Examples of inverse problems and data fitting

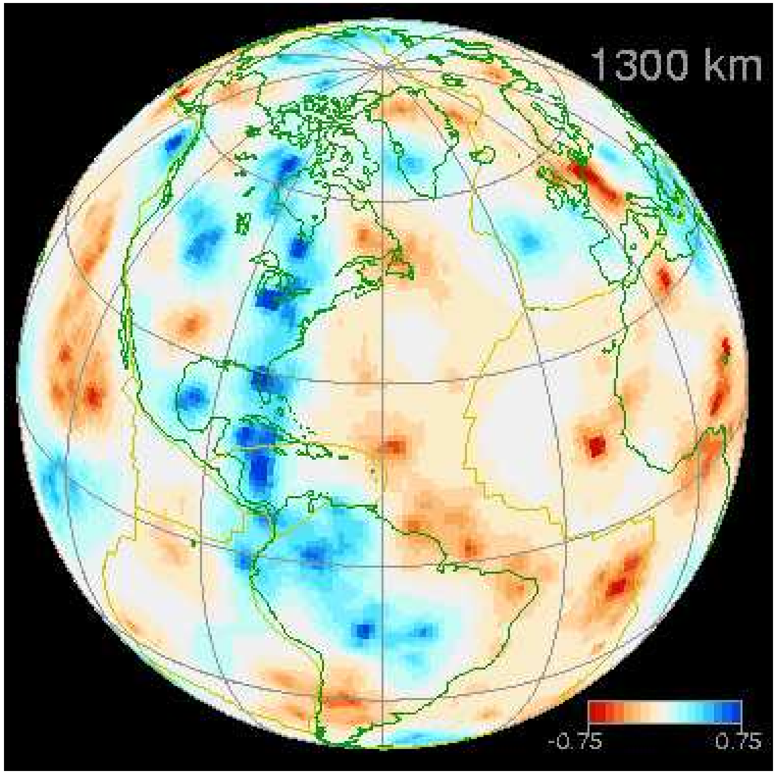
Malcolm Sambridge

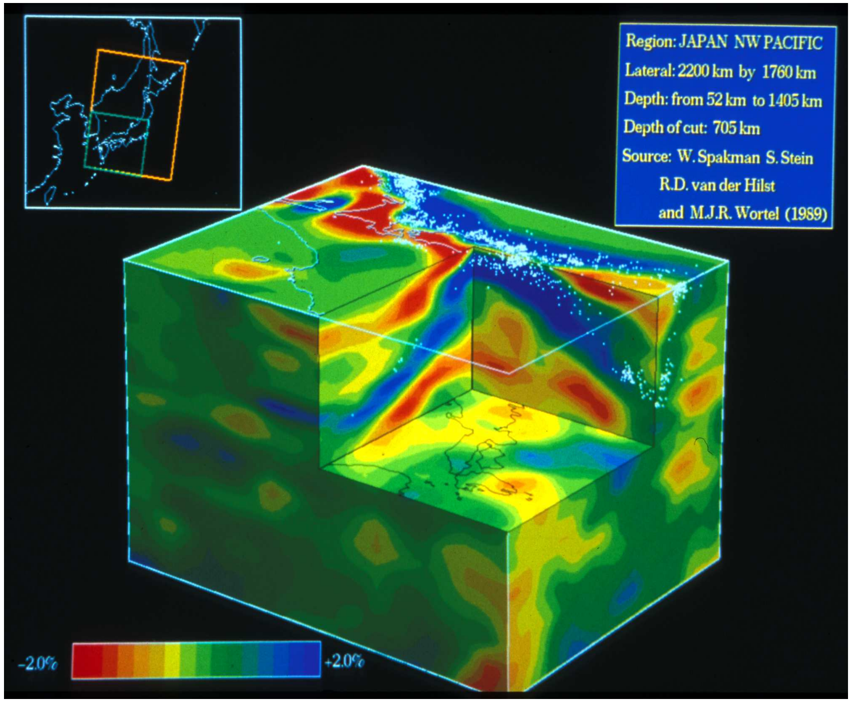
Research School of Earth Sciences

Australian National University

A complement to Inversion Tutorial slides

Constraining Earth's interior from the surface





A general nonlinear inverse problem

Nonlinear inverse problem

$$\boldsymbol{d} = g(\boldsymbol{m})$$

where $\boldsymbol{d}$ is the data vector and $\boldsymbol{m}$ is the model vector.

Choose a starting (or best guess model $\boldsymbol{m}_o$) and linearize about it,

$$\delta \boldsymbol{d} = G \delta \boldsymbol{m}$$

But G is not a square matrix. We could solve by minimizing,

$$\phi = (\delta \boldsymbol{d} - G\delta \boldsymbol{m})^T C_D^{-1} (\delta \boldsymbol{d} - G\delta \boldsymbol{m})$$

Where C_D^{-1} is a data covariance matrix.

A least squares solution

From

$$\delta \boldsymbol{d} = G\delta \boldsymbol{m}$$

we find $\delta \boldsymbol{m}$ which minimizes ϕ, $\ldots$ and get the normal equations

$$\delta \boldsymbol{m} = (G^T C_D^{-1} G)^{-1} G^T C_D^{-1} \delta \boldsymbol{d}$$

We introduce the generalized inverse as

$$\delta \boldsymbol{m} = G^{-g} \delta \boldsymbol{d}$$

Note that if data covariance matrix has the form

$$C_D^{-1} = \sigma^{-2} I$$

the estimated model is independent of the data errors !

Example

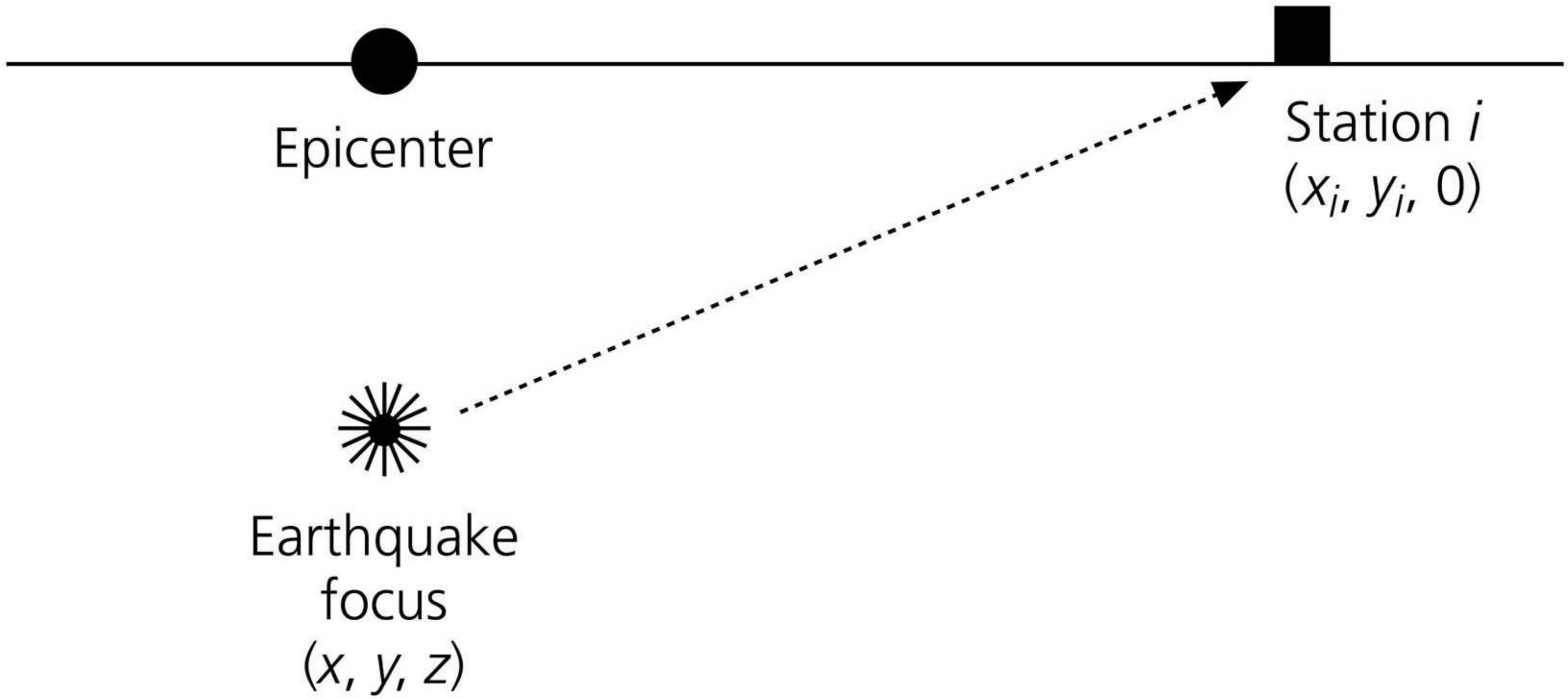
Data fitting in a discrete,

linearized over-determined problem.

$$\delta \boldsymbol{d} = G \delta \boldsymbol{m}$$

Earthquake location

Figure 7.2-1: Geometry for earthquake location in a homogeneous halfspace.



What are $\delta \boldsymbol{d}$, $\delta \boldsymbol{m}$ and G ?

