ETH zürich

Lecture 2: Model Selection

Prof Dagmar Iber, PhD DPhil

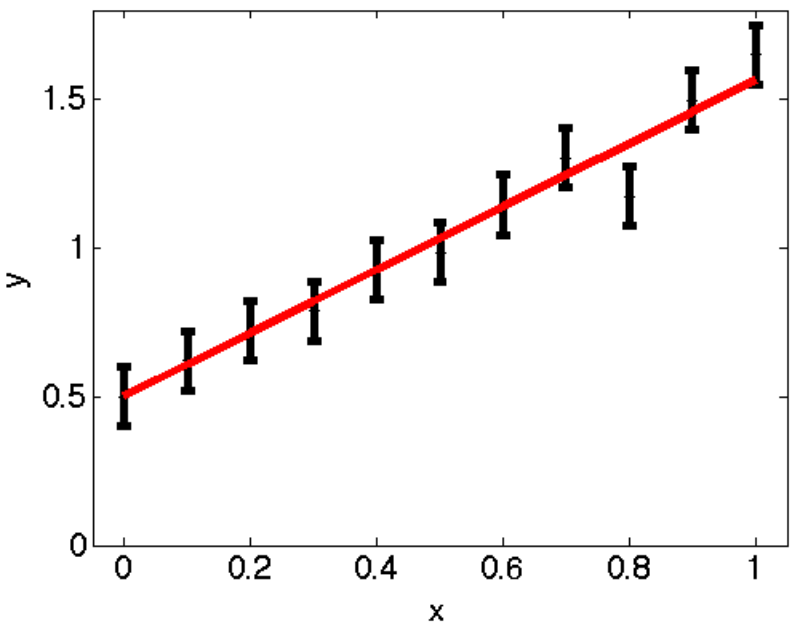
LMS Research School: PDEs in Mathematical Biology

Computational Biology Group (CoBI), D-BSSE, ETHZ

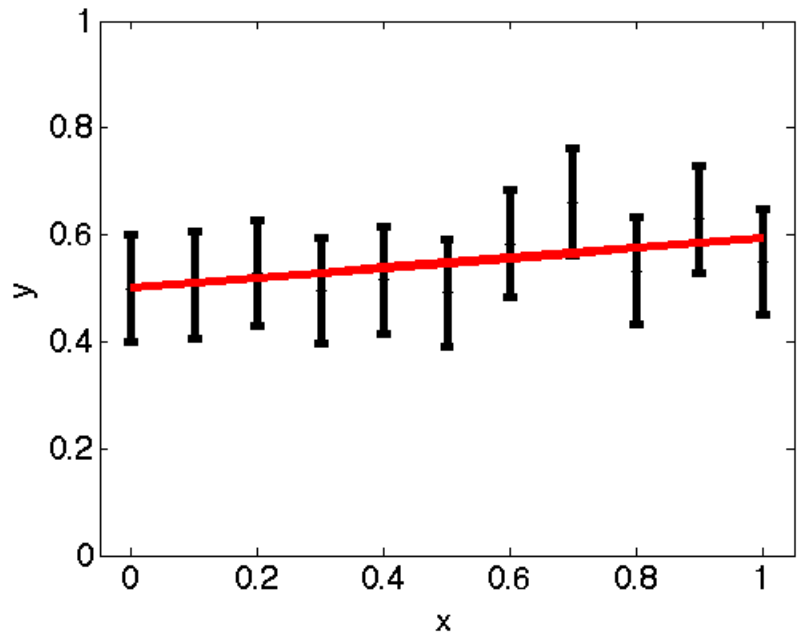
Contents

The problem: What model to select?

A fit of the function $y = ax + b$ (red) to the data (black).



A fit of the function $y = ax + b$ (red) to the data (black). Is this overfitting? Should we set $a = 0$?



