Statistical Models

Daniele Zago

July 14, 2022

Contents

Introduction

Instructor: Bruno Scarpa

This course will be focused on topics related to the practice of statistical modelling. Starting from some informal, though-provoking lectures, we will discuss with great depth different topics related to advanced techniques for modelling complex and dependent data. More specifically, the course contents will include the following macroscopic topics:

1. What is a statistical model?
2. Theory of nonparametric models.
3. Models for dependent observations:
    - random effects, multilevel and hierarchical models;
    - complex designs;
    - space-time dependence.
4. Model and variable selection procedures.

Lecture 1: What is a statistical model?

2022-02-23

These first lectures will be devoted to discussing broad topics, avoiding formulas and instead focusing on thought-provoking discussions. Most of the things we will examine in these first two weeks are topics that we already know very well, with the idea to look back at the things that we already have applied in order to better understand them and move forward.

1.1 The role of statistics

References: Cox and Donnelly (2011)

Suggested readings: Senn (2003)

In general, statistics is concerned with the development of methods for **quantitative reasoning**, and it has more in common with philosophy and epistemology than with the cruder methods of accounting.

The idea is that, in science, we want to both collect data and define relationships within the data itself for the general purpose of enriching knowledge.

In order to get formulas and models we need **intuition** about the type of model that we want to apply. Intuition is comprised of both *a*) an innate instinct and *b*) a layer of experience which is trainable through repeated exercises and practice.

When applying a data-analysis procedure, we need to be wary of the fallacies which appear at first to derive from the analysis, but are mere ideological conclusion (*post hoc ergo propter hoc*, "since event Y followed event X, event Y must have been caused by event X") which cannot be supported by the analysis itself; every conclusion has to be supported by facts.

Intuition and imagination are involved in many steps of an analysis that we want to perform:

- an expression of the **question** to be answered;
- the identification of the **variables** and their interactions to be taken into consideration;
- the definition of the **model** which links empirical observations to theoretical data.

A model may be functional to different goals:

1. to **describe** the evidence;
2. to clarify links and **interpret** the phenomenon;
3. to make **predictions** either within or without the limits of the data that we have observed.

> *A model is a simplified representation of the phenomenon of interest, which is functional to a specific objective.*
>
> (Azzalini and Scarpa, 2012)

In first instance, we have a link between variables and the causal relations with the observed data. Moreover, the "*simplified representation*" of the phenomenon can be interpreted under many dimensions: the choice in the elements to include, the nature of the relationship between the

variables, and more other subjective choices which may be hidden under the hood of a statistical model.

If we assume the above definition as a working definition of a statistical model, it naturally follows that a **true model** for a phenomenon of interest does not truly exist. In the words of Box, the fact that a model approximates reality implies that the model might be useful for scientific analysis.

1.2 What is statistics?

Statistics is not pure mathematics, since the primary concern is real-world data. Moreover, we are interested in developing methods which may possess different levels of replicability depending on the particular field of analysis. The theoretical framework of statistics is usually **weak** and it is under scrutiny due to the replication crisis (Ioannidis, 2005; Gelman and Vazire, 2021).

The crucial idea is that nature is hardly described by a manageable number of variables, and the observed data is "noisy" due to **unmeasured covariates**. Hence, even under the same experimental conditions we might observe different values of a variable at the time of measurement. The idea of **probabilistic** statistical modelling is that we **model** by pretending that the data looks "as though" it had been generated from the data-generating process.

Example (Coin tosses)
Coin tosses are the result of myriad of physical characteristics of the toss, such as how the coin was held, where the thumb hit the coin, the impulsive force, ...

We do not either have control or the means to observe all these variables, hence the idea of modelling the coin tosses with Bernoulli model.

Remarks The modelling approach has some important aspects

1. Nature is non-deterministic, but we are pretending that it is.
2. Probability theory: what does the generated data look like?

Statistical methods are employed to identify the model that would best approximate the data-generating process, given the observed data.

Example (Predictive variable)
Suppose that a model for (y_n, x_n) is given by

$$Y = 27X + \varepsilon,$$

where $X \sim \text{Unif}(0, 1)$ and $\varepsilon \sim \mathcal{N}(0, 1)$. Then, we can generate many values of (y, x) by

1. drawing $x \sim \text{Unif}(0, 1)$ and $\varepsilon \sim \mathcal{N}(0, 1)$;
2. set $y = 27x + \varepsilon$.

Remark The knowledge of the model completely specifies the effect of x on y.

In general notation, we define a statistical model for the data $(y_n, x_{n1}, \ldots, x_{np})$ through the specification of

$$\mathbb{P}(Y, X_1, \ldots, X_p) = P(Y|X_1, \ldots, X_p)\mathbb{P}(X_1, \ldots, X_p),$$

hence we have to define a) how to draw X for the n^{th} unit and b) the conditional probability of Y given X.

1.3 Statistical theory and applied statistics

The usual steps of designing a statistical analysis require the ability to:

1. **design** an experiment to gather data on (X, Y);
2. **select** an appropriate model from a family of probability models;
3. **use** the selected model to draw conclusions regarding X and Y.

Suggested readings: Barnett (1999)

Example
We try to fit a variable $x_n \in [-1, 1]$ using a linear model $y = \alpha + \beta x + z$ to get estimates $\widehat{\alpha} = 1.8$, $\widehat{\beta} = 0.1$ and run a hypothesis test to get β to be not significant. However, a mode careful analysis shows that the model should be

$$y = \alpha + \beta x^2 + z,$$

hence if β shows significance then there is a relationship which is not linear.

Remark The use of statistical tools tells us that our initial model was wrong.

Question
Assume a model $y = \alpha + \beta x + z$, with $z \sim \mathcal{N}(0, \sigma^2)$. How can you find out whether it is a reasonable mechanism for the data?

When we have a model, we can gather data to perform model selection, fitting, and checking, to get conclusion from our analysis.

Remark. In general, the world is not "i.i.d normal", and the data usually displays unwanted features such as missing information, corrupted values and outliers. Good statisticians "live" with the data and not with statistics textbooks.

David Cox: statistical analysis typical arises when there is unexplained and haphazard variation. Without variability there is no need for statistics, since this would simply become the realm of accounting. Such variability might be either a result of *natural variability*, *measurement errors*, or

other forms of variability.

While natural variability might be of interest, the measurement error is in principle a nuisance, although it can have an effect on the interpretation of results.

Asking the "right" question is a crucial step of the problem, which is desirable to pose a priori of the study. In other cases, we have that the research question of primary concern may emerge only as the study evolves. Major changes of focus need confirmation in supplementary investigations.

In the extreme case, there is a perception that data has to contain *generally useful* information, about something which is usually not clear. **Data mining** is often used in such contexts, where methods can uncover possible relationships, but conclusions are in most cases tentative and in need of independent confirmation.

Large amount of data $\neq$ *large amount of information*

Example (Negative example of big data)
Carpenter et al. (1997) analyzed 10^6 observations from UK cancer in order to investigate association between occupations and cancer rates in body sites.

The data was problematic since some categories were missing *not at random*, since multiple occupations and cancers at multiple sites were excluded for unknown reasons. Hence, this work used graphical approaches to obtain tentative conclusions with respect to previously well-established relationship.

Example (Conficting observational and experimental evidence)
A number of observational studies reviewed in Grady et al. (1992) suggested that women using hormone replacement therapy (HRT) for long period of time had a lower coronary heart disease rate to apparently comparable control groups.

However, the investigators had no way to control which specific woman did or did not use HRT. In a randomized experiment, women were assigned at random either to HRT or to inactive control.

The trial was stopped because of a possible adverse effect on total cardiovasculary events, and because of a strong evidence of no beneficial effects.

Lecture 2: What is a statistical model? (ii)

2022-02-25

In this lecture we discuss the problem of *measurement* and how the way we measure – or decide to measure – a phenomenon could affect the conclusions and the strength of our statistical analysis. The theory of applied statistics often neglects stressing the importance of correctly measure the quantities of interest, although the consequences of errors in this phase of the analysis are catastrophic

2.1 Principles of measurement

References: Cox and Donnelly (2011)

In any statistical analysis, we typically start from some measurement, which we assume to be **constructively valid**. With construct validity we mean that the measurements actually record the features of subject-matter concern, which derive from a careful examination of the problem at hand (see previous lecture). Specifically, we want to record a number of features which are enough to concisely capture the relevant aspects of the problem.

In general, the these features yield both **reliable** and **reasonably reproducible** results, and the **cost** of the measurements is commensurate with their importance. Moreover, we need to make sure that the measurement process does not **distort** the system under study, unless we want to risk our whole study to have possibly biased conclusions.

In general, the description of complex multidimensional phenomena by a limited number of summary measures requires the careful specification of objectives: summarizing using one-dimensional summaries has to be avoided except for highly specific contexts.

There are two major approaches to data analysis, each of which has its advantages and disadvantages:

a) do something as simple as possible and work your way from there;

b) start from something very complex, since there will be something useful.

What we do is to apply **feasible** and **transparent** methods, which are requirements of usability for other people to check. Indeed, transparent methods can show the pathway from the data to the conclusion, in order to see which aspects of the data has consequences on the outputs.

Data cleaning should be performed as soon as possible after data collection, since we might require further data if we observe problems in the collection process.

› anomalous values

› internal inconsistencies

› "sticking instruments" when there are thresholds in the instrument precision

› inspection for zero values, missing or irrelevant values.

The formal analysis is the most challenging part of the analysis, but we usually arrive there after a long process of data collection. Presentation is a very important part as well, which is something usually neglected and carried out with little professionality.

2.2 Formal analysis

Some methods of analysis may be described as algorithmic, i.e. relationships are recovered by a computer algorithm by minimizing a plausible criterion, which is determined by the user.

Example (Least squares)
The least squares algorithm was originally intended as a smoothing device for fitting a parametric curve to empirical measurements of trajectories. Only in later analysis there has been a statistical justification of the least squares algorithm under a probabilistic model for the data.

Typical statistical models are based on formal probability models, although we do not exclude purely algorithmic methods in the initial stages of the reduction of complex data.

Most of classical statistical models centers on analysis based on probability models for the data, leading – hopefully – to greater subject-matter understanding. Sometimes, such a probabilistic model is augmented by using more specific probabilistic theories of the system under investigation, in order to improve a more descriptive model. This is especially the case of epidemiologic and climate change models, which are representations of the dynamics of a complex systems via the specification of multiple partial differential equations.

The more tightly a model is specified, the more detailed are the conclusions that can be drawn from it.

One topic is building a workflow for checking and modifying our models

1. Conclusions if model is not true?
2. Check correctness
3. Use a simpler/different variation?

Most of classical statistical models are developed to interpreting data with the object of enhancing understanding. Other times, however, there is an aim towards **empirical predictions** of new unobserved features or new study individuals. Under this framework, we are interested in maximizing the agreement between the prediction and the ultimately realized value. When the predictive performance is the only criterion of usefulness, interpretation of the parameters in the model becomes irrelevant, and the choice between equally well-fitting models may be either based on convenience or cost.

Prediction is closely related to **decision problems** and **Bayesian statistics**, e.g. deciding whether units of production have to be accepted or not in an industrial inspection. The assessment of any prediction method is judged by its empirical success on data which is gathered independently from the data used to set up the prediction method.

When we are interested in prediction, an important but under-emphasized issue is the **stability** of the designed predictive method with respect to the variability of observed data.

Therefore, sometimes it is better to use a suboptimal procedure that works well in a wide range of settings.

Principles Some aspects which enter in the principles of applied statistics are

- formulation of **research questions**
- **design** of investigations
- production of effective and reliable **measurement procedures**
- **!!** development of methods of analysis with suitable **software**, which addresses the primary research question and give some assessment of uncertainty.

Theory and practice Somewhat in contrast, the emphasis of applied statistics is on the subject matter, rather than on the statistical techniques by themselves.

Lecture 3: What is a statistical model? (iii)

2022-03-02

Suggested readings: Breiman (2001), on the two cultures of statistics.

> *The great adventure of statistics is in gathering and using data to solve interesting and important real-world problems.*
>
> (Leo Breiman)

3.1 The two cultures of statistics

Leo Breiman raises the issue of statisticians belonging to two "cultures" when applying statistical modeling.

1. **Explanatory approach**: The data are generated by a stochastic data model: the model is validated using yes/no goodness-of-fit tests and residual examination.
2. **Predictive approach**: Algorithmic models where the data mechanism is unknown: model validation is based on predictive accuracy.

His conclusion is that the first approach has kept statistician (up to the year 2001 at least) from working on a wide range of interesting problems. Algorithmic modelling has instead developed rapidly in fields outside statistics, both in theory and practice.

When Breiman was in academia, he realized that all articles started and ended with assumptions over the data models. Data modelling has given many successes in analyzing data and getting information about the mechanisms producing the data.

Critique About the fact that statisticians start from assumptions about how the model correctly describes the world. There is a large misuse of models, leading to questionable conclusions about the underlying mechanism.

Problems The work of statisticians is often done under wrong assumptions and our relevance in science has shrunk over the years in favour of more algorithmic solutions.

Take-home If we want to solve problems, we need to move away from exclusive dependence on data models and adapt a more *diverse* set of tools, which may work better under different circumstances.

Statistical academia

- Focus on finding a good solution and live with **data before the model**, by trusting the model that we built beforehand. The perfect example of this is removing outliers from a linear model, since we take for granted that the model is correct a priori.
- There is a wide spectrum of opinions regarding the usefulness of what is published in the Annals of Statistics to the field of statistics as "science that deals with data". Breiman is at the low end of the spectrum.
- The idea of thinking about a model when faced with an applied problem is that the statistician can invent a reasonably good parametric model for a complex mechanism devised by nature.

When a model is fit to data, the conclusions are about the mechanisms of the model and not about the mechanism of nature. Hence, it's clear that

Poor model performance $\implies$ wrong conclusions.

Up until a few decades ago, belief in the model was almost religious and every article worked conditionally on the fitted particular model. Indeed, the theory of linear models was used with little consideration as to whether the data at hand could have been generated by a linear model.

Example (Gender discrimination)
Study on gender discrimination related to salaries in a particular faculty. The design of the study raises issues that enter before the consideration of the model:

1. Can the data gathered answer the question posed?
2. Is inference justified when the sample is the entire population?
3. Should a data model be used?

The focus was on the model, not on the problem itself. Indeed, the conclusions can be framed as

a) What does a "significant" $\widehat{\beta}_j$ mean? Note that the assumption $H_0 : \beta^* = 0$ includes all other assumption related to Gaussianity, linearity of the model, and the same linear combination of X_j's for all individuals.

b) Does a "significant" $\widehat{\beta}_j$ imply sexual discrimination?

c) Where does the randomness come from? Not on X, since we condition on X fixed. We don't even have a sample since we observe all salaries of the faculty.

