SHORT TOUR THROUGH MATHEMATICAL STATISTICS

Paulo Serra

VU Colloquium (23rd September 2020)

VRIJE UNIVERSITEIT AMSTERDAM

Content

Short tour through Mathematical Statistics

A BIT ABOUT MYSELF

A SHORT CV

- PhD in Eindhoven (Mathematical Statistics);
- Postdoc in Göttingen (Splines);
- Postdoc at UvA (Stats for networks);
- Assistant professor in Eindhoven;
- Since this July, I’m at the VU.

SHORT TOUR THROUGH MATHEMATICAL STATISTICS

Non-parametric Mathematical Statistics

STATISTICS IN PICTURES

Population

Statistical/Probability Model

Generate data from the population

Statistics

Learn about population from data

Data

Distributions

F

Sampling → ← **Inference**

$X \sim F \in \mathcal{F}$

Samples

Data, statistics,…

NOW SOME MATHEMATICAL STATISTICS

My field of research is (Non-parametric) Mathematical Statistics.

Statistics: we just saw what that is;

Mathematical: we use tools from Probability, Functional Analysis, Graph Theory, Linear Algebra, etc.;

Non-parametric: we work with *large* models;

Clarification: what is a *large* model?

We call this a parametrised model.

$$\mathcal{F} = \{F_\theta \colon \theta \in \Theta\}$$

$\Theta \subseteq \mathbb{R}^p, p \in \mathbb{N}$ 👉 parametric

otherwise 👉 non-parametric

IN WHAT WAYS CAN A MODEL BE NON-PARAMETRIC?

$$\mathcal{F} = \{F_\theta \colon \theta \in \Theta\}$$

A larger model is more flexible; more likely to explain the data.

- $\theta \in \mathbb{R}^{p_n}$, where $p_n \to \infty$, as $n \to \infty$; (n is e.g., *sample size*)

IN WHAT WAYS CAN A MODEL BE NON-PARAMETRIC?

$$\mathcal{F}^{(n)} = \left\{F_\theta^{(n)} \colon \theta \in \Theta\right\}, \qquad \boldsymbol{X} \sim F_\theta^{(n)}$$

(e.g., $\boldsymbol{X} = (X_1, \dots, X_n)$.)

- $\theta \in \mathbb{R}^{p_n}$, where $p_n \to \infty$, as $n \to \infty$ (we think $p_n < n$; $p_n > n$ would be silly)
- $\theta \in \mathbb{R}^{p}$, where $p \gg n$ $(\Theta = \{\theta \colon \#\{i \colon \theta_i \neq 0\} \leq q\}, q \ll n)$
- θ is a sequence
- θ is a function on $\mathbb{R}$
- θ is a function on $\mathbb{R}^d, d \in \mathbb{N}$
- θ is a graph
- …

WHAT ABOUT INFERENCE GOALS?

$$\mathcal{F}^{(n)} = \left\{F_\theta^{(n)} \colon \theta \in \Theta\right\}, \qquad \boldsymbol{X} \sim F_\theta^{(n)}$$

Since the model is parametrised, the goal is to *say something* about θ.

Say something like what?

- Construct an estimator $\hat{\theta}$;
- Construct a confidence set $\widehat{\Theta}$;
- Pick between $\theta \epsilon \Theta_0$ versus $\theta \epsilon \Theta_1$ (Θ_0 and Θ_1 disjoint subsets of Θ).

WHAT ABOUT INFERENCE GOALS?

$$\mathcal{F}^{(n)} = \left\{F_\theta^{(n)} \colon \theta \in \Theta\right\}, \qquad \boldsymbol{X} \sim F_\theta^{(n)}$$

How do we do *point estimation* (construct an estimator), for instance?

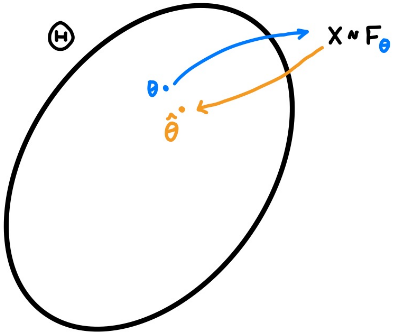
If the $F_\theta^{(n)}$ admit a density $f_\theta^{(n)}$ we can do *maximum likelihood estimation*.