Parameter Estimation

Bayes' Theorem for Parameter Estimation

According to Bayes' Theorem

$$\text{prob}(X|D,I) = \frac{\text{prob}(D|X,I) \times \text{prob}(X|I)}{\text{prob}(D|I)}$$

- $\text{prob}(X|D,I)$: **posterior probability density function (pdf)** that we want to determine.
- $\text{prob}(D|X,I)$: **likelihood function**
- $\text{prob}(X|I)$: **prior probability density function (pdf)** that reflects our knowledge about the system
- $\text{prob}(D|I)$: **evidence**, i.e. the likelihood of the data based on our knowledge. Here one could incorporate knowledge about the quality of different experimental techniques or experimental groups.

Maximum likelihood estimate

$$\begin{aligned} \text{prob}(X|D,I) &\propto \text{prob}(D|X,I) \\ \text{posterior pdf} &\propto \text{likelihood function.} \end{aligned} \tag{1}$$

Maximum likelihood estimate

Our best estimate X_0, given by the maximum of the posterior, is equivalent to the solution that yields the greatest value for the probability of the observed data.

Assume Gaussian Process

We can then write

$$\text{prob}(D_k|X, I) = \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left(-\frac{(F_k - D_k)^2}{2\sigma_k^2}\right). \quad (2)$$

We can rewrite this equation as

$$\text{prob}(D|X, I) \propto \exp\left(-\frac{\chi^2}{2}\right) \quad \text{with} \quad \chi^2 = \sum_k \left(\frac{F_k - D_k}{\sigma_k}\right)^2 = \sum_k \frac{R_k^2}{\sigma_k^2}.$$

Residuals

The $R_k = F_k - D_k$ are referred to as residuals.

Least-squares estimate

Take the logarithm of the likelihood function:

$$L = \ln\left(\mathrm{prob}(D|X, I)\right) = const - \frac{\chi^2}{2}. \tag{3}$$

Least-squares estimate

Since the maximum of the posterior will occur when χ^2 is smallest, the corresponding optimal solution X_0 is called least-squares estimate.

Local Optimization Methods: Gradient-based Methods

Iterative Linearization

We can expand the likelihood function around its maximum L_0 as

$$L = L(\theta_0)) + \underbrace{\sum_i \nabla_i L(\theta_i - \theta_{i0})}_{=0} + \underbrace{\frac{1}{2}\sum_{i,j}(\nabla\nabla L)_{ij}(\theta_i - \theta_{i0})(\theta_j - \theta_{0j})}_{\frac{1}{2}Q}$$

We will now start at some arbitrary point in parameter space θ_1 where the first derivative is no longer zero, i.e.

$$L = L(\theta_1) + \underbrace{\sum_i \nabla L(\theta - \theta_1)}_{\neq 0} + \underbrace{\frac{1}{2}\sum_{i,j}(\nabla\nabla L)_{ij}(\theta - \theta_1)(\theta - \theta_1)}_{\frac{1}{2}Q}$$

Iterative Linearization

We now develop a Taylor series for ∇L

$$\nabla L = \nabla L(\theta_1) + \nabla\nabla L(\theta_1)(\theta - \theta_1) + h.o.t.$$

At the optimum $\nabla L = 0$. We ignore the higher order terms and rearrange to find for the optimal parameter θ_0

$$\theta_0 \approx \theta_1 - [\nabla\nabla L(\theta_1)]^{-1}\nabla L(\theta_1)$$

The relationship will be exact if $\theta = \theta_0$, or if ∇L is truly linear.

Newton-Raphson Iterative Algorithm

Given a parameter (vector) θ and a likelihood function L:

Newton-Raphson Iterative Algorithm

Start with good estimate θ_1; evaluate the gradient vector ∇L at θ_1.

While $\nabla L > \epsilon$:

1. evaluate the second derivative matrix $\nabla\nabla L$ at $\theta = \theta_1$
2. calculate an improved estimate $\theta_2 = \theta_1 - [\nabla\nabla L(\theta_1)]^{-1}\nabla L(\theta_1)$
3. evaluate the gradient vector ∇L at θ_2

Comments

- If ∇L is linear only one iteration required, i.e. $\theta_2 = \theta_0$ independent of θ_1
- The algorithm will rapidly converge as long as θ_1 is reasonably close to θ_0

