Bayesian computation

Dr. Yoann Altmann

B39AX – Fall 2023
Heriot-Watt University

LEADERS IN IDEAS AND SOLUTIONS

Plan

- Likelihood and Maximum Likelihood
- Linear regression
- Bayesian modelling
- Bayesian estimation
  - MAP estimation
  - MMSE estimation

Bayesian “vs” frequentist

- Common problems
  - Decision between hypotheses given some observations $\boldsymbol{y}$
  - Choosing the “best” model able to explain the observations $\boldsymbol{y}$
  - Estimation of some parameters $\boldsymbol{x}$ given some observations $\boldsymbol{y}$
- Alternative notations
  - $\boldsymbol{x}$ for observations and $\boldsymbol{\theta}$ for parameters

Bayesian “vs” frequentist

- Different interpretations
  - Frequentist: the unknown parameters are deterministic
  - Bayesian: the unknown parameters are RVs
  - Pro/cons for both approaches
- Here, we will adopt a Bayesian approach and use prior distributions
- But first, let’s adopt a frequentist approach

PMF/PDF

- Until now, we made the difference between
  - pmf $\mathbb{p}(\boldsymbol{x})$ (discrete variables)
  - pdf $f_X(\boldsymbol{x})$ (continuous variables)
- But a vector $\boldsymbol{x}$ can also contain both discrete and continuous variables
- So the notations above can quickly become impractical

Bayes rule(s)

- $\boldsymbol{y}$ discrete, $\boldsymbol{x}$ discrete

  $$\mathbb{p}_{X|Y}(\boldsymbol{x}|\boldsymbol{y}) = \frac{\mathbb{p}_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\mathbb{p}_X(\boldsymbol{x})}{\sum_{\boldsymbol{x}'}\mathbb{p}_{Y|X}(\boldsymbol{y}|\boldsymbol{x}')\mathbb{p}_X(\boldsymbol{x}')} = \frac{\mathbb{p}_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\mathbb{p}_X(\boldsymbol{x})}{\mathbb{p}_Y(\boldsymbol{y})}$$
- $\boldsymbol{y}$ discrete, $\boldsymbol{x}$ continuous
- $$f_{X|Y}(\boldsymbol{x}|\boldsymbol{y}) = \frac{\mathbb{p}_{Y|X}(\boldsymbol{y}|\boldsymbol{x})f_X(\boldsymbol{x})}{\int_{\boldsymbol{x}'}\mathbb{p}_{Y|X}(\boldsymbol{y}|\boldsymbol{x}')f_X(\boldsymbol{x}')d\boldsymbol{x}'} = \frac{\mathbb{p}_{Y|X}(\boldsymbol{y}|\boldsymbol{x})f_X(\boldsymbol{x})}{\mathbb{p}_Y(\boldsymbol{y})}$$
- $\boldsymbol{y}$ continous, $\boldsymbol{x}$ discrete
- $$\mathbb{p}_{X|Y}(\boldsymbol{x}|\boldsymbol{y}) = \frac{\mathbb{p}_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\mathbb{p}_X(\boldsymbol{x})}{\sum_{\boldsymbol{x}'}f_{Y|X}(\boldsymbol{y}|\boldsymbol{x}')\mathbb{p}_X(\boldsymbol{x}')} = \frac{f_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\mathbb{p}_X(\boldsymbol{x})}{f_Y(\boldsymbol{y})}$$
- $\boldsymbol{y}$ continous, $\boldsymbol{x}$ continuous
- $$f_{X|Y}(\boldsymbol{x}|\boldsymbol{y}) = \frac{f_{Y|X}(\boldsymbol{y}|\boldsymbol{x})f_X(\boldsymbol{x})}{\int_{\boldsymbol{x}'}f_{Y|X}(\boldsymbol{y}|\boldsymbol{x}')f_X(\boldsymbol{x}')d\boldsymbol{x}'} = \frac{f_{Y|X}(\boldsymbol{y}|\boldsymbol{x})f_X(\boldsymbol{x})}{f_Y(\boldsymbol{y})}$$

Notations

- In the remainder of this chapter, we will use
  - $f_X(\boldsymbol{x})$ for the pmf/pdf of the RV $\boldsymbol{x}$ except in specific examples
  - Marginalisation will be denoted by integrals ($\int$) which can be replaced by sums for discrete RVs