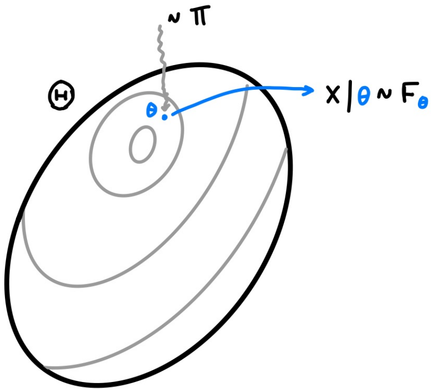
MLE: $f_\theta^{(n)}(\boldsymbol{x})$ tells us how likely sampling $\boldsymbol{x}$ is data comes from $F_\theta^{(n)}$ …
So $\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}} f_\theta^{(n)}(\boldsymbol{X})$ makes sense, and has all sorts of nice properties.

Only for parametric models though… for non-parametric models this estimator tends to be trivial.

WHAT ABOUT BAYES?

$$\mathcal{F}^{(n)} = \left\{F_\theta^{(n)} \colon \theta \in \Theta\right\}$$



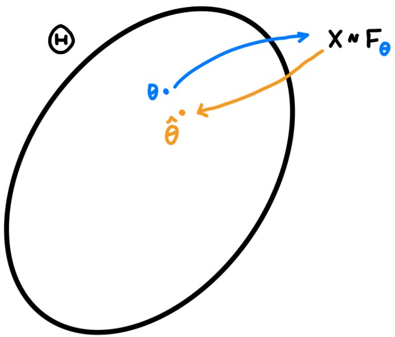


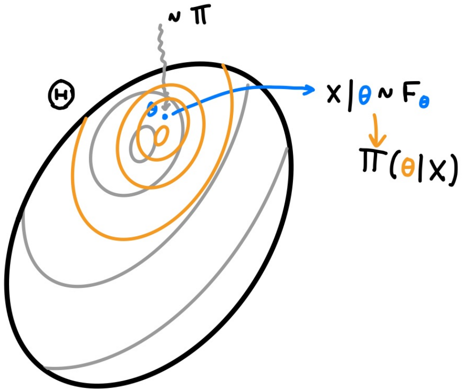
Frequentist: $\boldsymbol{X} \sim F_\theta^{(n)}$

Bayes: $\theta \sim \pi\ ,\ \ \boldsymbol{X}|\theta \sim F_\theta^{(n)}$

WHAT ABOUT BAYES?

$$\mathcal{F}^{(n)} = \left\{F_\theta^{(n)} : \theta \in \Theta\right\}$$





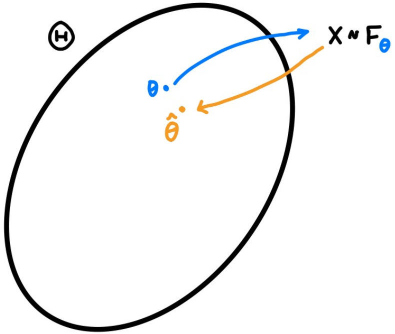
Frequentist: $\boldsymbol{X} \sim F_\theta^{(n)}$

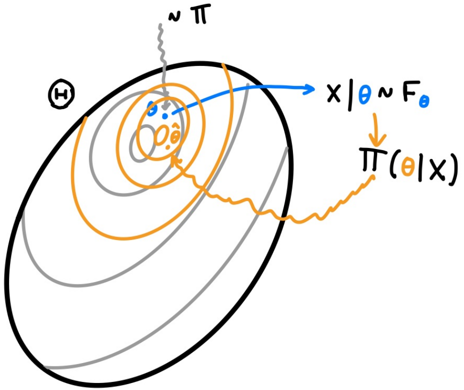
Bayes: $\theta \sim \pi$, $\boldsymbol{X}|\theta \sim F_\theta^{(n)}$

