TDA231
Linear Regression: Modelling the noise

Devdatt Dubhashi
dubhashi@chalmers.se

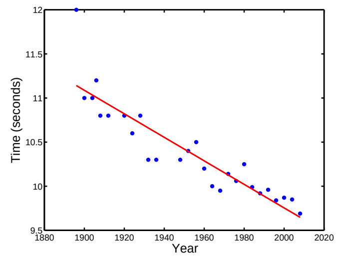
Dept. of Computer Science and Engg.
Chalmers University

March 26, 2018

What about the errors?

$$t = w_0 + w_1 x = \mathbf{w}^\mathsf{T}\mathbf{x}$$

$$t = w_0 + w_1 x + w_2 x^2 + w_3 x^2 + \ldots + w_K x^K = \sum_{k=0}^{K} w_k x^k = \mathbf{w}^\mathsf{T}\mathbf{x}$$



Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

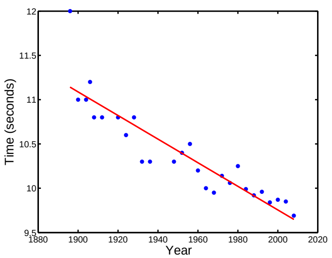
What about the errors?

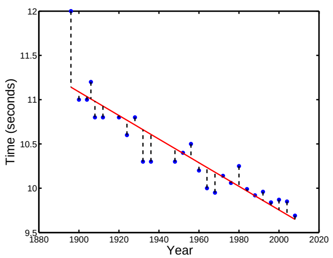
$$t = w_0 + w_1 x = \mathbf{w}^\mathsf{T}\mathbf{x}$$

$$t = w_0 + w_1 x + w_2 x^2 + w_3 x^2 + \ldots + w_K x^K = \sum_{k=0}^{K} w_k x^k = \mathbf{w}^\mathsf{T}\mathbf{x}$$

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n\right)^2$$

$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$





Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

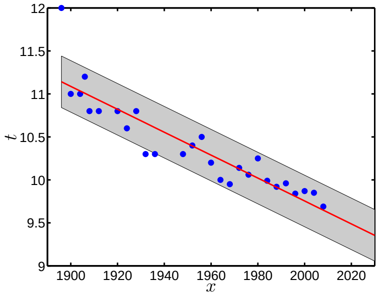
We **should** model the errors

- We know they're there - shouldn't ignore them.

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

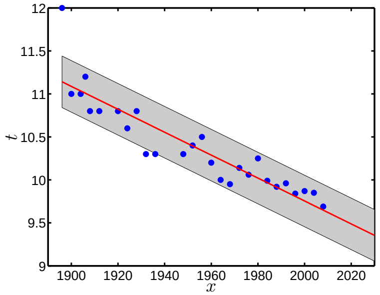
We **should** model the errors

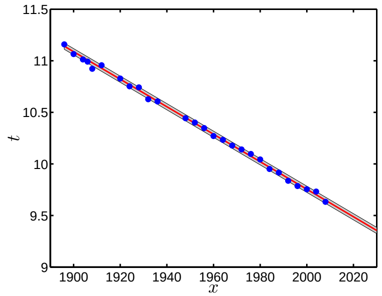
- We know they're there - shouldn't ignore them.
- They tell us how confident our predictions should be:



We **should** model the errors

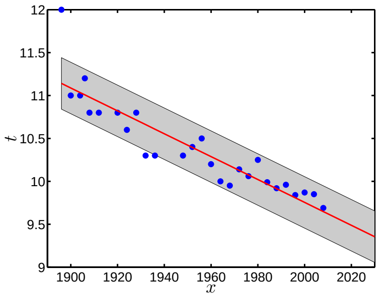
- We know they're there - shouldn't ignore them.
- They tell us how confident our predictions should be:

