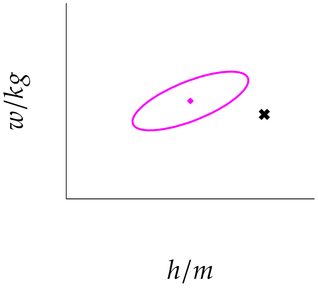


Marginal Distributions

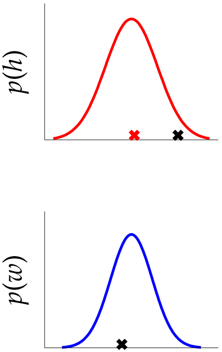


Sampling Two Dimensional Variables

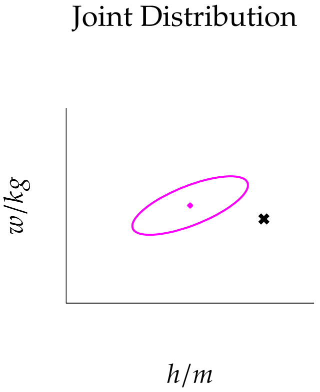
Joint Distribution

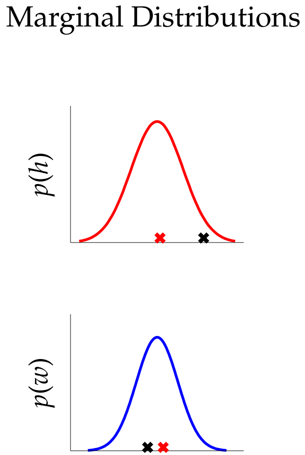


Marginal Distributions

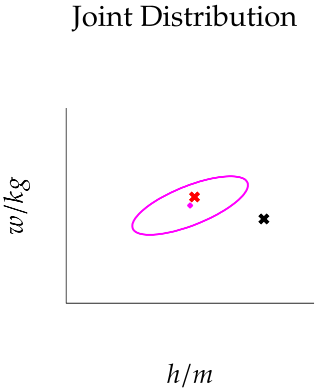


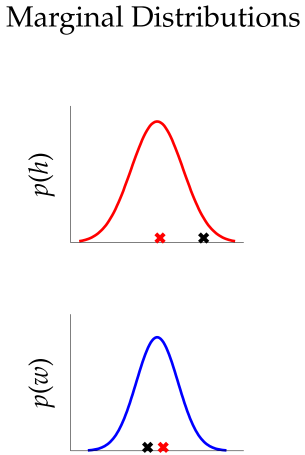
Sampling Two Dimensional Variables





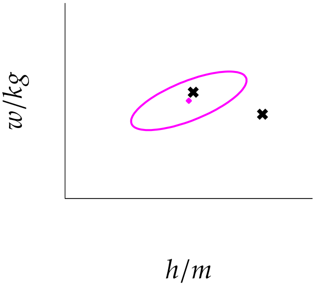
Sampling Two Dimensional Variables



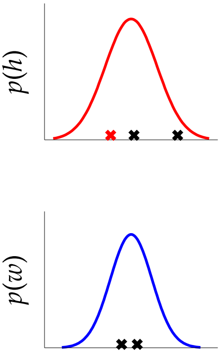


Sampling Two Dimensional Variables

Joint Distribution



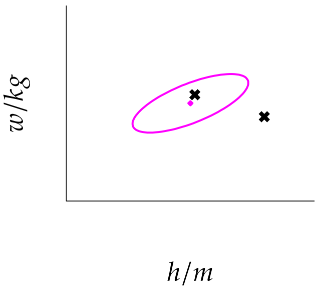
Marginal Distributions

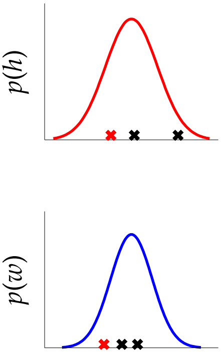


Sampling Two Dimensional Variables

Marginal Distributions

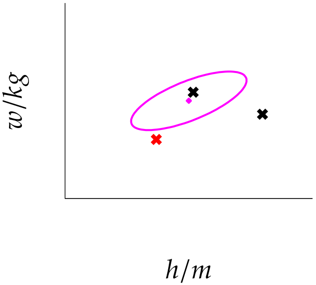
Joint Distribution



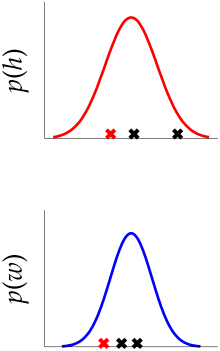


Sampling Two Dimensional Variables

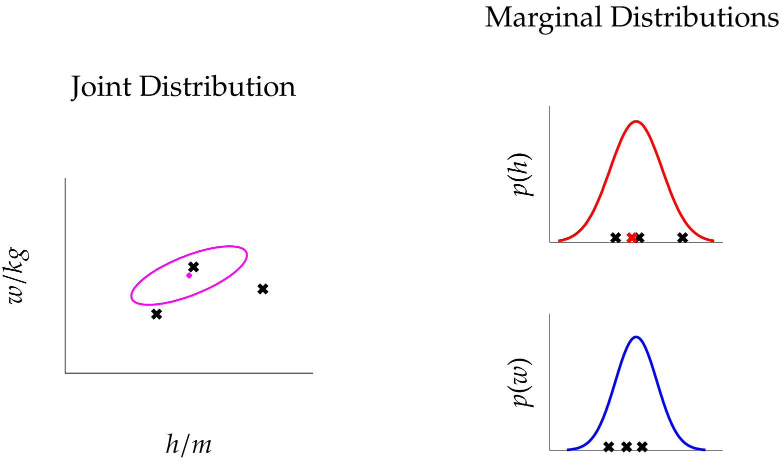
Joint Distribution



Marginal Distributions

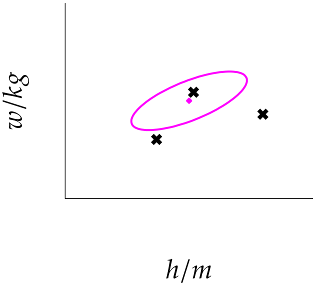


Sampling Two Dimensional Variables

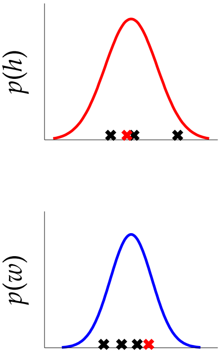


Sampling Two Dimensional Variables

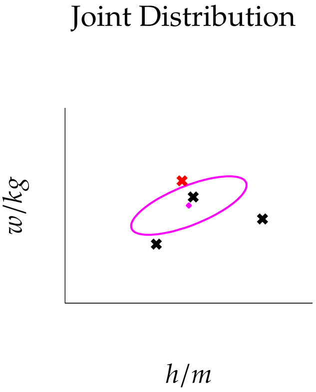
Joint Distribution

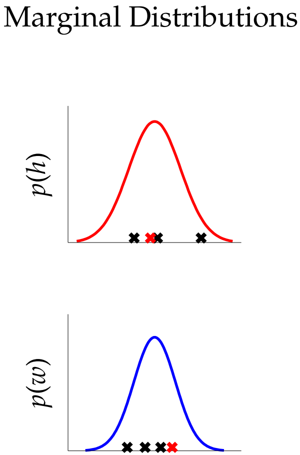


Marginal Distributions

