7. An Advanced Bayesian Toolbox – Part Two

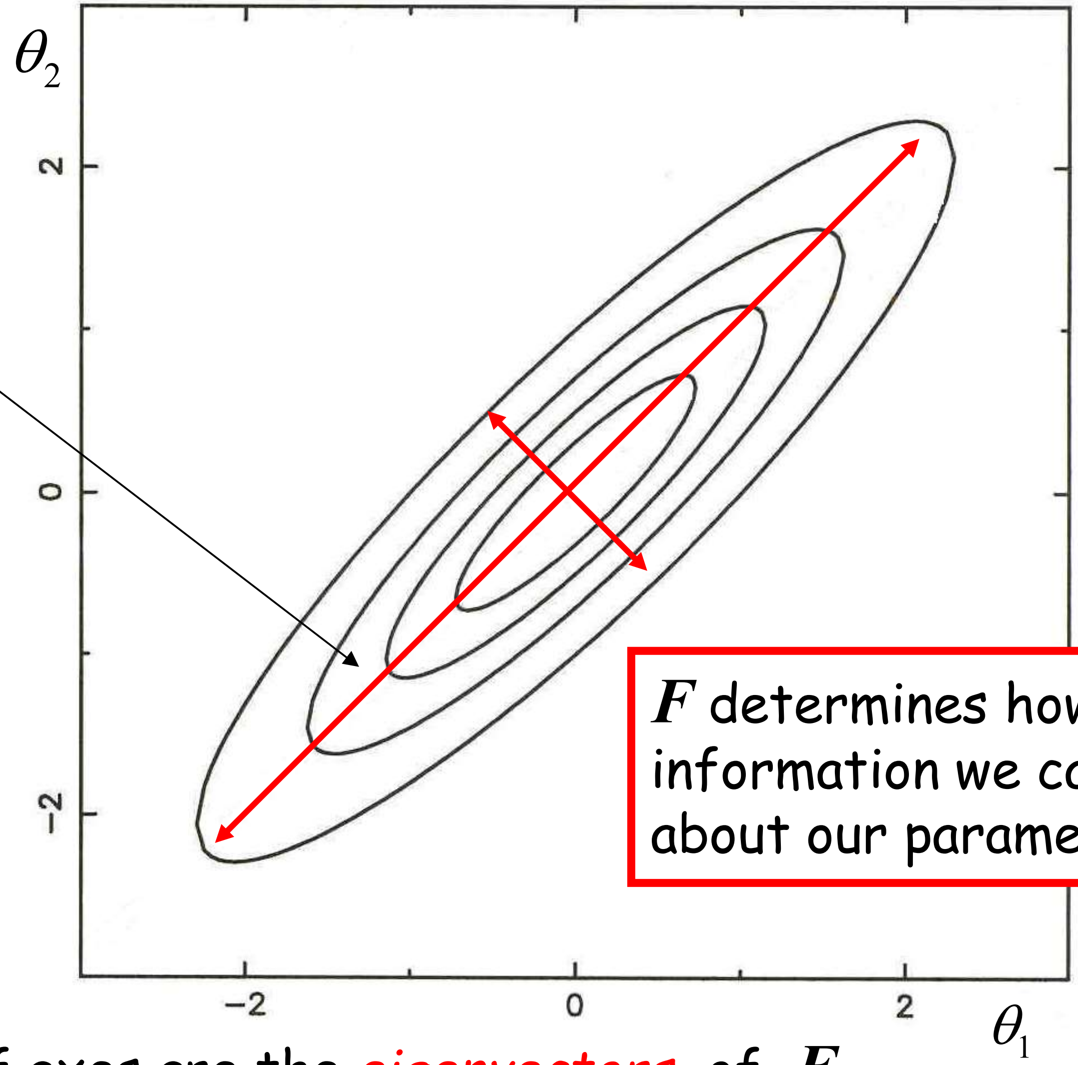
Parameter estimation: 2-D case

Linear combination of θ_1 and θ_2 well constrained by data

Length of axes determined by the **eigenvalues** of the Fisher information matrix

$$F_{ij} = \frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j} = \left[-\sigma_{ij}^2\right]^{-1}$$

$$\boxed{\boldsymbol{F}\,\boldsymbol{\theta} = \lambda\,\boldsymbol{\theta}}$$



$\boldsymbol{F}$ determines how much information we can learn about our parameters

Direction of axes are the *eigenvectors* of $\boldsymbol{F}$

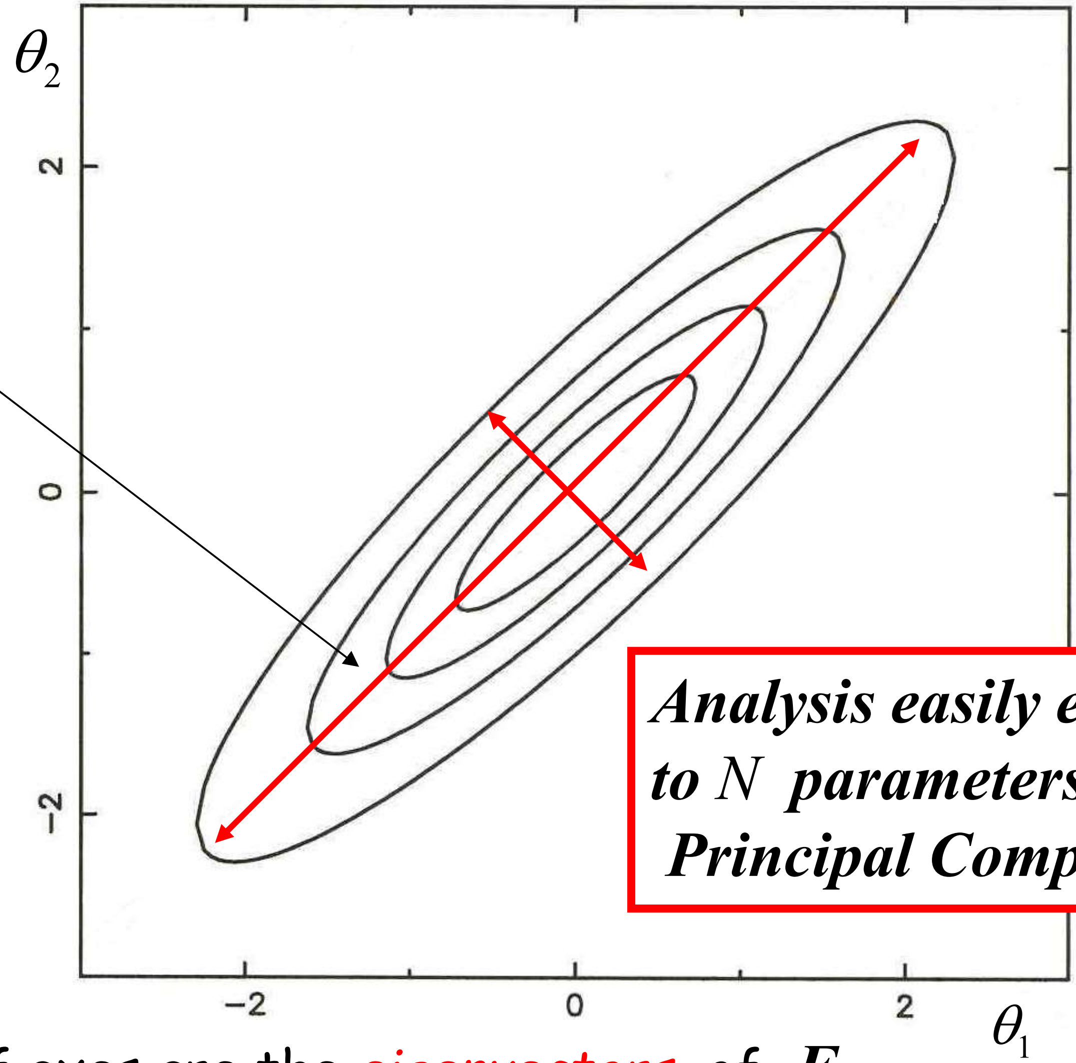
Parameter estimation: N-dimensional case