Posterior is $\theta|\boldsymbol{X} \sim \pi(\theta|\boldsymbol{X})$

WHAT ABOUT BAYES?

$$\mathcal{F}^{(n)} = \left\{F_\theta^{(n)}\colon \theta \in \Theta\right\}$$





Frequentist: $\boldsymbol{X} \sim F_\theta^{(n)}$

Bayes: $\theta \sim \pi$, $\boldsymbol{X} | \theta \sim F_\theta^{(n)}$

Posterior is $\theta | \boldsymbol{X} \sim \pi(\theta | \boldsymbol{X})$

WHAT ABOUT FREQUENTIST BAYES?

In *frequentist Bayes*:

We assume that $\boldsymbol{X} \sim F_\theta^{(n)}$ and essentially see the posterior $\pi(\theta|\boldsymbol{X}) \propto f_\theta^{(n)}(\boldsymbol{X})\, \pi(\theta)$ as a sampling distribution

From the posterior we can get:

- Estimators (posterior mode is essential a *weighted* MLE);
- Credible sets;
- Perform test;

I'll return to this later but the prior is crucial for non-parametric models.

SHORT TOUR THROUGH MATHEMATICAL STATISTICS

The Regression model

THE REGRESSION MODEL

We observe (X, Y) and want to say something about the relation between X (predictor) and Y (response).

$$Y = Y \pm f(X) = f(X) + Y - f(X) = f(X) + \varepsilon$$

We think of $\varepsilon = Y - f(X)$ as being small in some appropriate sense:

- For instance, if $\mathbb{E}\varepsilon = 0$, $\mathbb{V}\varepsilon \leq \infty$, then $f(X) = \mathbb{E}[Y|X]$;
- If $\varepsilon|X$ has τ-quantile 0, i.e., $\mathbb{P}(\varepsilon \leq 0|X) = \tau$, or, $\mathbb{P}(Y \leq f(X)|X) = \tau$, then $f(X) = Q_\tau(Y|X)$;

(Unsurprisingly,) what f represents depends on what we assume on ε.

THE REGRESSION MODEL

We observe (X, Y) and want to say something about the relation between X (predictor) and Y (response).

