Introduction to Bayesian methods and their use in fusion data analysis

Geert Verdoolaege

Department of Applied Physics, Ghent University, Ghent, Belgium

IAEA Technical Meeting on Fusion Data Processing, Validation and Analysis
Nov. 29, 2021

GHENT UNIVERSITY

EUROfusion

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014-2018 and 2019-2020 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

Overview

1. Classical probability and statistics
2. Principles of Bayesian probability theory
3. Applications
   - Classification
   - Regression analysis
   - Some other applications
4. Conclusions and references

Overview

1. Classical probability and statistics

2. Principles of Bayesian probability theory

3. Applications
    - Classification
    - Regression analysis
    - Some other applications

4. Conclusions and references

Frequentist probability

- Probability = frequency
- Straightforward:
  - Number of 1s in 60 dice throws ≈ 10: $p = 1/6$
  - Probability of plasma disruption $p \approx N_{\text{disr}}/N_{\text{tot}}$
- Less straightforward:
  - Probability of fusion electricity by 2050?
  - Probability of mass of Saturn $90\, m_A \leq m_S < 100\, m_A$?

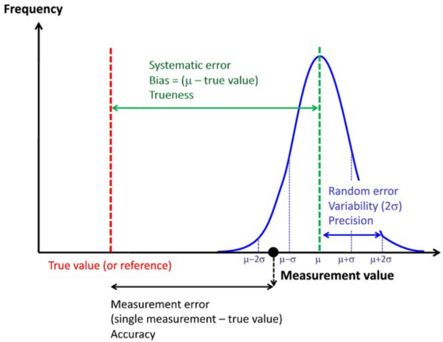
Flavors of uncertainty

Aleatoric/*statistical*/*random* **uncertainty**

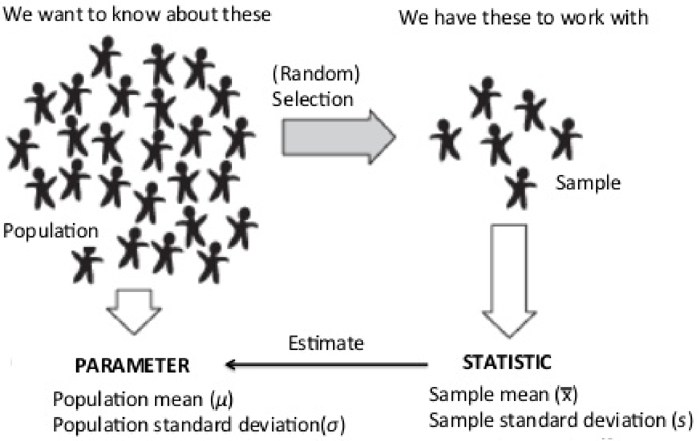
- Leads to different outcomes in multiple experimental trials
- Can be reduced by repeating measurement

Epistemic/*systematic* **uncertainty**

- 'Fixed' but unknown ('bias')
- Cannot be reduced through repeated measurement



Populations vs. sample



Statistics

- E.g. weight w of Belgian men: unknown but *fixed* for every individual
- Average weight μ_w in population?
- ***Random variable*** W
- Sample: $W_1, W_2, \ldots, W_n$
- Average weight: ***statistic*** (estimator) $\overline{W}$
- Central limit theorem:

$$W \sim p(W|\mu_{\mathrm{w}}, \sigma_{\mathrm{w}}) \Rightarrow \overline{W} \sim \mathcal{N}(\mu_{\mathrm{w}}, \sigma_{\mathrm{w}}/\sqrt{n})$$

Maximum likelihood parameter estimation

- ***Maximum likelihood*** (ML) principle:

$$\hat{\mu}_{\mathrm{w}} = \arg\max_{\mu_{\mathrm{w}} \in \mathbb{R}^+} p(W_1, \ldots, W_n | \mu_{\mathrm{w}}, \sigma_{\mathrm{w}})$$

$$\approx \arg\max_{\mu_{\mathrm{w}} \in \mathbb{R}^+} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{w}}} \exp\left[-\frac{(W_i - \mu_{\mathrm{w}})^2}{2\sigma_{\mathrm{w}}^2}\right]$$

$$= \arg\max_{\mu_{\mathrm{w}} \in \mathbb{R}^+} \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{w}}} \exp\left[-\sum_{i=1}^{n} \frac{(W_i - \mu_{\mathrm{w}})^2}{2\sigma_{\mathrm{w}}^2}\right]$$

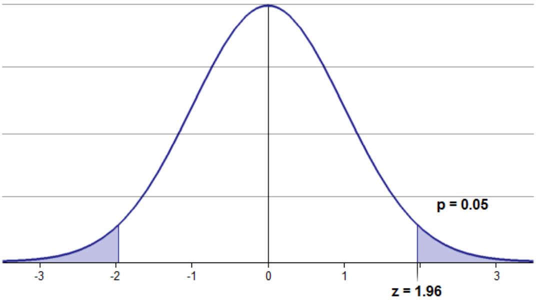
- ML estimator (known σ_{w}):

$$\hat{\mu}_{\mathrm{w}} = \overline{W} = \frac{1}{n} \sum_{i=1}^{n} W_i$$

Frequentist hypothesis testing

- Weight of Dutch men compared to Belgian men (populations)
- Observed sample averages $\overline{W}_{\mathrm{NL}}, \overline{W}_{\mathrm{BE}}$
- ***Null hypothesis*** H_0: $\mu_{\mathrm{w,NL}} = \mu_{\mathrm{w,BE}}$
- Test statistic:

$$\frac{\overline{W}_{NL} - \overline{W}_{BE}}{\sigma_{\overline{W}_{NL} - \overline{W}_{BE}}} \sim \mathcal{N}(0,1) \qquad \text{(under } H_0\text{)}$$



Overview

1. Classical probability and statistics

2. Principles of Bayesian probability theory

3. Applications
    - Classification
    - Regression analysis
    - Some other applications

4. Conclusions and references

Probability theory: quantifying uncertainty

- Every piece of information has uncertainty
- Uncertainty = lack of information
- Observation may reduce uncertainty
- Probability (distribution) *quantifies* uncertainty

Example: physical sciences

- Measurement of physical quantity
- Origin of stochasticity:
  - Apparatus
  - Microscopic fluctuations
- Systematic uncertainty is assigned a probability distribution
- E.g. coin tossing, voltage measurement, probability of hypothesis vs. another, ...
- Bayesian: no 'random' variables

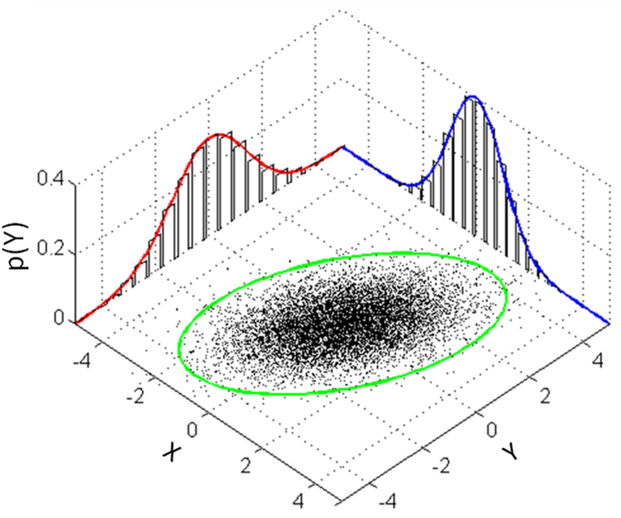
What is probability?

