Introduction to statistical methods in signal and image processing

Florence Forbes

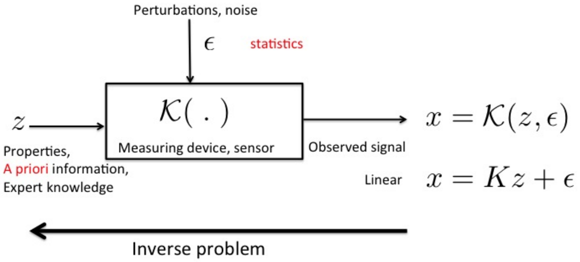
Introduction to statistical methods in signal and image processing

Florence Forbes

florence.forbes@inria.fr

INRIA Mistis team & Lab. Jean Kuntzman
Université Grenoble Alpes
http://mistis.inrialpes.fr

A methodological framework for inverse problems



Linear Models: convolution (image restoration), projection (tomography), mixtures (source separation), Laplace and Fourier transform (NMR, MRI)

Inversion: instability, non-unicity or existence of the solution

$\longrightarrow$ Ill-posed problem

Regularization: add constraints/hypothesis on the seek solution

- Bayesian inference: $p(z|x) \propto p(x|z)p(z)$
- Penalized criterion minimization: $F(z) = L(z, x) + \beta R(z)$

Overview: Part 1- Introduction to Bayesian tools

- Introduction
- Statistical inference
  - Learning and decision
  - Maximum likelihood
- Bayesian set up
  - prior, posterior, etc.
- Bayesian inference strategies
  - Point estimators
  - Fully Bayesian treatment
- Prior distributions
  - Conjugate priors and exponential family
  - Noninformative and Jeffreys' priors
- Tractability of posteriors

Overview: Part 2- Probabilistic graphical models

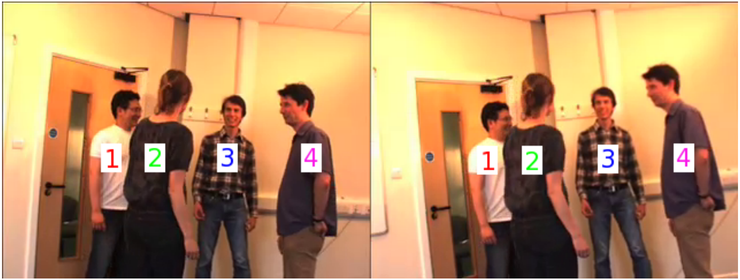
- Directed graphs: Bayesian networks
- Conditional independence and Markov properties
- Undirected graphs: Markov random fields
- Inference and learning
- Illustration: image segmentation

Introduction

Illustration: Audio-Visual Scene Analysis

- Estimate the number of audio-visual objects
- Localize and track every object
- Determine auditory activity and visibility

Illustration: Audio-Visual Scene Analysis



- Estimate the number of audio-visual objects
- Localize and track every object
- Determine auditory activity and visibility

Illustration: Audio-Visual Scene Analysis

- Estimate the number of audio-visual objects
- Localize and track every object
- Determine auditory activity and visibility

Illustration: Audio-Visual Scene Analysis

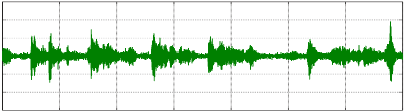
- Estimate the number of audio-visual objects
- Localize and track every object
- Determine auditory activity and visibility

Observed Data

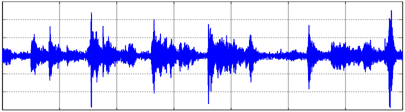
Right camera image:

Left camera image:

Left microphone signal:

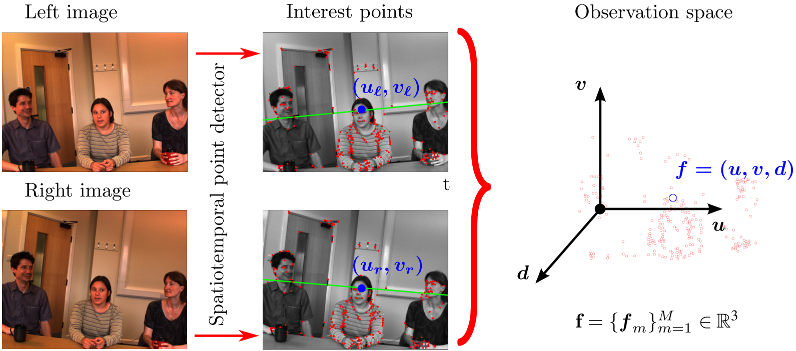


Right microphone signal:

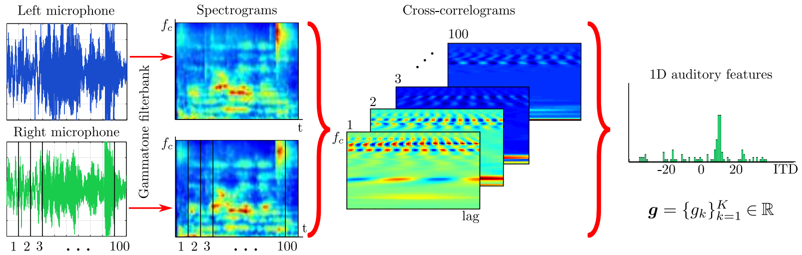


Visual Features Extraction

An image pair produces a set of visual observations
$\mathbf{f} = \{\boldsymbol{f}_m\}_{m=1}^{M} \in \mathbb{R}^3$;
$\boldsymbol{f} = (u, v, d)$: u, v - image coordinates, d - disparity