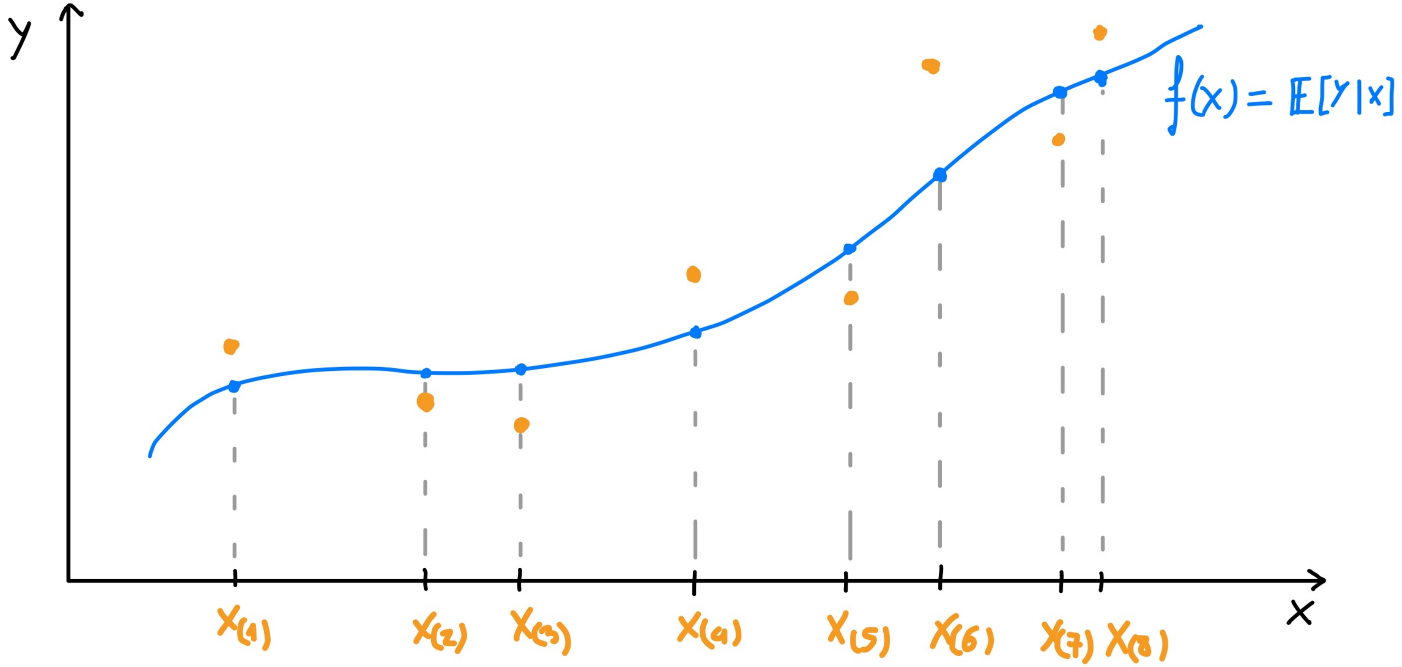
$$Y = f(X) + \varepsilon$$

We think of $\varepsilon = Y - f(X)$ as being small in some appropriate sense.

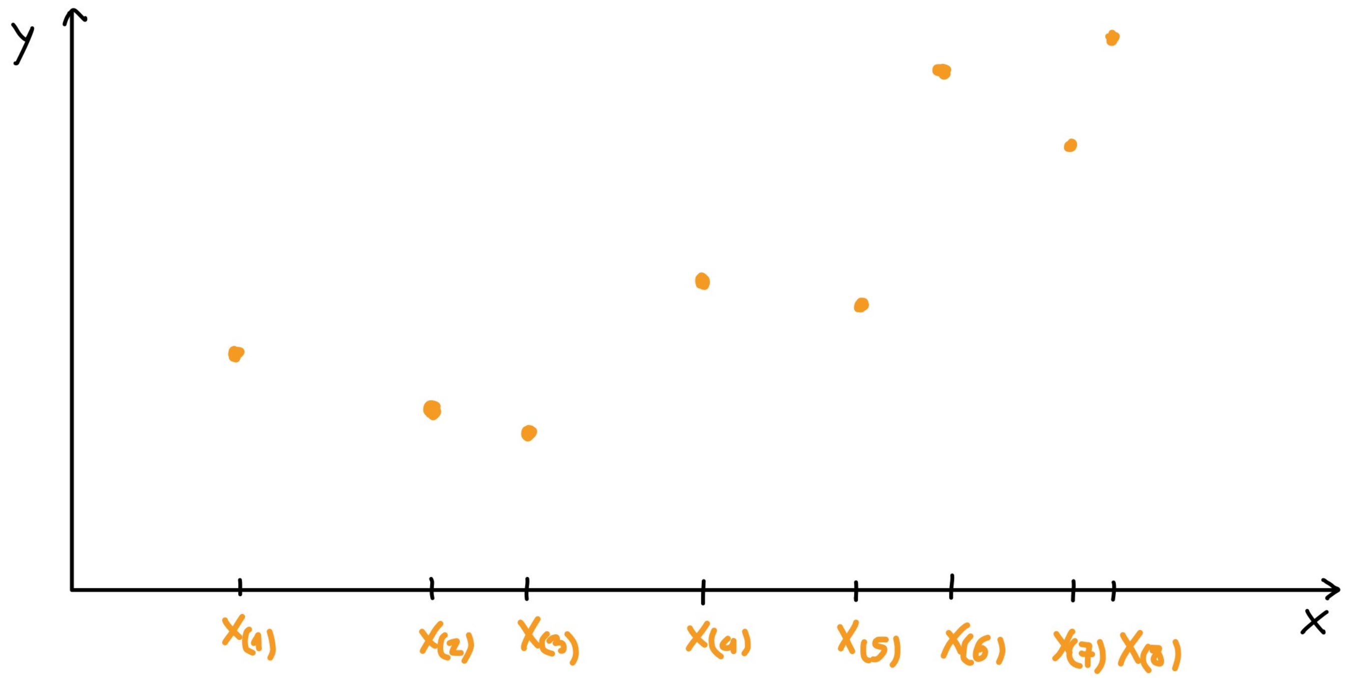
The problem: we observe independent copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of (X, Y) and want to infer f.