- Objective Bayesian view
- Probability = real number $\in [0, 1]$
- Always conditioned on known information
- Notation:

$$p(A|B) \quad \text{or} \quad p(A|I)$$

- Extension of logic: measure of degree to which *B implies A*
- Degree of plausibility, but subject to consistency
- Same information $\Rightarrow$ same probabilities
- ***Probability distribution:*** outcome $\rightarrow$ probability

Joint, marginal and conditional distributions



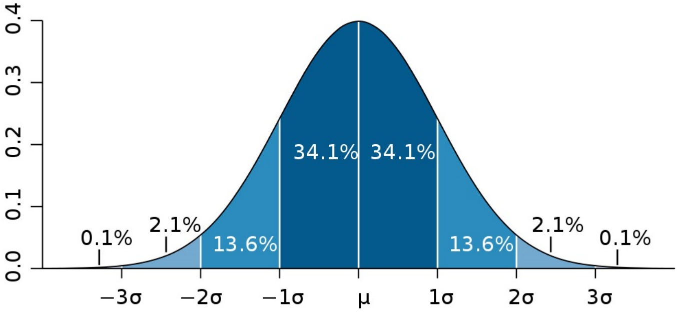
$$p(x,y), p(x), p(y), p(x|y), p(y|x)$$

Example: normal distribution

- Normal/Gaussian ***probability density function*** (PDF):

$$p(x|\mu,\sigma,I) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- Probability $x_1 \leq x < x_1 + \mathrm{d}x$
- Inverse problem: μ, σ given x?



Updating information states

Bayes' theorem

$$p(\boldsymbol{\theta}|\boldsymbol{x}, I) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)}{p(\boldsymbol{x}|I)}$$

$\boldsymbol{x}$ = data vector

$\boldsymbol{\theta}$ = vector of model parameters

I = implicit knowledge

- ***Likelihood***: misfit between model and data
- ***Prior*** distribution: 'expert' or diffuse knowledge
- ***Evidence***:

$$p(\boldsymbol{x}|I) = \int p(\boldsymbol{x}, \boldsymbol{\theta}|I)\,\mathrm{d}\boldsymbol{\theta} = \int p(\boldsymbol{x}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)\,\mathrm{d}\boldsymbol{\theta}$$

- ***Posterior*** distribution

Updating information states

Bayes' theorem

$$p(\boldsymbol{\theta}|\boldsymbol{x}, I) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)}{p(\boldsymbol{x}|I)}$$

$\boldsymbol{x}$ = data vector

$\boldsymbol{\theta}$ = vector of model parameters

I = implicit knowledge

- ***Likelihood***: misfit between model and data
- ***Prior*** distribution: 'expert' or diffuse knowledge
- ***Evidence***:

$$p(\boldsymbol{x}|I) = \int p(\boldsymbol{x}, \boldsymbol{\theta}|I)\, \mathrm{d}\boldsymbol{\theta} = \int p(\boldsymbol{x}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)\, \mathrm{d}\boldsymbol{\theta}$$

- ***Posterior*** distribution

Updating information states

Bayes' theorem

$$p(\boldsymbol{\theta}|\boldsymbol{x}, I) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)}{p(\boldsymbol{x}|I)}$$

$\boldsymbol{x}$ = data vector

$\boldsymbol{\theta}$ = vector of model parameters

I = implicit knowledge

- ***Likelihood***: misfit between model and data
- ***Prior*** distribution: 'expert' or diffuse knowledge
- ***Evidence***:

$$p(\boldsymbol{x}|I) = \int p(\boldsymbol{x}, \boldsymbol{\theta}|I)\, \mathrm{d}\boldsymbol{\theta} = \int p(\boldsymbol{x}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)\, \mathrm{d}\boldsymbol{\theta}$$

- ***Posterior*** distribution

Updating information states

Bayes' theorem

$$p(\boldsymbol{\theta}|\boldsymbol{x}, I) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)}{p(\boldsymbol{x}|I)}$$

$\boldsymbol{x}$ = data vector

$\boldsymbol{\theta}$ = vector of model parameters

I = implicit knowledge

- ***Likelihood***: misfit between model and data
- ***Prior*** distribution: 'expert' or diffuse knowledge
- ***Evidence***:

$$p(\boldsymbol{x}|I) = \int p(\boldsymbol{x}, \boldsymbol{\theta}|I)\, \mathrm{d}\boldsymbol{\theta} = \int p(\boldsymbol{x}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)\, \mathrm{d}\boldsymbol{\theta}$$

- ***Posterior*** distribution

Mean of a normal distribution: uniform prior

- n measurements $x_i \to \boldsymbol{x}$
- Independent and identically distributed x_i:

$$p(\boldsymbol{x}|\mu,\sigma,I) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

- Bayes' rule:

$$p(\mu,\sigma|\boldsymbol{x},I) \propto p(\boldsymbol{x}|\mu,\sigma,I)p(\mu,\sigma|I)$$

- Suppose $\sigma \equiv \sigma_\mathrm{e} \to$ delta function
- Assume $\mu \in [\mu_\mathrm{min}, \mu_\mathrm{max}] \to$ uniform prior:

$$p(\mu|I) = \begin{cases} \dfrac{1}{\mu_\mathrm{max}-\mu_\mathrm{min}}, & \text{if } \mu \in [\mu_\mathrm{min},\mu_\mathrm{max}] \\ 0, & \text{otherwise} \end{cases}$$

- Let $\mu_\mathrm{min} \to -\infty$, $\mu_\mathrm{max} \to +\infty \Rightarrow$ ***improper prior***
- Ensure proper posterior

Posterior for μ

- Posterior:

$$p(\mu|\boldsymbol{x}, I) \propto \exp\left[-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma_\mathrm{e}^2}\right]$$

- Define

$$\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{(\Delta x)^2} \equiv \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Adding and subtracting $2n\bar{x}^2$ ('completing the square'),

$$p(\mu|\boldsymbol{x}, I) \propto \exp\left\{-\frac{1}{2\sigma_\mathrm{e}^2/n}\left[(\mu - \bar{x})^2 + \overline{(\Delta x)^2}\right]\right\}$$

- Retaining dependence on μ,

$$p(\mu|\boldsymbol{x}, I) \propto \exp\left[-\frac{(\mu - \bar{x})^2}{2\sigma_\mathrm{e}^2/n}\right]$$

- $\mu \sim \mathcal{N}(\bar{x}, \sigma_\mathrm{e}^2/n)$

Mean of a normal distribution: normal prior

- Normal prior: $\mu \sim \mathcal{N}(\mu_0, \tau^2)$
- Posterior:

$$p(\mu|\boldsymbol{x}, I) \propto \exp\left[-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma_{\mathrm{e}}^2}\right] \times \exp\left[-\frac{(\mu - \mu_0)^2}{2\tau^2}\right]$$