Earthquake location example

Inversion for earthquake location and origin time (error free)

Parameter	True value	Solution at each iteration		
		0	1	2
X	0.0	3.0	-0.5	0.0
Y	0.0	4.0	-0.6	0.0
Z	10.0	20.0	10.1	10.0
T_0	0.0	2.0	0.2	0.0

Station	Arrival time residual		
1	-2.1	-0.4	0.0
2	-3.0	-0.2	0.0
3	-3.8	-0.1	0.0
4	-3.0	-0.2	0.0
5	-2.6	-0.3	0.0
6	-2.0	-0.3	0.0
7	-2.9	-0.2	0.0
8	-3.7	-0.2	0.0
9	-4.1	-0.2	0.0
10	-2.4	-0.4	0.0
Misfit	92.4	0.6	0.0

$$\delta \boldsymbol{m} = (G^T C_D^{-1} G)^{-1} G^T C_D^{-1} \delta \boldsymbol{d}$$

Propagating errors from data to model

Each set of observations $\boldsymbol{d}$ is only one realization of many possible that could have been observed,

$$\boldsymbol{d}^{(i)} \quad (i = 1, \ldots, K) \quad K \to \infty$$

The generalized inverse gives us an estimated model, $\boldsymbol{m}^{(i)}$ from each $\boldsymbol{d}^{(i)}$

$$\delta \boldsymbol{m}^{(i)} = G^{-g} \delta \boldsymbol{d}^{(i)}$$

This leads to the model covariance matrix

$$C_M = G^{-g} C_D (G^{-g})^T$$

$$\Rightarrow C_M = (G^T C_D^{-1} G)^{-1} \quad \text{(Least squares)}$$

$$\text{If} \quad C_D^{-1} = \sigma^{-2} I \quad \to \boxed{C_M = \sigma^2 (G^T G)^{-1}}$$

Earthquake location with noise

Inversion for earthquake location and origin time (σ = 0.1 s)

Parameter	True value	Solution at each iteration			
		0	1	2	3
X	0.0	3.0	-0.2	0.2	0.2
Y	0.0	4.0	-0.9	-0.4	-0.4
Z	10.0	20.0	12.2	12.2	12.2
T_0	0.0	2.0	0.0	-0.2	-0.2

X	Y	Z	T_0
0.06	0.01	0.01	0.00
0.01	0.08	-0.13	0.01
0.01	-0.13	1.16	-0.08
0.00	0.01	-0.08	0.01
0.25	0.28	1.08	0.10

Station	Arrival time residual			
1	-2.0	-0.1	0.1	0.1
2	-3.0	-0.1	0.0	0.0
3	-3.8	0.0	0.1	0.1
4	-3.2	-0.1	0.0	0.0
5	-2.8	-0.2	-0.1	-0.1
6	-2.1	-0.3	-0.1	-0.1
7	-2.9	-0.1	0.0	0.0
8	-3.7	-0.1	0.0	0.0
9	-4.0	-0.1	0.0	0.0
10	-2.5	-0.3	0.0	0.0
Misfit	93.74	0.33	0.04	0.04

$$\delta \boldsymbol{m} = (G^T C_D^{-1} G)^{-1} G^T C_D^{-1} \delta \boldsymbol{d}$$

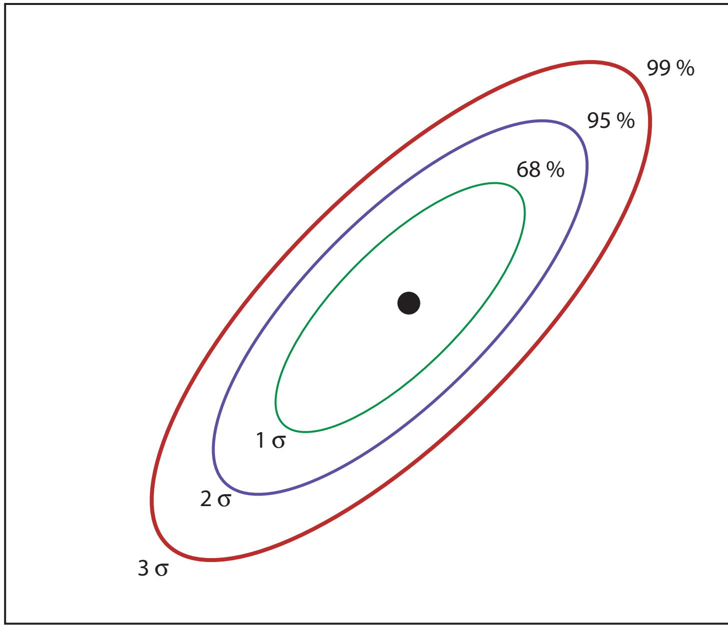
Confidence regions about the best fit model

How do we find confidence regions about best model, $\boldsymbol{m}^*$?
We could map out the data misfit function $\phi(\boldsymbol{m})$,

$$\phi(\boldsymbol{m}) = (\boldsymbol{d} - g(\boldsymbol{m}))^T C_D^{-1} (\boldsymbol{d} - g(\boldsymbol{m}))$$

It can be shown that for a linearized problem the confidence contours are quadratic and given by

$$\delta\phi(\boldsymbol{m}) = \delta\boldsymbol{m}^T C_M^{-1} \delta\boldsymbol{m}$$



Size and shape of the confidence regions determined by the inverse model covariance C_M^{-1}.

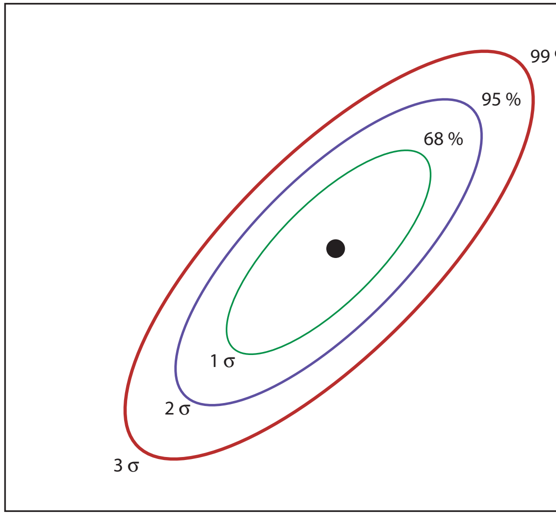
Confidence contours and goodness of fit

The confidence probability assigned to each contour and the $\phi(\boldsymbol{m}^*)$ is made with χ^2 statistics.

$$\chi^2(\boldsymbol{m}) = \sum_{i=1}^{N} \frac{(d_i - g_i(\boldsymbol{m}))^2}{\sigma_i^2}$$

Expected misfit for the best model $\boldsymbol{m}^*$,

$$\chi^2(\boldsymbol{m}^*) = ndf = N - k$$



Use statstical tables for a χ^2 distribution with $(N-k)$ degrees of freedom.
What if we don't know data errors ?

$$\sigma^2 = \frac{1}{N-k}\phi(\boldsymbol{m}^*)$$

Goodness of fit

ndf	$\chi^2(95\%)$	$\chi^2(50\%)$	$\chi^2(5\%)$
5	1.15	4.35	11.07
10	3.94	9.34	18.31
20	10.85	19.34	31.41
50	34.76	49.33	67.50
100	77.93	99.33	124.34

Percentage points of the χ^2 distribution.

- What happens if the χ^2 value is too small or too large ?

Goodness of fit: comparing two solutions

What if we have two solutions $\boldsymbol{m}_1^*$ and $\boldsymbol{m}_2^*$ with different numbers of unknowns, M_1 and M_2, and the second model fits the data better than the first.

$$\chi^2_{\nu_1} > \chi^2_{\nu_2}$$

where $\nu_1 = N - M_1$ and $\nu_2 = N - M_2$.

How can we tell if the improvement in data fit is significant ?

The F-ratio test can be performed,

$$F = \frac{\chi^2_{\nu_1}}{\chi^2_{\nu_2}}$$

Statistical tables give the probability distribution $P(F)$, i.e. that ratios greater than or equal to F occur 5% of the time.

Model resolution matrix

If we obtain a solution to a linearized inverse problem,

$$\delta \boldsymbol{m} = G^{-g} \delta \boldsymbol{d}$$

Then we have

$$\delta \boldsymbol{m} = G^{-g} G \delta \boldsymbol{m}_{true} = R \delta \boldsymbol{m}_{true}$$

This defines the model resolution matrix, R. For an over-determined problem we get

$$R = \left[(G^T C_D^{-1} G)^{-1} G^T C_D^{-1} \right] G = I$$

R measures the degree of blurring and does not depend on the errors in the data !