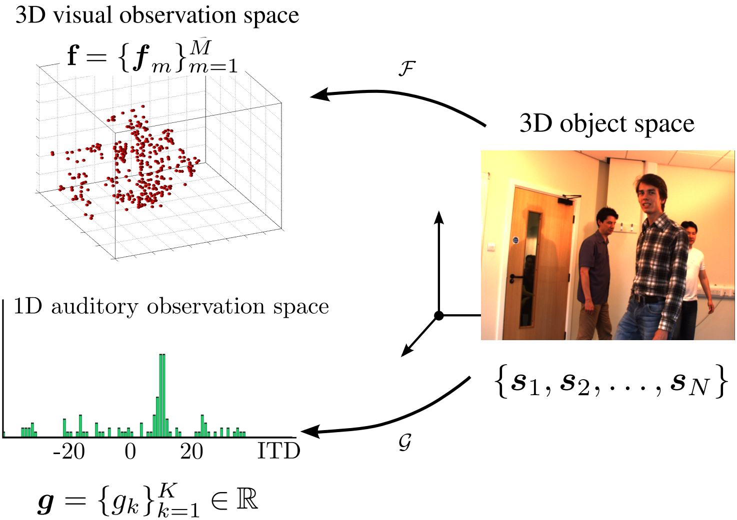
Auditory Features Extraction



ITD = interaural time difference

An ITD detection algorithm [H. Christensen, 2007] procudes for a 10ms interval of audio signals one auditory observation $g_k \in \mathbb{R}$

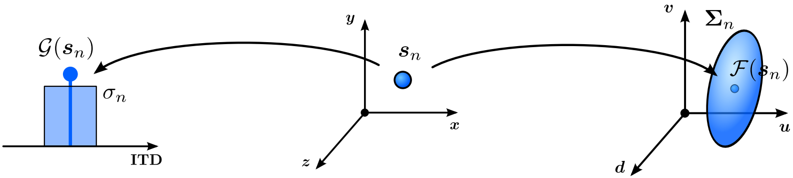
Audio-Visual Generative Model



Why statistical modelling in Audio-Visual Scene Analysis ?

- Observations are strongly affected by noise: detector errors, occlusions, reverberations, ambient sounds,

  can be accounted for with some probability distributions.

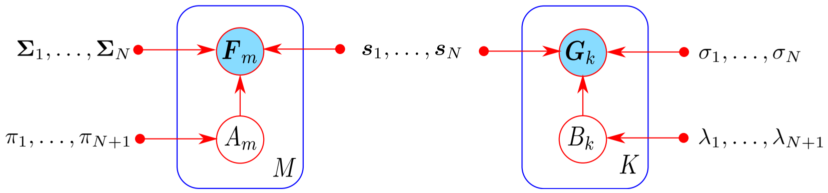


  - $P(\boldsymbol{f}_m \mid A_m = n; \boldsymbol{s}_n) = \mathcal{N}(\boldsymbol{f}_m;\ \mathcal{F}(\boldsymbol{s}_n), \boldsymbol{\Sigma}_n)$;
  - $P(g_k \mid B_k = n; \boldsymbol{s}_n) = \mathcal{N}(g_k;\ \mathcal{G}(\boldsymbol{s}_n), \boldsymbol{\Gamma}_n)$;

- Dynamically changing environment:

  can be accounted for with some prior knowledge, eg. on motion cues, trajectories are continuous, smooth, etc...

Statistical Model formulation



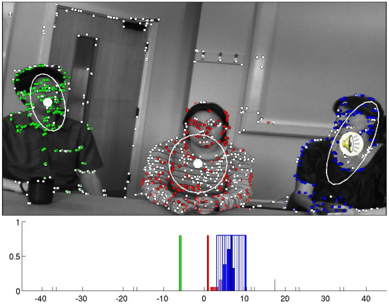
- $\boldsymbol{s} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n, \ldots, \boldsymbol{s}_N\}$ are tying parameters
- Simultaneous clustering in auditory and visual observation spaces

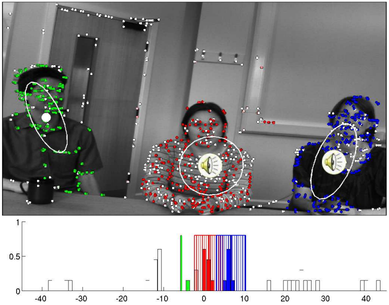
- Model parameters: $\boxed{\text{Determine } N \text{ and } \boldsymbol{s}_1, \ldots, \boldsymbol{s}_N}$

$$\boldsymbol{\theta} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N, \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_N, \pi_1, \ldots, \pi_{N+1}, \lambda_1, \ldots, \lambda_{N}$$

Real Data Results

Meeting scenario

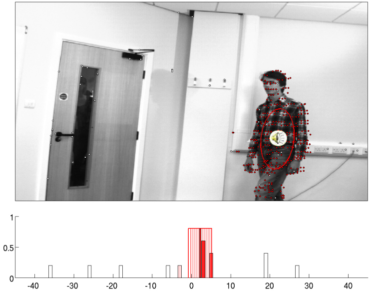


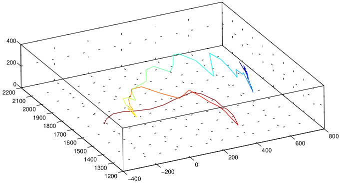


- Estimated speaker locations and their auditory activity for a quasi-stationary scene
- Error rates for auditory activity detection: 'missed target' = 0.16, 'false alarm' = 0.14
- Localization error: within 5cm

Real Data Results

Simple tracking scenario

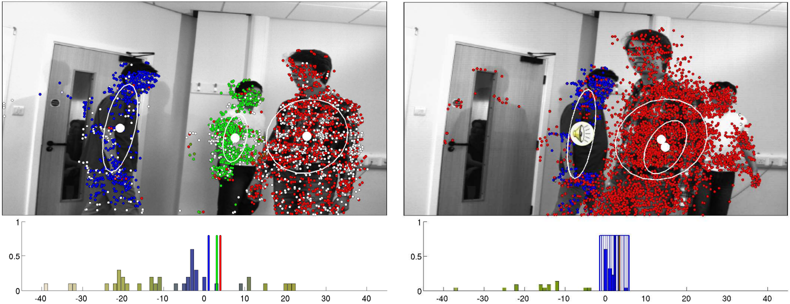




- Simple dynamic scene - results on the previous frame are used to initialize the model for the next frame
- Error rates for auditory activity detection: 'missed target' $=$ 0.13, 'false alarm' $=$ 0.43
- Localization error: within 10cm