- Expanding and completing the square,

$$\mu \sim \mathcal{N}(\mu_n, \sigma_n^2),$$

where

$$\mu_n \equiv \sigma_n^2 \left(\frac{n}{\sigma_{\mathrm{e}}^2}\bar{x} + \frac{1}{\tau^2}\mu_0\right) \quad \text{and} \quad \sigma_n^2 \equiv \left(\frac{n}{\sigma_{\mathrm{e}}^2} + \frac{1}{\tau^2}\right)^{-1}$$

- μ_n is weighted average of μ_0 and $\bar{x}$

Unknown mean and standard deviation

- Repeated measurements $\rightarrow$ information on σ
- Scale variable $\sigma \rightarrow$ ***Jeffreys' scale prior***:

$$p(\sigma|I) \propto \frac{1}{\sigma}, \qquad \sigma \in]0, +\infty[$$

- Posterior:

$$p(\mu, \sigma|\boldsymbol{x}, I) \propto \frac{1}{\sigma^n} \exp\left[-\frac{(\mu - \overline{x})^2 + \overline{(\Delta x)^2}}{2\sigma^2/n}\right] \times \frac{1}{\sigma}$$

Marginal posterior for μ (1)

- ***Marginalization*** = integrating out a (nuisance) parameter:

$$p(\mu|\boldsymbol{x},I) = \int_0^{+\infty} p(\mu,\sigma|\boldsymbol{x},I)\,\mathrm{d}\sigma$$
$$\propto \int_0^{+\infty} \frac{1}{2}\left[\frac{(\mu-\overline{x})^2+\overline{(\Delta x)^2}}{2/n}\right]^{-\frac{n}{2}} s^{\frac{n}{2}-1}\mathrm{e}^{-s}\,\mathrm{d}s$$
$$= \frac{1}{2}\Gamma\left(\frac{n}{2}\right)\left[\frac{(\mu-\overline{x})^2+\overline{(\Delta x)^2}}{2/n}\right]^{-\frac{n}{2}},$$

where

$$s \equiv \frac{(\mu-\overline{x})^2+\overline{(\Delta x)^2}}{2\sigma^2/n}$$

Marginal posterior for μ (2)

- After normalization:

$$p(\mu|\boldsymbol{x}, I) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi\overline{(\Delta x)^2}}\,\Gamma\left(\frac{n-1}{2}\right)} \left[1 + \frac{(\mu - \overline{x})^2}{\overline{(\Delta x)^2}}\right]^{-\frac{n}{2}}$$

- Changing variables,

$$t \equiv \frac{(\mu - \overline{x})^2}{\sqrt{\overline{(\Delta x)^2}/(n-1)}}, \quad \text{with} \quad p(t|\boldsymbol{x}, I)\mathrm{d}t \equiv p(\mu|\boldsymbol{x}, I)\mathrm{d}\mu,$$

$$p(t|\boldsymbol{x}, I) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\Gamma\left(\frac{n-1}{2}\right)} \left[1 + \frac{t^2}{n-1}\right]^{-\frac{n}{2}}$$

- Student's t-distribution with parameter $\nu = n - 1$
- If $n \gg 1$,

$$p(\mu|\boldsymbol{x}, I) \longrightarrow \frac{1}{\sqrt{2\pi\overline{(\Delta x)^2}/n}} \exp\left[-\frac{(\mu - \overline{x})^2}{2\overline{(\Delta x)^2}/n}\right]$$

Marginal posterior for σ

- Marginalization of μ:

$$p(\sigma|\boldsymbol{x}, I) = \int_{-\infty}^{+\infty} p(\mu, \sigma|\boldsymbol{x}, I)\,\mathrm{d}\mu$$
$$\propto \frac{1}{\sigma^n} \exp\left[-\frac{\overline{(\Delta x)^2}}{2\sigma^2/n}\right]$$

- Setting $X \equiv n\overline{(\Delta x)^2}/\sigma^2$,

$$p(X|\boldsymbol{x}, I) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} X^{\frac{k}{2}-1} \mathrm{e}^{-\frac{X}{2}}, \qquad k \equiv n-1$$

- χ^2 distribution with parameter k

The Laplace approximation (1)

- Laplace (saddle point) approximation of distributions around the mode (= maximum)
- E.g. marginal for μ:

$$p(\mu|\boldsymbol{x}, I) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi\overline{(\Delta x)^2}}\Gamma\left(\frac{n-1}{2}\right)}\left[1 + \frac{(\mu - \overline{x})^2}{\overline{(\Delta x)^2}}\right]^{-\frac{n}{2}}$$

- Taylor expansion around mode:

$$\begin{aligned}\ln[p(\mu|\boldsymbol{x}, I)] &\approx \ln[p(\overline{x}|\boldsymbol{x}, I)] + \frac{1}{2}\left.\frac{\mathrm{d}^2(\ln p)}{\mathrm{d}\mu^2}\right|_{\mu=\overline{x}}(\mu - \overline{x})^2 \\ &= \ln\left[\Gamma\left(\frac{n}{2}\right)\right] - \ln\left[\Gamma\left(\frac{n-1}{2}\right)\right] \\ &\quad - \frac{1}{2}\ln\left[\pi\overline{(\Delta x)^2}\right] - \frac{n}{2\overline{(\Delta x)^2}}(\mu - \overline{x})^2\end{aligned}$$

The Laplace approximation (2)

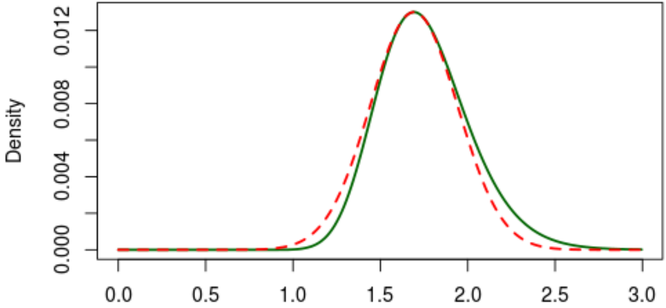
- On the original scale:

$$p(\mu|\boldsymbol{x}, I) \approx \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \frac{1}{\sqrt{\pi \overline{(\Delta x)^2}}} \exp\left[-\frac{(\mu - \bar{x})^2}{2\overline{(\Delta x)^2}/n}\right]$$

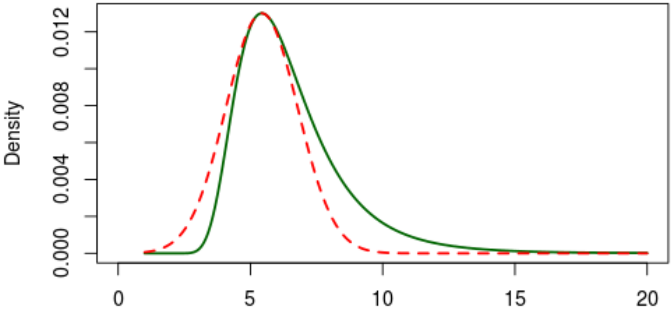
- Standard deviation $\sigma_{\mathrm{L}} \to$ curvature of $\ln p$:

$$\sigma_{\mathrm{L}} = \left[-\left.\frac{\mathrm{d}^2(\ln p)}{\mathrm{d}\mu^2}\right|_{\mu=\bar{x}}\right]^{-1/2}$$

