# Inverse problems in Finance and Human sciences

Ali Mohammad-Djafari
Laboratoire des Signaux et Systèmes (L2S)
UMR8506 CNRS-CentraleSupélec-UNIV PARIS SUD
SUPELEC, 91192 Gif-sur-Yvette, France
http://lss.centralesupelec.fr
Email: djafari@lss.supelec.fr
http://djafari.free.fr
http://publicationslist.org/djafari

Workshop Inverse problems in Finance and Human sciences, ATU, Tehran, Iran, September 24-25, 2016

# Contents

1. Examples of inverse problems
   - Low dimensional case
   - High dimensional case
2. Basics of Bayesian inference
3. Bayes for Inverse Problems and Machine Learning (Estimation, Prediction, Model Evaluation and selection)
4. Approximate Bayesian Computation (ABC)
   - Laplace approximation
   - Bayesian Information Criterion (BIC)
   - Variational Bayesian Approximation
   - Expectation Propagation (EP), Message Passing, MCMC, Exact Sampling, ...
5. Bayes for inverse problems
   - Traffic Management and Computed Tomography: A Linear problem
   - Differential Equations in Finance and Microwave imaging: A Bi-Linear or Non-Linear problem
6. Some canonical problems in Machine Learning
   - Regression, Classification and Model selection for classical and Big Data cases

# Examples of inverse problems

1. Discrete cases examples
2. Continuous cases examples

# Traffic management

$$\begin{array}{rcl} r_1\diamond & & \star c_1 \\ \vdots & & \vdots \\ \sum_j f_{ij} = r_i\diamond & \{f_{ij}\} & \star c_j = \sum_i f_{ij} \\ \vdots & & \vdots \\ r_I\diamond & & \star r_J \end{array}$$

- $M$ residential places each containing $r_i, i = 1, \cdots I$ cars
- $N$ working places each containing $c_j, j = 1, \cdots J$ parking lots
- We want to estimate the numbers $f_{i,j}$ of cars going from the residential place $i$ to working place $j$
- we know:

$$\sum_{j=1}^{J} f_{i,j} = r_i, i = 1, \cdots, I, \quad \sum_{i=1}^{I} f_{i,j} = c_j, j = 1, \cdots, J$$

- find $f_{i,j}$.

# Traffic management: A very low dimensional and simple example

- $I = 2,\ J = 2$

$$\sum_{j=1}^{2} f_{i,j} = r_i, i = 1, 2 \quad \sum_{i=1}^{2} f_{i,j} = c_j, j = 1, 2$$

$r_1 = 4, r_2 = 6, c_1 = 3, c_2 = 7$

- Writing it differently: find $f_{i,j}$

| | | |
|---|---|---|
| $f_{1,1}$ | $f_{1,2}$ | 4 |
| $f_{2,1}$ | $f_{2,2}$ | 6 |
| 3 | 7 | |

- A second example

| | | | |
|---|---|---|---|
| $f_{1,1}$ | $f_{1,2}$ | $f_{1,3}$ | 9 |
| $f_{2,1}$ | $f_{2,2}$ | $f_{2,3}$ | 10 |
| 3 | 7 | 11 | |

- Then, we extend it greater dimension

# Traffic management

Think about:

- Has this problem a solution?
- Is the solution unique?
- How to find all the possible solutions?
- How we can select one of these solution?
- Minimum Norm solutions? min $\sum_{i,j} f_{i,j}^2$ subject to the data constraints.
- What about minimizing:
  - $l_1$ norm: $\sum_{i,j} |f_{i,j}|$
  - $l_\alpha$ norm: $\sum_{i,j} |f_{i,j}|^\alpha$
  - Entropy: $-\sum_{i,j} f_{i,j} \log f_{i,j}$
- What if there are uncertainties in the data ?
- If you have to decide to construct for a fast road which one you propose?

# Continuous equivalent problem

- Given two functions $r(y)$ and $c(x)$, find a function and $f(x,y)$ of 2 variables $x$ and $y$ such that:

$$\begin{cases} \int f(x,y)\,\mathrm{d}x = r(y) \\ \int f(x,y)\,\mathrm{d}y = c(x) \end{cases}$$

- When $r(y)$ and $c(x)$ are the marginal probability distribution of a joint probability distribution $f(x,y)$, we see the Copula theory

# Prediction of gain

- Three days ago you gained 200 Euros, Tow days ago you gained 100 Euros, Yesterday 200 Euros, Today 100 Euros.
- How much are you expecting to gain tomorrow? after tomorrow ?

| $t_i$ | $-3$ | $-2$ | $-1$ | $0$ | $1$ | $2$ |
|---|---|---|---|---|---|---|
| $x_i$ | 200 | 100 | 160 | 200 | ? | ? |

- Think about the following models:
  - $x(t_i) = \theta_0 + \theta_1(t_i) + \epsilon_i$
  - $x(t_i) = \theta_0 + \theta_1(t_i) + \theta_2(t_i^2) + \epsilon_i$
  - $x(t_i) = \theta_0 + \theta_1(t_i) + \theta_2(t_i^2) + \theta_3(t_i^3) + \epsilon_i$
  - ...
  - $x_i = \theta_1 x_{i-1} + \epsilon_i$
  - $x_i = \theta_1 x_{i-1} + \theta_2 x_{i-2} + \epsilon_i$
  - $x_i = \sum_k^K \theta_k x_{i-k} + \epsilon_i$
  - ...
  - $x_i = \theta_0 + \theta_1 \sin(\pi t_i) + \epsilon_i$

# Prediction

- Write the equation $x(t_i) = \theta_0 + \theta_1(t_i) + \theta_2(t_i^2) + \theta_3(t_i^3) + \epsilon_i$ in Matrix form

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} 1 & t_1 & t_1^2 & t_1^3 \\ 1 & t_2 & t_2^2 & t_2^3 \\ \vdots & & & \\ 1 & t_N & t_N^2 & t_N^3 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

- Use $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ and solve for $\boldsymbol{\theta}$

$$\widehat{\boldsymbol{\theta}} = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{x}$$

- Do prediction for any $t_i$:
  $x(t_i) = \theta_0 + \theta_1(t_i) + \theta_2(t_i^2) + \theta_3(t_i^3) + \epsilon_i$

# Prediction

- Use any dictionary $\{h_k(t), k = 1, \cdots, K\}$
- Write the equation $x(t_i) = \sum_{k=1}^{K} \theta_k h_k(t_i) + \epsilon_i$ in matrix form

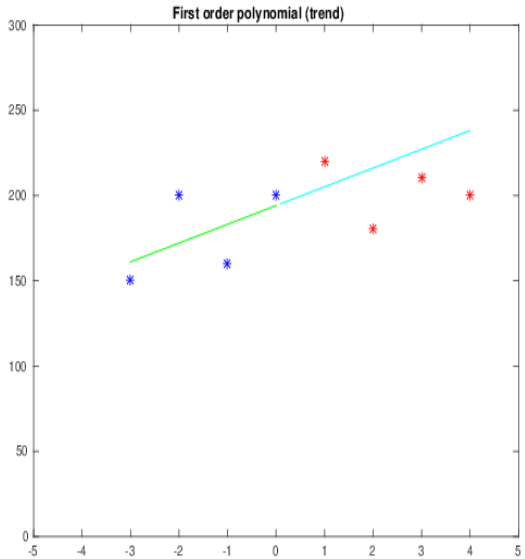
$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} h_1(t_1) & h_2(t_1) & \cdots & h_K(t_1) \\ h_1(t_2) & h_2(t_2) & \cdots & h_K(t_2) \\ h_1(t_3) & h_2(t_3) & \cdots & h_K(t_3) \\ \vdots & & & \\ h_1(t_N) & h_2(t_N) & \cdots & f_K(t_N) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_K \end{bmatrix}$$

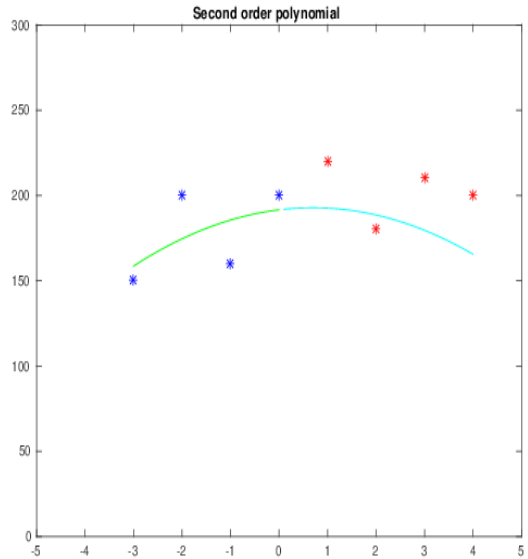
- Use $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ and solve for $\boldsymbol{\theta}$

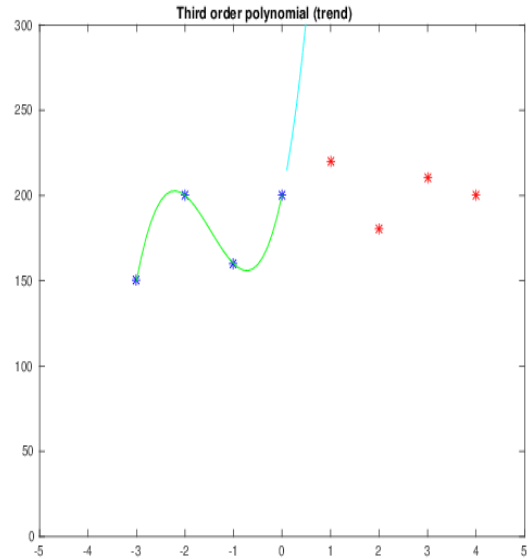
$$\widehat{\boldsymbol{\theta}} = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{x}$$

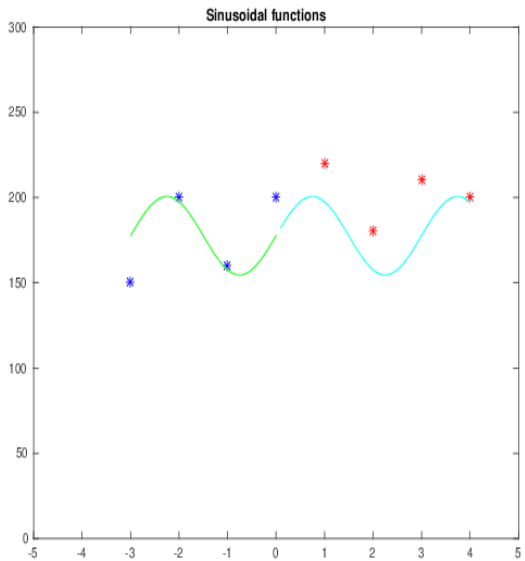
- Do prediction for any $t_i$: $x(t_i) = \sum_{k=1}^{K} \theta_k h_k(t_i) + \epsilon_i$

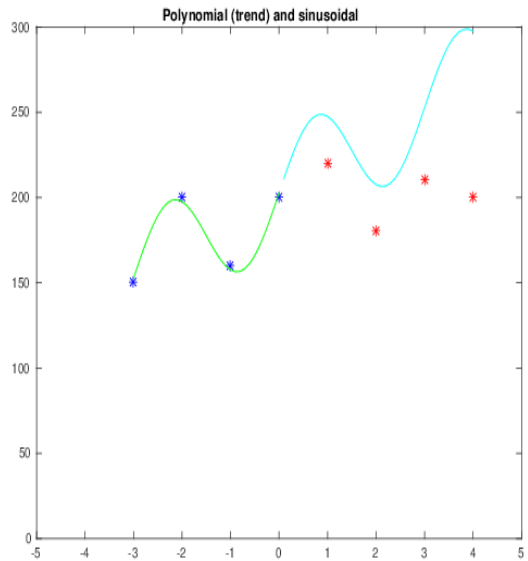
# Prediction examples

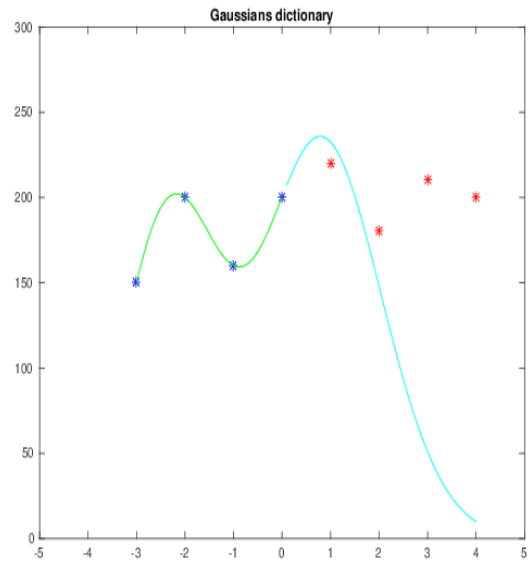


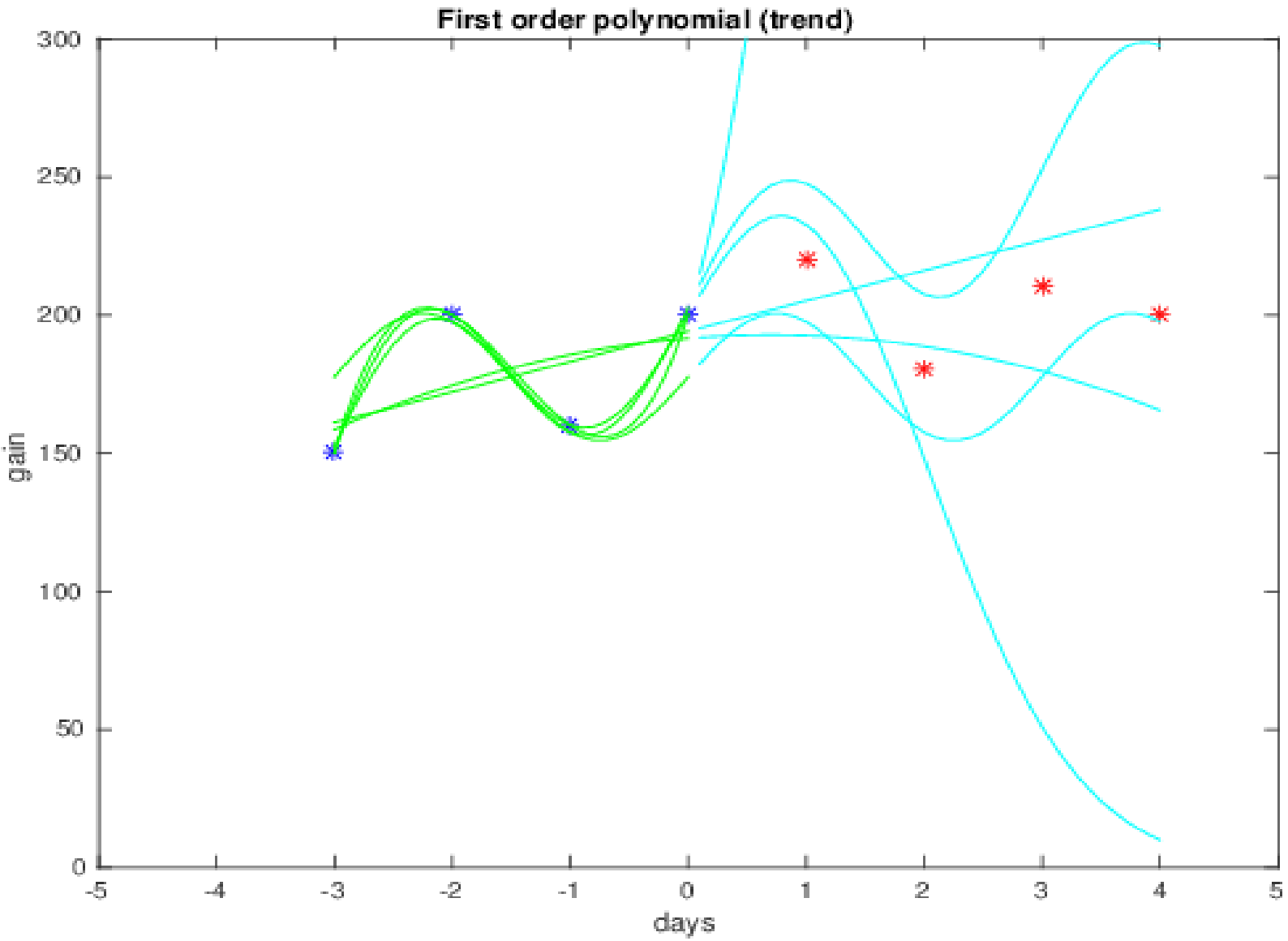



# Prediction examples

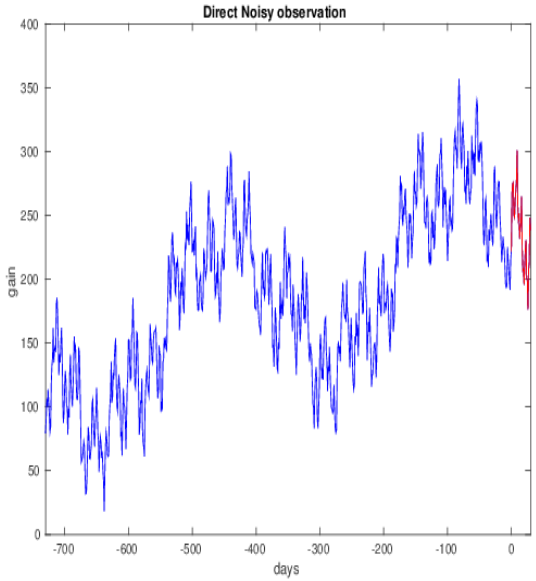


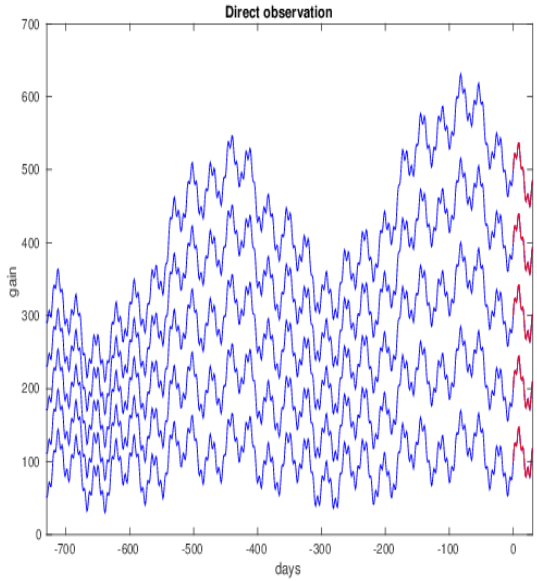
# Prediction of gain

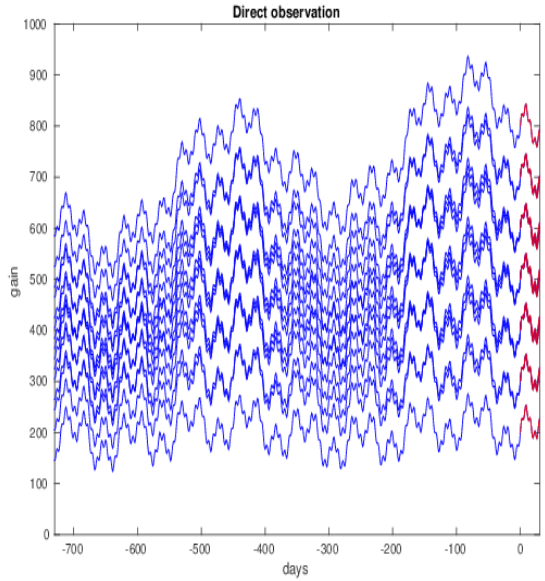
The same problem, but this time you have much more data

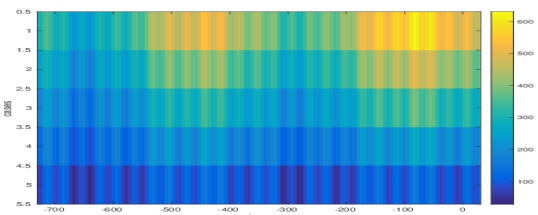
- regular daily since 2 years
- regular daily but with some missing since 2 years
- regular daily but with some outliers since 2 years
- regular daily of yourself and your colleagues with the same rank and positions
- regular daily of yourself and your colleagues with the same rank and positions, but also other colleagues in your company
- regular daily of yourself and your colleagues with the same rank and positions, but also other colleagues in your company and in many other companies