Notations

- In the remainder of this chapter, unless stated otherwise, we will use
  - $\boldsymbol{y}$: observations
  - $\boldsymbol{x}$: unknown parameters of interest
  - $\boldsymbol{\theta}$: additional (hyper)-parameters
- Other classical notations (not used here)
  - $\boldsymbol{x}$: observations
  - $\boldsymbol{\theta}$: unknown parameters of interest

Likelihood

- Goal: Estimation of some parameters $\boldsymbol{x}$ given some observations $\boldsymbol{y}$
- **Model for noisy measurements**
- $\boldsymbol{y}|\boldsymbol{x} \sim f_{Y|X}(\boldsymbol{y}|\boldsymbol{x})$
- Examples:

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}$$
$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$$
$$\boldsymbol{y} = g(\boldsymbol{x}) + \boldsymbol{n}$$

Likelihood

- AWGN noise model
  - $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n},\ \boldsymbol{y} \in \mathbb{R}^N$
  - $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{n}; \boldsymbol{0}, \sigma^2 \boldsymbol{I}) \ \rightarrow \boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{y}; \boldsymbol{x}, \sigma^2 \boldsymbol{I})$

- Likelihood function

$$f_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} e^{-\frac{\|\boldsymbol{y}-\boldsymbol{x}\|_2^2}{2\sigma^2}}$$

"how likely is the observation $\boldsymbol{y}$ given some value of $\boldsymbol{x}$"

Maximum likelihood estimator

$$\hat{x}_{MLE} = argmax_{\boldsymbol{x}} \;\; f_{Y|X}(\boldsymbol{y}|\boldsymbol{x})$$
$$= argmax_{\boldsymbol{x}} \;\; \log\left(f_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\right)$$

- Example: $y_n = x_0 + n_n, \forall n$ and $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{n}; \boldsymbol{0}, \sigma^2 \boldsymbol{I}_N)$ (σ^2 known, x_0 unknown)
- $f_{Y|X}(\boldsymbol{y}|x_0) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} e^{-\frac{\sum_{n=1}^{N}(y_n - x_0)^2}{2\sigma^2}}$
- $\hat{x}_{MLE} = \frac{1}{\mathrm{N}} \sum_{n=1}^{N} y_n$ (sample mean)

Maximum likelihood estimator

- Some properties
  - Consistency

    $\widehat{\boldsymbol{x}}_{MLE} \to \boldsymbol{x}$ in probability when $N \to \infty$

    $$\lim_{N\to\infty} \mathbb{p}\left(\|\widehat{\boldsymbol{x}}_{MLE} - \boldsymbol{x}_0\| > \epsilon\right) = 0, \forall \epsilon > 0$$

  - Asymptotic Normality

    $\sqrt{N}(\widehat{\boldsymbol{x}}_{MLE} - \boldsymbol{x}_0) \to \mathcal{N}(0, F^{-1})$ (in distribution)

    F: Fisher information matrix (advanced concept)

    Rq: $\widehat{\boldsymbol{x}}_{MLE}$ is asymptotically unbiased

Computation of the MLE

$$\hat{x}_{MLE} = argmax_{\boldsymbol{x}} \;\; f_{Y|X}(\boldsymbol{y}|\boldsymbol{x})$$

- If $f_{Y|X}(\boldsymbol{y}|\cdot)$ is differentiable (and $\boldsymbol{x}$ continuous)
  - Find analytically the zero(s) of the gradient $\nabla f_{Y|X}(\boldsymbol{y}|\cdot)$ (local extrema)
- If not possible, find a solution numerically (we will discuss this later)
- Sometimes it is easier to work with $\log\left(f_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\right)$.

Examples (revisited)

- AWGN noise model
  - $y_n = x_0 + n_n, \forall n = 1, \dots, N$
  - $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{n}; \boldsymbol{0}, \sigma^2 \boldsymbol{I}) \rightarrow \boldsymbol{y}|x_0 \sim \mathcal{N}(\boldsymbol{y}; x_0 \boldsymbol{1}, \sigma^2 \boldsymbol{I})$
  - If σ^2 is known but x_0 is unknown

$$\hat{x}_{MLE} = \frac{1}{\mathrm{N}} \sum_{n=1}^{N} y_n \,, \hat{x}_{MLE} \sim \mathcal{N}\left(x_0, \frac{\sigma^2}{N}\right)$$

$$\mathbb{E}[\hat{x}_{MLE}] = x_0$$

unbiased

Examples (revisited)