Moreover, we only observe 25 variables out of all the possible personal qualities of each individual, which might influence the individual salary outside the scope of the observed covariates.

Projections Also observe that the contribution to a x_j is the sum of all contributions correlated to x_j that are unobserved in the current model. Hence, $X_1, \ldots, X_p$ are **surrogates** for unmeasured variables that not in the equation.

Conclusions What we can say is that in the class of linear predictors, `sex` improves predictions. Perhaps we can find nonlinear predictors such that the performance is even better without the inclusion of `sex`.

Regression at best studies relationship between Y and $X_1, \ldots, X_p$ and this relationship may change if either

1. More X's are added
2. Some X's are deleted

Hence, how can one investigate for evidence of sex discrimination? The problem is that we are not discussing *experimental data*, where we do have high control over the variables.

Additional problems

- Omnibus goodness-of-fit tests, which test many directions simultaneously, will not reject until the lack of fit is extreme.
- In high dimensions, the interactions can produce passable residual plots for a variety of models.
- There may be several different models which are equally good in describing the data (Rashomon effect), and a slight perturbation in the data might cause a skip from one model to another. This might be a justification for the **stacking** approach.

Suggested readings: Shmueli (2010), on predictive and explanatory modeling.

Explanatory modeling and *predictive modeling* reflect the different processes related to using data. These approaches are usually considered separate approaches to data analysis:

1. *Descriptive modeling*, which is aimed at summarizing or compactly representing the data. Causal theory is absent, and the focus is at the measurement level.
2. *Explanatory modeling*, which are used for **testing causal theories** but are often association-based models applied to observational data. The justification is that the theory itself provides the causality mechanism.
3. *Predictive modelling*, where the goal is predicting Y given the input values X in terms of point predictions, prediction regions, and predictive distributions. Predictive modeling is nearly absent in many scientific fields, and the prediction problem is often considered unscientific.

Predictive modelling is often valued for its **applied utility**, although many still consider it unscientific.

Explaining and predicting are different Consider a theory that postulates $\mathcal{X} \to \mathcal{Y}$ via the relationship

$$\mathcal{Y} = \mathcal{F}(\mathcal{X}),$$

and the **operationalization** of $\mathcal{F}$ into a statistical model is to find

$$\mathbb{E}[Y] = f(X),$$

where $\mathcal{F}$ is considered in light of the current study design. In the *explanatory* context, we are interested in matching f to $\mathcal{F}$ as closely as possible, whereas in the *predictive* context we focus our attention on estimating f with the highest degree of accuracy.

$X \to Y$, whereas X is associated to Y
Theory-data: f is constructed to interpret the relationship, while in predictive modeling we construct f from the data.
Retrospective-prospective:
Bias-variance: the EPE is

$$\mathbb{V}[Y] + \text{Bias}^2 + \mathbb{V}[\widehat{f}(x)]$$

Example (Is the "true model" the best predictive model?)
Notice that the "wrong" model can sometimes predict better than the "true" model. Choosing a biased f^* in place of f can be undesirable from a theoretic-explanatory point of view. However, we can have a better predictive model using f^*.

There are some values of x_1 and x_2 such that the overall **expected prediction error** (EPE) is lower for the misspecified model. Specifically, leaving out q predictors has a lower EPE when the following inequality holds,

$$q\sigma^2 > \beta_2^\top X_2^\top (I - H) X_2 \beta_2.$$

This can happen in many practical situations, and provides a sort of "**paradox**" where the true model is not the best at yielding predictions.

There is a fundamental gap between these approaches, because an optimal model from a predictive point of view is not necessarily the best model for representing the underlying mechanism of nature.

Explanatory modelling

- **Goal**: Statistical power, bias reduction.
- **Design**: Experimental data.
- **Measurements**: Reliable instruments (item-response theory), factorial designs.
- **Preparing data**: Throw away data if unusable.

Predictive modelling

- **Goal**: Lowering overall MSE
- **Design**: Observational data.
- **Measurement**: Focus on measurement quality, response-surface methodology to estimate a nonlinear f.
- **Preparing data**: Using regression models with missingness in the dummy variables.

Data splitting. When focusing on prediction, we want to best predict data that we did not yet see, in order to minimize the combined bias and variance of the model. Data partitioning reduces statistical power and is usually applied for the purpose of preventing overfitting when considering multiple competing models. In these settings, the bias sacrifice is usually small when data is abundant.

References

Azzalini, A. and Scarpa, B. (2012). *Data Analysis and Data Mining: An Introduction*. Oxford University Press.

Barnett, V. (1999). *Comparative Statistical Inference*. John Wiley & Sons.

Breiman, L. (2001). «Statistical Modeling: The Two Cultures». In: *Statistical Science* 16.3, 199–231.

Carpenter, L. M. et al. (1997). «Examining Associations between Occupation and Health by Using Routinely Collected Data». In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 160.3, 507–521.

Cox, D. R. and Donnelly, C. A. (2011). *Principles of Applied Statistics*. Cambridge, UK ; New York: Cambridge University Press.

Gelman, A. and Vazire, S. (2021). «Why Did It Take So Many Decades for the Behavioral Sciences to Develop a Sense of Crisis Around Methodology and Replication?» In: *Journal of Methods and Measurement in the Social Sciences* 12.1.

Grady, D. et al. (1992). «Hormone Therapy to Prevent Disease and Prolong Life in Postmenopausal Women». In: *Annals of Internal Medicine* 117.12, 1016–1037.

Ioannidis, J. P. A. (2005). «Why Most Published Research Findings Are False». In: *PLOS Medicine* 2.8, e124.

Senn, S. (2003). *Dicing with Death: Chance, Risk And Health*. 1 edition. New York: Cambridge University Press.

Shmueli, G. (2010). «To Explain or to Predict?» In: *Statistical Science* 25.3, 289–310.

Part I

Nonparametric statistics

Instructor: Carlo Gaetan

References: Wasserman (2005)

> *Nonparametric statistics can and should be broadly defined to include all methodology that does not use a model based on a single parametric family.*
>
> (Handles, Hettmansperger and Casella)

> *The basic idea of nonparametric inference is to use data to infer an unknown quantity while making as few assumptions as possible.*
>
> (Wasserman, 2005)

This part of the course will focus on nonparametric statistics, and with particular focus on nonparametric estimation and regression. The topics of rank-based methods, such as rank statistical hypothesis tests and permutation methods will be left for other, more in-depth, courses. Here, our interest is in describing the statistical properties of nonparametric smoothers and nonparametric regression methods, with the aim of investigating their frequentist properties and generalizing confidence interval to functional confidence bands.

Lecture 4: Nonparametric statistics

2022-03-09

The main idea of nonparametric statistics is to make inferences about unknown quantities without resorting to simple parametric reductions of the problem.

4.1 Introduction to nonparametric statistics

Example (Parametric approach)
Suppose $Y \sim F$ and we wish to estimate $\mathbb{E}[Y]$ or $\mathbb{P}(Y > 1)$; the approach taken by parametric statistic is to assume F to belong to a family of distributions

$$\mathcal{F} = \left\{F(\cdot, \vartheta), \vartheta \in \Theta \subseteq \mathbb{R}^d\right\},$$

which can be described by a finite number of parameters, $d < \infty$. Inference about the quantities we were originally interested in ($\mathbb{E}[Y]$ or $\mathbb{P}(Y > 1)$) are carried out based on assuming $Y \sim F(\cdot; \widehat{\vartheta})$ for an estimated set of parameters $\widehat{\vartheta}$.

Another example is to assume a linear model

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X,$$

and we use the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to carry out inference about $\mathbb{E}[Y|X]$.

Both of the aforementioned parametric approach rely on a reduction of the original problem. They assume that all uncertainty regarding F, or $\mathbb{E}[Y|X]$, can be reduced to just two unknown numbers (i.e. parameters).

The perspective of nonparametric statistics is to make as few assumptions as possible about the data, for instance

- we could allow $F(y)$ to be any function that satisfies the properties of a cumulative distribution function;
- we could allow $\mathbb{E}[Y|X]$ to be any continuous function of X.

Obviously, this requires the development of a whole new set of tools, since instead of estimating parameters we will be estimating functions—which are much more complex objects to handle.

4.2 Estimating the CDF and functionals

Let $Y_1, Y_2, \ldots, Y_n \sim F$ be a random sample from Y with cumulative distribution function

$$F(y) = \mathbb{P}(Y \leq y).$$

Def. (Empirical CDF)

We define the **empirical cumulative distribution function** as the nonparametric estimator of the cumulative distribution function,

$$\widehat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,y]}(Y_i).$$

Remark. The above function is an average of random variables, $T(Y) := \mathbb{1}_{(-\infty,y]}(Y)$.

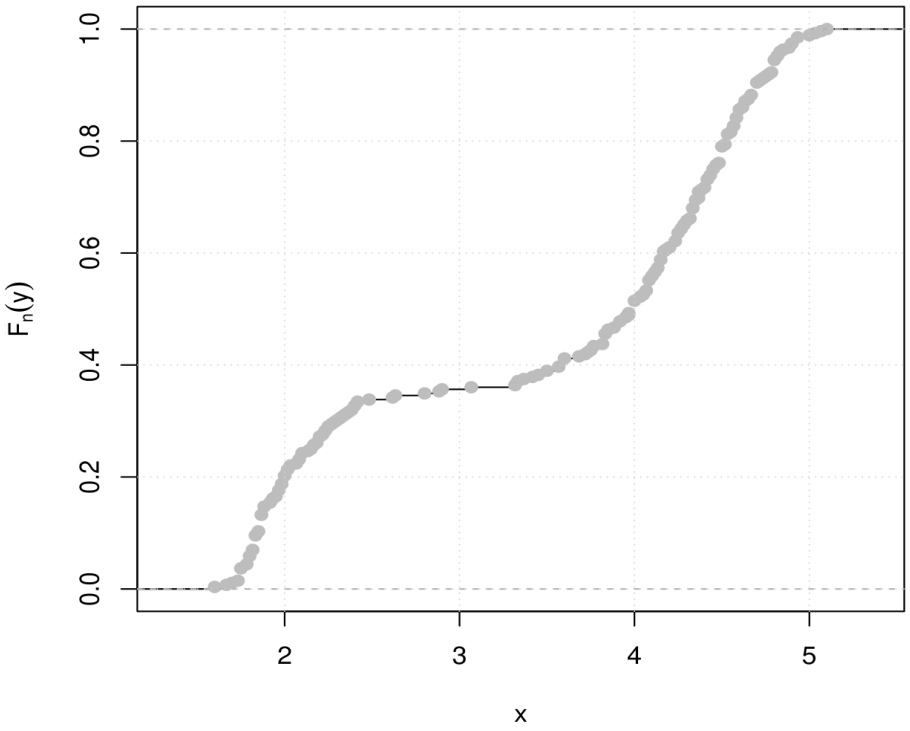


Figure 1: Estimated cumulative distribution function for the Old Faithful geyser dataset.

Prop. 1 (Properties of the ecdf)

Let $Y_1, Y_2, \ldots, Y_n \sim F$ and $\widehat{F}_n$ be the empirical cumulative distribution function, then for any fixed value y we have

1. $\mathbb{E}[\widehat{F}_n(y)] = F(y)$ *and* $\mathbb{V}[\widehat{F}_n(y)] = \frac{F(y)(1-F(y))}{n}$
2. $\widehat{F}_n(y) \xrightarrow{P} F(y)$
3. $\sqrt{n}\big(\widehat{F}_n(y) - F(y)\big) \xrightarrow{d} \mathcal{N}\big(0, F(y)(1-F(y))\big)$.

Proof.

Homework.

□

Theorem 1 (Glivenko-Cantelli)
If $\widehat{F}_n$ is the empirical cumulative distribution function, then

$$\sup_y |\widehat{F}_n(y) - F(y)| \xrightarrow{a.s.} 0.$$

Remark. This result is much more powerful, since it states a functional approximation result.

Theorem 2 (Dvoretzky-Kiefer-Wolfowitz)
For any $\varepsilon > 0$, we have

$$\mathbb{P}(\sup_y |\widehat{F}_n(y) - F(y)| > \varepsilon) \leq 2e^{-2n\varepsilon^2}.$$

Remark. The DKW inequality specifies the rate of convergence of the Glivenko-Cantelli theorem.

We discuss the notion of a confidence interval for a CDF F and compare its notions with those of **confidence bands**.

› A **pointwise** confidence interval finds a region $C(y)$ such that, for any F and fixed $y \in \mathbb{R}$,

$$\mathbb{P}\big(F(y) \in C(y)\big) \geq 1 - \alpha$$

› A different approach to inference is to find a **confidence band** $C(y)$ such that, for any CDF F

$$\mathbb{P}\big(\forall y : F(y) \in C(y)\big) \geq 1 - \alpha.$$

Prop. 2 (Confidence band)
For any distribution function F and all n,

$$\mathbb{P}\big(\forall y : L_n(y) \leq F(y) \leq U_n(y)\big) \geq 1 - \alpha,$$

where

$$L_n(y) = \max\left\{\widehat{F}_n(y) - \sqrt{\frac{1}{2n}\log(2/\alpha)}, 0\right\}$$

$$U_n(y) = \min\left\{\widehat{F}_n(y) + \sqrt{\frac{1}{2n}\log(2/\alpha)}, 1\right\}$$

Proof.
Homework.

□

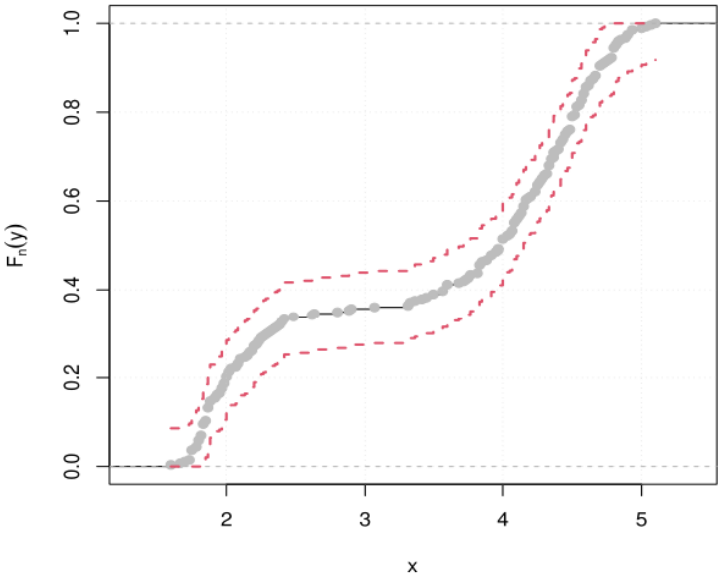


Figure 2: Nonparametric confidence band for the function $\widehat{F}_n$.