Real Data Results

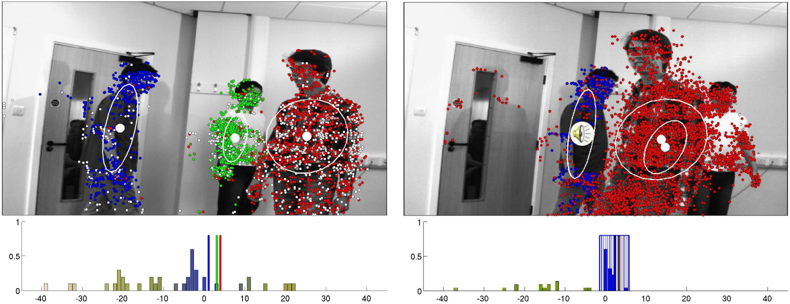
Cocktail party scenario



- Complex dynamic scene - may fail!
- Explicit dynamic model is required!

Real Data Results

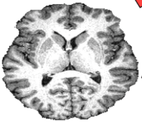
Cocktail party scenario

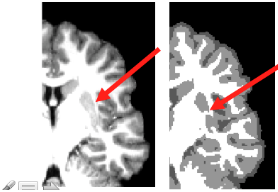


- Complex dynamic scene - may fail!
- Explicit dynamic model is required!

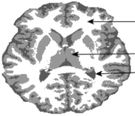
Illustration: MR Brain scan segmentation

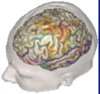
Assign each voxel to a class (label) (among K classes)





Tissue segmentation (WM, GM, CSF)





➔ Cortex 3D reconstruction

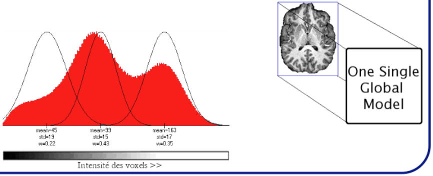
Structure segmentation

➔ Useful for :

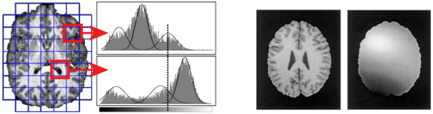
- Distinguishing Cortex GM from Nuclei GM
- volumetric studies
- ...

Tissue segmentation

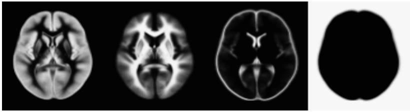
- **Global** estimation of **Gaussian** intensity models for CSF, WM, GM.



- Intensity inhomogeneities are modeled by a unique bias field (eg. Multiplicative) **to be estimated**



Constrain with an atlas



Statistical inference

Statistical inference

From a given set of observation $x = (x_1, \ldots, x_N)$, learn a model that best describes the data

- Probabilistic parametric model:
  $x = (x_1, \ldots, x_N)$ generated from a probability distribution $f(x|\theta)$

  $$x = (x_1, \ldots, x_N) \sim f(x|\theta)$$

  associated likelihood: $l(\theta|x) = f(x|\theta)$ viewed as a function of θ

- Learning: estimating θ
  e.g. that maximizes $l(\theta|x)$ (Maximum likelihood inference)

Decision

Once a model is learned, decide about:

- The occurence of an "event",
- Classify,
- Or find the value of a variable, etc.

Example 1: Linear model

Assume $x = Kz + \epsilon$

$z =$ clean signal, $z \sim f(z|\theta)$

$\epsilon =$ noise, $\epsilon \sim f(\epsilon|\phi)$

$x =$ noisy observed signal

Goal: obtain an estimate for z ($\hat{z}$)

Decision

Example 2: Classification

e.g. 2 groups of objects (people)

$$\theta_1 \longrightarrow f(x|\theta_1) \longrightarrow x \in g_1$$

$$\theta_2 \longrightarrow f(x|\theta_2) \longrightarrow x \in g_2$$

Training data: observations in g_1 and in $g_2 \longrightarrow \hat{\theta}_1, \hat{\theta}_2$

Goal: given x^{new}, decide to which group it belongs

(ie. compute $p(g|x^{new}, \hat{\theta}_1, \hat{\theta}_2)$)

Maximum likelihood estimation

- We observe N realizations $x_1, \dots, x_N$ of a variable X
- Decide on a parametric model for X: $f(x|\theta)$
- Estimate θ by maximizing $l(\theta|x)$ or $\log l(\theta|x)$

Example 1: Linear Gaussian model $z = Kx + \epsilon$ and $\epsilon \sim \mathcal{N}(\mu_\epsilon, \Sigma_\epsilon)$

$$\log f(z|\theta) = \log \mathcal{N}(Kx + \mu_\epsilon, \Sigma_\epsilon) \propto -(z - Kx - \mu_\epsilon)^T \Sigma_\epsilon^{-1} (z - Kx - \mu_\epsilon)$$

$$\hat{x}_{ML} = \arg\min_x (z - Kx - \mu_\epsilon)^T \Sigma_\epsilon^{-1} (z - Kx - \mu_\epsilon)$$

Normal equations: $(K^T \Sigma_\epsilon^{-1} K)\hat{x}_{ML} = K^T \Sigma_\epsilon^{-1}(z - \mu_\epsilon)$

- Least squares: $\mu_\epsilon = 0$ and $\Sigma_\epsilon = \sigma^2 Id$

$$\Longrightarrow \hat{x}_{ML} = \arg\min_x ||z - Kx||_2^2 = \hat{x}_{LS}$$

- Weighted least squares: $\mu_\epsilon = 0$ and $\Sigma_\epsilon = Diag(\sigma_1^2, \dots, \sigma_N^2)$

$$\Longrightarrow \hat{x}_{ML} = \arg\min_x \sum_n \frac{(z_n - [Kx]_n)^2}{\sigma_n^2} = \hat{x}_{WLS}$$