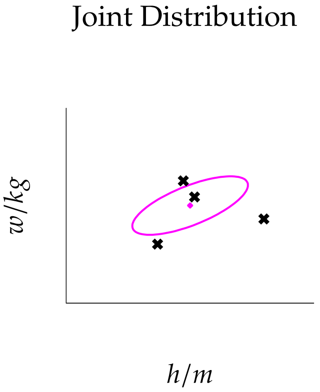


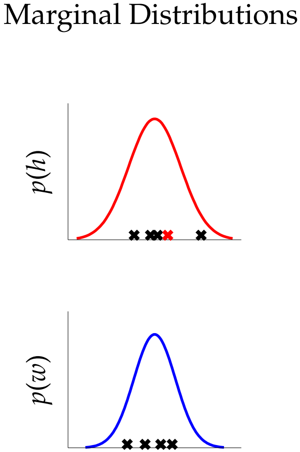
Sampling Two Dimensional Variables





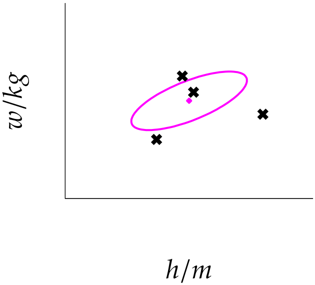
Sampling Two Dimensional Variables



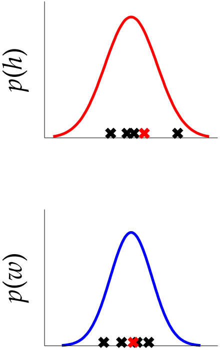


Sampling Two Dimensional Variables

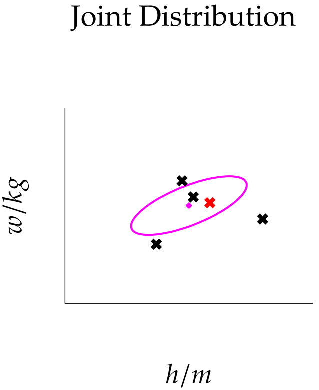
Joint Distribution

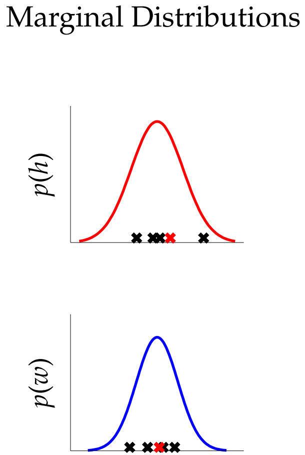


Marginal Distributions

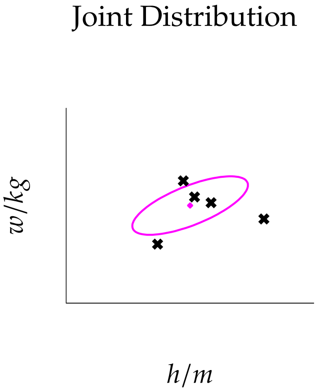


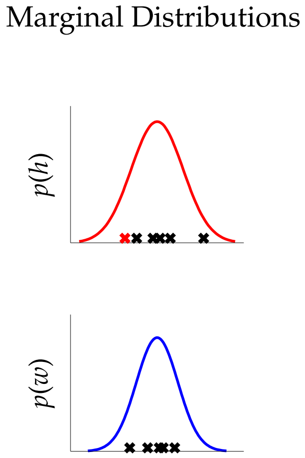
Sampling Two Dimensional Variables



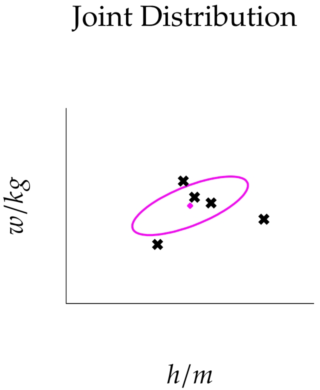


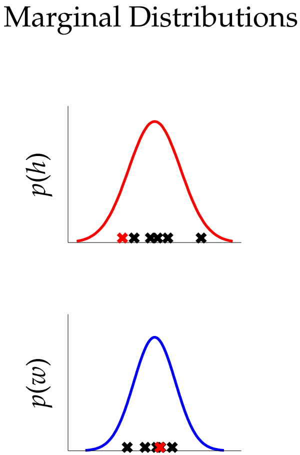
Sampling Two Dimensional Variables



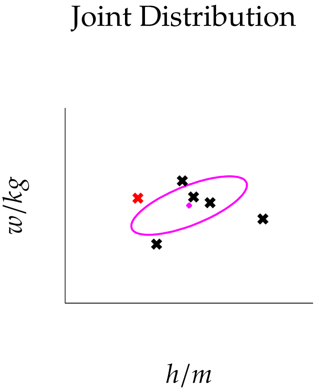


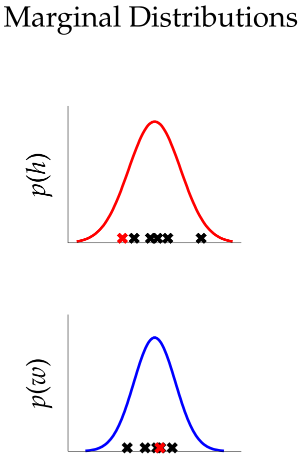
Sampling Two Dimensional Variables



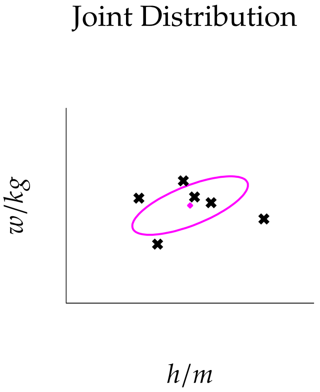


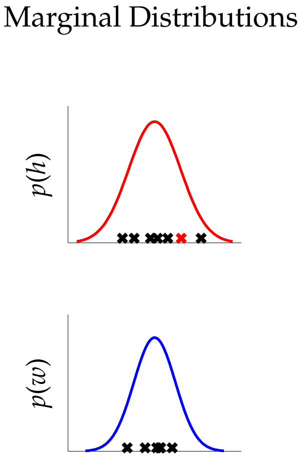
Sampling Two Dimensional Variables



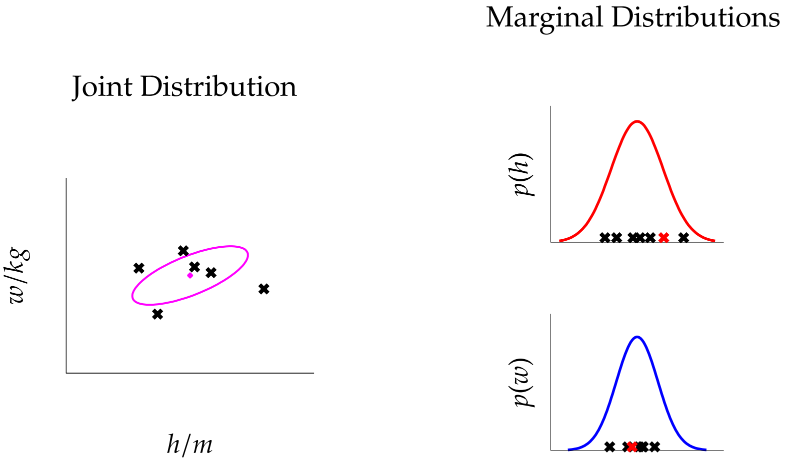


Sampling Two Dimensional Variables



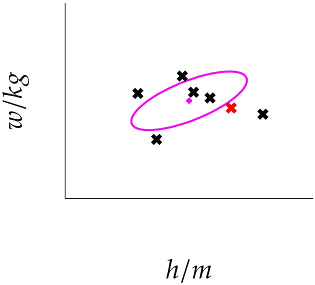


Sampling Two Dimensional Variables

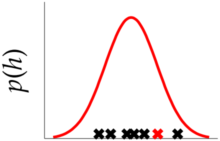


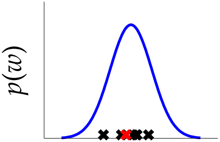
Sampling Two Dimensional Variables

Joint Distribution



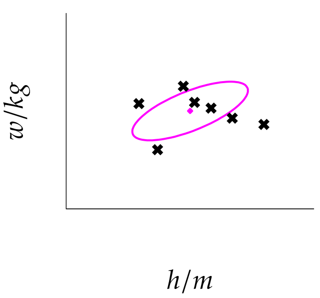
Marginal Distributions



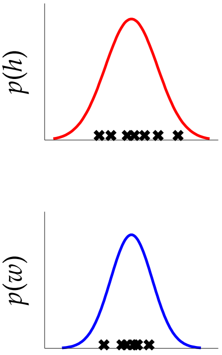