This important result now shows that empirical cumulative distribution function is a maximum likelihood estimator. Assuming that

$$f_n(y) = \sum_{i=1}^{n} p_i \mathbb{1}_{\{y\}}(Y_i),$$

then the nonparametric or **empirical likelihood** for $(p_1, \ldots, p_n)$ is

$$L(p_1, p_2, \ldots, p_n) = \prod_{i=1}^{n} f_n(Y_i) = \prod_{i=1}^{n} p_i.$$

Prop. 3 (Geometric mean is bounded by the arithmetic mean)
For any $p_1, \ldots, p_n$ it holds that

$$\left(\prod_{i=1}^{n} p_i \right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^{n} p_i = \frac{1}{n},$$

and the equality holds $\iff$ $p_1 = \ldots = p_n = 1/n$.

Proof.
By Jensen's inequality,

$$\log \left(\frac{\sum_{i=1}^{n} x_i}{n} \right) \geq \sum_{i=1}^{n} \frac{1}{n} \log x_i = \sum_{i=1}^{n} \left(\log x_i^{1/n} \right) = \log \prod_{i=1}^{n} x_i^{1/n}.$$

Applying the exponential function to both sides yields the result.

□

Theorem 3 (Nonparametric maximum likelihood)
The empirical cumulative distribution function $\widehat{F}_n$ maximizes the empirical likelihood

$$L(p_1, \dots, p_n) = \prod_{i=1}^{n} p_i.$$

Proof.
Using Prop. 3 and setting $\widehat{p}_i = 1/n$ to get the maximum, we have that

$$L(p_1, \dots, p_n) \leq L(\widehat{p}_1, \dots, \widehat{p}_n),$$

and the empirical cumulative distribution function $\widehat{F}_n$ corresponds to the cumulative distribution function of the density

$$\widehat{f}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_i\}}(y).$$

Hence, it maximizes the empirical likelihood.

□

4.3 Statistical functionals

Why do we put so much emphasis on estimating F? The reason is that $\widehat{F}_n$ will play the same role in nonparametric estimation played in parametric estimation by the MLE $\widehat{\vartheta}$. In general we are interested in estimating transformations of $\widehat{F}_n$, although nonparametric statistical methods do not easily provide such estimates.

Def. (Statistical functional)
A **statistical functional** $T(F)$ is any function of F.

Example (Functionals)
Some transformations of $\widehat{F}_n$ can be easily expressed as functionals

- *Mean*: $T(F) = \int y \, \mathrm{d}F(y)$.
- *Mean*: $T(F) = \int y^2 \, \mathrm{d}F(y) - \left(\int y \, \mathrm{d}F(y) \right)^2$.
- *Quantile*: $T(F) = F^{-1}(p)$, where

$$F^{-1}(p) = \inf\{y : F(y) \geq p\}, \quad p \in (0, 1).$$

Estimation of quantities such as the above is based on the nonparametric mle of F,

$$\widehat{T}(F) = T(\widehat{F}_n).$$

Example (Functionals (cont.))
The estimated functionals in the above example correspond to $\overline{Y}$, $\widehat{\sigma}^2$ and the sample quantile, respectively.

Is the plug-in estimator a good estimator in all cases? Not always: consider the estimation of a density, where the functional is

$$T(F) = \frac{\partial}{\partial y} F(y),$$

then we would like to characterize the convergence of $T(\widehat{F}_n) \to T(F)$ using the fact that from theorem 1 we have $\widehat{F}_n \xrightarrow{\text{a.s.}} F$. In order to do so, we need to introduce the concept of derivative of a statistical functional.

Def. (Gâteaux derivative)
The **Gâteaux derivative** of T at F in the direction G (G is a CDF) is defined by

$$\begin{aligned} L_F(T; G) &= \lim_{\varepsilon \to 0} \frac{T\big((1-\varepsilon)F + \varepsilon G\big) - T(F)}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \frac{T\big(F + \varepsilon D\big) - T(F)}{\varepsilon}, \end{aligned}$$

if we define $D = G - F$.

Remark. The Gâteaux derivative is a generalization of the concept of a directional derivative to the functional analysis setting.

Remark. From a statistical perspective, it represents the rate of change in a statistical functional upon a small amount (ε) of contamination by another distribution G (mixture of distributions). This has a long history in the **robust inference** literature.

Statisticians usually prefer to work with a particular Gâteaux derivative, which is a special case of the above definition when G places a point mass of 1 at the point y, i.e.

$$G_y(u) = \mathbb{1}_{(-\infty, y]}(u).$$

This yield the so-called **influence function** definition as a specialization of the Gâteaux derivative.

Def. (Influence function)
The **influence function** of T at F is defined by

$$L_F(y) = \lim_{\varepsilon \to 0} \frac{T\big((1-\varepsilon)F + \varepsilon G_y\big) - T(F)}{\varepsilon}, \tag{1}$$

where $G_y(u) = \mathbb{1}_{(-\infty, y]}(u)$.

Def. (Empirical influence function)

The **empirical influence function** is the estimate of $L_F(y)$ using the empirical cumulative distribution function,

$$\widehat{L}_n(y) = L_{\widehat{F}_n}(y) = \lim_{\varepsilon \to 0} \frac{T\big((1-\varepsilon)\widehat{F}_n + \varepsilon G_y\big) - T(\widehat{F}_n)}{\varepsilon}.$$

Example (Influence function)

Consider the functional $T(F) = \mu = \int y \, \mathrm{d}F(y)$, then the perturbed functional is

$$T\big((1-\varepsilon)F + \varepsilon G_y\big) = (1-\varepsilon)\mu + \varepsilon y,$$

so that

$$L_F(y) = y - \mu.$$

An important consequence is that, for $T(F) = \mu$ we have $\mathbb{E}[L_F(Y)] = 0$. Estimating it using the empirical influence function we have $T(\widehat{F}_n) = \bar{y}$ and so

$$\widehat{L}_n(y) = y - \bar{y}.$$

Linearity. Linear functionals such as the mean are particularly easy to work with, since linearity allows the straightforward characterization of the influence function.

Def. (Linear functional)

A functional T is a **linear functional** if it is defined as

$$T(F) = \int a(y) \, \mathrm{d}F(y),$$

for some function $a(y)$.

Remark. We have that the influence function has a very simple representation,

$$L_F(y) = a(y) - T(F)$$

$$\widehat{L}_n(y) = a(y) - T(\widehat{F}_n)$$

The Gâteaux derivative has many of the same properties as ordinary derivatives, in particular we have that:

Theorem 4 (Chain rule)

Suppose that a functional can be written as $T(F) = h\big(T_1(F), \ldots, T_m(F)\big)$, for some derivable function $h : \mathcal{F} \longrightarrow \mathbb{R}$, then

$$L_F(y) = \sum_{i=1}^{m} \frac{\partial h}{\partial t_i} L_i(y),$$

where $L_i(y)$ is the influence function of $T_i(F)$.

4.4 Functional delta method

In parametric statistics, we estimate ϑ by $\widehat{\vartheta}_n$ and we can use the delta method to obtain approximate distributional results for $g(\widehat{\vartheta}_n)$. For instance, if $\widehat{\theta}_n$ is an estimator and we let g be a smooth transformation with derivative g', then we have that

$$\sqrt{n}\big(g(\widehat{\vartheta}_n) - g(\vartheta)\big) \xrightarrow{d} \mathcal{N}\big(0, \sigma^2 g'(\vartheta)^2\big).$$

In nonparametric statistics we can use the **functional delta method** to obtain distributional results for $T(\widehat{F}_n)$.

Lemma 1 ("Functional theorem of calculus")
Assume that $T(F)$ is a linear functional, then for any function G it holds that

$$T(G) = T(F) + \int L_F(y)\,\mathrm{d}G(y). \tag{2}$$

Corollary 1 (Expectation of a linear functional)
It follows from the above lemma that, by setting $G = F$,

$$\int L_F(y)\,\mathrm{d}F(y) = 0. \tag{3}$$

Remark. Equation (3) states that the influence function – which is a sort of "derivative" – behaves like the score function in parametric estimation. Indeed, we know from Bartlett's identities that

$$\mathbb{E}_\vartheta\Big[\frac{\partial}{\partial\vartheta}\log\ell(\vartheta; Y)\Big] = 0.$$

Suppose that our plug in estimate of $T(F)$ is $T(\widehat{F}_n)$, then this plug-in estimate in the linear case can be written using (2) and setting $G = \widehat{F}_n$ as

$$T(\widehat{F}_n) = T(F) + \underbrace{\frac{1}{n}\sum_{i=1}^{n} L_F(Y_i)}_{\Longrightarrow \text{ CLT}}.$$

Proof.
Homework.

□

Lemma 2 (Central limit theorem)
Let $\tau^2 = \int L_F^2(y)\,\mathrm{d}F(y)$, then if $\tau^2 < \infty$ we have that

$$\sqrt{n}\frac{T(\widehat{F}_n) - T(F)}{\tau} \xrightarrow{d} \mathcal{N}(0,1).$$

Lemma 3 (estimator of τ)
Let $\widehat{\tau}_n^2 = n^{-1}\sum_{i=1}^n \widehat{L}_n^2(Y_i)$ be the plug-in estimator, then we have that

$$\widehat{\tau}_n^2 \xrightarrow{P} \tau^2$$

$$\frac{\widehat{SE}_n}{SE} \xrightarrow{P} 1,$$

where $\widehat{SE}_n = \widehat{\tau}_n/\sqrt{n}$ and $SE = \sqrt{\mathbb{V}[T(\widehat{F}_n)]}$.

Theorem 5 (functional delta method for a linear functional)
If $\widehat{\tau}_n^2$ is the plug-in estimator of τ, we have that

$$\sqrt{n}\frac{T(\widehat{F}_n) - T(F)}{\widehat{\tau}_n} \xrightarrow{d} \mathcal{N}(0,1).$$

Remark. In the general case of a non-linear functional T, the above theorem still holds by writing

$$T(\widehat{F}_n) = T(F) + \frac{1}{n}\sum_{i=1}^n L_F(Y_i) + o_p(1).$$

We can try to extend Taylor's theorem to the functional case. We recall that

Theorem 6 (Taylor's theorem)
Suppose f is a real function on $[a,b]$, $f^{(K-1)}$ is continuous on $[a,b]$, $f^{(K)}(x)$ is bounded for $y \in (a,b)$ then for any two distinct points $y_0 < y_1$ in $[a,b]$ there exists a point y between $y_0 < y < y_1$ such that

$$f(y_1) = f(y_0) + \sum_{k=1}^{K-1}\frac{f^{(k)}(y_0)}{k!}(y_1 - y_0)^k + \frac{f^{(k)}(y)}{K!}(y_1 - y_0)^K.$$

The question is whether or not there exists a functional extension to the above Taylor theorem. The answer is that yes, there exists, under the condition that T is Hadamard differentiable at F. Define $D = G - F$ and the Gâteaux derivative $L_F(D)$ of T as

$$\lim_{\varepsilon \to 0}\left(\frac{T(F + \varepsilon D) - T(F)}{\varepsilon} - L_F(D)\right) = 0.$$

Def. (Hadamard differentiability)
A functional T is **Hadamard differentiable** at F if, for any sequence $\varepsilon_n \to 0$ and D_n satisfying $\sup_y |D_n(y) - D(y)| \to 0$, we have

$$\lim_{\varepsilon_n \to 0}\left(\frac{T(F + \varepsilon_n D_n) - T(F)}{\varepsilon_n} - L_F(D_n)\right) = 0.$$

Prop. 4 (Properties)
The following properties hold:

- *if T is Hadamard differentiable, then $T(\widehat{F}_n) \xrightarrow{P} T(F)$;*
- *if T is Hadamard differentiable at F, then*

$$\sqrt{n}\frac{T(\widehat{F}_n) - T(F)}{\tau} \xrightarrow{d} \mathcal{N}(0,1),$$

  where $\tau^2 = \int L_F(y)^2 \, \mathrm{d}F(y)$;
- *also,*

$$\sqrt{n}\frac{T(\widehat{F}_n) - T(F)}{\widehat{\tau}} \xrightarrow{d} \mathcal{N}(0,1),$$

  where $\widehat{\tau}^2 = n^{-1}\sum_{i=1}^n \widehat{L}_n^2(Y_i)$.

Thus, under appropriate regularity conditions, a $1-\alpha$ **confidence interval** for $\vartheta = T(F)$ is

$$T(\widehat{F}_n) \pm z_{\alpha/2} n^{-1/2} \widehat{\tau}.$$

Example (Sample mean)
Using $\widehat{\vartheta} = T(\widehat{F}_n) = \bar{y}$, then

$$\widehat{L}_n(y_i) = y_i - \bar{y},$$

from which we obtain

$$\widehat{\tau}_n^2 = n^{-1}\sum_{i=1}^n \widehat{L}_n^2(y_i) = n^{-1}\sum_{i=1}^n (y_i - \bar{y})^2,$$

and the asymptotic confidence interval for ϑ is

$$y \pm z_{1-\alpha/2} n^{-1/2} \widehat{\tau}_n.$$

This is the usual confidence interval when the variance estimator is the biased version.

4.4.1 Score function and influence function

Let $\ell(\vartheta|y)$ be the log-likelihood for $\vartheta \in \mathbb{R}$ and $U_\vartheta(y) = \frac{\partial}{\partial\vartheta}\ell(\vartheta|y)$ be the score function, then we have for the maximum likelihood estimator $\widehat{\vartheta}$

$$\mathbb{E}_\vartheta[U_\vartheta(Y)] = 0$$

$$\mathbb{V}_\vartheta[\widehat{\vartheta}] \approx \frac{1}{n\mathbb{V}_\vartheta[U_\vartheta(Y)]} = \frac{1}{n\mathbb{E}_\vartheta[U_\vartheta^2(Y)]}$$

Whereas for the influence function, if we define $\vartheta = T(F)$ then

$$\mathbb{E}_\vartheta[L_F(Y)] = 0$$

$$\mathbb{V}_\vartheta[\widehat{\vartheta}] \approx \frac{\mathbb{V}[L_F(Y)]}{n} = \frac{\mathbb{E}[L_F^2(Y)]}{n}$$

where the approximation is exact for a linear functional T. This relationship is satisfied by deriving he influence function of a parametric model,

$$L_\vartheta(y) = I(\vartheta)^{-1} U_\vartheta(y),$$

where $I(\vartheta)$ is the Fisher information.