Laplace approximation: example 1



Laplace approximation: example 2



Multivariate Laplace approximation

- For $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^{\mathrm{t}}$,

$$p(\boldsymbol{\theta}|\boldsymbol{\theta}_0, I) \propto \exp\left[\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{t}}[\boldsymbol{\nabla}\boldsymbol{\nabla}(\ln p)]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right]$$

- $\boldsymbol{\nabla}\boldsymbol{\nabla}(\ln p)$: Hessian matrix, where

$$\Sigma_{\mathrm{L}} = -\left\{[\boldsymbol{\nabla}\boldsymbol{\nabla}(\ln p)]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right\}^{-1}$$

Model comparison (hypothesis testing)

- Let $\{H_i\}$ be complete set of hypotheses
- Data D to support or reject hypotheses
- Bayes' rule:

$$p(H_i|D,I) = \frac{p(D|H_i,I)p(H_i|I)}{p(D|I)}, \quad p(D|I) = \sum_i p(D|H_i,I)p(H_i|I)$$

- Assume single hypothesis H and complement $\overline{H}$
- ***Odds ratio*** o:

$$o \equiv \frac{p(H|D,I)}{p(\overline{H}|D,I)} = \underbrace{\frac{p(D|H,I)}{p(D|\overline{H},I)}}_{\text{Bayes factor}} \underbrace{\frac{p(H|I)}{p(\overline{H}|I)}}_{\text{Prior odds}}$$

- $p(D|H,I) =$ ***model evidence***

Testing a Gaussian mean (1)

- E.g. n measurements x_i of a quantity x
- Assume normal distribution with known variance σ^2
- Question: are the data compatible with mean $\mu = \mu_0$?
  - Yes: H
  - No: $\overline{H}$
- Under H:

$$p(\bar{x}|H, I) = C \exp\left[-\frac{1}{2\sigma^2/n}(\bar{x} - \mu_0)^2\right]$$

- Under $\overline{H}$:

$$p(\bar{x}|\overline{H}, I) = \int p(\bar{x}|\mu, \sigma, I) p(\mu|\overline{H}, I) \, \mathrm{d}\mu \tag{1}$$

Testing a Gaussian mean (2)

- Assume bounds $\mu_{\min}$ and $\mu_{\max}$:

$$p(\mu|\overline{H}, I) = \frac{1}{|\mu_{\max} - \mu_{\min}|} 1(\mu_{\min} \leq \mu \leq \mu_{\max})$$

- Then (1) becomes

$$p(\overline{x}|\overline{H}, I) = \frac{C}{|\mu_{\max} - \mu_{\min}|} \int_{\mu_{\min}}^{\mu_{\max}} \exp\left[-\frac{1}{2\sigma^2/n}(\overline{x} - \mu)^2\right] \mathrm{d}\mu$$

- Assume wide prior interval:

$$|\overline{x} - \mu_{\min}|, |\overline{x} - \mu_{\max}| \gg \text{SE},$$

$$\text{SE} = \text{standard error} \equiv \frac{\sigma}{\sqrt{n}}$$

Testing a Gaussian mean (3)

- Then

$$p(\overline{x}|\overline{H}, I) \approx \frac{C}{|\mu_{\text{max}} - \mu_{\text{min}}|} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2\sigma^2/n}(\overline{x} - \mu)^2\right] \, \mathrm{d}\mu$$

$$= \frac{C \ \text{SE} \ \sqrt{2\pi}}{|\mu_{\text{max}} - \mu_{\text{min}}|}$$

- Bayes factor BF:

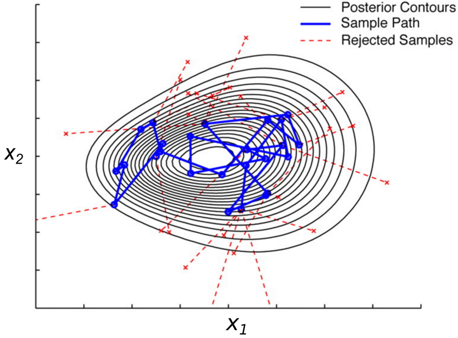
$$\text{BF} = \frac{p(\overline{x}|H, I)}{p(\overline{x}|\overline{H}, I)} = \frac{|\mu_{\text{max}} - \mu_{\text{min}}|}{\text{SE}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z^2},$$

$$z \equiv \frac{|\overline{x} - \mu_0|}{\text{SE}}$$

- Cf. frequentist hypothesis test

Sampling from distributions

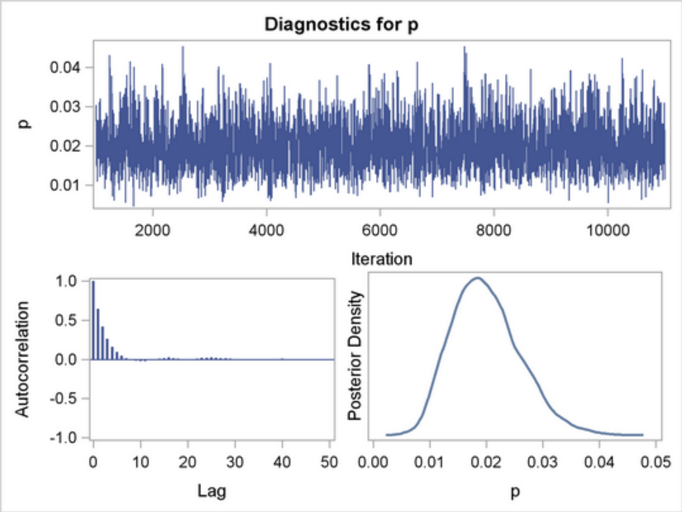
- Need samples from ***target distribution*** $p(\boldsymbol{\theta}|I)$
- ***Markov chain Monte Carlo*** sampling
- Sample from ***proposal distribution***



- Calculate Monte Carlo averages, e.g.

$$\overline{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} \theta_j^{(t_c+i)}, \qquad \overline{(\Delta\theta_j)^2} = \frac{1}{n}\sum_{i=1}^{n}\left(\theta_j^{(t_c+i)} - \overline{\theta}_j\right)^2$$

MCMC sampling



Overview

1. Classical probability and statistics

2. Principles of Bayesian probability theory

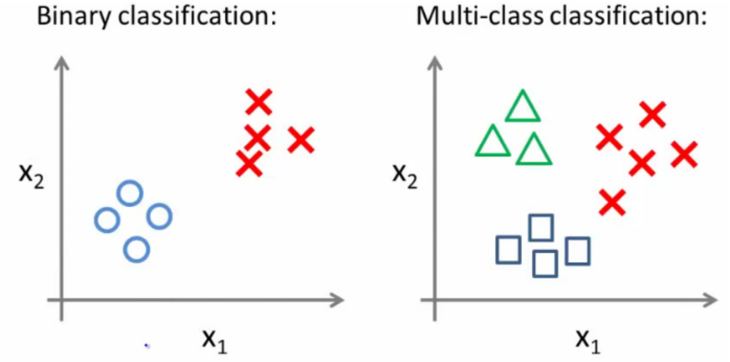
3. Applications
    - Classification
    - Regression analysis
    - Some other applications

4. Conclusions and references

Overview

1. Classical probability and statistics
2. Principles of Bayesian probability theory
3. Applications
   - Classification
   - Regression analysis
   - Some other applications
4. Conclusions and references