Maximum likelihood estimation

Example 2: Man-Woman classification problem

5 subjects in each class were asked if they like football and statistics

	Women g_1					Men g_2				
football	1	1	0	0	0	1	1	1	1	1
statistics	1	0	1	0	1	0	1	0	0	1

ratio	Positive answers	Negative answers	Positive answers	Negative answers
football	2/5=0.4	3/5=0.6	5/5=1	0/5=0
statistics	3/5=0.6	2/5=0.4	2/5=0.4	3/5=0.6

Observations and notation

- $N = 10$ responses $x_n = \begin{bmatrix} x_{1n} \\ x_{2n} \end{bmatrix} \in \{0,1\}^2$ (2 questions)

$$\mathbf{x} = \{x_1, \ldots, x_N\}$$

- $N = 10$ group assignments $g_n \in \{Woman, Man\} (= \{1,2\})$

$$\mathbf{g} = \{g_1, \ldots, g_N\}$$

$$\mathbf{x}^{g1} = \{x_n, g_n = 1\} = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$$

$$\mathbf{x}^{g2} = \{x_n, g_n = 2\} = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}$$

Model

- Independence: $f(\mathbf{g}) = \prod_{n=1}^{N} f(g_n)$
- $f(g_n = woman) = f(g_n = man) = 0.5$
- Conditional independence: $f(\mathbf{x}|\mathbf{g}) = \prod_{n=1}^{N} f(x_n|g_n)$
- Independence of the two questions: $f(x_n|g_n) = f(x_{1n}|g_n)\, f(x_{2n}|g_n)$
- $\forall n = 1 \ldots N, i = \{1,2\}, g = \{1,2\}$, Independent Bernouilli distributions ($\theta_i^g \in [0,1]$):

$$f(x_{in}|g_n = g) = \begin{cases} \theta_i^g & \text{if } x_{in} = 1 \\ 1 - \theta_i^g & \text{if } x_{in} = 0 \\ 0 & \text{otherwise} \end{cases}$$

or equivalently $f(x_{in}|g_n = g) = (\theta_i^g)^{x_{in}}\, (1 - \theta_i^g)^{1 - x_{in}}$

($\theta_i^g = 0.5 \longrightarrow$ *the coin is not biased*)

Likelihood for each group

- Learning task: Estimate (θ_1^g, θ_2^g) given $\mathbf{x}^g$ $(g = Woman, Man)$
- Likelihood function:

$$f(\mathbf{x}^g|\theta^g) = \prod_{n,\ g_n=g} f(x_{1n}|g)\ f(x_{2n}|g)$$

- Log-likelihood:

$$\log f(\mathbf{x}^g|\theta^g) = \sum_{n,\ g_n=g} \sum_{i=1,2} x_{in} \log \theta_i^g + (1 - x_{in}) \log(1 - \theta_i^g)$$

- Maximization: $\theta_i^g = \frac{\sum_{n, g_n=g} x_{in}}{N_g}$ (mean, frequencies of positive answers)

$$\theta^1 = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}, \theta^2 = \begin{bmatrix} 1 \\ 0.4 \end{bmatrix}$$

Decision: Naive Bayes classifier

Sum-rule: $P(B) = P(B, A) + P(B, A^c)$

Product-rule: $P(A, B) = P(B|A)P(A)$

It follows Bayes' theorem: $P(A|B) = P(B|A)P(A)/P(B)$

with normalization $P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$

Goal: Classify a person with $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ ie. $g =??$

- Bayes' rule : $f(g|x) = \frac{f(x|g)f(g)}{f(x)} = \frac{f(x|g)f(g)}{\sum_{g'} f(g')f(x|g')}$
- Assuming $f(woman) = f(man) = 0.5$,

$$\begin{aligned} f(woman|x) &= \frac{0.4 \times 0.6}{0.4 \times 0.6 + 1 \times 0.4} = 0.375 \\ f(man|x) &= \frac{1 \times 0.4}{0.4 \times 0.6 + 1 \times 0.4} = 0.625 = 1 - 0.375 \end{aligned}$$

Decision: Naive Bayes classifier

Goal: classify a person with $x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

$$
\begin{aligned}
f(woman|x) &= \frac{0.6 \times 0.6}{0.6 \times 0.6 + 0 \times 0.4} = 1 \\
f(man|x) &= \frac{0 \times 0.4}{0.6 \times 0.6 + 0 \times 0.4} = 0
\end{aligned}
$$

- Conclusion: if you don't like football and like statistics, you are almost surely a woman
- Overfitting effect to the small training set
- Priors over the parameters can avoid overfitting $\Longrightarrow$ Bayesian framework

Bayesian set up

Bayesian concepts

- Uncertainty on the parameters θ of a model modeled through a probability distribution on θ, called prior distribution

  The prior encoded the information available a priori, before observing x

- Inference based on the distribution of θ conditional on x, $f(\theta|x)$, called posterior distribution

Impact

- From unknown parameters to random
- Actualization of the information on θ by extracting the information on θ contained in the observations x
- Allows incorporation of imperfect information in the decision process
- Unique mathematical way to condition upon the observations (conditional perspective)
- Penalization factor

Three basic quantities in Bayesian inference

- Prior distribution $f(\theta)$
- likelihood $f(x|\theta)$
- Posterior distribution $f(\theta|x)$

Forward generative model:

$$f(\theta) \longrightarrow \theta \longrightarrow f(x|\theta) \longrightarrow x$$