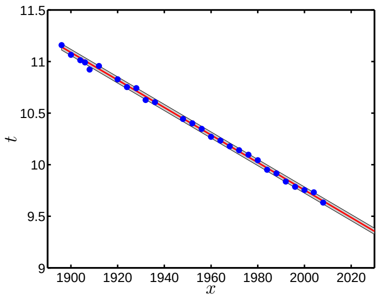




We **should** model the errors

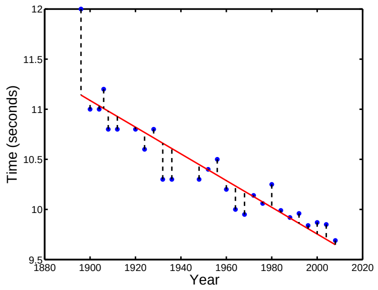
- We know they're there - shouldn't ignore them.
- They tell us how confident our predictions should be:





- ...and other reasons that we will get to later...

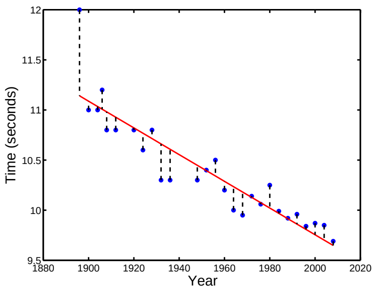
Additive errors



We'll assume that the noise is an additive term in the model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x} + \epsilon_n$$

Additive errors

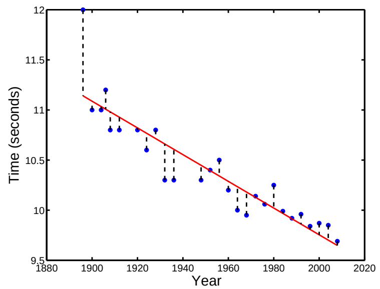


We'll assume that the noise is an additive term in the model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x} + \epsilon_n$$

What assumptions can we make about ϵ_n?

Additive errors



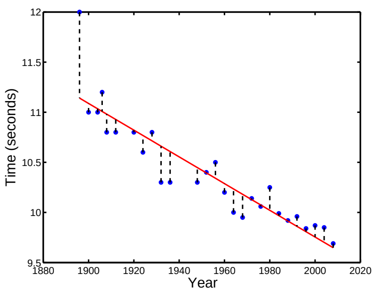
We'll assume that the noise is an additive term in the model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x} + \epsilon_n$$

What assumptions can we make about ϵ_n?

- It's different for each n.

Additive errors



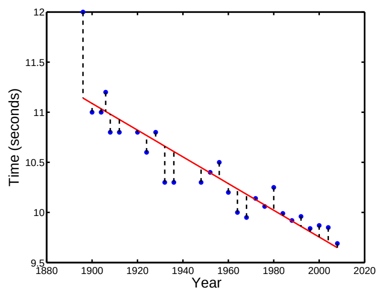
We'll assume that the noise is an additive term in the model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x} + \epsilon_n$$

What assumptions can we make about ϵ_n?

- It's different for each n.
- It's positive and negative.

Additive errors



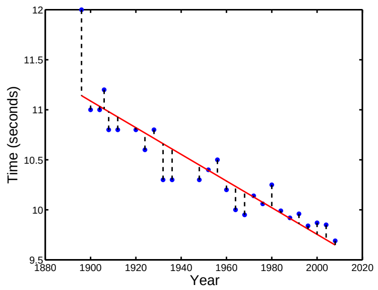
We'll assume that the noise is an additive term in the model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x} + \epsilon_n$$

What assumptions can we make about ϵ_n?

- It's different for each n.
- It's positive and negative.
- There doesn't seem to be any relationship between ϵ at different n.

Additive errors



We'll assume that the noise is an additive term in the model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x} + \epsilon_n$$

What assumptions can we make about ϵ_n?

- It's different for each n.
- It's positive and negative.
- There doesn't seem to be any relationship between ϵ at different n.
- Looks very hard to model exactly (if it were, it wouldn't be noise!)