Linear combination of θ_1 and θ_2 well constrained by data

Length of axes determined by the **eigenvalues** of the Fisher information matrix

$$F_{ij} = \frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j} = \left[-\sigma_{ij}^2\right]^{-1}$$

$$\boxed{\boldsymbol{F\,\theta = \lambda\,\theta}}$$



Analysis easily extends to N parameters – Principal Components

Direction of axes are the *eigenvectors* of $\boldsymbol{F}$

What is PCA?

Principal Component Analysis:

A method for transforming a multi-dimensional dataset, consisting of a number of statistically dependent (correlated) variables, into a set of uncorrelated variables: ***principal components***

What is PCA?

Principal Component Analysis:

1st principal component = *linear combination of the original variables that accounts for as much of the variation in the data as possible*

What is PCA?

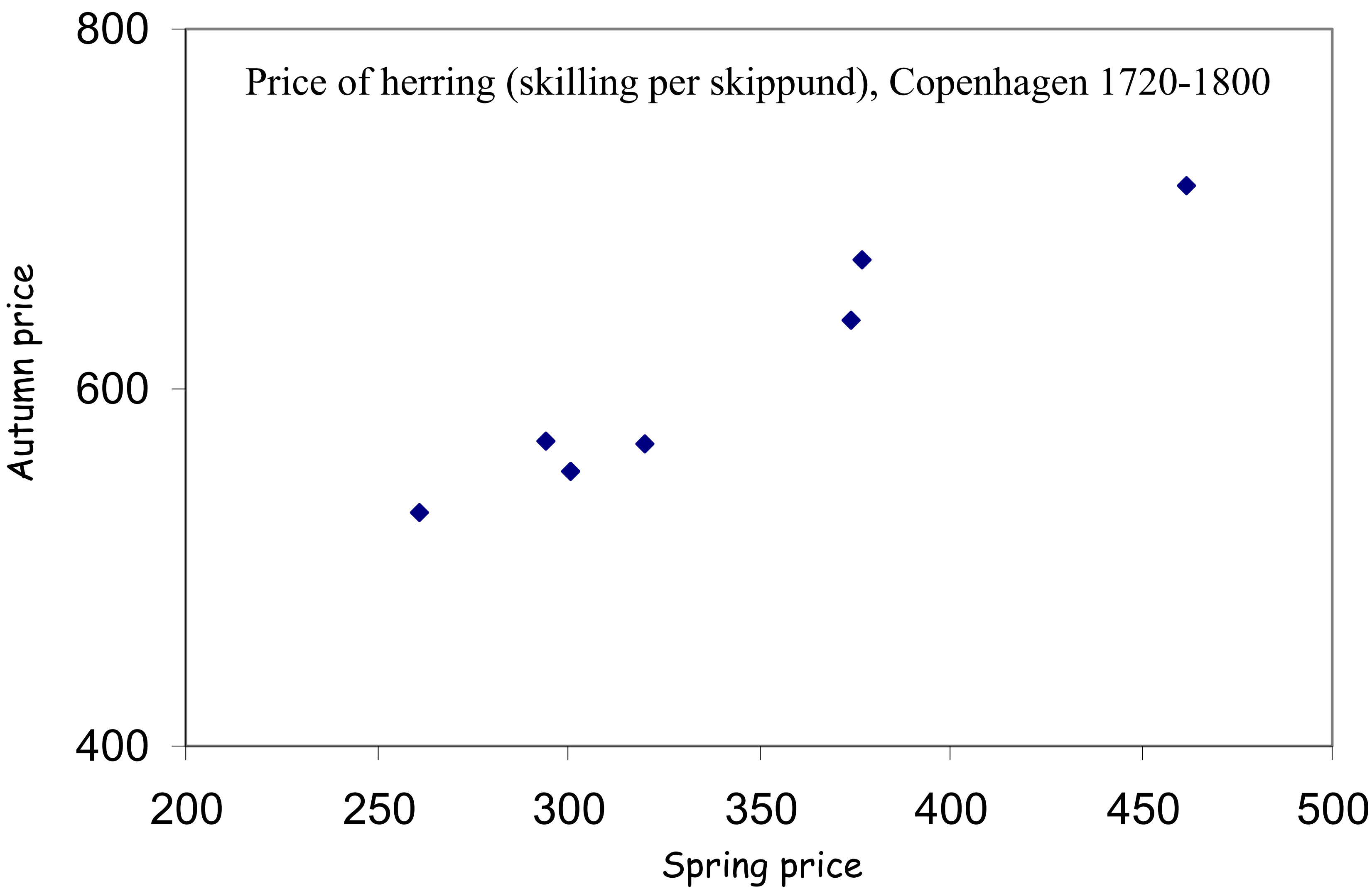
Principal Component Analysis:

1st principal component = *linear combination of the original variables that accounts for as much of the variation in the data as possible*