$\longrightarrow$ involves the prior and the likelihood

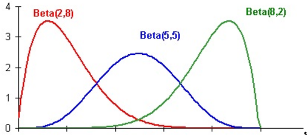
Inference is an inversion problem:

$$x \longrightarrow f(\theta|x) \longrightarrow \hat{\theta}$$

$\longrightarrow$ involves the posterior distribution

Classification example

Assume a Beta prior over the Bernouilli parameters: $\theta_i^g \in [0,1]$



$$f(\theta_i^g) = Beta(\alpha,\beta) = \mathcal{B}(\alpha,\beta)^{-1} {\theta_i^g}^{\alpha-1}(1-\theta_i^g)^{\beta-1}$$

$$E[\theta_i^g] = \alpha/(\alpha+\beta), \; Mode[\theta_i^g] = (\alpha-1)/(\alpha+\beta-2)$$

Compute the posterior distribution of θ_i^g:

$$f(\theta_i^g|\mathbf{x}^g) = \frac{f(\mathbf{x}^g|\theta_i^g)f(\theta_i^g)}{f(\mathbf{x}^g)} \propto f(\mathbf{x}^g|\theta_i^g)f(\theta_i^g)$$

$$\log f(\theta_i^g|\mathbf{x}^g) = cst + (A_i^g - 1)\log\theta_i^g + (B_i^g - 1)\log(1-\theta_i^g)$$

with $A_i^g = \sum_{n,g_n=g} x_{in} + \alpha$ and $B_i^g = \sum_{n,g_n=g}(1-x_{in}) + \beta$

so that posterior distribution:

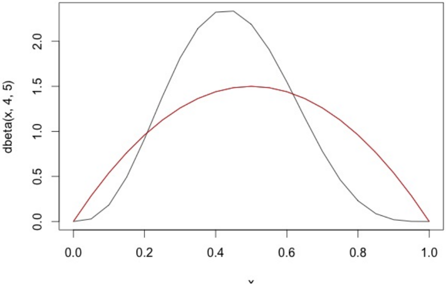
$$\begin{aligned} f(\theta_i^g|\mathbf{x}^g) &\propto {\theta_i^g}^{A_i^g-1}(1-\theta_i^g)^{B_i^g-1} \\ &= Beta(A_i^g, B_i^g) \end{aligned}$$

Classification example

	Women g_1					Men g_2				
football	1	1	0	0	0	1	1	1	1	1
statistics	1	0	1	0	1	0	1	0	0	1

Example: $\alpha = \beta = 2 \Longrightarrow \quad A_1^1 = \alpha + 2 = 4, \quad B_1^1 = \beta + 3 = 5$

$\downarrow \mathcal{B}(4,5)$



$\nwarrow \mathcal{B}(2,2)$

Bayesian inference strategies

Point estimators

Goal: provide an estimation of θ

The two most common Bayesian estimators are:

- Maximum a posteriori (MAP) estimator

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\theta} f(\theta|x) \\ &= \arg\max_{\theta} f(x|\theta)f(\theta) \\ &= \arg\max_{\theta} \log f(x|\theta) + \log f(\theta)\end{aligned}$$

  Note: if $f(\theta) = constant$ then $\hat{\theta}_{MAP} = \hat{\theta}_{MLE}$

- Posterior Mean Estimator

$$\hat{\theta}_{PM} = E_{\theta}[\theta|x] = \int \theta f(\theta|x) d\theta$$

  Note: $f(\theta|x)$ requires the normalizing term $f(x) = \int f(x|\theta)f(\theta)d\theta$.

$\hat{\theta}_{MAP}$ usually easier to obtain, it involves optimization rather than integration

Posterior mean and Bayesian MSE

The Bayesian Mean Square Error (MSE) is

$$E_{\theta,X}[||\hat{\theta} - \theta||_2^2] = \int\int ||\hat{\theta}(x) - \theta||_2^2 \; f(\theta, x) \; d\theta \; dx$$

Minimum Mean Square Error (MMSE) estimator:

Definition: $\hat{\theta}_{MMSE} = \arg\min_{\hat{\theta}} E_{\theta,X}[||\hat{\theta} - \theta||_2^2]$

Solution: $\hat{\theta}_{MMSE} = E_\theta[\theta|X] = \hat{\theta}_{PM}$

since $E_{\theta,X}[||\hat{\theta} - \theta||_2^2] \;=\; E_X[E_\theta[||\hat{\theta} - \theta||_2^2|X]]$

and $E_\theta[||\hat{\theta} - \theta||_2^2|X]$ is minimum when $\hat{\theta} = E_\theta[\theta|X]$

MAP and 0-1 Loss

The MSE quadratic cost (loss) can be replaced by a 0-1 cost

$$E_{\theta,X}[1 - \delta_\theta(\hat{\theta})]$$

where $1 - \delta_\theta(\hat{\theta}) = 0$ if $\hat{\theta} = \theta$ (no loss) and 1 otherwise (max loss)

$$\min E_{\theta,X}[1 - \delta_\theta(\hat{\theta})] \quad = \quad \max E_X[E_\theta[\delta_\theta(\hat{\theta})|X]]$$

and $E_\theta[\delta_\theta(\hat{\theta})|X] = p(\theta = \hat{\theta}|X)$ which is max at the MAP

Linear Minimum MSE

Assume $E[\theta] = E[X] = 0$ and consider an estimator of the form $\hat{\theta} = A^T X$