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Gaussian noise model

- Our model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x} + \epsilon_n$$

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Gaussian noise model

- Our model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x} + \epsilon_n$$

- ϵ_n is continuous.
- We need to choose $p(\epsilon)$.

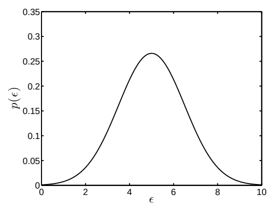
Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Gaussian noise model

- Our model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x} + \epsilon_n$$

- ϵ_n is continuous.
- We need to choose $p(\epsilon)$.
- Gaussian:



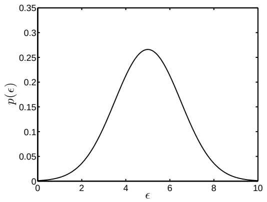
$$p(\epsilon|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\epsilon-\mu)^2\right\}$$

Gaussian noise model

- Our model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x} + \epsilon_n$$

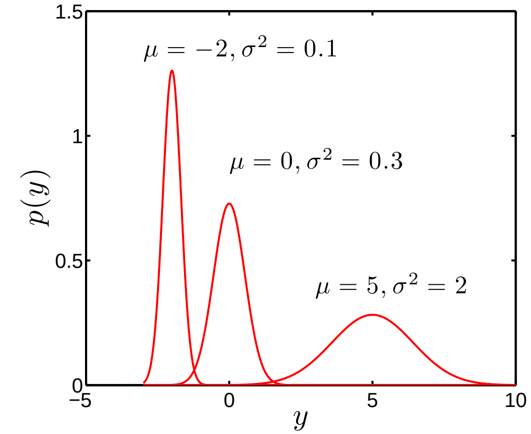
- ϵ_n is continuous.
- We need to choose $p(\epsilon)$.
- Gaussian:



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

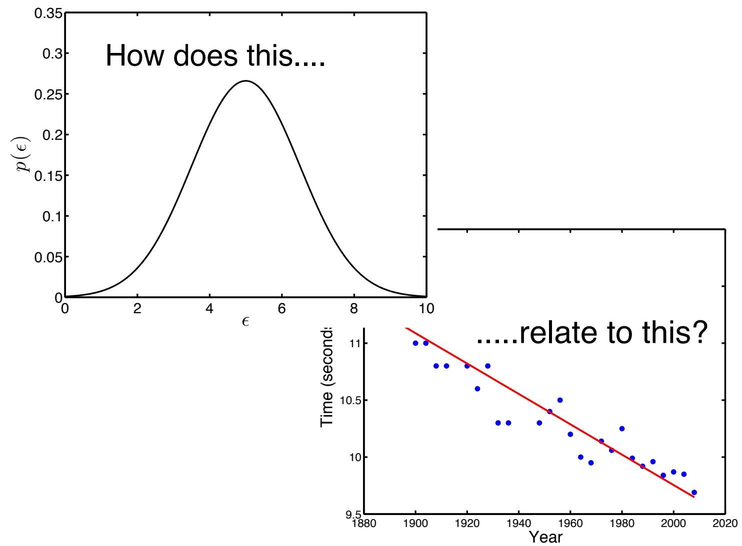
- 2 parameters: Mean μ and Variance σ^2.

Gaussian examples



Effect of varying the mean (μ) and variance (σ^2) parameters of the Gaussian.

Generating data



Likelihood

- Evaluate the density:

$$p(t|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

- at $t = t_n$ is called for the Likelihood.

Likelihood

- Evaluate the density:

$$p(t|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

- at $t = t_n$ is called for the Likelihood.
- The higher the value, the more likely t_n is given the model....