Example

Data fitting in a discrete, linearized under and over determined problem.

$$\delta \boldsymbol{d} = G \delta \boldsymbol{m}$$

Travel time tomography

Travel time equation

$$t = \int_{R_o} \frac{1}{v(\boldsymbol{x})} dl = \int_{R_o} s(\boldsymbol{x}) dl$$

If we choose a reference slowness field $s_o(\boldsymbol{x})$ and linearize the relationship about it, we get

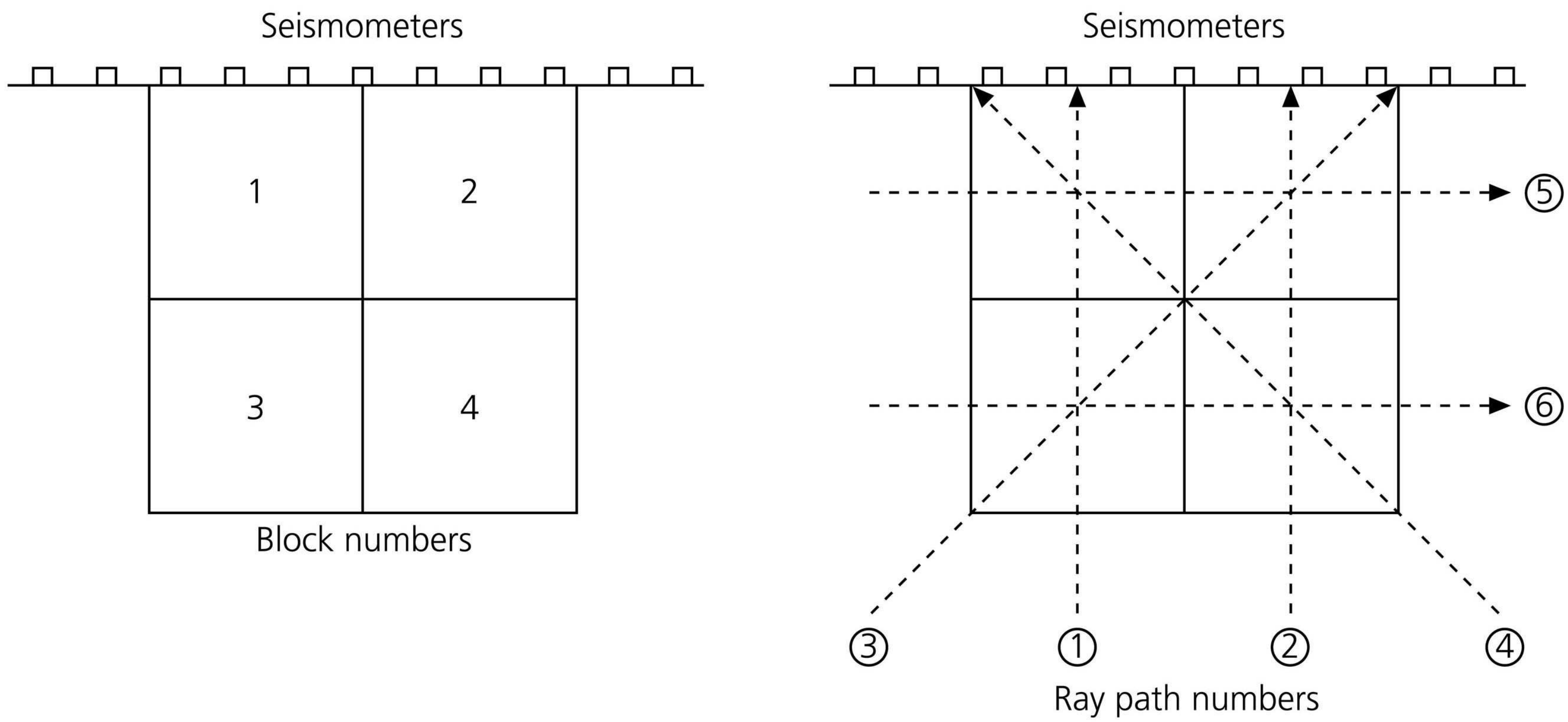
$$\delta t = \int_{R_o} \delta s(\boldsymbol{x}) dl$$

The basis of all travel time tomography.
Discretization: Choose a set of basis functions

$$\delta s(\boldsymbol{x}) = \sum_{j=1}^{M} m_j \phi_j(\boldsymbol{x}) \quad \Rightarrow \quad \boxed{\delta \boldsymbol{d} = G \delta \boldsymbol{m}}$$

Idealized tomographic experiment

Figure 7.3-2: Ray path and block geometry for an idealized tomographic experiment.



Idealized tomographic experiment

Using rays $1 \longrightarrow 4$ we get

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & \sqrt{2} & \sqrt{2} & 0 \\ \sqrt{2} & 0 & 0 & \sqrt{2} \end{pmatrix} \begin{pmatrix} \delta m_1 \\ \delta m_2 \\ \delta m_3 \\ \delta m_4 \end{pmatrix} = \begin{pmatrix} \delta d_1 \\ \delta d_2 \\ \delta d_3 \\ \delta d_4 \end{pmatrix}$$

which gives

$$G^T G = \begin{pmatrix} 3 & 0 & 1 & 2 \\ 0 & 3 & 2 & 1 \\ 1 & 2 & 3 & 0 \\ 2 & 1 & 0 & 3 \end{pmatrix}$$

which has eigenvalues 0,2,4,6 and hence is singular !

$$\delta \boldsymbol{m} = (G^T C_D^{-1} G)^{-1} G^T C_D^{-1} \delta \boldsymbol{d}$$

Singular value decomposition

We can write $G^T G$ in terms of eigenvectors and eigenvalues

$$G^T G = V \Lambda V^T \quad V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p : \boldsymbol{v}_{p+1}, \ldots, \boldsymbol{v}_r)$$
$$G G^T = U \Lambda U^T \quad U = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q : \boldsymbol{u}_{q+1}, \ldots, \boldsymbol{u}_d)$$

where $\Lambda = diag(\lambda_1, \ldots, \lambda_p, 0, \ldots)$. This gives the Lanczos decomposition of the generalized inverse

$$G^{-p} = V_p \Lambda_p^{-1} U_p^T$$

$$R = G^{-p} G = (V_p \Lambda_p^{-1} U_p^T)(U_p \Lambda_p V_p^T) = V_p V_p^T$$

Inadequate ray resolution causes blurring !

Null space = resolution blurring

The resolution matrix becomes

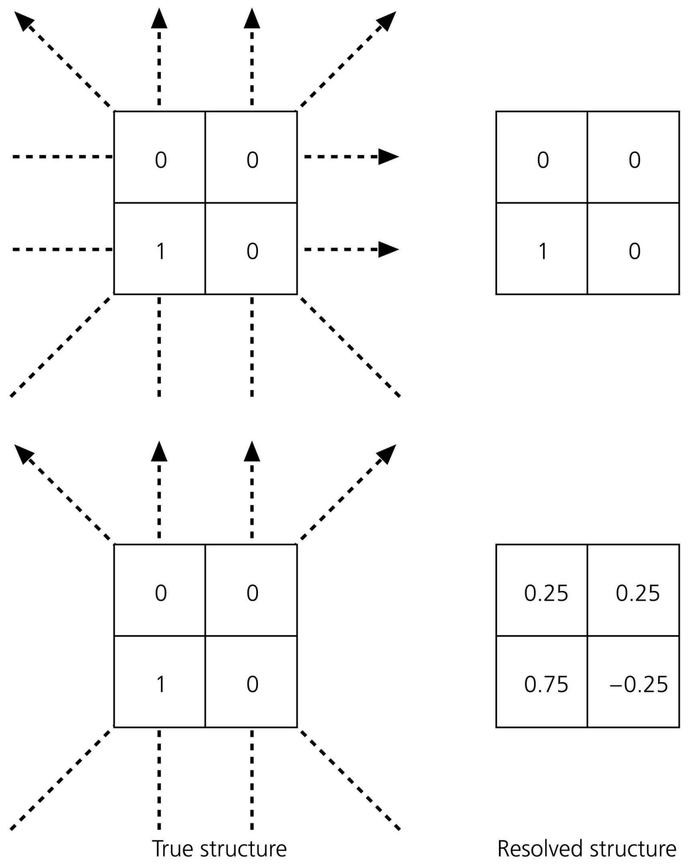
$$\delta \boldsymbol{m} = \begin{pmatrix} 0.75 & -0.25 & 0.25 & 0.25 \\ -0.25 & 0.75 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.75 & -0.25 \\ 0.25 & 0.25 & -0.25 & 0.75 \end{pmatrix} \delta \boldsymbol{m}_{true}$$

The ray distribution cannot resolve equal slowness perturbations in blocks 1 and 2, with opposite perturbations in 3 and 4.

The zero eigenvalues create a null space !

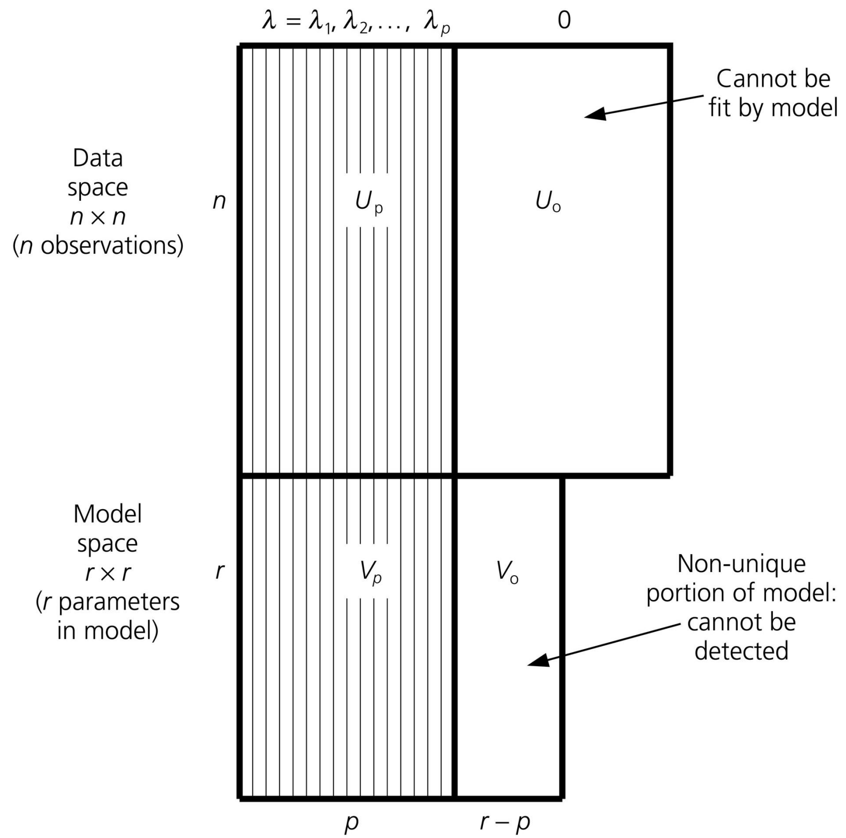
Tomography blurring

Figure 7.3-3: Illustration of "blurring" due to incomplete ray coverage.