For now assume we go for assuming $\mathbb{E}\varepsilon = 0$ so that $f(X) = \mathbb{E}[Y|X]$.

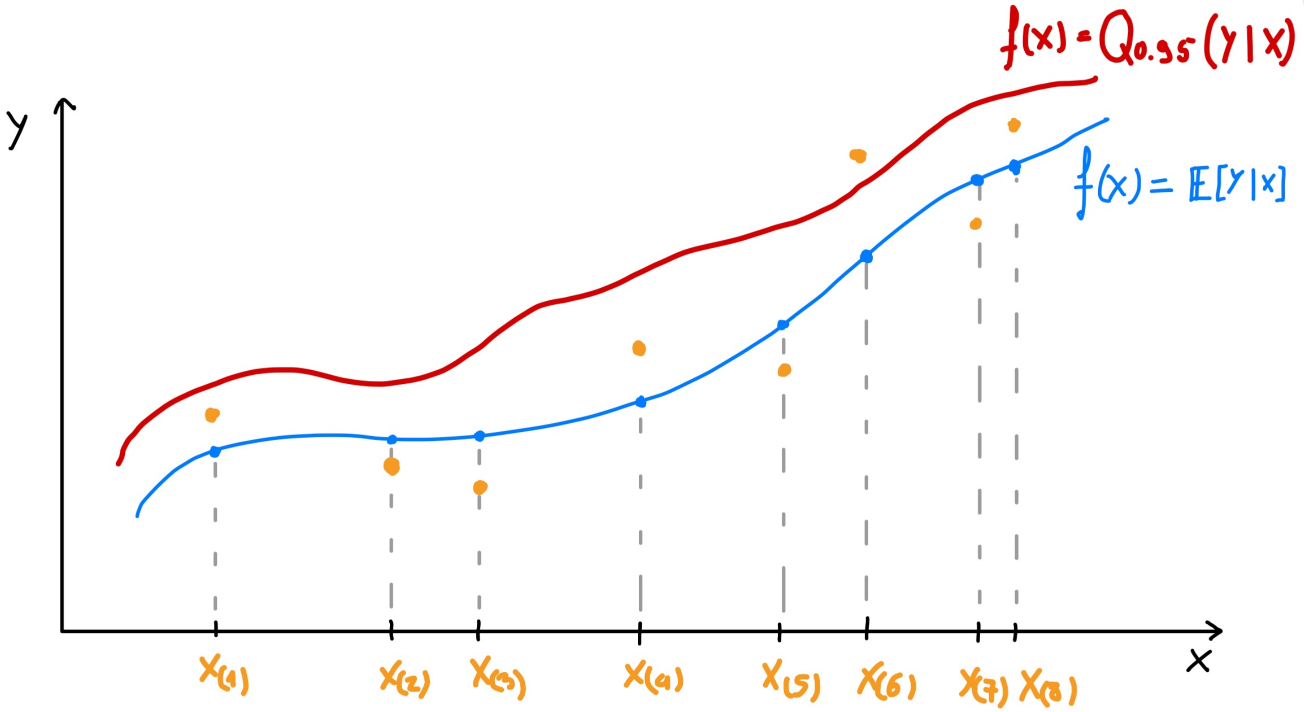
THE REGRESSION MODEL



THE REGRESSION MODEL



THE REGRESSION MODEL



LEAST SQUARES

$(X_1, Y_1), \dots, (X_n, Y_n)$ independent copies of (X, Y), $\mathbb{E}\varepsilon = 0$, $\mathbb{V}\varepsilon \leq \infty$.