Example 1: Fitting a straight line

Suppose we have N data $\{Y_k\}$, with associated error-bars $\{\sigma_k\}$ at positions $\{x_k\}$ and we wish to estimate the best estimate of the slope of a straight line given the intercept. For a straight line the kth ideal datum is given by

$$y_k = mx_k + c \tag{4}$$

where m is the slope and c is the intercept. Substituting $F_k = y_k$, and $D_k = Y_k$ we obtain

$$\chi^2 = \sum_k R_k^2 = \sum_k \left(\frac{F_k - D_k}{\sigma_k} \right)^2 = \sum_k \left(\frac{mx_k + c - Y_k}{\sigma_k} \right)^2 . \tag{5}$$

Example 1: Fitting one parameter

In the first instance we assume that c is known and that we need to determine only a best estimate for m.

Given an initial estimate for $m = m_1$ we iterate according to

$$m_{n+1} = m_n - [\nabla\nabla L]^{-1}\nabla L(m_N). \tag{6}$$

where m_N is our estimate after $N - 1$ iterations.

Example 1: Fitting one parameter

$$\chi^2 = \sum_k R_k^2 = \sum_k \left(\frac{F_k - D_k}{\sigma_k} \right)^2 = \sum_k \left(\frac{mx_k + c - Y_k}{\sigma_k} \right)^2 \tag{7}$$

By differentiating χ^2 in Eq. 7 with respect to m we obtain

$$\nabla L(m_N) = -\frac{1}{2}\nabla\chi^2 = -\sum_k \frac{(m_N x_k + c - Y_k)x_k}{\sigma_k^2} \tag{8}$$

$$\nabla\nabla L = -\frac{1}{2}\nabla\nabla\chi^2 = -\sum_k \frac{x_k^2}{\sigma_k^2} \tag{9}$$

$\nabla\nabla L < 0$ so that we indeed obtain the maximal likelihood.

Example 1: Fitting two parameter of a straight line

In the next step we want to estimate both m and c. We use vector-matrix formulation and write

$$\begin{pmatrix} m_{N+1} \\ c_{N+1} \end{pmatrix} = \begin{pmatrix} m_N \\ c_N \end{pmatrix} - [\nabla\nabla L_N]^{-1}\nabla L_N. \tag{10}$$

where

$$\begin{aligned} \nabla L_N &= -\begin{pmatrix} \sum_k \frac{(m_N x_k + c_N - Y_k)x_k}{\sigma_k^2} \\ \sum_k \frac{m_N x_k + c_N - Y_k}{\sigma_k^2} \end{pmatrix}, \\ \nabla\nabla L &= -\begin{bmatrix} \sum_k \frac{x_k^2}{\sigma_k^2} & \sum_k \frac{x_k}{\sigma_k^2} \\ \sum_k \frac{x_k}{\sigma_k^2} & \sum_k \frac{1}{\sigma_k^2} \end{bmatrix}. \end{aligned} \tag{11}$$

Example: Fitting parameters in an ODE model

We now consider a simple ODE of the form

$$\frac{dy}{dt} = f(y) = -py \qquad y(0) = 1 \tag{12}$$

where we have time-dependent data Y_k with variance σ_k at the time points t_k.

We seek to estimate the decay rate p.

As before we write

$$L = const - \frac{\chi^2}{2} \qquad \text{with} \quad \chi^2 = \sum_k R_k^2 = \sum_k \left(\frac{y(t_k) - Y_k}{\sigma_k} \right)^2 . \tag{13}$$

Example: Fitting parameters in an ODE model

We now seek to maximise the likelihood,

$$L = const - \frac{\chi^2}{2} \tag{14}$$

which is equivalent to minimise

$$\chi^2 = \sum_k R_k^2 = \sum_k \left(\frac{y(t_k) - Y_k}{\sigma_k} \right)^2 . \tag{15}$$

As discussed in the previous lecture, we can use the **Newton-Raphson algorithm**: given an initial estimate for $p = p_1$ we iterate according to

$$p_{N+1} = p_N - [\nabla\nabla L]^{-1} \nabla L, \tag{16}$$

where the RHS is evaluated at p_N, the estimate after $N - 1$ iterations.