Inversion: null spaces

Figure 7.3-4: Illustration of the relation between the data and model spaces.



Features of inverse problems

- Linearization is an approximation
- Parametrization is a choice
- Unknowns of different types (e.g. velocity and hypocentres)
- Nonuniqueness can occur
  - over determined
  - even determined
  - under determined
- More data reduces input noise but independent data matters most.
- Trade-off between model variance and resolution (spread)

Underdetermined inversion: Regularization

When the problem is under or mixed-determined we can minimize a combination of data fit and model control.

$$\Psi(\boldsymbol{m}) = \phi(\boldsymbol{d}, \boldsymbol{m}) + \lambda^2 \psi(\boldsymbol{m}) \tag{1}$$

λ is a trade-off parameter that must be chosen. It adds stability but decreases resolution.
If the regularization is chosen $\psi(\boldsymbol{m}) = \delta \boldsymbol{m}^T \delta \boldsymbol{m}$

$$\Rightarrow G^{-g} = (G^T C_D^{-1} G + \lambda^2 I)^{-1} G^T C_D^{-1}$$

This gives a minimum variance solution. The poorly constrained parts of the model are damped towards the reference model.

Distrust struture on the scale length of the blocks !

Underdetermined inversion: Regularization

An alternative is a Laplacian operator

$$\psi(\boldsymbol{m}) = |||L\boldsymbol{m}||^2 = \boldsymbol{m}^T L^T L\boldsymbol{m}$$

L is a finite difference approximation to ∇^2.

Model roughness (or flatness) is minimized.

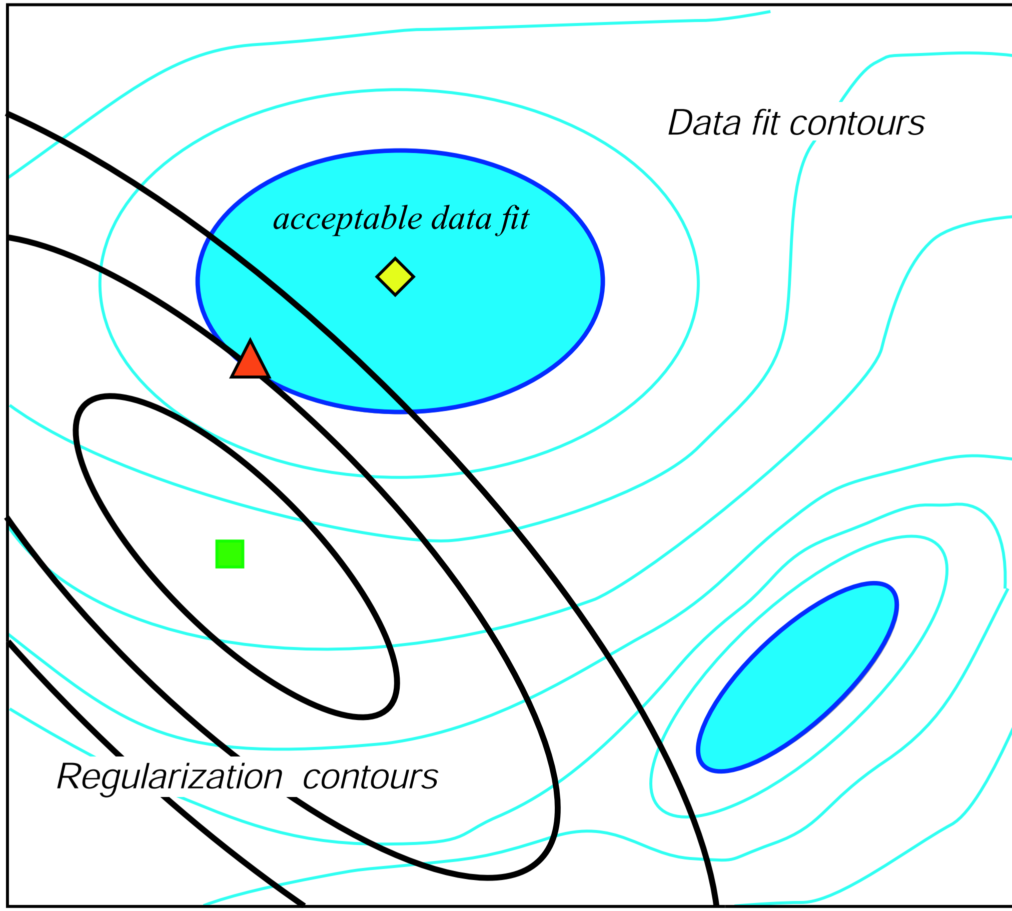
- Resulting models will be smooth but not of minimum variance
- Blocks not sampled will be smoothed.

Distrust large amplitude anomalies in areas with few data !

For large numbers of unknowns ($> 10^4$) iterative methods are needed to solve the resulting system of equations, e.g. conjugate gradients. $\rightarrow$ High performance computation.

Solutions to inverse problems

- *Optimal data fit solution (c.f. MAP)*
- *Extremal solution*
- *Data acceptable solutions*