4.4.2 Misspecified models

Note that $\mathbb{E}_\vartheta[U_\vartheta^2(Y)] = I(\vartheta)$ is correct only if the parametric model is not misspecified. In general, we can estimate $\mathbb{V}_\vartheta[\widehat{\vartheta}]$ using the nonparametric delta method,

$$\widehat{\mathbb{V}}[\widehat{\vartheta}] = \frac{1}{n}\sum_{i=1}^n \widehat{L}(y_i)^2,$$

and our estimate would be

$$\frac{\frac{1}{n}\sum_{i=1}^n \widehat{L}(y_i)^2}{n} = \frac{\frac{1}{n}\sum_{i=1}^n U_{\widehat{\vartheta}}(y_i)^2}{nI(\vartheta)^2},$$

to get the so-called **sandwich estimator**,

$$\mathbb{V}[\widehat{\vartheta}] = \frac{1}{n} I(\widehat{\vartheta})^{-1} \left[\frac{1}{n}\sum_{i=1}^n U_{\widehat{\vartheta}}(y_i)^2\right] I(\widehat{\vartheta})^{-1}.$$

Lecture 5: Simulation-based inference

2022-03-11

We described influence functions as a tool for assessing the standard error of statistical functionals and obtaining nonparametric confidence intervals. Today, we will introduce other ideas for estimating standard errors in a nonparametric way, as well as estimates of the bias.

5.1 Jackknife

We will see that the jackknife is built on essentially the same idea as the influence function, although the jackknife was proposed much earlier (1949).
Suppose we have an estimator T_n of $\vartheta = T(F)$ which can be computed from a sample $(Y_1, \dots, Y_n)$.

Def. (Jackknife estimator)
The **Jackknife estimator** of T is

$$\overline{T}_n = \frac{1}{n} \sum_{i=1}^{n} T_{(-i)},$$

where $T_{(i)}$ is T computed on $(Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n)$.

Remark. If T_n is unbiased, then

$$\mathbb{E}[\overline{T}_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[T_{(-i)}] = 0.$$

On the other hand, if T_n is asymptotically unbiased so that $\lim_{n\to\infty} \mathbb{E}[T_n] = \vartheta$ with

$$\mathbb{E}[T_n] = \vartheta + \frac{a}{n} + \frac{b}{n^2} + O(n^{-3}), \tag{4}$$

then

$$\mathbb{E}[T_{(-i)}] = \vartheta + \frac{a}{n-1} + \frac{b}{(n-1)^2} + O(n^{-3}),$$

and we can estimate the bias using a quantity defined as

$$b_{\text{jack}} = (n-1)(\overline{T}_n - T_n).$$

We have that

$$\mathbb{E}[b_{\text{jack}}] = \frac{a}{n} + \frac{(2n-1)b}{n^2(n-1)} + O(n^{-2}).$$

Def. (Bias-corrected jacknife estimator)
The **bias-corrected jacknife estimator** of T is defined as

$$T_{\text{jack}} = T_n - b_{\text{jack}}. \tag{5}$$

Remark. This is an unbiased estimate of ϑ up to *second order*, which is an improved from the first-order of (4),

$$\text{Bias}(T_{\text{jack}}) = -\frac{b}{n(n-1)} + O(n^{-2}).$$

Example (Estimator of variance)
We can do even more: consider the plug-in estimate of the variance ϑ, one possibility would be to use

$$T(F) = \vartheta = \int y^2 dF(y) - \left(\int y dF(y)\right)^2,$$

which has sample analogue

$$T(\widehat{F}_n) = n^{-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

and the expected value of $T(F_n)$

$$\mathbb{E}[T(\widehat{F}_n)] = \frac{n-1}{n}\vartheta.$$

Hence, the expected value of the jackknife estimate of bias is

$$\mathbb{E}[b_{\text{jack}}] = -\frac{\vartheta}{n} = \text{Bias}\big(T(\widehat{F}_n)\big),$$

and the bias-corrected estimate is

$$T_{\text{jack}} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

Another way to think about the jackknife is in terms of the **pseudo-values**,

$$\tilde{T}_i = nT_n - (n-1)T_{(-i)},$$

then we have that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\tilde{T}_i &= nT_n - (n-1)\overline{T}_n \\ &= T_n - b_{\text{jack}} \\ &= T_{\text{jack}}.\end{aligned}$$

The idea behind pseudo-values is that it allows us to think of the bias-corrected estimate as simply the mean of n "independent" data values. This allows us to study properties of T_{jack} in terms of central limit theorems, although we need some care since these random variables are not independent.

Remark. The pseudo-values $\tilde{T}_i$ are not in general independent, although note that for the special case of a linear statistic we have that

$$T_n = \frac{1}{n}\sum_{i=1}^{n}a(Y_i) \implies \tilde{T}_i = a(Y_i) \implies \tilde{T}_j \perp\!\!\!\perp T_i, \quad i \neq j.$$

A reasonable idea, therefore, is to treat $\tilde{T}_i$ as linear approximations to i.i.d observations and approach inference for T_{jack} as we would the sample mean. Thus, we can calculate the sample variance of the pseudo-values,

$$\tilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{T}_i - T_{\text{jack}})^2 \implies \widehat{\mathbb{V}}[T_n] = v_{\text{jack}} = \frac{\tilde{s}^2}{n}.$$

Def. (Jackknife variance estimator)
The **jackknife variance estimator** is defined as

$$\tilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{T}_i - T_{\text{jack}})^2 \implies \widehat{\mathbb{V}}[T_n] = v_{\text{jack}} = \frac{\tilde{s}^2}{n}. \tag{6}$$

Hence, we have

- unbiased estimate up to second order;
- an estimate of the variance of the estimator;
- confidence intervals which are similar to those obtained by the functional delta method (later), using

  $$T_{\text{jack}} \pm t_{1-\frac{\alpha}{2},n-1}\sqrt{v_{\text{jack}}}.$$

  The above CI is approximately correct since observations are neither Gaussian nor independent.

Consistency. Until now there are no distributional assumptions about Y_i, but we must investigate the conditions under which v_{jack} is a good estimator of $\mathbb{V}[T_n]$.
In general, if g is a continuously differentiable function, then v_{jack} is a consistent estimator of $g(\overline{Y})$,

$$\frac{v_{\text{jack}}}{\mathbb{V}[g(\overline{Y})]} \xrightarrow{P} 1.$$

In particular, v_{jack} can be shown to perform poorly when the estimator is not a smooth function of the data, for example the *median*. It can be shown (Efron, 1982) that the jackknife variance estimate is inconsistent for all F and all quantiles of order p, and for the median, $p = 1/2$ and

$$\frac{v_{\text{jack}}}{\mathbb{V}[T_n]} \xrightarrow{d} \left(\frac{1}{2}\chi^2\right)^2.$$

Influence function. There is a close connection between the jackknife and influence functions. In particular, the jackknife can be seen as a plug-in estimator, which calculates n estimates based on a perturbed version of the empirical distribution function and compares these altered estimates to the plug-in estimate in order to assess variability of the estimate.

Removing a single value, we obtain $T_{(-i)} = T\big(F_n^{(-i)}(y)\big)$, where

$$\begin{aligned} F_n^{(-i)} = \frac{1}{n-1}\sum_{j\neq i}^{n}\mathbb{1}_{(-\infty,y]}(Y_j) &= \frac{n}{n-1}\widehat{F}_n(y) - \frac{1}{n-1}\mathbb{1}_{(-\infty,y]}(Y_i) \\ &= \frac{n}{n-1}\widehat{F}_n(y) - \frac{1}{n-1}G_{Y_i}(y). \end{aligned} \tag{7}$$

And here, $G_{Y_i}(y)$ is the cumulative distribution function that puts unitary mass at $y = Y_i$. In fact, we can use this quantity to approximate the influence function, by setting $F = \widehat{F}_n$ and $\varepsilon = -\frac{1}{n-1}$,

$$L_F(Y_i) \approx \frac{\overbrace{T\left(\frac{n}{n-1}\widehat{F}_n - \frac{1}{n-1}G_{Y_i}\right)}^{T_{(-i)}} - \overbrace{T(\widehat{F}_n)}^{T_n}}{-1/(n-1)} = (n-1)(T_n - T_{(-i)}).$$

In a sense, then, the jackknife is a numerical approximation to the functional delta, and this justifies the alternative name of **infinitesimal jackknife** for the functional delta method.

Differences. There is an important difference, however, between the jackknife and the functional delta method: the delta method adds point mass to Y_i, while the jackknife in (7) takes point mass away.

Positive jackknife. Another take on the jackknife, then, is to compute n estimates $T(i)$ by adding an observation at Y_i instead of taking one away (i.e., $\varepsilon = 1/(n+1)$)). This method is called the **positive jackknife**, which however is not commonly used.

Delete-*d* jackknife. Another variation on the jackknife that has been proposed is called the **delete-*d* jackknife**, which leaves out d observations for each estimate. This can be an improvement for reducing dependence between the Jackknife estimates, and if d is appropriately chosen then the estimate of the variance is consistent for the median. However, it has the drawback that instead of calculating n leave-one-out estimates, we now have to calculate $\binom{n}{d}$ leave-d-out estimates - a much larger number, often bordering on the computationally infeasible

5.2 Bootstrap

The bootstrap was introduced by Efron (1979) as a computer-based method to estimate the variance and the distribution of an estimate and more generally of a statistic $T_n = T(Y_1, Y_2, \ldots, Y_n)$. In general, we can use it to construct confidence intervals.

The advantages of the bootstrap are multiple:

- Completely automatic.
- Requires no theoretical calculations.
- Not based on asymptotic results.
- Available no matter how complicated the statistic is.

Suppose that we want to calculate the variance of an estimator,

$$\mathbb{V}_F[T_n] = \int T_n^2 dF(y_1)\cdots dF(y_n) - \left(\int T_n dF(y_1)\cdots dF(y_n)\right)^2,$$

and the **ideal bootstrap** estimates $\mathbb{V}_F[T_n]$ with $\mathbb{V}_{\widehat{F}_n}[T_n]$ using a plug-in estimator of the variance.

Problem. A possible drawback is that $\mathbb{V}_{\widehat{F}_n}[T_n]$ might be difficult to compute. Indeed, the above integrals have to be performed over $\mathbb{R}^n$ and, by plugin, the sums become $\underbrace{\sum_{i=1}^{n} \cdots \sum_{i=1}^{n}}_{n \text{ times}}$, hence it is computationally $O(n^n)$.

Solution. The idea is to approximate it using a simulation, i.e. by using a subset of size B of the n^n terms of the summation.

Algorithm 1 Bootstrap

1: Sample $\boldsymbol{Y}_1^*, \ldots, \boldsymbol{Y}_B^*$ **with replacement** from $\widehat{F}_n$, $Y_{b,i}^* \overset{\text{iid}}{\sim} \widehat{F}_n, \quad b = 1, \ldots, B$.
2: Calculate the bootstrap replication $T^* = g(\boldsymbol{Y}_b^*)$ for $b = 1, \ldots, B$
3: Estimate the variance using

$$\mathbb{V}[T_n] = \frac{1}{B-1} \sum_{b=1}^{B} (T_b^* - \overline{T}^*)^2,$$

with $\overline{T}^* = B^{-1} \sum_{b=1}^{B} T_b^*$.

By the law of large numbers, then

$$\mathbb{V}_{\text{boot}}[T_n] \xrightarrow{B \to \infty} \mathbb{V}_{\widehat{F}_n}[T_n].$$

We can also use it to estimate the bias, or any aspect of T_n. Indeed, we can write the bias of ϑ as

$$\text{bias}_{\text{boot}} = \overline{\vartheta}^* - \widehat{\vartheta} = \frac{1}{B} \sum_{B=1}^{B} \vartheta_b^* - \widehat{\vartheta},$$

or by setting $G = \mathbb{P}(T_n \leq t)$, then the bootstrap approximation to G is

$$\widehat{G}_n^*(t) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{(-\infty, t]}(T_b^*)$$

Theorem 7 (Consistent estimator of G)
If $T_n = T(F)$ is Hadamard differentiable, then $\widehat{G}_n$ is a consistent estimator of G.

Proof.
Wasserman (2005)

□

Example (Failure of the bootstrap)
Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from F and that $\mathbb{E}[Y] = \mu$, $\mathbb{V}[Y] = 1$. Consider

$\vartheta = |\int y dF(y)|$ with plug-in estimator $\widehat{\vartheta}_n = |\overline{Y}|$.

If $\vartheta = 0$, then the bootstrap is not consistent for estimating the distribution of

$$E_n = \sqrt{n}(|\overline{Y}_n| - |\mu|) \xrightarrow[\mu=0]{d} |Z|, \quad Z \sim \mathcal{N}(0,1)$$

It can be shown however that

$$\left(\sqrt{n}(\overline{Y}_n - \mu), \sqrt{n}(\overline{Y}_n^* - \overline{Y})\right) \xrightarrow{d} (Z_1, Z_2),$$

where Z_1 and Z_2 are independent $\mathcal{N}(0,1)$ random variables. In practice, we have that

$$\begin{aligned} E_n^* &= \sqrt{n}(\overline{Y}^* - \overline{Y}) \\ &= |\sqrt{n}(\overline{Y}_n^* - \overline{Y}) + \sqrt{n}\overline{Y}_n| - \sqrt{n}\overline{Y}_n \\ &\xrightarrow{d} |Z_1 + Z_2| - |Z_1|. \end{aligned}$$

Hence, E_n^* does not converge to the absolute value of a $N(0,1)$ random variable.

5.2.1 Confidence intervals

There are several ways of constructing confidence intervals, and the bootstrap allows us to do so for a general statistic.

