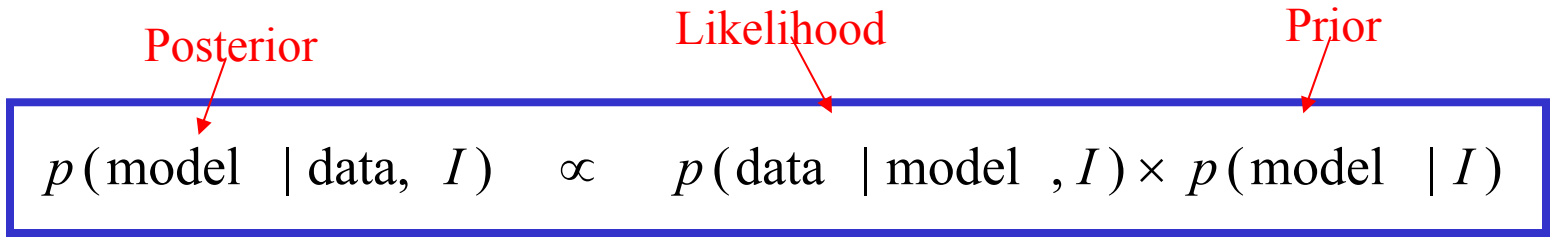
4. An Advanced Bayesian Toolbox

Parameter estimation:



Parameter estimation: the Gaussian approximation

$$p(\theta \mid \text{data}, I) \propto p(\text{data} \mid \theta, I) \times p(\theta \mid I)$$

'Best' estimator: $\left.\frac{\partial p(\theta \mid \text{data}, I)}{\partial \theta}\right|_{\theta=\theta_0} = 0$ ← Maximise posterior likelihood

Equivalently, we can define $\ell = \log p(\theta \mid \text{data}, I)$ and compute $\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta=\theta_0} = 0$

Parameter estimation: the Gaussian approximation

$$p(\theta \mid \text{data}, I) \propto p(\text{data} \mid \theta, I) \times p(\theta \mid I)$$

'Best' estimator: $\left.\frac{\partial p(\theta \mid \text{data}, I)}{\partial \theta}\right|_{\theta=\theta_0} = 0$ ← Maximise posterior likelihood

Equivalently, we can define $\ell = \log p(\theta \mid \text{data}, I)$ and compute $\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta=\theta_0} = 0$

Taylor expand $\ell(\theta)$ around $\theta=\theta_0$:

$$\ell(\theta) = \ell(\theta_0) + \left.\frac{\partial \ell}{\partial \theta}\right|_{\theta=\theta_0} (\theta-\theta_0) + \frac{1}{2}\left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta=\theta_0} (\theta-\theta_0)^2 + \ldots$$

Parameter estimation: the Gaussian approximation

$$p(\theta \mid \text{data}, I) \propto p(\text{data} \mid \theta, I) \times p(\theta \mid I)$$

'Best' estimator: $\left.\frac{\partial p(\theta \mid \text{data}, I)}{\partial \theta}\right|_{\theta=\theta_0} = 0$ ← Maximise posterior likelihood

Equivalently, we can define $\ell = \log p(\theta \mid \text{data}, I)$ and compute $\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta=\theta_0} = 0$

Taylor expand $\ell(\theta)$ around $\theta=\theta_0$:

$$\ell(\theta) = \ell(\theta_0) + \left.\frac{\partial \ell}{\partial \theta}\right|_{\theta=\theta_0} (\theta-\theta_0) + \frac{1}{2}\left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta=\theta_0} (\theta-\theta_0)^2 + \ldots$$

Parameter estimation: the Gaussian approximation

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

Neglecting higher order terms in $\ell(\theta)$

$$p(\theta \mid \text{data}, I) \propto \exp\left(-\frac{A}{2}(\theta - \theta_0)^2\right)$$

where $A = -\left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta=\theta_0}$

This is equivalent to a normal distribution, with $\sigma^{-2} = A = -\left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta=\theta_0}$

Parameter estimation: the Gaussian approximation

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

Neglecting higher order terms in $\ell(\theta)$ ← Gaussian approximation

$$p(\theta \mid \text{data}, I) \propto \exp\left(-\frac{A}{2}(\theta - \theta_0)^2\right)$$

where $A = -\left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta=\theta_0}$

This is equivalent to a normal distribution, with $\sigma^{-2} = A = -\left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta=\theta_0}$

Parameter estimation: the Gaussian approximation

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

Neglecting higher order terms in $\ell(\theta)$ ← Gaussian approximation

$$p(\theta \mid \text{data}, I) \propto \exp\left(-\frac{A}{2}(\theta - \theta_0)^2\right)$$

where $A = -\left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta=\theta_0}$

This is equivalent to a normal distribution, with $\sigma^{-2} = A = -\left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta=\theta_0}$

Can summarise inference from posterior by

$$\theta = \theta_0 \pm \sigma$$

Question 9: Neglecting the higher order terms in the log likelihood expansion produces a posterior which can be written as a normal pdf because

A The higher order moments of a Gaussian are all zero

B The Gaussian pdf is uniquely specified by its mean and variance

C The logarithm of a Gaussian pdf can be written in the form of a quadratic

D All of the above

Question 9: Neglecting the higher order terms in the log likelihood expansion produces a posterior which can be written as a normal pdf because

A The higher order moments of a Gaussian are all zero

B The Gaussian pdf is uniquely specified by its mean and variance

C The logarithm of a Gaussian pdf can be written in the form of a quadratic

D All of the above

Parameter estimation: 2-D case

Recall our definition of *variance*

$$\mathrm{var}[x] = \int_{-\infty}^{\infty} \left(x - \langle x \rangle\right)^2 p(x \mid I)\, dx$$

Extends to 2 variables – *covariance*

$$\mathrm{cov}[x, y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(x - \langle x \rangle\right)\left(y - \langle y \rangle\right) p(x, y \mid I)\, dx\, dy$$

Parameter estimation: 2-D case

Recall our definition of *variance*

$$\mathrm{var}[x] = \int_{-\infty}^{\infty} \left(x - \langle x \rangle\right)^2 p(x \mid I)\, dx$$

Extends to 2 variables - *covariance*

$$\mathrm{cov}[x, y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(x - \langle x \rangle\right)\left(y - \langle y \rangle\right) p(x, y \mid I)\, dx dy$$

If x and y are independent, $\mathrm{cov}[x, y] = 0$

This is because $p(x, y \mid I) = p(x \mid I)\, p(y \mid I)$

Parameter estimation: 2-D case

$$p(\theta_1, \theta_2 \mid \text{data}, I) \quad \propto \quad p(\text{data} \mid \theta_1, \theta_2, I) \times p(\theta_1, \theta_2 \mid I)$$

'Best' estimator: $\left.\dfrac{\partial p(\theta_1, \theta_2 \mid \text{data}, I)}{\partial \theta_j}\right|_{\theta_j = \theta_{0j}} = 0$

Compute $\left.\dfrac{\partial \ell}{\partial \theta_j}\right|_{\theta_j = \theta_{0j}} = 0$ where $\ell = \log p(\theta_1, \theta_2 \mid \text{data}, I)$

Parameter estimation: 2-D case

$$p(\theta_1, \theta_2 \mid \text{data}, I) \quad \propto \quad p(\text{data} \mid \theta_1, \theta_2, I) \times p(\theta_1, \theta_2 \mid I)$$

'Best' estimator: $\left.\dfrac{\partial p(\theta_1, \theta_2 \mid \text{data}, I)}{\partial \theta_j}\right|_{\theta_j = \theta_{0j}} = 0$

Compute $\left.\dfrac{\partial \ell}{\partial \theta_j}\right|_{\theta_j = \theta_{0j}} = 0$ where $\ell = \log p(\theta_1, \theta_2 \mid \text{data}, I)$

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

Parameter estimation: 2-D case

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

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

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

Parameter estimation: 2-D case

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

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

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

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

$$\propto \exp\left[-\tfrac{1}{2}Q\right] \leftarrow \text{Gaussian approximation}$$

$$\chi^2$$

Parameter estimation: 2-D case

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

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

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

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

$$\propto \exp\left[-\tfrac{1}{2} Q\right] \leftarrow \text{Gaussian approximation}$$

χ^2

Maximising likelihood
≡ Minimising χ^2

Parameter estimation: 2-D case

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

$$Q = \begin{pmatrix} \theta_1 - \theta_{10} & \theta_2 - \theta_{20} \end{pmatrix} \begin{bmatrix} A & C \\ C & B \end{bmatrix} \begin{pmatrix} \theta_1 - \theta_{10} \\ \theta_2 - \theta_{20} \end{pmatrix}$$

← Quadratic form

where $A = \left. \frac{\partial^2 \ell}{\partial \theta_1^{\,2}} \right|_{\theta_j = \theta_{0j}}$ $B = \left. \frac{\partial^2 \ell}{\partial \theta_2^{\,2}} \right|_{\theta_j = \theta_{0j}}$ $C = \left. \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} \right|_{\theta_j = \theta_{0j}}$

Parameter estimation: 2-D case

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

$$Q = \begin{pmatrix} \theta_1 - \theta_{10} & \theta_2 - \theta_{20} \end{pmatrix} \begin{bmatrix} A & C \\ C & B \end{bmatrix} \begin{pmatrix} \theta_1 - \theta_{10} \\ \theta_2 - \theta_{20} \end{pmatrix}$$

← Quadratic form

where $A = \left. \frac{\partial^2 \ell}{\partial \theta_1^{\,2}} \right|_{\theta_j = \theta_{0j}}$ $B = \left. \frac{\partial^2 \ell}{\partial \theta_2^{\,2}} \right|_{\theta_j = \theta_{0j}}$ $C = \left. \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} \right|_{\theta_j = \theta_{0j}}$

This is a bivariate normal distribution with covariance matrix

$$\sigma_{ij}^2 = \mathrm{cov}_{ij} = \left\langle (\theta_i - \theta_{i0})(\theta_j - \theta_{j0}) \right\rangle = \left[- \frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} \right]^{-1}$$

Fisher information matrix

$$\mathbf{F} \equiv F_{ij} = \frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}$$ is known as the **Fisher information matrix**

It provides a measure of how much information a given dataset can yield about the parameters of a model.

We can see this most easily in the case where the Fisher matrix is **diagonal**.

Then $$\mathbf{F} = -\mathrm{diag}\left(\sigma_1^{-2}, \ldots, \sigma_n^{-2}\right)$$

If the i^{th} element of the Fisher matrix is large (negative), the **variance** of parameter θ_i is small (and positive).

In general the Fisher matrix (and covariance matrix) will **not** be diagonal; the Fisher matrix then tells us which **combinations** of the parameters are well constrained by the data. (see later).

So if, for our model:

- the likelihood is Gaussian in shape (or if we can approximate it as Gaussian – i.e. if the higher order terms in the Taylor expansion of the log likelihood can be neglected);
- the parameters have broad, uniform priors;

then the posterior will also be Gaussian.

If we can evaluate the first and second partial derivatives of the log likelihood, we can:

- compute the **Fisher Information Matrix;**
- compute the **Covariance Matrix** of the posterior.

We can also compute **credible regions** for the parameters (in fact for this we don't need the derivatives – see Section 6)

We can write the log likelihood as

$$\ell(\theta_1,\theta_2) \;=\; \text{const} \;-\tfrac{1}{2}\chi^2(\theta_1,\theta_2)$$

Now $\chi^2 = \chi^2_{\min}$ when $(\theta_1,\theta_2)=(\theta_{01},\theta_{02})$

Maximising likelihood ≡ Minimising χ^2

so we can write, for $\Delta\chi^2(\theta_1,\theta_2)= \chi^2(\theta_1,\theta_2)- \chi^2_{\min}$

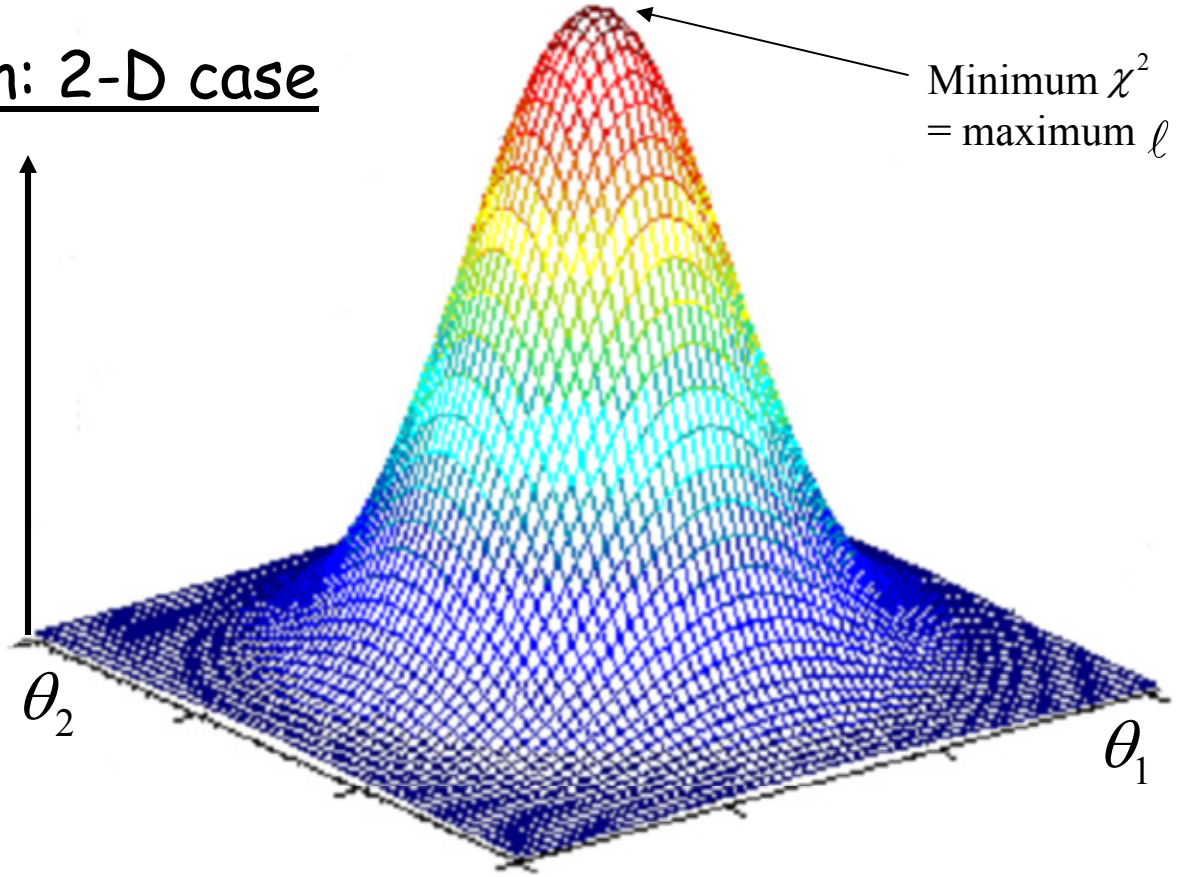
$$\ell(\theta_1,\theta_2) \;=\; \ell(\theta_{01},\theta_{02}) \;-\tfrac{1}{2}\Delta\chi^2(\theta_1,\theta_2)$$

So that

$$p(\theta_1,\theta_2 \mid \text{data},\ I) \;=\; p(\theta_{01},\theta_{02} \mid \text{data},\ I)\ \exp\left[-\tfrac{1}{2}\Delta\chi^2(\theta_1,\theta_2)\right]$$