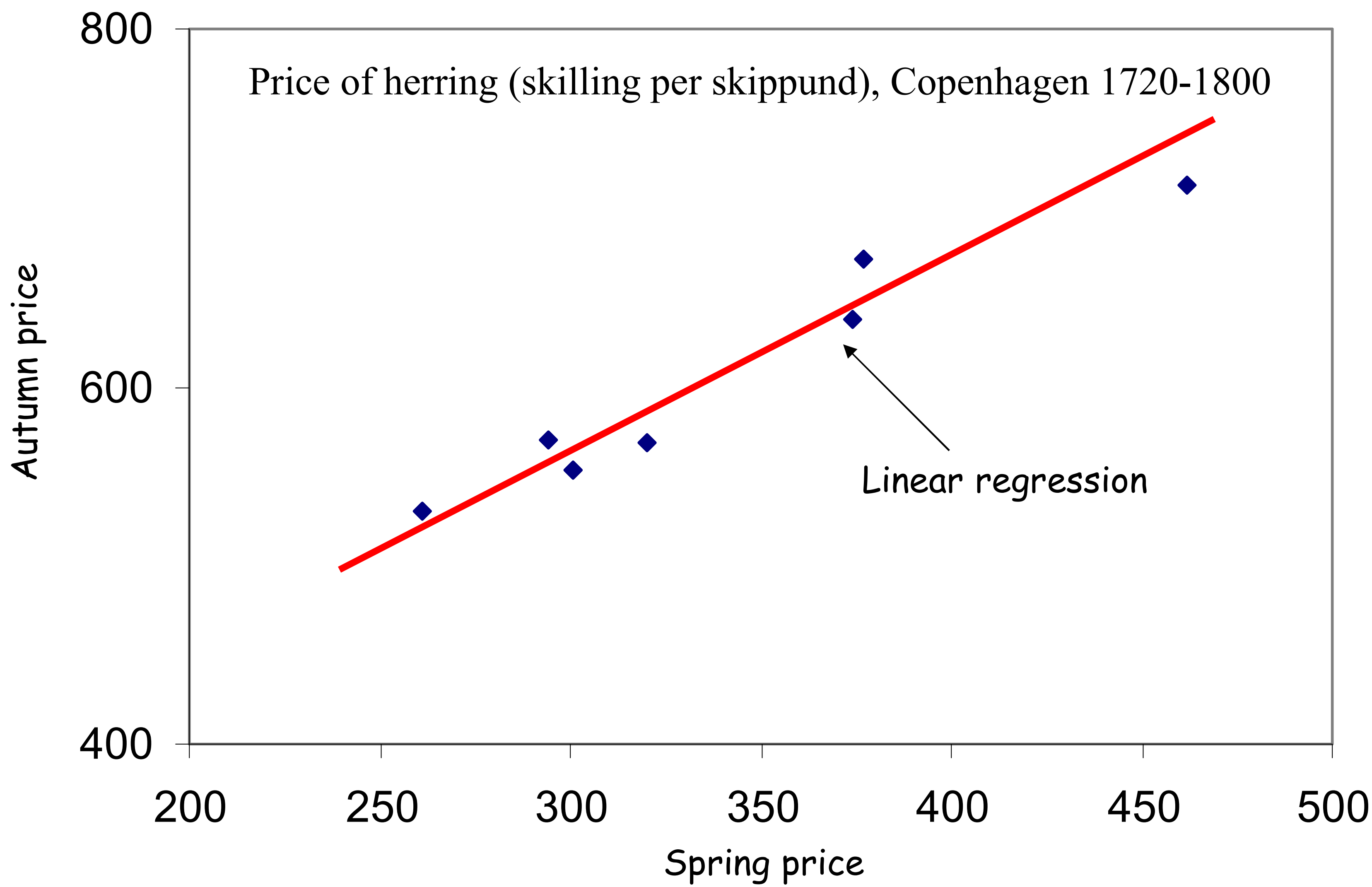
2nd principal component = *linear combination that accounts for as much of the remaining variation as possible, and is orthogonal to PC1*