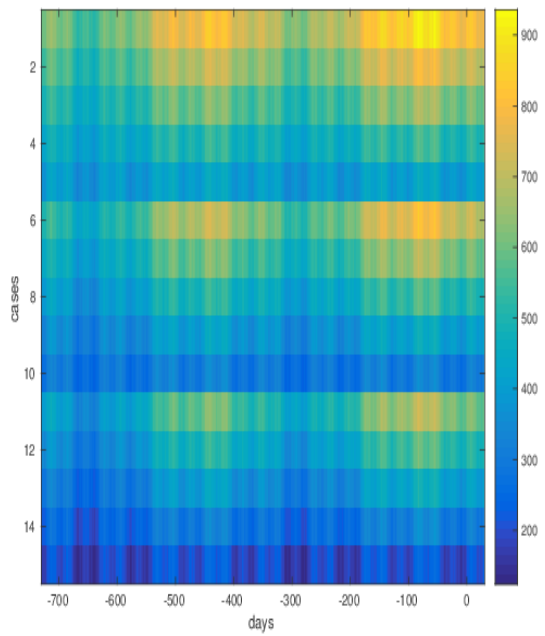
# Prediction problems examples



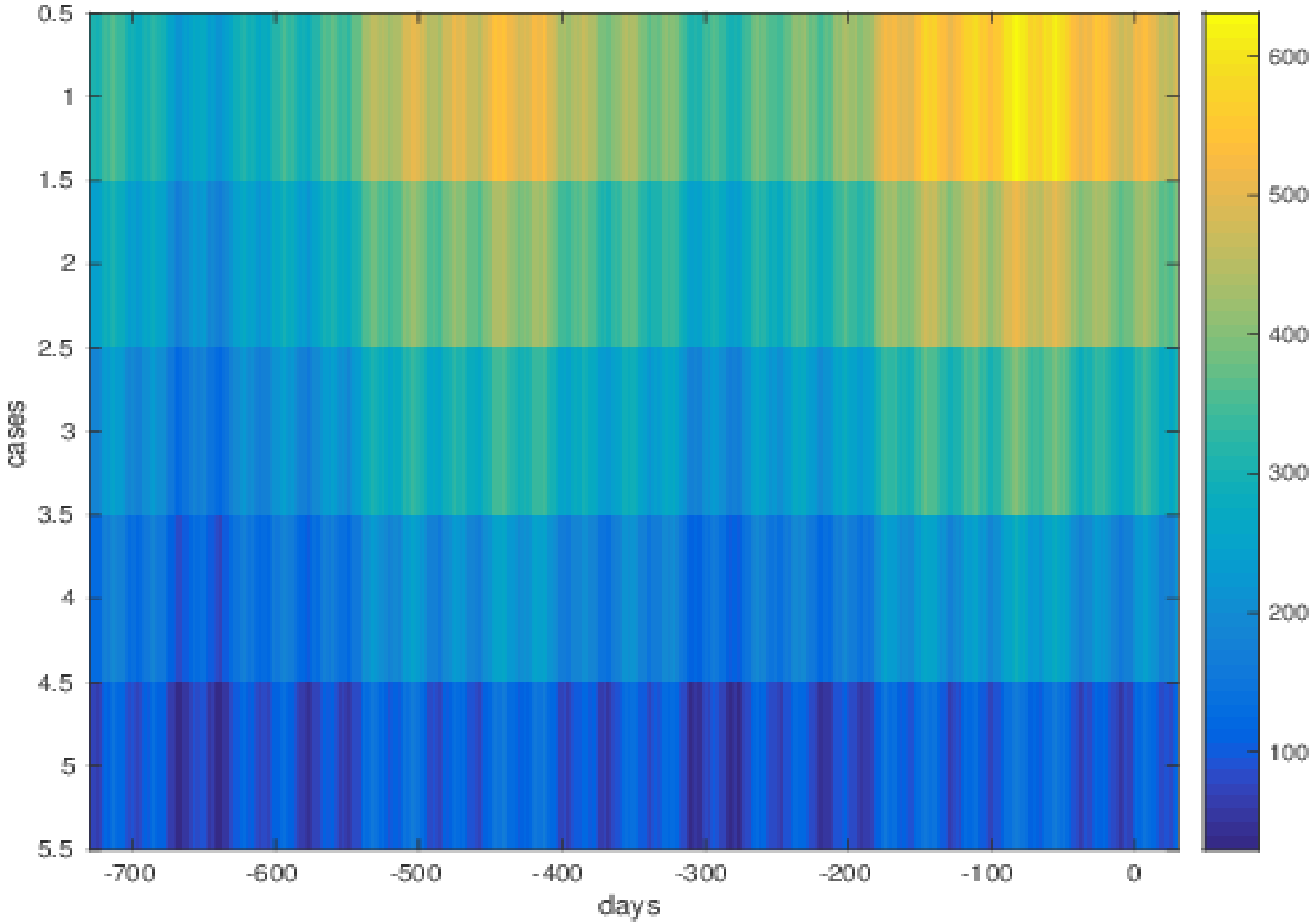



# Prediction examples



# Prediction with direct or indirect observation

The same problem, but this time you want to do daily prediction, but you have the data either weakly or monthly.

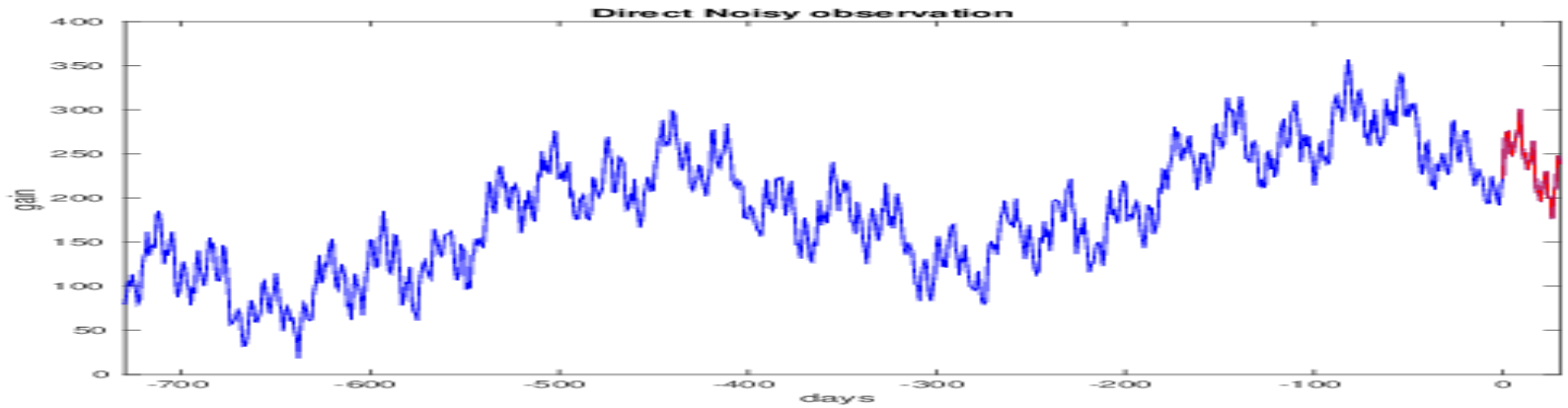
- Each week, you have the mean value of the week
- Each month, you have the mean value of the month

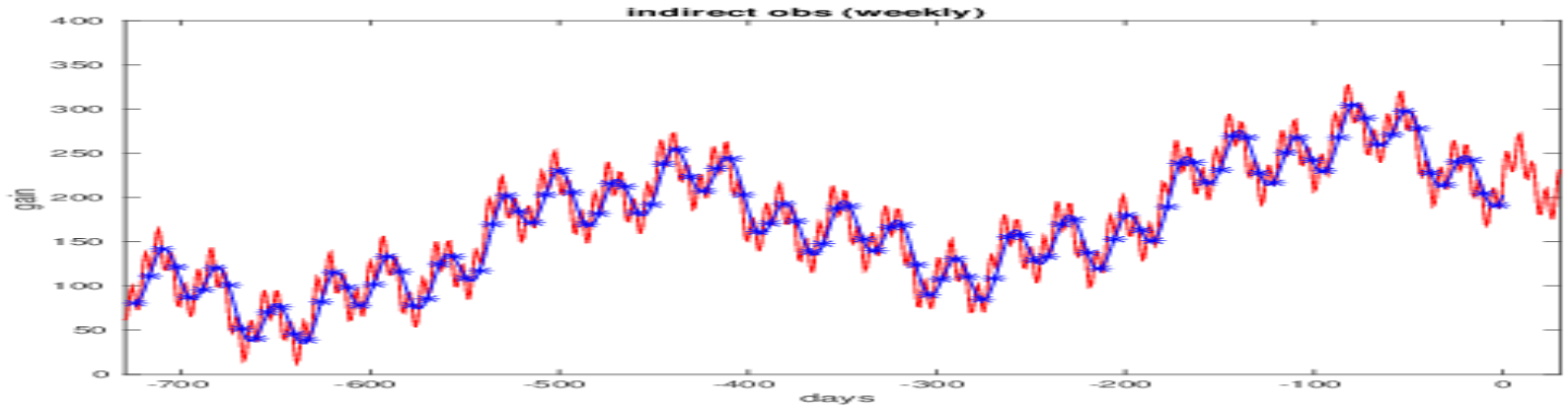
If we note $f(n)$ the daily data and $g(m)$ the observed data, then we have

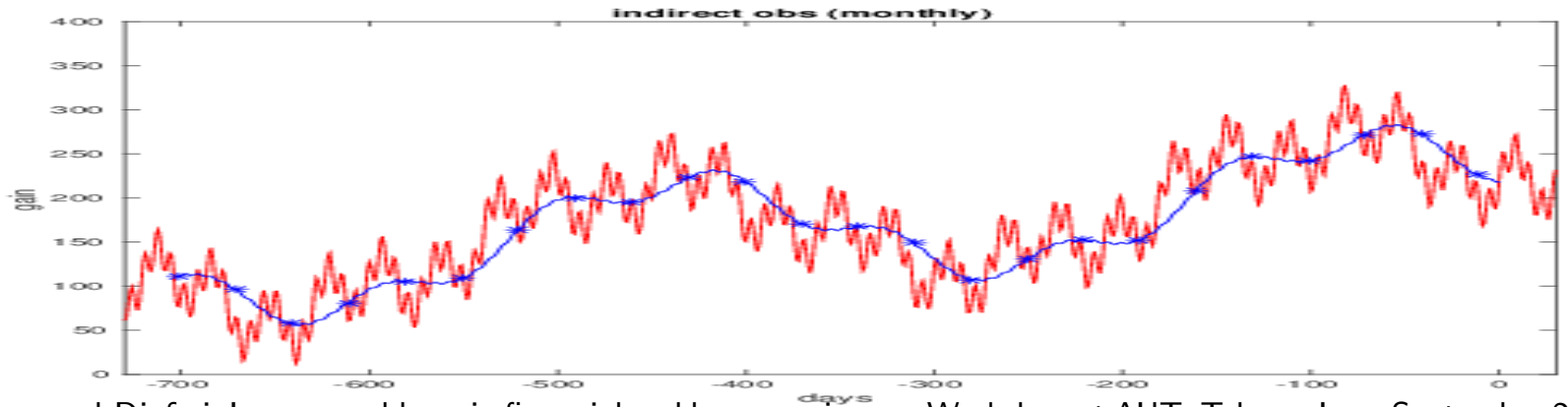
- last day of the week: $g(m) = f(n = 7 * m)$
- last day of the month: $g(m) = f(n = 30 * m)$
- mean value of the week: $g(m) = \sum_{k=1}^{7} f(7 * m - k + 1)$
- mean value of the month: $g(m) = \sum_{k=1}^{30} f(7 * m - k + 1)$
- Uncertain data

$$g(m) = \sum_{k=1}^{K} f(7 * m - k + 1) + \epsilon(m)$$

# Prediction problems examples



# Prediction examples

# Population observation, modelling and evolution

We know approximate numbers of population in some of the cities

$$g(x_i, y_i), i = 1, \cdots, M$$

(probably every 4 years since 40 years

$$g_i(t_n), n = -40 : 4 : 0)$$

and we want to know

- The distribution of the population in the whole country $\{f_{i,j}, i = 1, \cdots, I, j = 1, \cdots, J\}$
- The evolution of this distribution year by year
- Prediction of this distribution of the population in the future years

But also

- To model the evolution of this distribution and its correlation with some other external events

# Population observation, modelling and evolution

Think about modelling:

- Discrete writing

$$g(x_i, y_i) = \sum_{(k,l):(x_k,y_l)\in\mathcal{R}_r} f(x_i - x_k, y_i - y_l) + \epsilon(x_i, y_i)$$

- Continuous writing

$$g(x_i, y_i) = \int_{\mathcal{R}_r} f(x_i - x, y_i - y)\,\mathrm{d}x\,\mathrm{d}y + \epsilon(x_i, y_i)$$

- Differential forms

$$\frac{\partial f(x,y)}{\partial x} + f(x,y) = 0 \text{ with initial condition } f(x,y) = g(x,y)$$

# Inverse problems scientific communities

Two communities working on Inverse problems:

- Mathematical departments:
  Analytical methods: Existance and Uniqueness
  Differential equations, PDE

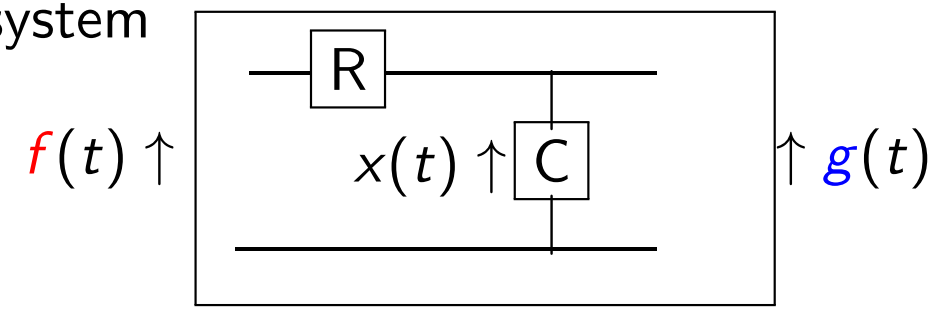
- Engineering and Computer sciences:
  Algebraic methods: Discretization, Uniqueness and Stability
  Integral equations, Discretization using Moments method, Galerkin, ...

Two examples:

- Deconvolution: Inverse filtering and Wiener filtering
- X ray Computed Tomography: Radon transform:
  Direct Inversion or Filtered Backprojection methods

# Differential Equation, State Space and Input-Output

A simple electric system




$$f(t) = R\,i(t) + v_c(t) = RC\,\frac{\partial x(t)}{\partial t} + x(t), \quad RC = 1$$

- Differential Equation Modelling

$$\frac{\partial x(t)}{\partial t} + x(t) = f(t), \qquad x(t) = g(t)$$

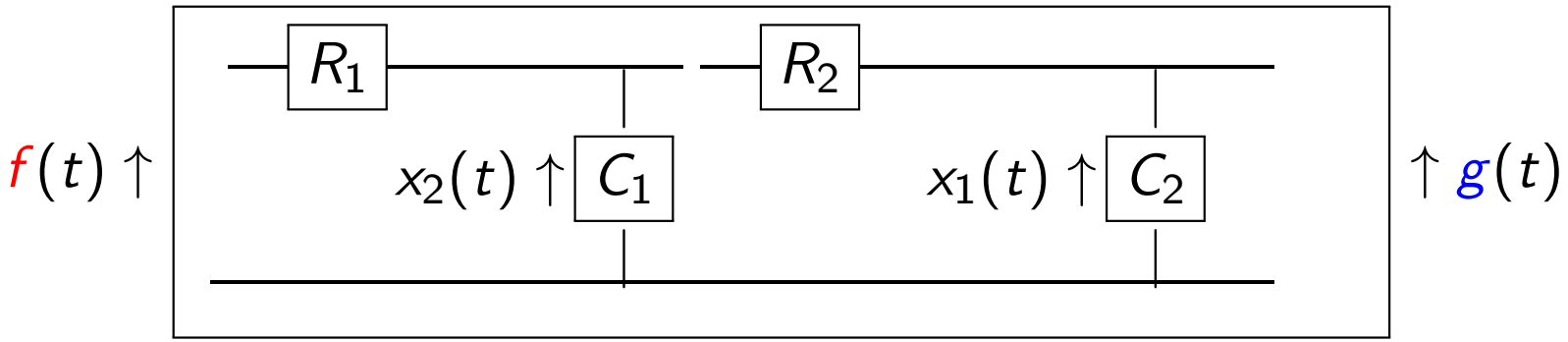
- State Space Modelling

$$\begin{cases} \frac{\partial x(t)}{\partial t} & = -x(t) + f(t) \\ g(t) & = x(t) \end{cases}$$

- Input-Output Modelling

$$\begin{cases} \frac{\partial x(t)}{\partial t} = -x(t) + f(t) \\ g(t) = x(t) \end{cases} \rightarrow \begin{cases} pX(p) = -X(p) + F(p) \rightarrow X(p) = \frac{1}{p+1}F(p) \\ g(t) = x(t) = h(t) * f(t), \quad h(t) = \exp[-t] \end{cases}$$

# A more complex electric system example




$$f(t) = \frac{\partial x_2(t)}{\partial t} + x_2(t), \quad x_2(t) = \frac{\partial x_1(t)}{\partial t} + x_1(t)$$

- Differential Equation model: $\frac{\partial^2 x_1(t)}{\partial t^2} + 2\frac{\partial x_1(t)}{\partial t} + x_1(t) = f(t)$
- State space model

$$\begin{cases} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} f(t) \\ g(t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} x_1(t) \end{cases}$$

- Input-Output Model: $g(t) = h(t) * f(t)$

# Design/Control Inverse problems examples

Simple Electrical system:

$$a\frac{\partial x(t)}{\partial t} + x(t) = f(t), \quad x(0) = x_0, \quad g(t) = x(t)$$

- Design: $\theta = a = RC$
  - Forward: Given $\boldsymbol{\theta} = a$ and $f(t), t > 0$, find $x(t), t > 0$
  - Inverse: Given $x(t)$ and $f(t)$ find $\theta = a$
- Control: $f(t)$
  - Forward: Given $\boldsymbol{\theta} = a$ and $f(t), t > 0$, find $x(t), t > 0$
  - Inverse: Given $\boldsymbol{\theta} = a$ and $x(t), t > 0$, find $f(t)$

More complex Electrical system:

$$f(t) = b\frac{\partial x_2(t)}{\partial t} + x_2(t), \quad x_2(t) = a\frac{\partial x_1(t)}{\partial t} + x_1(t), \quad g(t) = x_1(t)$$

$$\theta = (a = R_1C_1, b = R_2C_2)$$

# Design/Control Inverse problems examples

Mass-spring-dashpot system

$$m\frac{\partial^2 x(t)}{\partial t^2} + c\frac{\partial x(t)}{\partial t} + k = F(t), \quad x(0) = x_0, \quad \frac{\partial x}{\partial t}(0) = v_0$$

- Design: $\theta = (m, c, k)$
  - Forward: Given $\boldsymbol{\theta} = (m, c, k)$, $x_0, v_0$ and $F(t), t > 0$, find $x(t), t > 0$
  - Inverse: Given $x(t)$ for $t > 0$, $v_0$, $F(t)$ find $\theta = (m, c, k)$
- Control: $F(t)$
  - Forward: Given $\boldsymbol{\theta} = (m, c, k)$, $x_0, v_0$ and $F(t), t > 0$, find $x(t), t > 0$
  - Inverse: Given $\boldsymbol{\theta} = (m, c, k)$, $v_0$ and $x(t), t > 0$, find $F(t)$

# Input-Output model

- Linear Systems
  - Single Input Single Output (SISO) systems

$$y(t) = \int h(t,\tau)\, u(\tau)\, \mathrm{d}\tau$$

  - Multi Input Multi Output (MIMO) systems

$$\mathbf{y}(t) = \int \mathbf{H}(t,\tau)\, \mathbf{u}(\tau)\, \mathrm{d}\tau$$

- Linear Time Invariant System
  - SISO Convolution

$$y(t) = h(t) * u(t) = \int h(t-\tau)\, u(\tau)\, \mathrm{d}\tau$$

  - MIMO Convolution

$$\mathbf{y}(t) = \int \mathbf{H}(t-\tau)\, \mathbf{u}(\tau)\, \mathrm{d}\tau$$

  - Impulse response $h(t)$ or $\mathbf{H}(t) = \begin{bmatrix} . & . & . \\ . & h_{ij}(t) & . \\ . & . & . \end{bmatrix}$

# State space model: Continuous case

Dynamic systems:

- Single Input Single Output (SISO) system:

$$\begin{cases} \dot{x}(t) &= A\,x(t) + B\,u(t) & \text{State equation} \\ y(t) &= C\,x(t) + D\,v(t) & \text{Observation equation} \end{cases}$$

- Multiple Input Multiple Output (MIMO) system:

$$\begin{cases} \dot{\mathbf{x}}(t) &= \mathbf{H}\,\mathbf{x}(t) + \mathbf{B}\,\mathbf{u}(t) & \text{State equation} \\ \mathbf{y}(t) &= \mathbf{C}\,\mathbf{x}(t) + \mathbf{D}\,\mathbf{v}(t) & \text{Observation equation} \end{cases}$$

$\mathbf{H}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ are the matrices of the system.