In this case we want f to run through the observations so we solve

$$\min_{f \in L_2} \sum_{i=1}^{n} \left(Y_i - f(X_i)\right)^2$$

$$\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(X_i)\right)^2 \quad \xrightarrow{\text{a.s.}} \quad \mathbb{E}\left(Y - f(X)\right)^2, n \to \infty.$$

The function that minimises this limit is $f(X) = \mathbb{E}[Y|X]$ so makes sense to do this.

LEAST SQUARES

$(X_1, Y_1), \dots, (X_n, Y_n)$ independent copies of (X, Y), $\mathbb{E}\varepsilon = 0$, $\mathbb{V}\varepsilon = 0$.

Our estimator of f optimises

$$\min_{f \in L_2} \sum_{i=1}^{n} \left(Y_i - f(X_i)\right)^2$$

The solution to this is *silly*, though… *any function in L_2 that interpolates the data solves the above.*

We could instead solve over $\{a + bx : a, b \in \mathbb{R}\}$ but this is parametric…

Seems like too many options…

PENALISED LEAST SQUARES

$(X_1, Y_1), \dots, (X_n, Y_n)$ independent copies of (X, Y), $\mathbb{E}\varepsilon = 0$, $\mathbb{V}\varepsilon^2 = 0$.

Our estimator of f optimises

We introduce a (positive) penalty.

$$\min_{f \in L_2} \sum_{i=1}^{n} \left(Y_i - f(X_i)\right)^2 + \lambda\, P(f)$$

- If $P(f) = P(f')$, then we pick the one that fits the data best;
- If $\sum_{i=1}^{n} \left(Y_i - f(X_i)\right)^2 = \sum_{i=1}^{n} \left(Y_i - f'(X_i)\right)^2$, then we pick the function with the smallest penalty.
- The $\lambda > 0$ parameter controls the trade-off between the two.

This is what we want but can we actually solve this?

PENALISED LEAST SQUARES

Our estimator of f solves

$$\min_{f \in C} \sum_{i=1}^{n} \left(Y_i - f(X_i)\right)^2 + \lambda\, P(f)$$

Usually we work with spaces of functions that admit a nice representation, say

$$f(x) = \sum_{i=1}^{p} f_i\, \varphi_i(x)\,, \text{with} \left\langle \varphi_i, \varphi_j \right\rangle = \delta_{ij}$$

and we pick penalties like $P(f) = \int f(x)^2 dx = \boldsymbol{f}^T \boldsymbol{f}$.

PENALISED LEAST SQUARES

Our estimator of f comes from optimising

$$\min_{f \in \mathbb{R}^p} (\boldsymbol{Y} - \boldsymbol{\Phi} \boldsymbol{f})^T (\boldsymbol{Y} - \boldsymbol{\Phi} \boldsymbol{f}) + \lambda\, \boldsymbol{f}^T \boldsymbol{f}$$

where $\boldsymbol{\Phi} = \left[\, \varphi_j(X_i) \right]_{ij}$. So we get something quadratic in $\boldsymbol{f}$.

This is solved by $\widehat{\boldsymbol{f}} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \lambda\, \boldsymbol{I})^{-1} \boldsymbol{\Phi}^T \boldsymbol{Y}$ giving

$$\hat{f}(x) = \sum_{i=0}^{p} \widehat{f_i}\, \varphi_i(x)$$

There are many variations of this corresponding to different penalties…

PENALISED LEAST SQUARES (PRIMAL/DUAL)

Our estimator of f optimises

$$\min_{\boldsymbol{f}\in\mathbb{R}^p} (\boldsymbol{Y}-\boldsymbol{\Phi f})^T(\boldsymbol{Y}-\boldsymbol{\Phi f}) + \lambda\, P(\boldsymbol{f}), \quad (\text{dual})$$

or equivalently

$$\min_{\boldsymbol{f}\in\mathbb{R}^p : P(\boldsymbol{f})\le r_\lambda} (\boldsymbol{Y}-\boldsymbol{\Phi f})^T(\boldsymbol{Y}-\boldsymbol{\Phi f}), \quad (\text{primal})$$

where $\boldsymbol{\Phi} = \left[\,\varphi_j(X_i)\right]_{ij}$.

What is the story for quantiles?