Sampling Two Dimensional Variables

Joint Distribution



Marginal Distributions



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

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

Independent Gaussians

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

Correlated Gaussian

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

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

Correlated Gaussian

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

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

Correlated Gaussian

Form correlated from original by rotating the data space using matrix **R**.

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

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 **R**.

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

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

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

Multivariate Consequence

- If

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

- And

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

Multivariate Consequence

- If

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

- And

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

- Then

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{W}\boldsymbol{\mu}, \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^\top\right)$$

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:

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

Posterior Density

- Once again we want to know the posterior:

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})$$

- And we can compute by completing the square.

Posterior Density

- Once again we want to know the posterior:

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})$$

- And we can compute by completing the square.

$$\log p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i \mathbf{x}_{i,:}^\top \mathbf{w}$$
$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \mathbf{w}^\top \mathbf{x}_{i,:}\mathbf{x}_{i,:}^\top \mathbf{w} - \frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w} + \text{const.}$$

Computing the Posterior

- By inspection we extract the inverse covariance

$$\log p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i \mathbf{x}_{i,:}^\top \mathbf{w}$$
$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \mathbf{w}^\top \mathbf{x}_{i,:}\mathbf{x}_{i,:}^\top \mathbf{w} - \frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w} + \text{const.}$$

- Completing the square allows us to compute the mean.