# Modelling with Partial Differential Equations

- Different PDE

$$\frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2} + f(x,y) = 0$$

$$\frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2} + \frac{\partial f(x,y)}{\partial x}\frac{\partial f(x,y)}{\partial y} f(x,y) = 0$$

with initial cond. $f(x,y) = g(x,y)$

# Prediction with indirect observation

- Data available every $K$ days with uncertainty

$$g(m) = \sum_{k=1}^{K} f(K * m - k + 1) + \epsilon(m)$$

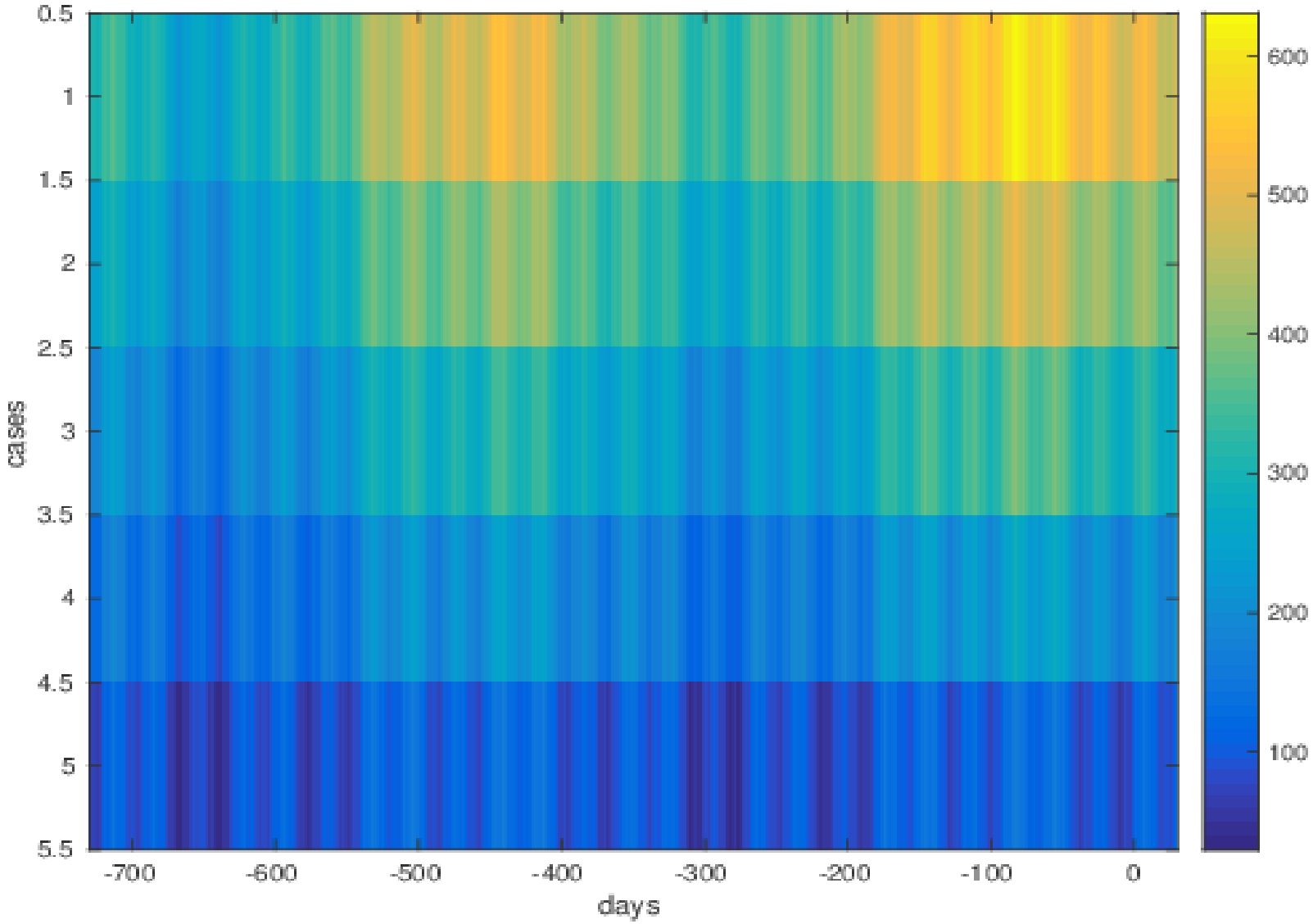
- more general: Convolution

$$g(m) = \sum_{k=1}^{K} h(k) f(n - k + 1) + \epsilon(m)$$

- One can show easily that both can be written as

$$\mathbf{g} = \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon}$$

# Prediction examples



# Simple Examples

| | | | | |
|---|---|---|---|---|
| $f_1$ | $f_5$ | $f_9$ | $f_{13}$ | $g_8$ |
| $f_2$ | $f_6$ | $f_{10}$ | $f_{14}$ | $g_7$ |
| $f_3$ | $f_7$ | $f_{11}$ | $f_{15}$ | $g_6$ |
| $f_4$ | $f_8$ | $f_{12}$ | $f_{16}$ | $g_5$ |
| $g_1$ | $g_2$ | $g_3$ | $g_4$ | |

| | | | | |
|---|---|---|---|---|
| $f_{11}$ | $f_{12}$ | $f_{13}$ | $f_{14}$ | $g_{24}$ |
| $f_{21}$ | $f_{22}$ | $f_{23}$ | $f_{24}$ | $g_{23}$ |
| $f_{31}$ | $f_{32}$ | $f_{33}$ | $f_{34}$ | $g_{22}$ |
| $f_{41}$ | $f_{42}$ | $f_{43}$ | $f_{44}$ | $g_{21}$ |
| $g_{11}$ | $g_{12}$ | $g_{13}$ | $g_{14}$ | |

Noting also by:
$\mathbf{g}_1 = [g_1, \cdots, g_4]^t = [g_{11}, \cdots, g_{14}]^t$,
$\mathbf{g}_2 = [g_5, \cdots, g_8]^t = [g_{21}, \cdots, g_{24}]^t$
and the matrices $\mathbf{H}_1$, $\mathbf{H}_2$ and $\mathbf{H}$ such that:

$$\mathbf{g}_1 = \mathbf{H}_1\mathbf{f}, \quad \mathbf{g}_2 = \mathbf{H}_2\mathbf{f}, \quad \mathbf{g} = \mathbf{H}\mathbf{f} = \left[\frac{\mathbf{H}_1}{\mathbf{H}_2}\right]\mathbf{f}$$

# Inversion, Generalized inversion

- Forward problem: Given $\mathbf{f}$ compute $\mathbf{g}$:

$$f = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \longrightarrow \begin{cases} \mathbf{g}_1 = \mathbf{H}_1\mathbf{f} = [0\,2\,2\,0]^t, \\ \mathbf{g}_2 = \mathbf{H}_2\mathbf{f} = [0\,2\,2\,0]^t, \\ \mathbf{g} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{H}_1 \\ \hline \mathbf{H}_2 \end{bmatrix} \mathbf{f} \\ \mathbf{g} = \mathbf{H}\mathbf{f} = [0\,2\,2\,0\,0\,2\,2\,0]^t \end{cases}$$

- Inverse problem: Given $\mathbf{g}$ find $\mathbf{f}$.
  Many possible solutions:

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} -.5 & 0 & 0 & .5 \\ 1 & 2 & 0 & -1 \\ -1 & 0 & 2 & 1 \\ 0.5 & 0 & 0 & -.5 \end{bmatrix} \quad \begin{bmatrix} -.5 & 0 & 0 & .5 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ .5 & 0 & 0 & -.5 \end{bmatrix}$$

# MN, LS and MNLS solutions

- Minimum Norme (MN) solution:

$$\widehat{\mathbf{f}} = \underset{\mathbf{Hf}=\mathbf{g}}{\arg\min} \left\{ \|\mathbf{f}\|^2 \right\}$$

and if $\mathbf{H}\mathbf{H}^t$ was invertible, then we had: $\widehat{\mathbf{f}} = \mathbf{H}^t(\mathbf{H}\mathbf{H}^t)^{-1}\mathbf{g}$.
But

$$\text{svd}(\mathbf{H}\mathbf{H}^t) = [8\,4\,4\,4\,4\,4\,4\,0]$$

- Least Squares (LS) solution:

$$\widehat{\mathbf{f}} = \underset{\mathbf{f}}{\arg\min} \left\{ \|\mathbf{g} - \mathbf{H}\mathbf{f}\|^2 \right\},$$

and if $\mathbf{H}^t\mathbf{H}$ was invertible, then we had: $\widehat{\mathbf{f}} = (\mathbf{H}^t\mathbf{H})^{-1}\mathbf{H}^t\mathbf{g}$.
But

$$\text{svd}(\mathbf{H}^t\mathbf{H}) = [8\,4\,4\,4\,4\,4\,4\,0\,0\,0\,0\,0\,0\,0\,0\,0]$$

# SVD and MNLS solutions

- Truncation of the singular values to define an unique Generalized Inverse solution

$$\widehat{\mathbf{f}} = \sum_{k=1}^{k} \frac{< \mathbf{g}, \mathbf{u}_k >}{\lambda_k} \mathbf{v}_k$$

where $\mathbf{u}_k$ and $\mathbf{v}_k$ are, respectively, the eigenvectors of $\mathbf{A}\mathbf{A}^t$ and $\mathbf{A}^t\mathbf{A}$ and $\lambda_k$ their corresponding eigen values.

$$\widehat{\mathbf{f}} = \begin{bmatrix} -0.2500 & 0.2500 & 0.2500 & -0.2500 \\ 0.2500 & 0.7500 & 0.7500 & 0.2500 \\ 0.2500 & 0.7500 & 0.7500 & 0.2500 \\ -0.2500 & 0.2500 & 0.2500 & -0.2500 \end{bmatrix}$$

- MNLS

$$\widehat{\mathbf{f}} = \arg\min_{\mathbf{f}} \left\{ \|\mathbf{g} - \mathbf{H}\mathbf{f}\|^2 + \|\mathbf{f}\|^2 \right\},$$

# Regularization theory

Inverse problems = Ill posed problems

$\longrightarrow$ Need for prior information

Functional space (Tikhonov):

$$g = \mathcal{H}(f) + \epsilon \longrightarrow J(f) = ||g - \mathcal{H}(f)||_2^2 + \lambda ||\mathcal{D}f||_2^2$$

Finite dimensional space (Philips & Towmey): $\mathbf{g} = \mathbf{H}(\mathbf{f}) + \boldsymbol{\epsilon}$

- Minimum norme LS (MNLS): $J(\mathbf{f}) = ||\mathbf{g} - \mathbf{H}(\mathbf{f})||^2 + \lambda ||\mathbf{f}||^2$
- Classical regularization: $J(\mathbf{f}) = ||\mathbf{g} - \mathbf{H}(\mathbf{f})||^2 + \lambda ||\mathbf{D}\mathbf{f}||^2$
- More general regularization:

$$J(\mathbf{f}) = \mathcal{Q}(\mathbf{g} - \mathbf{H}(\mathbf{f})) + \lambda \Omega(\mathbf{D}\mathbf{f})$$

or

$$J(\mathbf{f}) = \Delta_1(\mathbf{g}, \mathbf{H}(\mathbf{f})) + \lambda \Delta_2(\mathbf{f}, \mathbf{f}_\infty)$$

**Limitations:**

- Errors are considered implicitly white and Gaussian
- Limited prior information on the solution
- Lack of tools for the determination of the hyperparameters

# Basic Bayes

- Product rules

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A) \rightarrow P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Sum rule

$$P(B) = \sum_A P(A|B)P(B)$$

- Bayes rule (discrete events)

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_A P(B|A)P(A)}$$

- $P(\textit{hypothesis}|\textit{data}) = \frac{P(\textit{data}|\textit{hypothesis})P(\textit{hypothesis})}{P(\textit{data})}$
- Bayes rule (Continuous variables with finite parametric models)

$$p(\boldsymbol{\theta}|\mathbf{d}) = \frac{p(\mathbf{d}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{p(\mathbf{d})} = \frac{p(\mathbf{d}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int p(\mathbf{d}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} \propto p(\mathbf{d}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})$$

# Basic Bayes

- $P(\textit{hypothesis}|\textit{data}) = \frac{P(\textit{data}|\textit{hypothesis})P(\textit{hypothesis})}{P(\textit{data})}$
- Bayes rule tells us how to do inference about hypotheses from data.
- Finite parametric models:

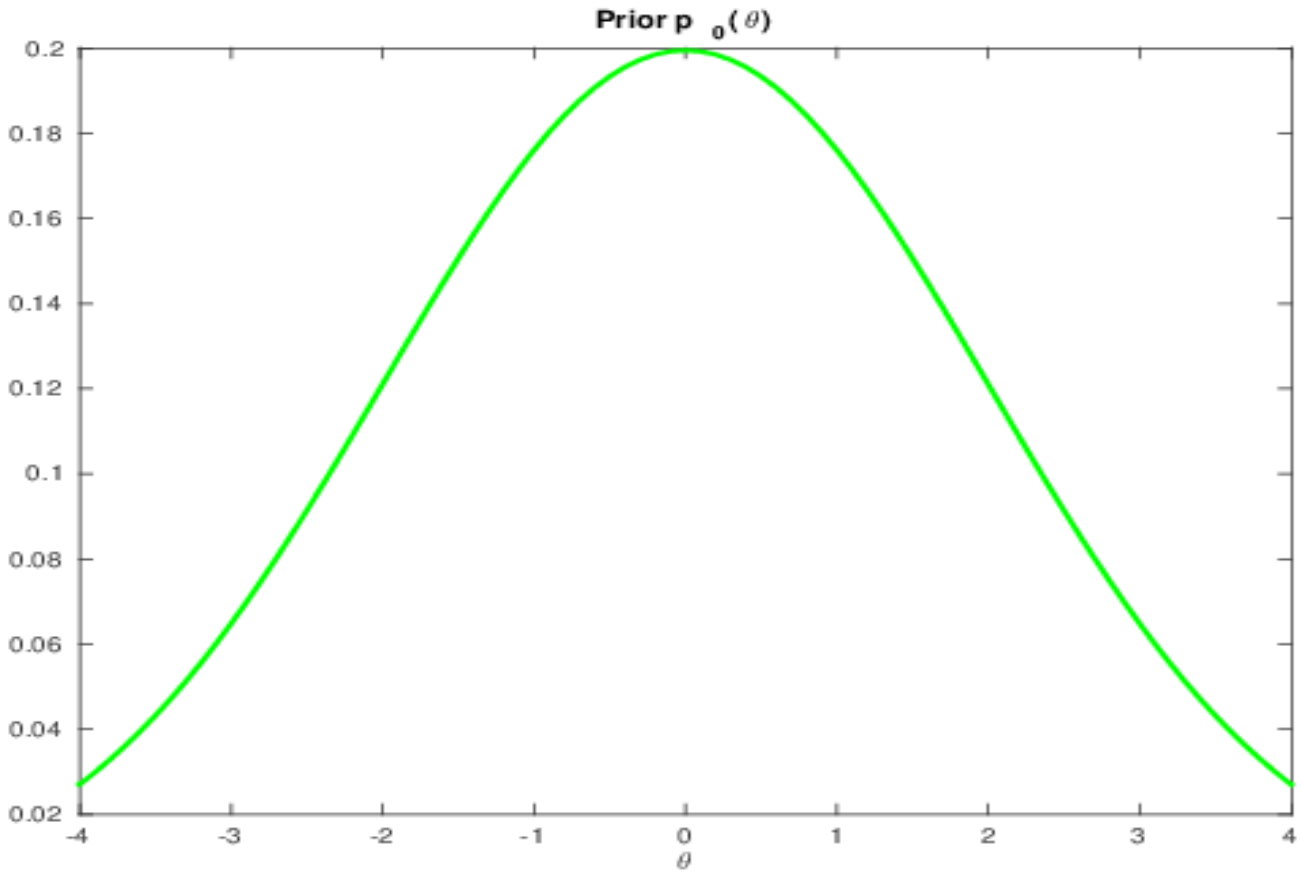
$$p(\boldsymbol{\theta}|\mathbf{d}) = \frac{p(\mathbf{d}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{d})}$$

- Forward model: $p(\mathbf{d}|\boldsymbol{\theta})$
  called also likelihood of the parameters in data $\mathcal{L}(\boldsymbol{\theta}) = p(\mathbf{d}|\boldsymbol{\theta})$
- Prior knowledge: $p(\boldsymbol{\theta})$
- Posterior knowledge: $p(\boldsymbol{\theta}|\mathbf{d})$
- Evidence: $p(\mathbf{d}) = \int p(\mathbf{d}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$

# Bayesian inference: simple one parameter case

$$p(\theta), \mathcal{L}(\theta) = p(\mathbf{d}|\theta) \longrightarrow p(\theta|\mathbf{d}) \propto \mathcal{L}(\theta)\, p(\theta)$$

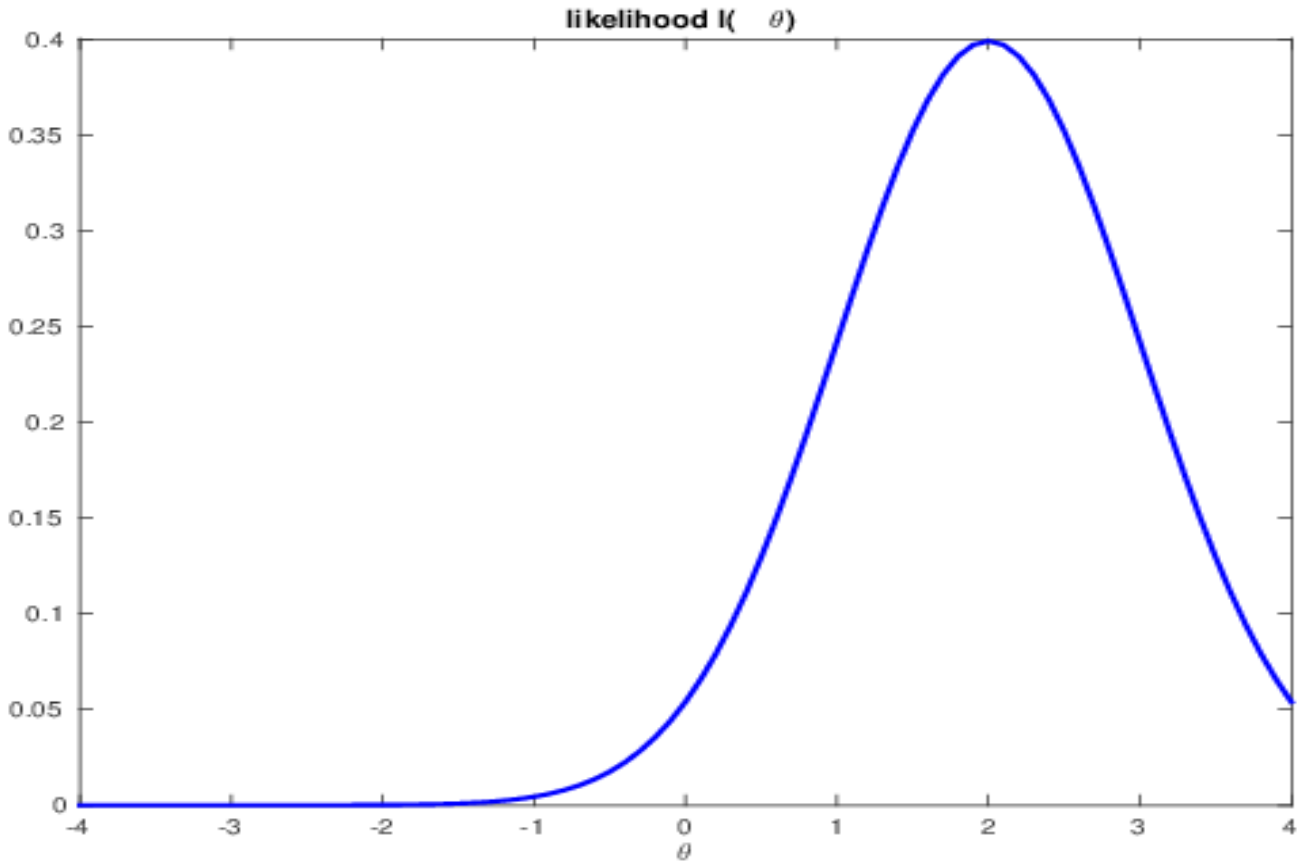
Prior: $p(\theta)$



# Bayesian inference: simple one parameter case

$$p(\theta), \mathcal{L}(\theta) = p(\mathbf{d}|\theta) \longrightarrow p(\theta|\mathbf{d}) \propto \mathcal{L}(\theta)\, p(\theta)$$

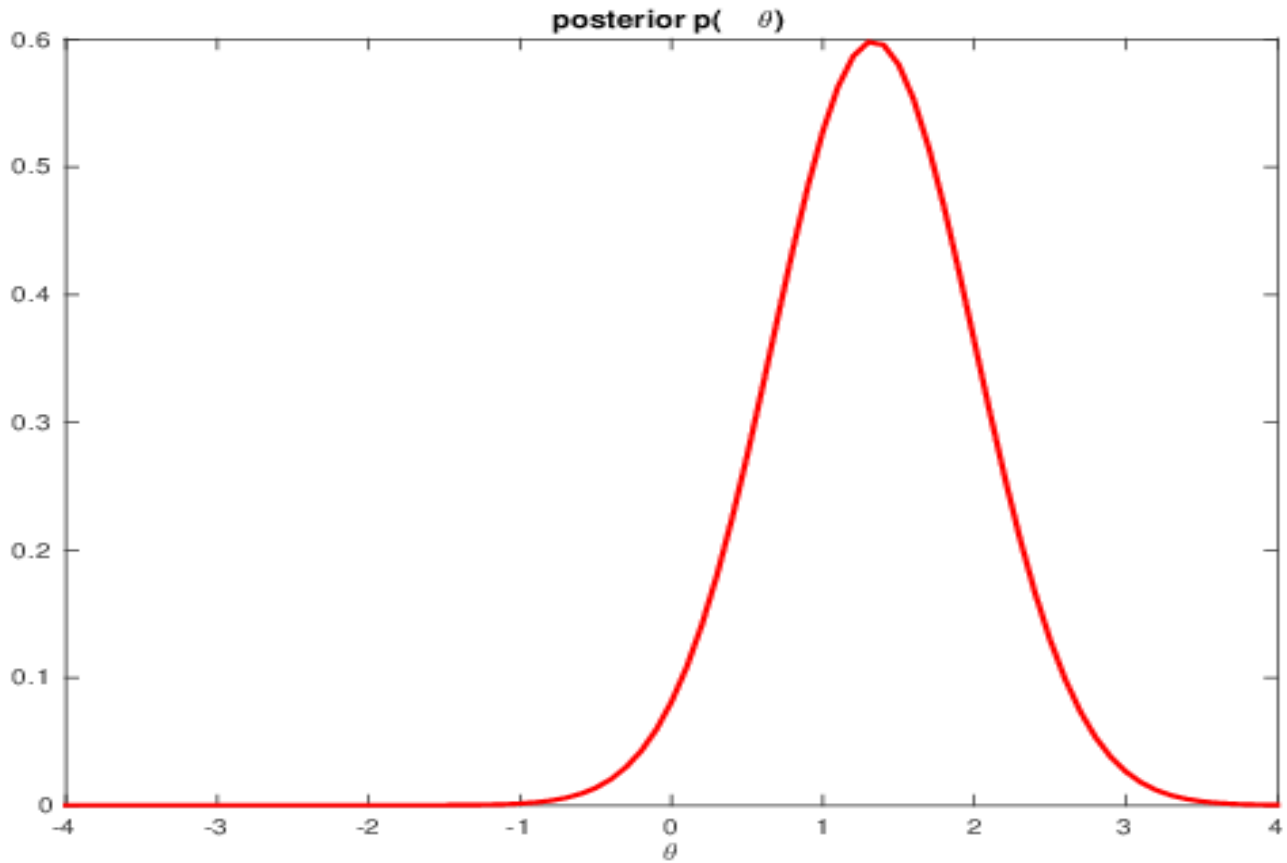
Likelihood: $\mathcal{L}(\theta) = p(\mathbf{d}|\theta)$



# Bayesian inference: simple one parameter case

$$p(\theta), \mathcal{L}(\theta) = p(\mathbf{d}|\theta) \longrightarrow p(\theta|\mathbf{d}) \propto \mathcal{L}(\theta)\, p(\theta)$$