Classification or clustering



Simple Bayesian classification

- M clusters of in total n data points $\boldsymbol{x}_i$ in P-dimensional space
- Known class labels ω_j ($j = 1, \ldots, M$) of $\boldsymbol{x}_i$
- Bayes' rule for new point $\boldsymbol{x}$:

$$p(\omega_j | \boldsymbol{x}, I) = \frac{p(\boldsymbol{x} | \omega_j, I) p(\omega_j | I)}{p(\boldsymbol{x} | I)}$$

- ***Maximum a posteriori*** (MAP) classification rule for $\boldsymbol{x}$:

$$\text{Assign } \boldsymbol{x} \text{ to } \omega_i = \arg\max_{\omega_j} p(\omega_j | \boldsymbol{x}, I) = \arg\max_{\omega_j} p(\boldsymbol{x} | \omega_j, I) p(\omega_j | I)$$

Examples of priors and likelihoods

- Examples of prior probabilities (indifference):
  - $p(\omega_i|I) = p(\omega_j|I), \forall i, j$
  - Count class membership:

$$p(\omega_i|I) \equiv \frac{n_i}{n}, \qquad i = 1, \ldots, M$$

- Examples of likelihoods:
  - *Naive Bayesian classifier:*

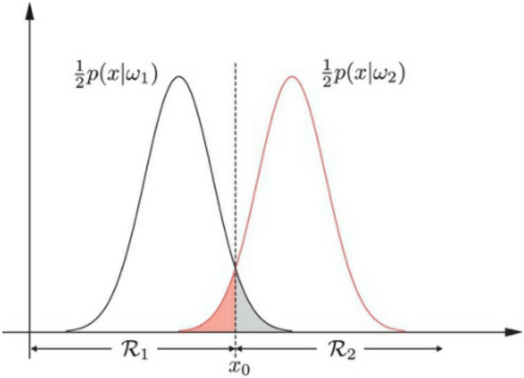
$$p(\boldsymbol{x}|\omega_i) = \prod_{k=1}^{P} p(x_k|\omega_i), \qquad i = 1, \ldots, M$$

  - Multivariate Gaussian:

$$p(\boldsymbol{x}|\omega_j, I) = \frac{1}{(2\pi)^{p/2}|\Sigma_j|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^{\mathrm{t}} \Sigma_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j)\right]$$

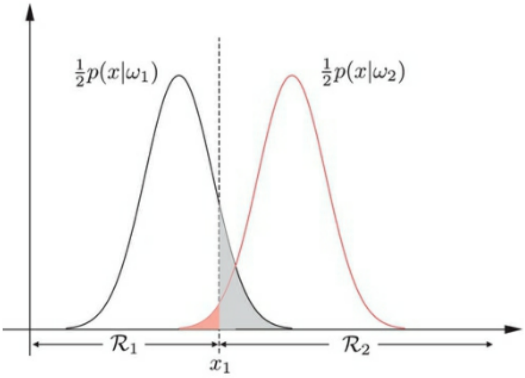
Optimality of Bayesian classifier

Bayesian classifier minimizes probability of misclassification



Optimality of Bayesian classifier

Bayesian classifier minimizes probability of misclassification



MAP classification: decision surfaces

- MAP: maximize w.r.t. ω_j:

$$\ln p(\omega_j|\boldsymbol{x}, I) = \ln p(\boldsymbol{x}|\omega_j, I) + \ln p(\omega_j|I)$$

- Define ($M = 2$):

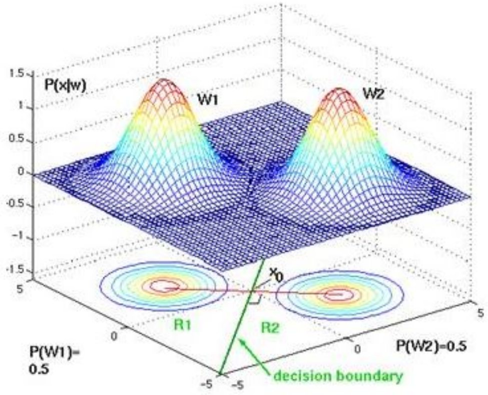
$$g(\boldsymbol{x}) \equiv \ln p(\omega_1|\boldsymbol{x}, I) - \ln p(\omega_2|\boldsymbol{x}, I)$$

- For normal likelihood:

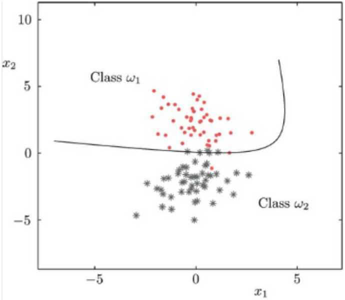
$$g(\boldsymbol{x}) = \overbrace{\frac{1}{2}\left(\boldsymbol{x}^{\mathrm{t}}\Sigma_2^{-1}\boldsymbol{x} - \boldsymbol{x}^{\mathrm{t}}\Sigma_1^{-1}\boldsymbol{x}\right)}^{\text{Quadratic}} + \overbrace{\boldsymbol{\mu}_1^{\mathrm{t}}\Sigma_1^{-1}\boldsymbol{x} - \boldsymbol{\mu}_2^{\mathrm{t}}\Sigma_2^{-1}\boldsymbol{x}}^{\text{Linear}}$$
$$\underbrace{-\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{t}}\Sigma_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{t}}\Sigma_2^{-1}\boldsymbol{\mu}_2 + \frac{1}{2}\ln\frac{|\Sigma_2|}{|\Sigma_1|} + \ln\frac{p(\omega_1|I)}{p(\omega_2|I)}}_{\text{Constant}}$$

- $g(\boldsymbol{x})$ separates classes: ***decision hypersurface***

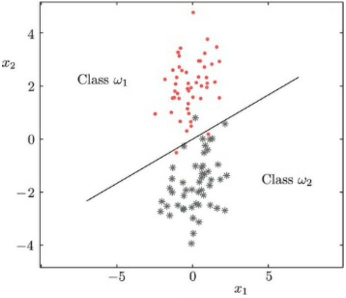
Discriminant analysis



QDA and LDA

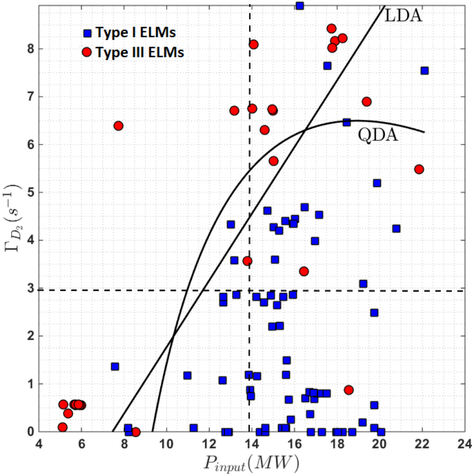


Quadratic discriminant analysis (QDA)



Linear discriminant analysis (LDA): $\Sigma_1 = \Sigma_2$

ELM type classification



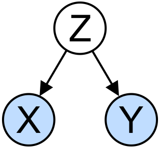
$$P_{\text{input}} - 1.41\Gamma_{D_2} = 7.47$$