Computing the Posterior

- By inspection we extract the inverse covariance

$$\log p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = -\frac{1}{2\sigma^2}\mathbf{y}^\top\mathbf{y} + \frac{1}{\sigma^2}\mathbf{y}^\top\mathbf{X}\mathbf{w} - \frac{1}{2\sigma^2}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - \frac{1}{2\alpha}\mathbf{w}^\top\mathbf{w} + \text{const.}$$

- Completing the square allows us to compute the mean.

Computing the Posterior

- By inspection we extract the inverse covariance

$$\log p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = -\frac{1}{2\sigma^2}\mathbf{y}^\top\mathbf{y} + \frac{1}{\sigma^2}\mathbf{y}^\top\mathbf{X}\mathbf{w} \\ -\frac{1}{2}\mathbf{w}^\top\left[\sigma^{-1}\mathbf{X}^\top\mathbf{X} + \alpha^{-1}\mathbf{I}\right]\mathbf{w} + \text{const.}$$

- Completing the square allows us to compute the mean.

Making Predictions

- Giving a Gaussian density

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_w, \mathbf{C}_w)$$

$$\mathbf{C}_w = \left[\sigma^{-2}\mathbf{X}^\top\mathbf{X} + \alpha^{-1}\mathbf{I}\right]^{-1} \qquad \boldsymbol{\mu}_w = \mathbf{C}_w\sigma^{-2}\mathbf{X}^\top\mathbf{y}$$

- Posterior is combined with 'test data' likelihood to make future predictions:

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \int p(y_*|\mathbf{x}_*, \mathbf{w})p(\mathbf{w}|\mathbf{X}, \mathbf{y})\mathrm{d}\mathbf{w}$$

Bayesian vs Maximum Likelihood

- Note the similarity between posterior mean

$$\boldsymbol{\mu}_w = (\sigma^{-2}\mathbf{X}^\top\mathbf{X} + \alpha^{-1}\mathbf{I})^{-1}\sigma^{-2}\mathbf{X}^\top\mathbf{y}$$

- and Maximum likelihood solution

$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Marginal Likelihood

- In some sense though the *real* model is now the marginal likelihood.
- Marginalization of $\mathbf{W}$ follows sum rule

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}$$

  giving

$$p(\mathbf{y}|\mathbf{X}) = \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \alpha\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}\right)$$

- Often the integral is intractable.
- Leads to variational approximations, MCMC (Michael Betancourt, Mark Girolami), Laplace approximation (Harvard Rue).
- For the case of Gaussians it's trivial!!

Marginal Likelihood

- Can compute the marginal likelihood as:

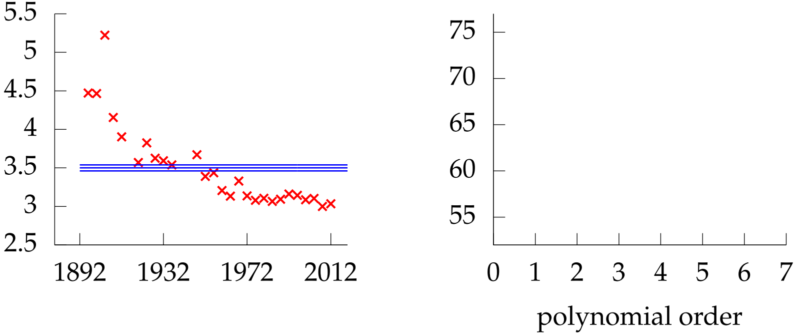
$$p(\mathbf{y}|\mathbf{X}, \alpha, \sigma) = \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \alpha\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}\right)$$

- Or if we use a basis set we have

$$p(\mathbf{y}|\mathbf{X}, \alpha, \sigma) = \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \alpha\mathbf{\Phi}\mathbf{\Phi}^\top + \sigma^2\mathbf{I}\right)$$

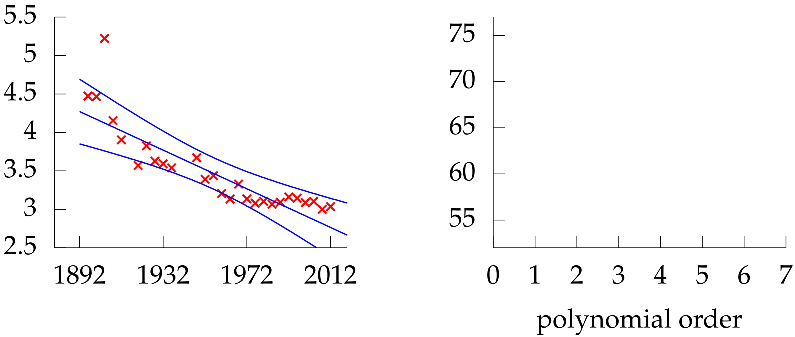
- This Gaussian is no longer i.i.d. across data and *this* is where things get interesting.