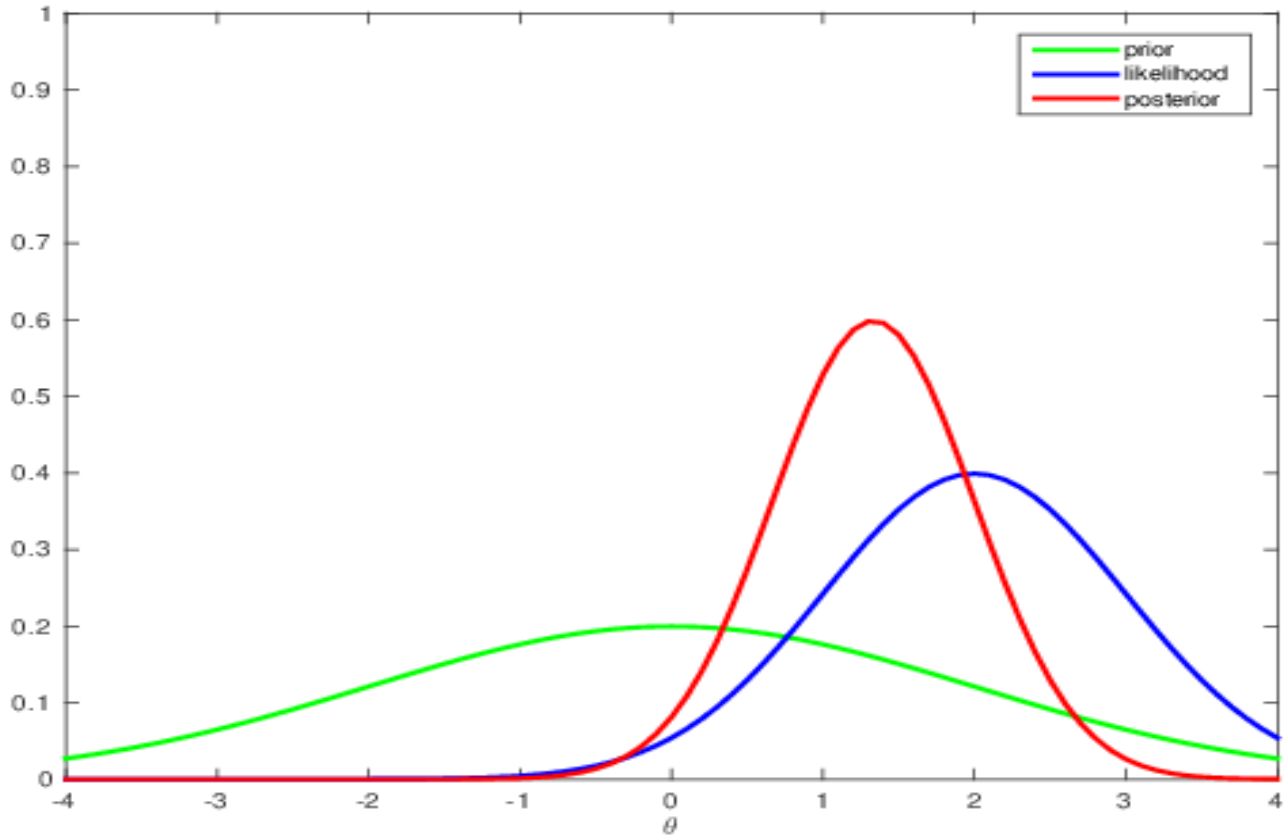
Posterior: $p(\theta|\mathbf{d}) \propto p(\mathbf{d}|\theta)\, p(\theta)$



# Bayesian inference: simple one parameter case

$$p(\theta), \mathcal{L}(\theta) = p(\mathbf{d}|\theta) \longrightarrow p(\theta|\mathbf{d}) \propto \mathcal{L}(\theta)\, p(\theta)$$

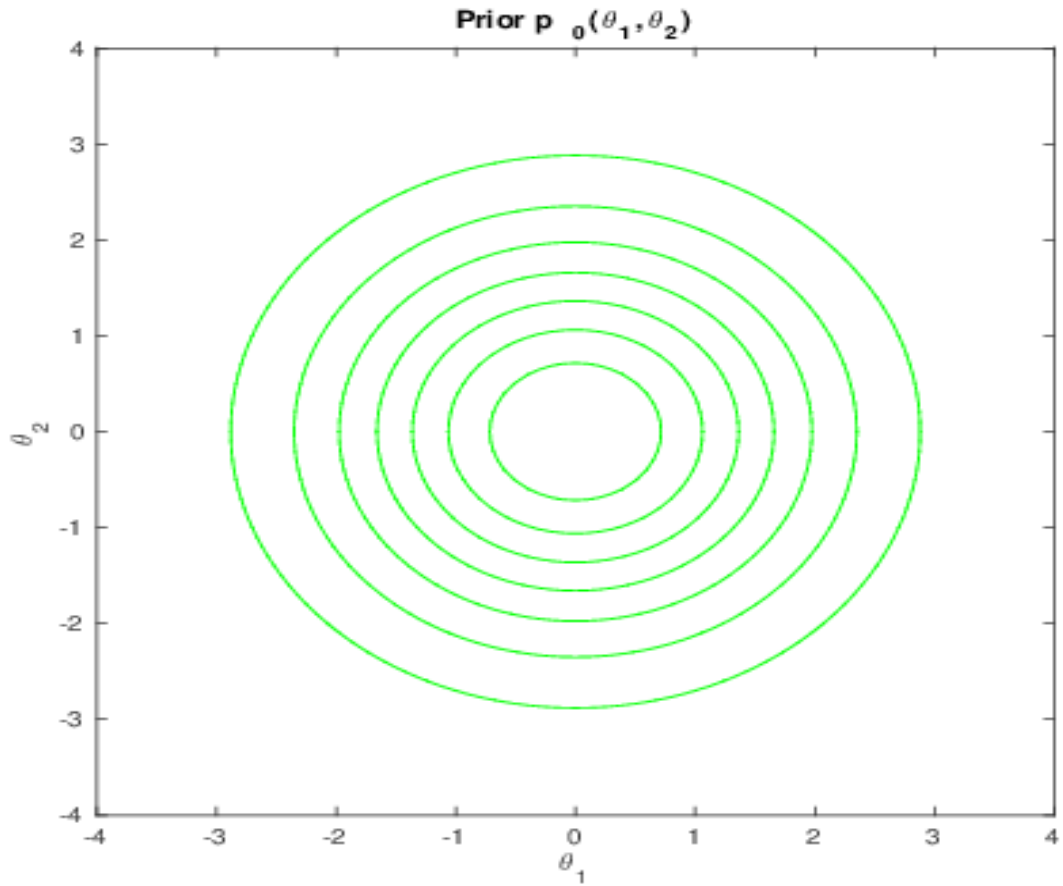
Prior, Likelihood and Posterior:



# Bayesian inference: simple two parameters case

$$p(\theta_1, \theta_2), \mathcal{L}(\theta_1, \theta_2) = p(\mathbf{d}|\theta_1, \theta_2) \longrightarrow p(\theta_1, \theta_2|\mathbf{d}) \propto \mathcal{L}(\theta_1, \theta_2)\, p(\theta_1, \theta_2)$$

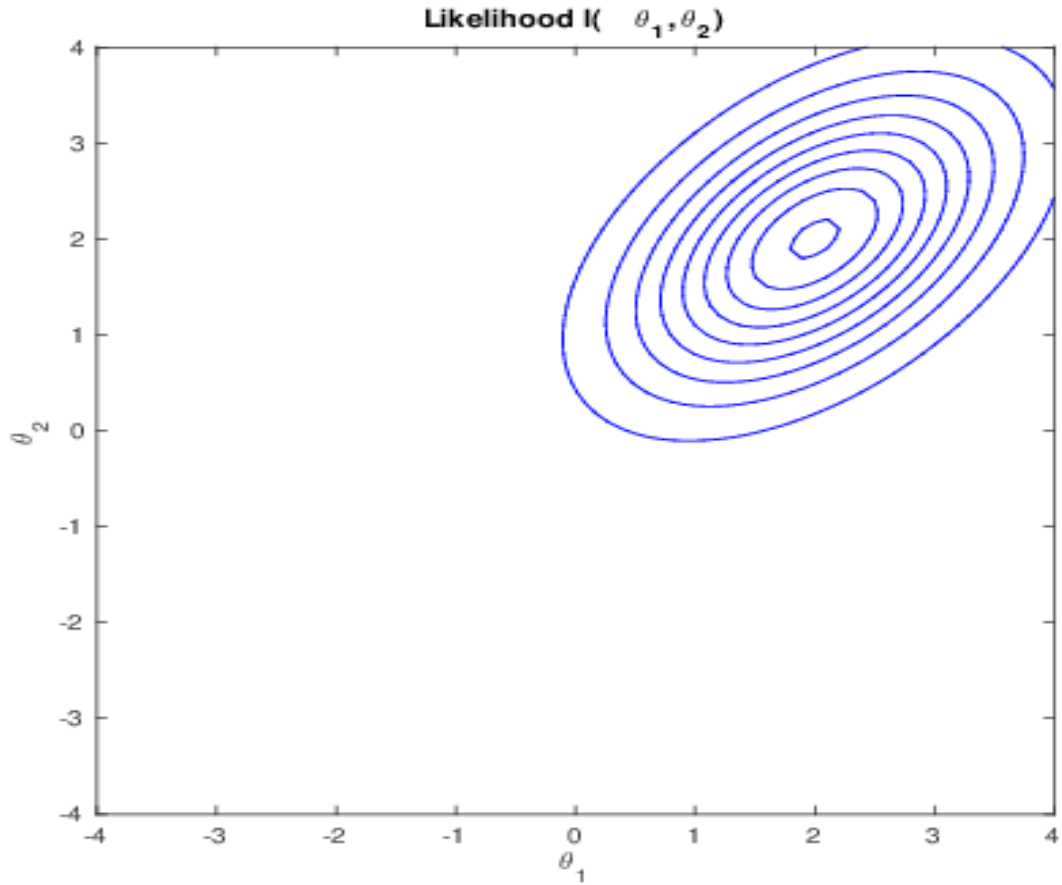
Prior: $p(\theta_1, \theta_2)$



# Bayesian inference: simple one parameter case

$$p(\theta_1, \theta_2), \mathcal{L}(\theta_1, \theta_2) = p(\mathbf{d}|\theta_1, \theta_2) \longrightarrow p(\theta_1, \theta_2|\mathbf{d}) \propto \mathcal{L}(\theta_1, \theta_2)\, p(\theta_1, \theta_2)$$

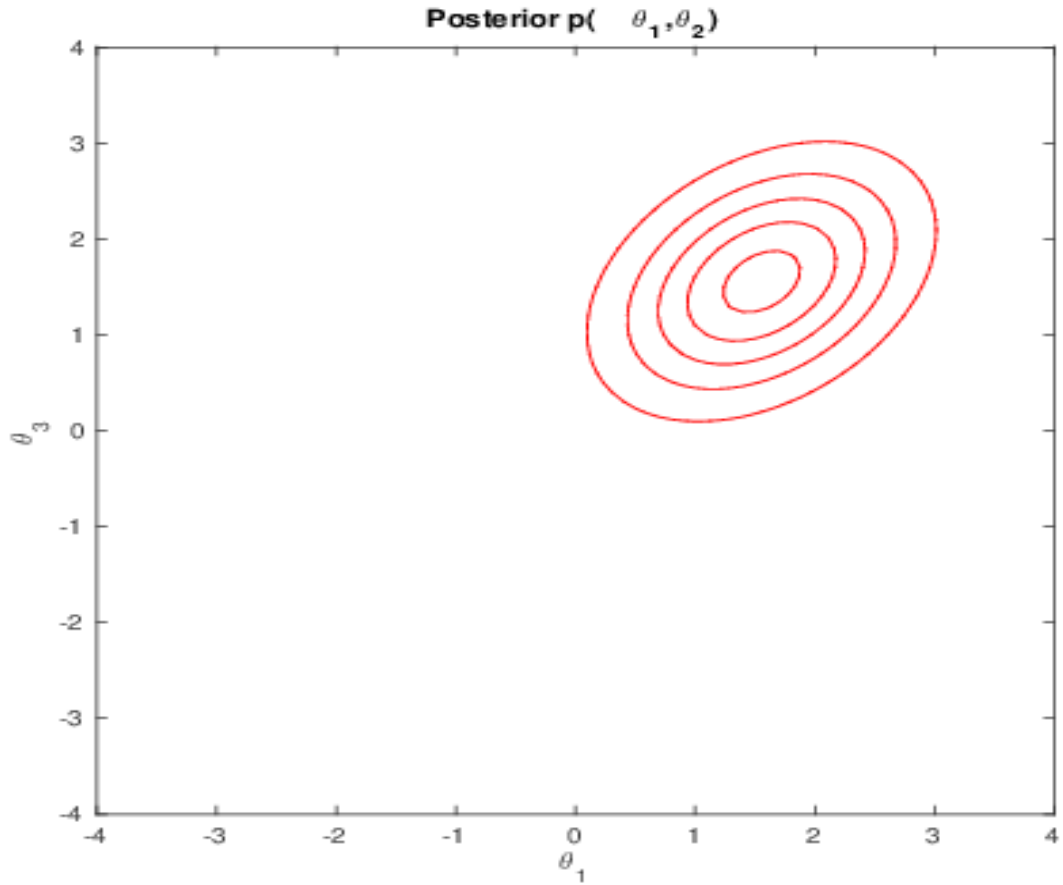
Likelihood: $\mathcal{L}(\theta_1, \theta_2) = p(\mathbf{d}|\theta_1, \theta_2)$



# Bayesian inference: simple one parameter case

$$p(\theta_1, \theta_2), \mathcal{L}(\theta_1, \theta_2) = p(\mathbf{d}|\theta_1, \theta_2) \longrightarrow p(\theta_1, \theta_2|\mathbf{d}) \propto \mathcal{L}(\theta_1, \theta_2)\, p(\theta_1, \theta_2)$$

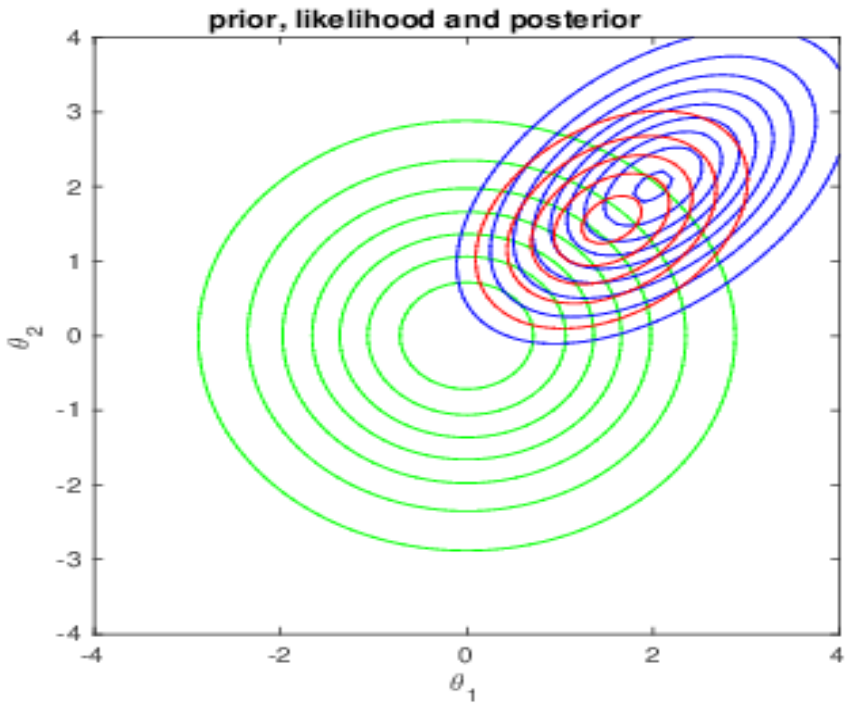
Posterior: $p(\theta_1, \theta_2|\mathbf{d}) \propto p(\mathbf{d}|\theta_1, \theta_2)\, p(\theta_1, \theta_2)$



# Bayesian inference: simple one parameter case

$$p(\theta_1, \theta_2), \mathcal{L}(\theta_1, \theta_2) = p(\mathbf{d}|\theta_1, \theta_2) \longrightarrow p(\theta_1, \theta_2|\mathbf{d}) \propto \mathcal{L}(\theta_1, \theta_2)\, p(\theta_1, \theta_2)$$

Prior, Likelihood and Posterior:



# Bayes: 1D case

$$p(\theta|\mathbf{d}) = \frac{p(\mathbf{d}|\theta)\, p(\theta)}{p(\mathbf{d})} \propto p(\mathbf{d}|\theta)\, p(\theta)$$

- Maximum A Posteriori (MAP)

$$\widehat{\theta} = \arg\max_{\theta} \{p(\theta|\mathbf{d})\} = \arg\max_{\theta} \{p(\mathbf{d}|\theta)\, p(\theta)\}$$

- Posterior Mean

$$\widehat{\theta} = \mathrm{E}_{p(\theta|\mathbf{d})} \{\theta\} = \int \theta p(\theta|\mathbf{d})\, \mathrm{d}\theta$$

- Region of high probabilities

$$[\widehat{\theta}_1, \widehat{\theta}_2] : \int_{\widehat{\theta}_1}^{\widehat{\theta}_2} p(\theta|\mathbf{d})\, \mathrm{d}\theta = 1 - \alpha$$

- Sampling and exploring

$$\theta \sim p(\theta|\mathbf{d})$$

# Bayesian inference: great dimensional case

- Simple Linear case: $\mathbf{d} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\epsilon}$
- Gaussian priors:

$$p(\mathbf{d}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{d}|\mathbf{H}\boldsymbol{\theta}, v_\epsilon \mathbf{I})$$
$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|0, v_\theta \mathbf{I})$$

- Gaussian posterior:

$$p(\boldsymbol{\theta}|\mathbf{d}) = \mathcal{N}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{V}})$$
$$\widehat{\boldsymbol{\theta}} = [\mathbf{H}'\mathbf{H} + \lambda \mathbf{I}]^{-1}\mathbf{H}'\mathbf{d}, \quad \lambda = \frac{v_\epsilon}{v_\theta}$$
$$\widehat{\mathbf{V}} = [\mathbf{H}'\mathbf{H} + \lambda \mathbf{I}]^{-1}$$

- Computation of $\widehat{\boldsymbol{\theta}}$ can be done via optimization of: $J(\boldsymbol{\theta}) = -\ln p(\boldsymbol{\theta}|\mathbf{d}) = \frac{1}{2v_\epsilon}\|\mathbf{d} - \mathbf{H}\boldsymbol{\theta}\|^2 + \frac{1}{2v_\theta}\|\boldsymbol{\theta}\|^2 + c$
- Computation of $\widehat{\mathbf{V}} = [\mathbf{H}'\mathbf{H} + \lambda \mathbf{I}]^{-1}$ needs great dimensional matrix inversion.

# Bayesian inference: great dimensional case

- Gaussian posterior:

$$p(\boldsymbol{\theta}|\mathbf{d}) = \mathcal{N}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{V}}),$$
$$\widehat{\boldsymbol{\theta}} = [\mathbf{H}'\mathbf{H} + \lambda\mathbf{I}]^{-1}\mathbf{H}'\mathbf{d}, \quad \widehat{\mathbf{V}} = [\mathbf{H}'\mathbf{H} + \lambda\mathbf{I}]^{-1}, \ \lambda = \frac{v_\epsilon}{v_\theta}$$

- Computation of $\widehat{\boldsymbol{\theta}}$ can be done via optimization of:
$$J(\boldsymbol{\theta}) = -\ln p(\boldsymbol{\theta}|\mathbf{d}) = c + \|\mathbf{d} - \mathbf{H}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$$
- Gradient based methods:
$\nabla J(\boldsymbol{\theta}) = -2\mathbf{H}'(\mathbf{d} - \mathbf{H}\boldsymbol{\theta}) + 2\lambda\boldsymbol{\theta}$
- constant step, Steepest descend, ...

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \alpha^{(k)}\nabla J(\boldsymbol{\theta}^{(k)}) = \boldsymbol{\theta}^{(k)} + 2\alpha^{(k)}\left[\mathbf{H}'(\mathbf{d} - \mathbf{H}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}\right]$$

- Conjugate Gradient, ...
- At each iteration, we need to be able to compute:
  - Forward operation: $\widehat{\mathbf{d}} = \mathbf{H}\boldsymbol{\theta}^{(k)}$
  - Backward (Adjoint) operation: $\mathbf{H}^t(\mathbf{d} - \widehat{\mathbf{d}})$

# Bayesian inference: great dimensional case

- Computation of $\widehat{\mathbf{V}} = [\mathbf{H}'\mathbf{H} + \lambda\mathbf{I}]^{-1}$ needs great dimensional matrix inversion.
- Almost impossible except in particular cases of Toeplitz, Circulante, TBT, CBC,... where we can diagonalize it via Fast Fourier Transform (FFT).
- Recursive use of the data and recursive update of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\mathbf{V}}$ leads to Kalman Filtering which are still computationally demanding for High dimensional data.
- We also need to generate samples from this posterior: There are many special sampling tools.
- Mainly two categories: Using the covariance matrix $\mathbf{V}$ or its inverse (Precision matrix) $\mathbf{\Lambda} = \mathbf{V}^{-1}$

# Bayesian inference: non Gaussian priors case

- Linear forward model: $\mathbf{d} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\epsilon}$
- Gaussian noise model:

$$p(\mathbf{d}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{d}|\mathbf{H}\boldsymbol{\theta}, v_\epsilon \mathbf{I}) \propto \exp\left[-\frac{1}{2v_\epsilon}\|\mathbf{d} - \mathbf{H}\boldsymbol{\theta}\|_2^2\right]$$

- Sparsity enforcing prior:

$$p(\boldsymbol{\theta}) \propto \exp\left[\alpha\|\boldsymbol{\theta}\|_1\right]$$

- Posterior:

$$p(\boldsymbol{\theta}|\mathbf{d}) \propto \exp\left[-\frac{1}{2v_\epsilon}J(\boldsymbol{\theta})\right] \text{ with } J(\boldsymbol{\theta}) = \|\mathbf{d}-\mathbf{H}\boldsymbol{\theta}\|_2^2+\lambda\|\boldsymbol{\theta}\|_1,\ \lambda = 2v_\epsilon\alpha$$

- Computation of $\widehat{\boldsymbol{\theta}}$ can be done via optimization of $J(\boldsymbol{\theta})$
- Other computations are much more difficult.