Maximum of the posterior

Parameter estimation: 2-D case

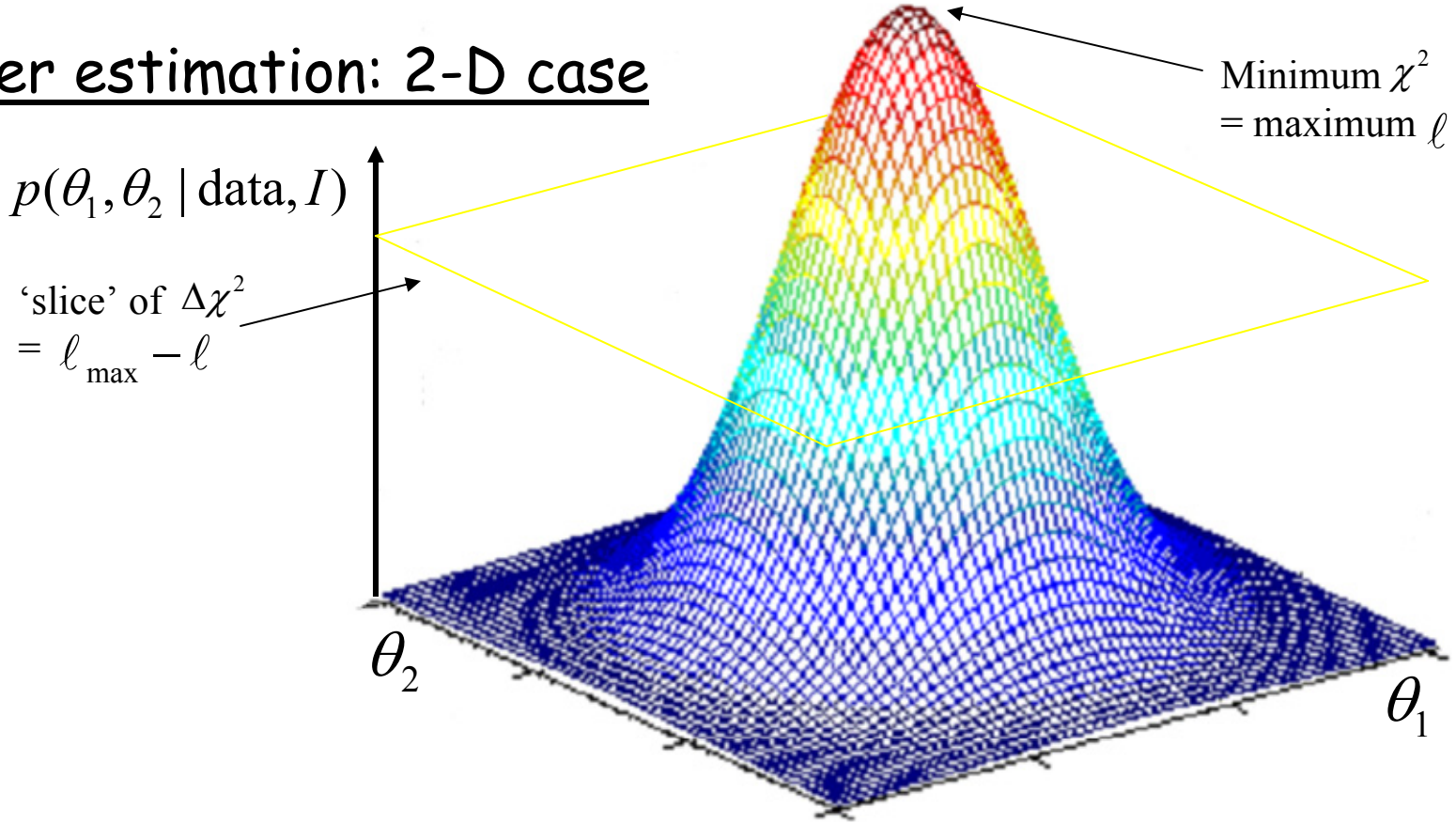


This is a bivariate normal distribution with covariance matrix

$$\sigma_{ij}^2 = \text{cov}_{ij} = \left\langle (\theta_i - \theta_{i0})(\theta_j - \theta_{j0}) \right\rangle = \left[-\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} \right]^{-1}$$

Fisher information matrix

Parameter estimation: 2-D case

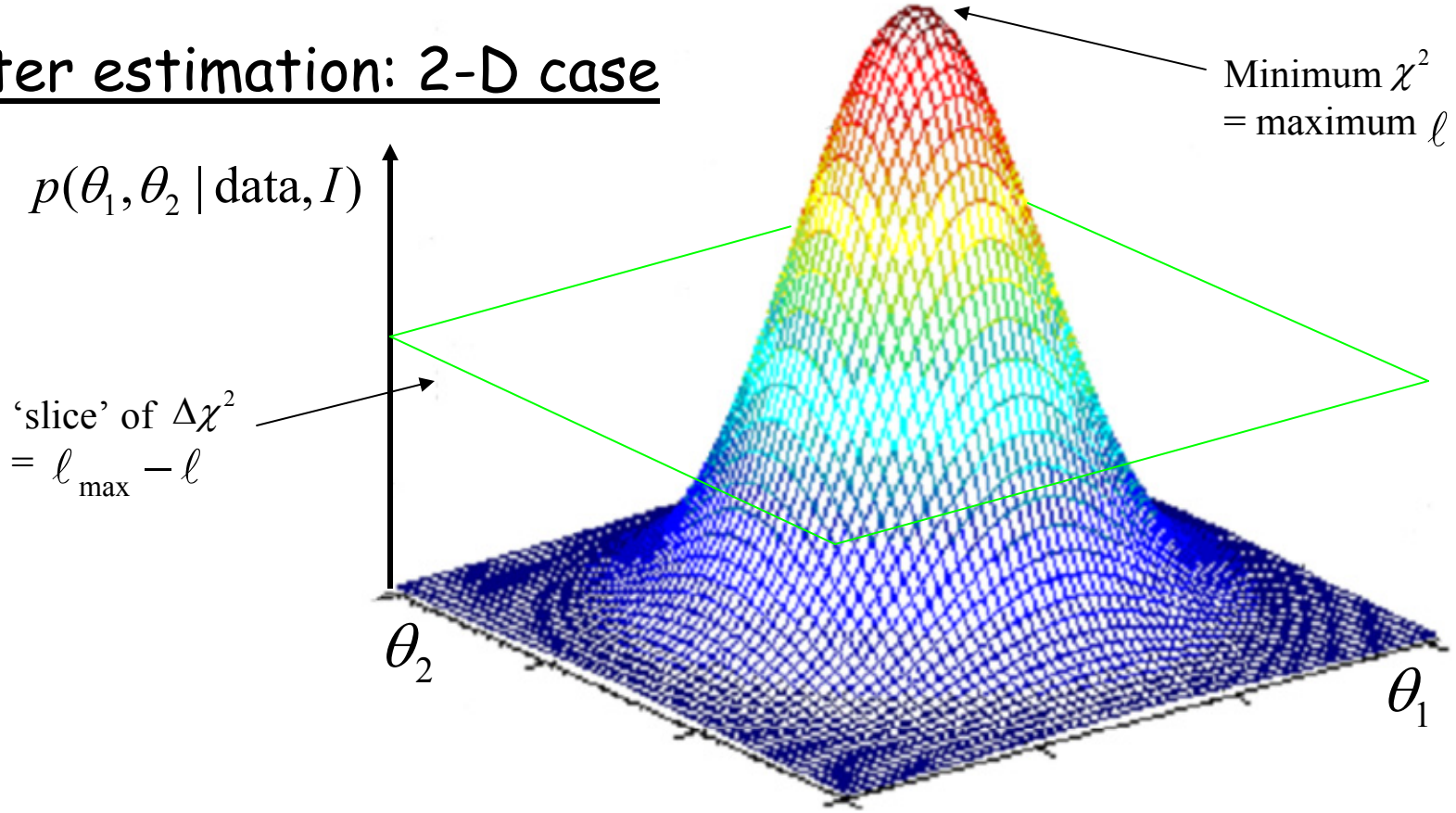


This is a bivariate normal distribution with covariance matrix

$$\sigma_{ij}^2 \quad = \quad \text{cov}_{ij} \quad = \quad \left\langle (\theta_i - \theta_{i0})(\theta_j - \theta_{j0}) \right\rangle \quad = \quad \left[-\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j} \right]^{-1}$$

Fisher information matrix

Parameter estimation: 2-D case

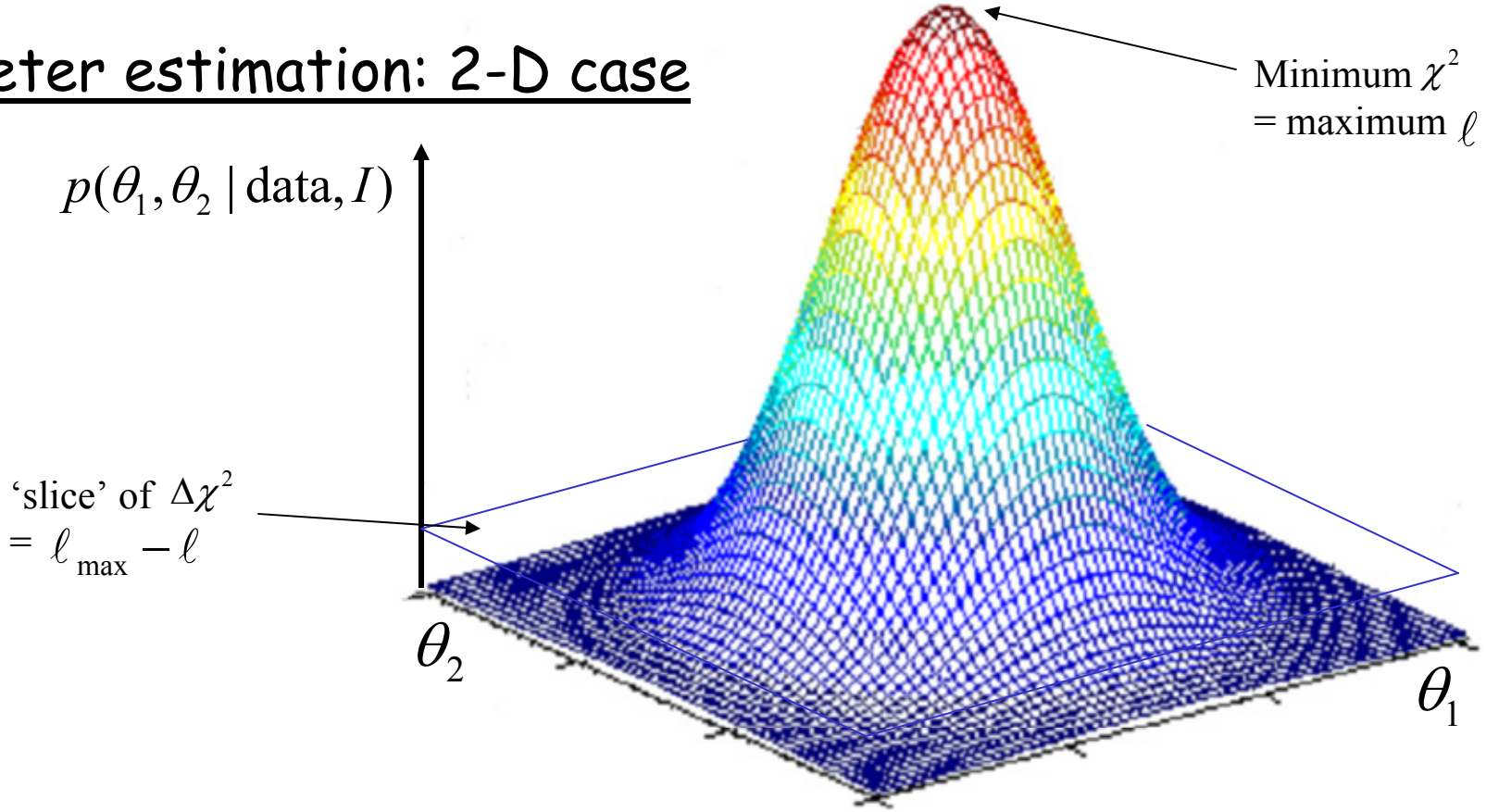


This is a bivariate normal distribution with covariance matrix

$$\sigma_{ij}^2 \quad = \quad \text{cov}_{ij} \quad = \quad \left\langle (\theta_i - \theta_{i0})(\theta_j - \theta_{j0}) \right\rangle \quad = \quad \left[-\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} \right]^{-1}$$

Fisher information matrix

Parameter estimation: 2-D case



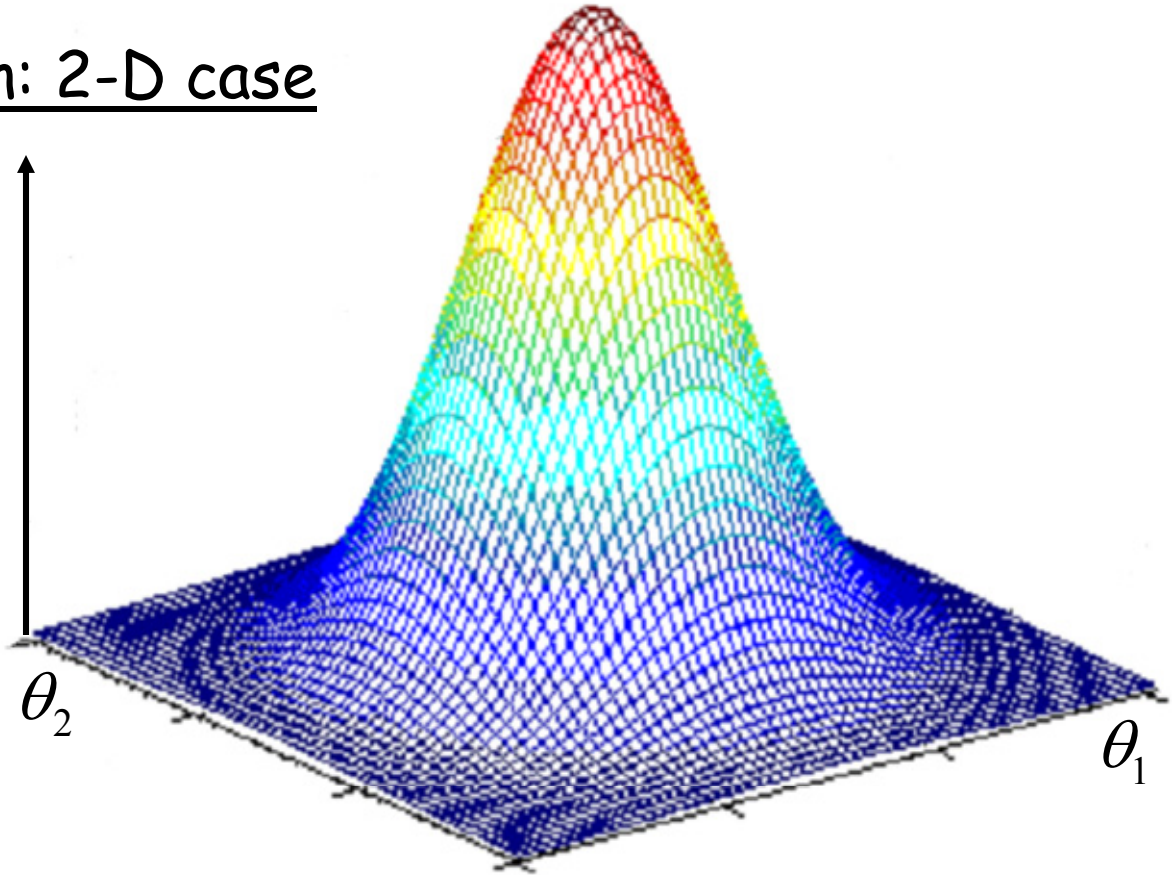
This is a bivariate normal distribution with covariance matrix

$$\sigma_{ij}^2 \quad = \quad \text{cov}_{ij} \quad = \quad \left\langle (\theta_i - \theta_{i0})(\theta_j - \theta_{j0}) \right\rangle \quad = \quad \left[-\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} \right]^{-1}$$

Fisher information matrix

Parameter estimation: 2-D case

$p(\theta_1, \theta_2 \,|\, \text{data}, I)$



We can compute the $\Delta\chi^2$ that corresponds to e.g. 68%, 95%, 99% of the posterior pdf.