Example: Fitting parameters in an ODE model

By differentiating

$$\chi^2 = \sum_k R_k^2 = \sum_k \left(\frac{y(t_k) - Y_k}{\sigma_k} \right)^2 . \tag{17}$$

with respect to p we obtain

$$\nabla L(p_N) = -\frac{1}{2} \nabla \chi^2 = -\sum_k \frac{(y(t_k) - Y_k) \frac{\partial y(t_k)}{\partial p}}{\sigma_k^2} \tag{18}$$

$$\nabla \nabla L = -\frac{1}{2} \nabla \nabla \chi^2 = -\sum_k \frac{(\frac{\partial y(t_k)}{\partial p})^2}{\sigma_k^2} . \tag{19}$$

$\nabla \nabla L < 0$ so that we indeed obtain the maximal likelihood.

Example: Fitting parameters in an ODE model

$$\begin{aligned}\nabla L(p_N) &= -\frac{1}{2}\nabla\chi^2 = -\sum_k \frac{(y(t_k) - Y_k)\frac{\partial y(t_k)}{\partial p}}{\sigma_k^2} \\ \nabla\nabla L &= -\frac{1}{2}\nabla\nabla\chi^2 = -\sum_k \frac{(\frac{\partial y(t_k)}{\partial p})^2}{\sigma_k^2}.\end{aligned}$$

So how can we determine the sensitivities $S_k = \frac{\partial y(t_k)}{\partial p}$? We notice that

$$\frac{d\frac{dy}{dp}}{dt} = \frac{d\frac{dy}{dt}}{dp} = \frac{df(y)}{dp} = \frac{\partial f(y)}{\partial y}\frac{dy}{dp} + \frac{\partial f(y)}{\partial p} = J\frac{dy}{dp} + \frac{\partial f(y)}{\partial p} \quad (20)$$

Example: Fitting parameters in an ODE model

$$\begin{aligned}
\nabla L(p_N) &= -\frac{1}{2}\nabla\chi^2 = -\sum_k \frac{(y(t_k) - Y_k)\frac{\partial y(t_k)}{\partial p}}{\sigma_k^2} \\
\nabla\nabla L &= -\frac{1}{2}\nabla\nabla\chi^2 = -\sum_k \frac{(\frac{\partial y(t_k)}{\partial p})^2}{\sigma_k^2}. \\
\frac{d(\frac{dy}{dp})}{dt} &= J\frac{dy}{dp} + \frac{\partial f(y)}{\partial p}
\end{aligned}$$

The Jacobian J and $\frac{\partial f(y)}{\partial p}$ can be calculated ahead of time and the entire equation can then be integrated alongside the integration of the model ODE.

Example: Fitting parameters in an ODE model

$$\begin{aligned} \frac{dy}{dt} &= f(y) = -py; \qquad y(0) = 1 \\ \frac{dS}{dt} &= JS + \frac{\partial f(y)}{\partial p} = -pS - y; \qquad S(0) = 0 \end{aligned}$$

with Jacobian $J = \frac{\partial f(y)}{\partial y} = -p$, $\frac{\partial f(y)}{\partial p} = -y$, and sensitivity $S = \frac{\partial y}{\partial p}$.

$$\text{OPTIMIZATION:} \qquad p_{N+1} = p_N - [\nabla\nabla L]^{-1}\nabla L$$

$$\begin{aligned} \nabla L(p_N) &= -\frac{1}{2}\nabla\chi^2 = -\sum_k \frac{(y(t_k) - Y_k)\frac{\partial y(t_k)}{\partial p}}{\sigma_k^2} \\ \nabla\nabla L &= -\frac{1}{2}\nabla\nabla\chi^2 = -\sum_k \frac{(\frac{\partial y(t_k)}{\partial p})^2}{\sigma_k^2}. \end{aligned}$$

Example: Fitting 2 parameters in an ODE model

We now estimate two parameter values for a simple ODE of the form

$$\frac{dy}{dt} = f(y) = \rho - dy; \qquad y(0) = 10. \tag{21}$$

The parameter vector is then given by $\vec{p} = [\rho, d]^T$.