# Bayes Rule for Machine Learning (Simple case)

- Inference on the parameters: Learning from data $\mathbf{d}$:

$$p(\boldsymbol{\theta}|\mathbf{d}, \mathcal{M}) = \frac{p(\mathbf{d}|\boldsymbol{\theta}, \mathcal{M})\, p(\boldsymbol{\theta}|\mathcal{M})}{p(\mathbf{d}|\mathcal{M})}$$

- Model Comparison:

$$p(\mathcal{M}_k|\mathbf{d}) = \frac{p(\mathbf{d}|\mathcal{M}_k)\, p(\mathcal{M}_k)}{p(\mathbf{d})}$$

with

$$p(\mathbf{d}|\mathcal{M}_k) = \int p(\mathbf{d}|\boldsymbol{\theta}, \mathcal{M}_k)\, p(\boldsymbol{\theta}|\mathcal{M})\, \mathrm{d}\boldsymbol{\theta}$$

- Prediction with selected model:

$$p(z|\mathcal{M}_k) = \int p(z|\boldsymbol{\theta}, \mathcal{M}_k) p(\boldsymbol{\theta}|\mathbf{d}, \mathcal{M}_k)\, \mathrm{d}\boldsymbol{\theta}$$

# Approximation methods

- Laplace approximation
- Bayesian Information Criterion (BIC)
- Variational Bayesian Approximations (VBA)
- Expectation Propagation (EP)
- Markov chain Monte Carlo methods (MCMC)
- Exact Sampling

# Laplace Approximation

- Data set $\mathbf{d}$, models $\mathcal{M}_1, \cdots, \mathcal{M}_K$, parameters $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K$
- Model Comparison:

$$p(\boldsymbol{\theta}, \mathbf{d}|\mathcal{M}) = p(\mathbf{d}|\boldsymbol{\theta}, \mathcal{M})\, p(\boldsymbol{\theta}|\mathcal{M})$$
$$p(\boldsymbol{\theta}|\mathbf{d}, \mathcal{M}) = p(\boldsymbol{\theta}, \mathbf{d}|\mathcal{M})/p(\mathbf{d}|\mathcal{M})$$
$$p(\mathbf{d}|\mathcal{M}) = \int p(\mathbf{d}|\boldsymbol{\theta}, \mathcal{M})\, p(\boldsymbol{\theta}|\mathcal{M})\, \mathrm{d}\boldsymbol{\theta}$$

- For large amount of data (relative to number of parameters, $m$), $p(\theta|\mathbf{d}, \mathcal{M})$ is approximated by a Gaussian around its maximum (MAP) $\widehat{\boldsymbol{\theta}}$:

$$p(\boldsymbol{\theta}|\mathbf{d}, \mathcal{M}) \approx (2\pi)^{-m/2} |\mathbf{A}|^{1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})' \mathbf{A} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\right]$$

where $A_{i,j} = \frac{d^2}{\theta_i \theta_j} \ln p(\boldsymbol{\theta}|\mathbf{d}, \mathcal{M})$ is the $m \times m$ Hessian matrix.

- $p(\mathbf{d}|\mathcal{M}) = p(\boldsymbol{\theta}, \mathbf{d}|\mathcal{M})/p(\boldsymbol{\theta}|\mathbf{d}, \mathcal{M})$ and evaluating it at $\widehat{\boldsymbol{\theta}}$:

$$\ln p(\mathbf{d}|\mathcal{M}_k) \approx \ln p(\mathbf{d}|\widehat{\boldsymbol{\theta}}, \mathcal{M}_k) + \ln p(\widehat{\boldsymbol{\theta}}|\mathcal{M}_k) + \frac{m}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{A}|$$

- Needs computation of $\widehat{\boldsymbol{\theta}}$ and $|\mathbf{A}|$.

# Bayesian Information Criteria (BIC)

- BIC is obtained from the Laplace approximation

$$\ln p(\mathbf{d}|\mathcal{M}_k) \approx \ln p(\widehat{\boldsymbol{\theta}}|\mathcal{M}_k) + p(\mathbf{d}|\widehat{\boldsymbol{\theta}}, \mathcal{M}_k) + \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}|$$

  by taking the large sample limit ($n \mapsto \infty$) where $n$ is the number of data points:

$$\ln p(\mathbf{d}|\mathcal{M}_k) \approx p(\mathbf{d}|\widehat{\boldsymbol{\theta}}, \mathcal{M}_k) - \frac{d}{2}\ln(n)$$

- Easy to compute
- It does not depend on the prior
- It is equivalent to MDL criterion
- Assumes that as ($n \mapsto \infty$), all the parameters are identifiable.
- Danger: counting parameters can be deceiving (sinusoid, infinite dim models)

# Bayes Rule for Machine Learning with hidden variables

- Data: $\mathbf{d}$, Hidden Variables: $\mathbf{x}$, Parameters: $\boldsymbol{\theta}$, Model: $\mathcal{M}$
- Bayes rule

$$p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{d}, \mathcal{M}) = \frac{p(\mathbf{d}|\mathbf{x}, \boldsymbol{\theta}, \mathcal{M})\, p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M}))p(\boldsymbol{\theta}|\mathcal{M})}{p(\mathbf{d}|\mathcal{M})}$$

- Model Comparison

$$p(\mathcal{M}_k|\mathbf{d}) = \frac{p(\mathbf{d}|\mathcal{M}_k)\, p(\mathcal{M}_k)}{p(\mathbf{d})}$$

with

$$p(\mathbf{d}|\mathcal{M}_k) = \int\int p(\mathbf{d}|\mathbf{x}, \boldsymbol{\theta}, \mathcal{M}_k)\, p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M}))p(\boldsymbol{\theta}|\mathcal{M})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\boldsymbol{\theta}$$

- Prediction with a new data $z$

$$p(z|\mathcal{M}) = \int\int p(z|\mathbf{x}, \boldsymbol{\theta}, \mathcal{M})p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M}))\, \mathrm{d}\mathbf{x}\, \mathrm{d}\boldsymbol{\theta}$$

# Lower Bounding the Marginal Likelihood

Jensen's inequality:

$$\begin{aligned}
\ln p(\mathbf{d}|\mathcal{M}_k) &= \ln \int\int p(\mathbf{d},\mathbf{x},\boldsymbol{\theta}|\mathcal{M}_k)\,\mathrm{d}\mathbf{x}\,\mathrm{d}\boldsymbol{\theta} \\
&= \ln \int\int q(\mathbf{x},\boldsymbol{\theta})\frac{p(\mathbf{d},\mathbf{x},\boldsymbol{\theta}|\mathcal{M}_k)}{q(\mathbf{x},\boldsymbol{\theta})}\,\mathrm{d}\mathbf{x}\,\mathrm{d}\boldsymbol{\theta} \\
&\geq \int\int q(\mathbf{x},\boldsymbol{\theta})\ln\frac{p(\mathbf{d},\mathbf{x},\boldsymbol{\theta}|\mathcal{M}_k)}{q(\mathbf{x},\boldsymbol{\theta})}\,\mathrm{d}\mathbf{x}\,\mathrm{d}\boldsymbol{\theta}
\end{aligned}$$

Using a factorised approximation for $q(\mathbf{x},\boldsymbol{\theta}) = q_1(\mathbf{x})q_2(\boldsymbol{\theta})$:

$$\begin{aligned}
\ln p(\mathbf{d}|\mathcal{M}_k) &\geq \int\int q_1(\mathbf{x})q_2(\boldsymbol{\theta})\ln\frac{p(\mathbf{d},\mathbf{x},\boldsymbol{\theta}|\mathcal{M}_k)}{q_1(\mathbf{x})q_2(\boldsymbol{\theta})}\,\mathrm{d}\mathbf{x}\,\mathrm{d}\boldsymbol{\theta} \\
&= \mathcal{F}_{\mathcal{M}_k}(q_1(\mathbf{x}), q_2(\boldsymbol{\theta}), \mathbf{d})
\end{aligned}$$

Maximising this free energy leads to VBA.

# Variational Bayesian Learning

$$\begin{aligned}\mathcal{F}_{\mathcal{M}}(q_1(\mathbf{x}), q_2(\boldsymbol{\theta}), \mathbf{d}) &= \int\int q_1(\mathbf{x}) q_2(\boldsymbol{\theta}) \ln \frac{p(\mathbf{d}, \mathbf{x}, \boldsymbol{\theta}|\mathcal{M})}{q_1(\mathbf{x}) q_2(\boldsymbol{\theta})} \, \mathrm{d}\mathbf{x} \, \mathrm{d}\boldsymbol{\theta} \\ &= \mathcal{H}(q_1) + \mathcal{H}(q_2) + \langle \ln p(\mathbf{d}, \mathbf{x}, \boldsymbol{\theta}|\mathcal{M}) \rangle_{q_1 q_2}\end{aligned}$$

Minimising this lower bound with respect to $q_1$ and then $q_2$ leads to EM-like iterative update

$$\begin{aligned} q_1^{(t+1)}(\mathbf{x}) &\propto \exp\left[\langle \ln p(\mathbf{d}, \mathbf{x}, \boldsymbol{\theta}|\mathcal{M}) \rangle_{q_2^{(t)}(\boldsymbol{\theta})}\right] \quad \text{E-like step} \\ q_2^{(t+1)}(\boldsymbol{\theta}) &\propto \exp\left[\langle \ln p(\mathbf{d}, \mathbf{x}, \boldsymbol{\theta}|\mathcal{M}) \rangle_{q_1^{(t+1)}(\mathbf{x})}\right] \quad \text{M-like step}\end{aligned}$$

which can also be written as:

$$\begin{aligned} q_1^{(t+1)}(\mathbf{x}) &\propto \exp\left[\langle \ln p(\mathbf{d}, \mathbf{x}|\boldsymbol{\theta}, \mathcal{M}) \rangle_{q_2^{(t)}(\boldsymbol{\theta})}\right] \quad \text{E-like step} \\ q_2^{(t+1)}(\boldsymbol{\theta}) &\propto p(\boldsymbol{\theta}|\mathcal{M}) \exp\left[\langle \ln p(\mathbf{d}, \mathbf{x}|\boldsymbol{\theta}, \mathcal{M}) \rangle_{q_1^{(t+1)}(\mathbf{x})}\right] \quad \text{M-like step}\end{aligned}$$

# EM and VBEM algorithms

**EM for Marginal MAP estimation**

Goal: maximize $p(\boldsymbol{\theta}|\mathbf{d},\mathcal{M})$ w.r.t. $\boldsymbol{\theta}$

**E Step:** Compute

$q_1^{(t+1)}(\mathbf{x}) = p(\mathbf{x}|\mathbf{d},\boldsymbol{\theta}^{(t)})$ and

$Q(\boldsymbol{\theta}) = \langle \ln p(\mathbf{d},\mathbf{x},\boldsymbol{\theta}|\mathcal{M})\rangle_{q_1^{(t+1)}(\mathbf{x})}$

**M Step:** Maximize

$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \{Q(\boldsymbol{\theta})\}$

**Variational Bayesian EM**

Goal: lower bound $p(\mathbf{d}|\mathcal{M})$

**VB-E Step:** Compute

$q_1^{(t+1)}(\mathbf{x}) = p(\mathbf{x}|\mathbf{d},\boldsymbol{\phi}^{(t)})$ and

$Q(\boldsymbol{\theta}) = \langle \ln p(\mathbf{d},\mathbf{x},\boldsymbol{\theta}|\mathcal{M})\rangle_{q_1^{(t+1)}(\mathbf{x})}$

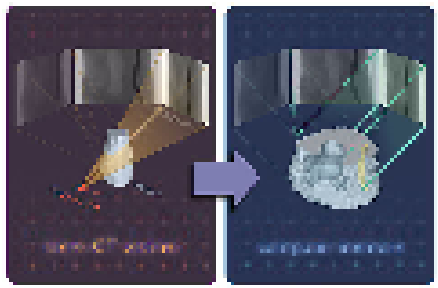
**M Step:** Maximize

$q_2^{(t+1)}(\boldsymbol{\theta}) = \exp\left[Q(\boldsymbol{\theta})\right]$

Properties:

- VB-EM reduces to EM if $q_2(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})$
- VB-EM has the same complexity than EM
- If we choose $q_2(\boldsymbol{\theta})$ in the conjugate family of $p(\mathbf{d},\mathbf{x}|\boldsymbol{\theta})$, then $\boldsymbol{\phi}$ becomes the expected natural parameters
- The main computational part of both methods is in the E-step. We can use belief propagation, Kalman filter, etc. to do it. In VB-EM, $\boldsymbol{\phi}$ replaces $\boldsymbol{\theta}$.

# Computed Tomography: Seeing inside of a body



- $f(x, y)$ a section of a real 3D body $f(x, y, z)$
- $g_\phi(r)$ a line of observed radiography $g_\phi(r, z)$
- Forward model:
  Line integrals or Radon Transform

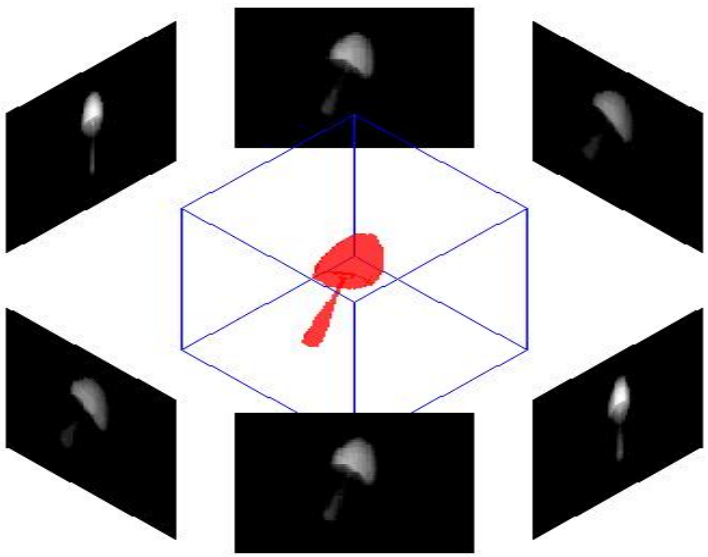
$$
\begin{aligned}
g_\phi(r) &= \int_{L_{r,\phi}} f(x, y)\, \mathrm{d}l + \epsilon_\phi(r) \\
&= \iint f(x, y)\, \delta(r - x\cos\phi - y\sin\phi)\, \mathrm{d}x\, \mathrm{d}y + \epsilon_\phi(r)
\end{aligned}
$$

- Inverse problem: Image reconstruction

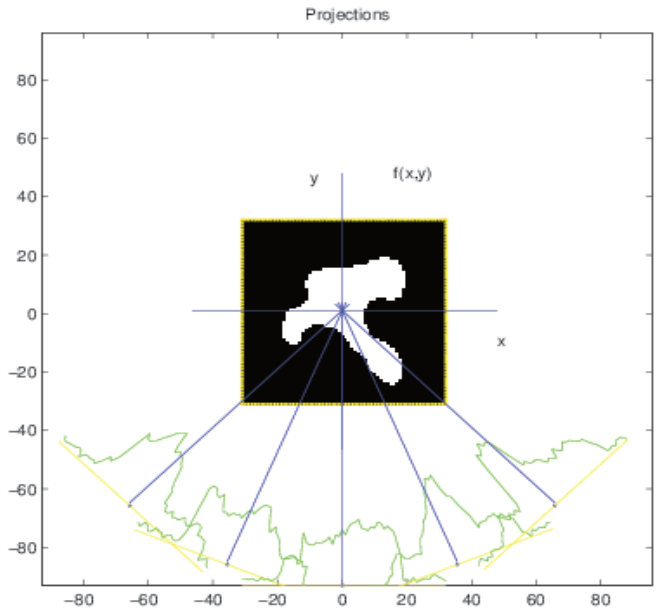
  Given the forward model $\mathcal{H}$ (Radon Transform) and
  a set of data $g_{\phi_i}(r), i = 1, \cdots, M$
  find $f(x, y)$

# 2D and 3D Computed Tomography

3D



2D

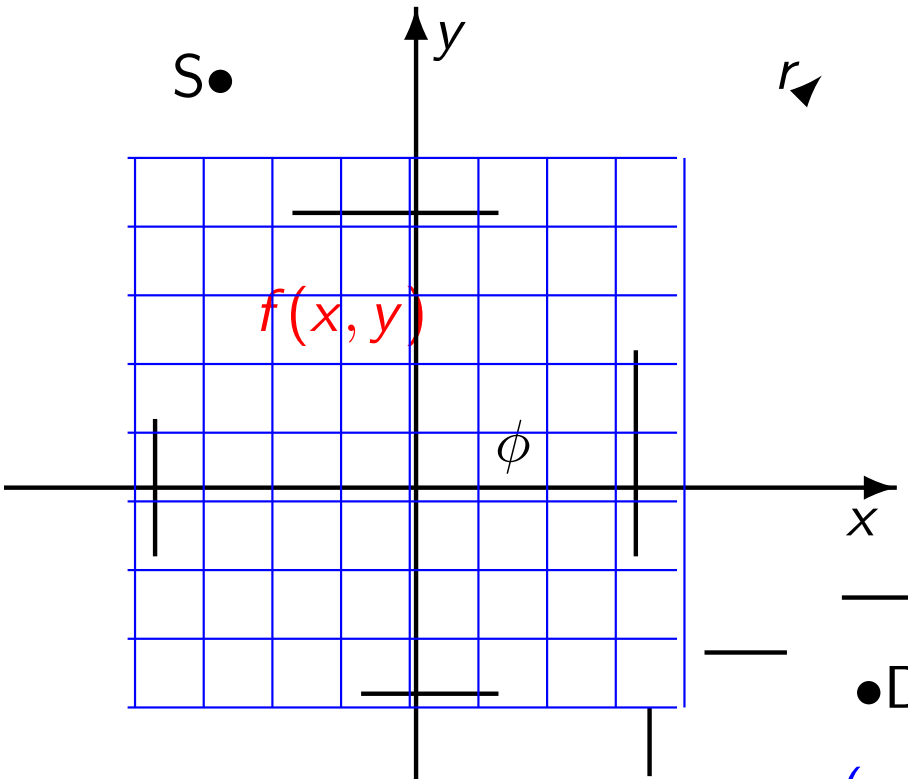


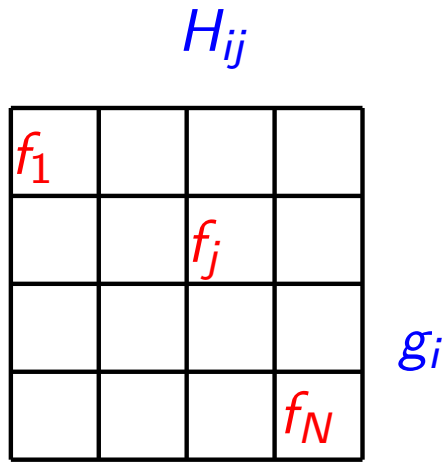

$$g_\phi(r_1, r_2) = \int_{\mathcal{L}_{r_1,r_2,\phi}} f(x, y, z)\, \mathrm{d}l \qquad g_\phi(r) = \int_{\mathcal{L}_{r,\phi}} f(x, y)\, \mathrm{d}l$$

Forward probelm: $f(x, y)$ or $f(x, y, z) \longrightarrow g_\phi(r)$ or $g_\phi(r_1, r_2)$

Inverse problem: $g_\phi(r)$ or $g_\phi(r_1, r_2) \longrightarrow f(x, y)$ or $f(x, y, z)$

# Algebraic methods: Discretization




$$g(r,\phi) = \int_L f(x,y)\, dl$$

$$f(x,y) = \sum_j f_j\, b_j(x,y)$$

$$b_j(x,y) = \begin{cases} 1 & \text{if } (x,y) \in \text{ pixel } j \\ 0 & \text{else} \end{cases}$$

$$g_i = \sum_{j=1}^{N} H_{ij}\, f_j + \epsilon_i \rightarrow \mathbf{g} = \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon}$$

- **H** is huge dimensional: 2D: $10^6 \times 10^6$, 3D: $10^9 \times 10^9$.
- **Hf** corresponds to forward projection
- $\mathbf{H}^t\mathbf{g}$ corresponds to Back projection (BP)

# Bayesian approach for linear inverse problems

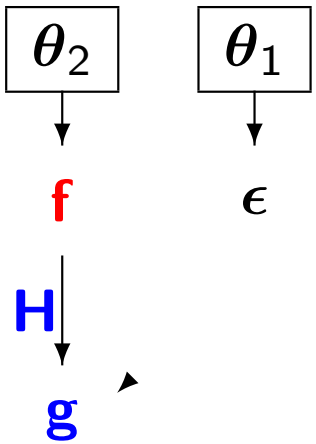
$$\mathcal{M}: \qquad \mathbf{g} = \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon}$$