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Likelihood

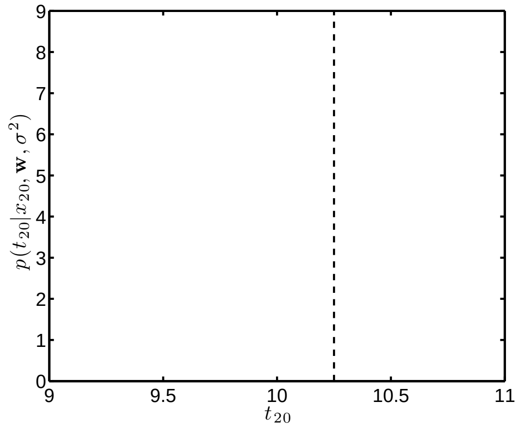
- Evaluate the density:

$$p(t|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

- at $t = t_n$ is called for the Likelihood.
- The higher the value, the more likely t_n is given the model....
  - ....the better the model is.

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

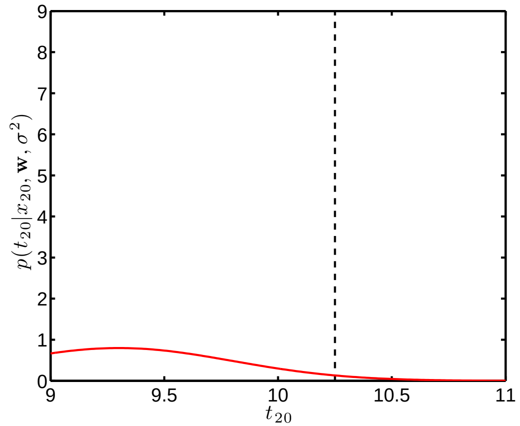
Likelihood



Lets look at the 1980 Olympics ($n = 20$).
Dashed line shows t_{20}.

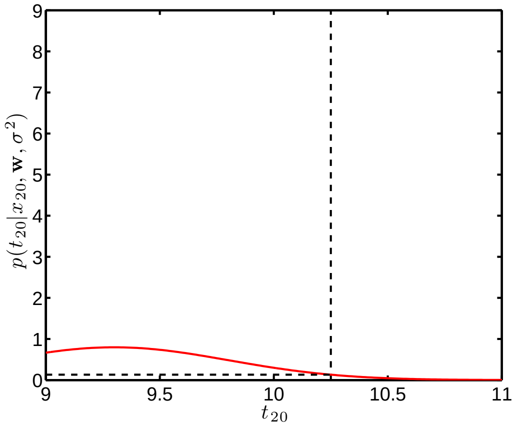
Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Likelihood



Model 1. Red line shows $\mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$

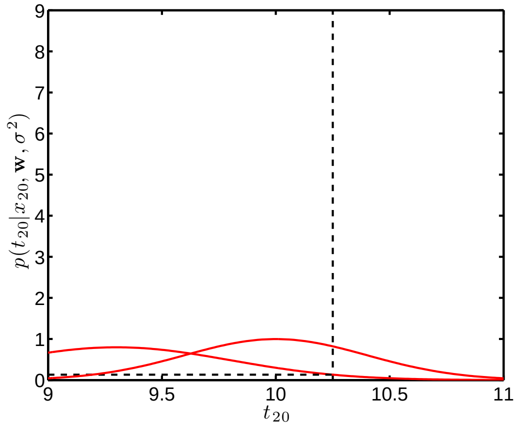
Likelihood



$p(t_{20}|\dots) \approx 0.1.$

Likelihood

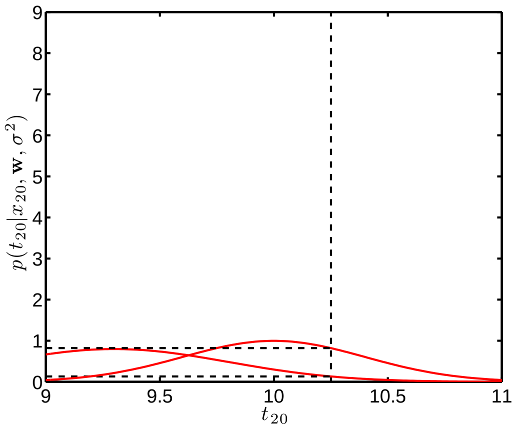
Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary



Model 2. Red line shows $\mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$ for a different $\mathbf{w}$

Likelihood

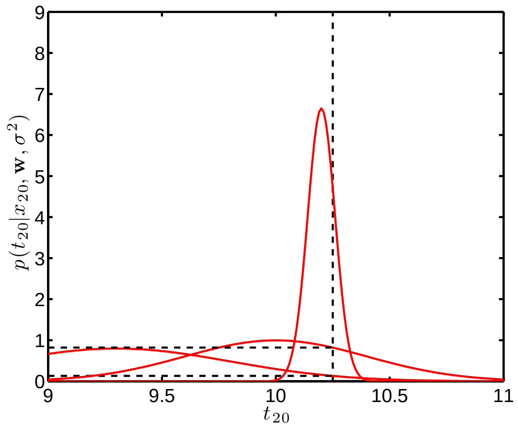
Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary



$p(t_{20}|\dots) \approx 0.9.$

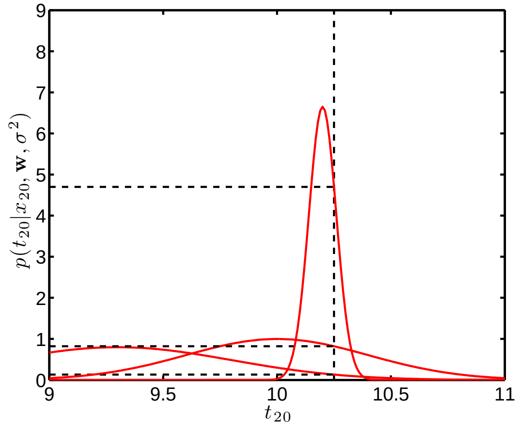
Likelihood

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary



Model 3.

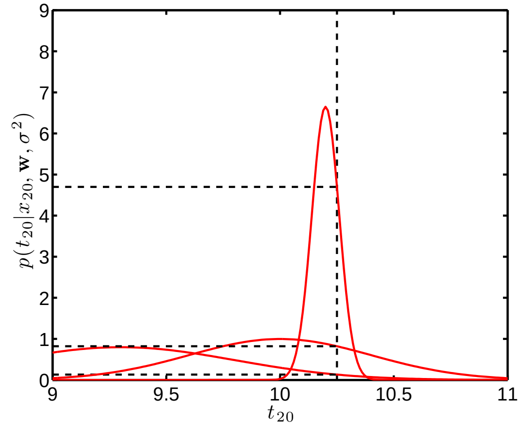
Likelihood



Model 3.

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Likelihood



Model 3.

Model 3 looks best.

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Likelihood

- The value we get when we evaluate the density function is called the **likelihood**.

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Likelihood

- The value we get when we evaluate the density function is called the **likelihood**.
- i.e.
  - The likelihood for model 1 was 0.1.
  - The likelihood for model 2 was 0.9.
  - The likelihood for model 3 was 4.8.
- For continuous random variables, it is **not** a probability!

Likelihood

- The value we get when we evaluate the density function is called the **likelihood**.
- i.e.
  - The likelihood for model 1 was 0.1.
  - The likelihood for model 2 was 0.9.
  - The likelihood for model 3 was 4.8.
- For continuous random variables, it is **not** a probability!
- As t_n is fixed, we can find the values of $\mathbf{w}$ and σ^2 that maximise the likelihood.
  - ...just like we found them that minimised the loss.