- AWGN noise model
  - $y_n = x_0 + n_n, \forall n = 1, \ldots, N$
  - $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{n}; \boldsymbol{0}, \sigma^2 \boldsymbol{I}) \rightarrow \boldsymbol{y}|x_0 \sim \mathcal{N}(\boldsymbol{y}; x_0 \boldsymbol{1}, \sigma^2 \boldsymbol{I})$
  - If σ^2 is unknown but x_0 is known

$$\hat{\sigma}_{MLE}{}^2 = \frac{\sum_{n=1}^{N} (y_n - x_0)^2}{N}$$

$$\mathbb{E}\left[\hat{\sigma}_{MLE}{}^2\right] = \sigma^2$$

unbiased

Examples (revisited)

- AWGN noise model
  - $y_n = x_0 + n_n, \forall n = 1, \dots, N$
  - $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{n}; \boldsymbol{0}, \sigma^2 \boldsymbol{I}) \rightarrow \boldsymbol{y}|x_0 \sim \mathcal{N}(\boldsymbol{y}; x_0 \boldsymbol{1}, \sigma^2 \boldsymbol{I})$
  - If (x_0, σ^2) is unknown

$$\hat{x}_{MLE} = \frac{1}{\mathrm{N}} \sum_{n=1}^{N} y_n, \qquad \hat{\sigma}_{MLE}{}^2 = \frac{\sum_{n=1}^{N} (y_n - \hat{x}_{MLE})^2}{N}$$

$$\mathbb{E}[\hat{x}_{MLE}] = x_0, \qquad \mathbb{E}\left[\hat{\sigma}_{MLE}{}^2\right] = \frac{N-1}{N} \sigma^2$$

unbiased

biased but asymptotically unbiased

Verification using Matlab

- Demo1.m

Linear regression

- AWGN noise model (known variance)
  - $y_n = \boldsymbol{a}_n^T \boldsymbol{x} + n_n, \forall n = 1, \dots, N$
  - $\{\boldsymbol{a}_n\}_n \in \mathbb{R}^D$ with $D < N$
  - $\boldsymbol{A} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_N]^T$ is a known $N \times D$ matrix (full-rank)
  - How to estimate $\boldsymbol{x}$ via MLE?

$$f_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} e^{-\frac{\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2}{2\sigma^2}}$$

Linear regression

- Least-square regression

$$\hat{\boldsymbol{x}}_{MLE} = argmax_{\boldsymbol{x}} \ \log\left(f_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\right)$$

$$\hat{\boldsymbol{x}}_{MLE} = argmin_{\boldsymbol{x}} \ \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$$

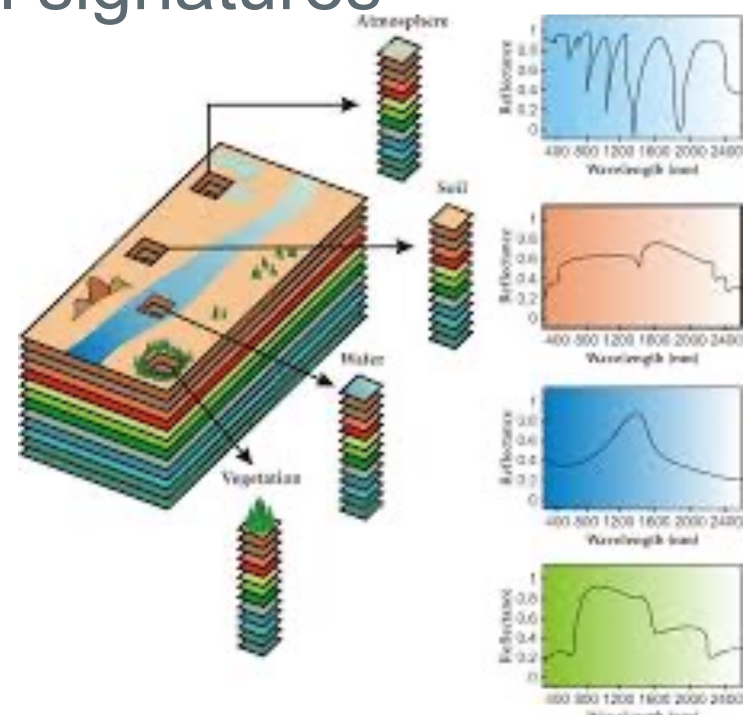
- Using the pseudo inverse

$\boldsymbol{A}^{+} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$, $\mathrm{D}\times N$ matrix