- Observation model $\mathcal{M}$ + Information on the noise $\boldsymbol{\epsilon}$:
  $$p(\mathbf{g}|\mathbf{f}, \theta_1; \mathcal{M}) = p_{\boldsymbol{\epsilon}}(\mathbf{g} - \mathbf{H}\mathbf{f}|\theta_1)$$
- A priori information $p(\mathbf{f}|\theta_2; \mathcal{M})$
- Basic Bayes :
  $$p(\mathbf{f}|\mathbf{g}, \theta_1, \theta_2; \mathcal{M}) = \frac{p(\mathbf{g}|\mathbf{f}, \theta_1; \mathcal{M})\, p(\mathbf{f}|\theta_2; \mathcal{M})}{p(\mathbf{g}|\theta_1, \theta_2; \mathcal{M})}$$
- Unsupervised:
  $$p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{g}, \boldsymbol{\alpha}_0) = \frac{p(\mathbf{g}|\mathbf{f}, \theta_1)\, p(\mathbf{f}|\theta_2)\, p(\boldsymbol{\theta}|\boldsymbol{\alpha}_0)}{p(\mathbf{g}|\boldsymbol{\alpha}_0)}, \quad \boldsymbol{\theta} = (\theta_1, \theta_2)$$
- Hierarchical prior models:

$$p(\mathbf{f}, \mathbf{z}, \boldsymbol{\theta}|\mathbf{g}, \boldsymbol{\alpha}_0) = \frac{p(\mathbf{g}|\mathbf{f}, \theta_1)\, p(\mathbf{f}|\mathbf{z}, \theta_2)\, p(\mathbf{z}|\theta_3)\, p(\boldsymbol{\theta}|\boldsymbol{\alpha}_0)}{p(\mathbf{g}|\boldsymbol{\alpha}_0)}, \quad \boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$$

# Bayesian inference for inverse problems

Simple case:




$$\mathbf{g} = \mathbf{H}\mathbf{f} + \epsilon$$

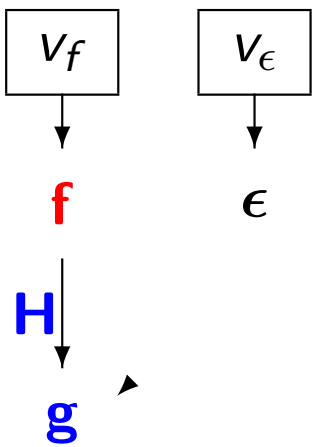
$$p(\mathbf{f}|\mathbf{g}, \boldsymbol{\theta}) \propto p(\mathbf{g}|\mathbf{f}, \boldsymbol{\theta}_1)\, p(\mathbf{f}|\boldsymbol{\theta}_2)$$

– Objective: Infer $\mathbf{f}$

– MAP: $\widehat{\mathbf{f}} = \arg\max_{\mathbf{f}} \{p(\mathbf{f}|\mathbf{g}, \boldsymbol{\theta})\}$

– Posterior Mean (PM): $\widehat{\mathbf{f}} = \int \mathbf{f}\, p(\mathbf{f}|\mathbf{g}, \boldsymbol{\theta})\, \mathrm{d}\mathbf{f}$

Example: Caussian case:




$$\begin{cases} p(\mathbf{g}|\mathbf{f}, v_\epsilon) = \mathcal{N}(\mathbf{g}|\mathbf{H}\mathbf{f}, v_\epsilon \mathbf{I}) \\ p(\mathbf{f}|v_f) = \mathcal{N}(\mathbf{f}|0, v_f \mathbf{I}) \end{cases} \rightarrow p(\mathbf{f}|\mathbf{g}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\widehat{\mathbf{f}}, \widehat{\boldsymbol{\Sigma}})$$

– MAP: $\widehat{\mathbf{f}} = \arg\min_{\mathbf{f}} \{J(\mathbf{f})\}$ with

$$J(\mathbf{f}) = \frac{1}{v_\epsilon}\|\mathbf{g} - \mathbf{H}\mathbf{f}\|^2 + \frac{1}{v_f}\|\mathbf{f}\|^2$$

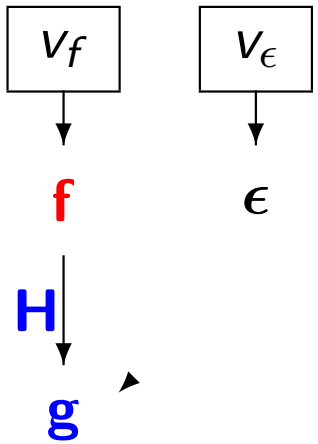
– Posterior Mean (PM)=MAP:

$$\begin{cases} \widehat{\mathbf{f}} = (\mathbf{H}^t\mathbf{H} + \lambda\mathbf{I})^{-1}\mathbf{H}^t\mathbf{g} \\ \widehat{\boldsymbol{\Sigma}} = (\mathbf{H}^t\mathbf{H} + \lambda\mathbf{I})^{-1} \end{cases} \quad \text{with } \lambda = \frac{v_\epsilon}{v_f}.$$

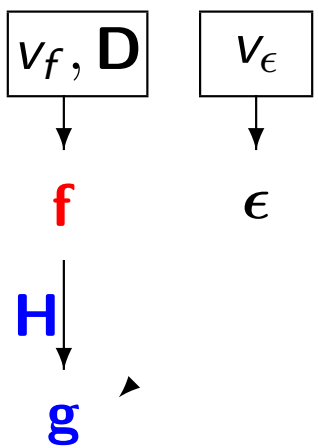
# Gaussian model: Simple separable and Markovian

$$\mathbf{g} = \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon}$$

Separable Gaussian




Gauss-Markov




$$\mathbf{g} = \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon}$$

$$\begin{cases} p(\mathbf{g}|\mathbf{f}, \theta_1) = \mathcal{N}(\mathbf{g}|\mathbf{H}\mathbf{f}, v_\epsilon \mathbf{I}) \\ p(\mathbf{f}|v_f) = \mathcal{N}(\mathbf{f}|0, v_f \mathbf{I}) \end{cases} \rightarrow p(\mathbf{f}|\mathbf{g}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\widehat{\mathbf{f}}, \widehat{\boldsymbol{\Sigma}})$$

– MAP: $\widehat{\mathbf{f}} = \arg\min_{\mathbf{f}} \{J(\mathbf{f})\}$ with

$J(\mathbf{f}) = \frac{1}{v_\epsilon}\|\mathbf{g} - |\mathbf{H}\mathbf{f}\|^2 + \frac{1}{v_f}\|\mathbf{f}\|^2$

– Posterior Mean (PM)=MAP:

$$\begin{cases} \widehat{\mathbf{f}} = (\mathbf{H}^t\mathbf{H} + \lambda \mathbf{I}]^{-1}\mathbf{H}^t\mathbf{g} \\ \widehat{\boldsymbol{\Sigma}} = v_\epsilon(\mathbf{H}^t\mathbf{H} + \lambda \mathbf{I}]^{-1} \end{cases} \quad \text{with } \lambda = \frac{v_\epsilon}{v_f}.$$

Markovian case:

$p(\mathbf{f}|v_f, \mathbf{D}) = \mathcal{N}(\mathbf{f}|0, v_f(\mathbf{D}\mathbf{D}^t)^{-1})$

– MAP: $J(\mathbf{f}) = \frac{1}{v_\epsilon}\|\mathbf{g} - |\mathbf{H}\mathbf{f}\|^2 + \frac{1}{v_f}\|\mathbf{D}\mathbf{f}\|^2$

– Posterior Mean (PM)=MAP:

$$\begin{cases} \widehat{\mathbf{f}} = (\mathbf{H}^t\mathbf{H} + \lambda \mathbf{D}^t\mathbf{D}]^{-1}\mathbf{H}^t\mathbf{g} \\ \widehat{\boldsymbol{\Sigma}} = v_\epsilon(\mathbf{H}^t\mathbf{H} + \lambda \mathbf{D}^t\mathbf{D}]^{-1} \end{cases} \quad \text{with } \lambda = \frac{v_e}{v_f}.$$

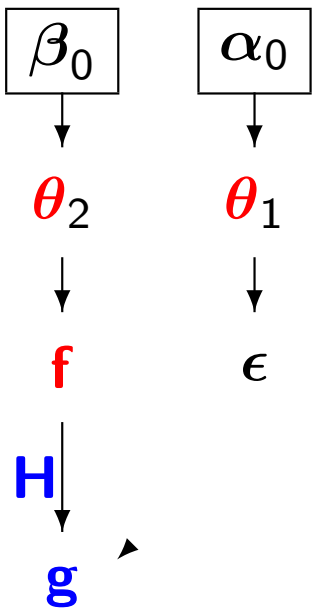
# Bayesian inference (Unsupervised case)

Unsupervised case: Hyper parameter estimation

$$p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{g}) \propto p(\mathbf{g}|\mathbf{f}, \boldsymbol{\theta}_1)\, p(\mathbf{f}|\boldsymbol{\theta}_2)\, p(\boldsymbol{\theta})$$

– Objective: Infer $(\mathbf{f}, \boldsymbol{\theta})$

JMAP: $(\widehat{\mathbf{f}}, \widehat{\boldsymbol{\theta}}) = \arg\max_{(\mathbf{f}, \boldsymbol{\theta})} \{p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{g})\}$




– Marginalization 1:

$$p(\mathbf{f}|\mathbf{g}) = \int p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{g})\, \mathrm{d}\boldsymbol{\theta}$$

– Marginalization 2:

$$p(\boldsymbol{\theta}|\mathbf{g}) = \int p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{g})\, \mathrm{d}\mathbf{f} \text{ followed by:}$$

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \{p(\boldsymbol{\theta}|\mathbf{g})\} \rightarrow \widehat{\mathbf{f}} = \arg\max_{\mathbf{f}} \left\{p(\mathbf{f}|\mathbf{g}, \widehat{\boldsymbol{\theta}})\right\}$$

– MCMC Gibbs sampling:

$\mathbf{f} \sim p(\mathbf{f}|\boldsymbol{\theta}, \mathbf{g}) \rightarrow \boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\mathbf{f}, \mathbf{g})$ until convergence

Use samples generated to compute mean and variances

– VBA: Approximate $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{g})$ by $q_1(\mathbf{f})\, q_2(\boldsymbol{\theta})$

Use $q_1(\mathbf{f})$ to infer $\mathbf{f}$ and $q_2(\boldsymbol{\theta})$ to infer $\boldsymbol{\theta}$

# JMAP, Marginalization, VBA

- JMAP:

$$\boxed{\begin{array}{c} p(\mathbf{f},\boldsymbol{\theta}|\mathbf{g}) \\ \text{optimization} \end{array}} \longrightarrow \widehat{\mathbf{f}}, \quad \longrightarrow \widehat{\boldsymbol{\theta}}$$

- Marginalization

$$\boxed{p(\mathbf{f},\boldsymbol{\theta}|\mathbf{g})} \longrightarrow \boxed{p(\boldsymbol{\theta}|\mathbf{g})} \longrightarrow \widehat{\boldsymbol{\theta}} \longrightarrow \boxed{p(\mathbf{f}|\widehat{\boldsymbol{\theta}},\mathbf{g})} \longrightarrow \widehat{\mathbf{f}}$$

Joint Posterior  Marginalize over $\mathbf{f}$

- Variational Bayesian Approximation

$$\boxed{p(\mathbf{f},\boldsymbol{\theta}|\mathbf{g})} \longrightarrow \boxed{\begin{array}{c} \text{Variational} \\ \text{Bayesian} \\ \text{Approximation} \end{array}} \begin{array}{l} \longrightarrow q_1(\mathbf{f}) \longrightarrow \widehat{\mathbf{f}} \\ \longrightarrow q_2(\boldsymbol{\theta}) \longrightarrow \widehat{\boldsymbol{\theta}} \end{array}$$

# Variational Bayesian Approximation

- Approximate $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{g})$ by $q(\mathbf{f}, \boldsymbol{\theta}) = q_1(\mathbf{f})\, q_2(\boldsymbol{\theta})$
  and then use them for any inferences on $\mathbf{f}$ and $\boldsymbol{\theta}$ respectively.
- Criterion $\mathrm{KL}(q(\mathbf{f}, \boldsymbol{\theta}|\mathbf{g}) : p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{g}))$
  $\mathrm{KL}(q : p) = \int\int q \ln \frac{q}{p} = \int\int q_1 q_2 \ln \frac{q_1 q_2}{p}$
- Iterative algorithm $q_1 \longrightarrow q_2 \longrightarrow q_1 \longrightarrow q_2, \cdots$

$$
\begin{cases}
\widehat{q}_1(\mathbf{f}) & \propto \exp\left[\langle \ln p(\mathbf{g}, \mathbf{f}, \boldsymbol{\theta}; \mathcal{M})\rangle_{\widehat{q}_2(\boldsymbol{\theta})}\right] \\
\widehat{q}_2(\boldsymbol{\theta}) & \propto \exp\left[\langle \ln p(\mathbf{g}, \mathbf{f}, \boldsymbol{\theta}; \mathcal{M})\rangle_{\widehat{q}_1(\mathbf{f})}\right]
\end{cases}
$$

$$
\boxed{p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{g})} \longrightarrow \boxed{\begin{array}{c}\text{Variational} \\ \text{Bayesian} \\ \text{Approximation}\end{array}} \begin{array}{l} \longrightarrow q_1(\mathbf{f}) \longrightarrow \widehat{\mathbf{f}} \\ \\ \longrightarrow q_2(\boldsymbol{\theta}) \longrightarrow \widehat{\boldsymbol{\theta}} \end{array}
$$

# Variational Bayesian Approximation

$$p(\mathbf{g}, \mathbf{f}, \boldsymbol{\theta}|\mathcal{M}) = p(\mathbf{g}|\mathbf{f}, \boldsymbol{\theta}, \mathcal{M})\, p(\mathbf{f}|\boldsymbol{\theta}, \mathcal{M})\, p(\boldsymbol{\theta}|\mathcal{M})$$

$$p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{g}, \mathcal{M}) = \frac{p(\mathbf{g}, \mathbf{f}, \boldsymbol{\theta}|\mathcal{M})}{p(\mathbf{g}|\mathcal{M})}$$

$$\mathrm{KL}(q : p) = \iint q(\mathbf{f}, \boldsymbol{\theta}) \ln \frac{p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{g}; \mathcal{M})}{q(\mathbf{f}, \boldsymbol{\theta})} \, \mathrm{d}\mathbf{f}\, \mathrm{d}\boldsymbol{\theta}$$

$$\begin{aligned} p(\mathbf{g}|\mathcal{M}) &= \iint q(\mathbf{f}, \boldsymbol{\theta}) \frac{p(\mathbf{g}, \mathbf{f}, \boldsymbol{\theta}|\mathcal{M})}{q(\mathbf{f}, \boldsymbol{\theta})} \, \mathrm{d}\mathbf{f}\, \mathrm{d}\boldsymbol{\theta} \\ &\geq \iint q(\mathbf{f}, \boldsymbol{\theta}) \ln \frac{p(\mathbf{g}, \mathbf{f}, \boldsymbol{\theta}|\mathcal{M})}{q(\mathbf{f}, \boldsymbol{\theta})} \, \mathrm{d}\mathbf{f}\, \mathrm{d}\boldsymbol{\theta} \end{aligned}$$

Free energy:

$$\mathcal{F}(q) = \iint q(\mathbf{f}, \boldsymbol{\theta}) \ln \frac{p(\mathbf{g}, \mathbf{f}, \boldsymbol{\theta}|\mathcal{M})}{q(\mathbf{f}, \boldsymbol{\theta})} \, \mathrm{d}\mathbf{f}\, \mathrm{d}\boldsymbol{\theta}$$

Evidence of the model $\mathcal{M}$:

$$p(\mathbf{g}|\mathcal{M}) = \mathcal{F}(q) + \mathrm{KL}(q : p)$$

# VBA: Separable Approximation

$$p(\mathbf{g}|\mathcal{M}) = \mathcal{F}(q) + \mathrm{KL}(q:p)$$

$$q(\mathbf{f},\boldsymbol{\theta}) = q_1(\mathbf{f})\, q_2(\boldsymbol{\theta})$$

Minimizing $\mathrm{KL}(q:p)$ = Maximizing $\mathcal{F}(q)$

$$(\widehat{q}_1,\widehat{q}_2) = \underset{(q_1,q_2)}{\arg\min}\left\{\mathrm{KL}(q_1q_2:p)\right\} = \underset{(q_1,q_2)}{\arg\max}\left\{\mathcal{F}(q_1q_2)\right\}$$

$\mathrm{KL}(q_1q_2:p)$ is convexe wrt $q_1$ when $q_2$ is fixed and vise versa:

$$\left\{\begin{array}{ll}\widehat{q}_1 & = \arg\min_{q_1}\left\{\mathrm{KL}(q_1\widehat{q}_2:p)\right\} = \arg\max_{q_1}\left\{\mathcal{F}(q_1\widehat{q}_2)\right\}\\ \widehat{q}_2 & = \arg\min_{q_2}\left\{\mathrm{KL}(\widehat{q}_1q_2:p)\right\} = \arg\max_{q_2}\left\{\mathcal{F}(\widehat{q}_1q_2)\right\}\end{array}\right.$$

$$\left\{\begin{array}{ll}\widehat{q}_1(\mathbf{f}) & \propto \exp\left[\left\langle \ln p(\mathbf{g},\mathbf{f},\boldsymbol{\theta};\mathcal{M})\right\rangle_{\widehat{q}_2(\boldsymbol{\theta})}\right]\\ \widehat{q}_2(\boldsymbol{\theta}) & \propto \exp\left[\left\langle \ln p(\mathbf{g},\mathbf{f},\boldsymbol{\theta};\mathcal{M})\right\rangle_{\widehat{q}_1(\mathbf{f})}\right]\end{array}\right.$$

# BVA: Choice of family of laws $q_1$ and $q_2$

- Case 1 : $\longrightarrow$ Joint MAP

$$\left\{\begin{array}{ll}\widehat{q}_1(\mathbf{f}|\widetilde{\mathbf{f}}) & = \delta(\mathbf{f}-\widetilde{\mathbf{f}}) \\ \widehat{q}_2(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}}) & = \delta(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})\end{array}\right. \longrightarrow \left\{\begin{array}{l}\widetilde{\mathbf{f}} = \arg\max_{\mathbf{f}}\left\{p(\mathbf{f}, r\widetilde{\boldsymbol{\theta}}|\mathbf{g};\mathcal{M})\right\} \\ \widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}}\left\{p(\widetilde{\mathbf{f}}, \boldsymbol{\theta}|\mathbf{g};\mathcal{M})\right\}\end{array}\right.$$

- Case 2 : $\longrightarrow$ EM

$$\left\{\begin{array}{ll}\widehat{q}_1(\mathbf{f}) & \propto p(\mathbf{f}|\boldsymbol{\theta},\mathbf{g}) \\ \widehat{q}_2(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}}) & = \delta(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})\end{array}\right. \longrightarrow \left\{\begin{array}{l}Q(\boldsymbol{\theta},\widetilde{\boldsymbol{\theta}}) = \langle \ln p(\mathbf{f},\boldsymbol{\theta}|\mathbf{g};\mathcal{M})\rangle_{q_1(\mathbf{f}|\widetilde{\boldsymbol{\theta}})} \\ \widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}}\left\{Q(\boldsymbol{\theta},\widetilde{\boldsymbol{\theta}})\right\}\end{array}\right.$$

- Appropriate choice for inverse problems

$$\left\{\begin{array}{ll}\widehat{q}_1(\mathbf{f}) & \propto p(\mathbf{f}|\widetilde{\boldsymbol{\theta}},\mathbf{g};\mathcal{M}) \\ \widehat{q}_2(\boldsymbol{\theta}) & \propto p(\boldsymbol{\theta}|\mathbf{f},\mathbf{g};\mathcal{M})\end{array}\right. \longrightarrow \left\{\begin{array}{l}\text{Accounts for the uncertainties of} \\ \widehat{\boldsymbol{\theta}} \text{ for } \widehat{\mathbf{f}} \text{ and vice versa.}\end{array}\right.$$

Exponential families, Conjugate priors

# JMAP, EM and VBA

JMAP Alternate optimization Algorithm:

$$\boxed{\widetilde{\mathbf{f}} = \arg\max_{\mathbf{f}} \left\{ p(\mathbf{f}, \widetilde{\boldsymbol{\theta}} | \mathbf{g}) \right\}} \longrightarrow \widetilde{\mathbf{f}} \longrightarrow \widehat{\mathbf{f}}$$

$$\widehat{\boldsymbol{\theta}} \longleftarrow \widetilde{\boldsymbol{\theta}} \longleftarrow \boxed{\widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{ p(\widetilde{\mathbf{f}}, \boldsymbol{\theta} | \mathbf{g}) \right\}} \longleftarrow \widetilde{\mathbf{f}}$$

$$\boldsymbol{\theta}^{(0)} \longrightarrow \widetilde{\boldsymbol{\theta}} \longrightarrow \boxed{\widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{ Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) \right\}} \longleftarrow q_1(\mathbf{f})$$

$$\boxed{q_1(\mathbf{f}) = p(\mathbf{f} | \widetilde{\boldsymbol{\theta}}, \mathbf{g})} \longrightarrow q_1(\mathbf{f}) \longrightarrow \widehat{\mathbf{f}}$$

EM:

$$\boldsymbol{\theta}^{(0)} \longrightarrow \widetilde{\boldsymbol{\theta}} \longrightarrow \boxed{q_1(\mathbf{f}) = p(\mathbf{f} | \widetilde{\boldsymbol{\theta}}, \mathbf{g})} \longrightarrow q_1(\mathbf{f}) \longrightarrow \widehat{\mathbf{f}}$$

$$\widehat{\boldsymbol{\theta}} \longleftarrow \widetilde{\boldsymbol{\theta}} \longleftarrow \boxed{Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) = \left\langle \ln p(\mathbf{f}, \boldsymbol{\theta} | \mathbf{g}) \right\rangle_{q_1(\mathbf{f})}} \longleftarrow q_1(\mathbf{f})$$

VBA:

$$\boldsymbol{\theta}^{(0)} \longrightarrow q_2(\boldsymbol{\theta}) \longrightarrow \boxed{q_1(\mathbf{f}) \propto \exp\left[ \left\langle \ln p(\mathbf{f}, \boldsymbol{\theta} | \mathbf{g}) \right\rangle_{q_2(\boldsymbol{\theta})} \right]} \longrightarrow q_1(\mathbf{f}) \longrightarrow \widehat{\mathbf{f}}$$

$$\widehat{\boldsymbol{\theta}} \longleftarrow q_2(\boldsymbol{\theta}) \longleftarrow \boxed{q_2(\boldsymbol{\theta}) \propto \exp\left[ \left\langle \ln p(\mathbf{f}, \boldsymbol{\theta} | \mathbf{g}) \right\rangle_{q_1(\mathbf{f})} \right]} \longleftarrow q_1(\mathbf{f})$$