Likelihood optimisation

- For each input-response pair, we have a Gaussian likelihood:

$$p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2) = \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Likelihood optimisation

- For each input-response pair, we have a Gaussian likelihood:

$$p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2) = \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

- To combine them all, we want the joint likelihood:

$$p(t_1, \ldots, t_N|\mathbf{w}, \sigma^2, \mathbf{x}_1, \ldots, \mathbf{x}_N)$$

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Likelihood optimisation

- For each input-response pair, we have a Gaussian likelihood:

$$p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2) = \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

- To combine them all, we want the joint likelihood:

$$p(t_1, \ldots, t_N|\mathbf{w}, \sigma^2, \mathbf{x}_1, \ldots, \mathbf{x}_N)$$

- Assume that the t_n are independent:

$$p(t_1, \ldots, t_N|\mathbf{w}, \sigma^2, \mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{n=1}^{N} p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2)$$

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Likelihood optimisation

Finding the parameters that maximise the likelihood is expressed mathematically as:

$$\underset{\mathbf{w},\sigma^2}{\text{argmax}} \prod_{n=1}^{N} p(t_n|\mathbf{w},\mathbf{x}_n,\sigma^2)$$

Likelihood optimisation

Finding the parameters that maximise the likelihood is expressed mathematically as:

$$\underset{\mathbf{w},\sigma^2}{\text{argmax}} \prod_{n=1}^{N} p(t_n|\mathbf{w},\mathbf{x}_n,\sigma^2)$$

In fact, we'll optimise the (natural) log likelihood because it's easier.

- If we increase z, $\log(z)$ increases, if we decrease z, $\log(z)$ decreases. So, at a maximum of z, $\log(z)$ will also be at a maximum.

Likelihood optimisation

Finding the parameters that maximise the likelihood is expressed mathematically as:

$$\underset{\mathbf{w},\sigma^2}{\text{argmax}} \prod_{n=1}^{N} p(t_n|\mathbf{w},\mathbf{x}_n,\sigma^2)$$

In fact, we'll optimise the (natural) log likelihood because it's easier.

- If we increase z, $\log(z)$ increases, if we decrease z, $\log(z)$ decreases. So, at a maximum of z, $\log(z)$ will also be at a maximum.

$$\underset{\mathbf{w},\sigma^2}{\text{argmax}} \log \prod_{n=1}^{N} p(t_n|\mathbf{w},\mathbf{x}_n,\sigma^2)$$

Some re-arranging...

$$\begin{aligned} p(t_n|\mathbf{w},\mathbf{x}_n,\sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2\right\} \\ \log L &= \log \prod_{n=1}^{N} p(t_n|\mathbf{w},\mathbf{x}_n,\sigma^2) \end{aligned}$$

Some re-arranging...

$$
\begin{aligned}
p(t_n|\mathbf{w},\mathbf{x}_n,\sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(t_n-\mathbf{w}^\mathsf{T}\mathbf{x}_n)^2\right\}\\
\log L &= \log\prod_{n=1}^{N} p(t_n|\mathbf{w},\mathbf{x}_n,\sigma^2)\\
&= \sum_{n=1}^{N}\log p(t_n|\mathbf{w},\mathbf{x}_n,\sigma^2)\\
&= \sum_{n=1}^{N}\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)-\sum_{n=1}^{N}\frac{1}{2\sigma^2}(t_n-\mathbf{w}^\mathsf{T}\mathbf{x}_n)^2\\
&= -N\log(\sigma\sqrt{2\pi})-\frac{1}{\sigma^2}\sum_{n=1}^{N}(t_n-\mathbf{w}^\mathsf{T}\mathbf{x}_n)^2
\end{aligned}
$$

Looks familiar!