Def. (Confidence interval)
A **confidence interval** of level $1-\alpha$ is a random interval which contains ϑ_0 with probability $1-\alpha$, i.e.

$$\mathbb{P}(\vartheta \in C) = 1 - \alpha.$$

Def. (Pivotal interval)
Let $\vartheta = T(F)$, $\widehat{\vartheta}_n = T(\widehat{F}_n)$, $R_n = \widehat{\vartheta}_n - \vartheta$ be the **pivot** and $H(r) = \mathbb{P}(R_n \leq r)$. Then, the **pivotal interval** is defined as the interval such that

$$\mathbb{P}(\widehat{\vartheta}_n - H^{-1}(1 - \alpha/2) \leq \vartheta \leq \widehat{\vartheta}_n - H^{-1}(\alpha/2)) = 1 - \alpha.$$

Remark. Since H is unknown, the bootstrap estimate of H is

$$\widehat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{(-\infty, r]}(R_b^*), \quad \text{where } R_b^* = \widehat{\vartheta}_b^* - \widehat{\vartheta}_n,$$

from which we obtain

$$\begin{aligned} 1 - \alpha &= \mathbb{P}(\widehat{\vartheta}_n - H^{-1}(1 - \alpha/2) \leq \vartheta \leq \widehat{\vartheta}_n - H^{-1}(\alpha/2)) \\ &\overset{\text{boot}}{=} \dots \end{aligned}$$

Theorem 8 (Pivot interval)
If $T(F)$ is Hadamard differentiable, then

$$C_n = \big(2\widehat{\vartheta}_n - \vartheta^*_{1-\frac{\alpha}{2}}, 2\widehat{\vartheta}_n - \vartheta^*_{\frac{\alpha}{2}}\big)$$

is such that

$$\mathbb{P}\big(T(F) \in C_n\big) \xrightarrow{n\to\infty} 1-\alpha.$$

Def. (Studentized pivotal interval)
Let $Z_n = (\widehat{\vartheta}_n - \vartheta)/\widehat{\text{se}}_{\text{boot}}$, and

$$Z^*_b = \frac{\widehat{\vartheta}^*_b - \widehat{\vartheta}_n}{\widehat{\text{se}}^*_b},$$

where $\widehat{\text{se}}^*_b$ is an estimate of the standard error of $\widehat{\vartheta}^*_b$. The **studentized pivotal interval** is defined as

$$C_n = \big(\widehat{\vartheta}_n - z^*_{1-\frac{\alpha}{2}}\widehat{\text{se}}_{\text{boot}}, \widehat{\vartheta}_n + z^*_{1-\frac{\alpha}{2}}\widehat{\text{se}}_{\text{boot}}\big).$$

Remark. In order to compute $\widehat{\text{se}}^*_{\text{boot}}$ we need to nest the bootstraps.

Def. (Percentile interval)
The **percentile interval** is defined as the interval

$$C = \big(G^{-1}(\alpha/2), G^{-1}(1-\alpha/2)\big),$$

which is estimated using the α quantiles of ϑ^*_b,

$$C_n = \big(t^*_{\frac{\alpha}{2}}, t^*_{1-\frac{\alpha}{2}}\big).$$

Suppose that we have a one-sided interval of the form $[\widehat{\vartheta}_\alpha, \infty)$, then we would like our confidence interval to be such

$$\mathbb{P}(\vartheta > \widehat{\vartheta}_\alpha) = 1-\alpha \iff \mathbb{P}(\vartheta \leq \widehat{\vartheta}_\alpha) = \alpha.$$

- If $\mathbb{P}(\vartheta \leq \widehat{\vartheta}_\alpha) = \alpha + O(n^{-1/2})$, then the interval is **first-order accurate**.
- If $\mathbb{P}(\vartheta \leq \widehat{\vartheta}_\alpha) = \alpha + O(n^{-1})$, then the interval is **second-order accurate**.

Prop. 5 (Accuracy)
For the approximated confidence intervals, we have that

- *Normal interval:* $\mathbb{P}(\vartheta \leq \widehat{\vartheta}_\alpha) = \alpha + O(n^{-1/2})$
- *Pivotal interval:* $\mathbb{P}(\vartheta \leq \widehat{\vartheta}_\alpha) = \alpha + O(n^{-1/2})$
- *Studentized interval:* $\mathbb{P}(\vartheta \leq \widehat{\vartheta}_\alpha) = \alpha + O(n^{-1})$
- *Percentile interval:* $\mathbb{P}(\vartheta \leq \widehat{\vartheta}_\alpha) = \alpha + O(n^{-1/2})$

Remark. The percentile interval has more justifications in terms of invariance with respect to reparametrization, hence a popular extension is the **bias-corrected and accelerated percentile interval**.

5.3 Geometry of the bootstrap

In this last part of the lecture, we will present a unifying framework between the jackknife, the bootstrap, and the Delta method that we discussed so far.

Let $\boldsymbol{w}^* = (w_1^*, \ldots, w_n^*)$ denote a vector of weights such that $w_i^* \in [0,1]$ for all i and $\sum_{i=1}^n w_i^* = 1$. Let now $\widehat{F}(\boldsymbol{w}^*)$ denote the cumulative distribution function that places point mass w_i^* at Y_i, and define

$$\widehat{\vartheta}^* = T\big(\widehat{F}(\boldsymbol{w}^*)\big). \tag{8}$$

The statistic represented in (8) is a function defined on the n-dimensional simplex

$$\Delta = \{(w_1, \ldots, w_n) : \sum_{i=1}^n w_i = 1,\ 0 \leq w_i \leq 1\}.$$

a) The influence function defined in (1) studies the behavior of $\widehat{\vartheta}^*$ in the infinitesimal region around $\widehat{w}$.

b) The jackknife in (5) studies the behavior of $\widehat{\vartheta}^*$ as $\boldsymbol{w}$ is moved away from $\widehat{\boldsymbol{w}}$ by an amount equal to $1/n$ in the direction opposite to the i^{th} vertex.

c) The nonparametric bootstrap draws $\boldsymbol{w}^*$ from a multinomial distribution,

$$\boldsymbol{w}^* \sim \text{Multinomial}(n, \widehat{\boldsymbol{w}}),$$

in order to form a **weighted representation** of the distribution on the simplex.

By specifying a different mechanism for obtaining the weights $\boldsymbol{w}^*$, we can obtain other approximations to the distribution of the statistic. For instance, the **Bayesian bootstrap** places a prior distribution $\text{Dir}_n(\boldsymbol{\alpha})$ onto the n-dimensional simplex Δ and draws the weights $\boldsymbol{w}^*$ through the posterior distribution

$$\boldsymbol{w}|y_1, \ldots, y_n \sim \text{Dir}_n(\boldsymbol{\alpha} + \mathbf{1}_n).$$

As $\boldsymbol{\alpha} \to \mathbf{0}_n$, we obtain the standard nonparametric bootstrap, interpreted by applying a "noninformative" prior distribution on the observed data points,

$$\text{Dir}_n(\boldsymbol{\alpha} + \mathbf{1}_n) \xrightarrow{\alpha \to 0} \frac{1}{n}\text{Mult}(n, \widehat{\boldsymbol{w}}).$$

Lecture 6: Nonparametric density estimation

2022-03-16

Suppose we observe $Y_1, Y_2, \ldots, Y_n$ where F has density $f = dF/dy$ with respect to Lebesgue measure on the real line. Some observations about the estimation of f is that, although $\widehat{F}_n(y)$ is often a good estimator of F, $d\widehat{F}_n(y)/dy$ is usually not a good estimator of f. The estimator $d\widehat{F}(n)/dy$ can be represented simply as the empirical histogram using a set of contiguous intervals B_k, $k = 1, \ldots, K$.

6.1 Kernel density estimator

Def. (Histogram)
Formally, we can define the **histogram** as the estimator

$$\widehat{f}_n(y) = \sum_{k=1}^{K} \frac{1}{w_k} \mathbb{1}_{B_k}(y)\widehat{p}_k,$$

where $w_k = \operatorname{diam}(B_k)$ and $\widehat{p}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{B_k}(Y_i)$.

Remark. The estimator is not continuous and depends heavily on the choice of bins. In general, it is a rudimentary estimator which gives basic information about the true population.

In general, we can define an estimator of f as the limit for $\operatorname{diam}(B_k) \to 0$.

Def. (Naive estimator)
Given a sample of n observations $Y_1, Y_2, \ldots, Y_n$, the **naive estimator** for f is

$$\widehat{f}_n(y) = \frac{\widehat{F}_n(y+h) - \widehat{F}_n(y-h)}{2h}$$

Remark. The estimator is simply the plug-in estimator using the definition of a differentiable F,

$$f(y) = \frac{\partial}{\partial y} F(y) = \lim_{h \to 0} \frac{F(y+h) - F(y-h)}{2h}$$

Remark. Starting from what we defined before, we can rewrite it as

$$\begin{aligned}
\widehat{f}_n(y) &= \frac{\sum_{i=1}^{n} \mathbb{1}_{(-\infty, y+h}(Y_i) - \sum_{i=1}^{n} \mathbb{1}_{(-\infty, y-h)}(Y_i)}{2nh} \\
&= \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} \mathbb{1}_{(-1,1]} \left(\frac{y - Y_i}{h} \right) \\
&= \frac{1}{nh} \sum_{i=1}^{n} K \left(\frac{y - Y_i}{h} \right),
\end{aligned}$$

where $K(u) = \frac{\mathbb{1}_{(-1,1]}(u)}{2}$ is the density function of $U \sim \operatorname{Unif}(-1, 1)$. Hence, the naive estimator is the average of density functions scaled by some width h.

Properties. In general, we can state the following properties about the naive estimator:

- $\widehat{f}$ is more flexible than the simple histogram.
- $\widehat{f}$ depends heavily on the value of the **bandwidth** h, although is still rough for large h.
- It is a density for any value of h.

The idea of the general kernel density estimator is to place a small density around each observation.

Def. (Kernel density estimator)
The **kernel density estimator** of f is

$$\widehat{f}_n(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right), \tag{9}$$

where K is a symmetric density centered in zero.

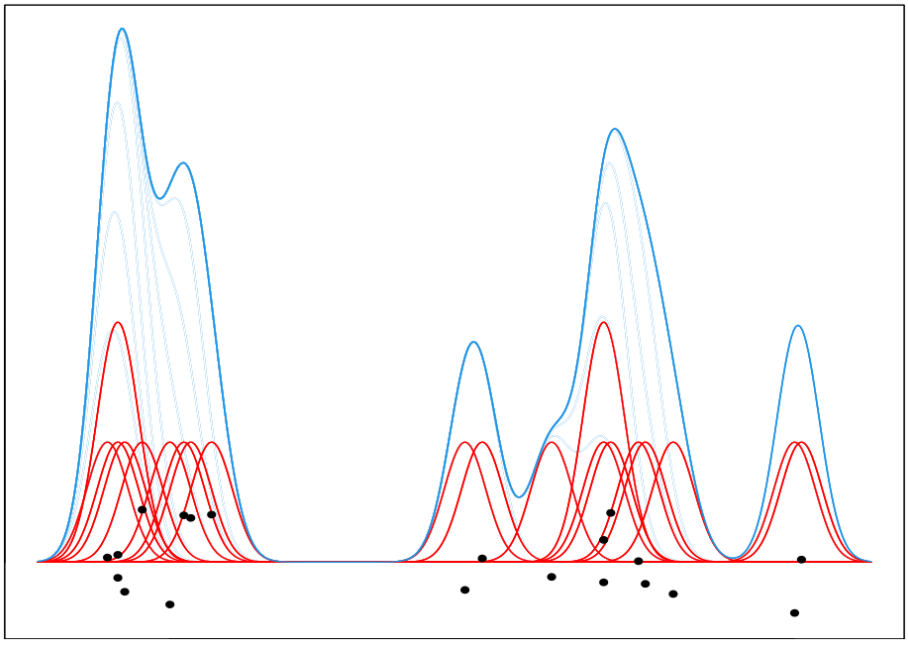


Figure 3: General idea for the kernel density estimator of the density of f.

Kernel	$K(u)$
Uniform (Rectangular)	$\frac{1}{2} I_{[-1,1]}(\lvert u\rvert)$
Triangle	$(1-\lvert u\rvert) I_{[-1,1]}(\lvert u\rvert)$
Triweight	$\frac{35}{32}(1-u^2)^3 I_{[-1,1]}(\lvert u\rvert)$
Quartic (Biweight)	$\frac{15}{16}(1-u^2)^2 I_{[-1,1]}(\lvert u\rvert)$
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-u^2/2}$
Epanechnikov	$\frac{3}{4}(1-u^2) I_{[-1,1]}(\lvert u\rvert)$
Cosine	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right) I_{[-1,1]}(\lvert u\rvert)$

Figure 4: Most common kernels in kernel density estimates for f, see `density` function in `R`. Note that most kernels are bounded density, whereas the Gaussian is unbounded.

Remark. The shape of the kernel doesn't really affect the asymptotic properties of $\widehat{f}$. In smoothing in general there is a fundamental trade-off between the bias and variance of the estimate $\widehat{f}_n$, and this trade-off is governed by the smoothing parameter h.

Def. (Mean-squared error)
We define the **mean-squared error** of an estimator $\widehat{f}_n$ as the risk

$$\mathrm{MSE}(\widehat{f}_n) = \mathbb{E}[L(\widehat{f}_n, f(y))] = \mathbb{E}\left[\left(\widehat{f}_n(y) - f(y)\right)^2\right].$$

Bias-variance. In general, we can highlight the mean-squared error as a combination of bias and variance of the estimator, since

$$\begin{aligned}\mathrm{MSE}(\widehat{f}_n) &= \Big(\mathbb{E}[\widehat{f}_n(Y)] - f(Y)\Big)^2 + \mathbb{V}[\widehat{f}_n(Y)] \\ &= \mathrm{Bias}_Y^2\big(\widehat{f}_n(Y)\big) + \mathbb{V}_Y[\widehat{f}_n(Y)].\end{aligned}$$

Most of the time we instead consider averaged versions of the mean-squared error.

Def. (Integrated mean-squared error)
The **mean-integrated -squared error** (MISE) is

$$\mathrm{MISE}(\widehat{f}_n, f) = \int R(\widehat{f}_n(y), f(y))dy. \tag{10}$$

Def. (Average mean-squared error)
The **average mean-squared error** (AMSE) is

$$\mathrm{AMSE}(\widehat{f}_n, f) = \frac{1}{n}\sum_{i=1}^{n} R(\widehat{f}_n(y_i), f(y_i)). \tag{11}$$

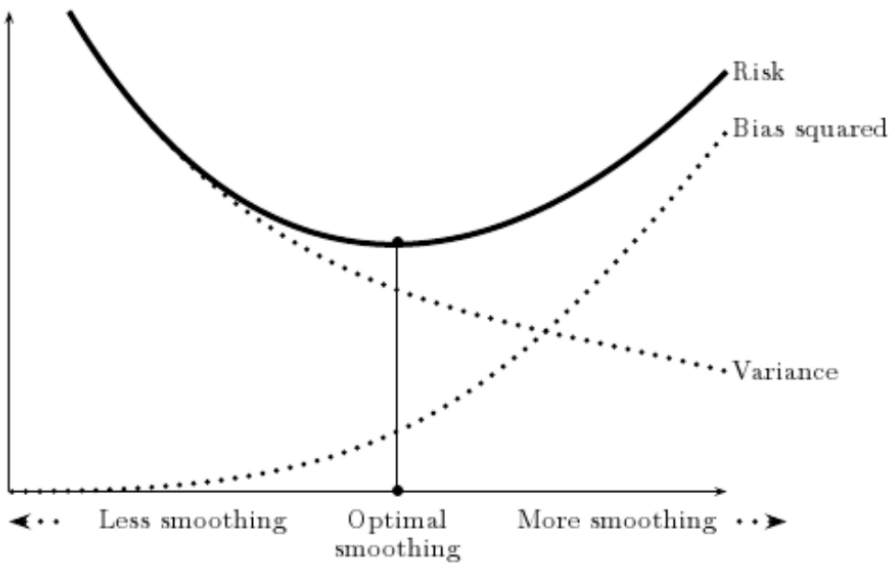


Figure 5: Example of bias-variance trade-off in kernel density estimation.