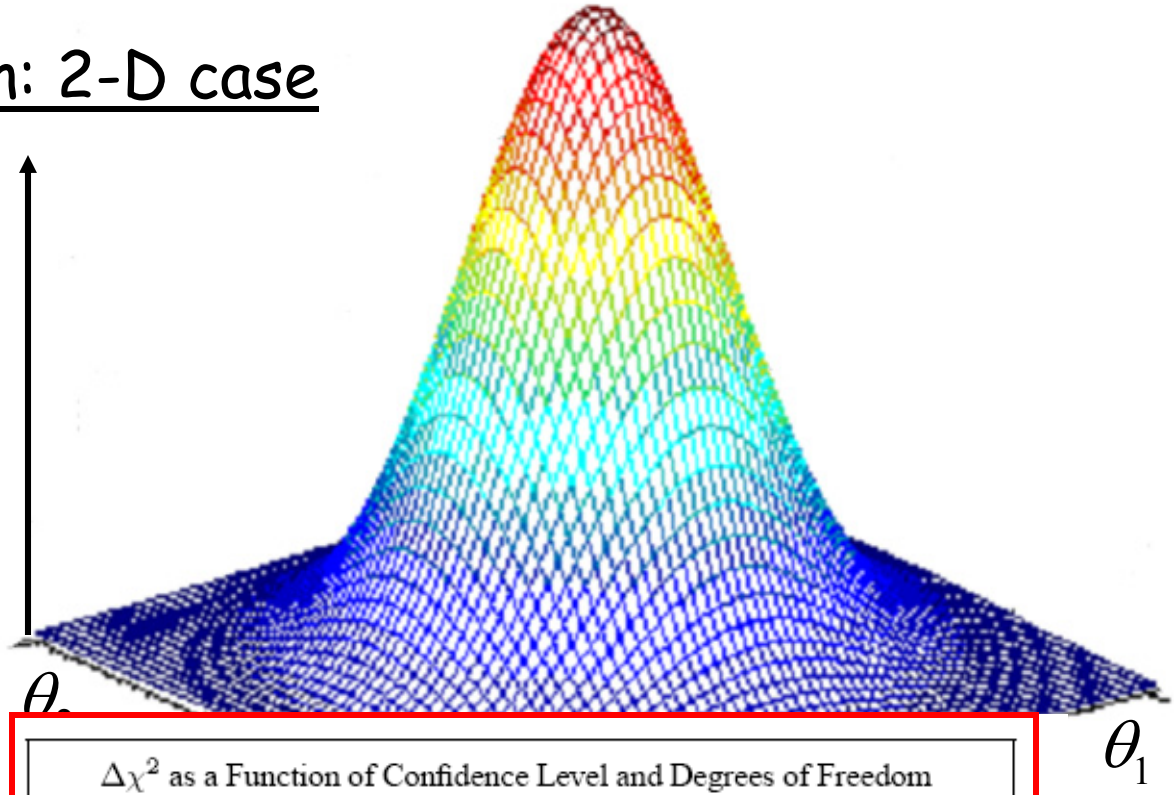
We can draw contours of equal probability

$\Rightarrow$ **Credible regions for the parameters**

Extends easily to N parameters – or *degrees of freedom*

Parameter estimation: 2-D case

$p(\theta_1, \theta_2 \,|\, \text{data}, I)$



We can compute the $\Delta\chi^2$ that corresponds to e.g. 68%, 95%, 99% of the posterior pdf.

We can draw contours of equal probability

⇒ **Credible regions for the parameters**

Extends easily to N parameters – or *degrees of freedom*

$\Delta\chi^2$ as a Function of Confidence Level and Degrees of Freedom						
	ν					
p	1	2	3	4	5	6
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8

From Numerical Recipes

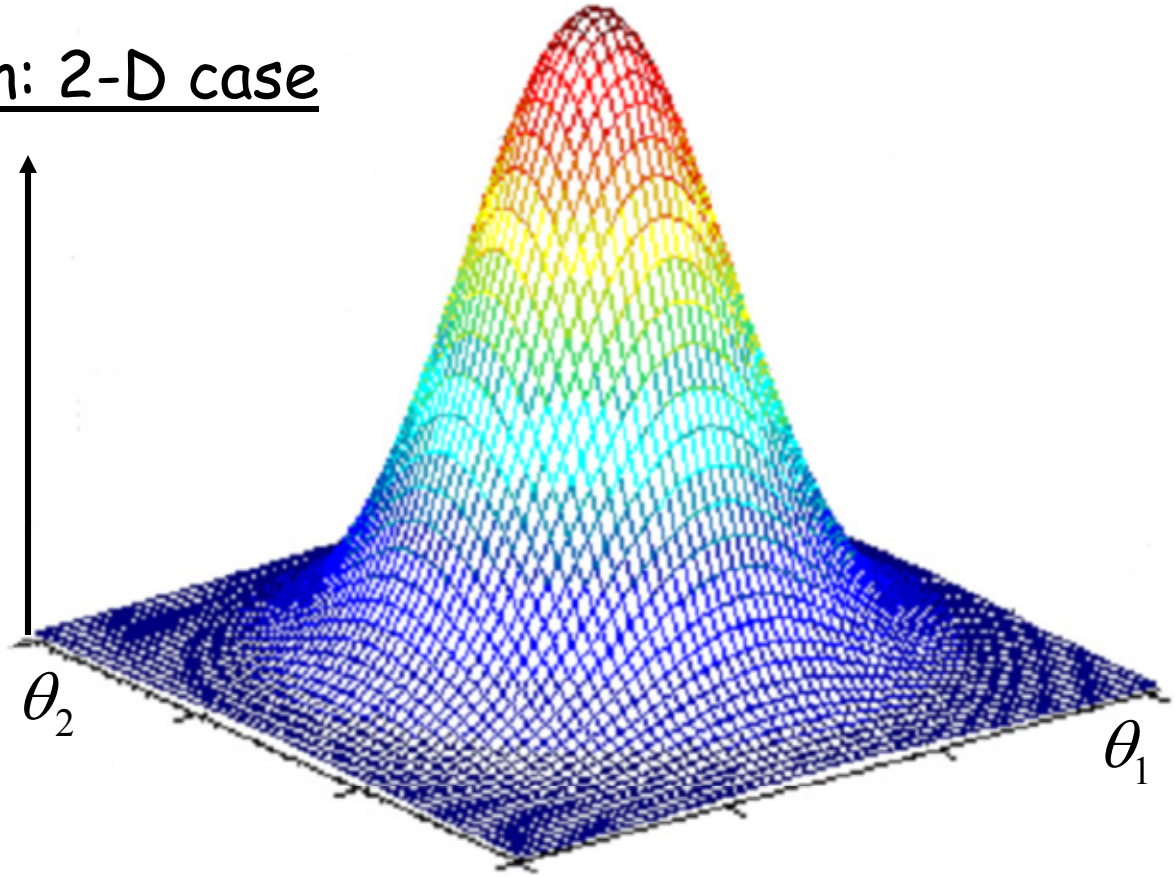
Parameter estimation: 2-D case

$p(\theta_1, \theta_2 \mid \text{data}, I)$

Contours of constant probability are ellipses.

Covariance matrix is *not* in general diagonal

$\Rightarrow$ What we infer about θ_1 and θ_1 is *not* independent



Parameter estimation: 2-D case

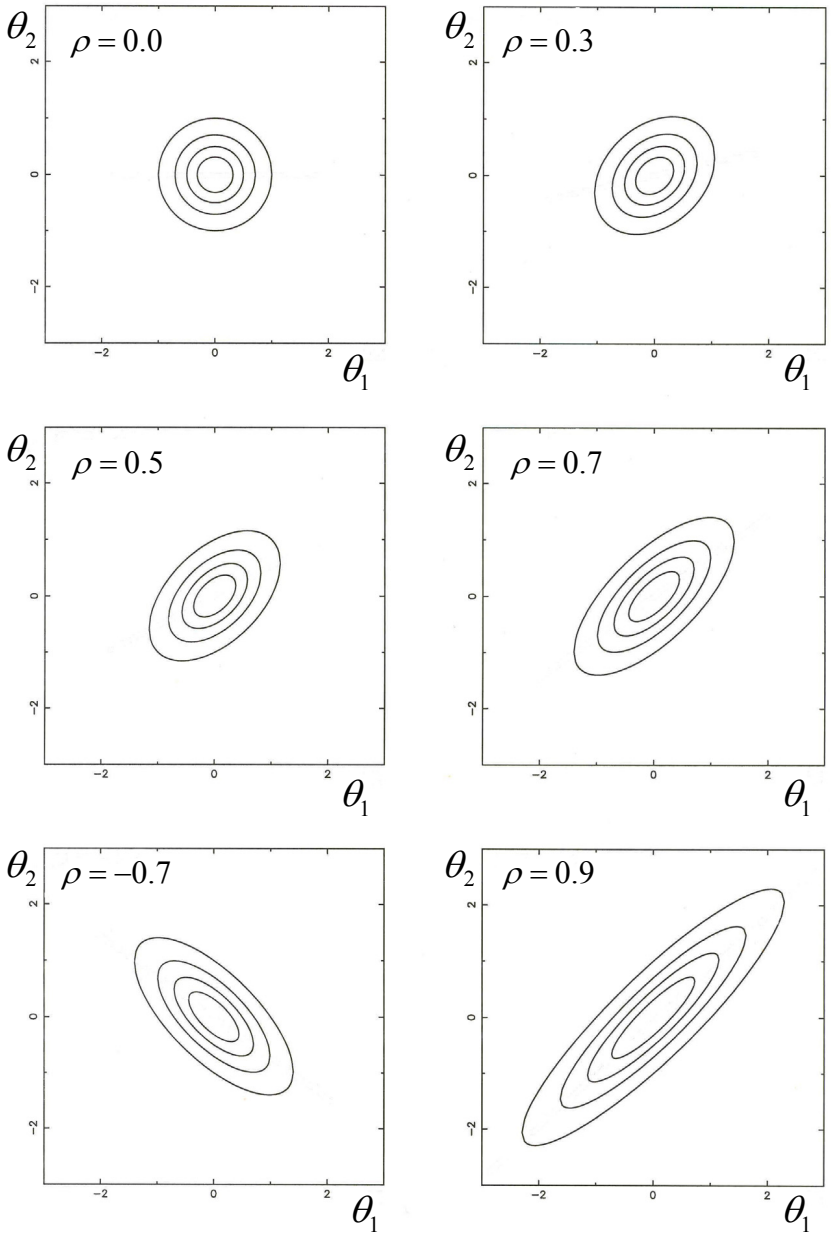
Can define *correlation coefficient*

$$\rho = \frac{\mathrm{cov}[\theta_1, \theta_2]}{\sqrt{\mathrm{var}[\theta_1]}\sqrt{\mathrm{var}[\theta_2]}} \qquad -1 \le \rho \le 1$$

Covariance matrix becomes less diagonal

$\Rightarrow$ $|\rho|$ increases

$\Rightarrow$ isoprobability contours elongate



Parameter estimation: 2-D case

Can define *correlation coefficient*

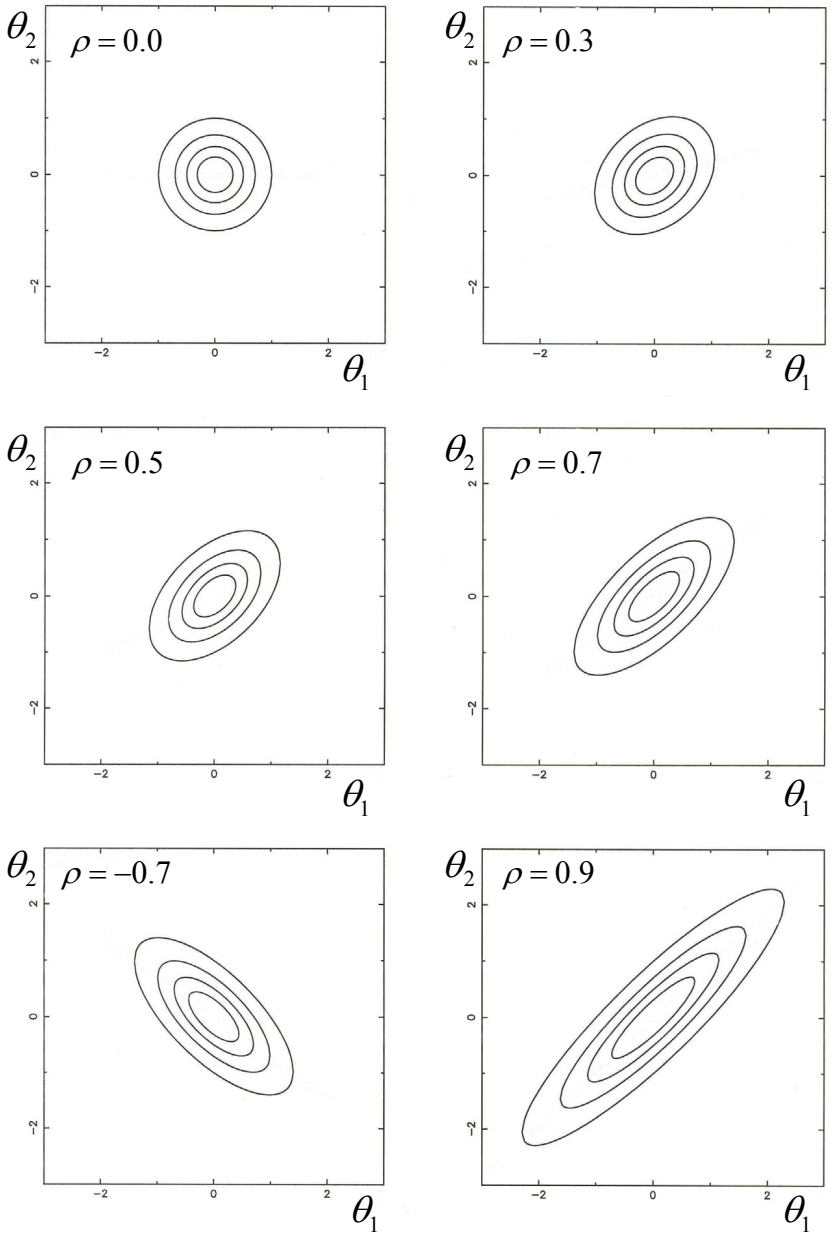
$$\rho = \frac{\mathrm{cov}[\theta_1, \theta_2]}{\sqrt{\mathrm{var}[\theta_1]}\sqrt{\mathrm{var}[\theta_2]}} \qquad -1 \le \rho \le 1$$

Covariance matrix becomes less diagonal

$\Rightarrow$ $|\rho|$ increases

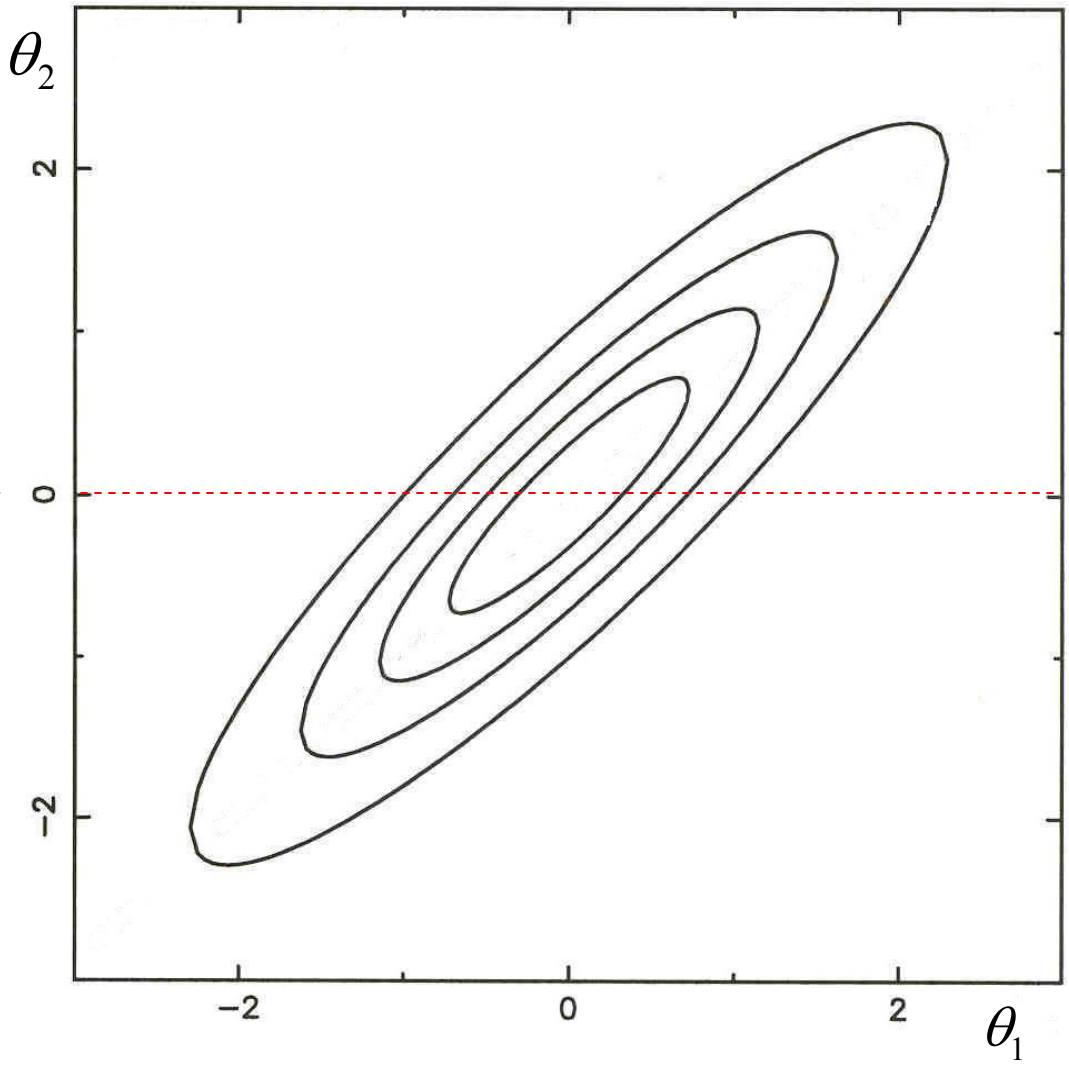
$\Rightarrow$ isoprobability contours elongate