Some re-arranging...

$$
\begin{aligned}
p(t_n|\mathbf{w},\mathbf{x}_n,\sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(t_n-\mathbf{w}^\mathsf{T}\mathbf{x}_n)^2\right\}\\
\log L &= \log\prod_{n=1}^{N} p(t_n|\mathbf{w},\mathbf{x}_n,\sigma^2)\\
&= \sum_{n=1}^{N}\log p(t_n|\mathbf{w},\mathbf{x}_n,\sigma^2)\\
&= \sum_{n=1}^{N}\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)-\sum_{n=1}^{N}\frac{1}{2\sigma^2}(t_n-\mathbf{w}^\mathsf{T}\mathbf{x}_n)^2\\
&= -N\log(\sigma\sqrt{2\pi})-\frac{1}{\sigma^2}\sum_{n=1}^{N}(t_n-\mathbf{w}^\mathsf{T}\mathbf{x}_n)^2
\end{aligned}
$$

Looks familiar! To continue (good exercise):

$$
\frac{\partial\log L}{\partial\mathbf{w}}=0,\quad \frac{\partial\log L}{\partial\sigma^2}=0
$$

A shortcut

The multi-variate Gaussian

$$
\mathbf{y}=\begin{bmatrix} y_1\\ y_2\end{bmatrix},\quad p(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})
$$

$$
\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}
$$

D is number of variables, $|\boldsymbol{\Sigma}|$ is the determinant.

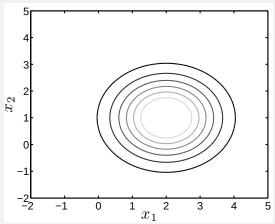
A shortcut

The multi-variate Gaussian

$$
\mathbf{y}=\begin{bmatrix} y_1\\ y_2\end{bmatrix},\quad p(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})
$$

$$
\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}
$$

D is number of variables, $|\boldsymbol{\Sigma}|$ is the determinant.



$$
\boldsymbol{\mu}=\begin{bmatrix}2\\1\end{bmatrix},\quad \boldsymbol{\Sigma}=\begin{bmatrix}1&0\\0&1\end{bmatrix}
$$

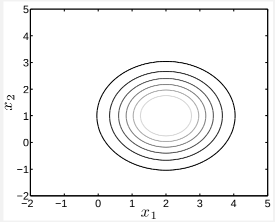
A shortcut

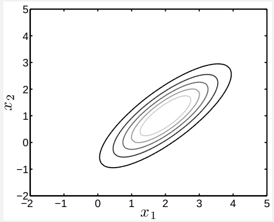
The multi-variate Gaussian

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \; p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right\}$$

D is number of variables, $|\boldsymbol{\Sigma}|$ is the determinant.





$$\boldsymbol{\mu} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \; \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

A shortcut

The multi-variate Gaussian

A special case:

$$\prod_{n=1}^{N} \mathcal{N}(\mu_n, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}, \; \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

So, in our model:

$$\log L = \log \prod_{n=1}^{N} p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2) = \log \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}) = \log p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)$$

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Maximising the multi-variate log-likelihood

- Partial derivative w.r.t. $\mathbf{w}$, set to zero and solve:

$$\begin{aligned} \log L &= \log \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}) \\ \frac{\partial \log L}{\partial \mathbf{w}} &= -\frac{1}{2\sigma^2}(2\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w} - 2\mathbf{X}^{\mathsf{T}}\mathbf{t}) = 0 \\ \mathbf{w} &= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t} \end{aligned}$$

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Maximising the multi-variate log-likelihood

- Partial derivative w.r.t. $\mathbf{w}$, set to zero and solve:

$$\begin{aligned} \log L &= \log \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}) \\ \frac{\partial \log L}{\partial \mathbf{w}} &= -\frac{1}{2\sigma^2}(2\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w} - 2\mathbf{X}^{\mathsf{T}}\mathbf{t}) = 0 \\ \mathbf{w} &= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t} \end{aligned}$$

- This is the same expression we've seen before!