⋮

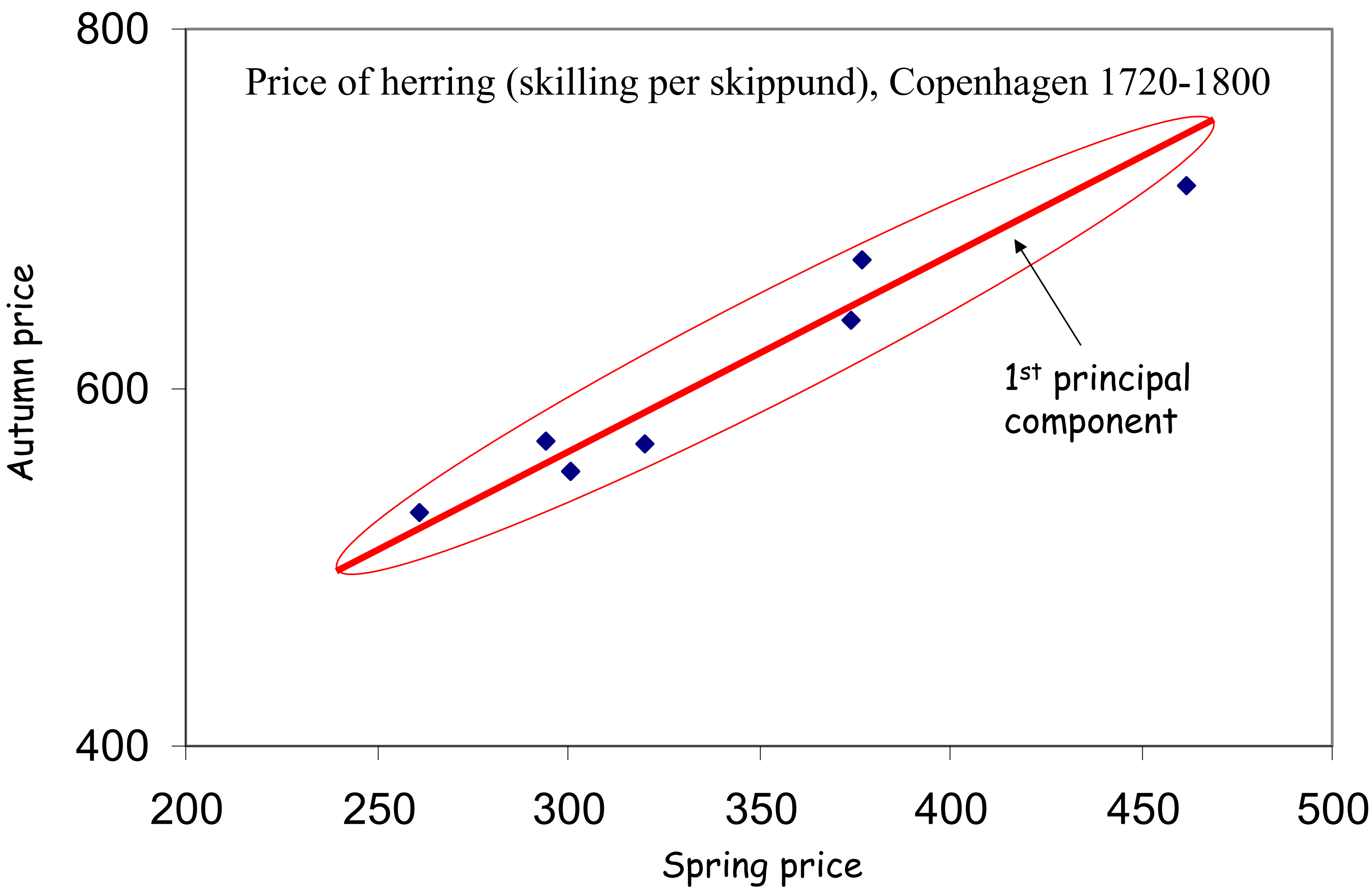
The price of fish…



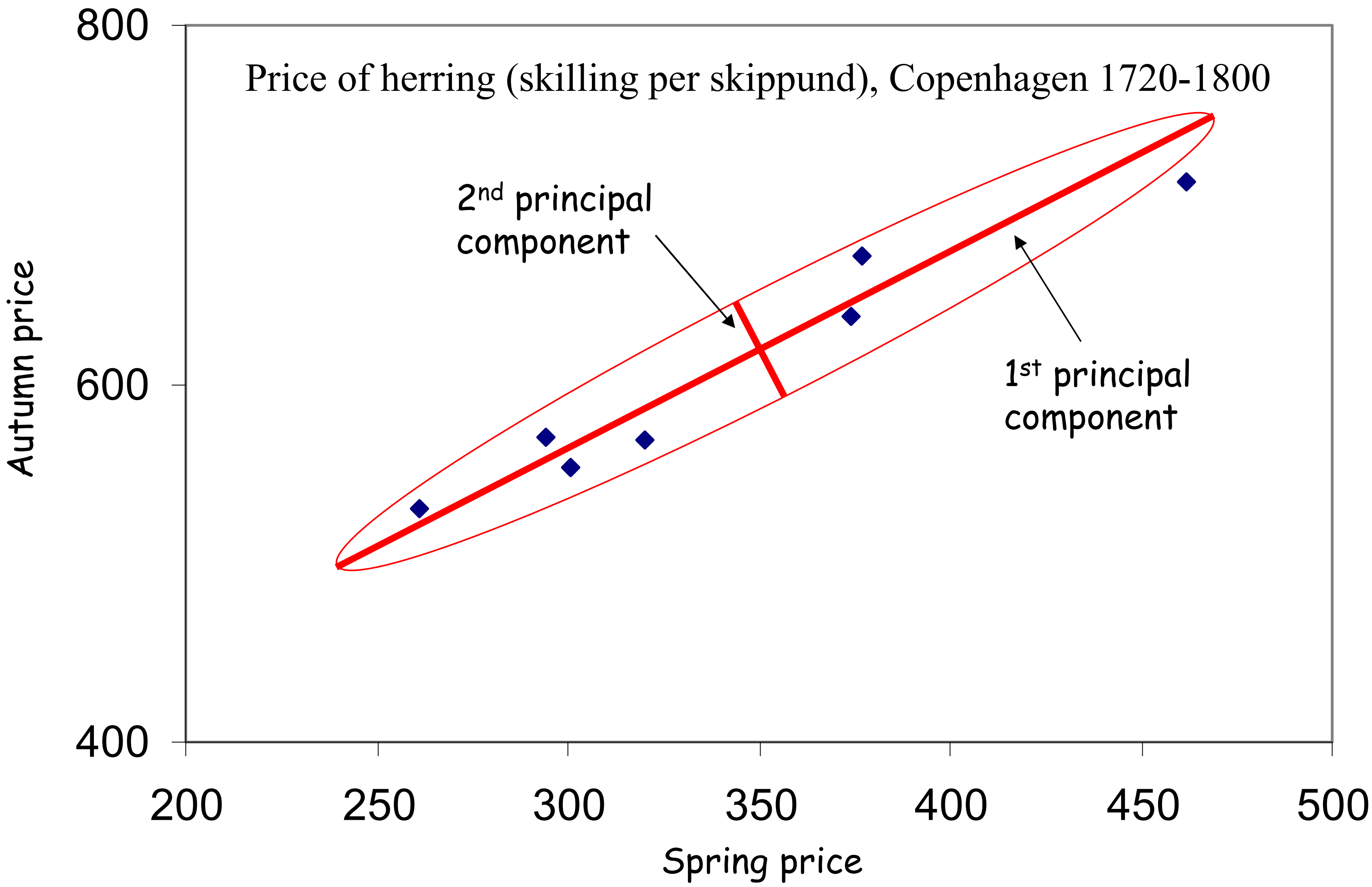
The price of fish…



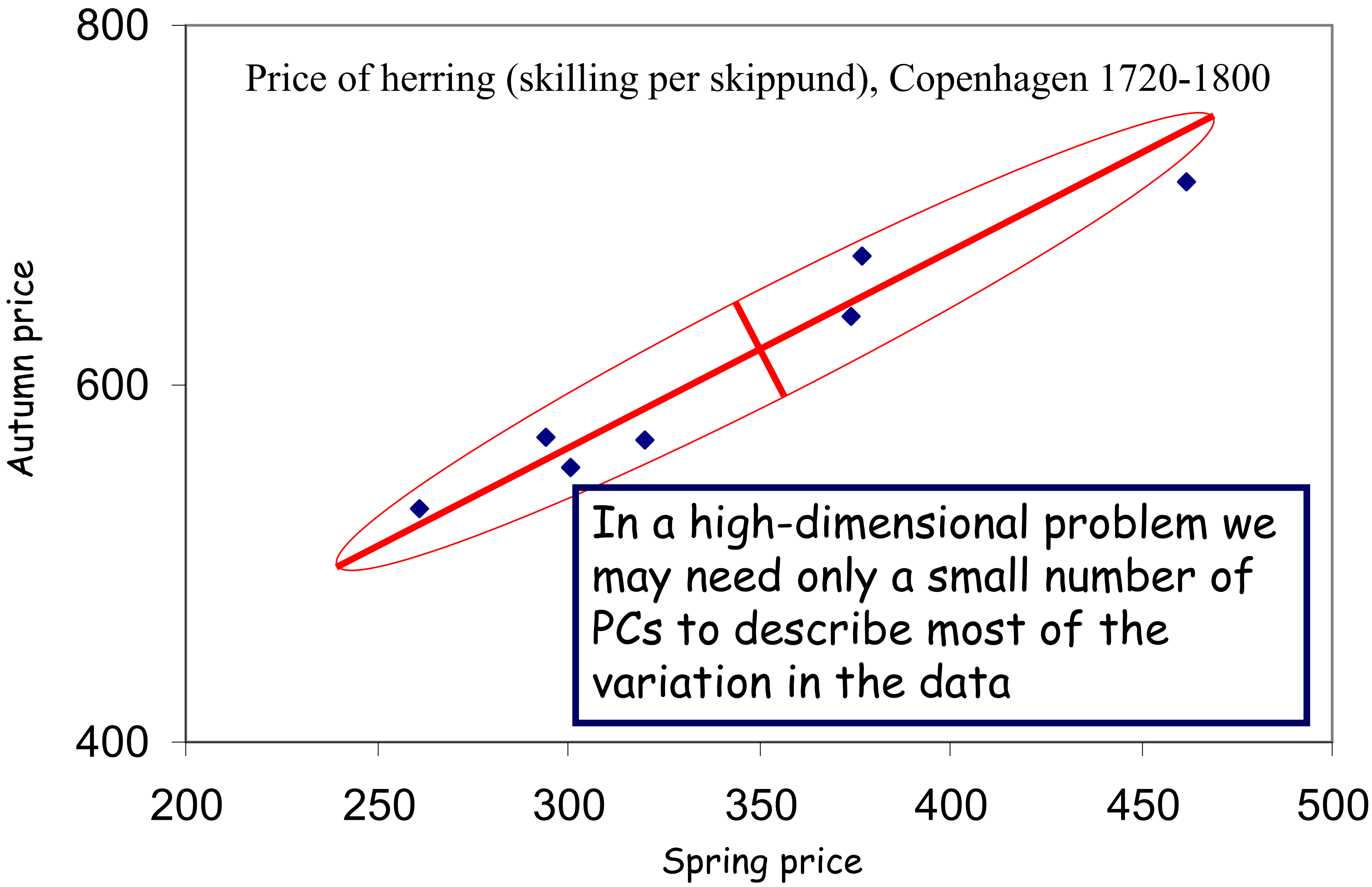
The price of fish…



The price of fish…



The price of fish…



In a high-dimensional problem we may need only a small number of PCs to describe most of the variation in the data