$$\text{OPTIMIZATION:} \qquad \vec{p}_{N+1} = \vec{p}_N - [\nabla\nabla L]^{-1}\nabla L$$

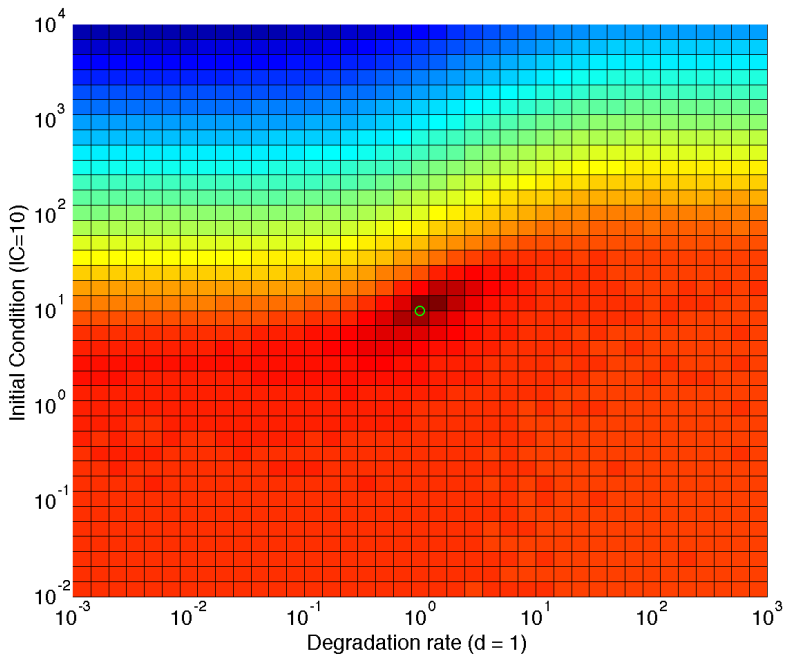
$$\nabla L = -\frac{1}{2}\nabla\chi^2 = -\sum_k \frac{(y(t_k) - Y_k)}{\sigma_k^2} \begin{pmatrix} \frac{\partial y(t_k)}{\partial \rho} \\ \frac{\partial y(t_k)}{\partial d} \end{pmatrix}$$

$$\nabla\nabla L = -\frac{1}{2}\nabla\nabla\chi^2 = -\begin{pmatrix} \sum_k \frac{1}{\sigma_k^2}\left(\frac{\partial y(t_k)}{\partial \rho}\right)^2 & \sum_k \frac{1}{\sigma_k^2}\frac{\partial y(t_k)}{\partial \rho}\frac{\partial y(t_k)}{\partial d} \\ \sum_k \frac{1}{\sigma_k^2}\frac{\partial y(t_k)}{\partial \rho}\frac{\partial y(t_k)}{\partial d} & \sum_k \frac{1}{\sigma_k^2}\left(\frac{\partial y(t_k)}{\partial d}\right)^2 \end{pmatrix}.$$

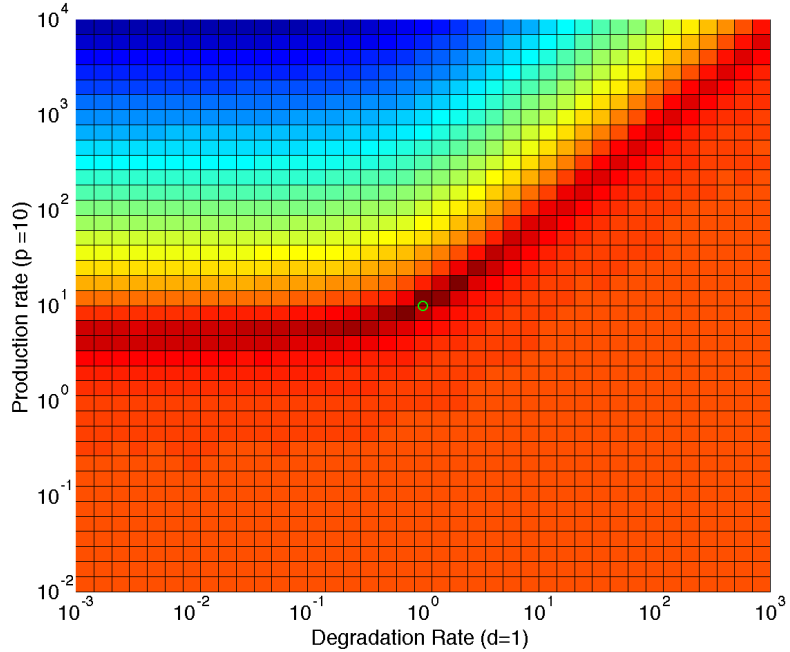
Example: Likelihood Function dependent on parameter values