$$\hat{\boldsymbol{x}}_{MLE} = \boldsymbol{A}^{+}\boldsymbol{y}$$

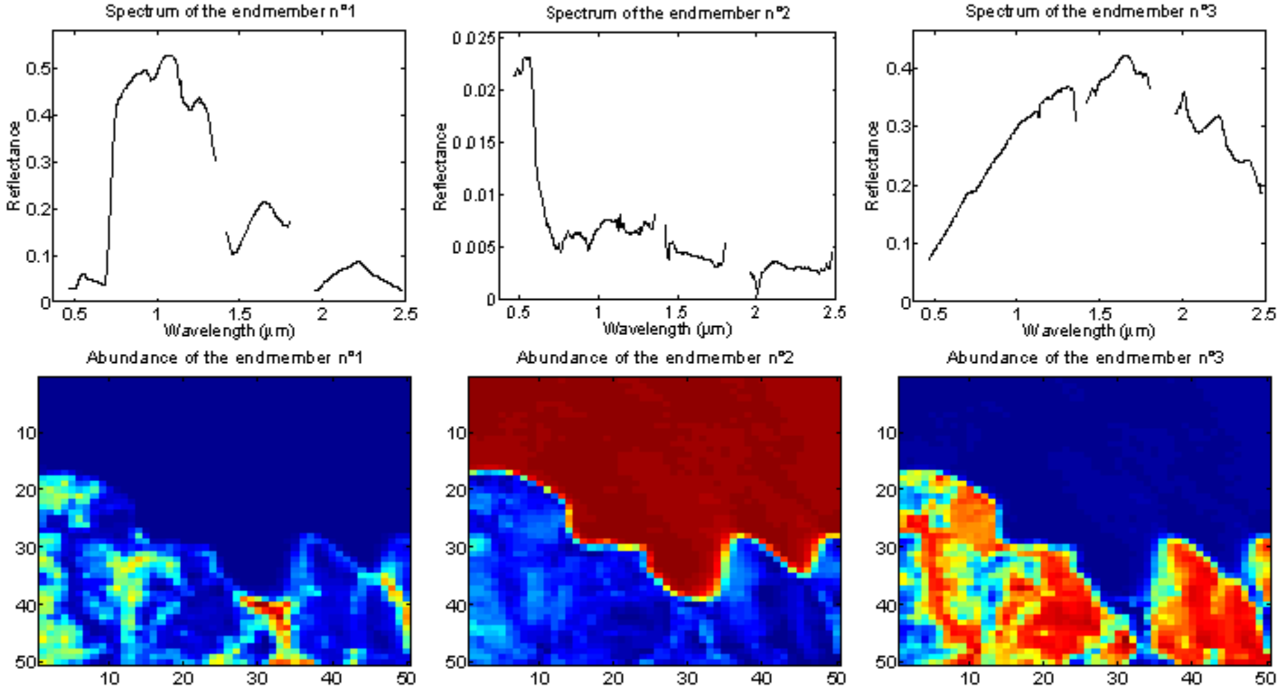
Example of linear regression

- Spectral unmixing for earth observation
  - $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N]^T$: spectral signatures
  - $\boldsymbol{y}$: observed spectrum
  - $\boldsymbol{x}$: material fractions or abundances



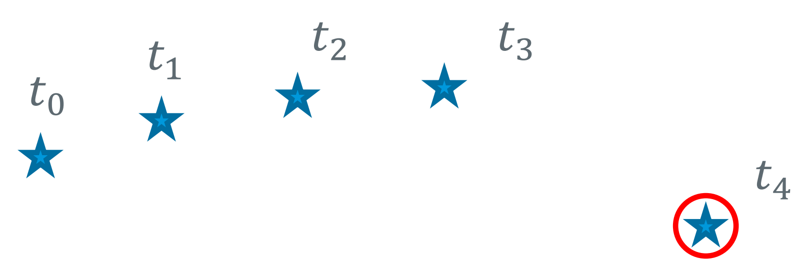
Example of linear regression

- Spectral unmixing for earth observation



Beyond MLE

- MLE: “Good” estimator if N is large
- But … large variance if N “small”
- What if we have additional information about $\boldsymbol{x}_0$?
- Can we improve the estimation of $\boldsymbol{x}$?
- Example: noisy trajectory



Should we use/trust this measurement?