Relationship between PCA and SVD

Yesterday we discussed briefly how we could obtain stable solutions to the general linear model $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{a} + \boldsymbol{e}$ via the SVD of matrix $\boldsymbol{A}$:

M parameters

N observations

Diagonal matrix of singular values

$$\begin{pmatrix} \\ \\ \mathbf{A} \\ \\ \\ \end{pmatrix} = \begin{pmatrix} \\ \\ \mathbf{U} \\ \\ \\ \end{pmatrix} \cdot \begin{pmatrix} w_1 & & & \\ & w_2 & & \\ & & \cdots & \\ & & \cdots & \\ & & & w_M \end{pmatrix} \cdot \begin{pmatrix} \mathbf{V}^T \end{pmatrix}$$

From which the LS solution is:

$$\boxed{\hat{\boldsymbol{a}}_{LS} = \sum_{i=1}^{M} \left(\frac{\mathbf{U}_{(i)} \cdot \mathbf{b}}{w_i} \right) \mathbf{V}_{(i)}}$$

Relationship between PCA and SVD

So in PCA, we find the eigenvalues and eigenvectors of the data covariance matrix – in our notation $\boldsymbol{A}\,\boldsymbol{A}^{\boldsymbol{T}}$

From linear algebra we can decompose this as $\boxed{\boldsymbol{A}\,\boldsymbol{A}^{\boldsymbol{T}} \;=\; QDQ^T}$

Here D is a diagonal matrix containing the eigenvalues of $\boldsymbol{A}\,\boldsymbol{A}^{\boldsymbol{T}}$ and Q is an orthogonal matrix the columns of which are the eigenvectors.

In SVD we construct the matrix $\boldsymbol{A} \;=\; UWV^T$

It then follows that $\boldsymbol{A}\,\boldsymbol{A}^{\boldsymbol{T}} = \left(UWV^T\right)\left(UWV^T\right)^T \;=\; UW^2U^T$

So we see that the eigenvalues of the covariance matrix are equal to the **squares** of the singular values of the original data matrix $\boldsymbol{A}$.

Relationship between PCA and SVD

Recall from yesterday

$$\hat{a}_{LS} = \sum_{i=1}^{M} \left(\frac{\mathbf{U}_{(i)} \cdot \mathbf{b}}{w_i} \right) \mathbf{V}_{(i)}$$

Very small values of w_i will amplify any round-off errors in $\mathbf{b}$

Solution: for these very small singular values, set $\dfrac{1}{w_i} = 0$.

This suppresses their noisy contribution to the least-squares solution for the parameters $\hat{a}_{LS}$.

So it is often useful to perform an SVD and "switch off" the smallest singular values *before* applying PCA .

[Image processing example]

Beyond PCA: ICA

- PCA works by diagonalising the covariance matrix of a dataset, producing new linear combinations of the data which are uncorrelated.

- If the data are Gaussian then PCA will produce combinations which are statistically independent (all higher moments are zero for a Gaussian).

- For non-Gaussian data, uncorrelated does not in general imply independent. Alternative method called independent component analysis: linear transformation of data vector to minimise statistical dependence of components.

Interesting applications to time series data

COCKTAIL PARTY PROBLEM

Imagine you're at a cocktail party. For you it is no problem to follow the discussion of your neighbours, even if there are lots of other sound sources in the room: other discussions in English and in other languages, different kinds of music, etc.. You might even hear a siren from the passing-by police car.

It is not known exactly how humans are able to separate the different sound sources. Independent component analysis is able to do it, if there are at least as many microphones or 'ears' in the room as there are different simultaneous sound sources. In this demo, you can select which sounds are present in your cocktail party. ICA will separate them without knowing anything about the different sound sources or the positions of the microphones.

ORIGINAL SOUND SOURCES

By clicking the icons you can listen to the original sound sources.

SAMPLES AT THE COCKTAIL PARTY

Listen to the mixtures by clicking the microphones.

FOUND SOUND SOURCES

Below are the the sound sources separated by ICA. Note that they might be in different order than the original ones.

http://research.ics.aalto.fi/ica/cocktail/cocktail_en.cgi

https://www.cs.helsinki.fi/u/ahyvarin/papers/NN00new.pdf

Parameter estimation: Gaussian approximation

Taylor expand $\ell(\theta_1, \theta_2)$ around θ_{0j} :

$$\ell(\theta_1,\theta_2) = \ell(\theta_{01},\theta_{02}) + \cancel{\left.\frac{\partial \ell}{\partial \theta_1}\right|_{\theta_j=\theta_{0j}} (\theta_1 - \theta_{01})} + \cancel{\left.\frac{\partial \ell}{\partial \theta_2}\right|_{\theta_j=\theta_{0j}} (\theta_2 - \theta_{02})} +$$

$$\frac{1}{2}\left[\left.\frac{\partial^2 \ell}{\partial \theta_1^2}\right|_{\theta_j=\theta_{0j}} (\theta_1 - \theta_{01})^2 + \left.\frac{\partial^2 \ell}{\partial \theta_2^2}\right|_{\theta_j=\theta_{0j}} (\theta_2 - \theta_{02})^2 + 2\left.\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2}\right|_{\theta_j=\theta_{0j}} (\theta_1 - \theta_{01})(\theta_2 - \theta_{02}) \right] + \ldots$$

$$p(\theta_1, \theta_2 \mid \text{data}, \; I) = \exp\left[\ell(\theta_1, \theta_2)\right]$$

$$= \exp\left[-\tfrac{1}{2} Q\right]$$ ← Gaussian approximation

Is the Gaussian approximation a good idea?