We could use other loss functions apart from the L_2 loss, such as:

› $\boldsymbol{L^p}$ **loss**, using

$$\left\{\int |\widehat{f}_n(y) - f(y)|^p dy\right\}^{1/p},$$

for which L_2 yields results that are easier to achieve.

› **Kullback-Leibler loss**, used especially in the machine learning community,

$$L(\widehat{f}_n, f) = \int f(y) \log\left(\frac{f(y)}{\widehat{f}_n(y)}\right) dy,$$

which is sensitive to the tails of the distribution (Hall, 1987).

6.1.1 Bias of the estimator

We have that the estimator in (9) has bias given by the following expression,

$$\begin{aligned}
\mathbb{E}[\widehat{f}_n(y)] &= \frac{1}{nh}\sum_{i=1}^{n} \mathbb{E}\left[K\Big(\frac{y - Y_i}{h}\Big)\right] \\
&= \mathbb{E}\left[\frac{1}{h}K\Big(\frac{y - Y}{h}\Big)\right] \\
&= \int K_h(y-u) f(u) du \\
&= K_h * f(y),
\end{aligned}$$

which is the **convolution** between $K_h(u) = h^{-1}K(u/h)$ and f. In particular, we have that

$$\text{Bias}\big(\widehat{f}_n(y)\big) = \mathbb{E}[\widehat{f}_n(y) - f(y)] = K_h * f(y) - f(y).$$

6.1.2 Variance of the estimator

As for the variance, we have that in the case the data is i.i.d,

$$\begin{aligned}\mathbb{V}[\widehat{f}_n(y)] &= \frac{1}{n}\,\mathbb{V}\left[\frac{1}{h}K\Big(\frac{y-Y_i}{h}\Big)\right] \\ &= \frac{1}{n}\Big(\mathbb{E}[K_h(h-Y_i)^2] - \mathbb{E}[K_h(h-Y_i) - f(y)]^2\Big) \\ &= \frac{1}{n}\Big(K_h^2 * f(y) - [K_h * f(y)]^2\Big)\end{aligned}$$

6.1.3 Mean-squared error of the estimator

In general, we can conclude that

$$\begin{aligned}\mathrm{MSE}(\widehat{f}_n(y)) &= \mathrm{Bias}\big(\widehat{f}_n(y)\big)^2 + \mathbb{V}[\widehat{f}_n(y)] \\ &= (K_h * f(y) - f(y))^2 + \frac{1}{n}\Big(K_h^2 * f(y) - [K_h * f(y)]^2\Big)\end{aligned}$$

hence the bias for a fixed $h > 0$ does not tend to zero as as n increases. The only way to have a vanishing bias is to work on the bandwidth h of the kernel, for instance by choosing the kernel K_h such that

$$K_h * f(y) \xrightarrow{n\to\infty} f(y),$$

and this can be done if n, h are chosen so that they jointly satisfy

$$\begin{aligned}h = h(n) &\xrightarrow{n\to\infty} 0 \\ n \cdot h &\xrightarrow{n\to\infty} \infty.\end{aligned}$$

We will consider the mean integrated squared error (10) to study the asymptotic properties of the kernel density estimator.

Suppose that we are under the following assumptions for the unknown function f and the kernel K:

1. **Conditions on f**:
   - f is three-times differentiable with $|f^{(j)}| \leq C$ for $j = 0, \ldots, 3$.
2. **Conditions on K**:
   - K is positive and symmetric.
   - $\int K(u)du = 1$ and $\int uK(u)du$
   - $\int u^2K(u)du < \infty$
   - $\int |u|^3K(u)du < \infty$
   - $\int K^2(u)du < \infty$
3. **Conditions on h**
   - $h \xrightarrow{n\to\infty} 0$ and $nh \xrightarrow{n\to\infty} \infty$.

Bias. By using a Taylor expansion of order 2 of f around y, we find that

$$\begin{aligned}\mathbb{E}[\widehat{f}_n(y)] &= \int K(y)[f(y) - huf'(y) + \frac{(hu)^2}{2}f''(y) + o\big((hu)^2\big)]du \\ &= f(y)\underbrace{\int K(u)du}_{=1} - hf'(y)\underbrace{\int uK(u)du}_{=0} + \frac{h^2}{2}f''(y)\underbrace{\int u^2K(u)du}_{=1} + o(h^2) \\ &= f(y) + \frac{h^2}{2}f''(y)\mu_{K,2} + o(h^2),\end{aligned}$$

and it is clear that as $h \to 0$ the asymptotic bias disappears. In general, the asymptotic bias depends on $f''(y)$, hence it is more pronounced in the peaks and troughs of f.

Variance. By applying a Taylor expansion of order 1 for $\mathbb{E}[K_h^2(y - Y_i)]$ we can write write

$$\begin{aligned}\mathbb{E}[K_h^2(y - Y_i)] &= h^{-1}\int K^2(u)f(y - hu)du \\ &= k^{-1}\int K^2(y)[f(y) - huf'(y) + o(h, u)]du \\ &= h^{-1}\dots \\ &= \frac{1}{h}f(y)\int K^2(u)du + o(1),\end{aligned}$$

hence by combining the above with $\mathbb{E}[\widehat{f}_n(y)]^2$ we obtain

$$\begin{aligned}\mathbb{V}[\widehat{f}_n(y)] &= \frac{1}{nh}f(y)\int K^2(u)du + O(n^{-1}) \\ &= \frac{1}{nh}f(y)R(K) + O(n^{-1})\end{aligned}$$

Remark. The variance increases depending on $R(K) = \int K^2(u)du$, and this is the only main contribution given by choice of the kernel. In general, we can find bandwidths $h_1, h_2, \dots$ such that different kernels $K_1, K_2, \dots$ yield practically equivalent result in terms of mean-squared error.

Combining the two results above yields

$$\mathrm{AMSE}\big(\widehat{f}_n(y), h\big) = \underbrace{\frac{h^4}{4}\mu_{K,2}^2 f''(y)^2}_{\text{Bias}^2} + \underbrace{\frac{R(K)}{nh}f(y)}_{\text{Variance}}, \tag{12}$$

and we observe that the estimator is consistent if both $h \xrightarrow{n\to\infty} 0$ and $nh \xrightarrow{n\to\infty} \infty$. If we integrate the mean-squared error we obtain

$$\begin{aligned}\mathrm{MISE}(\widehat{f}_n, h) &= \int \mathrm{MSE}(\widehat{f}_n(y))dy \\ &= \dots \\ &= \mathrm{AMSE}(\widehat{f}_n, h) + O(1/n) + o(h^4).\end{aligned}$$

6.1.4 Optimal global bandwidth

Although we have an expression for the asymptotic mean-squared error, we are interested in finding the optimal bandwidth h_{opt} for a finite value of n. From (12), we can focus on the leading terms to write

$$h_{\mathrm{opt}} \approx \operatorname*{argmin}_h \mathrm{AMISE}(\widehat{f}_n, h) = \left(\frac{R(K)}{\mu_{K,2}^2 R(f'')}\right)^{1/5} n^{-1/5},$$

although this depends in practice by f'' which is unknown. We can replace it by an estimate based on the sample using some empirical rules.

Normal reference rule Consider f to be a Gaussian density, then

$$R(f'') = \int f''(y)^2 dy = \frac{3}{8\sigma^5\sqrt{\pi}},$$

hence we can estimate $\widehat{\sigma} = \min\{s, r\}$, where s is the sample estimate of the standard error and

$$r = \frac{\widehat{F}_n^{-1}(0.75) - \widehat{F}_n^{-1}(0.25)}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)} \approx \sigma \quad \text{if } f = \mathcal{N}(\cdot|\mu, \sigma^2).$$

Plug-in bandwidth. Another reference rule could be obtain by using the **plug-in bandwidth**, which defines an estimator

$$\widehat{R}(f'') = \int \widehat{f}_n''(y)^2 dy,$$

where $\widehat{f}_n''$ is estimated on the observed data based on the chosen kernel itself.

Leave-one-out Another estimator is based on the **leave-one-out cross-validation** procedure. Starting from MISE, we have

$$\begin{aligned}\mathrm{MISE}(\widehat{f}_n, h) &= \int \mathrm{MSE}(y; h)dy \\ &= \mathbb{E}\int \widehat{f}_n(y)^2 dy - 2\,\mathbb{E}\int \widehat{f}_n f(y)dy + \int f^2(y)dy,\end{aligned}$$

which can be minimized only by considering the terms that depend on $\widehat{f}_n$,

$$\mathbb{E}\int \widehat{f}_n(y)^2 dy - 2\,\mathbb{E}\int \widehat{f}_n(y) f(y)dy,$$

and the first item can be estimated using the unbiased estimate $\int \widehat{f}_n(y)^2 dy$. The second term is more complicated, and can be estimated using the cross-validation procedure to obtain

$$\int \widehat{f}_n(y) f(y) dy \approx \frac{1}{n} \sum_{i=1}^{n} \widehat{f}_n^{(-i)}(Y_i),$$

where $\widehat{f}_n^{(-i)}(y)$ is the kernel density estimator with the i^{th} observation removed. We have that

$$\mathbb{E}[\widehat{f}_n^{(-i)}(Y_i)] = \mathbb{E}\Big[\int \widehat{f}_n(y) f(u) du\Big],$$

hence we can define the following criterion for choosing h.

Theorem 9 (Cross-validation criterion)

We have that

$$CV(\widehat{f}_n, h) = \int \widehat{f}_n(y)^2 dy - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_n^{(-i)}(Y_i)$$

is an unbiased estimator of

$$MISE(\widehat{f}_n, h) - \int f(u)^2 dy.$$

Def. (Data-driven bandwidth)

The **data-driven bandwidth** criterion for h is

$$\widehat{h}_{\text{opt}} = \underset{h}{\operatorname{argmin}}\, \text{CV}(\widehat{f}_n, h).$$

Remarks.

› Since the procedure is computationally intensive, we only use a grid of bandwidths.

› The quality of the resulting estimate is very variable.

› The optimal solution might not be unique.

6.1.5 Kernel density estimator with finite support

The results we have seen are only valid when the density f is continuous on its support. If instead f is the density of $Y \sim \text{Exp}(\vartheta)$, we have that f is supported on $[0, \infty)$ with finite right limit as $x \to 0$.

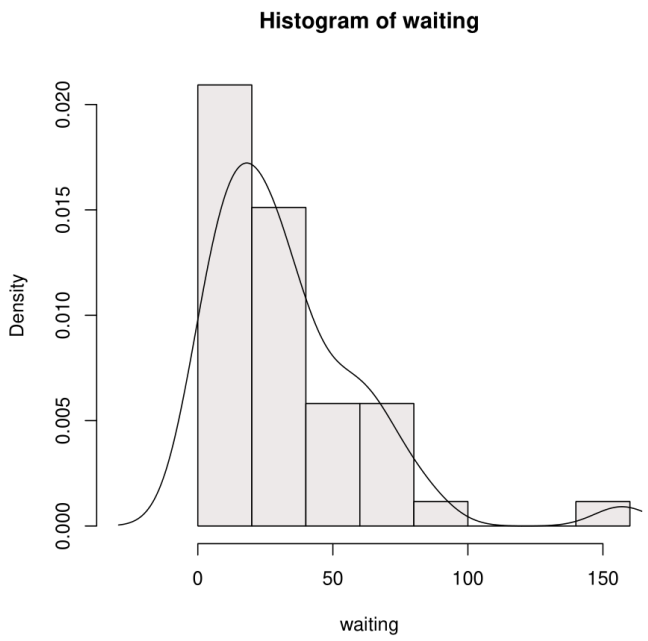


Figure 6: Example of a wrong kernel density estimator output when applied to positive data.

Remarks. Some naive solutions would be truncating $\widehat{f}_n$ for $y < 0$, but the density would not integrate to 1. Rescaling the truncation yields an insufficient estimator, as we will see in a moment.

Suppose that $\operatorname{supp} f = [0, \infty]$ and that f is two-times continuously differentiable. If K is symmetric with support $[-1, 1]$, then consider $y = ph$ with $p < 1$. For $p \geq 1$ we are in the interior and the usual properties apply, whereas if $p < 1$ we have that

$$\mathbb{E}[\widehat{f}_n(y)] = a_0(p)f(y) - ha_1(p)f'(y) + o(h),$$

so we are not able to remove the first part of the bias. The kernel density estimator is not consistent at the boundary because

$$\mathbb{E}[\widehat{f}_n(0)] \approx a_0(p)f(0) = f(0)\int_{-1}^{0} K(u)du = \frac{1}{2}f(0).$$

Suppose that we have an estimate $\widehat{a}_0(p)$, then if we consider a rescaled version we have that

$$\mathbb{E}\left[\frac{\widehat{f}(y)}{a_0(p)}\right] = 1.$$

Def. (Boundary correction via renormalization)
The **boundary correction** of the kernel density estimator is the **renormalized** version of $\widehat{f}_n$,

$$\tilde{f}_n(y) = \frac{\widehat{f}_n(y)}{a_0(p)},$$

and $a_0(p) = 1$ for $p \geq 1$, hence it is valid in the interior.

Remark. Calculating the bias and variance (slides), we find that $\tilde{f}_n$ is consistent but the bias is of order $O(h)$ near the boundary.

Remark. The optimal MSE is $O(n^{-2/3})$ at the boundary and $O(n^{-4/5})$ elsewhere.

Consider instead the augmented dataset

$$Y_1, -Y_1, Y_2, -Y_2, \ldots, Y_n, -Y_n,$$

then we could construct a consistent kernel density estimator for $f(y)$ based on the augmented dataset.

Def. (Boundary correction by reflection)
The **boundary correction by reflection** is given by

$$\widehat{f}_n^R(y) = \begin{cases} 2\tilde{f}(y) & \text{if } y \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{f}(y)$ is the simple kernel density estimator based on the augmented dataset, $Y_1, -Y_1, Y_2, -Y_2, \ldots, Y_n, -Y_n$.

Remark. The estimate corresponds to replacing the kernel with a modified kernel,

$$\tilde{K}_h(y - Y_i) = K_h(y - Y_i) + K_h(-y - Y_i) \implies \widehat{f}_n^R(y) = \widehat{f}_n(y) + \widehat{f}_n(-y).$$

Hence, we again have explicit formulas for the bias and variance (slides).