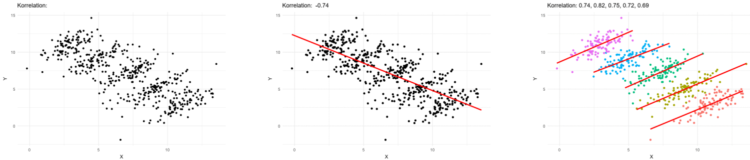
A. Shabbir *et al.*, Fusion Eng. Des. **123**, 717, 2017

Overview

Uncertainties in regression analysis

- Measurement uncertainty
- Percentage errors from database
- Model uncertainty:
  - Power law
  - Missing variables
  - Confounding variables
- Predictor correlations (e.g. $I_{\mathrm{p}} \propto B_{\mathrm{t}}$)
- Heterogeneity: multi-machine database
- Simpson's paradox:





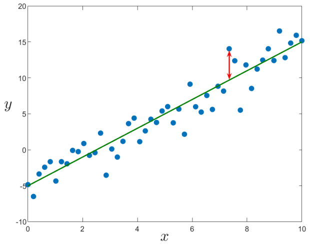
Multilinear regression and simple least squares

- Regression model (Gauss-Markov):

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p + \epsilon \qquad \text{Often loglinear!}$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2),\ \sigma \text{ known}$$

- Take n measurements:

$$\boldsymbol{y} \equiv [y_1, \ldots, y_n]^{\mathrm{t}},$$
$$X \equiv \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$



- Ordinary least squares (OLS):

$$\boldsymbol{\alpha}_{\mathrm{OLS}} = \arg\min_{\boldsymbol{\alpha}} \left[(\boldsymbol{y} - X\boldsymbol{\alpha})^{\mathrm{t}} (\boldsymbol{y} - X\boldsymbol{\alpha}) \right] = \underbrace{(X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}}_{\text{Moore-Penrose pseudoinverse}} \boldsymbol{y}$$

Maximum likelihood solution

- Likelihood:

$$p(y|\boldsymbol{x},\boldsymbol{\alpha},\sigma,I) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}\left(y-\alpha_0-\sum_{j=1}^{p}\alpha_j x_j\right)^2\right],$$

$$\boldsymbol{\alpha} \equiv [\alpha_0, \boldsymbol{\alpha}_p^{\mathrm{t}}]^{\mathrm{t}}, \quad \boldsymbol{\alpha}_p \equiv [\alpha_1, \ldots, \alpha_p]^{\mathrm{t}}$$

- Conditional independence:

$$p(\boldsymbol{y}|X,\boldsymbol{\alpha},\sigma,I) = (2\pi)^{-n/2}\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-X\boldsymbol{\alpha})^{\mathrm{t}}(\boldsymbol{y}-X\boldsymbol{\alpha})\right]$$

- ML solution:

$$0 = \boldsymbol{\nabla}_{\boldsymbol{\alpha}}(\boldsymbol{y}-X\boldsymbol{\alpha})^{\mathrm{t}}(\boldsymbol{y}-X\boldsymbol{\alpha}) = -2X^{\mathrm{t}}\boldsymbol{y} + 2X^{\mathrm{t}}X\boldsymbol{\alpha}$$

$$\Rightarrow \quad \boldsymbol{\alpha}_{\mathrm{ML}} = (X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}\boldsymbol{y} = \boldsymbol{\alpha}_{\mathrm{OLS}}$$

MAP solution and posterior

- Uniform priors on α_j (not the most uninformative!):

$$p(\boldsymbol{\alpha}|\boldsymbol{y}, X, \sigma, I) \propto \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\alpha})^{\mathrm{t}}(\boldsymbol{y} - X\boldsymbol{\alpha})\right]$$

- Due to linearity and Gaussianity: $\boldsymbol{\alpha}_{\mathrm{MAP}} = \boldsymbol{\alpha}_{\mathrm{ML}} = \boldsymbol{\alpha}_{\mathrm{LS}}$
- Taylor expansion (exact!):

$$\begin{aligned}(\boldsymbol{y} - X\boldsymbol{\alpha})^{\mathrm{t}}(\boldsymbol{y} - X\boldsymbol{\alpha}) = (\boldsymbol{y} &- X\boldsymbol{\alpha}_{\mathrm{MAP}})^{\mathrm{t}}(\boldsymbol{y} - X\boldsymbol{\alpha}_{\mathrm{MAP}}) \\ &+ \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\mathrm{MAP}})^{\mathrm{t}}\, 2X^{\mathrm{t}}X\,(\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\mathrm{MAP}})\end{aligned}$$

- Posterior distribution:

$$\begin{aligned}p(\boldsymbol{\alpha}|\boldsymbol{y}, X, \sigma, I) = (2\pi)^{-n/2}|\Sigma|^{-1/2} & \\ \times \exp\left[-\frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\mathrm{MAP}})^{\mathrm{t}}\Sigma^{-1}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\mathrm{MAP}})\right], & \\ \Sigma \equiv \sigma^2(X^{\mathrm{t}}X)^{-1} &\end{aligned}$$

Posterior predictive distribution (1)

- New predictions by the model?

- ***Posterior predictive distribution***:

$$p(y_{\text{new}}|\boldsymbol{x}_{\text{new}}, \boldsymbol{y}, X, \sigma, I) = \int_{\mathbb{R}^{p+1}} p(y_{\text{new}}, \boldsymbol{\beta}|\boldsymbol{x}_{\text{new}}, \boldsymbol{y}, X, \sigma, I)\,\mathrm{d}\boldsymbol{\beta}$$
$$= \int_{\mathbb{R}^{p+1}} p(y_{\text{new}}|\boldsymbol{x}_{\text{new}}, \boldsymbol{\beta}, I)\, p(\boldsymbol{\beta}|\boldsymbol{y}, X, \sigma, I)\,\mathrm{d}\boldsymbol{\beta}$$

- But

$$p(y_{\text{new}}|\boldsymbol{x}_{\text{new}}, \boldsymbol{\beta}, I) = \delta(y_{\text{new}} - \boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{x}_{\text{new}})$$

- Fix $\beta_0 = y_{\text{new}} - \beta_1 x_{\text{new},1} - \ldots - \beta_p x_{\text{new},p}$

Posterior predictive distribution (2)

- Marginalize over $\boldsymbol{\beta}_p$ with flat priors:

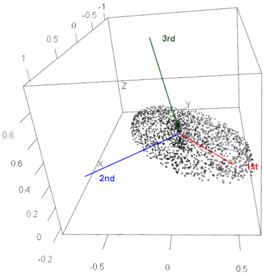
$$p(y_{\text{new}}|\boldsymbol{x}_{\text{new}}, \boldsymbol{y}, X, \sigma, I)$$
$$\propto \sigma^{-n} \int_{\mathbb{R}^p} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[y_i - y_{\text{new}} + \sum_{j=1}^{p} \beta_j (x_{\text{new},j} - x_{ij}) \right]^2 \right\} \mathrm{d}\boldsymbol{\beta}_p$$

- After (quite some) algebra, one finds simply

$$y_{\text{new,MAP}} = \sum_{j=1}^{p} x_{\text{new},j} \beta_{\text{MAP},j} + \beta_{\text{MAP},0}$$

- Simpler derivation based on properties of $\mathbb{E}$ and Var
- General posterior more complicated!

Multicollinearity



Detection

- Correlation matrices
- Belsley collinearity diagnostics

Remediation

- Eliminate predictor variables
- Principal component regression
- Regularization: ridge, lasso, elastic net, . . .