Very important if we are interested only in *one* parameter



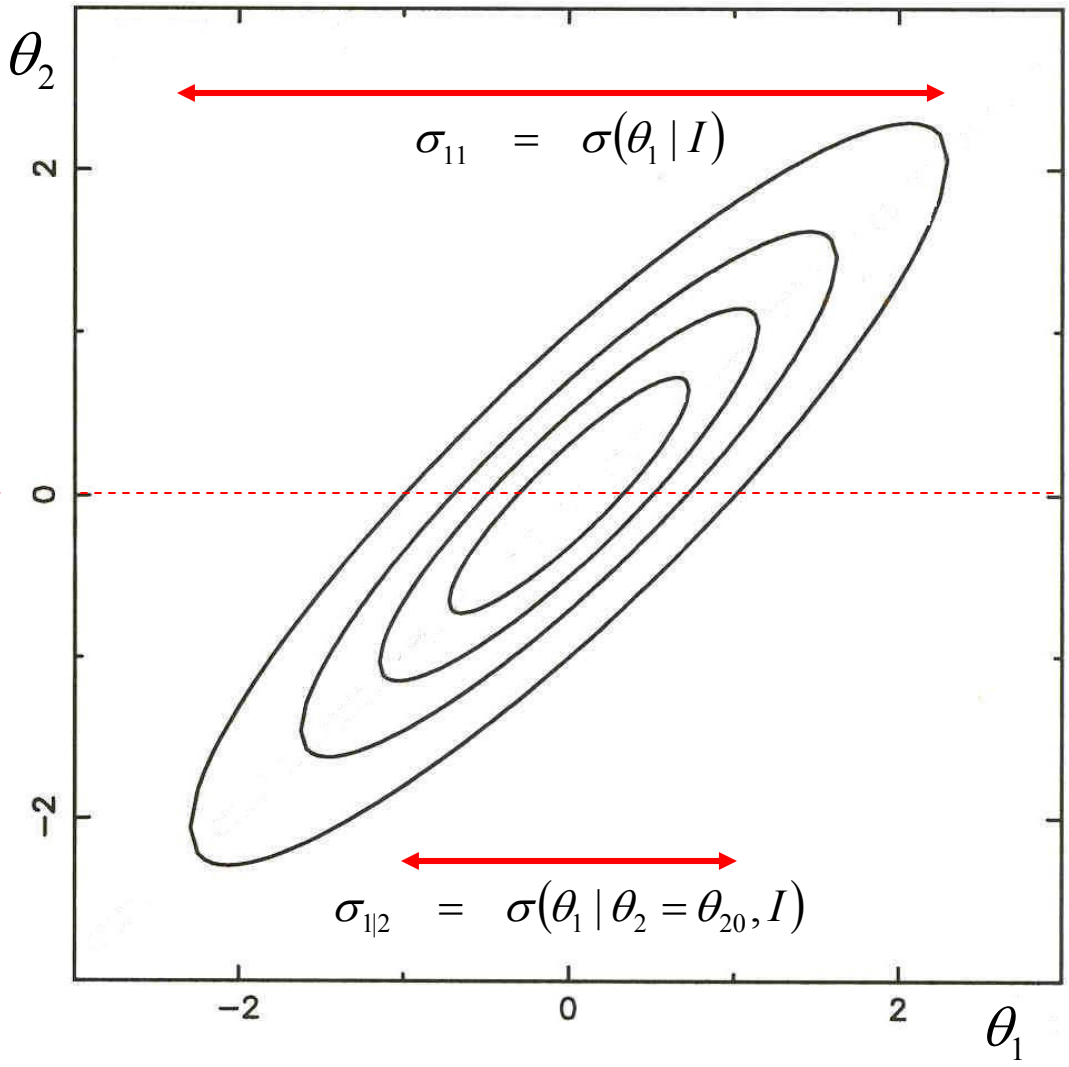
Parameter estimation: 2-D case

'Best-fit' value of θ_2 , found from $\left.\frac{\partial \ell}{\partial \theta_j}\right|_{\theta_j = \theta_{0j}} = 0$



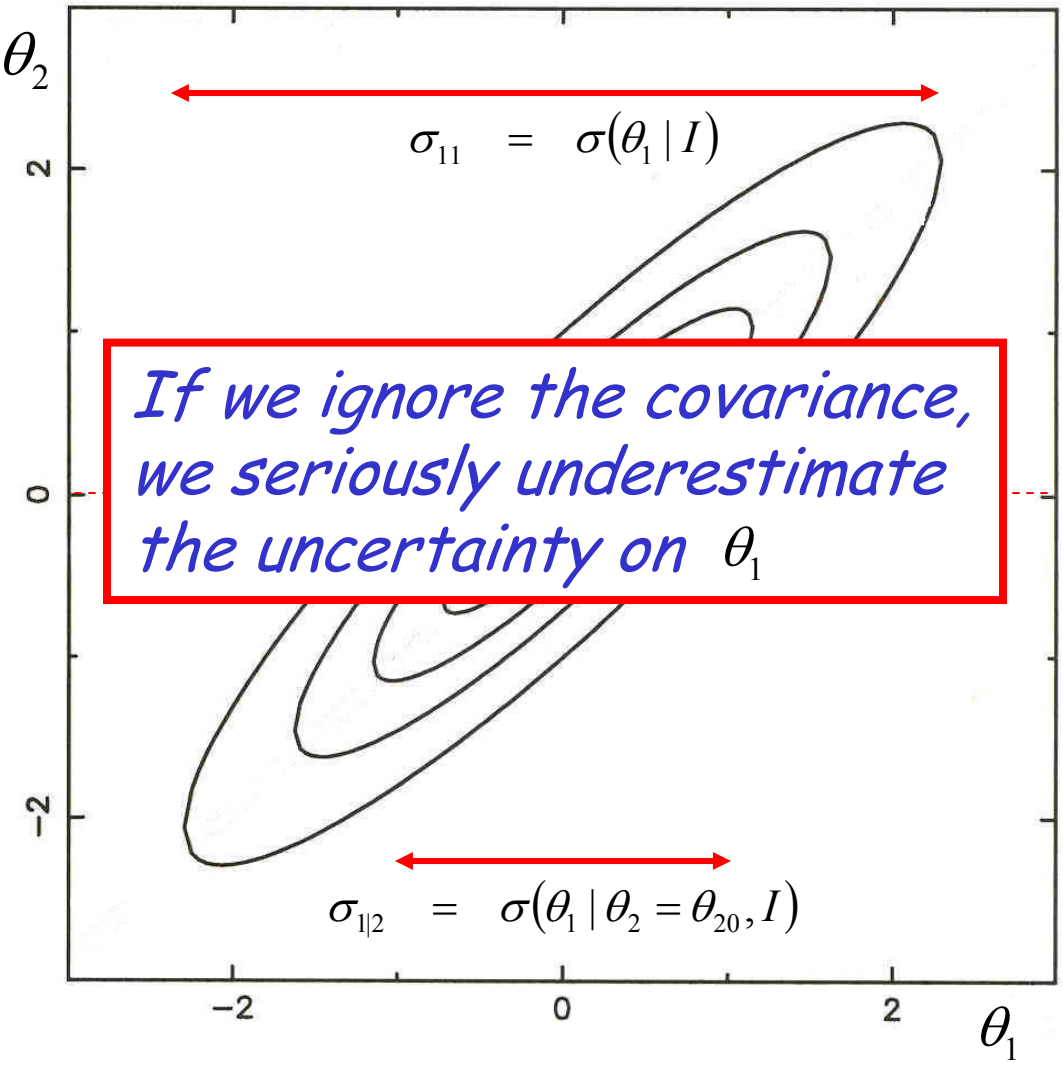
Parameter estimation: 2-D case

'Best-fit' value of θ_2 , found from $\left.\dfrac{\partial \ell}{\partial \theta_j}\right|_{\theta_j=\theta_{0j}} = 0$



Parameter estimation: 2-D case

'Best-fit' value of θ_2 , found from $\left.\dfrac{\partial \ell}{\partial \theta_j}\right|_{\theta_j=\theta_{0j}} = 0$



If we ignore the covariance, we seriously underestimate the uncertainty on θ_1

Question 10: The marginal and conditional error bars on θ_1 will be equal provided

A $\text{cov}[\theta_1, \theta_2] = 0$

B $\text{cov}[\theta_1, \theta_2] = 1$

C $\text{cov}[\theta_1, \theta_2] = -1$

D None of the above

Question 10: The marginal and conditional error bars on θ_1 will be equal provided

A $\text{cov}[\theta_1, \theta_2] = 0$

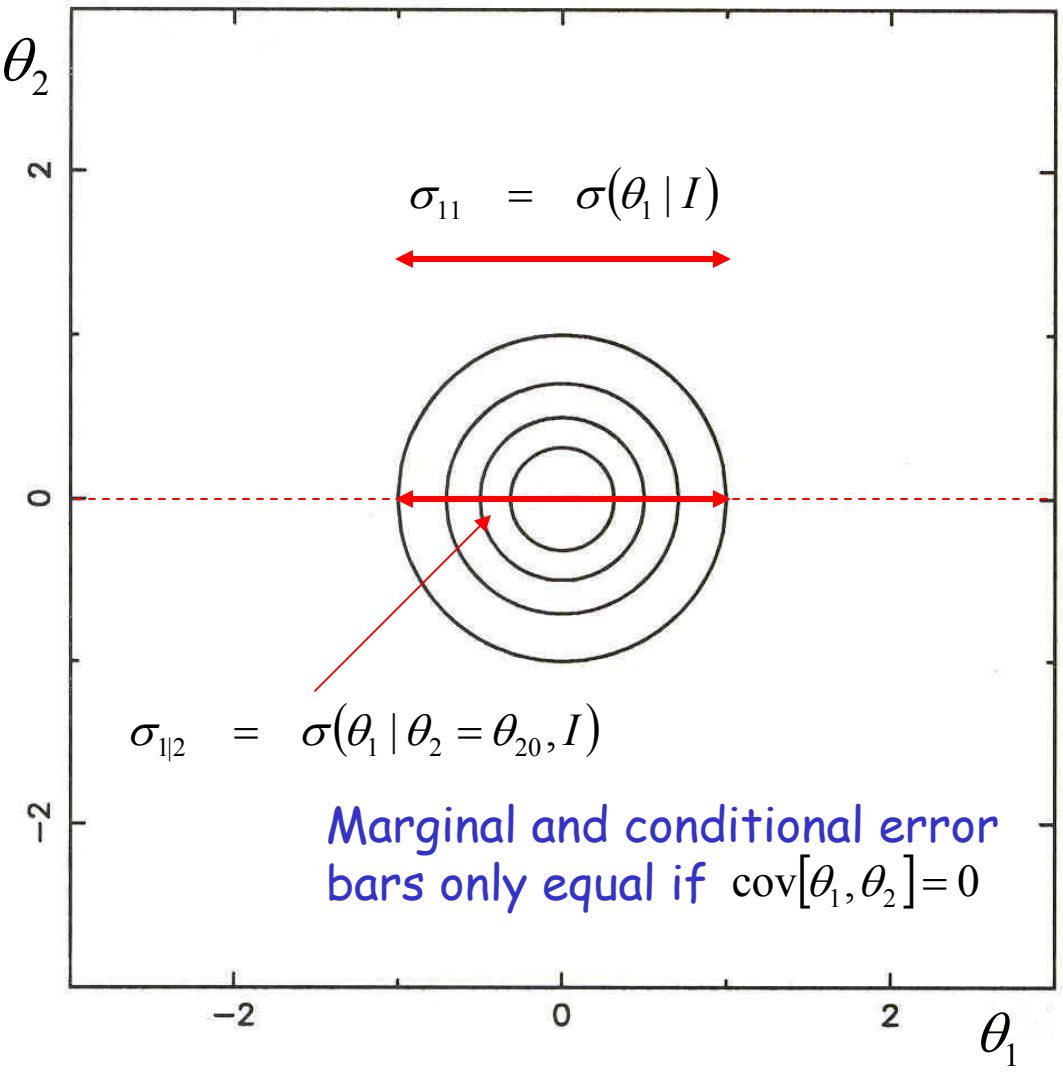
B $\text{cov}[\theta_1, \theta_2] = 1$

C $\text{cov}[\theta_1, \theta_2] = -1$

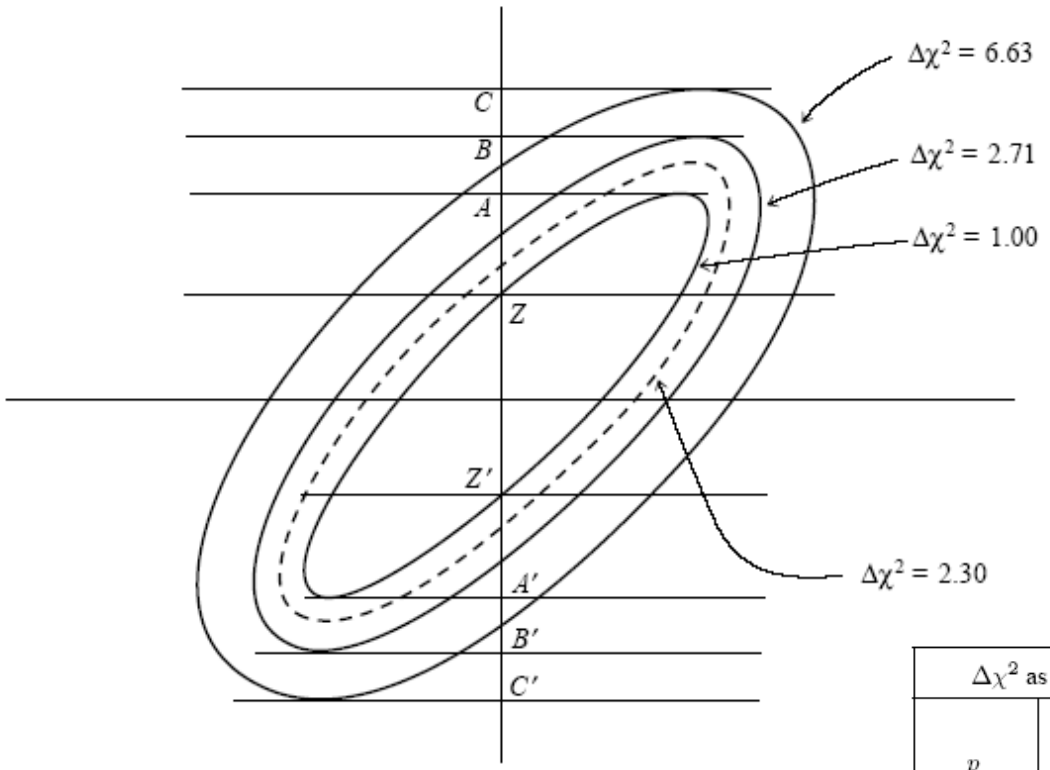
D None of the above

Parameter estimation: 2-D case

'Best-fit' value of θ_2 , found from $\left.\dfrac{\partial \ell}{\partial \theta_j}\right|_{\theta_j=\theta_{0j}} = 0$