Bayesian estimation

- Instead of treating $\boldsymbol{x}$ as a deterministic but unknown variable, it is seen as a RV
- Additional information expressed as a distribution, called prior distribution $f_X(\boldsymbol{x})$ (or $\mathbb{p}(\boldsymbol{x})$)
- $f_X(\boldsymbol{x})$: what we know about $\boldsymbol{x}$ *prior to observing* $\boldsymbol{y}$
  - Range (e.g., positivity)
  - Mean
  - Smoothness, sparsity,…

Bayes rule revisited

- Likelihood: $f_{Y|X}(\boldsymbol{y}|\boldsymbol{x})$
- Prior distribution: $f_X(\boldsymbol{x})$
- Posterior distribution: $f_{X|Y}(\boldsymbol{x}|\boldsymbol{y}) = \frac{f_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) f_X(\boldsymbol{x})}{f_Y(\boldsymbol{y})}$

What we know about $\boldsymbol{x}$ after having observed $\boldsymbol{y}$

- Evidence: $f_Y(\boldsymbol{y})$ (also marginal likelihood)

Bayesian estimation

$$f_{X|Y}(\boldsymbol{x}|\boldsymbol{y}) = \frac{f_{Y|X}(\boldsymbol{y}|\boldsymbol{x})f_X(\boldsymbol{x})}{f_Y(\boldsymbol{y})}$$

- What we know about $\boldsymbol{x}$ after having observed $\boldsymbol{y}$
- Combines the data and our prior knowledge

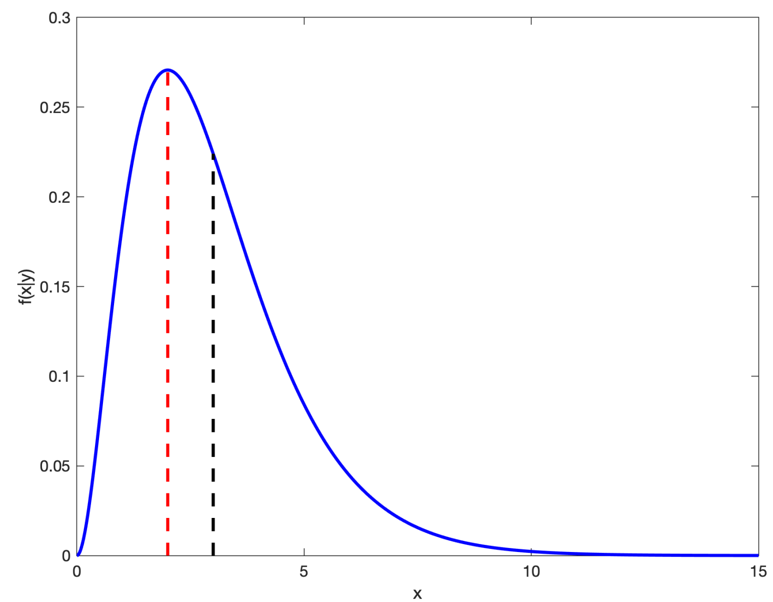
- How can we use this information?

Bayesian estimation

- Example:

$f_{X|Y}(\boldsymbol{x}|\boldsymbol{y})$: gamma distribution $G(x; 3,1)$

Which point estimator or summary statistics should we use?



Maximum a posteriori (MAP)

$$\hat{\boldsymbol{x}}_{MLE} = argmax_{\boldsymbol{x}} \;\; f_{Y|X}(\boldsymbol{y}|\boldsymbol{x})$$
$$\hat{\boldsymbol{x}}_{MAP} = argmax_{\boldsymbol{x}} \;\; f_{X|Y}(\boldsymbol{x}|\boldsymbol{y})$$
$$= argmax_{\boldsymbol{x}} \;\; \log\left(f_{X|Y}(\boldsymbol{x}|\boldsymbol{y})\right)$$

$\hat{\boldsymbol{x}}_{MLE}$: most likely $\boldsymbol{x}$ given the data $\boldsymbol{y}$ only

$\hat{\boldsymbol{x}}_{MAP}$: most likely $\boldsymbol{x}$ given the data $\boldsymbol{y}$ and our prior knowledge

Maximum a posteriori (MAP)

- Important remark (for model selection):

If x is discrete and takes only a finite number of values, the MAP rule/estimator minimizes (over all decision rules) the probability of selecting an incorrect hypothesis.