Goal: find matrix A that minimizes the Bayesian MSE

$$
\begin{aligned}
MSE(A) &= E_{\theta,X}[||A^T X - \theta||_2^2] \\
&= E_{\theta,X}[\text{trace}\,((A^T X - \theta)(A^T X - \theta)^T)] \\
&= \text{trace}\,(E_{\theta,X}[(A^T X - \theta)(A^T X - \theta)^T]) \\
&= \text{trace}\,(E[\theta\theta^T] - A^T E[X\theta^T] - E[\theta X^T]A + A^T E[XX^T]A) \\
&= \text{trace}\,(\Sigma_\theta - A^T \Sigma_{x\theta} - \Sigma_{\theta x} A + A^T \Sigma_x A)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial A} MSE(A) &= -2\Sigma_{x\theta} + 2\Sigma_x A = 0 \\
\hat{A} &= \Sigma_x^{-1} \Sigma_{x\theta} \quad \text{Wiener-Hopf equation}
\end{aligned}
$$

$$
\begin{aligned}
\hat{\theta}_{LMMSE} &= \Sigma_{\theta x} \Sigma_x^{-1} X \quad \text{Wiener filter} \\
(\hat{\theta}_{LMMSE} &= \Sigma_{\theta x} \Sigma_x^{-1} (X - \mu_x) + \mu_\theta \quad \text{in the non centered case})
\end{aligned}
$$

Linear Minimum MSE: Linear model example

Assume $x = K\theta + \epsilon$ where $\theta \sim \mathcal{N}(0, \sigma_\theta^2 I)$ and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$ are independent

Then $x \sim \mathcal{N}(0, \sigma_\theta^2 KK^T + \sigma_\epsilon^2 I)$ and

$$\begin{aligned} \Sigma_x &= \sigma_\theta^2 KK^T + \sigma_\epsilon^2 I \\ \Sigma_{\theta x} &= E[X\theta^T] = E[K\theta\theta^T + \epsilon\theta^T] = K\Sigma_\theta = \sigma_\theta^2 K \end{aligned}$$

$$\hat{\theta}_{LMMSE} = \sigma_\theta^2 K^T(\sigma_\theta^2 KK^T + \sigma_\epsilon^2 I)^{-1} X = K^T(KK^T + \frac{\sigma_\epsilon^2}{\sigma_\theta^2} I)^{-1} X$$

Note: when SNR increases, $\frac{\sigma_\epsilon^2}{\sigma_\theta^2} \rightarrow 0$, $\hat{\theta}_{LMMSE} \rightarrow \hat{\theta}_{MLE} = (KK^T)^{-1} K^T X$

Classification example

MMSE estimator: Since $f(\theta_i^g|\mathbf{x}^g)$ is a Beta distribution

$$E[\theta_i^g|\mathbf{x}^g] = \frac{\sum_{n,g_n=g} x_{in} + \alpha}{\alpha + \beta + N_g}$$

With $\alpha = \beta = 2$ (mode and mean at 0.5), we get

$$\theta^1 = \begin{bmatrix} 4/9 \\ 5/9 \end{bmatrix} \text{ and } \theta^2 = \begin{bmatrix} 7/9 \\ 4/9 \end{bmatrix}$$

Then for $x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, it comes $f(man|x) = 8/33 = 0.242$

MAP estimator: using the mode of the posterior we get instead :

$$\theta^1 = \begin{bmatrix} 3/7 \\ 4/7 \end{bmatrix} \text{ and } \theta^2 = \begin{bmatrix} 6/7 \\ 3/7 \end{bmatrix}$$

and for $x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, it comes $f(man|x) = 3/19 = 0.158$

Predictive distributions

Use the full posterior rather than a point estimate.

Other distributions of interest are:

Prior predictive (marginal):

Before we observe the data, what do we expect the distribution of observations to be?

$$f(x) = \int f(x|\theta) f(\theta) \, d\theta$$

- What we would predict for x given no data
- Useful for assessing whether choice of prior distribution does capture prior beliefs.

Predictive distributions

Posterior predictive

What is the predictive distribution of a new observation x^{new} given the current data $\mathbf{x}$?

$$
\begin{aligned}
f(x^{new}|\mathbf{x}) &= \int f(x^{new}, \theta|\mathbf{x}) \; d\theta \\
&= \int f(x^{new}|\theta) f(\theta|\mathbf{x}) d\theta
\end{aligned}
$$

Use the assumption that x^{new} is independent of $\mathbf{x}$ given θ.

Classification example

In each group, the posterior predictive is:

$$
\begin{aligned}
f(x^{new}|\mathbf{x}^g) &= \int f(x^{new}|\theta^g) f(\theta^g|\mathbf{x}^g) d\theta^g \; = \; f(x_1^{new}|\mathbf{x}_1^g) f(x_2^{new}|\mathbf{x}_2^g) \\
f(x_i^{new}|\mathbf{x}_i^g) &= \int f(x_i^{new}|\theta_i^g) f(\theta_i^g|\mathbf{x}_i^g) d\theta_i^g \\
&= \frac{\mathcal{B}(x_i^{new} + A_i^g, 1 - x_i^{new} + B_i^g)}{\mathcal{B}(A_i^g, B_i^g)}
\end{aligned}
$$

Then using Bayes' rule:

$$
\begin{aligned}
f(g^{new}|x^{new}, \mathbf{x}, \mathbf{g}) &\propto f(x^{new}|\mathbf{x}^{g^{new}}) f(g^{new}) \\
&\propto f(x_1^{new}|\mathbf{x}_1^{g^{new}}) f(x_2^{new}|\mathbf{x}_2^{g^{new}}) \times 0.5
\end{aligned}
$$

For $x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, it comes $f(man|x) = 8/33 = 0.242$

Classification example: all results for $f(man|x = [0,1]^T)$

Estimator	$f(man\|x)$
Maximum likelihood	0
Bayesian MMSE	0.242
Bayesian MAP	0.158
Fully Bayesian	0.242

Prior distributions

From prior information to prior distributions

- All computations depends on the prior choice
- The prior is a tool summarizing available information as well as uncertainty related with this information
- The prior distribution is the key to Bayesian inference but the available prior information is usually not precise enough to lead to an exact determination

Different strategies are possible:

- Conjugate priors
- Noninformative priors
- Jeffreys prior
- Hierarchical modelling, etc.