Parameter estimation: 2-D case

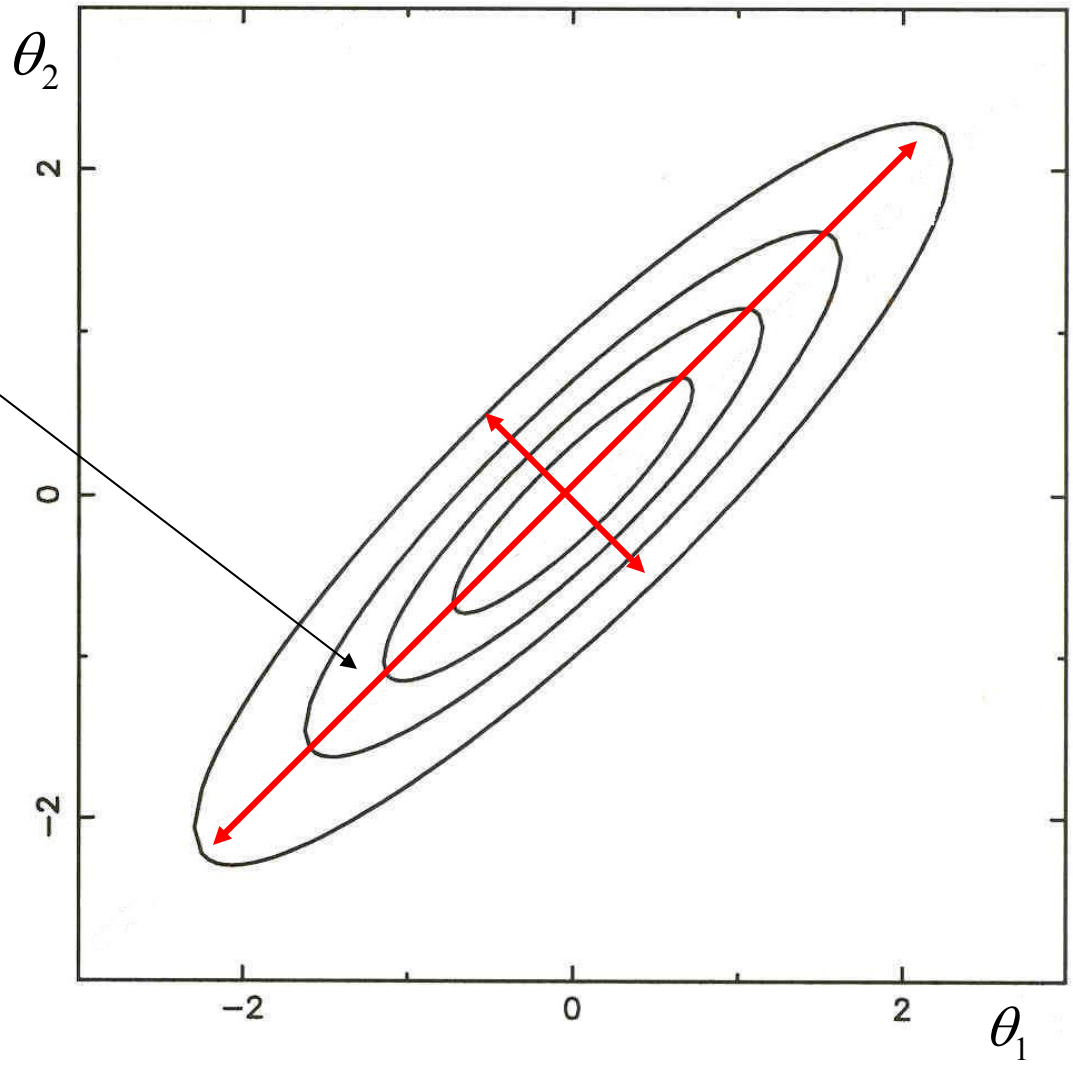


From Numerical Recipes

$\Delta\chi^2$ as a Function of Confidence Level and Degrees of Freedom						
	ν					
p	1	2	3	4	5	6
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8

Parameter estimation: 2-D case

Linear combination of θ_1 and θ_2 well constrained by data



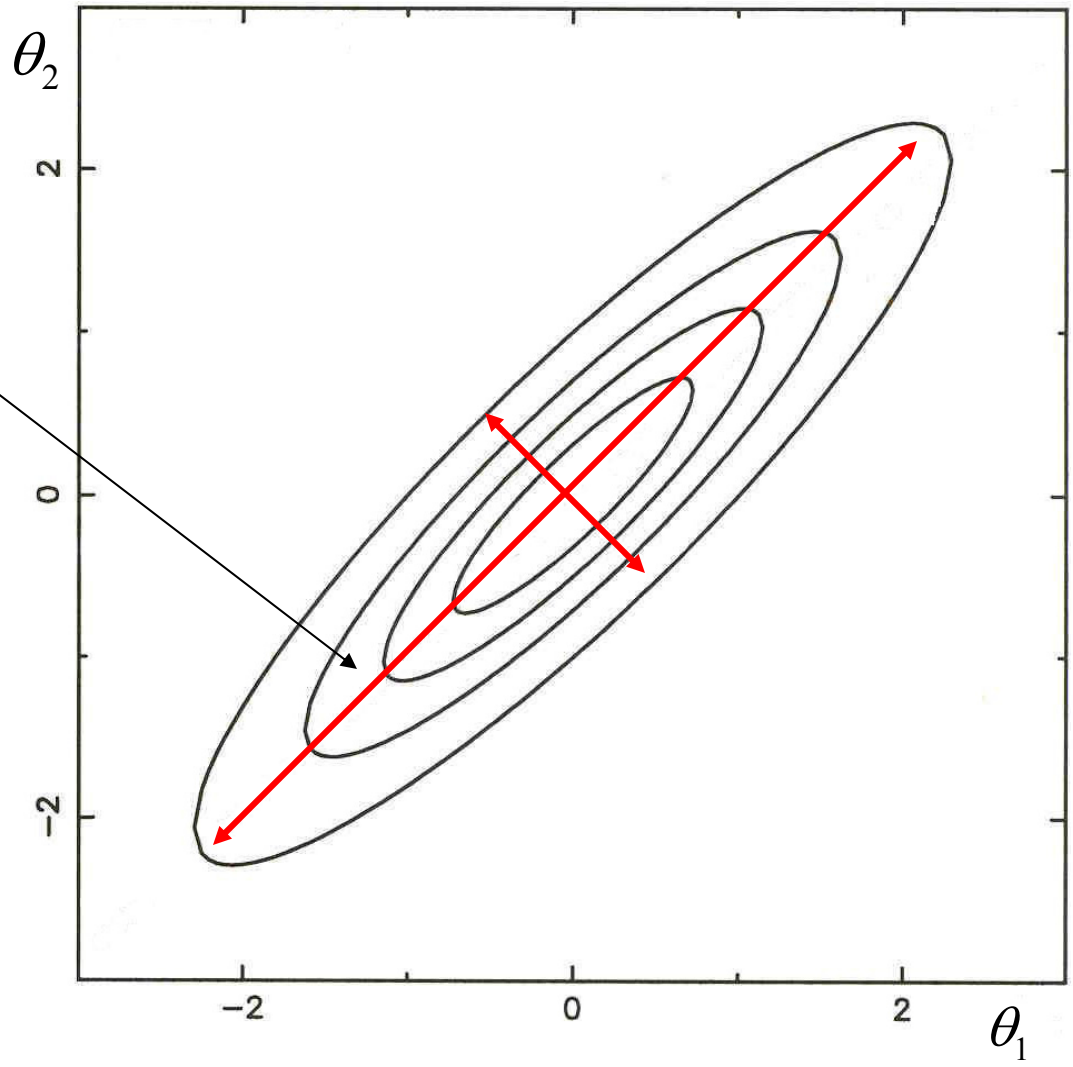
Parameter estimation: 2-D case

Linear combination of θ_1 and θ_2 well constrained by data

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

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

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



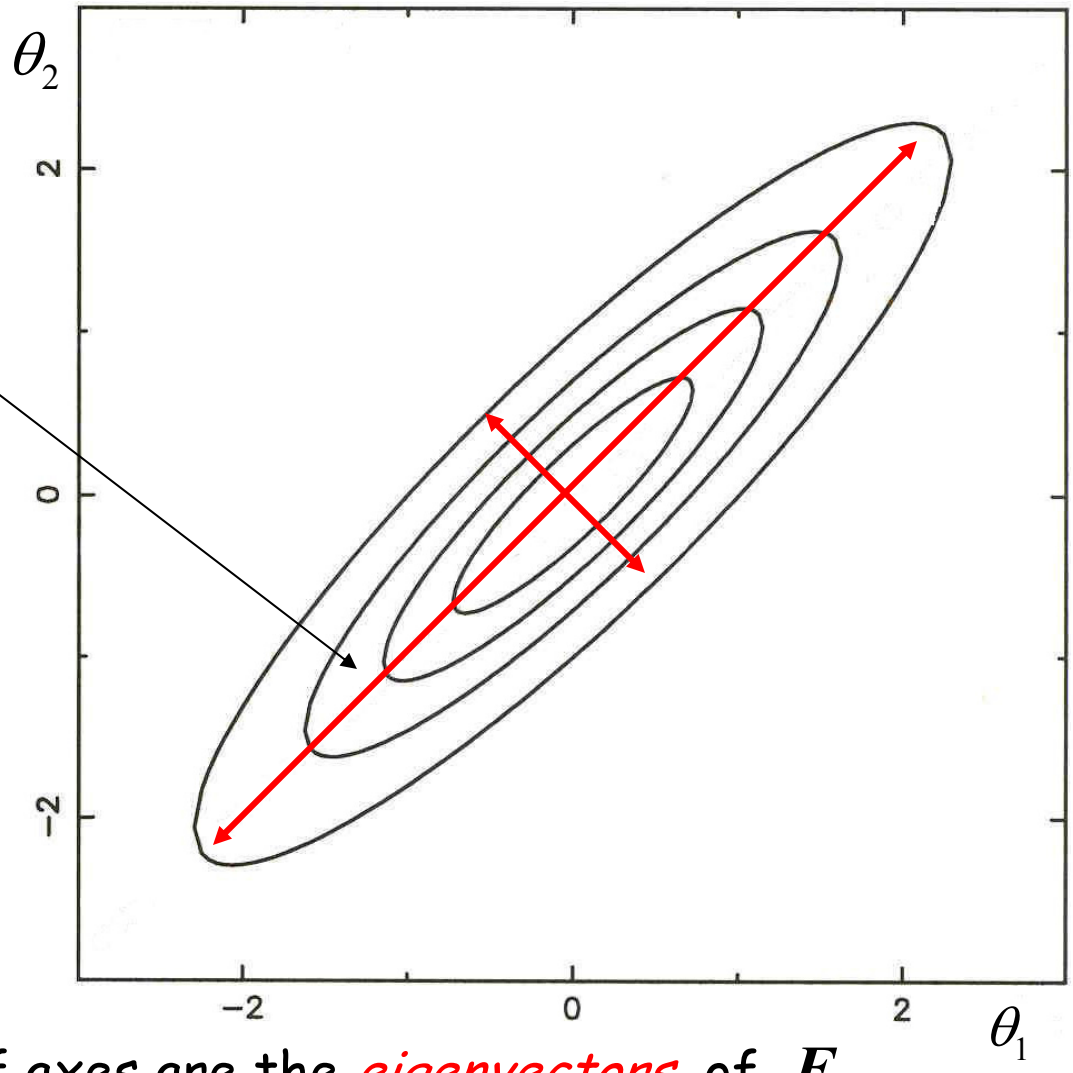
Parameter estimation: 2-D case

Linear combination of θ_1 and θ_2 well constrained by data

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

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

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



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

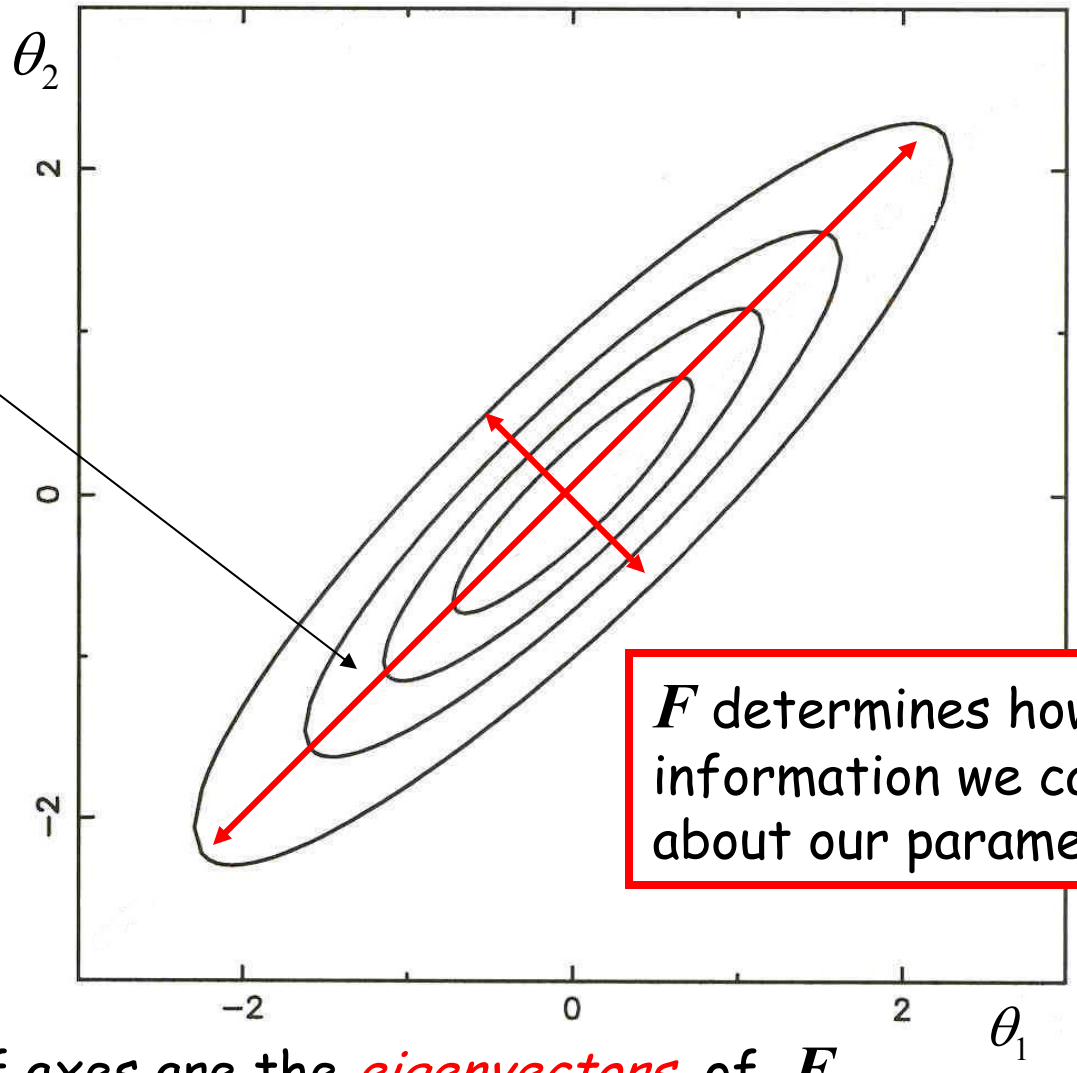
Parameter estimation: 2-D case

Linear combination of θ_1 and θ_2 well constrained by data

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

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

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



F determines how much information we can learn about our parameters

Direction of axes are the *eigenvectors* of ***F***

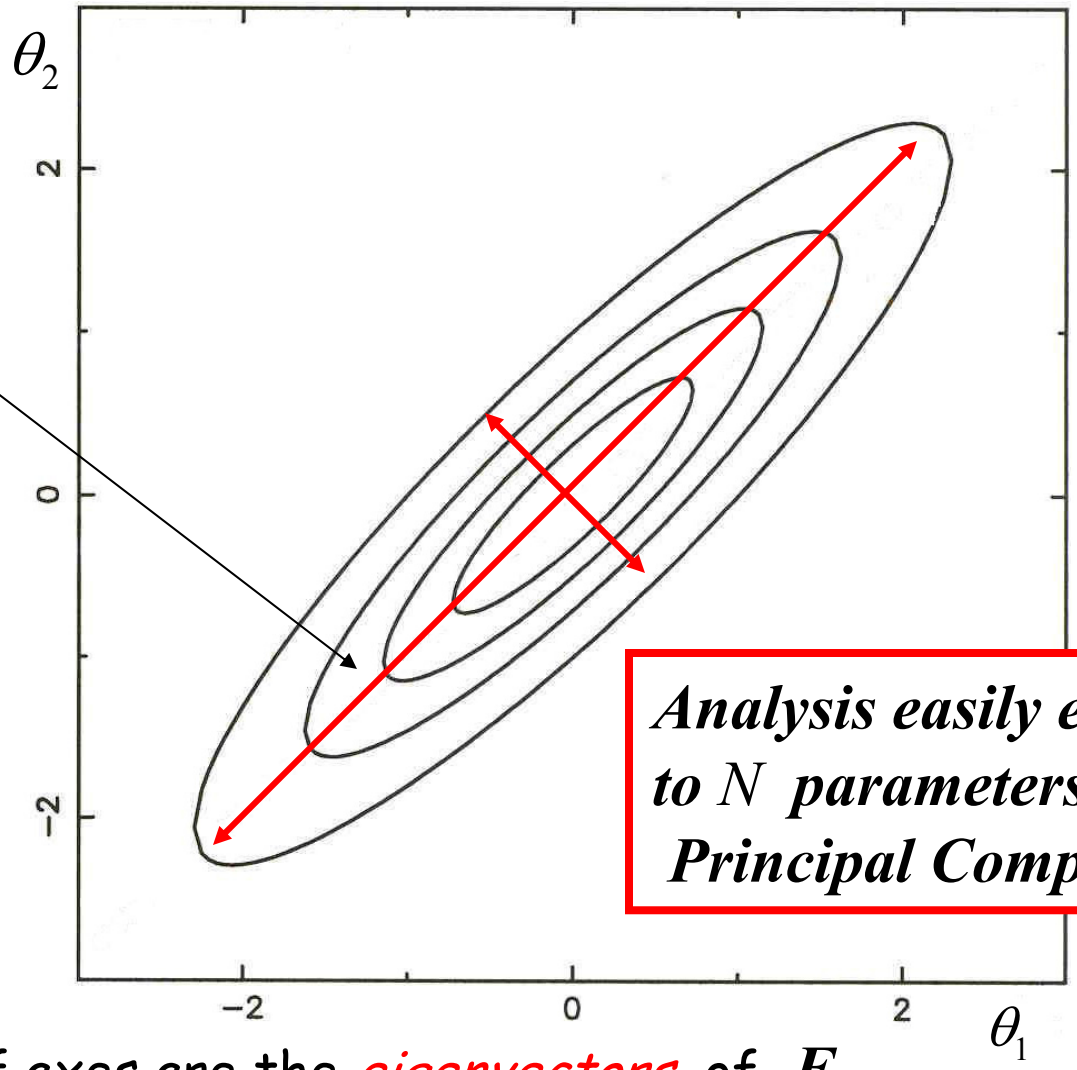
Parameter estimation: N-dimensional case