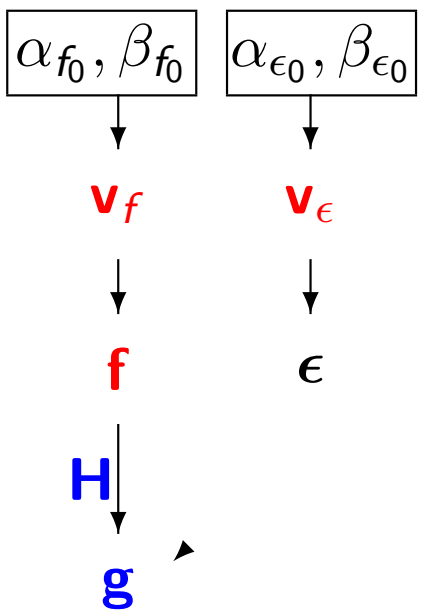
# Non stationary noise and sparsity enforcing model

– Non stationary noise:

$\mathbf{g} = \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon},\ \epsilon_i \sim \mathcal{N}(\epsilon_i|0, v_{\epsilon_i}) \rightarrow \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|0, \mathbf{V}_\epsilon = \text{diag}\,[v_{\epsilon 1}, \cdots, v_{\epsilon M}])$

– Student-t prior model and its equivalent IGSM :

$f_j|v_{f_j} \sim \mathcal{N}(f_j|0, v_{f_j})$ and $v_{f_j} \sim \mathcal{IG}(v_{f_j}|\alpha_{f_0}, \beta_{f_0}) \rightarrow f_j \sim \mathcal{S}t(f_j|\alpha_{f_0}, \beta_{f_0})$




$$\begin{cases} p(\mathbf{g}|\mathbf{f}, \mathbf{v}_\epsilon) = \mathcal{N}(\mathbf{g}|\mathbf{H}\mathbf{f}, \mathbf{V}_\epsilon), & \mathbf{V}_\epsilon = \text{diag}\,[\mathbf{v}_\epsilon] \\ p(\mathbf{f}|\mathbf{v}_f) = \mathcal{N}(\mathbf{f}|0, \mathbf{V}_f), & \mathbf{V}_f = \text{diag}\,[\mathbf{v}_f] \end{cases}$$

$$\begin{cases} p(\mathbf{v}_\epsilon) = \prod_i \mathcal{IG}(v_{\epsilon_i}|\alpha_{\epsilon_0}, \beta_{\epsilon_0}) \\ p(\mathbf{v}_f) = \prod_i \mathcal{IG}(v_{f_j}|\alpha_{f_0}, \beta_{f_0}) \end{cases}$$

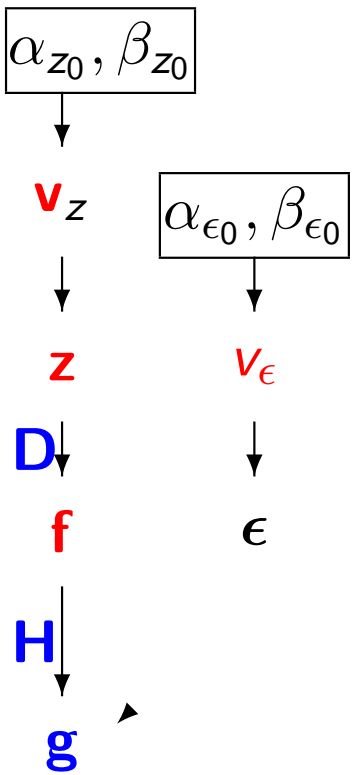
$$p(\mathbf{f}, \mathbf{v}_\epsilon, \mathbf{v}_f|\mathbf{g}) \propto p(\mathbf{g}|\mathbf{f}, \mathbf{v}_\epsilon)\, p(\mathbf{f}|\mathbf{v}_f)\, p(\mathbf{v}_\epsilon)\, p(\mathbf{v}_f)$$

Objective: Infer $(\mathbf{f}, \mathbf{v}_\epsilon, \mathbf{v}_f)$

– VBA: Approximate $p(\mathbf{f}, \mathbf{v}_\epsilon, \mathbf{v}_f|\mathbf{g})$ by $q_1(\mathbf{f})\, q_2(\mathbf{v}_\epsilon)\, q_3(\mathbf{v}_f)$

# Sparse model in a Transform domain 1

$$\mathbf{g} = \mathbf{H}\mathbf{f} + \epsilon, \quad \mathbf{f} = \mathbf{D}\mathbf{z}, \quad \mathbf{z} \text{ sparse}$$




$$\begin{cases} p(\mathbf{g}|\mathbf{z}, v_\epsilon) = \mathcal{N}(\mathbf{g}|\mathbf{H}\mathbf{D}\mathbf{f}, v_\epsilon \mathbf{I}) \\ p(\mathbf{z}|\mathbf{v}_z) = \mathcal{N}(\mathbf{z}|0, \mathbf{V}_z), \qquad \mathbf{V}_z = \text{diag}\,[\mathbf{v}_z] \end{cases}$$

$$\begin{cases} p(v_\epsilon) = \mathcal{IG}(v_\epsilon|\alpha_{\epsilon_0}, \beta_{\epsilon_0}) \\ p(\mathbf{v}_z) = \prod_i \mathcal{IG}(v_{z_j}|\alpha_{z_0}, \beta_{z_0}) \end{cases}$$

$$p(\mathbf{z}, v_\epsilon, \mathbf{v}_z, v_\xi|\mathbf{g}) \propto p(\mathbf{g}|\mathbf{z}, \mathbf{v}_\epsilon)\, p(\mathbf{z}|\mathbf{v}_z)\, p(v_\epsilon)\, p(\mathbf{v}_z)\, p(v_\xi)$$

– JMAP:

$$(\widehat{\mathbf{z}}, \hat{v}_\epsilon, \widehat{\mathbf{v}}_z) = \arg\max_{(\mathbf{z}, v_\epsilon, \mathbf{v}_z)} \{p(\mathbf{z}, v_\epsilon, \mathbf{v}_z|\mathbf{g})\}$$

Alternate optimization:

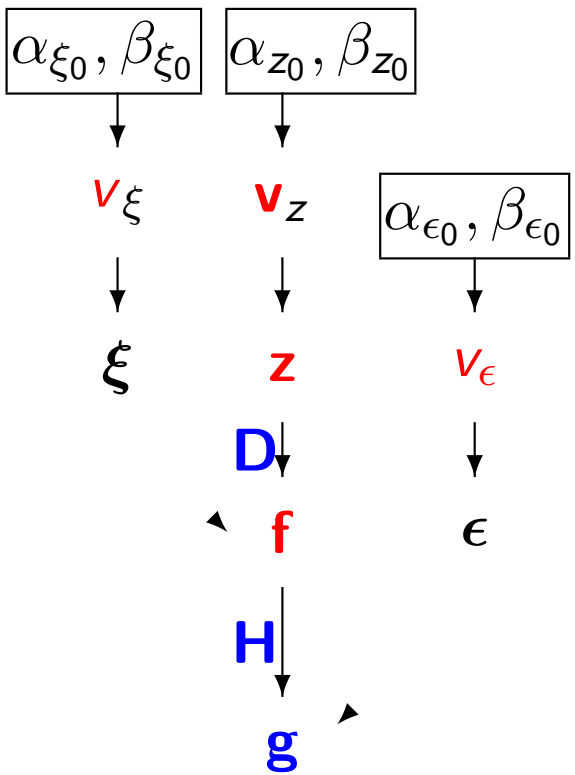
$$\begin{cases} \widehat{\mathbf{z}} = \arg\min_{\mathbf{z}} \{J(\mathbf{z})\} \text{ with:} \\ J(\mathbf{z}) = \frac{1}{2\hat{v}_\epsilon}\|\mathbf{g} - \mathbf{H}\mathbf{D}\mathbf{z}\|^2 + \|\mathbf{V}_z{}^{-1/2}\mathbf{z}\|^2 \\ \widehat{v}_{z_j} = \frac{\beta_{z_0} + \widehat{z}_j^2}{\alpha_{z_0} + 1/2} \\ \hat{v}_\epsilon = \frac{\beta_{\epsilon_0} + \|\mathbf{g} - \mathbf{H}\mathbf{D}\widehat{\mathbf{z}}\|^2}{\alpha_{\epsilon_0} + M/2} \end{cases}$$

– VBA: Approximate

$p(\mathbf{z}, v_\epsilon, \mathbf{v}_z, v_\xi|\mathbf{g})$ by $q_1(\mathbf{z})\, q_2(v_\epsilon)\, q_3(\mathbf{v}_z)$

Alternate optimization.

# Sparse model in a Transform domain 2




$$\mathbf{g} = \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon}, \quad \mathbf{f} = \mathbf{D}\mathbf{z} + \boldsymbol{\xi}, \quad \mathbf{z} \text{ sparse}$$

$$\begin{cases} p(\mathbf{g}|\mathbf{f}, v_\epsilon) = \mathcal{N}(\mathbf{g}|\mathbf{H}\mathbf{f}, v_\epsilon \mathbf{I}) \\ p(\mathbf{f}|\mathbf{z}) = \mathcal{N}(\mathbf{f}|\mathbf{D}\mathbf{z}, v_\xi \mathbf{I}), \\ p(\mathbf{z}|\mathbf{v}_z) = \mathcal{N}(\mathbf{z}|0, \mathbf{V}_z), \qquad \mathbf{V}_z = \text{diag}\,[\mathbf{v}_z] \end{cases}$$

$$\begin{cases} p(v_\epsilon) = \mathcal{IG}(v_\epsilon|\alpha_{\epsilon_0}, \beta_{\epsilon_0}) \\ p(\mathbf{v}_z) = \prod_i \mathcal{IG}(v_{z_j}|\alpha_{z_0}, \beta_{z_0}) \\ p(v_\xi) = \mathcal{IG}(v_\xi|\alpha_{\xi_0}, \beta_{\xi_0}) \end{cases}$$

$$p(\mathbf{f}, \mathbf{z}, v_\epsilon, \mathbf{v}_z, v_\xi|\mathbf{g}) \propto p(\mathbf{g}|\mathbf{f}, \mathbf{v}_\epsilon)\, p(\mathbf{f}|\mathbf{z}_f)\, p(\mathbf{z}|\mathbf{v}_z)\, p(v_\epsilon)\, p(\mathbf{v}_z)\, p(v_\xi)$$

– JMAP:

$$(\widehat{\mathbf{f}}, \widehat{\mathbf{z}}, \widehat{v}_\epsilon, \widehat{\mathbf{v}}_z, \widehat{v}_\xi) = \underset{(\mathbf{f}, \mathbf{z}, v_\epsilon, \mathbf{v}_z, v_\xi)}{\arg\max} \{p(\mathbf{f}, \mathbf{z}, v_\epsilon, \mathbf{v}_z, v_\xi|\mathbf{g})\}$$

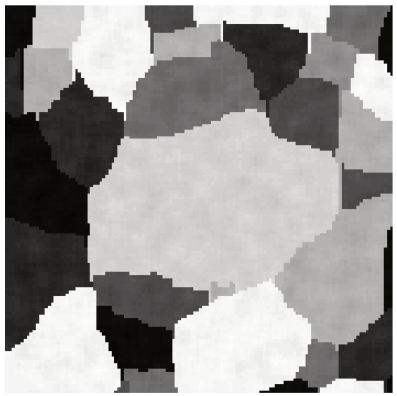
Alternate optimization.

– VBA: Approximate

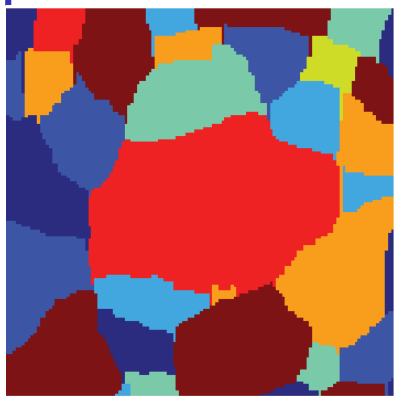
$p(\mathbf{f}, \mathbf{z}, v_\epsilon, \mathbf{v}_z, v_\xi|\mathbf{g})$ by $q_1(\mathbf{f})\, q_2(\mathbf{z})\, q_3(v_\epsilon)\, q_4(\mathbf{v}_z)\, q_5(v_\xi)$

Alternate optimization.

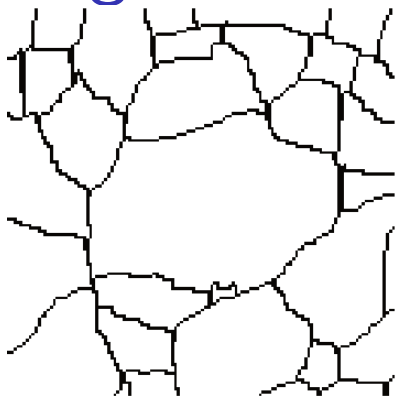
# Gauss-Markov-Potts prior models for images



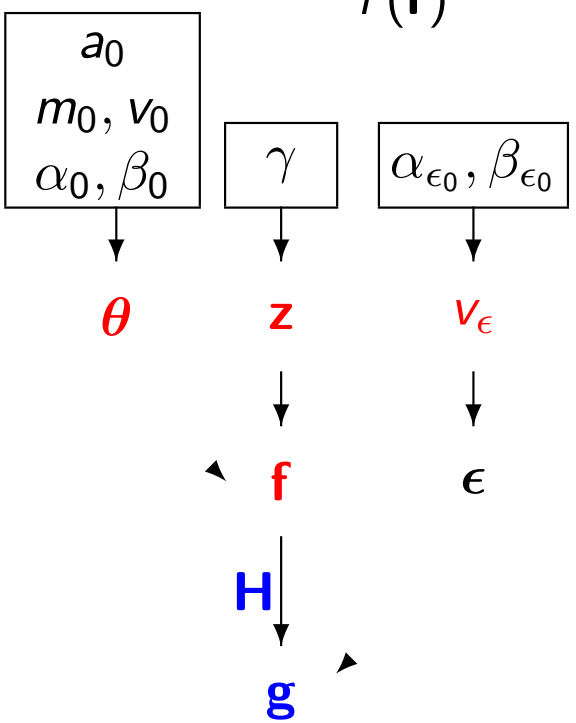
$f(\mathbf{r})$



$z(\mathbf{r})$



$c(\mathbf{r}) = 1 - \delta(z(\mathbf{r}) - z(\mathbf{r}'))$




$$
\begin{cases}
\mathbf{g} = \mathbf{H}\mathbf{f} + \epsilon \\
p(\mathbf{g}|\mathbf{f}, v_\epsilon) = \mathcal{N}(\mathbf{g}|\mathbf{H}\mathbf{f}, v_\epsilon \mathbf{I}) \\
p(v_\epsilon) = \mathcal{IG}(v_\epsilon|\alpha_{\epsilon_0}, \beta_{\epsilon_0})
\end{cases}
$$

$$
\begin{cases}
p(f(\mathbf{r})|z(\mathbf{r}) = k, m_k, v_k) = \mathcal{N}(f(\mathbf{r})|m_k, v_k) \\
p(\mathbf{f}|\mathbf{z}, \boldsymbol{\theta}) = \sum_k \prod_{\mathbf{r} \in \mathcal{R}_k} a_k \mathcal{N}(f(\mathbf{r})|m_k, v_k), \\
\qquad \boldsymbol{\theta} = \{(a_k, m_k, v_k), k = 1, \cdots, K\} \\
p(\boldsymbol{\theta}) = \mathcal{D}(\mathbf{a}|a_0)\mathcal{N}(\mathbf{a}|m_0, v0)\mathcal{IG}(\mathbf{v}|\alpha_0, \beta_0) \\
p(\mathbf{z}|\gamma) \propto \exp\left[\gamma \sum_{\mathbf{r}} \sum_{\mathbf{r}' \in \mathcal{N}(\mathbf{r})} \delta(z(\mathbf{r}) - z(\mathbf{r}'))\right] \quad \text{Potts MRF}
\end{cases}
$$

$$
p(\mathbf{f}, \mathbf{z}, \boldsymbol{\theta}|\mathbf{g}) \propto p(\mathbf{g}|\mathbf{f}, v_\epsilon)\, p(\mathbf{f}|\mathbf{z}, \boldsymbol{\theta})\, p(\mathbf{z}|\gamma)
$$

MCMC: Gibbs Sampling

VBA: Alternate optimization.

# Mixture Models

1. Mixture models
2. Different problems related to classification and clustering
   - Training
   - Supervised classification
   - Semi-supervised classification
   - Clustering or unsupervised classification
3. Mixture of Gaussian (MoG)
4. Mixture of Student-t (MoSt)
5. Variational Bayesian Approximation (VBA)
6. VBA for Mixture of Gaussian
7. VBA for Mixture of Student-t
8. Conclusion

# Mixture models

- General mixture model

$$p(\mathbf{x}|\mathbf{a},\mathbf{\Theta},K)=\sum_{k=1}^{K}a_k\,p_k(\mathbf{x}_k|\boldsymbol{\theta}_k),\quad 0<a_k<1,\ \sum_{k=1}^{K}a_k=1$$

- Same family $p_k(\mathbf{x}_k|\boldsymbol{\theta}_k)=p(\mathbf{x}_k|\boldsymbol{\theta}_k),\ \forall k$
- Gaussian $p(\mathbf{x}_k|\boldsymbol{\theta}_k)=\mathcal{N}(\mathbf{x}_k|\boldsymbol{\mu}_k,\mathbf{V}_k)$ with $\boldsymbol{\theta}_k=(\boldsymbol{\mu}_k,\mathbf{V}_k)$
- Data $\mathbf{X}=\{\mathbf{x}_n,n=1,\cdots,N\}$ where each element $\mathbf{x}_n$ can be in one of the $K$ classes $c_n$.
- $a_k=p(c_n=k)$, $\mathbf{a}=\{a_k,k=1,\cdots,K\}$, $\mathbf{\Theta}=\{\boldsymbol{\theta}_k,k=1,\cdots,K\}$, $\mathbf{c}=\{c_n,n=1,\cdots,N\}$

$$p(\mathbf{X},\mathbf{c}|\mathbf{a},\mathbf{\Theta})=\prod_{n=1}^{N}p(\mathbf{x}_n,c_n=k|a_k,\boldsymbol{\theta}_k)$$

# Different problems

- Training:
  Given a set of (training) data $\mathbf{X}$ and classes $\mathbf{c}$, estimate the parameters $\mathbf{a}$ and $\mathbf{\Theta}$.
- Supervised classification:
  Given a sample $\mathbf{x}_m$ and the parameters $K$, $\mathbf{a}$ and $\mathbf{\Theta}$ determine its class

$$k^* = \arg\max_k \{p(c_m = k|\mathbf{x}_m, \mathbf{a}, \mathbf{\Theta}, K)\}.$$

- Semi-supervised classification (Proportions are not known):
  Given sample $\mathbf{x}_m$ and the parameters $K$ and $\mathbf{\Theta}$, determine its class

$$k^* = \arg\max_k \{p(c_m = k|\mathbf{x}_m, \mathbf{\Theta}, K)\}.$$

- Clustering or unsupervised classification (Number of classes K is not known):
  Given a set of data $\mathbf{X}$, determine $K$ and $\mathbf{c}$.

# Training

- Given a set of (training) data $\mathbf{X}$ and classes $\mathbf{c}$, estimate the parameters $\mathbf{a}$ and $\mathbf{\Theta}$.
- Maximum Likelihood (ML):

$$(\widehat{\mathbf{a}}, \widehat{\mathbf{\Theta}}) = \underset{(\mathbf{a},\mathbf{\Theta})}{\arg\max} \{p(\mathbf{X}, \mathbf{c}|\mathbf{a}, \mathbf{\Theta}, K)\}.$$

- Bayesian: Assign priors $p(\mathbf{a}|K)$ and $p(\mathbf{\Theta}|K) = \prod_{k=1}^{K} p(\theta_k)$ and write the expression of the joint posterior laws:

$$p(\mathbf{a}, \mathbf{\Theta}|\mathbf{X}, \mathbf{c}, K) = \frac{p(\mathbf{X}, \mathbf{c}|\mathbf{a}, \mathbf{\Theta}, K)\, p(\mathbf{a}|K)\, p(\mathbf{\Theta}|K)}{p(\mathbf{X}, \mathbf{c}|K)}$$

where

$$p(\mathbf{X}, \mathbf{c}|K) = \iint p(\mathbf{X}, \mathbf{c}|\mathbf{a}, \mathbf{\Theta}|K) p(\mathbf{a}|K)\, p(\mathbf{\Theta}|K)\, \mathrm{d}\mathbf{a}\, \mathrm{d}\mathbf{\Theta}$$

- Infer on $\mathbf{a}$ and $\mathbf{\Theta}$ either as the Maximum A Posteriori (MAP) or Posterior Mean (PM).