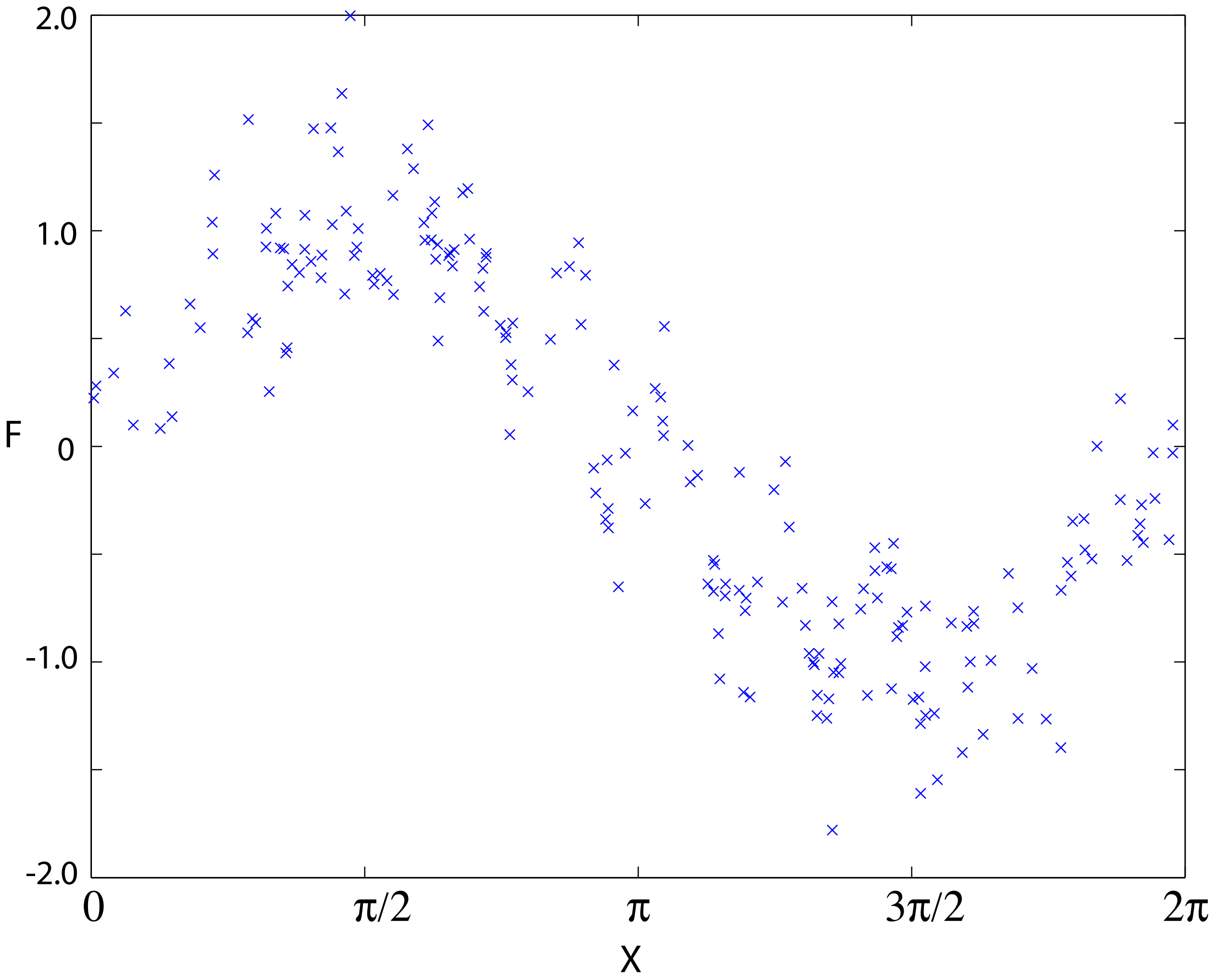


Example

Fitting data and smoothing models

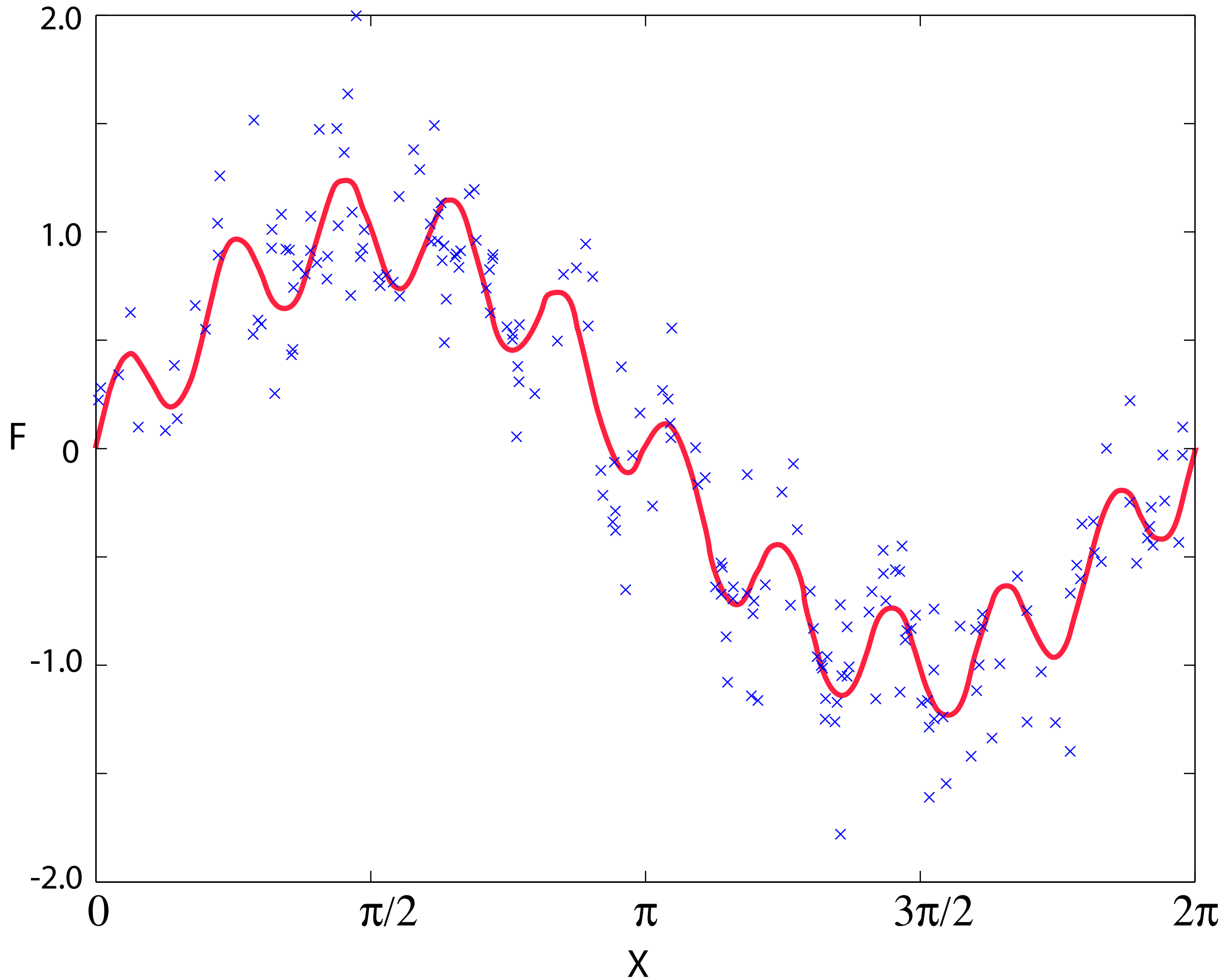
$$\Psi(\boldsymbol{m}) = \phi(\boldsymbol{d}, \boldsymbol{m}) + \lambda^2 \psi(\boldsymbol{m})$$

Example: smoothing data



We want to fit the data and find the curve which generated it.

Example: smoothing data



We want to fit the data and find the curve which generated it.

This is the curve that generated the data

Constructing smooth models - theory

Typically we would want to fit the data and regularize or smooth the model at the same time.

$$\psi(\boldsymbol{d},\boldsymbol{m}) = \sum_{i=1}^{N} (d_i - s(\boldsymbol{x}_i,\boldsymbol{m}))^2 + \mu J(s)$$

Where,

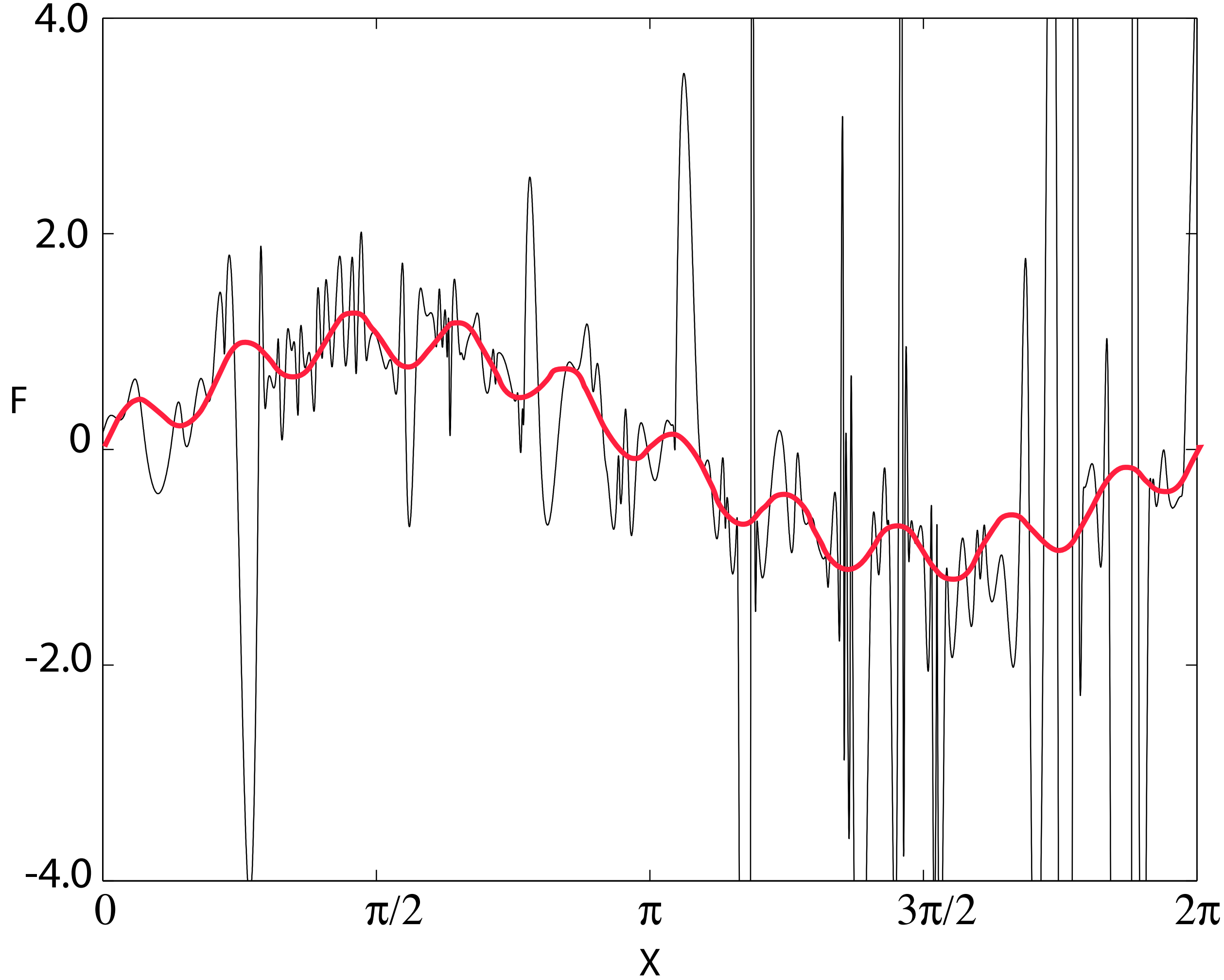
$$J(s) = \int \left[\left(\frac{\partial^2 s}{\partial x^2}^{2} \right) + 2 \left(\frac{\partial^2 s}{\partial x \partial y}^{2} \right) + \left(\frac{\partial^2 s}{\partial y^2}^{2} \right) \right] d\boldsymbol{x}$$

Can we find a smooth model that fits the data exactly ?

$$s(\boldsymbol{x},\boldsymbol{m}) = p(\boldsymbol{x}) + \sum_{i=1}^{N} \lambda_i \phi(\boldsymbol{x} - \boldsymbol{x}_i)$$

Yes ! use *Thin Plate Splines* for $\phi(\boldsymbol{x})$ (Duchon, 1976)

Smooth models - practice

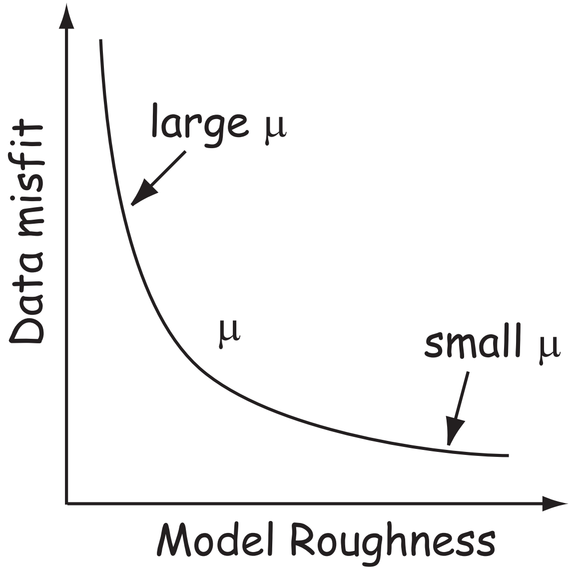


Relaxing the fit to data

We do not want to fit noisy data exactly !

$$\psi(\boldsymbol{d}, \boldsymbol{m}) = \sum_{i=1}^{N} (d_i - s(\boldsymbol{x}_i, \boldsymbol{m}))^2 + \mu J(s)$$

In order to relax the requirement to fit the data we must find a value of the trade-off parameter μ.