SHORT TOUR THROUGH MATHEMATICAL STATISTICS

Quantile Regression

OTHER LOSSES

$(X_1, Y_1), \ldots, (X_n, Y_n)$ independent copies of (X, Y), ε has τ-quantile 0.

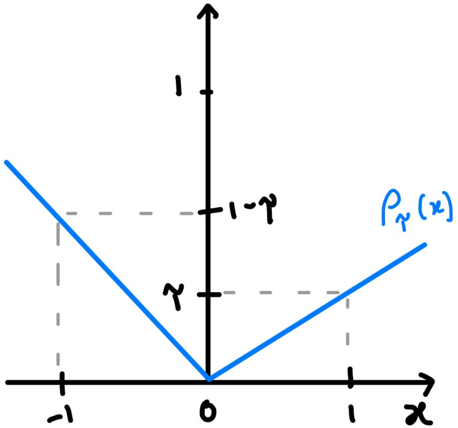
In this case there is asymmetry in terms of under- of over-predicting Y.

We solve

$$\min_{f \in L_2} \sum_{i=1}^{n} \rho_\tau\big(Y_i - f(X_i)\big),$$

where $\rho_\tau(x) = x(\tau - 1\{x < 0\}) = (\tau - 1)x\, 1\{x < 0\} + \tau\, x\, 1\{x \geq 0\}$.

As before, the function that minimizes $\mathbb{E}\rho_\tau(Y - f(X))$ is $f(X) = Q_\tau(Y|X)$.



OTHER LOSSES

$(X_1, Y_1), \dots, (X_n, Y_n)$ independent copies of (X, Y), ε has τ-quantile 0.

If f admits a similar representation as before, then we can equivalently solve

$$\min_{\boldsymbol{f} \in \mathbb{R}^p, \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^n} \sum_{i=1}^{n} \tau \, \mathbf{1}^T \boldsymbol{u} + (1-\tau) \, \mathbf{1}^T \boldsymbol{v}, \qquad s.t., \qquad \boldsymbol{\Phi} \boldsymbol{f} + \boldsymbol{u} - \boldsymbol{v} = \boldsymbol{Y}.$$

This is a linear program which can be solved efficiently.

OTHER LOSSES

$(X_1, Y_1), \ldots, (X_n, Y_n)$ independent copies of (X, Y), ε has τ-quantile 0.

If f admits a similar representation as before, then we can equivalently solve

We introduce a (positive) penalty.

$$\min_{\boldsymbol{f} \in \mathbb{R}^p, \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^n} \sum_{i=1}^{n} \tau \, \mathbf{1}^T \boldsymbol{u} + (1 - \tau) \, \mathbf{1}^T \boldsymbol{v}, \qquad s.t., \qquad \boldsymbol{\Phi f} + \boldsymbol{u} - \boldsymbol{v} = \boldsymbol{Y}, \qquad P(\boldsymbol{f}) \leq r.$$

Penalties (linear, quadratic, other) can also be added here.

QUANTILE CROSSING

Quantile crossing is also a problems sometimes:

- This can be due to low sample size;
- Can be due to inappropriate modelling of f.

Solution is to estimate several quantile curves at the same time and introduce constraint:

$$\min_{f_{\tau_1},\ldots,f_{\tau_q} \in L_2} \sum_{j=1}^{q} w_j \sum_{i=1}^{n} \rho_{\tau_j}\left(Y_i - f_{\tau_j}(X_i)\right), \qquad s.t. \qquad f_{\tau_j}(x) \leq f_{\tau_k}(x), \qquad \text{j} < \text{k}.$$

for $\tau_1 < \cdots < \tau_j < \cdots < \tau_q$.

Penalties can also be added here.

SHORT TOUR THROUGH MATHEMATICAL STATISTICS

Bayesian Inference in Regression

VRIJE UNIVERSITEIT AMSTERDAM

WHAT ABOUT BAYES? (AGAIN)

It turns out that penalisation is closely connected with Bayes.