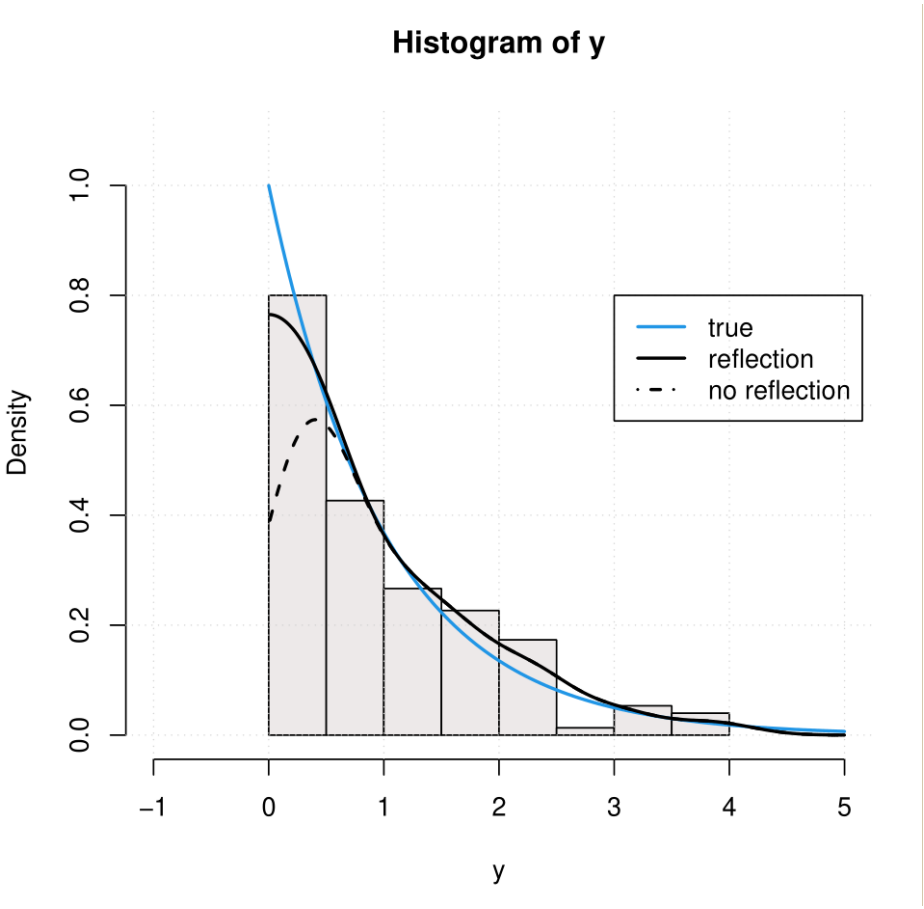


Figure 7: modifiedDensityEstimators

A third possibility is transforming the data and then moving them back to the original scale. Using the properties of transformation of the density, $Z = g(Y)$, we have that

$$f_Y(y) = f_Z\big(g(y)\big)g'(y).$$

Def. (Boundary correction by transformation)
The **boundary correction by transformation** uses the estimate

$$\widehat{f}_n(y) = \frac{g'(y)}{nh_Z}\sum_{i=1}^{n} K\left(\frac{g(y) - g(Y_i)}{h_Z}\right),$$

where h_Z is chosen based on the Z scale.

6.2 Local polynomials

Other methods of estimating the density can for example be based on local polynomials by using

$$\ell(f) = \sum_{i=1}^{n} \log f(y_i),$$

or by using the more general definition

$$\mathcal{L}(f) = \sum_{i=1}^{n} \log f(Y_i) - n\left(\int f(y)dy - 1\right),$$

where the second term is zero when f integrates to one. Including this term allows us to maximize over all non-negative functions f while imposing the constraint that f is a density.

Def. (Local likelihood)
The **local likelihood** is a weighted likelihood such that

$$\mathcal{L}_y(f) = \sum_{i=1}^{n} K\left(\frac{Y_i - y}{h}\right) \log f(Y_i) - n\left(\int K\left(\frac{u-y}{h}\right) f(u)du - 1\right).$$

Remark. Since f is unknown, we replace $\log f(u)$ by an approximation

$$\log f(u) = P_Y(\alpha, u; q) = \alpha_0 + \alpha_1(y-u) + \ldots + \frac{\alpha_q}{q!}(y-u)^q.$$

Note that

$$f(y) \in (0,1) \implies \log f(y) \in \mathbb{R}.$$

Def. (Local polynomial likelihood)
The **local polynomial likelihood** is defined as the approximated local likelihood using P_Y,

$$\mathcal{L}_y(f) = \sum_{i=1}^{n} K\left(\frac{Y_i - y}{h}\right) P_Y(\alpha, u; q) - n\left(\int K\left(\frac{u-y}{h}\right) \exp\left\{P_Y(\alpha, u; q)\right\} du - 1\right).$$

Let $\widehat{\alpha} = \text{argmax}_\alpha \, \mathcal{L}_Y(\alpha)$, then the local likelihood density estimate is

$$\widehat{f}_n(y) = e^{\widehat{\alpha}_0(y)}.$$

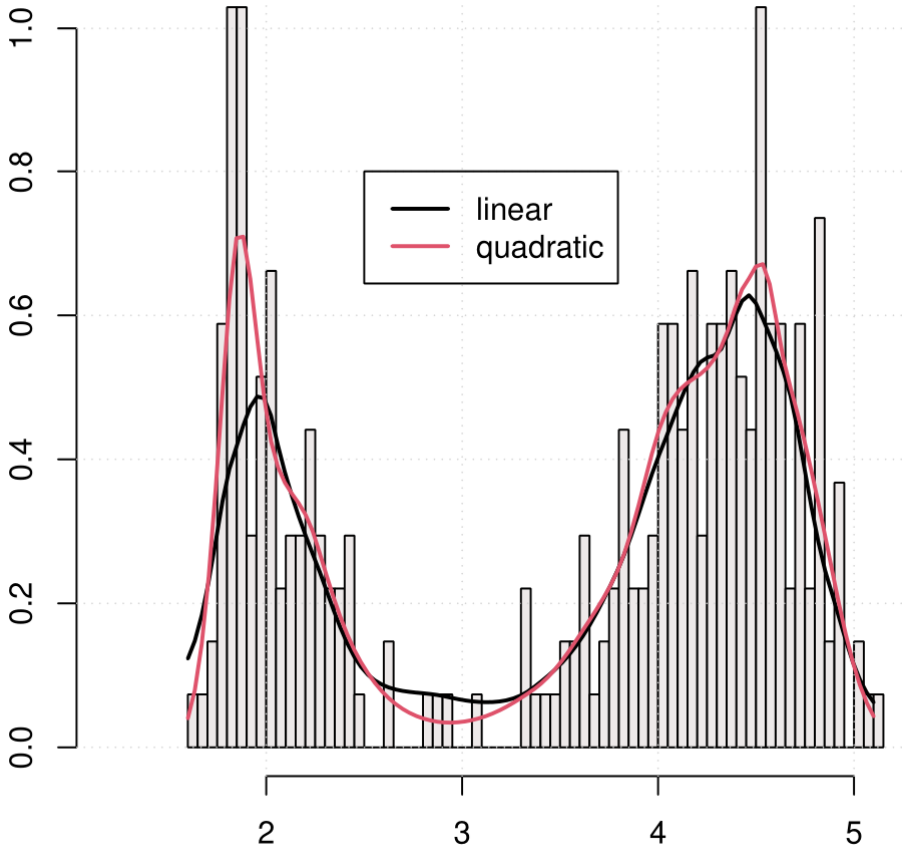


Figure 8: Local linear smoothing using linear and quadratic approximations.

6.3 Gaussian mixture models

Using a Gaussian mixture model we can approximate a wide range of densities simply by controlling the number of components K of the mixture.

Def. (Gaussian mixture model)
We define the **gaussian mixture model** for the density f of Y as

$$f(y;\vartheta) = \sum_{k=1}^{K} p_k \varphi(y;\mu_k,\sigma_k^2),$$

where $\varphi(\cdot;\mu,\sigma^2)$ is the Gaussian density with mean μ and variance σ^2.

Remark. The parameters are $\vartheta = (p_1, p_2, \ldots, p_k, \mu_1, \mu_2, \ldots, \mu_k, \sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$.

Estimation. We can find the maximum likelihood estimate of the parameters using the **expectation-maximization algorithm** (EM), which is useful for maximizing likelihoods in the presence of unobserved latent variables Z, in this case the indicators of the mixture.

Suppose that we have a mixture with $k = 2$, define the complete-data log-likelihood for (Y, Z) as

$$\begin{aligned}\mathcal{L}_{Y,Z}(\vartheta) &= \sum_{i:Z_i=1}^{n} \log\big(p\varphi(Y_i;\mu_1,\sigma_1^2)\big) + \sum_{i:Z_i=0} \log\big((1-p)\varphi(Y_i;\mu_2,\sigma_2^2)\big),\\ &= \sum_{i=1}^{n} Z_i \log\big(p\varphi(Y_i;\mu_1,\sigma_1^2)\big) + (1-Z_i)\log\big((1-p)\varphi(Y_i;\mu_2,\sigma_2^2)\big),\end{aligned}$$

Algorithm 2 EM algorithm for Gaussian mixture, $k = 2$

1: **E-step:** given a current value $\tilde{\vartheta}$, evaluate $Q(\vartheta, \tilde{\vartheta}) = \mathbb{E}_{\tilde{\vartheta}}[\mathcal{L}_{Y,Z}(\vartheta)|Y]$, i.e.

$$\begin{aligned}Q(\vartheta,\tilde{\vartheta}) &= \sum_{i=1}^{n} \mathbb{E}_{\tilde{\vartheta}}[Z_i|Y_i]\log\big(p\varphi(Y_i;\mu_1,\sigma_1^2)\big) + (1-\mathbb{E}_{\tilde{\vartheta}}[Z_i|Y_i])\log\big((1-p)\varphi(Y_i;\mu_2,\sigma_2^2)\big),\\ &= \sum_{i=1}^{n} w_i(\tilde{\vartheta})\log\big(p\varphi(Y_i;\mu_1,\sigma_1^2)\big) + (1-w_i(\tilde{\vartheta}))\log\big((1-p)\varphi(Y_i;\mu_2,\sigma_2^2)\big),\end{aligned}$$

where

$$w_i(\tilde{\vartheta}) = \mathbb{E}_{\tilde{\vartheta}}[Z_i|Y_i] = \frac{\cdots}{\cdots}.$$

2: **M-step**: maximize $Q(\vartheta, \tilde{\vartheta})$ to get the updated value of $\tilde{\vartheta}$. This is especially convenient for mixture models, since we have explicit forms for the updated parameters.

Remark. Maximization of $Q(\vartheta, \tilde{\vartheta})$ can also be substituted with a single Newton-Raphson update step without full maximization of Q.

Remark. In practice we need to estimate the number of components K. In general, we can minimize the following criteria:

- **AIC**: $k^* = \operatorname{argmin}_k \mathrm{AIC}(k) = \operatorname{argmin}_k -2\mathcal{L} + 2k$.
- **BIC**: $k^* = \operatorname{argmin}_k \mathrm{BIC}(k) = \operatorname{argmin}_k -2\mathcal{L} + k\log n$.

Sensitivity Since the Gaussian is sensitive to outliers, we could replace it by a t distribution. We exchange more robustness by paying the price of not having explicit estimates for the parameters, hence a slower estimation routine.

Bandwidth. If $\sigma_k^2 = \sigma^2 > 0$ is fixed and $K \to n$, then $\widehat{p}_k \to \frac{1}{n}$ and the MLE converges to the kernel density estimate.

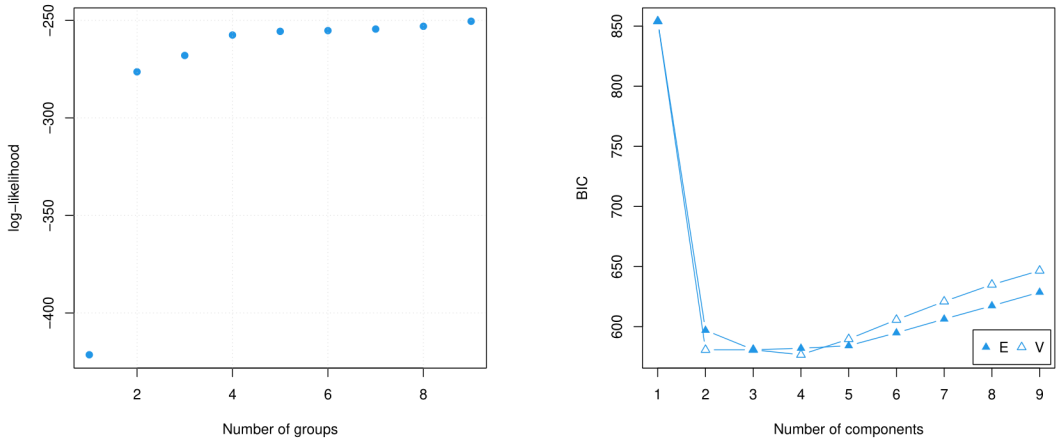


Figure 9: Mean-squared error for the Gaussian mixture model using equal variance "E" and different variances "V".

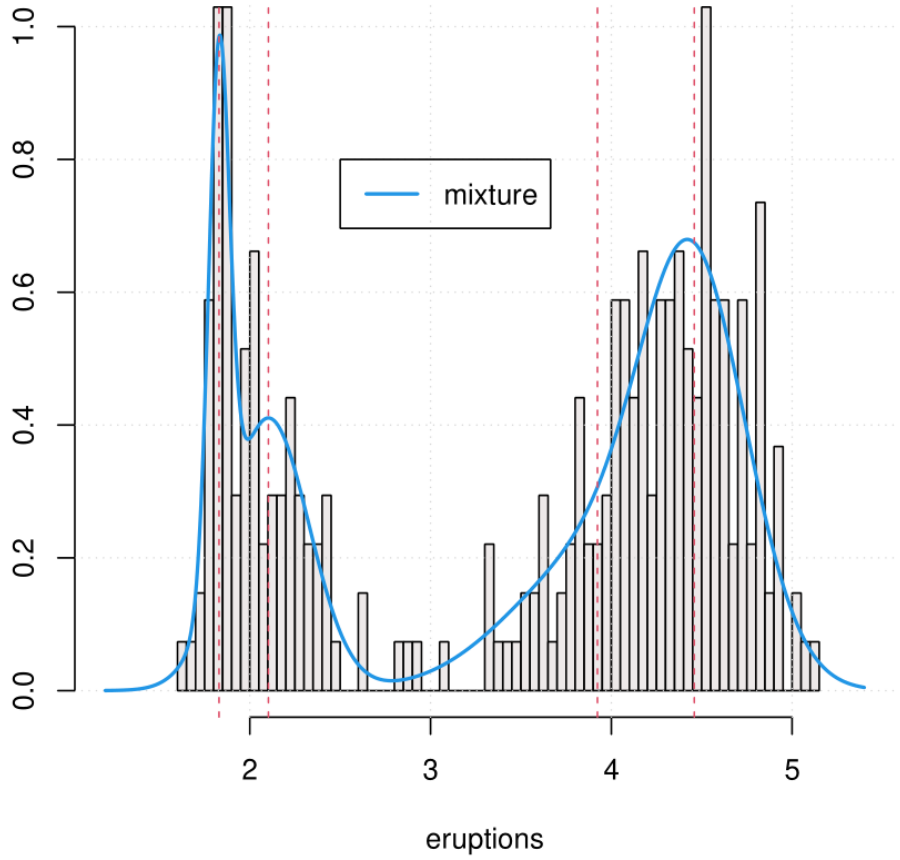


Figure 10: Example of an estimated density using four groups.