Choosing trade-off parameter

One way of finding a balance between data fit and model smoothness is Generalized Cross Validation - which essentially means use the data to find a value for μ.

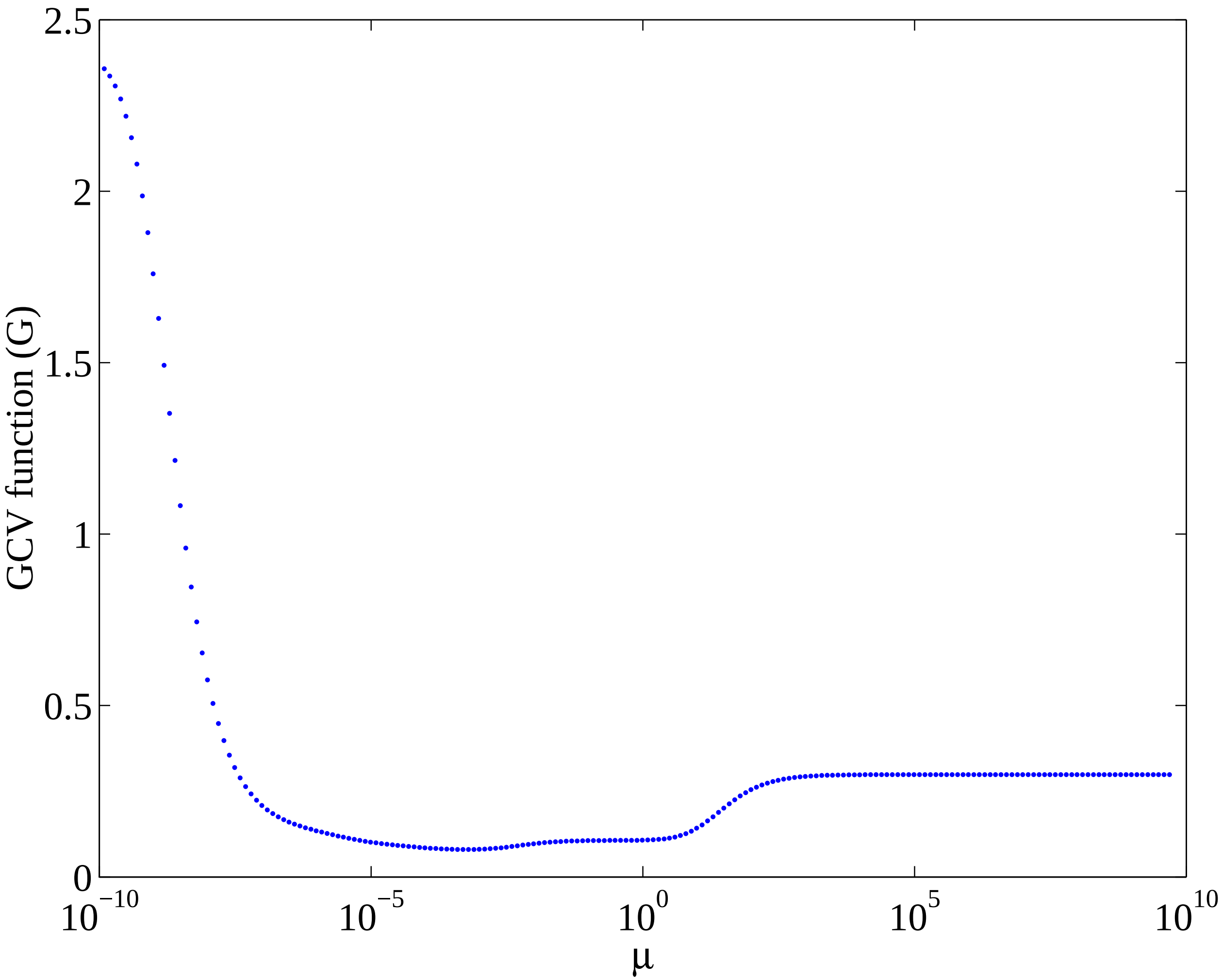
$$G(\mu) = \sum_{i=1}^{N} (d_i - s_i(\boldsymbol{x}_i, \boldsymbol{m}))^2$$

Where $s_i(\boldsymbol{x}, \boldsymbol{m})$ is the TPS interpolant produced when the ith datum is removed. Find μ that minimizes $G(\mu)$. Note

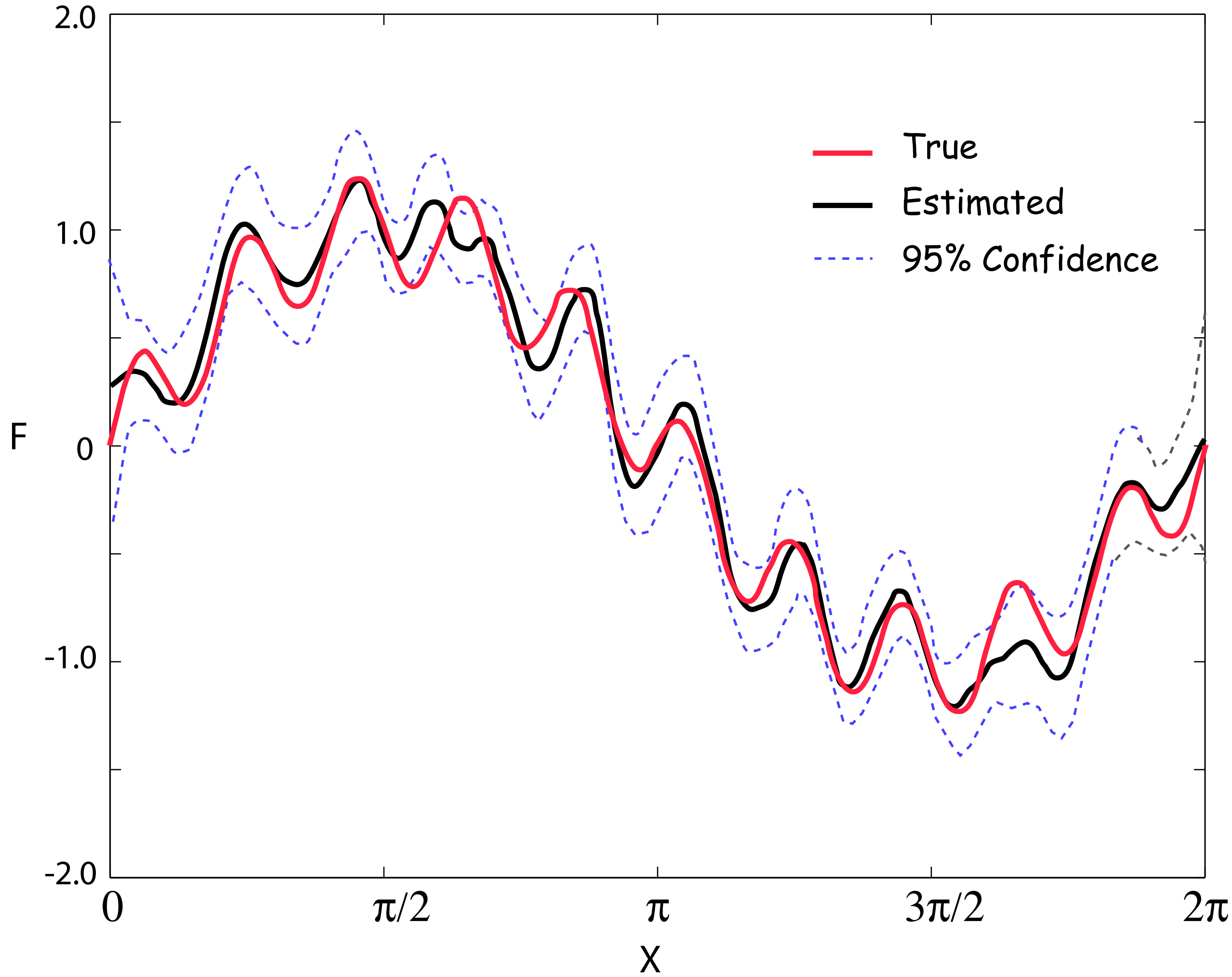
$$\mu \rightarrow \infty \Rightarrow G(\mu) \uparrow$$
$$\mu \rightarrow 0 \Rightarrow G(\mu) \uparrow$$

$G(\mu)$ is a bootstrap measure of interpolation error.