Linear combination of θ_1 and θ_2 well constrained by data

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

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

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



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*

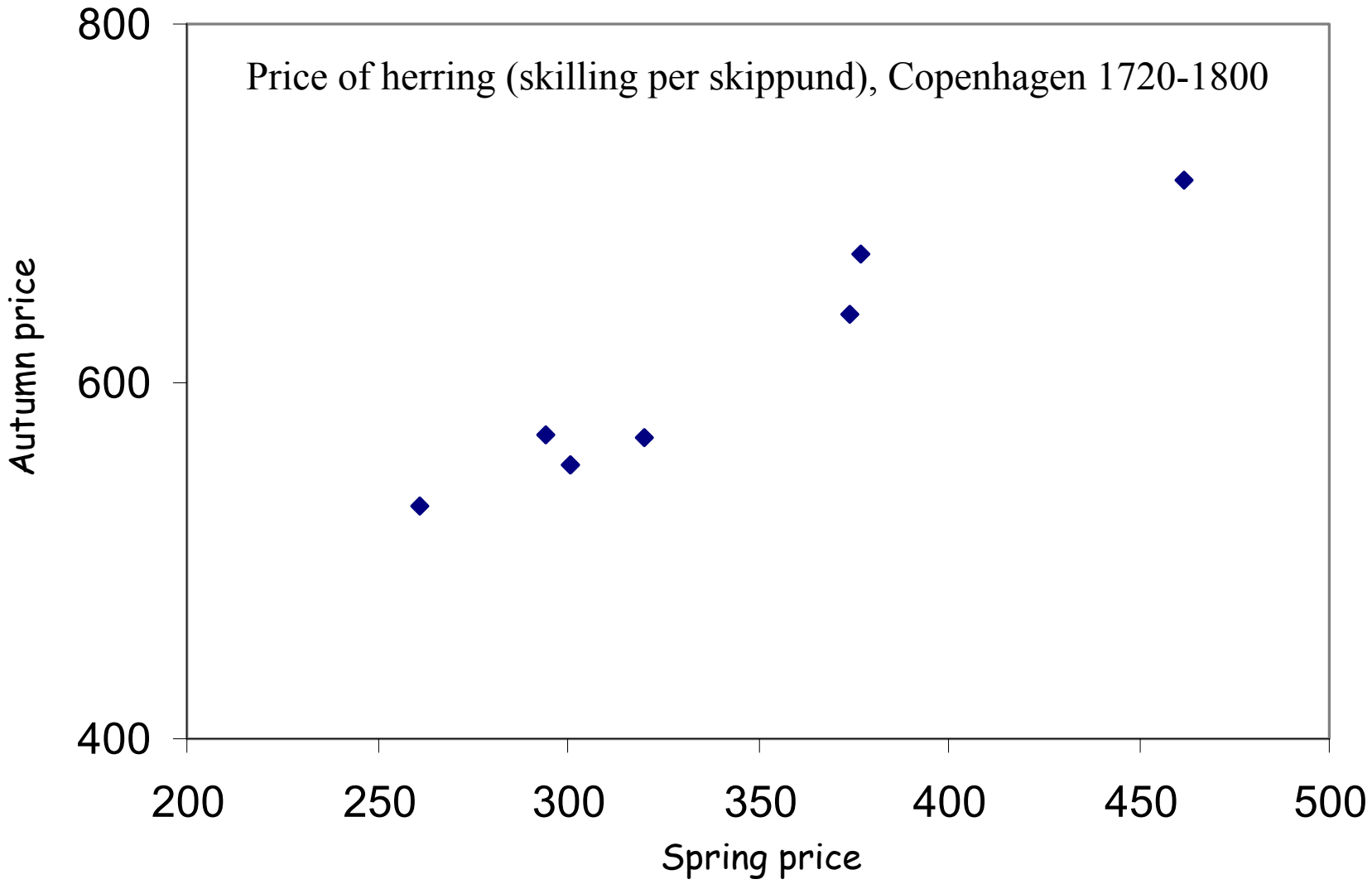
What is PCA?

Principal Component Analysis:

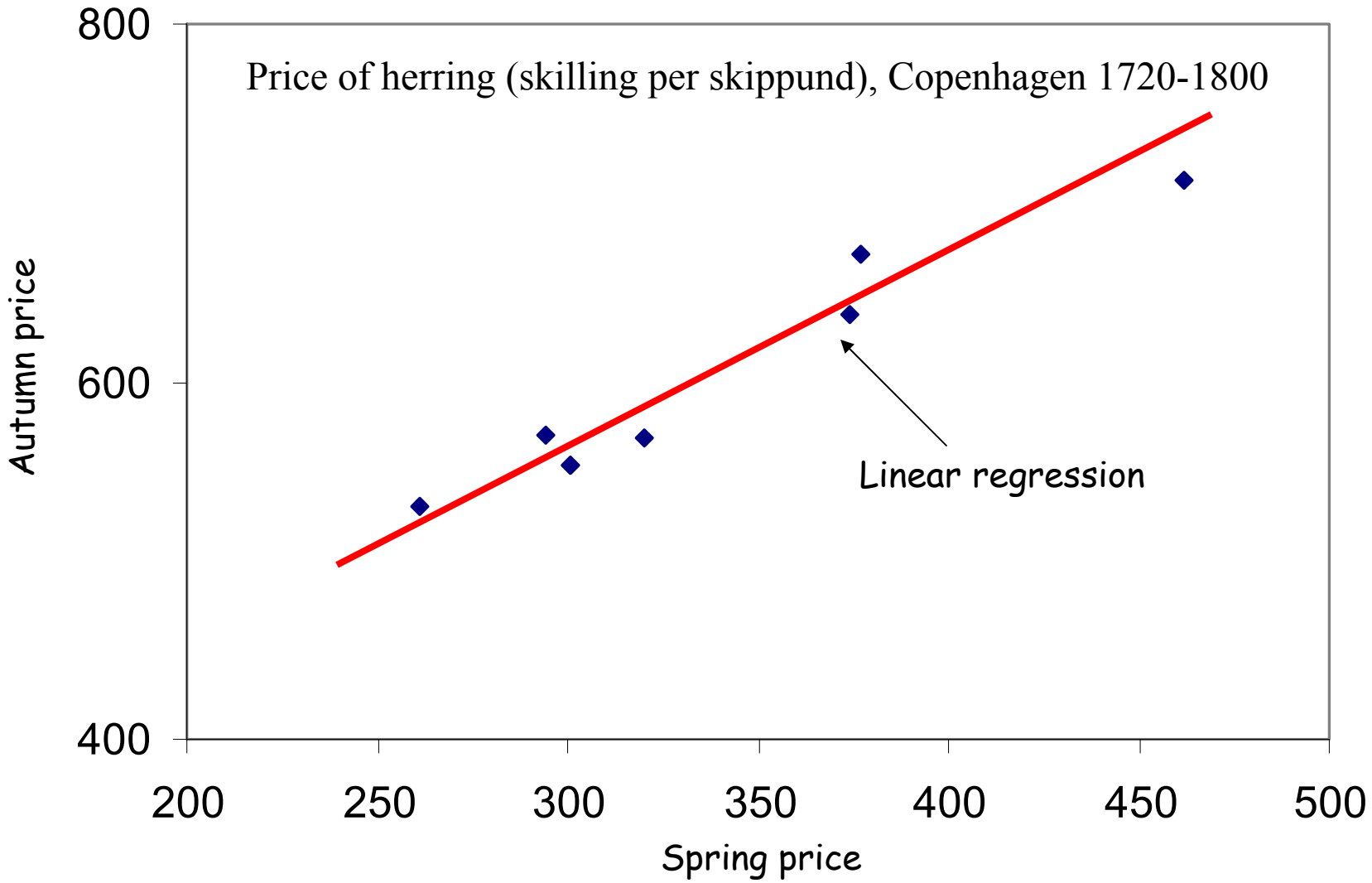
1st principal component = *linear combination of the original variables that accounts for as much of the variation in the data as possible*