Polynomial Fits to Olympics Data



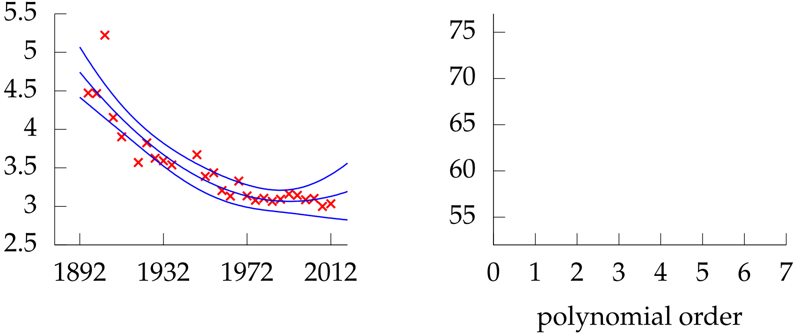
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 0, model error 29.757, $\sigma^2 = 0.286$, $\sigma = 0.535$.

Polynomial Fits to Olympics Data



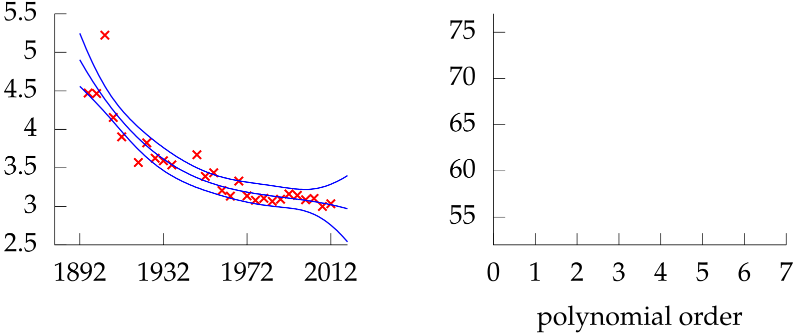
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 1, model error 14.942, $\sigma^2 = 0.0749$, $\sigma = 0.274$.

Polynomial Fits to Olympics Data



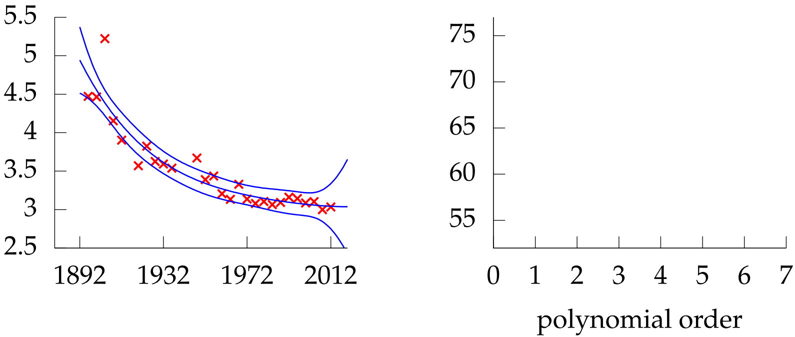
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 2, model error 9.7206, $\sigma^2 = 0.0427$, $\sigma = 0.207$.

Polynomial Fits to Olympics Data



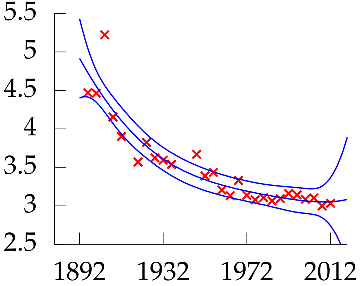
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 3, model error 10.416, $\sigma^2 = 0.0402$, $\sigma = 0.200$.