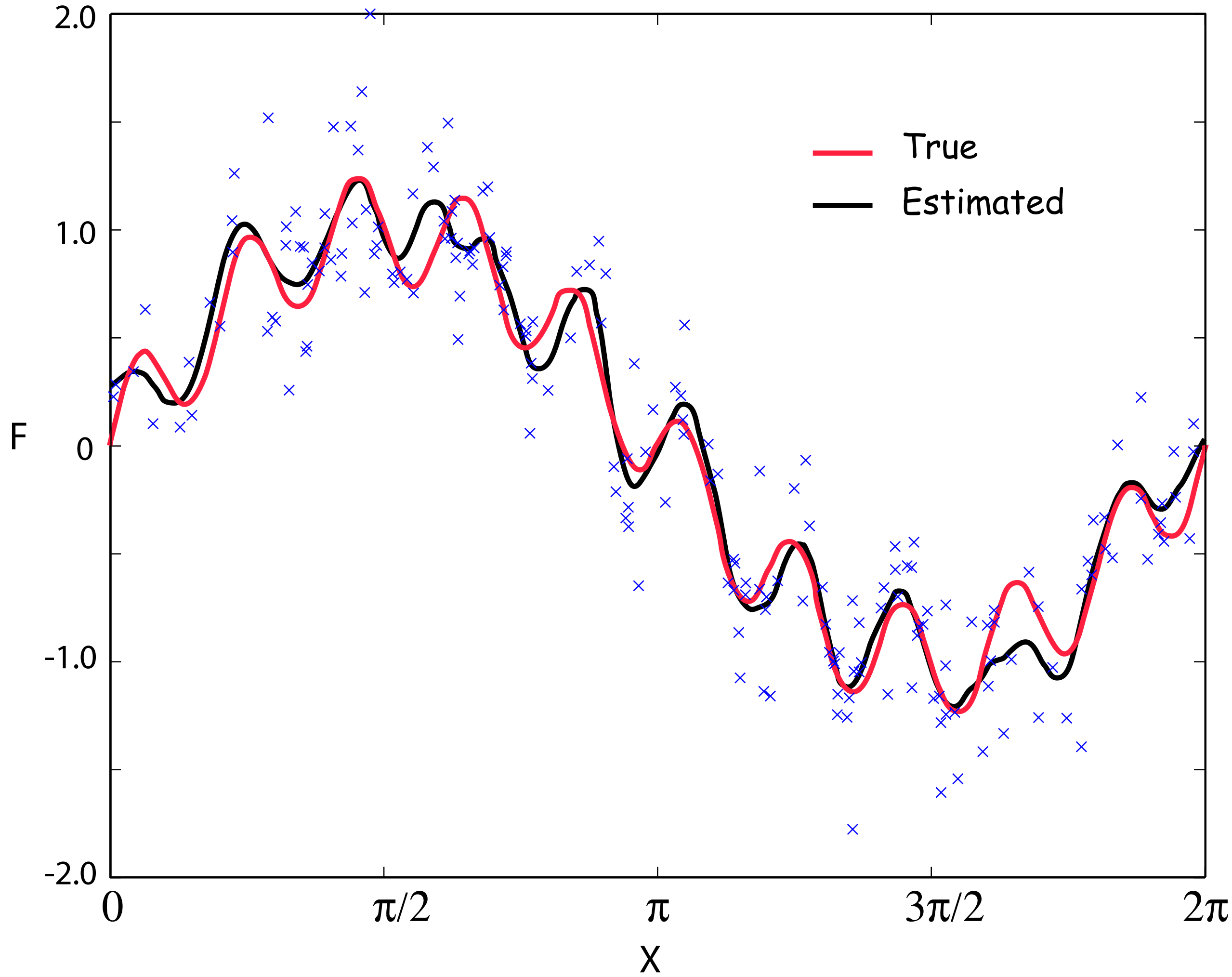
Minimizing GCV to find μ



Generalized cross validation



Generalized cross validation

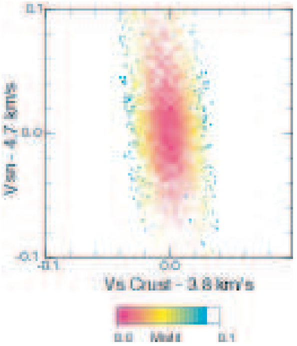


Bayesian inference
model comparison

Bayesian inference

Bayesian inference can be applied to:

- The model inference problem
  Estimating the unknowns



- The model comparision problem
  Hypothesis testing
  When the number of unknowns is one of your unknowns !

Bayesian inference

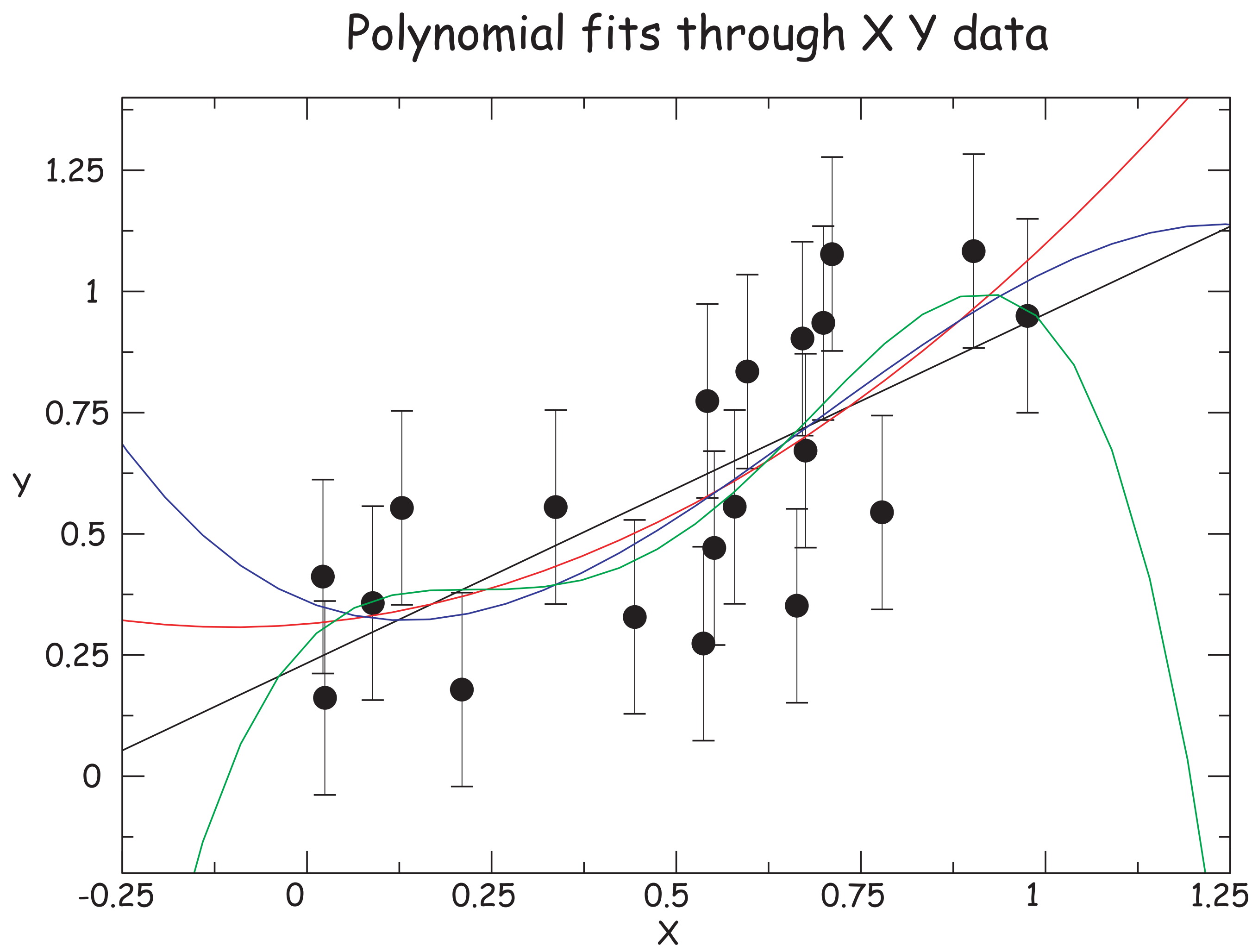
$$\frac{p(\mathcal{H}_1|\boldsymbol{d})}{p(\mathcal{H}_2|\boldsymbol{d})} = \frac{p(\boldsymbol{d}|\mathcal{H}_1)\; p(\mathcal{H}_1)}{p(\boldsymbol{d}|\mathcal{H}_2)\; p(\mathcal{H}_2)}$$