$$\frac{dy}{dt} = f(y) = \rho - dy; \qquad y(0) = 10.$$

Optimizing the IC and d.



Optimizing the correlated production and decay rates.



Parameter Estimation for ODE Models

Consider a dynamical system with N state variables which we describe by a set of ordinary differential equations:

$$\frac{d\vec{x}(t)}{dt} = f(\vec{x}(t), t, \vec{k}), \qquad \vec{x}(t_0) = \vec{x_0}. \tag{22}$$

- $\vec{x}(t)$: vector with all state variables
- $\vec{k}$: vector with all parameters
- $\vec{x_0}$: vector for all initial expressions

Observables

Often, the state variables cannot be directly observed.

We specify an observation function $g : \mathbb{R}^N \rightarrow \mathbb{R}^M$ which maps the state variables $\vec{x}$ to a set of M observables,

$$\vec{y}(t) \quad = \quad g(\vec{x}(t), \vec{s}) \tag{23}$$

We require both $f(\cdot)$ and $g(\cdot)$ to be continuously differentiable functions with respect to their parameters.

Note that we may be able to only partially observe the system such that $M < N$.

The parameter vector $\vec{p}$ now comprises the kinetic parameters, $\vec{k}$, the initial conditions $\vec{x_0}$, and the parameters of the observation function, $\vec{s}$, such that

$$\vec{p} = \{\vec{x_0}, \vec{k}, \vec{s}\}. \tag{24}$$

Maximum Likelihood

The optimal parameter set is the one with the highest probability of observing the data and can be determined by maximizing the likelihood, $\mathrm{prob}(\vec{y}|\vec{p})$ of the data y_{ij} with respect to the parameter set p

$$\mathrm{prob}(\vec{y}|\vec{p}) = \prod_{i=1}^{T}\prod_{j=1}^{M}\frac{1}{\sigma_{ij}\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(g_j(\vec{x}(t_i,\vec{p}),\vec{p}))-y_{ij})^2}{\sigma_{ij}^2}\right).$$

Log-likelihood

In practical terms, to find the maximum of the likelihood function the negative log-likelihood, L, is minimized

$$L = -\log[\mathrm{prob}(\vec{y}|\vec{p})] = \sum_{i=1}^{T}\sum_{j=1}^{M}\frac{1}{2}R_{ij}(\vec{p})^2 + c_{ij},$$

$$R_{ij}(\vec{p}) = \frac{g_j(\vec{x}(t_i,\vec{p}),\vec{p})) - y_{ij}}{\sigma_{ij}}, \qquad c_{ij} = \log[\sigma_{ij}\sqrt{2\pi}].$$

R_{ij} is called **residual**. The term c_{ij} is independent of $\vec{p}$, and can be left out of the minimization.

The maximum likelihood estimator

The maximum likelihood estimator for the model parameters is thus given by

$$-\log[L(\vec{y}|\vec{p})] \quad \propto \quad \sum_{i=1}^{T}\sum_{j=1}^{M}\frac{1}{2}R_{ij}(\vec{p})^2. \tag{25}$$

In the background of independent Gaussian measurement errors the parameters $\vec{p}$ can therefore be determined by least squares minimization.