Parameter estimation: Gaussian approximation

Taylor expand $\ell(\theta_1, \theta_2)$ around θ_{0j} :

$$\ell(\theta_1,\theta_2) = \ell(\theta_{01},\theta_{02}) + \cancel{\left.\frac{\partial \ell}{\partial \theta_1}\right|_{\theta_j=\theta_{0j}} (\theta_1-\theta_{01})} + \cancel{\left.\frac{\partial \ell}{\partial \theta_2}\right|_{\theta_j=\theta_{0j}} (\theta_2-\theta_{02})} +$$

$$\frac{1}{2}\left[\left.\frac{\partial^2 \ell}{\partial \theta_1^{\,2}}\right|_{\theta_j=\theta_{0j}} (\theta_1-\theta_{01})^2 + \left.\frac{\partial^2 \ell}{\partial \theta_2^{\,2}}\right|_{\theta_j=\theta_{0j}} (\theta_2-\theta_{02})^2 + 2\left.\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2}\right|_{\theta_j=\theta_{0j}} (\theta_1-\theta_{01})(\theta_2-\theta_{02}) \right] + \ldots$$

Is the Gaussian approximation a good idea?

- Greatly simplifies calculations – only need to compute the elements of the Fisher matrix (covariance matrix)

Parameter estimation: Gaussian approximation

Taylor expand $\ell(\theta_1, \theta_2)$ around θ_{0j} :

$$\ell(\theta_1,\theta_2) = \ell(\theta_{01},\theta_{02}) + \cancel{\left.\frac{\partial \ell}{\partial \theta_1}\right|_{\theta_j=\theta_{0j}} (\theta_1-\theta_{01})} + \cancel{\left.\frac{\partial \ell}{\partial \theta_2}\right|_{\theta_j=\theta_{0j}} (\theta_2-\theta_{02})} +$$
$$\frac{1}{2}\left[\left.\frac{\partial^2 \ell}{\partial \theta_1^{\,2}}\right|_{\theta_j=\theta_{0j}} (\theta_1-\theta_{01})^2 + \left.\frac{\partial^2 \ell}{\partial \theta_2^{\,2}}\right|_{\theta_j=\theta_{0j}} (\theta_2-\theta_{02})^2 + 2\left.\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2}\right|_{\theta_j=\theta_{0j}} (\theta_1-\theta_{01})(\theta_2-\theta_{02}) \right] + \ldots$$

Is the Gaussian approximation a good idea?

- Greatly simplifies calculations – only need to compute the elements of the Fisher matrix (covariance matrix)

- Nowadays, however, we can compute **full posterior** pdf. Not too hard with present-day computers, even for large N

Parameter estimation: Gaussian approximation

Taylor expand $\ell(\theta_1, \theta_2)$ around θ_{0j} :

$$\boxed{\begin{aligned} \ell(\theta_1,\theta_2) \;=\; & \ell(\theta_{01},\theta_{02}) + \cancel{\left.\frac{\partial \ell}{\partial \theta_1}\right|_{\theta_j=\theta_{0j}}(\theta_1-\theta_{01})} + \cancel{\left.\frac{\partial \ell}{\partial \theta_2}\right|_{\theta_j=\theta_{0j}}(\theta_2-\theta_{02})} + \\ & \frac{1}{2}\left[\left.\frac{\partial^2 \ell}{\partial \theta_1^{\,2}}\right|_{\theta_j=\theta_{0j}}(\theta_1-\theta_{01})^2 + \left.\frac{\partial^2 \ell}{\partial \theta_2^{\,2}}\right|_{\theta_j=\theta_{0j}}(\theta_2-\theta_{02})^2 + 2\left.\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2}\right|_{\theta_j=\theta_{0j}}(\theta_1-\theta_{01})(\theta_2-\theta_{02})\right] + \ldots \end{aligned}}$$

Is the Gaussian approximation a good idea?

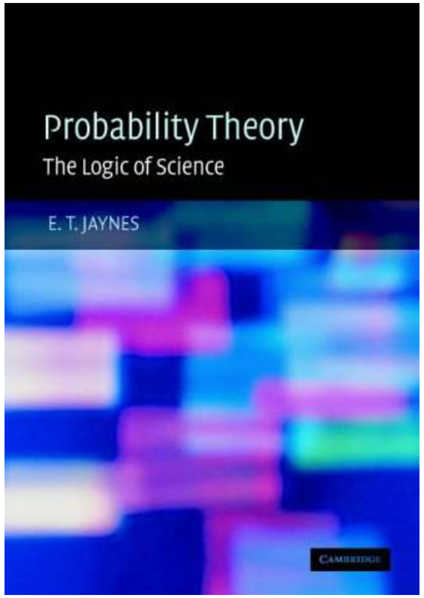
- Greatly simplifies calculations – only need to compute the elements of the Fisher matrix (covariance matrix)
- Nowadays, however, we can compute **full posterior** pdf. Not too hard with present-day computers, even for large N

$$\boxed{\textit{Markov Chain Monte Carlo Methods – see later}}$$

Defining Probabilities

Bayesian versus Frequentist statistics: Who is right?

Frequentists are correct to worry about subjectiveness of assigning probabilities – Bayesians worry about this too!!!



Ed Jaynes
(1922 – 1998)

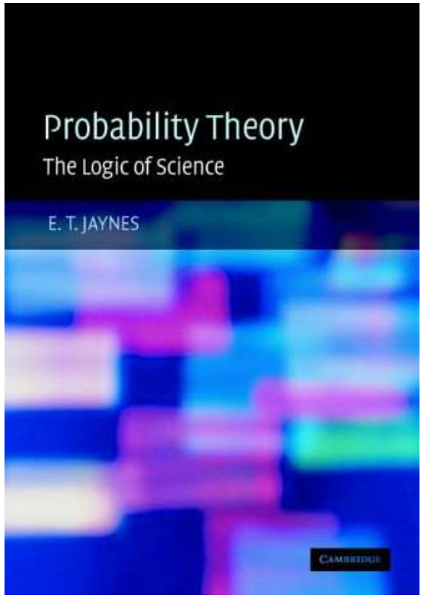
Probability *is* subjective; it depends on the available information

Subjective $\neq$ arbitrary

Given the *same* background information, two observers should assign the *same* probabilities

Bayesian versus Frequentist statistics: Who is right?

Frequentists are correct to worry about subjectiveness of assigning probabilities – Bayesians worry about this too!!!



Ed Jaynes
(1922 – 1998)

Probability *is* subjective; it depends on the available information

Subjective ≠ arbitrary

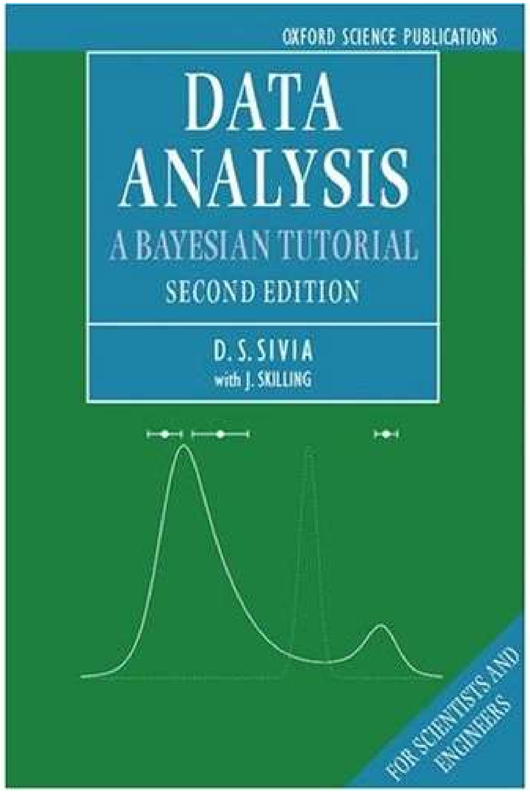
Given the *same* background information, two observers should assign the *same* probabilities

But what should they be?...