$$\text{Posterior} = \text{Likelihood} \; \times \; \text{Prior}$$

where

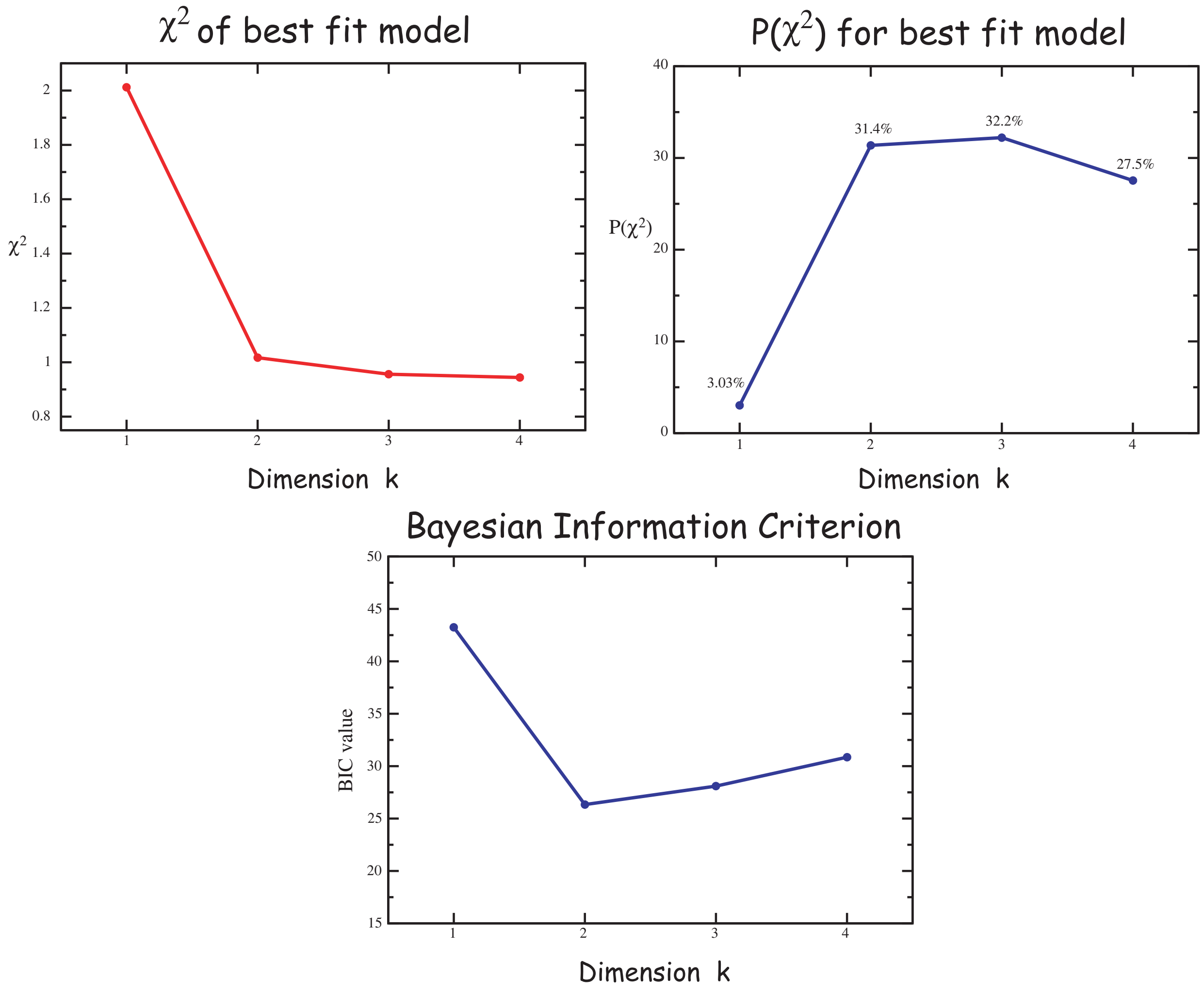
$$\mathcal{H}_1 = \text{Hypothesis1}$$
$$\mathcal{H}_2 = \text{Hypothesis2}$$

A regression example



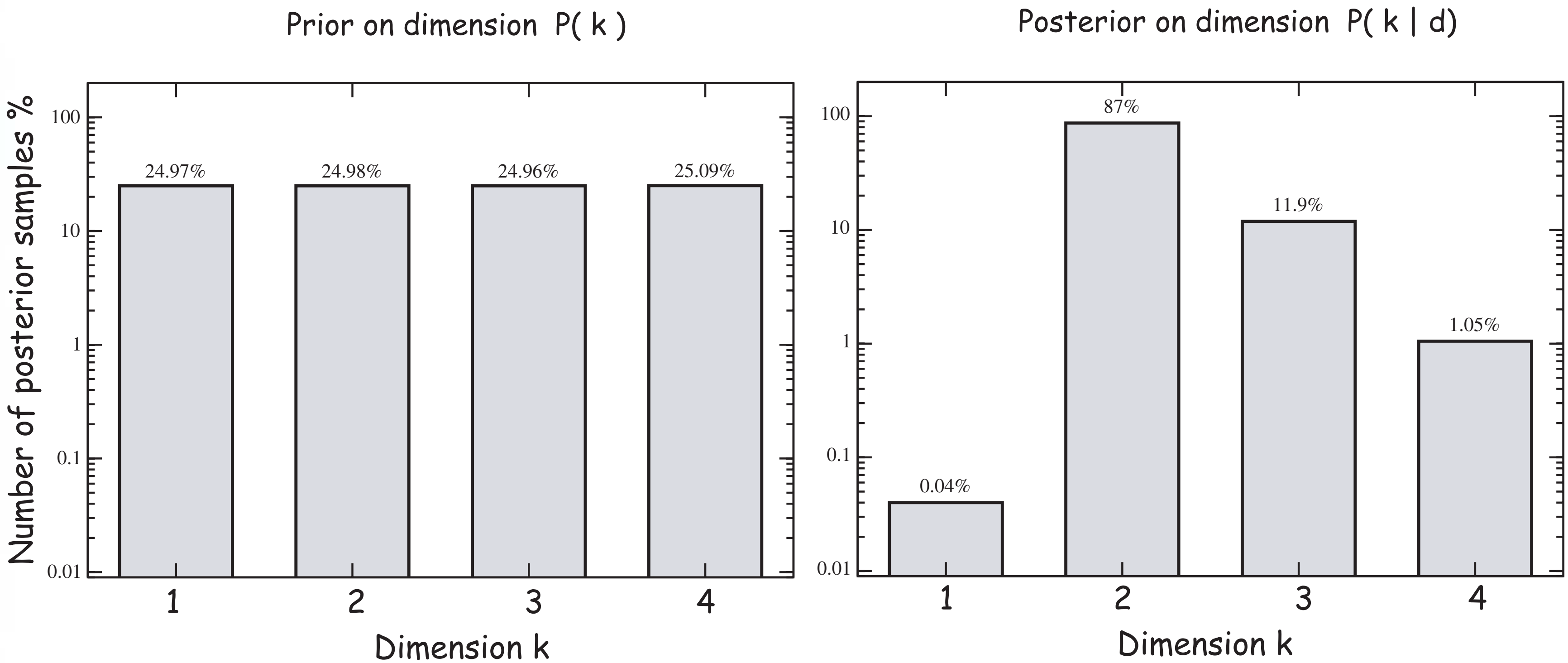
Lets fit the data with a polynomial, $y = a_0 + \sum_{i=1}^{k-1} a_i x^i$ and let k be one of the unknowns !

Best fit solutions



Statistical measures of significance of fit.

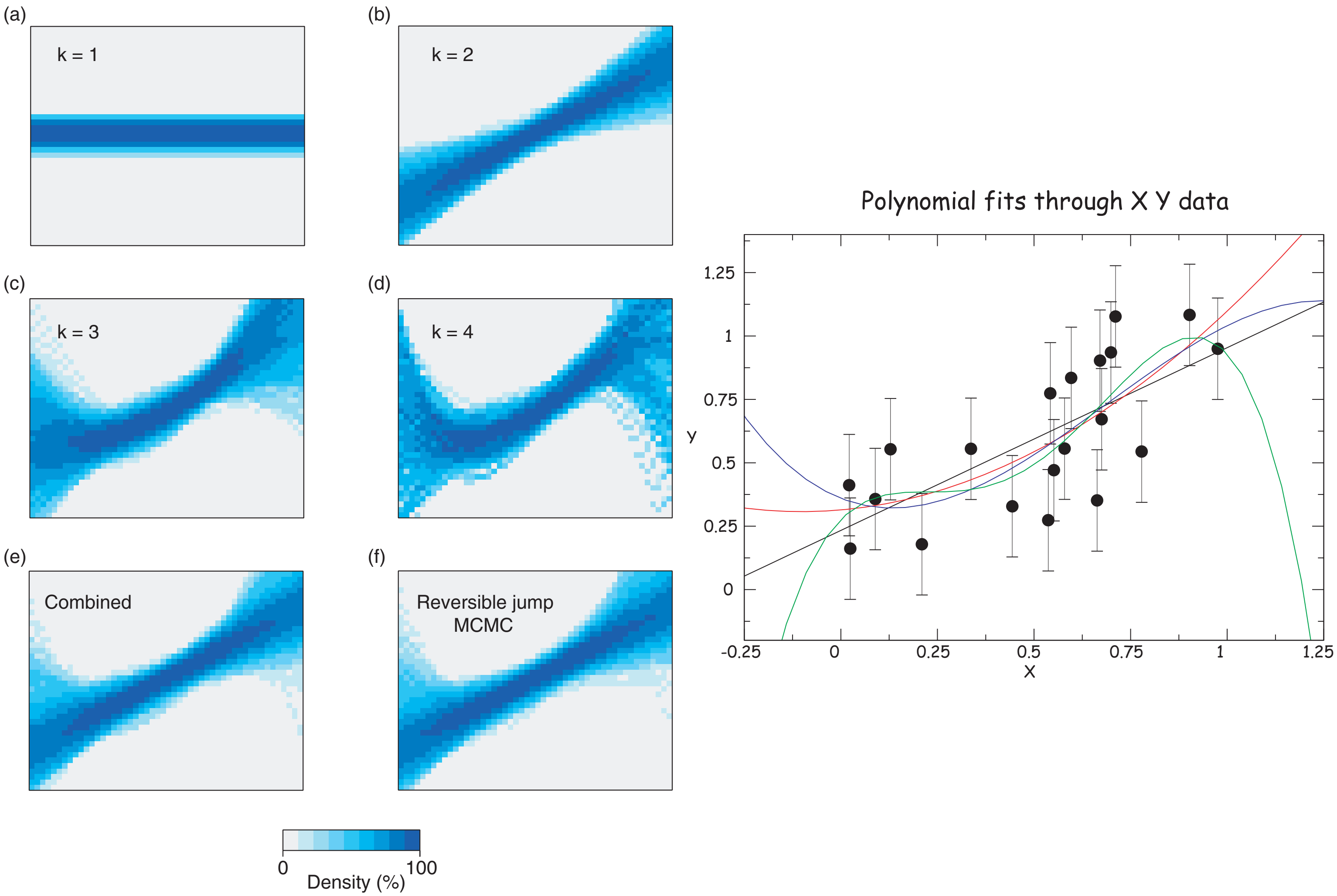
The number of unknowns



Bayesian Inference is parsimonious !

Occams razor is incorporated naturally

Posterior predictions



Samples produced by MCMC and the original data.