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Maximising the multi-variate log-likelihood

- Partial derivative w.r.t. $\mathbf{w}$, set to zero and solve:

$$\begin{aligned} \log L &= \log \mathcal{N}(\mathbf{Xw}, \sigma^2 \mathbf{I}) \\ \frac{\partial \log L}{\partial \mathbf{w}} &= -\frac{1}{2\sigma^2}(2\mathbf{X}^\mathsf{T}\mathbf{Xw} - 2\mathbf{X}^\mathsf{T}\mathbf{t}) = 0 \\ \mathbf{w} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} \end{aligned}$$

- This is the same expression we've seen before!
- Same for σ^2:

$$\begin{aligned} \frac{\partial \log L}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{(\sigma^2)^2}(\mathbf{t} - \mathbf{Xw})^\mathsf{T}(\mathbf{t} - \mathbf{Xw}) = 0 \\ \sigma^2 &= \frac{1}{N}(\mathbf{t} - \mathbf{Xw})^\mathsf{T}(\mathbf{t} - \mathbf{Xw}) \end{aligned}$$

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Optimum parameters

- Compute optimum $\widehat{\mathbf{w}}$ from:

$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$

- Use this to compute optimum $\widehat{\sigma^2}$ from:

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})$$

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Optimum parameters

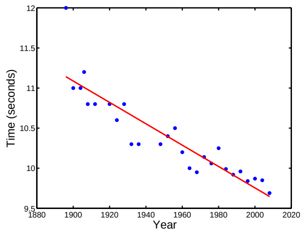
- Compute optimum $\widehat{\mathbf{w}}$ from:

$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$

- Use this to compute optimum $\widehat{\sigma^2}$ from:

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})$$

- e.g. Olympic 100 m data (again!)



$$\widehat{\mathbf{w}} = \begin{bmatrix} 36.416 \\ -0.0133 \end{bmatrix}, \quad \widehat{\sigma^2} = 0.0503$$

Introduction
D. Dubhashi
Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary

Optimum parameters

- We have point estimates of our parameters.
- How confident should we be in them?
  - If we changed them a little bit, would the model still be good?

Confidence in parameter estimates

- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.

Confidence in parameter estimates

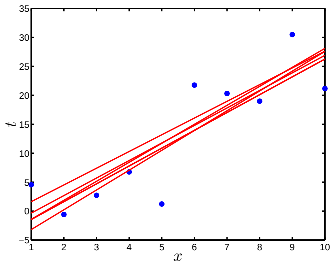
- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.
- How good our our estimates $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$?
  - Are they correct (on average)?
  - If we could keep adding data, would we converge on the true value?

Confidence in parameter estimates

- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.
- How good our our estimates $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$?
  - Are they correct (on average)?
  - If we could keep adding data, would we converge on the true value?
- How confident should we be in our estimates?
  - Could we change parameters a little bit and still have a good model?

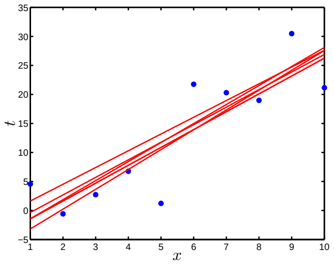
Confidence in parameter estimates

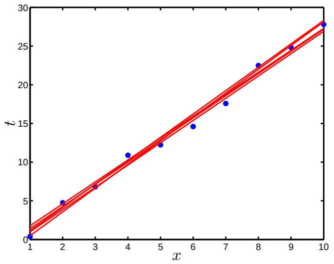
- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.
- How good our our estimates $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$?
  - Are they correct (on average)?
  - If we could keep adding data, would we converge on the true value?
- How confident should we be in our estimates?
  - Could we change parameters a little bit and still have a good model?



Confidence in parameter estimates

- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.
- How good our our estimates $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$?
  - Are they correct (on average)?
  - If we could keep adding data, would we converge on the true value?
- How confident should we be in our estimates?
  - Could we change parameters a little bit and still have a good model?





Summary

- Modelled the error as a random variable.
- Used a Gaussian random variable.
- Maximized the likelihood

Introduction

D. Dubhashi

Introduction
An error term
Adding noise to the model
Likelihood
Confidence in parameter estimates
Summary