Bernoulli (1713) 'Principle of insufficient reason'

Keynes (1921) 'Principle of indifference'

If we can enumerate a set of basic mutually exclusive possibilities, and we have no reason to believe that any one of these is more likely to be true than another, then we should assign the same probability to all.



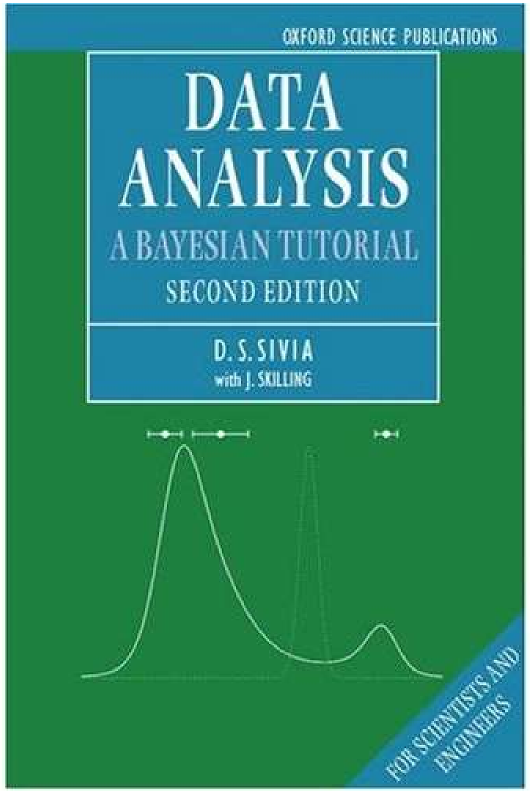
Bernoulli (1713) 'Principle of insufficient reason'

Keynes (1921) 'Principle of indifference'

If we can enumerate a set of basic mutually exclusive possibilities, and we have no reason to believe that any one of these is more likely to be true than another, then we should assign the same probability to all.

$$X_i \equiv \text{face on top has } i \text{ dots}$$

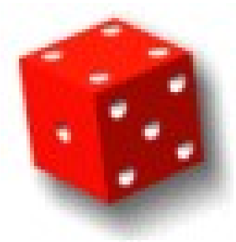
$$p(X_i \mid I) = \tfrac{1}{6} \quad \text{for all } i$$



Bernoulli (1713) 'Principle of insufficient reason'

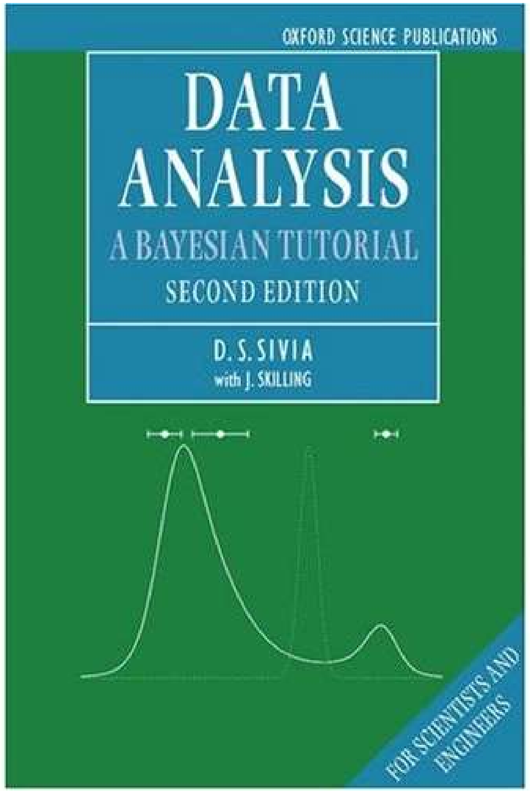
Keynes (1921) 'Principle of indifference'

If we can enumerate a set of basic mutually exclusive possibilities, and we have no reason to believe that any one of these is more likely to be true than another, then we should assign the same probability to all.



$$X_i \equiv \text{face on top has } i \text{ dots}$$

$$p(X_i \mid I) = \tfrac{1}{6} \quad \text{for all } i$$



Agrees with common sense, but can we justify more fundamentally?

Bernoulli (1713) 'Principle of insufficient reason'

Keynes (1921) 'Principle of indifference'

If we can enumerate a set of basic mutually exclusive possibilities, and we have no reason to believe that any one of these is more likely to be true than another, then we should assign the same probability to all.

X_i are just labels, e.g. suppose we define

$$X_i \equiv \text{face on top has } 7-i \text{ dots}$$

Should still have $p(X_i \mid I) = \frac{1}{6} \quad \text{for all } i$

Bernoulli (1713) 'Principle of insufficient reason'

Keynes (1921) 'Principle of indifference'

If we can enumerate a set of basic mutually exclusive possibilities, and we have no reason to believe that any one of these is more likely to be true than another, then we should assign the same probability to all.

X_i are just labels

Should still have $p(X_i \mid I) = \frac{1}{6}$ for all i

Extending to continuum case,

Let x be a location parameter.

Principle of indifference means we should have

$$\boxed{p(x \mid I)dx \quad = \quad p(x+\Delta \mid I)d(x+\Delta)}$$

where Δ is a constant

Since $dx \quad = \quad d(x+\Delta)$ we must have

$$\boxed{p(x \mid I) = \text{constant}}$$

Similarly,

Let L be a scale parameter.

Principle of indifference means we should have

$$\boxed{p(L \mid I)dL \quad = \quad p(\beta L \mid I)d(\beta L)}$$

where β is a positive constant

Since $d(\beta L) = \beta dL$ we must have $\boxed{p(L \mid I) \propto 1/\mathrm{L}}$

Jeffreys' prior

A Jeffreys' prior represents complete ignorance about the value of a scale parameter.

It is equivalent to a uniform pdf for the logarithm of L

i.e.

$$\boxed{p(\log L \mid I)dL \quad = \quad \text{constant}}$$

A Jeffreys' prior represents complete ignorance about the value of a scale parameter.

It is equivalent to a uniform pdf for the logarithm of L

i.e. $$\boxed{p(\log L \mid I)\,dL \;=\; \text{constant}}$$

In fact what is referred to as a Jeffreys prior $p(L \mid I) \propto 1/\mathrm{L}$ is just the special case of a more general result.

Suppose our inference problem is described by a likelihood with parameter(s) $\vec{\theta}$.