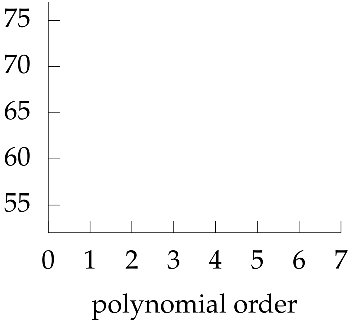
Polynomial Fits to Olympics Data



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

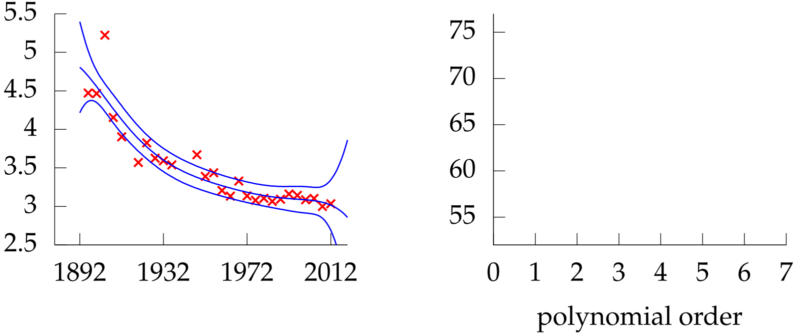
Polynomial Fits to Olympics Data





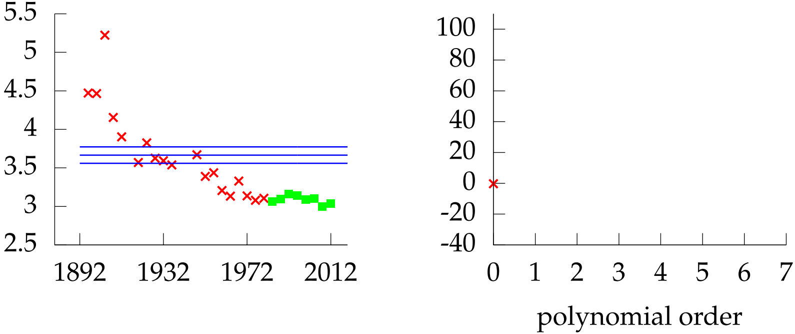
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 5, model error 11.986, $\sigma^2 = 0.0399$, $\sigma = 0.200$.

Polynomial Fits to Olympics Data



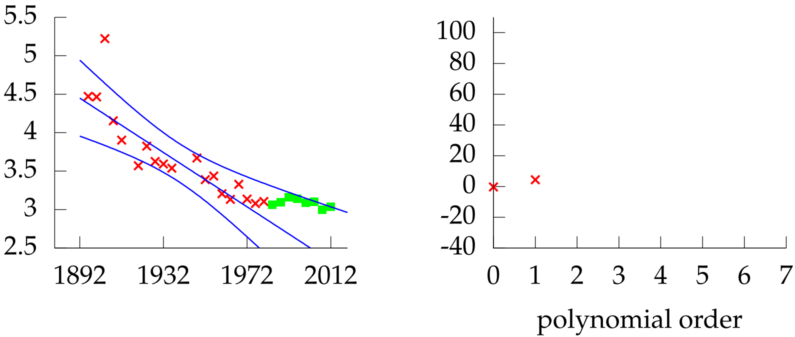
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 6, model error 12.369, $\sigma^2 = 0.0384$, $\sigma = 0.196$.

Validation Set



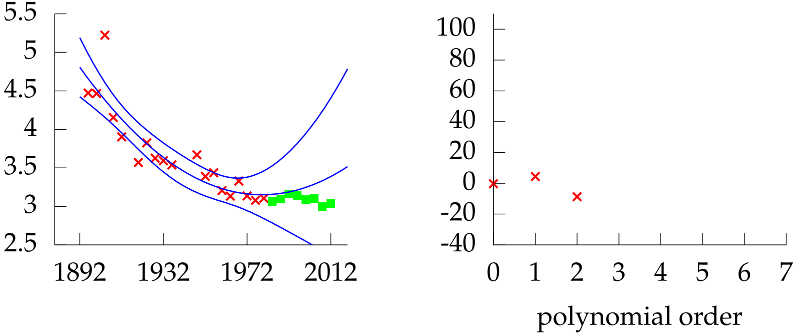
Left: fit to data, *Right*: model error. Polynomial order 0, training error 29.757, validation error -0.29243, $\sigma^2 = 0.302$, $\sigma = 0.550$.