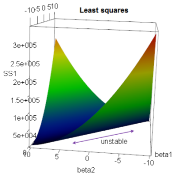
Ridge regression and lasso

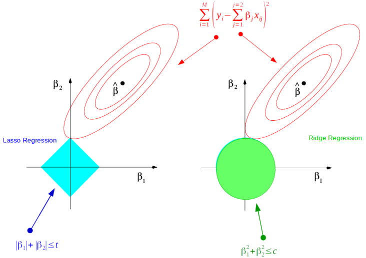
- Ridge regression (Tikhonov regularization) or zero-mean normal prior:

$$\boldsymbol{\alpha}_{\text{ridge}} = \arg\min_{\boldsymbol{\alpha}} \left[(\boldsymbol{y} - X\boldsymbol{\alpha})^{t}(\boldsymbol{y} - X\boldsymbol{\alpha}) + \lambda \sum_{j=0}^{p} \alpha_j^2 \right]$$

- Lasso or zero-mean Laplace prior:

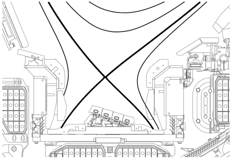
$$\boldsymbol{\alpha}_{\text{lasso}} = \arg\min_{\boldsymbol{\alpha}} \left[(\boldsymbol{y} - X\boldsymbol{\alpha})^{t}(\boldsymbol{y} - X\boldsymbol{\alpha}) + \lambda \sum_{j=0}^{p} |\alpha_j| \right]$$



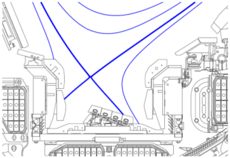


Divertor configurations at JET

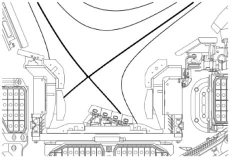
CC (corner)



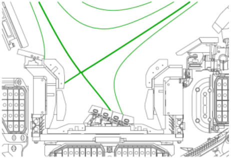
HT3R



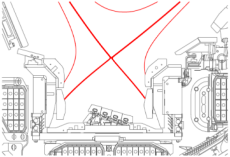
V5



V5L



VT



Categorical variables

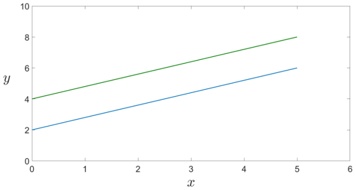
- Loglinear model:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & x_{B_t,1} & x_{n_e,1} \\ 1 & 0 & 0 & 0 & 0 & x_{B_t,2} & x_{n_e,2} \\ 0 & 1 & 0 & 0 & 0 & x_{B_t,3} & x_{n_e,3} \\ 0 & 1 & 0 & 0 & 0 & x_{B_t,4} & x_{n_e,4} \\ 0 & 1 & 0 & 0 & 0 & x_{B_t,5} & x_{n_e,5} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & x_{B_t,n-1} & x_{n_e,n-1} \\ 0 & 0 & 0 & 0 & 1 & x_{B_t,n} & x_{n_e,n} \end{bmatrix} \begin{bmatrix} \alpha_{0,\mathrm{CC}} \\ \alpha_{0,\mathrm{HT3R}} \\ \alpha_{0,\mathrm{V5}} \\ \alpha_{0,\mathrm{V5L}} \\ \alpha_{0,\mathrm{VT}} \\ \alpha_B \\ \alpha_n \end{bmatrix} \begin{matrix} \left.\right\} \text{Intercepts} \\ \left.\right\} \text{Slopes} \end{matrix}$$

Columns: $x_{0,\mathrm{CC}}$ $x_{0,\mathrm{HT3R}}$ $x_{0,\mathrm{V5}}$ $x_{0,\mathrm{V5L}}$ $x_{0,\mathrm{VT}}$ x_{B_t} x_{n_e}

- Statistical model:

$$\boldsymbol{y} = X\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$$



Global confinement scaling

- Power law + log-transform:

$$\tau_{\mathrm{E,th}} = \mathrm{e}^{\alpha_0} I_{\mathrm{p}}^{\alpha_I} B_{\mathrm{t}}^{\alpha_B} \bar{n}_{\mathrm{e}}^{\alpha_n} P_{\mathrm{l,th}}^{\alpha_P} R_{\mathrm{geo}}^{\alpha_R} (1+\delta)^{\alpha_\delta} \kappa_{\mathrm{a}}^{\alpha_\kappa} \epsilon^{\alpha_\epsilon} M_{\mathrm{eff}}^{\alpha_M}$$

$$\eta = \ln \tau_{\mathrm{E,th}}, \quad \xi_1 = \ln I_{\mathrm{p}}, \quad \ldots, \quad \xi_p = \ln M_{\mathrm{eff}}$$

- Errors in all variables:

$$\eta = \alpha_0 + \sum_{j=1}^{p} \alpha_j \xi_j$$

$$y = \eta + \epsilon_y, \quad x_1 = \xi_1 + \epsilon_{x_1}, \quad \ldots, \quad x_p = \xi_p + \epsilon_{x_p}$$

$$\epsilon_y \sim \mathcal{N}\left(0, \sigma_y^2\right), \quad \epsilon_{x_1} \sim \mathcal{N}\left(0, \sigma_{x_1}^2\right), \quad \ldots, \quad \epsilon_{x_p} \sim \mathcal{N}\left(0, \sigma_{x_p}^2\right)$$

$$\sigma_{\mathrm{mod}}^2 = \sigma_y^2 + \sum_{j=1}^{p} \alpha_j^2 \sigma_{x_j}^2$$

Robust Bayesian regression

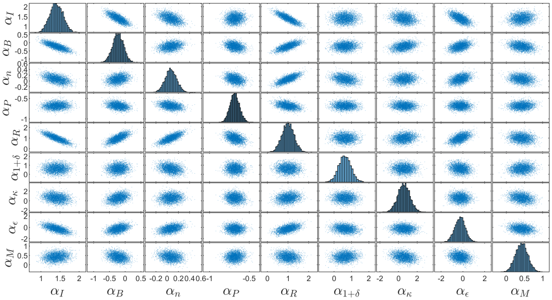
- Robust Bayesian regression ***RBAYES***

$$p(\{y_{i_k,k}\},\{x_{i_k,j,k}\}|\{\alpha_0,\alpha_j\},\{\gamma_k\})$$

$$= \prod_k \prod_{i_k} \frac{1}{\sqrt{2\pi\gamma_k^2\sigma^2_{\mathrm{mod},i_k,k}}} \exp\left[-\frac{1}{2}\frac{\left(y_{i_k,k}-\eta_{i_k,k}\right)^2}{\gamma_k{}^2\sigma^2_{\mathrm{mod},i_k,k}}\right]$$

1 for each device → γ_k

- Sensitivity analysis → practical error bars:



Geodesic least squares

- Geodesic least squares: ***GLS***

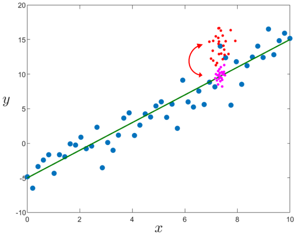
$$\prod_k \prod_{i_k} \frac{1}{\sqrt{2\pi\sigma^2_{\text{tot},i_k,k}}} \exp\left[-\frac{1}{2}\frac{\left(y_{i_k,k}-\eta_{i_k,k}\right)^2}{\sigma^2_{\text{mod},i_k,k}}\right]$$