Example: Bayesian model selection

- Likelihood: $y_n = x_0 + n_n, \forall n = 1, \dots, N$
  - $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{n}; \boldsymbol{0}, \sigma^2 \boldsymbol{I})$ with σ^2 known
  - $\mathbb{p}(x_0 = 1) = \pi$, $\mathbb{p}(x_0 = 2) = 1 - \pi$
- Given N observations $y_1, \dots, y_N$ and our prior knowledge, we would like to decide between $x_0 = 1$ and $x_0 = 2$
- We decide $x_0 = 1$ if

$$\mathbb{p}(x_0 = 1|\boldsymbol{y}) > \mathbb{p}(x_0 = 2|\boldsymbol{y})$$

i.e.,

$$\mathbb{p}(x_0 = 1) f_{Y|X_0}(\boldsymbol{y}|x_0 = 1) > \mathbb{p}(x_0 = 2) f_{Y|X_0}(\boldsymbol{y}|x_0 = 2)$$

Computation of the MAP estimate

$$\widehat{\boldsymbol{x}}_{MAP} = argmax_{\boldsymbol{x}} \;\; f_{X|Y}(\boldsymbol{x}|\boldsymbol{y})$$

- If $f_{X|Y}(\cdot\,|\boldsymbol{y})$ is differentiable (and $\boldsymbol{x}$ continuous)
  - – Find analytically the zero(s) of the gradient $\nabla f_{X|Y}(\cdot\,|\boldsymbol{y})$ (local extrema)
- If not possible, find a solution numerically
  - Using optimization algorithms
  - Using simulation methods (later)
- Sometimes, it is easier to work with $\log\left(f_{X|Y}(\boldsymbol{x}|\boldsymbol{y})\right)$

Computation of the MAP estimate

$$\widehat{\boldsymbol{x}}_{MAP} = argmax_{\boldsymbol{x}} \;\; f_{X|Y}(\boldsymbol{x}|\boldsymbol{y}) = argmax_{\boldsymbol{x}} \, f_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) f_X(\boldsymbol{x})$$

$$\widehat{\boldsymbol{x}}_{MAP} = argmin_{\boldsymbol{x}} \; -\log\left(f_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\right) - \log\left(f_X(\boldsymbol{x})\right)$$

- Often easier to solve as nearly quadratic functions
- MAP estimation as the maximization of a penalised likelihood
- The term $-\log\left(f_X(\boldsymbol{x})\right)$ acts as a penalty or regularisation

Example

- AWGN noise model
  - $y_n = x_0 + n_n, \forall n = 1, \dots, N$
  - $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{n}; \boldsymbol{0}, \sigma^2 \boldsymbol{I}) \ \rightarrow \boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{y}; \boldsymbol{x}, \sigma^2 \boldsymbol{I})$
  - If σ^2 is known but x_0 is unknown
  - Prior distribution: $\mathcal{N}(x_0; m, s^2)$

$\hat{x}_{MLE} = \frac{1}{\mathrm{N}} \sum_{n=1}^{N} y_n$, $\hat{x}_{MAP} =?$

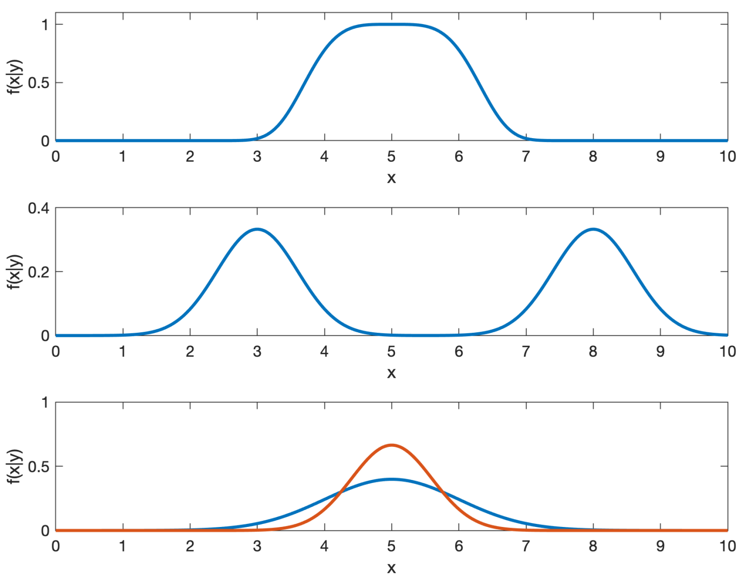
$\hat{x}_{MAP} = \gamma^2 \left(\frac{\sum_{n=1}^{N} y_n}{\sigma^2} + \frac{m}{s^2} \right)$ with $\gamma^2 = \frac{\sigma^2 s^2}{\sigma^2 + N s^2}$