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

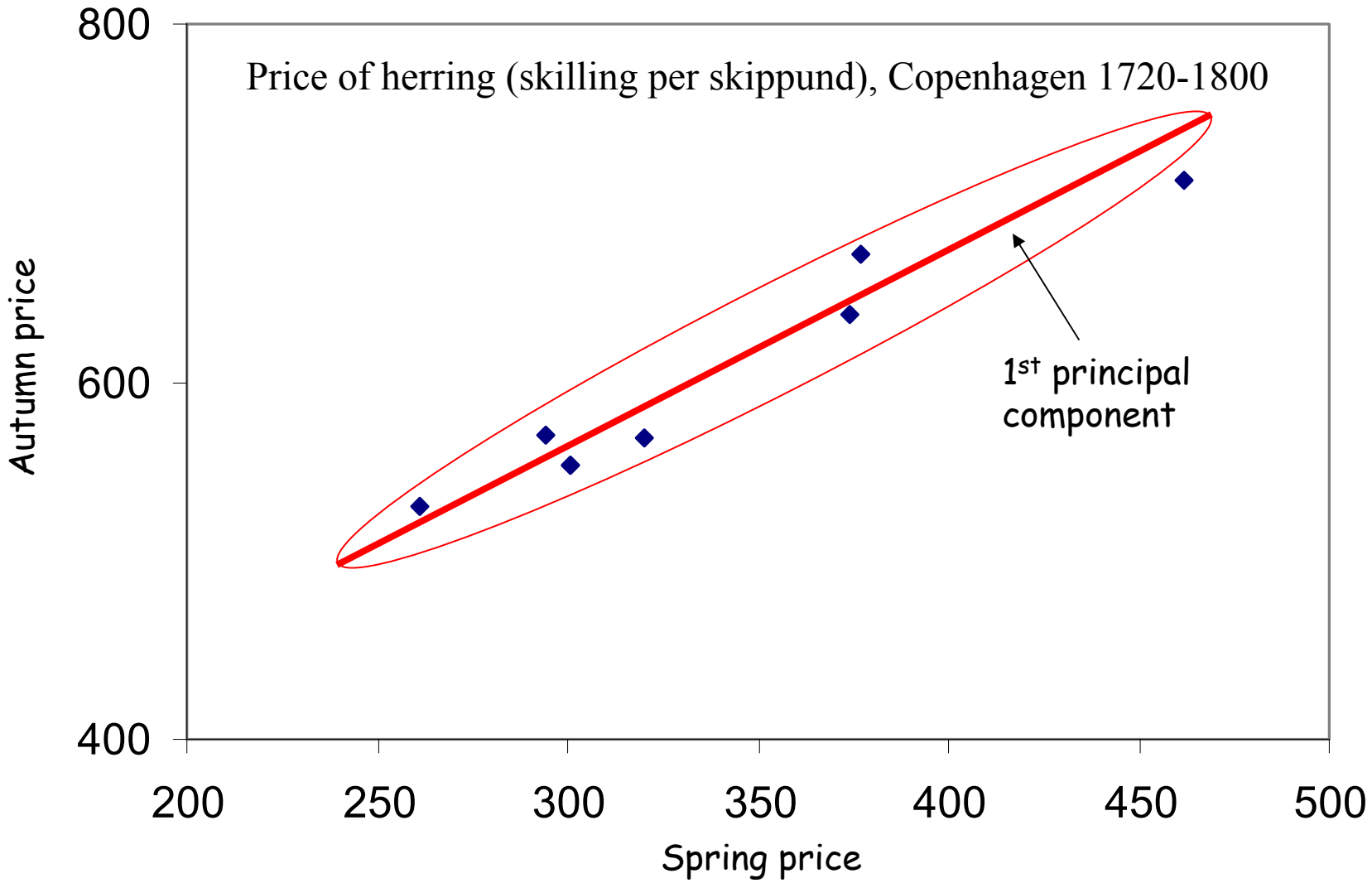
The price of fish…



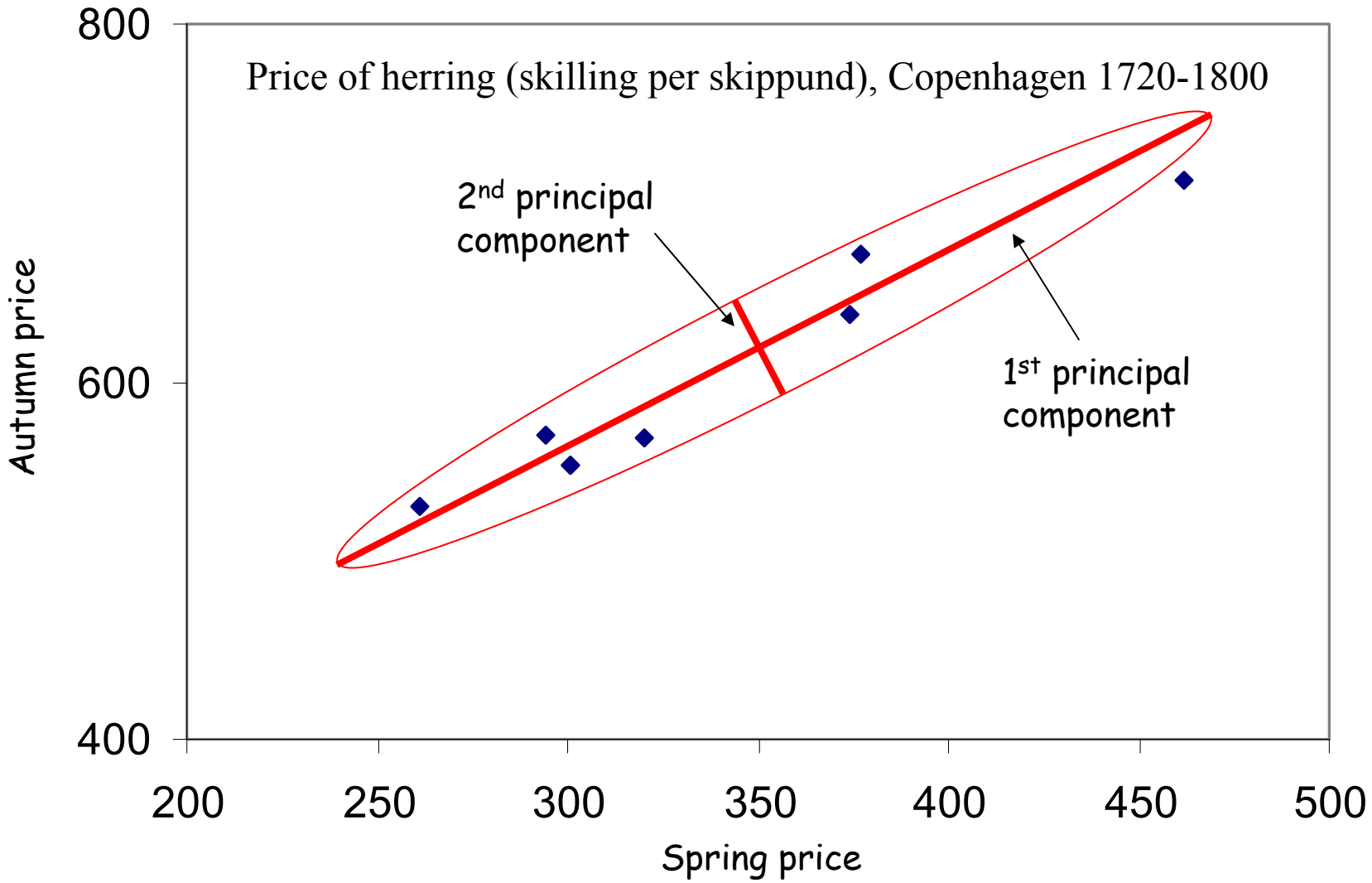
The price of fish…



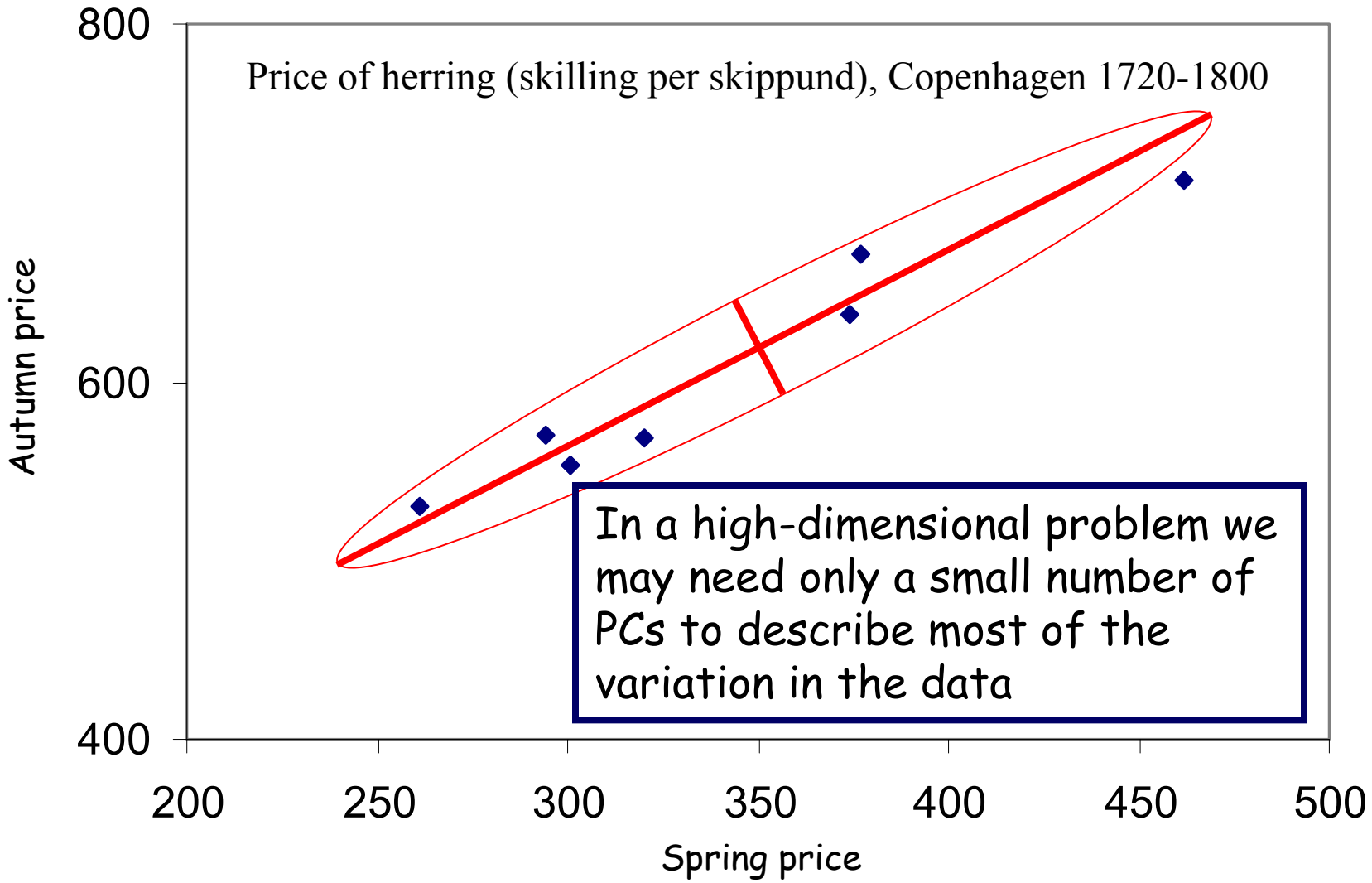
The price of fish…



The price of fish…



The price of fish…



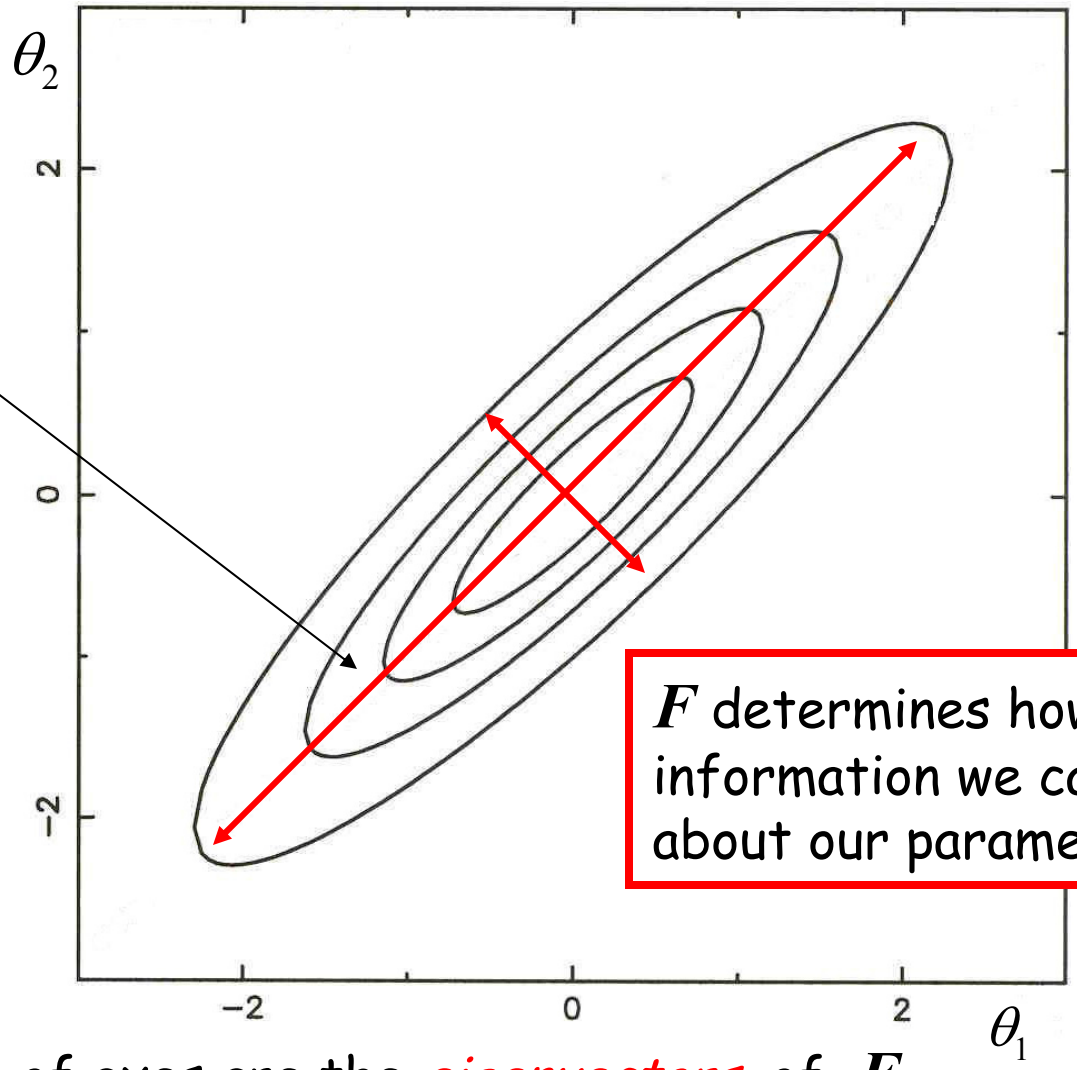
Parameter estimation: 2-D case

Linear combination of θ_1 and θ_2 well constrained by data

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

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

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



F determines how much information we can learn about our parameters

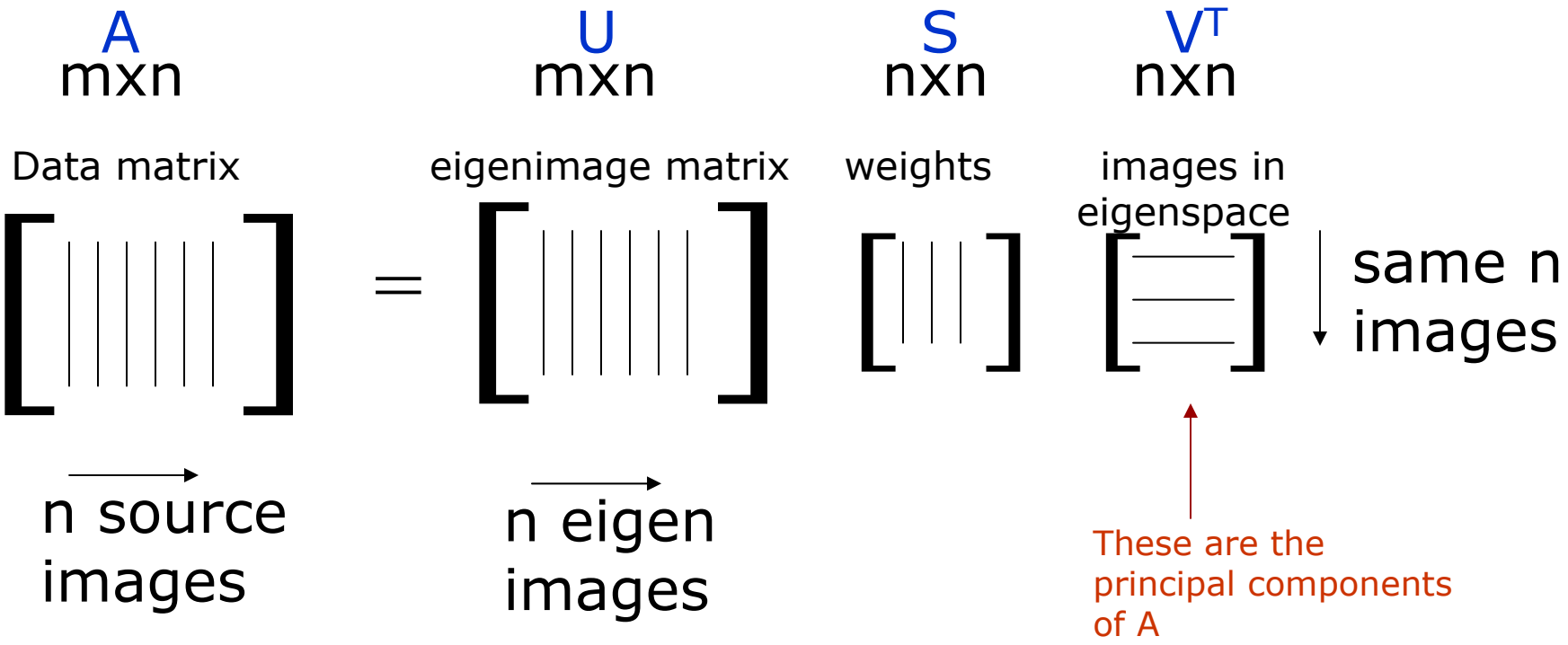
Direction of axes are the *eigenvectors* of ***F***

More advanced PCA

- PCA can be applied to any number of datasets of any dimension
- Widely used for data compression and data characterisation by only considering the highest principal components, and throwing away those that contain little correlation information.
- E.g., “eigenfaces”…

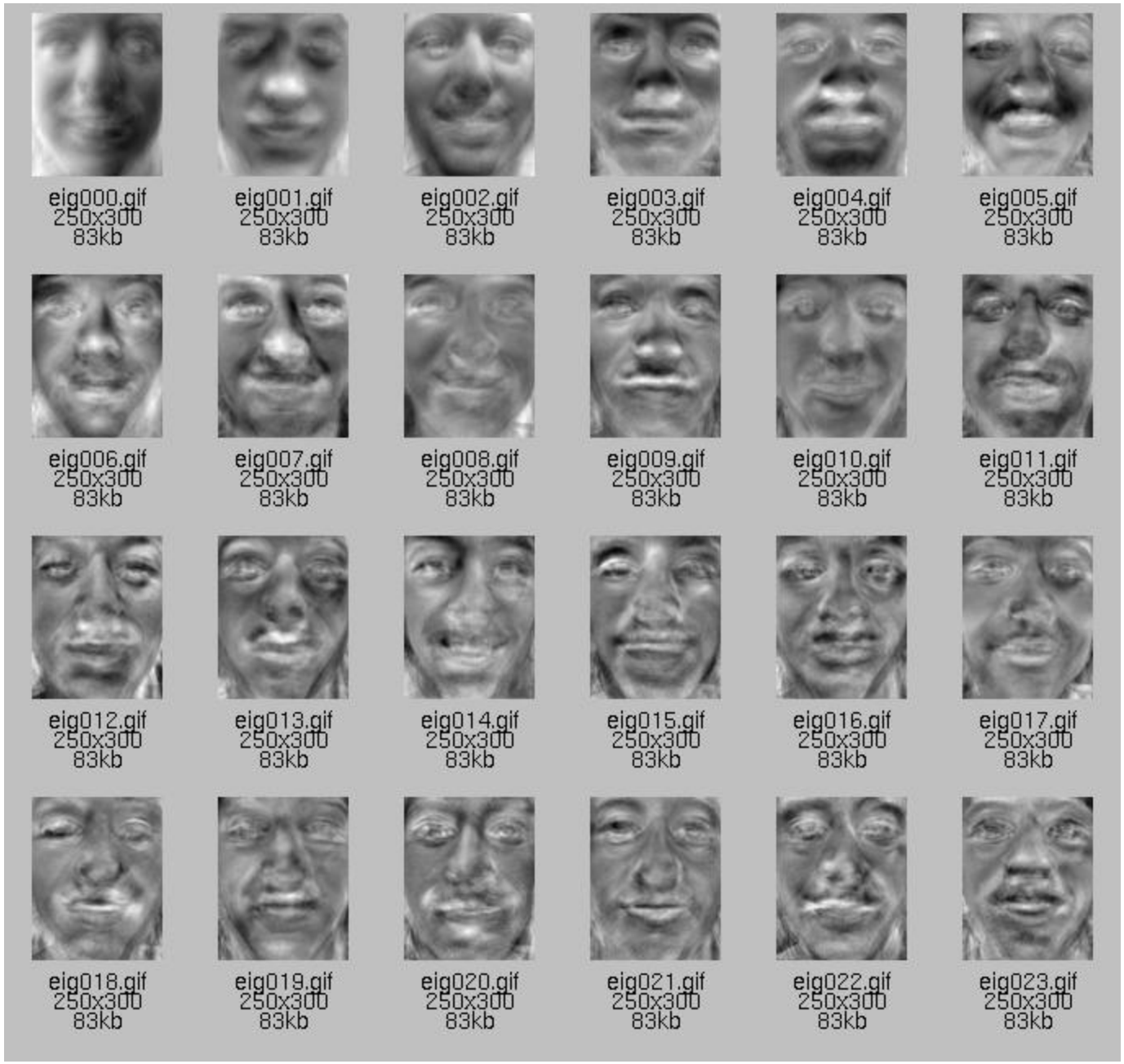
Generating eigenvectors by SVD

- A useful way of generating the principal components is by carrying out a singular value decomposition (SVD) on the data matrix:



$$A = USV^T$$

Rice University Eigenfaces Group



Rice University Eigenfaces Group

Biophysical Journal

HOME HELP FEEDBACK SUBSCRIPTIONS ARCHIVE SEARCH TABLE OF CONTENTS

Biophysical Journal 84:2112-2129 (2003)

Application of Singular Value Decomposition to the Analysis of Time-Resolved Macromolecular X-Ray Data

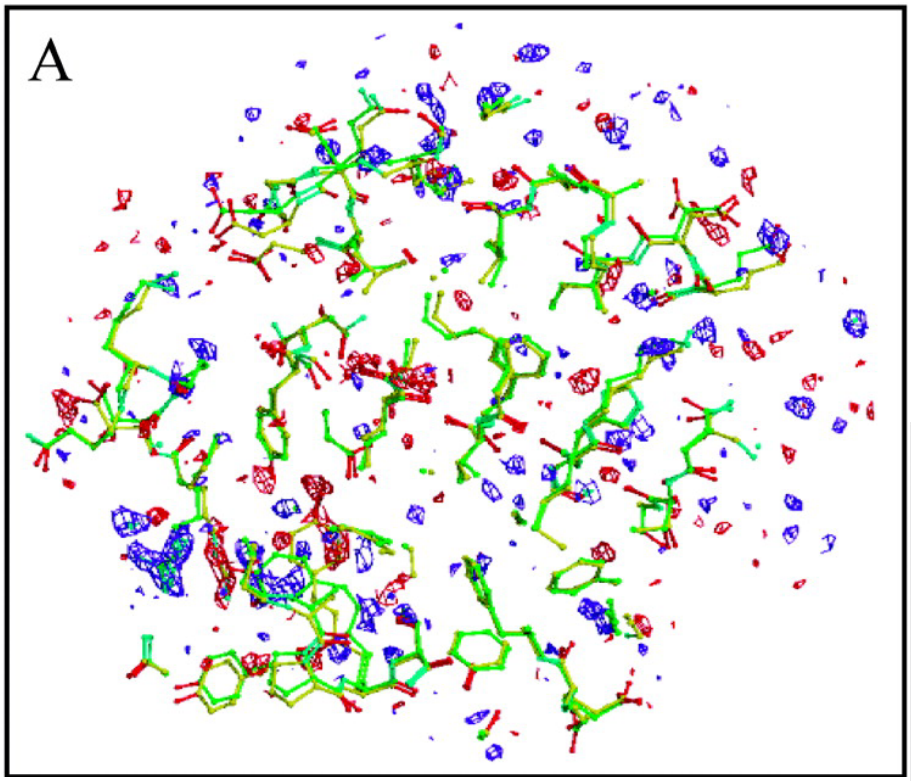
Marius Schmidt[*,†], **Sudarshan Rajagopal**[†], **Zhong Ren**[†,‡,§] **and Keith Moffat**[†,‡]