# Supervised classification

- Given a sample $\mathbf{x}_m$ and the parameters $K$, $\mathbf{a}$ and $\mathbf{\Theta}$ determine

$$p(c_m = k|\mathbf{x}_m, \mathbf{a}, \mathbf{\Theta}, K) = \frac{p(\mathbf{x}_m, c_m = k|\mathbf{a}, \mathbf{\Theta}, K)}{p(\mathbf{x}_m|\mathbf{a}, \mathbf{\Theta}, K)}$$

where $p(\mathbf{x}_m, c_m = k|\mathbf{a}, \mathbf{\Theta}, K) = a_k p(\mathbf{x}_m|\boldsymbol{\theta}_k)$ and

$$p(\mathbf{x}_m|\mathbf{a}, \mathbf{\Theta}, K) = \sum_{k=1}^{K} a_k \, p(\mathbf{x}_m|\boldsymbol{\theta}_k)$$

- Best class $k^*$:

$$k^* = \arg\max_k \{p(c_m = k|\mathbf{x}_m, \mathbf{a}, \mathbf{\Theta}, K)\}$$

# Semi-supervised classification

- Given sample $\mathbf{x}_m$ and the parameters $K$ and $\boldsymbol{\Theta}$ (not the proportions $\mathbf{a}$), determine the probabilities

$$p(c_m = k|\mathbf{x}_m, \boldsymbol{\Theta}, K) = \frac{p(\mathbf{x}_m, c_m = k|\boldsymbol{\Theta}, K)}{p(\mathbf{x}_m|\boldsymbol{\Theta}, K)}$$

where

$$p(\mathbf{x}_m, c_m = k|\boldsymbol{\Theta}, K) = \int p(\mathbf{x}_m, c_m = k|\mathbf{a}, \boldsymbol{\Theta}, K)p(\mathbf{a}|K)\,\mathrm{d}\mathbf{a}$$

and

$$p(\mathbf{x}_m|\boldsymbol{\Theta}, K) = \sum_{k=1}^{K} p(\mathbf{x}_m, c_m = k|\boldsymbol{\Theta}, K)$$

- Best class $k^*$, for example the MAP solution:

$$k^* = \arg\max_k \{p(c_m = k|\mathbf{x}_m, \boldsymbol{\Theta}, K)\}.$$

# Clustering or non-supervised classification

- Given a set of data $\mathbf{X}$, determine $K$ and $\mathbf{c}$.
- Determination of the number of classes:

$$p(K = L|\mathbf{X}) = \frac{p(\mathbf{X}, K = L)}{p(\mathbf{X})} = \frac{p(\mathbf{X}|K = L)\, p(K = L)}{p(\mathbf{X})}$$

and

$$p(\mathbf{X}) = \sum_{L=1}^{L_0} p(K = L)\, p(\mathbf{X}|K = L),$$

where $L_0$ is the a priori maximum number of classes and

$$p(\mathbf{X}|K = L) = \int \int \prod_{n} \prod_{k=1}^{L} a_k p(\mathbf{x}_n, c_n = k|\theta_k) p(\mathbf{a}|K)\, p(\mathbf{\Theta}|K)\, \mathrm{d}\mathbf{a}\, \mathrm{d}\mathbf{\Theta}.$$

- When $K$ and $\mathbf{c}$ are determined, we can also determine the characteristics of those classes $\mathbf{a}$ and $\mathbf{\Theta}$.

# Mixture of Gaussian and Mixture of Student-t

$$p(\mathbf{x}|\mathbf{a},\mathbf{\Theta},K)=\sum_{k=1}^{K}a_k\,p(\mathbf{x}_k|\boldsymbol{\theta}_k),\quad 0<a_k<1,\ \sum_{k=1}^{K}a_k=1$$

- Mixture of Gaussian (MoG)

$$p(\mathbf{x}_k|\boldsymbol{\theta}_k)=\mathcal{N}(\mathbf{x}_k|\boldsymbol{\mu}_k,\mathbf{V}_k),\quad \boldsymbol{\theta}_k=(\boldsymbol{\mu}_k,\mathbf{V}_k)$$

$$\mathcal{N}(\mathbf{x}_k|\boldsymbol{\mu}_k,\mathbf{V}_k)=(2\pi)^{-\frac{p}{2}}\,|\mathbf{V}_k|^{-\frac{1}{2}}\exp\left[\frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu}_k)'\mathbf{V}_k^{-1}(\mathbf{x}_k-\boldsymbol{\mu}_k)\right]$$

- Mixture of Student-t (MoSt)

$$p(\mathbf{x}_k|\boldsymbol{\theta}_k)=\mathcal{T}(\mathbf{x}_k|\nu_k,\boldsymbol{\mu}_k,\mathbf{V}_k),\quad \boldsymbol{\theta}_k=(\nu_k,\boldsymbol{\mu}_k,\mathbf{V}_k)$$

$$\mathcal{T}(\mathbf{x}_k|\nu,\boldsymbol{\mu}_k,\mathbf{V}_k)=\frac{\Gamma\left[\frac{(\nu_k+p)}{2}\right]}{\Gamma(\frac{\nu_k}{2})\nu^{\frac{p}{2}}\pi^{\frac{p}{2}}}\,|\mathbf{V}_k|^{-\frac{1}{2}}\left[1+\frac{1}{\nu}(\mathbf{x}_k-\boldsymbol{\mu}_k)'\mathbf{V}_k^{-1}(\mathbf{x}_k-\boldsymbol{\mu}_k)\right]^{-\frac{(\nu+p)}{2}}$$

# Mixture of Student-t model

- Student-t and its Infinite Gaussian Scaled Model (IGSM):

$$\mathcal{T}(\mathbf{x}|\nu,\boldsymbol{\mu},\mathbf{V}) = \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},u^{-1}\mathbf{V})\,\mathcal{G}(u|\frac{\nu}{2},\frac{\nu}{2})\,\mathrm{d}z$$

where

$$\begin{aligned}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\mathbf{V}) &= |2\pi\mathbf{V}|^{-\frac{1}{2}}\exp\left[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]\\ &= |2\pi\mathbf{V}|^{-\frac{1}{2}}\exp\left[-\tfrac{1}{2}\mathrm{Tr}\left\{(\mathbf{x}-\boldsymbol{\mu})\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})'\right\}\right]\end{aligned}$$

and

$$\mathcal{G}(u|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}u^{\alpha-1}\exp\left[-\beta u\right].$$

- Mixture of generalized Student-t: $\mathcal{T}(\mathbf{x}|\alpha,\beta,\boldsymbol{\mu},\mathbf{V})$

$$p(\mathbf{x}|\{a_k,\boldsymbol{\mu}_k,\mathbf{V}_k,\alpha_k,\beta_k\},K) = \sum_{k=1}^{K} a_k\,\mathcal{T}(\mathbf{x}_n|\alpha_k,\beta_k,\boldsymbol{\mu}_k,\mathbf{V}_k).$$

# Mixture of Gaussian model

- Introducing $z_{nk} \in \{0,1\}$, $\mathbf{z}_k = \{z_{nk}, n = 1, \cdots, N\}$, $\mathbf{Z} = \{z_{nk}\}$ with $P(z_{nk} = 1) = P(c_n = k) = a_k$, $\boldsymbol{\theta}_k = \{a_k, \boldsymbol{\mu}_k, \mathbf{V}_k\}$, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_k, k = 1, \cdots, K\}$
- Assigning the priors $p(\boldsymbol{\Theta}) = \prod_k p(\boldsymbol{\theta}_k)$, we can write:

$$p(\mathbf{X}, \mathbf{c}, \mathbf{Z}, \boldsymbol{\Theta}|K) = \prod_n \sum_k a_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{V}_k)\,(1-\delta(z_{nk}))\, \prod_k p(\boldsymbol{\theta}_k)$$

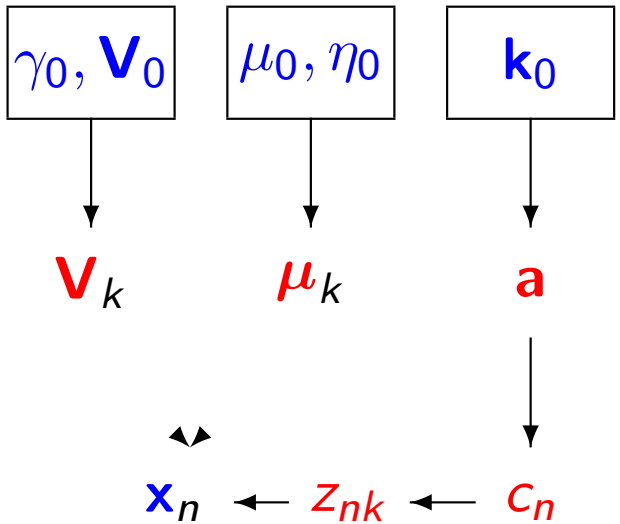
$$p(\mathbf{X}, \mathbf{c}, \mathbf{Z}, \boldsymbol{\Theta}|K) = \prod_n \prod_k [a_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{V}_k)]^{z_{nk}}\, p(\boldsymbol{\theta}_k)$$

- Joint posterior law:

$$p(\mathbf{c}, \mathbf{Z}, \boldsymbol{\Theta}|\mathbf{X}, K) = \frac{p(\mathbf{X}, \mathbf{c}, \mathbf{Z}, \boldsymbol{\Theta}|K)}{p(\mathbf{X}|K)}.$$

- The main task now is to propose some approximations to it in such a way that we can use it easily in all the above mentioned tasks of classification or clustering.

# Hierarchical graphical model for Mixture of Gaussian




$$\begin{cases} p(\mathbf{a}) = \mathcal{D}(\mathbf{a}|\mathbf{k}_0) \\ p(\mu_k|\mathbf{V}_k) = \mathcal{N}(\mu_k|\mu_0 1, \eta_0{}^{-1}\mathbf{V}_k) \\ p(\mathbf{V}_k) = \mathcal{IW}(\mathbf{V}_k|\gamma_0, \mathbf{V}_0) \\ P(z_{nk} = 1) = P(c_n = k) = a_k \end{cases}$$

$$p(\mathbf{X}, \mathbf{c}, \mathbf{Z}, \mathbf{\Theta}|K) = \prod_n \prod_k [a_k \mathcal{N}(\mathbf{x}_n|\mu_k, \mathbf{V}_k)]^{z_{nk}}$$

$$p(a_k)p(\mu_k|\mathbf{V}_k)p(\mathbf{V}_k)$$

# Mixture of Student-t model

- Introducing $\mathbf{U} = \{u_{nk}\}$
  $\boldsymbol{\theta}_k = \{\alpha_k, \beta_k, \mathbf{a}_k, \boldsymbol{\mu}_k, \mathbf{V}_k\}$, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_k, k = 1, \cdots, K\}$
- Assigning the priors $p(\boldsymbol{\Theta}) = \prod_k p(\boldsymbol{\theta}_k)$, we can write:

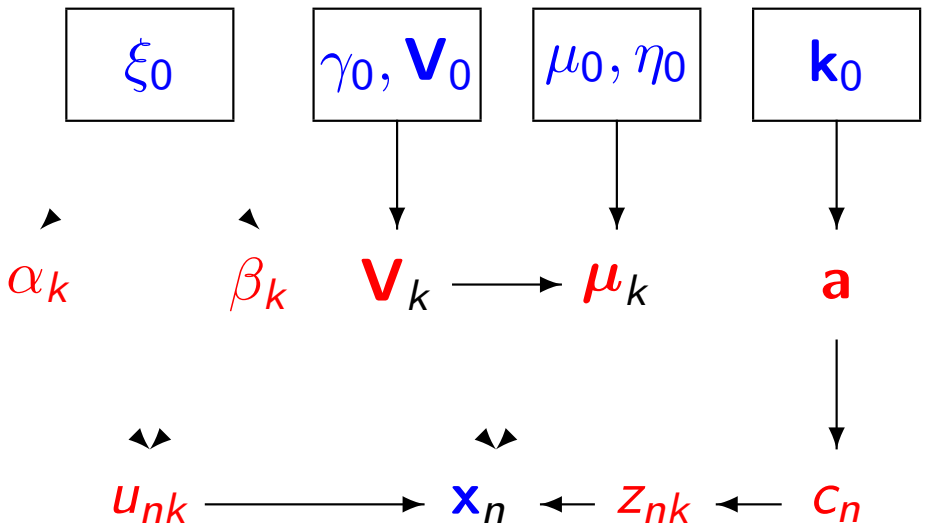
$$p(\mathbf{X}, \mathbf{c}, \mathbf{Z}, \mathbf{U}, \boldsymbol{\Theta}|K) = \prod_n \prod_k \left[a_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, u_{n,k}{}^{-1}\mathbf{V}_k)\, \mathcal{G}(u_{nk}|\alpha_k, \beta_k)\right]^{z_{nk}} p(\boldsymbol{\theta}_k)$$

- Joint posterior law:

$$p(\mathbf{c}, \mathbf{Z}, \mathbf{U}, \boldsymbol{\Theta}|\mathbf{X}, K) = \frac{p(\mathbf{X}, \mathbf{c}, \mathbf{Z}, \mathbf{U}, \boldsymbol{\Theta}|K)}{p(\mathbf{X}|K)}.$$

- The main task now is to propose some approximations to it in such a way that we can use it easily in all the above mentioned tasks of classification or clustering.

# Hierarchical graphical model for Mixture of Student-t




$$
\begin{cases}
p(\mathbf{a}) = \mathcal{D}(\mathbf{a}|\mathbf{k}_0) \\
p(\boldsymbol{\mu}_k|\mathbf{V}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mu_0 1, \eta_0{}^{-1}\mathbf{V}_k) \\
p(\mathbf{V}_k) = \mathcal{IW}(\mathbf{V}_k|\gamma_0, \mathbf{V}_0) \\
p(\alpha_k) = \mathcal{E}(\alpha_k|\zeta_0) = \mathcal{G}(\alpha_k|1, \zeta_0) \\
p(\beta_k) = \mathcal{E}(\beta_k|\zeta_0) = \mathcal{G}(\beta_k|1, \zeta_0) \\
P(z_{nk} = 1) = P(c_n = k) = a_k \\
p(u_{nk}) = \mathcal{G}(u_{nk}|\alpha_k, \beta_k)
\end{cases}
$$

$$
p(\mathbf{X}, \mathbf{c}, \mathbf{Z}, \mathbf{U}, \boldsymbol{\Theta}|K) = \prod_n \prod_k [a_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{V}_k)\mathcal{G}(u_{nk}|\alpha_k, \beta_k)]^{z_{nk}}
$$

$$
p(a_k)p(\boldsymbol{\mu}_k|\mathbf{V}_k)p(\mathbf{V}_k)p(\alpha_k)p(\beta_k)
$$

# Variational Bayesian Approximation (VBA)

- Main idea: to propose easy computational approximations: $q(\mathbf{c}, \mathbf{Z}, \mathbf{\Theta}) = q(\mathbf{c}, \mathbf{Z})q(\mathbf{\Theta})$ for $p(\mathbf{c}, \mathbf{Z}, \mathbf{\Theta}|\mathbf{X}, K)$ for MoG model, or
  $q(\mathbf{c}, \mathbf{Z}, \mathbf{U}, \mathbf{\Theta}) = q(\mathbf{c}, \mathbf{Z}, \mathbf{U})q(\mathbf{\Theta})$ for $p(\mathbf{c}, \mathbf{Z}, \mathbf{U}, \mathbf{\Theta}|\mathbf{X}, K)$ for MoSt model.
- Criterion:

$$\mathrm{KL}(q : p) = -\mathcal{F}(q) + \ln p(\mathbf{X}|K)$$

  where

$$\mathcal{F}(q) = \langle -\ln p(\mathbf{X}, \mathbf{c}, \mathbf{Z}, \mathbf{\Theta}|K)\rangle_q$$

  or

$$\mathcal{F}(q) = \langle -\ln p(\mathbf{X}, \mathbf{c}, \mathbf{Z}, \mathbf{U}, \mathbf{\Theta}|K)\rangle_q$$

- Maximizing $\mathcal{F}(q)$ or minimizing $\mathrm{KL}(q : p)$ are equivalent and both give un upper bound to the evidence of the model $\ln p(\mathbf{X}|K)$.
- When the optimum $q^*$ is obtained, $\mathcal{F}(q^*)$ can be used as a criterion for model selection.

# Proposed VBA for Mixture of Student-t priors model

- Dirichlet

$$\mathcal{D}(\mathbf{a}|\mathbf{k}) = \frac{\Gamma(\sum_l k_k)}{\prod_l \Gamma(k_l)} \prod_l a_l^{k_l - 1}$$

- Exponential

$$\mathcal{E}(t|\zeta_0) = \zeta_0 \exp\left[-\zeta_0 t\right]$$

- Gamma

$$\mathcal{G}(t|a,b) = \frac{b^a}{\Gamma(a)} t^{a-1} \exp\left[-bt\right]$$

- Inverse Wishart

$$\mathcal{IW}(\mathbf{V}|\gamma,\gamma\mathbf{\Delta}) = \frac{|\frac{1}{2}\mathbf{\Delta}|^{\gamma/2} \exp\left[-\frac{1}{2}\mathrm{Tr}\left\{\mathbf{\Delta}\mathbf{V}^{-1}\right\}\right]}{\Gamma_D(\gamma/2)|\mathbf{V}|^{\frac{\gamma+D+1}{2}}}.$$

# Expressions of $q$

$$
\begin{aligned}
q(\mathbf{c}, \mathbf{Z}, \mathbf{\Theta}) =& q(\mathbf{c}, \mathbf{Z})\, q(\mathbf{\Theta}) \\
&= \prod_n \prod_k [q(c_n = k|z_{nk})\, q(z_{nk})] \\
&\prod_k [q(\alpha_k)\, q(\beta_k)\, q(\boldsymbol{\mu}_k|\mathbf{V}_k)\, q(\mathbf{V}_k)]\, q(\mathbf{a}).
\end{aligned}
$$

with:

$$
\begin{cases}
q(\mathbf{a}) = \mathcal{D}(\mathbf{a}|\tilde{\mathbf{k}}), \quad \tilde{\mathbf{k}} = [\tilde{k}_1, \cdots, \tilde{k}_K] \\
q(\alpha_k) = \mathcal{G}(\alpha_k|\tilde{\xi}_k, \tilde{\eta}_k) \\
q(\beta_k) = \mathcal{G}(\beta_k|\tilde{\xi}_k, \tilde{\eta}_k) \\
q(\boldsymbol{\mu}_k|\mathbf{V}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\tilde{\boldsymbol{\mu}}, \tilde{\eta}^{-1}\mathbf{V}_k) \\
q(\mathbf{V}_k) = \mathcal{IW}(\mathbf{V}_k|\tilde{\gamma}, \tilde{\Sigma})
\end{cases}
$$

With these choices, we have

$$
\mathcal{F}(q(\mathbf{c}, \mathbf{Z}, \mathbf{\Theta})) = \langle \ln p(\mathbf{X}, \mathbf{c}, \mathbf{Z}, \mathbf{\Theta}|K) \rangle_{q(\mathbf{c}, \mathbf{Z}, \mathbf{\Theta})} = \prod_k \prod_n \mathcal{F}_{1_{kn}} + \prod_k \mathcal{F}_{2_k}
$$

$$
\begin{aligned}
\mathcal{F}_{1_{kn}} &= \langle \ln p(\mathbf{x}_n, c_n, z_{nk}, \boldsymbol{\theta}_k) \rangle_{q(c_n=k|z_{nk})q(z_{nk})} \\
\mathcal{F}_{2_k} &= \langle \ln p(\mathbf{x}_n, c_n, z_{nk}, \boldsymbol{\theta}_k) \rangle_{q(\boldsymbol{\theta}_k)}
\end{aligned}
$$

# VBA Algorithm step

Expressions of the updating expressions of the tilded parameters are obtained by following three steps:

- E step: Optimizing $\mathcal{F}$ with respect to $q(\mathbf{c}, \mathbf{Z})$ when keeping $q(\mathbf{\Theta})$ fixed, we obtain the expression of $q(c_n = k|z_{nk}) = \tilde{a}_k$, $q(z_{nk}) = \mathcal{G}(z_{nk}|\widetilde{\alpha}_k, \widetilde{\beta}_k)$.
- M step: Optimizing $\mathcal{F}$ with respect to $q(\mathbf{\Theta})$ when keeping $q(\mathbf{c}, \mathbf{Z})$ fixed, we obtain the expression of $q(\mathbf{a}) = \mathcal{D}(\mathbf{a}|\tilde{\mathbf{k}})$, $\quad \tilde{\mathbf{k}} = [\tilde{k}_1, \cdots, \tilde{k}_K]$, $q(\alpha_k) = \mathcal{G}(\alpha_k|\tilde{\zeta}_k, \tilde{\eta}_k)$, $q(\beta_k) = \mathcal{G}(\beta_k|\tilde{\zeta}_k, \tilde{\eta}_k)$, $q(\boldsymbol{\mu}_k|\mathbf{V}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\widetilde{\boldsymbol{\mu}}, \tilde{\eta}^{-1}\mathbf{V}_k)$, and $q(\mathbf{V}_k) = \mathcal{IW}(\mathbf{V}_k|\tilde{\gamma}, \tilde{\gamma}\tilde{\Sigma})$, which gives the updating algorithm for the corresponding tilded parameters.
- $\mathcal{F}$ evaluation: After each E step and M step, we can also evaluate the expression of $\mathcal{F}(q)$ which can be used for stopping rule of the iterative algorithm.
- Final value of $\mathcal{F}(q)$ for each value of $K$, noted $\mathcal{F}_k$, can be used as a criterion for model selection, i.e.; the determination of the number of clusters.

# VBA: choosing the good families for $q$

- Main question: We approximate $p(X)$ by $q(X)$. What are the quantities we have conserved?
  - a) Modes values: $\arg\max_x \{p(X)\} = \arg\max_x \{q(X)\}$ ?
  - b) Expected values: $E_p(X) = E_q(X)$ ?
  - c) Variances: $V_p(X) = V_q(X)$ ?
  - d) Entropies: $H_p(X) = H_q(X)$ ?
- Recent works shows some of these under some conditions.
- For example, if $p(x) = \frac{1}{Z}\exp\left[-\phi(x)\right]$ with $\phi(x)$ convex and symetric, properties a) and b) are satisfied.
- Unfortunately, this is not the case for variances or other moments.
- If $p$ is in the exponential family, then choosing appropriate conjugate priors, the structure of $q$ will be the same and we can obtain appropriate fast optimization algorithms.

# Conclusions

- Bayesian approach with Hierarchical prior model with hidden variables are very powerful tools for inverse problems and Machine Learning.
- The computational cost of all the sampling methods (MCMC and many others) are too high to be used in practical high dimensional applications.
- We explored VBA tools for effective approximate Bayesian computation.
- Application in different inverse problems in imaging system (3D X ray CT, Microwaves, PET, Ultrasound, Optical Diffusion Tomography (ODT), Acoustic source localization,...)
- Clustering and classification of a set of data are between the most important tasks in statistical researches for many applications such as data mining in biology.
- Mixture models are classical models for these tasks.
- We proposed to use a mixture of generalised Student-t distribution model for more robustness.
- To obtain fast algorithms and be able to handle large data