Optimization Algorithms

1. Local Methods
   - Gradient-based Methods
     - Newton-Raphson Iterative Algorithm
     - Levenberg-Marquard
   - Direct, derivative-free Methods
     - Simplex Methods
     - Nelder-Mead Method
     - Conjugate Gradient Method
2. Global Methods
   - Simulated Annealing
   - Evolutionary Algorithms

Calculation of the Sensitivities

The sensitivities can be computed by an integration of the sensitivity equations (as discussed above) in parallel with the ODE model.

$$\frac{dS_{p_l}^n}{dt} = \frac{d}{dt}\frac{dx_n}{dp_l} = \frac{d}{dp_l}\frac{dx_n}{dt} = \frac{df(t,\vec{x}(t),\vec{k})}{dp_l} = \sum_{q=1}^{N}\frac{\partial f_n}{\partial x_q}\frac{dx_q}{dp_l} + \frac{\partial f_n}{\partial p_l}$$

$$S_{p_l}^n(0) = \left(\frac{dx_q}{dp_l}\right)(0) = \begin{cases} 1 & : \quad p_l \in \{x_0\} \\ 0 & : \quad p_l \in \{s,k\} \end{cases}$$

Gradient of the weighted residuals R_{ij}

$$\begin{aligned}\frac{\partial R_{ij}(\vec{p})}{\partial p_l} &= -\frac{1}{\sigma_{ij}} \frac{dg_j(\vec{x}(t_i, p_l), p_l))}{dp_l} \\ &= -\frac{1}{\sigma_{ij}} \left(\sum_{n=1}^{N} \frac{\partial g_j}{\partial x_n}\bigg|_{t_i} \frac{dx_n}{dp_l}\bigg|_{t_i} + \frac{\partial g_j}{\partial p_l}\bigg|_{t_i} \right) \end{aligned} \tag{26}$$

$\frac{\partial g_j}{\partial x_n}$ and $\frac{\partial g_j}{\partial p_l}$ are the Jacobians of the differential equation system with respect to the state variables and with respect to the parameters.

$S^n_{p_l} = (dx_n)/(dp_l)$ are the sensitivities of the state variables to changes in the parameter values that we discussed above.

Workflow of gradient based minimization procedures

Initialize model system and parameters
LOOP
    Integrate ODE and sensitivity equations
    Calculate Jacobian of residuals
    Calculate residuals
    IF change in norm(residuals) $<$ threshold
        BREAK
    ELSE
        Update parameter values using current parameter values and Jacobian
    ENDIF
ENDLOOP
Calculate fit statistics, parameter variances and confidence limits

Structural Identifiability

Structural Identifiability

A model with M **state variables ($\vec{y}$)** and P **parameters ($\vec{p}$)** is structurally identifiable if its sensitivity matrix

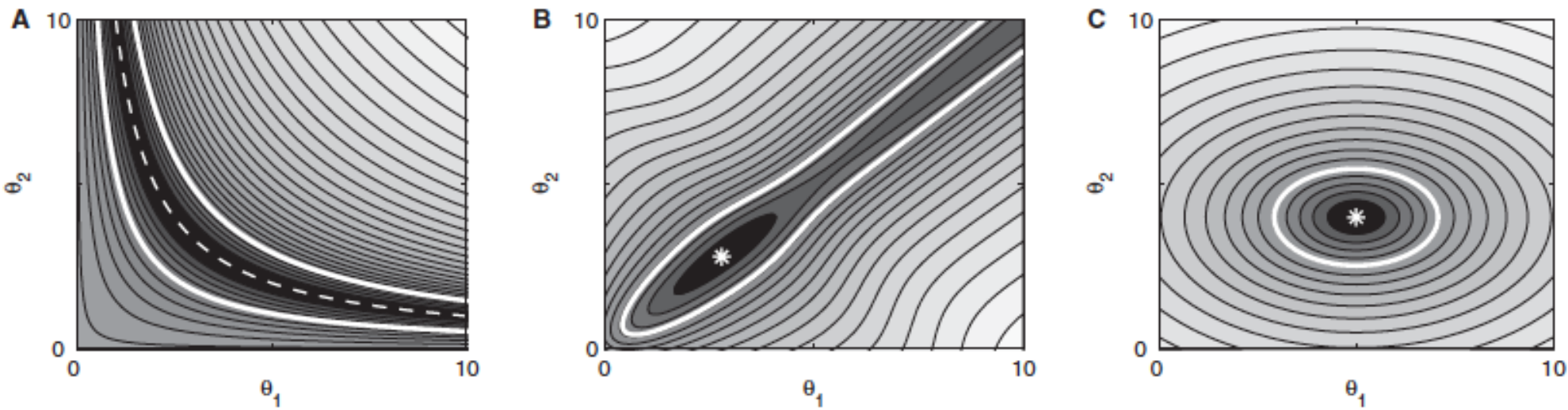
$$S_{p_j}^{m_i} = \frac{\partial y_i}{\partial p_j}, \quad i = 1, ..., M \quad j = 1, ..., P. \tag{27}$$