Verification using Matlab

- Demo2.m

Limitations of the MAP estimator

- Useful estimator, obtained via optimization
- But…might be difficult to compute
  - Flat gradient / numerical errors
- Might not be unique
  - Multimodal/flat distribution
- Provides limited information
  - Other solutions almost as likely?



Alternative estimator

- Posterior mean or minimum mean square error (MMSE) estimator

$$\hat{\boldsymbol{x}}_{MMSE} = \mathbb{E}_{f(\boldsymbol{x}|\boldsymbol{y})}[\boldsymbol{x}] = \int \boldsymbol{x} f(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x}$$

under weak conditions on $f(\boldsymbol{x}|\boldsymbol{y})$ (e.g., existence of mean and variance)

Properties of the MMSE estimator

By definition, it minimizes the mean square error

$$MSE = \mathbb{E}[(\hat{\boldsymbol{x}} - \boldsymbol{x})^T(\hat{\boldsymbol{x}} - \boldsymbol{x})]$$

where the expectation is taken over $\boldsymbol{x}$ and $\boldsymbol{y}$.

The MMSE estimator is unbiased and asymptotically Gaussian

Computation of the MMSE estimator

- Analytically,

$$\widehat{\boldsymbol{x}}_{MMSE} = \mathbb{E}_{f(\boldsymbol{x}|\boldsymbol{y})}[\boldsymbol{x}] = \int \boldsymbol{x} f(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x}.$$

- Often the integral is not tractable
  - One possibility is to approximate the expectations using simulation tools (Monte Carlo sampling)
  - Another possibility is to simply the estimation by imposing additional constraints.

Principle of Monte Carlo sampling

To approximate

$$\mathbb{E}[g(\boldsymbol{x})] = \int g(\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x}$$

Monte Carlo sampling consists of generating a large number N_{MC} of random variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_{MC}}$ from the distribution $f(\boldsymbol{x}|\boldsymbol{y})$. We then obtain

$$\frac{1}{N_{MC}} \sum_{n=1}^{N_{MC}} g(\boldsymbol{x}_n) \approx \mathbb{E}[g(\boldsymbol{x})]$$

Efficient methods to sample for complex distributions $f(\boldsymbol{x}|\boldsymbol{y})$ is still an open domain of research !

Principle of Monte Carlo sampling

- Estimation of mean and variance of arbitrary distribution
- Demo3.m

Linear MMSE estimator

- If we enforce $\hat{\boldsymbol{x}}_{MMSE} = \boldsymbol{W}\boldsymbol{y} + \boldsymbol{b}$ (linear function of $\boldsymbol{y}$), the problem

$$\min_{\hat{\boldsymbol{x}}} MSE, \quad \text{s.t. } \hat{\boldsymbol{x}}_{MMSE} = \boldsymbol{W}\boldsymbol{y} + \boldsymbol{b}$$

can become easier than the original problem, with a negligible performance degradation.

MAP vs MMSE estimation

- For discrete RVs, the MMSE estimator might not be meaningful

Ex1: $x \in \{0,1\}, \mathbb{E}[x] \in [0,1]$

can make sense.

Ex2: $x \in \{cat, dog\}$, $\mathbb{E}[x]$ does not make sense

In this case, the MAP estimator is more adapted

MAP vs MMSE estimation

- For continuous variables, MAP and MMSE estimation can be complementary
  - MAP: most probable solution
  - MMSE: minimizes the MSE
    - Flat, skewed or multimodal distributions
    - Sometimes, the posterior covariance can also be computed (e.g., using MC sampling)
  - If the mean and mode of the posterior distribution coincide, the two estimators are the same
    - Example: Gaussian distribution

Summary

- Likelihood and Maximum Likelihood
- Linear regression
- Bayesian modelling
- Bayesian estimation
  - MAP estimation
  - MMSE estimation

- Next chapter: Introduction to Information Theory