Conjugate priors: a starting point

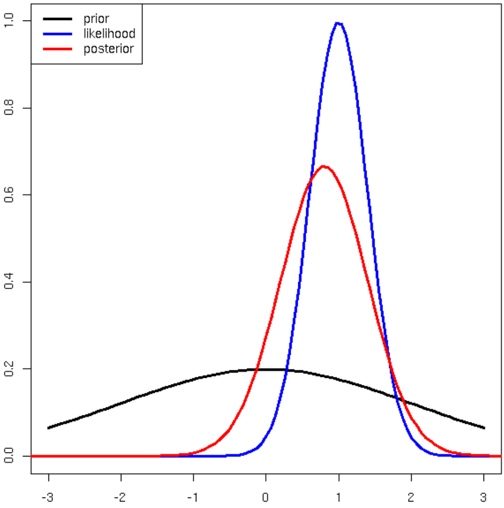
Specific parametric family with convenient analytical properties

Definition: *A family $\mathcal{F}$ of probability distributions on θ is conjugate for a likelihood function $f(x|\theta)$ if, for every $\pi \in \mathcal{F}$, the posterior distribution $f(\theta|x) \propto f(x|\theta)\pi(\theta)$ also belongs to $\mathcal{F}$.*

Main interest is when $\mathcal{F}$ is parametric: computing the posterior distribution reduces then to an updating of the corresponding parameters of the prior.

- The prior "structure" on θ is propagated to the posterior (actualisation)
- Tractability and simplicity
- First approximations to adequate priors

Conjugate priors: Gaussian case



Exponential families

Conjugate priors are usually associated with exponential families of distributions.

Definition: *C, h are positive functions, R, T are functions in R^k*

The family of distributions

$$f(x|\theta) = C(\theta)h(x)\exp(R(\theta)T(x))$$

is called an exponential family of dimension k.

When

$$f(x|\theta) = C(\theta)h(x)\exp(\theta\ x) = \mathbf{h}(\mathbf{x})\exp(\theta\ \mathbf{x} - \mathbf{\Psi}(\theta))$$

the family is said to be natural.

Exponential families

Interesting analytical properties :

- Sufficient statistics of constant dimension exist
- Include common distributions (normal, binomial, Poisson, Wishart, etc.)
- Availability of the moments:

$$E_X[X|\theta] = \nabla\Psi(\theta),\ cov(X_i, X_j) = \frac{\partial^2\Psi}{\partial\theta_i\partial\theta_j}(\theta).$$

- Allow for conjugate priors

Conjugate distributions for exponential families

If $f(x|\theta) = h(x)\exp(\theta\, x - \Psi(\theta))$ then

$$f(\theta|\mu,\lambda) = K(\mu,\lambda)\exp(\theta\mu - \lambda\Psi(\theta))$$

where $K(\mu,\lambda)$ is the normalizing constant, **is conjugate** for $f(x|\theta)$.

The posterior is then $f(\theta|\mu + x, \lambda + 1)$.

It follows an "automatic" way to derive prior from $f(x|\theta)$ BUT μ, λ have still to be specified.

Linearity of the posterior mean

$f(x|\theta)$ in the natural exponential family: $f(x|\theta) = h(x)\exp(\theta\, x - \Psi(\theta))$

$$E_X[X] = m(\theta) = \nabla\Psi(\theta)$$

$f(\theta)$ has a conjugate prior: $f(\theta) \propto \exp(\mu\, x - \lambda\Psi(\theta))$

$$E_\theta[m(\theta)] = \int m(\theta) f(\theta) d\theta = \frac{\mu}{\lambda}$$

If $x_1, \ldots x_N$ i.i.d $f(x|\theta)$ then

$$f(\theta|x_1,\ldots,x_N) \propto f(\theta|x_1) f(x_2|\theta)\ldots f(x_N|\theta) = f(\theta|\mu + \sum_{n=1}^{N} x_n, \lambda + N)$$

.

$$E_\theta[m(\theta)|x_1,\ldots,x_n] = \frac{\mu + \sum_{n=1}^{N} x_n}{\lambda + N}$$

Common conjugate priors

$f(x\|\theta)$	$f(\theta)$	$f(\theta\|x)$
Normal $\mathcal{N}(\theta,\sigma^2)$	Normal $\mathcal{N}(\mu,\tau^2)$	Normal $\mathcal{N}(\frac{\sigma^2\mu+\tau^2 x}{\sigma^2+\tau^2},(\frac{1}{\sigma^2}+\frac{1}{\tau^2})^{-1})$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	Gamma $\mathcal{G}(\alpha+x,\beta+1)$
Gamma $\mathcal{G}(\nu,\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	Gamma $\mathcal{G}(\alpha+\nu,\beta+x)$
Binomial $Bin(n,\theta)$	Beta $B(\alpha,\beta)$	Beta $B(\alpha+x,\beta+n-x)$
Multinomial $\mathcal{M}(\theta_1,\ldots,\theta_K)$	Dirichlet $\mathcal{D}(\alpha_1,\ldots,\alpha_K)$	Dirichlet $\mathcal{D}(\alpha_1+x_1,\ldots,\alpha_K+x_K)$
Normal $\mathcal{N}(\mu,\frac{1}{\theta})$	Gamma $\mathcal{G}(\alpha,\beta)$	Gamma $\mathcal{G}(\alpha+1/2,\beta+(x-\mu)^2/2)$

Non informative priors

How to encode absence of prior knowledge?

Is there such a thing as a default prior when prior information is missing?

In the absence of prior information, prior distributions solely derived from the sample distribution $f(x|\theta)$

Uniform priors (Laplace's priors)

Equiprobability of elementary events: the same likelihood to each value of θ

$$\theta \in \{\theta_1, \ldots, \theta_p\} \quad \longrightarrow f(\theta_i) = \frac{1}{p}$$

Extensions to continuous spaces:

$$f(\theta) \propto 1 \quad (= \text{constant})$$

Examples:

Location parameters: $f(x|\theta) = f(x - \theta) \longrightarrow \quad f(\theta) \propto 1$

Scale parameters: $f(x|\theta) = \frac{1}{\theta} f(\frac{x}{\theta}) \longrightarrow f(\theta) \propto \frac{1}{\theta} \quad (f(\log\theta) \propto 1)$

Some drawbacks

Lack of invariance through reparameterization: $\theta \longrightarrow \eta = g(\theta)$

$$f(\theta) \propto 1 \longrightarrow f(\eta) \propto \left|\frac{dg^{-1}(\eta)}{d\eta}\right| \neq constant \quad \text{(Jacobian formula)}$$

Information is not missing anymore !!