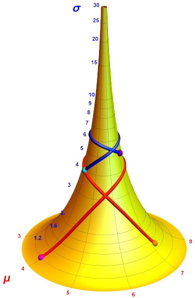
↕ Rao geodesic distance (GD)

$$\frac{1}{\sqrt{2\pi}\,\sigma_{\text{obs}}} \exp\left[-\frac{1}{2}\frac{(y-y_i)^2}{\sigma_{\text{obs}}{}^2}\right]$$

G. Verdoolaege *et al.*, Nucl. Fusion, **55**, 113019, 2015

G. Verdoolaege *et al.*, Entropy, **17**, 4602–4626, 2015





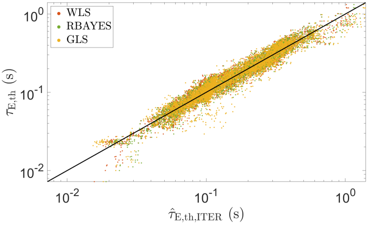
Multi-machine engineering scaling

STD5-IL ELMy H-mode (error bars from Bayesian analysis)

Engineering scaling *ITPA20-IL*

$$\tau_{E,th} = (0.067 \pm 0.060)\, I_p^{1.29\pm0.17}\, B_t^{-0.13\pm0.17}\, \bar{n}_e^{0.15\pm0.10}\, P_{l,th}^{-0.644\pm0.060}\, R_{geo}^{1.19\pm0.29} \times (1+\delta)^{0.56\pm0.35}\, \kappa_a^{0.67\pm0.65}\, M_{eff}^{0.30\pm0.17} \rightarrow \mathbf{H_{20}}$$

$$\hat{\tau}_{E,th,ITER} = 2.79 \pm 0.44\,\mathrm{s}$$



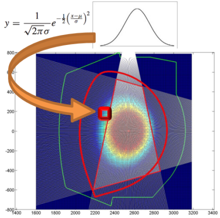
G. Verdoolaege *et al.*, Nucl. Fusion, **61**, 076006, 2021

G. Verdoolaege *et al.*, 27th IAEA Fusion Energy Conference, EX/P7-1, Gandhinagar, India, 2018

S. Kaye *et al.*, 60th Annual Meeting of the APS Division of Plasma Physics, TP11.00104, Portland, OR, USA, 2018

Overview

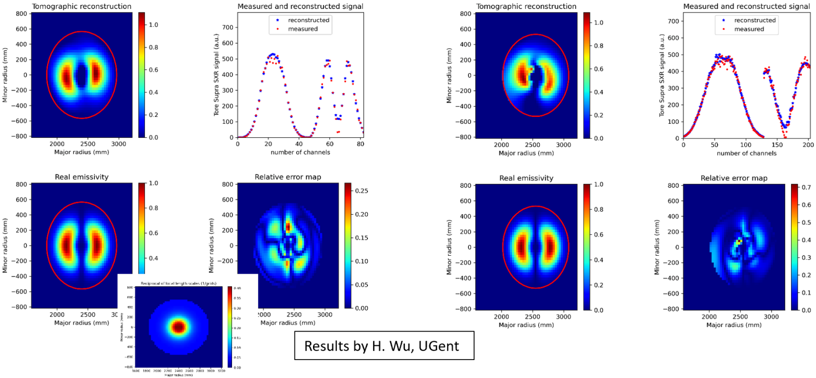
Gaussian process tomography



$$\Sigma = \begin{pmatrix} k(\vec{r}_1, \vec{r}_1) & \cdots & k(\vec{r}_1, \vec{r}_n) \\ \vdots & \ddots & \vdots \\ k(\vec{r}_n, \vec{r}_1) & \cdots & k(\vec{r}_n, \vec{r}_n) \end{pmatrix}, \quad k(\vec{r}_i, \vec{r}_j) = \sigma_f{}^2 \exp\left[-\left(\frac{\left\|\vec{r}_i - \vec{r}_j\right\|^2}{2\sigma_l{}^2}\right)\right]$$

J. Svensson, EFDA-ET-PR(11)24, 2011
D. Li et al., Rev. Sci. Instrum. **84**, 083506, 2013
T. Wang et al., Rev. Sci. Instrum. **89**, 063505, 2018

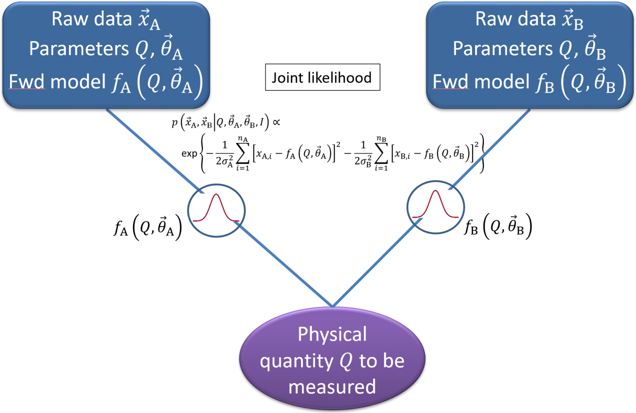
SXR tomography



Results by H. Wu, UGent

Information integration

- Data fusion / sensor fusion / integrated data analysis (IDA)



$$\underbrace{p(Q, \boldsymbol{\theta}_\mathrm{A}, \boldsymbol{\theta}_\mathrm{B} | \boldsymbol{x}_\mathrm{A}, \boldsymbol{x}_\mathrm{B}, I)}_{\text{Posterior}} \propto \underbrace{p(\boldsymbol{x}_\mathrm{A}, \boldsymbol{x}_\mathrm{B} | Q, \boldsymbol{\theta}_\mathrm{A}, \boldsymbol{\theta}_\mathrm{B}, I)}_{\text{likelihood}} \underbrace{p(Q, \boldsymbol{\theta}_\mathrm{A}, \boldsymbol{\theta}_\mathrm{B} | I)}_{\text{Prior}}$$

See IDA session on Friday!

Overview

1. Classical probability and statistics
2. Principles of Bayesian probability theory
3. Applications
   - Classification
   - Regression analysis
   - Some other applications
4. Conclusions and references

Conclusions

- Frequentist vs. Bayesian methods: interpretation of probability
- Bayesian probability: extension of logic to situations with uncertainty
- Posterior probability of parameters or hypotheses
- Numerical approach in general
- Underlies or explains many machine learning techniques

References

- D.S. Sivia and J. Skilling, *Data Analysis: A Bayesian Tutorial*, 2nd edition, Oxford University Press, 2006
- W. von der Linden, V. Dose and U. von Toussaint, *Bayesian Probability Theory: Applications in the Physical Sciences*, Cambridge University Press, 2014
- P. Gregory, *Data Analysis: A Bayesian Tutorial*, 2nd edition, Oxford University Press, 2006
- S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, Academic Press (Elsevier), 2015
- E.T. Jaynes (G.L. Bretthorst, ed.), *Probability Theory: The Logic of Science*, Cambridge University Press, 2003
- S.B. McGrayne, *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*, Yale University Press, 2011