satisfies two conditions:

- each column has at least one large entry (i.e. each parameter has a large impact on at least one experimental measurement)
- the matrix has full rank (i.e. all columns must be linearly independent, which means that the effects of the parameters on the measurements must be independent of each other.)

Structural & Practical Identifiability

If parameters are correlated then only relative values can be determined for the correlated parameters since their effects compensate for each other.



Model Selection

The Problem

Mr A has a theory; Mr B also has a theory, but with an adjustable parameter λ.

Whose theory should we prefer on the basis of data D?

The Basic Solution

Evaluate the ratio of the posteriors:

$$\text{posterior ratio} = \frac{\text{prob}(A|D,I)}{\text{prob}(B|D,I)} \tag{28}$$

If it is greater than 1 we prefer theory A, and vice versa.

The Basic Solution

According to Bayes' Theorem

$$\underbrace{\frac{\text{prob}(A|D,I)}{\text{prob}(B|D,I)}}_{\text{Posterior odds}} = \underbrace{\frac{\text{prob}(D|A,I)}{\text{prob}(D|B,I)}}_{\text{Bayes Factor (BF)}} \times \underbrace{\frac{\text{prob}(A|I)}{\text{prob}(B|I)}}_{\text{Prior odds}}$$

The last term factors in what we thaught about the two theorists in the first place... Let's be fair and put it to one:

$$\frac{\text{prob}(A|D,I)}{\text{prob}(B|D,I)} = \frac{\text{prob}(D|A,I)}{\text{prob}(D|B,I)} \tag{29}$$

[A harsher treatment would be to factor in the track records of the two theorists...]

Bayes' Factor

$$\underbrace{\frac{\text{prob}(A|D,I)}{\text{prob}(B|D,I)}}_{\text{Posterior odds}} = \underbrace{\frac{\text{prob}(D|A,I)}{\text{prob}(D|B,I)}}_{\text{Bayes Factor (BF)}} \times \underbrace{\frac{\text{prob}(A|I)}{\text{prob}(B|I)}}_{\text{Prior odds}}$$

Rules of Thumb:

$0 \leq 2\ln(BF) \leq 2.2$	very weak evidence for A
$2.2 \leq 2\ln(BF) \leq 5$	weak to moderate evidence for A
$5 \leq 2\ln(BF) \leq 10$	moderate to strong evidence for A
$2\ln(BF) > 10$	decisive evidence for A

Approaches to determine Bayes' Factor

- AIC - Akaike information criterion
- BIC - Bayesian information criterion
- Markov Chain Monte Carlo (MCMC)

AIC & BIC

We now consider m models $M_1, \ldots, M_m$, where usually $m > 2$.

Definitions:
ML_i: maximum likelihood over the ith model
$MLL_i = \ln(MLi)$: the maximum log likelihood over the ith model
d_i: dimension of the ith model M_i, i.e. number of free parameters
n: number of fitted data points
Different penalties have been proposed to be subtracted from MLL_i to avoid overfitting.

$$AIC_i = MLL_i - d_i \tag{30}$$

$$BIC_i = MLL_i - d_i \ln n \tag{31}$$

For either AIC or BIC, one would select the model with the largest value of the criterion.

Thanks!!

Thanks for your attention!

Slides for this talk will be available at:
`http://www.bsse.ethz.ch/cobi/education`