May generate improper posterior:

$$x \sim \mathcal{N}(\theta, \sigma^2) \quad \text{with } f(\theta, \sigma^2) \propto 1$$

Then

$$f(\theta, \sigma^2|x) \propto f(x|\theta) \propto \sigma^{-1} \exp(\frac{(x-\theta^2)^2}{2\sigma^2})$$

$$\Longrightarrow f(\sigma^2|x) \propto 1 \quad \text{is improper, paradoxes occur}$$

$\Longrightarrow$ Invariant priors

$\Longrightarrow$ Jeffreys' priors as an alternative

The Jeffreys' priors

Based on Fisher information

Univariate case:

$$I(\theta) = E_X\left[\left(\frac{\partial \log f(X|\theta)}{\partial \theta}\right)^2\right] = -E_X\left[\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2}\right]$$

Multivariate case:

$$I(\theta)_{ij} = -E_X\left[\frac{\partial^2 \log f(X|\theta)}{\partial \theta_i \partial \theta_j}\right]$$

The Jeffreys' prior distribution is $f(\theta) \propto |I(\theta)|^{1/2}$

where $|I(\theta)|$ is the determinant of the Fisher Information matrix

Exponential family: if $f(x|\theta) = h(x)\exp(\theta\, x - \Psi(\theta))$ then

$I(\theta) = \nabla^2\Psi(\theta)$ and $f(\theta) \propto \left(\frac{\partial^2\Psi(\theta)}{\partial\theta_i^2}\right)^{1/2}$

Key feature: Reparameterization invariance

Assume $f(\theta) \propto |I(\theta)|^{1/2}$ and $\eta = g(\theta)$ for a 1-to-1 mapping g

$$
\begin{aligned}
f(\eta) &= f(\theta)\left|\frac{\partial\theta}{\partial\eta}\right| \propto \sqrt{|I(\theta)|\left(\frac{\partial\theta}{\partial\eta}\right)^2} \\
&\propto \sqrt{E_X\left[\left(\frac{\partial \log f(X|\theta)}{\partial\theta}\right)^2 (\frac{\partial\theta}{\partial\eta})^2\right]} \\
&\propto \sqrt{E_X\left[\left(\frac{\partial \log f(X|\theta)}{\partial\eta}\right)^2\right]} \\
&\propto |I(\eta)|^{1/2}
\end{aligned}
$$

Other features

- Information based: $I(\theta)$ corresponds to the amount of information brought by the model on θ.
  Noninformative: Minimize the effect of the prior which is in accordance with the model.
- Violates the likelihood principle
- Usually improper
- May lead to incoherences in multidimensional case
- Have been generalized into reference priors (Berger and Bernardo) by distinguishing between nuisance and interest parameters

Example: $x \sim \mathcal{N}(\mu, \sigma)$

- $\theta = (\mu, \sigma)$ unknown: $f(\theta) \propto 1/\sigma^2$

because
$$I(\theta) = E_X\left[\begin{pmatrix} 1/\sigma^2 & 2(x-\mu)/\sigma^3 \\ 2(x-\mu)/\sigma^3 & 3(x-\mu)^2/\sigma^4 - 1/\sigma^2 \end{pmatrix}\right] = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$$

- $\theta = \mu$, σ fixed: $f(\mu) \propto 1$
- $\theta = \sigma$, μ fixed: $f(\sigma) \propto 1/\sigma$
- μ and σ a priori independent: $f(\theta) = f(\mu)f(\sigma) \;\propto\; 1/\sigma$

Hierarchical modelling

Consider a conjugate prior for $f(x|\theta)$: $f_1(\theta|\lambda)$

$f_1(\theta|\lambda)$ may be too restrictive and require specification of λ.

λ unknown $\longrightarrow$ add a noninformative prior on λ:

$$\begin{array}{rcl} \lambda & \sim & f_2(\lambda) \\ \theta|\lambda & \sim & f_1(\theta|\lambda) \\ x|\theta & \sim & f(x|\theta) \end{array}$$

The prior on θ is then $f(\theta) = \int f_1(\theta|\lambda) f_2(\lambda)\, d\lambda$

- not conjugate anymore
- heavier tails (eg. Student distributions or Gaussian scale mixtures)
- Computationaly flexible

Posterior distributions

Computing posterior distributions

Posteriors are not always tractable...

Observed data: $x = \{x_1, \ldots, x_N\}$ eg. a discretized signal

Hidden variables: $z = \{z_1, \ldots, z_M\}$. eg. a segmentation or a clean version of x

Add prior knowledge on z but if the dependence structure in z is too complex (eg an image), $f(z|x)$ can't be obtained analytically

Solution: "Approximate" the dependence structure

- Sampling methods (Gibbs sampler, MCMC)
- Approximations (Laplace, Variational Bayes, EP)

Conclusion

- Maximum likelihood for large training data. Risk of overfitting for small data set.
- Bayesian framework to incorporate prior information (eg temporal dynamics, spatial relationships) and prevent overfitting
- MMSE and MAP provide point estimates that use prior information
- For fully Bayesian treatment, use predictive distributions
- If posterior distributions are not tractable, use sampling methods (eg MCMC) or approximate inference (eg Variational Bayes)

Main references

- C. P. Robert. The Bayesian Choice. Springer texts in statistics.
- A. Gelman, J. B. Carlin, H. S. Stern, D. Rubin. Bayesian data analysis. Texts in statistical science. Chapman & Hall CRC.