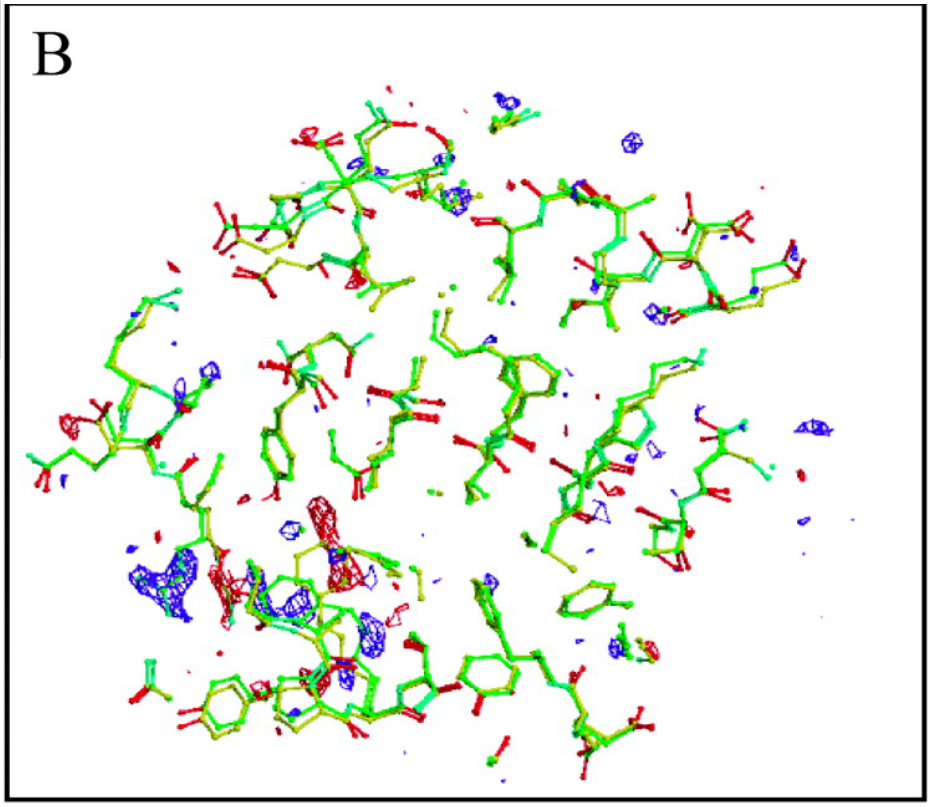
[*] Physik-Department E17, Technische Universitaet Muenchen, 85747 Garching, Germany; [†] Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, Illinois 60637 USA; [‡] BioCARS, Argonne National Laboratory, Argonne, Illinois 60439 USA; and [§] Renz Research, Des Plaines, Illinois 60018 USA

Correspondence: Address reprint requests to Marius Schmidt, E-mail: marius@hexa.e17.physik.tu-muenchen.de.

ABSTRACT

Singular value decomposition (SVD) is a technique commonly used in the analysis of spectroscopic data that both acts as a noise filter and reduces the dimensionality of subsequent least-squares fits. To establish the applicability of SVD to crystallographic data, we applied SVD to calculated difference Fourier maps simulating those to be obtained in a time-resolved crystallographic study of photoactive yellow protein. The atomic structures of one dark state and three intermediates were used in


A


B

Parameter estimation: Gaussian approximation

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

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

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

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

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

Is the Gaussian approximation a good idea?

Parameter estimation: Gaussian approximation

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

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

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

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} :

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

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

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} :

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

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

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

Markov Chain Monte Carlo Methods – see later

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. New 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

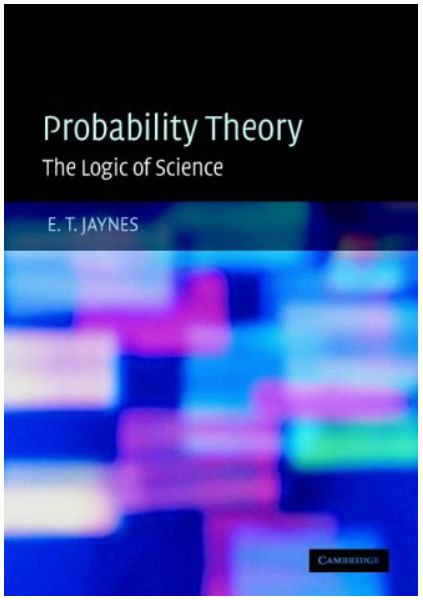
Below are the the sound sources separated by ICA. Note that they might be in different order than the original ones.

http://www.cis.hut.fi/projects/ica/cocktail/cocktail_en.cgi

4b. 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)

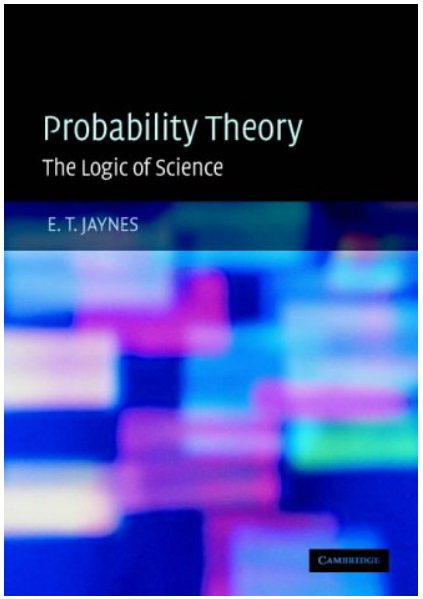
Probability *is* subjective; it depends on the available information

Subjective ≠ 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

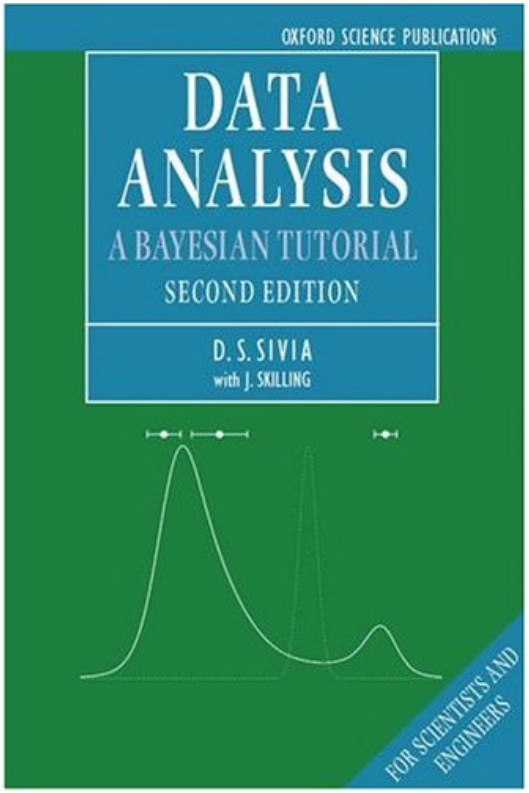
Given the *same* background information, two observers should assign the *same* probabilities

But what should they be?...

Bernoulli (1713) 'Principle of insufficient reason'

Keynes (1921) 'Principle of indifference'

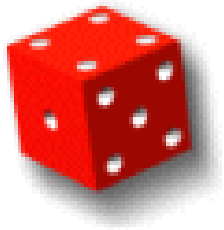
If we can enumerate a set of basic mutually exclusive possibilities, and we have no reason to believe that any one of these is more likely to be true than another, then we should assign the same probability to all.



Bernoulli (1713) 'Principle of insufficient reason'

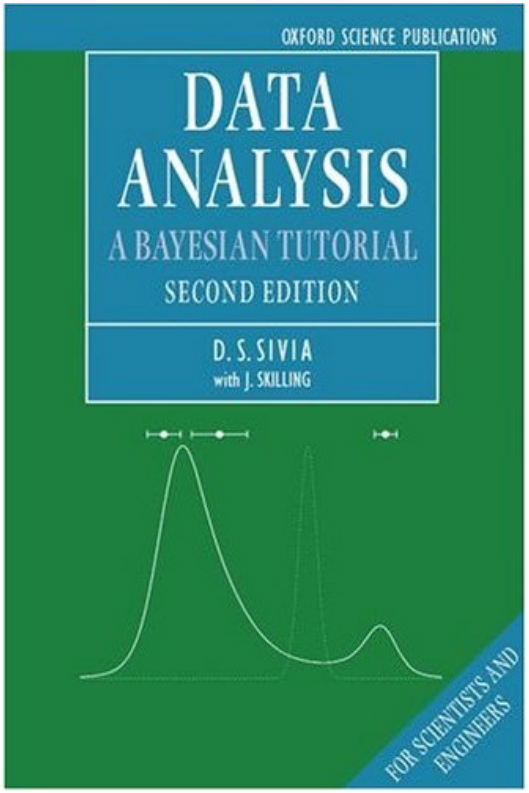
Keynes (1921) 'Principle of indifference'

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.



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

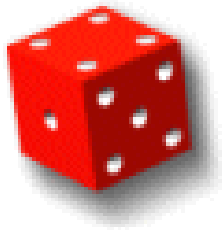
$$p(X_i \mid I) = \tfrac{1}{6} \quad \text{for all } i$$



Bernoulli (1713) 'Principle of insufficient reason'

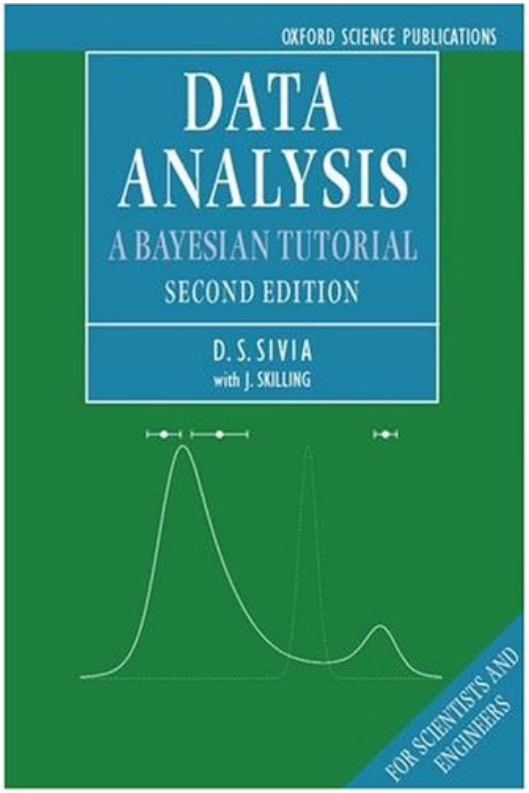
Keynes (1921) 'Principle of indifference'

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.



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

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

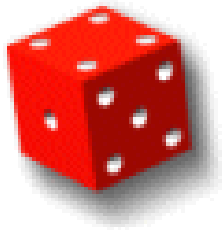


Agrees with common sense, but can we justify more fundamentally?

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

Extending to continuum case,

Let x be a location parameter.

Principle of indifference means we should have

$$p(x \mid I)dx = p(x+\Delta \mid I)d(x+\Delta)$$

where Δ is a constant

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

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

Similarly,

Let L be a scale parameter.

Principle of indifference means we should have

$$p(L \mid I)dL = p(\beta L \mid I)d(\beta L)$$

where β is a positive constant

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

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

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. $$p(\log L \,|\, I)dL \quad = \quad \text{constant}$$

In fact what is referred to as a Jeffreys prior $p(L \,|\, I) \propto 1/\mathrm{L}$ 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

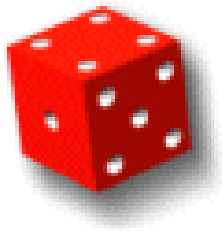
$$I(\vec{\theta})_{i,j} = E\left[\frac{\partial}{\partial\theta_i}\ln L(\vec{\theta})\frac{\partial}{\partial\theta_j}\ln L(\vec{\theta})\right]$$

[Note this expression reduces to that for the Fisher matrix given in Section 4a 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 $p(X_i \mid I)$?

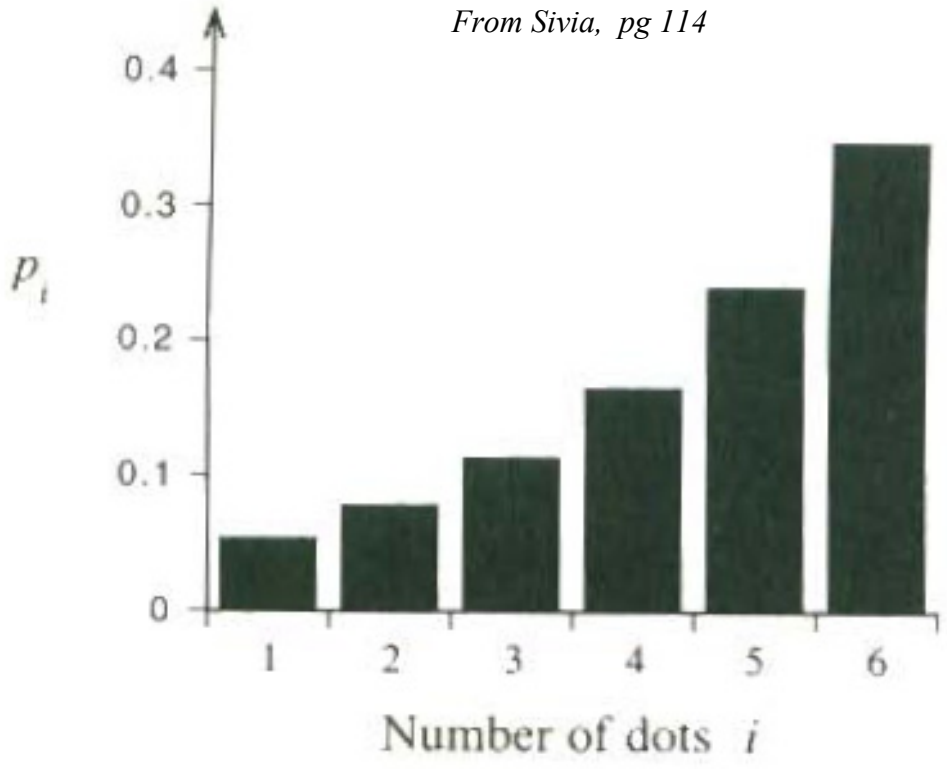
Jaynes (1957) suggests maximising the **Entropy**

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

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 (see notes on website)

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 11: Assuming that eye-colour and handedness are uncorrelated 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

Question 11: Assuming that eye-colour and handedness are uncorrelated 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-Eyed \ *Left-Handed*	True	False
True	p_1	p_2
False	p_3	p_4

Blue-Eyed \ *Left-Handed*	True	False
True	$0 \le x \le \frac{1}{3}$	$\frac{1}{3} - x$
False	$\frac{1}{3} - x$	$\frac{1}{3} + x$

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 x ?

Independence arguments favour $x = 1/9$

Blue-Eyed \ Left-Handed	True	False
True	p_1	p_2
False	p_3	p_4

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

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 x ?

Independence arguments favour $x = 1/9$

	Variational function	Optimal x	Implied correlation
MAXENT →	$-\Sigma p_i \log_e(p_i)$	0.1111	None
	$-\Sigma p_i^2$	0.0833	Negative
	$\Sigma \log_e(p_i)$	0.1301	Positive
	$\Sigma \sqrt{p_i}$	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

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

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

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

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 $p(x\,|\,I)$?

Using MAXENT it can be shown that

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

Normal 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 $p(x\,|\,I)$?

Using MAXENT it can be shown that

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

Normal distribution

MAXENT justifies the relevance of common pdfs