The **Jeffreys prior** is a non-informative (objective) prior defined as:

$$p(\vec{\theta}) \propto \left[\det I(\vec{\theta})\right]^{1/2}$$

Here $I(\vec{\theta})$ is the **Fisher Information** defined as

$$I(\vec{\theta})_{i,j} = E\left[\frac{\partial}{\partial\theta_i}\ln L(\vec{\theta})\frac{\partial}{\partial\theta_j}\ln L(\vec{\theta})\right]$$

[Note this expression reduces to that for the Fisher matrix given in Section 6 for the special case of a Gaussian likelihood.]

Key feature: the Jeffreys prior is **invariant** under ***any*** re-parameterisation of $\vec{\theta}$

Testable information

How do we deal with more complicated situations?

e.g. suppose we know that, when our die was rolled many times, the average result was 4.5 (and not 3.5)

How do we use this information to constrain $p(X_i \mid I)$?

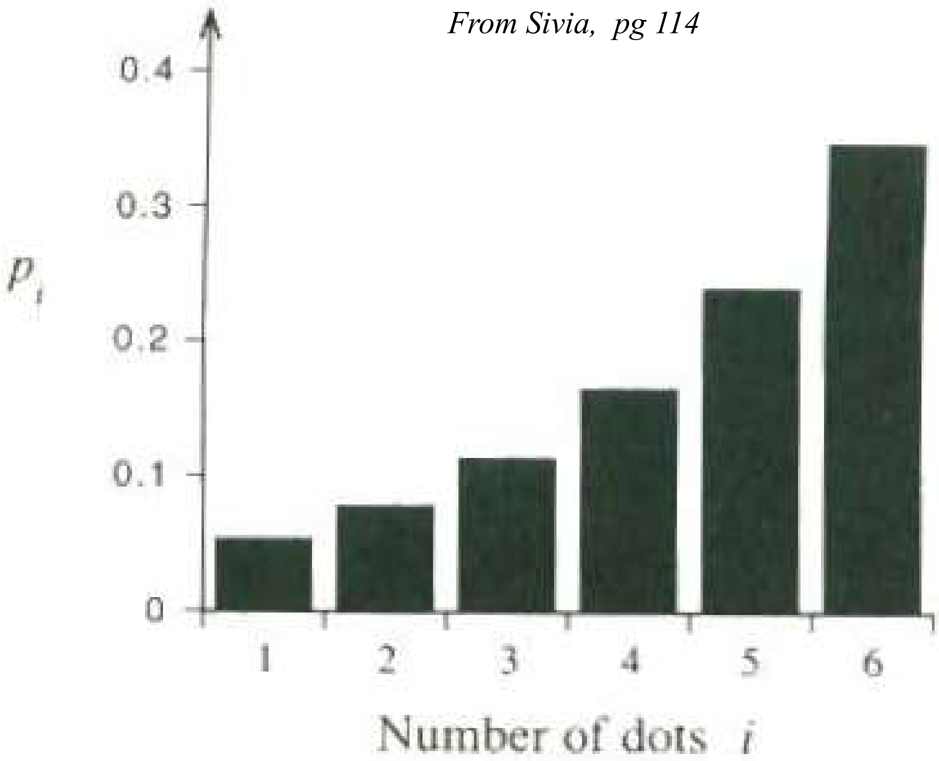
Jaynes (1957) suggests maximising the **Entropy**

$$S = -\sum_{i=1}^{6} p_i \log[p_i] \quad \text{subject to} \quad \sum_{i=1}^{6} p_i = 1 \quad \text{and} \quad \sum_{i=1}^{6} i\, p_i = 4.5$$

We can solve for the p_i using Lagrange Multipliers.

But why MAXENT?

From Sivia, pg 114



We can justify the importance of MAXENT via two approaches:

1) Independence argument (the kangaroo problem)

2) Shannon's Theorem and multiplicity

Consider the Kangaroo problem!

Information:	*1/3 of all kangaroos have blue eyes;* *1/3 of all kangaroos are left-handed*
Question:	*On the basis of the above information alone, what proportion of kangaroos are both blue eyed and left-handed?*

Question 15: Assuming that eye-colour and handedness are independent for kangaroos (humans?), we expect the proportion of kangaroos that are both blue-eyed and left-handed to be:

A zero

B 100%

C 1/9

D 1/3

Consider the Kangaroo problem!

Information: *1/3 of all kangaroos have blue eyes;*
1/3 of all kangaroos are left-handed

Question: *On the basis of the above information alone, what proportion of kangaroos are both blue eyed and left-handed?*

Blue eyes	Left-Handed	
	True	False
True	p_1	p_2
False	p_3	p_4

Blue eyes	Left-Handed	
	True	False
True	$0 \le z \le \frac{1}{3}$	$\frac{1}{3} - z$
False	$\frac{1}{3} - z$	$\frac{1}{3} + z$

We know that: $p_1 + p_2 + p_3 + p_4 = 1$

$$p_1 + p_2 = 1/3 \qquad p_1 + p_3 = 1/3$$

What is z ?

Independence arguments favour $z = 1/9$

Blue eyes	Left-Handed	
	True	False
True	p_1	p_2
False	p_3	p_4

Blue eyes	Left-Handed	
	True	False
True	$0 \le z \le \frac{1}{3}$	$\frac{1}{3} - z$
False	$\frac{1}{3} - z$	$\frac{1}{3} + z$

We know that: $p_1 + p_2 + p_3 + p_4 = 1$

$$p_1 + p_2 = 1/3 \qquad p_1 + p_3 = 1/3$$

What is z ?

Independence arguments favour $z = 1/9$

	Variation Function	Optimal z	Implied Correlation
MAXENT ⟶	$-\sum p_i \ln p_i$	$1/9 = 0.1111$	uncorrelated
	$-\sum p_i^2$	$1/12 = 0.0833$	negative
	$\sum \ln p_i$	0.1303	positive
	$\sum p_i^{1/2}$	0.1218	positive

MAXENT and common pdfs

Suppose we only know the expected value, μ , of a continuous physical quantity, x

What should we assign as $p(x \mid I)$?

Using MAXENT it can be shown that

$$\boxed{p(x \mid \mu) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu} \right)}$$

Exponential distribution

MAXENT and common pdfs

Suppose we only know the expected value, μ , of a

discrete physical quantity, N

What should we assign as $p(x \mid I)$?

Using MAXENT it can be shown that

$$\boxed{p(N \mid \mu) = \frac{\mu^N e^{-\mu}}{N!}}$$

Poisson distribution

MAXENT and common pdfs

Suppose we only know the expected value, μ , and $\langle x-\mu \rangle^2 = \sigma^2$ of a continuous physical quantity, x

What should we assign as $p(x \mid I)$?

Using MAXENT it can be shown that

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

Normal distribution

University of Glasgow

SUPA

MAXENT and common pdfs

Suppose we only know the expected value, μ , and

$\langle x-\mu \rangle^2 = \sigma^2$ of a continuous physical quantity, x

What should we assign as $p(x \mid I)$?

Using MAXENT it can be shown that

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

Normal distribution

MAXENT justifies the relevance of common pdfs

University of Glasgow

SUPA