Consider the model $\boldsymbol{Y} = \boldsymbol{\Phi f} + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, and put prior on $\boldsymbol{f} \sim N\left(\boldsymbol{0}, \frac{\sigma^2}{\lambda}\boldsymbol{\Omega}\right)$; then the posterior is

$$\propto e^{-\frac{1}{\sigma^2}(\boldsymbol{Y}-\boldsymbol{\Phi f})^T(\boldsymbol{Y}-\boldsymbol{\Phi f})} \times e^{-\frac{\lambda}{\sigma^2}\boldsymbol{f}^T\boldsymbol{\Omega}^{-1}\boldsymbol{f}} = e^{-\frac{1}{\sigma^2}(\boldsymbol{f}-\widehat{\boldsymbol{f}})^T(\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\lambda\boldsymbol{\Omega}^{-1})(\boldsymbol{f}-\widehat{\boldsymbol{f}})+Y^T(I+\Phi\Omega\Phi^T)^{-1}Y}$$

$$\propto e^{-\frac{1}{\sigma^2}(\boldsymbol{f}-\widehat{\boldsymbol{f}})^T(\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\lambda\boldsymbol{\Omega}^{-1})(\boldsymbol{f}-\widehat{\boldsymbol{f}})}$$

where $\widehat{\boldsymbol{f}} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{\Omega}^{-1})^{-1}\boldsymbol{\Phi}^T\boldsymbol{Y}$ (looks familiar); we see also that the posterior is Normal.

Note that maximizing the posterior is the same as minimizing (also looks familiar)

$$(\boldsymbol{Y} - \boldsymbol{\Phi f})^T(\boldsymbol{Y} - \boldsymbol{\Phi f}) + \lambda\, \boldsymbol{f}^T\boldsymbol{\Omega}^{-1}\boldsymbol{f}.$$

WHAT ABOUT BAYES? (AGAIN)

It turns out that penalization is closely connected with Bayes.

Consider the model $\boldsymbol{Y} = \boldsymbol{\Phi f} + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$, and put prior on $\boldsymbol{f} \sim N\left(\mathbf{0}, \frac{\sigma^2}{\lambda}\boldsymbol{\Omega}\right)$; then the posterior is $N\left(\widehat{\boldsymbol{f}}, (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{\Omega}^{-1})^{-1}\right)$ and maximizing the posterior is the same as minimizing

$$(\boldsymbol{Y} - \boldsymbol{\Phi f})^T(\boldsymbol{Y} - \boldsymbol{\Phi f}) + \lambda\,\boldsymbol{f}^T\boldsymbol{\Omega}^{-1}\boldsymbol{f}.$$

The posterior is centered at $\widehat{\boldsymbol{f}}$; there is also a certain amount of concentration around the estimator.

Is there something similar going on with *quantiles*?

WHAT ABOUT BAYES FOR QUANTILE REGRESSION?

We can reverse-engineer a likelihood and a prior that leads to the minimization of

$$\sum_{i=1}^{n} \rho_\tau(Y_i - \{\boldsymbol{\Phi f}\}_i)\,.$$

We are still free to to use other priors (penalties.)

We model the likelihood of $\boldsymbol{Y}$ and the prior are as being respectively $\propto e^{-\alpha \sum_{i=1}^{n} \rho_\tau(Y_i - \{\boldsymbol{\Phi f}\}_i)}$ and $\propto 1$.
This likelihood corresponds to an Asymmetric Laplace distribution, and the prior is uniform.

This is an *improper* prior but the respective posterior is *proper* (but not a named distribution.)

We don't know the posterior distribution but we do know it's *concentered around the QR estimate*.

Open questions: how does UQ work for QR.

SHORT TOUR THROUGH MATHEMATICAL STATISTICS

Closing

CLOSING

- This was a very high level tour through Mathematical statistics;
- Please keep in mind that:
    - I omitted a lot of details;
    - I made everything sound more general that it is;
    - I focused more on things that are closer to my work;
- There is a new PhD student starting in 3 weeks (Alexandra Vegelien) working on similar problems;
- At some point we may have some questions for some of you.

THE END

Questions?

VRIJE UNIVERSITEIT AMSTERDAM

WHITEBOARD