Validation Set



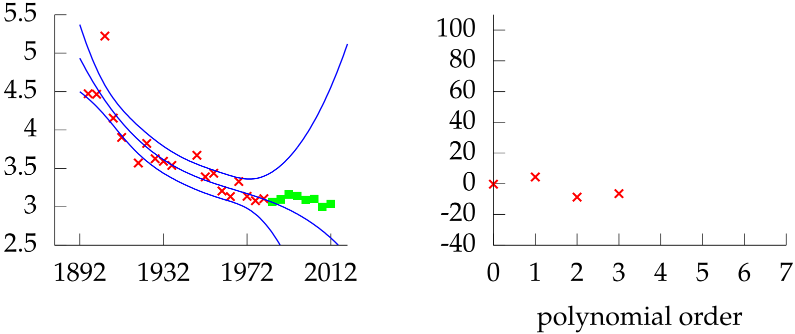
Left: fit to data, *Right*: model error. Polynomial order 1, training error 14.942, validation error 4.4027, $\sigma^2 = 0.0762$, $\sigma = 0.276$.

Validation Set



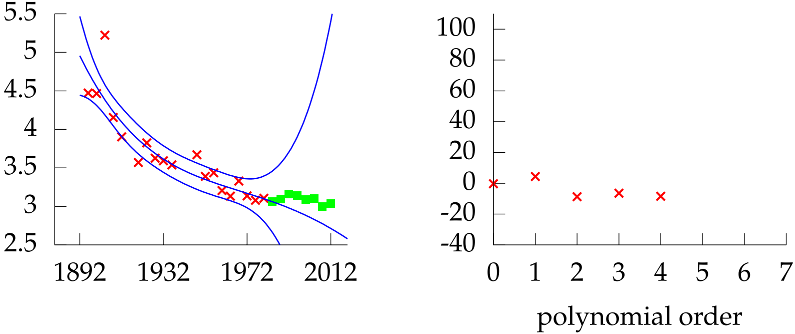
Left: fit to data, *Right*: model error. Polynomial order 2, training error 9.7206, validation error -8.6623, $\sigma^2 = 0.0580$, $\sigma = 0.241$.

Validation Set



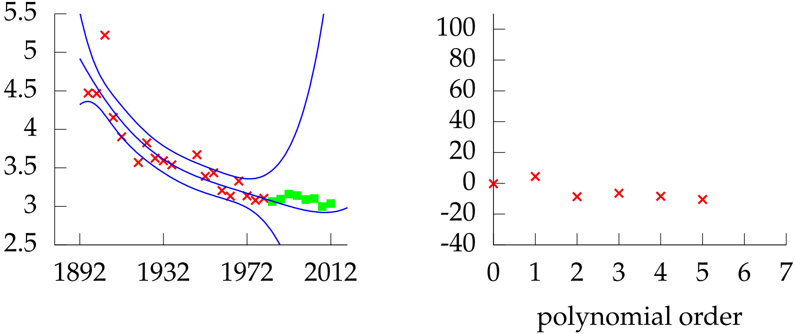
Left: fit to data, *Right*: model error. Polynomial order 3, training error 10.416, validation error -6.4726, $\sigma^2 = 0.0555$, $\sigma = 0.236$.

Validation Set



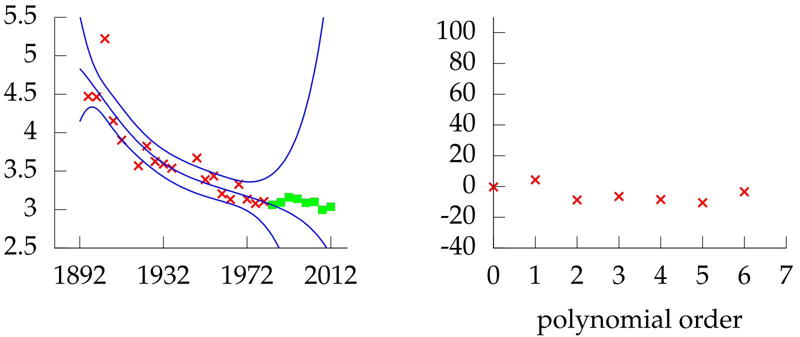
Left: fit to data, *Right*: model error. Polynomial order 4, training error 11.34, validation error -8.431, $\sigma^2 = 0.0555$, $\sigma = 0.236$.

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, $\sigma^2 = 0.0537$, $\sigma = 0.232$.

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).