Extensions.

› Nearest neighbours

› Orthogonal series estimators

› Maximum penalized likelihood estimators

› Wavelet estimators

6.4 Kernel density estimation in d-dimensions

Suppose that $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are d-dimensional random vectors, then we can easily extend the kernel density estimator to higher dimension using straightforward generalizations of the univariate case.

Def. (Multivariate kernel density estimator)
The **multivariate kernel density estimator** of f_Y at $y \in \mathbb{R}^d$ is

$$\widehat{f}_n(\boldsymbol{y}) = \frac{1}{n|\boldsymbol{H}|^{1/2}} \sum_{i=1}^{n} K\big(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{Y}_i)\big),$$

where $\boldsymbol{H} = (h_{ij})$ is a matrix of bandwidths.

Remark. We usually use a matrix bandwidth $\boldsymbol{H}$ proportional to the covariance matrix of the data. Other choices could be $\boldsymbol{H} = \text{diag}(h_1, \ldots, h_d)$ or $\boldsymbol{H} = \text{diag}(h, \ldots, h)$.

Remark. The kernel also has common choices:

1. **Multiplicative**: $K(\boldsymbol{y}) = \prod_{j=1}^{d} \kappa(y_j)$
2. **Spherical/Radial-symmetric**: $K(\boldsymbol{y}) = \dfrac{\kappa(\sqrt{\boldsymbol{y}^\top \boldsymbol{y}})}{\int \kappa\sqrt{\boldsymbol{u}^\top \boldsymbol{u}}\, d\boldsymbol{u}}$

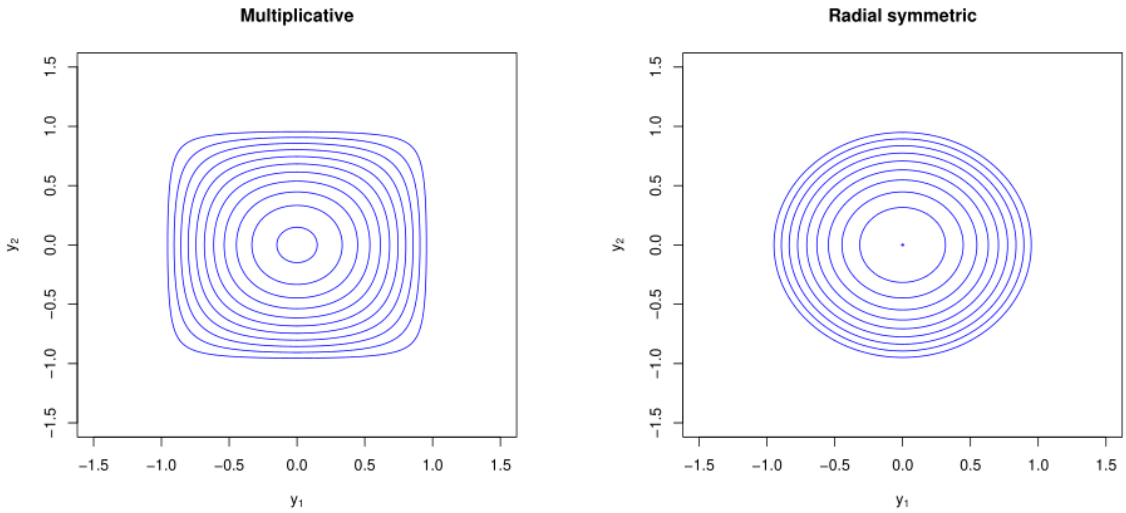


Figure 11: Example of multiplicative and radial-symmetric multivariate kernels.

Properties. Suppose that K is a kernel such that:

1. *Density*: $\int K(\boldsymbol{u}) d\boldsymbol{u} = 1$ and $K(\boldsymbol{u}) \geq 0$,

2. *Symmetric*: $\int \nu K(\boldsymbol{u}) d\boldsymbol{u} = \mathbf{0}$,
3. *Second moment*: $\int \nu \boldsymbol{u}\boldsymbol{u}^\top K(\boldsymbol{u}) d\boldsymbol{u} = \mu_{K,2} \cdot I_d$,

Then, if $\mathcal{H}_f(\boldsymbol{y})$ is the Hessian matrix of f at $\boldsymbol{y}$ and $R(K) = \int K(\boldsymbol{u})^2 d\boldsymbol{u}$, we have that

$$\text{Bias}(\widehat{f}_n(\boldsymbol{y})) = \frac{1}{2}\mu_{K,2} \operatorname{tr}\left\{\boldsymbol{H}^\top \mathcal{H}_f(\boldsymbol{y}) \boldsymbol{H}\right\}$$

$$\mathbb{V}[\widehat{f}_n(\boldsymbol{y})] \approx \frac{R(K) f(\boldsymbol{y})}{n|\boldsymbol{H}|},$$

and

$$\text{AMISE}(\boldsymbol{H}) = \frac{R(k)}{n|\boldsymbol{H}|} + \frac{\mu_{K,2}^2}{4} \int \operatorname{tr}\left\{\boldsymbol{H}^\top \mathcal{H}_f(\boldsymbol{y}) \boldsymbol{H}\right\}^2 d\boldsymbol{y}.$$

Therefore, we can also find an **asymptotically optimal bandwidth**

$$h_{\text{opt}} \sim n^{-1/(4+d)} \implies \text{AMISE}(h_{\text{opt}} \boldsymbol{H}_0) \sim n^{-4/(4+d)},$$

which means that the multivariate kernel density estimator has a slower rate of convergence compared to the univariate one.

Lecture 7: Nonparametric regression

2022-03-18

A common problem in applied statistics is that one has an dependent variable or outcome Y and various independent variable or covariates $X_1, X_2, \ldots, X_p$. The goal of this lecture is to provide some models for estimating a regression model when the response is nonparametrically dependent on X.

1. Y and X can be random variables, i.e. $m(x_1, \ldots, x_p) = \mathbb{E}[Y|X_1, \ldots, X_p]$ is the **regression function**.

> **Def. (Random design model)**
> The **random design model** for the regression of Y on X (both random variables) is defined as
> $$Y = m(X) + \varepsilon,$$
> where $\mathbb{E}[Y|X = x] = m(x)$ and $\varepsilon \perp\!\!\!\perp X$.

2. For some **designed experiments** we have complete control over the values of X, hence there is no reason to assume it to be a random variable, but rather we assume that $x_i = (x_{i1}, \ldots, x_{ip})$ are fixed design points.

> **Def. (Fixed design model)**
> The **fixed design model** instead assumes X to be known and fixed to the observed covariates x, i.e.
> $$Y = m(x) + \varepsilon,$$
> where $\mathbb{E}[Y] = m(x)$.

7.1 Linear regression

A common procedure for statistical modelling is to use a linear regression, i.e.

$$\mathbb{E}[Y|X = x] = \sum_{i=1}^{p} \beta_j x_j,$$

under the assumption that $Y|X$ follows a normal distribution. On he other hand, if $Y|X$ comes from a general dispersion family (Pace and Salvan, 1997), we can specify

$$g\big(\mathbb{E}[Y|X = x]\big) = \sum_{j=1}^{p} \beta_j x_j,$$

and $g : \text{conv}(\mathcal{Y}) \longrightarrow \mathbb{R}$ is called the **link function**, which is usually chosen for convenience reasons.

Remark 1. The parameters β usually have a direct scientific interpretation.

Remark 2. If the model is appropriate the estimates have many desirable statistical properties.

Problem. Linearity and additivity are two very strong assumptions, and sometimes we might want to relax them at the cost of less efficiency in the estimate, by specifying a more complex regression function

$$m(x) = \mathbb{E}[Y|X = x], \quad x \in \mathbb{R}.$$

7.2 Smoothing

Suppose that we have a relationship such as that in Figure 12, then we would like to estimate a model of the form

$$Y_i = m(X_i) + \varepsilon_i,$$

hence a simple estimator would be to take expectations under the empirical cumulative distribution function $\widehat{F}_n$, from which we have the plug-in estimate

$$\mathbb{E}_{\widehat{F}_n}[Y|X = x] = \frac{\sum_{i=1}^n Y_i \mathbb{1}_x(x_i)}{\sum_{j=1}^n \mathbb{1}_x(x_j)}. \tag{13}$$

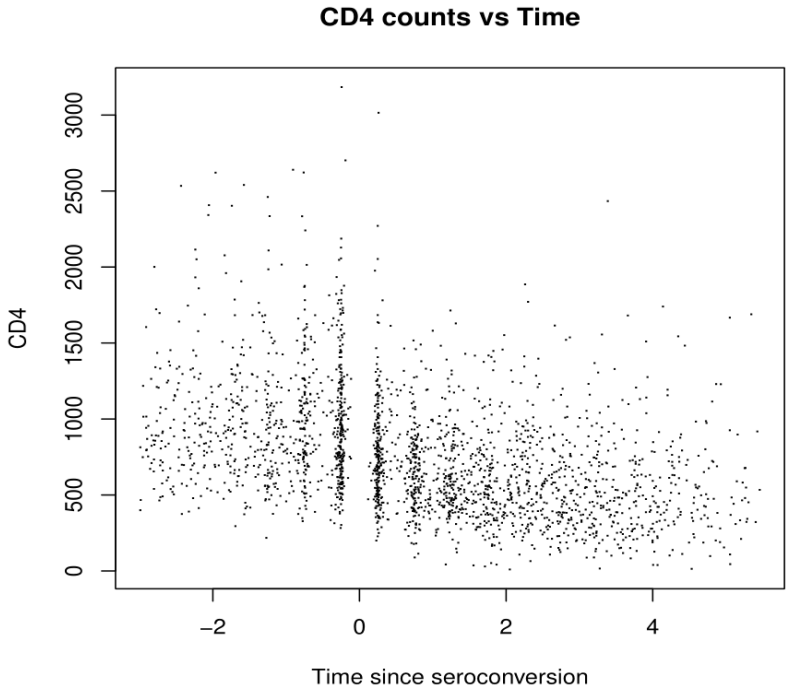


Figure 12: Scatterplot of a bivariate relationship between two continuous variables.

Some problems with the estimator in (13) is that the estimates are not smooth, and if we do not observe the particular value of x we are interested in, then we cannot calculate this expectation. Therefore, we need some approaches which allow extrapolating the nonparametric function $m(x)$ to unobserved values of $\boldsymbol{x}$, much like linear regression.

7.2.1 Parametric smoother

The parametric smoother is simply a linear regression using a polynomial basis $X = (1, x, x^2, \ldots, x^p)$ as covariates. We define a function defined by "few" parameters on the data and use least squares to

find the most appropriate estimates for the parameters,

$$\widehat{m}(x) = X(X^\top X)^{-1}X^\top Y = \sum_{i=1}^{n} W_i Y_i,$$

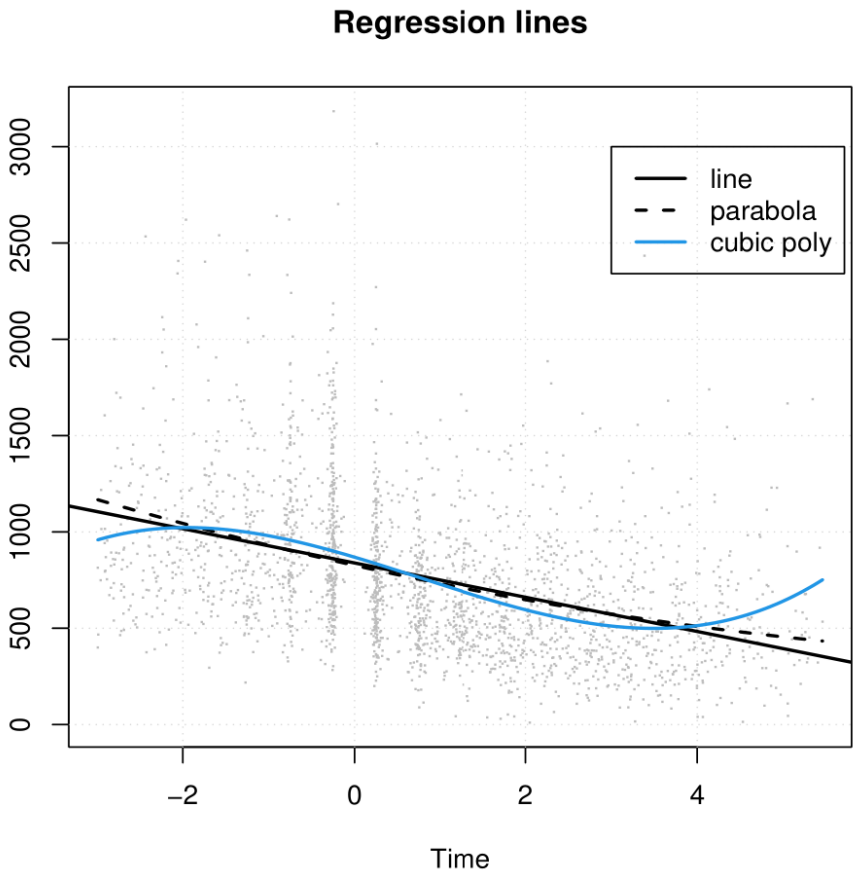


Figure 13: Example of parametric smoothing using polynomials of order 1, 2, 3. Note the problematic behaviour at the extremes of the covariate space.

7.2.2 Bin smoothers

A bin smoother, also known as a regressogram, mimics a categorical smoother by partitioning the predicted value into disjoint regions, $R_k = \{i : c_k \leq x_i < c_{k+1}\}$ for $k = 0, \ldots, K$, and then averaging the response Y in each region.

Def. (Bin smoother (regressogram) estimator)

The **bin smoother** or **regressogram estimator** is defined as

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} y_i \mathbb{1}_{R_k}(x_i)}{\sum_{j=1}^{n} \mathbb{1}_{R_k}(x_j)}, \quad x \in R_k = \{i : c_k \leq x_i < c_{k+1}\}.$$

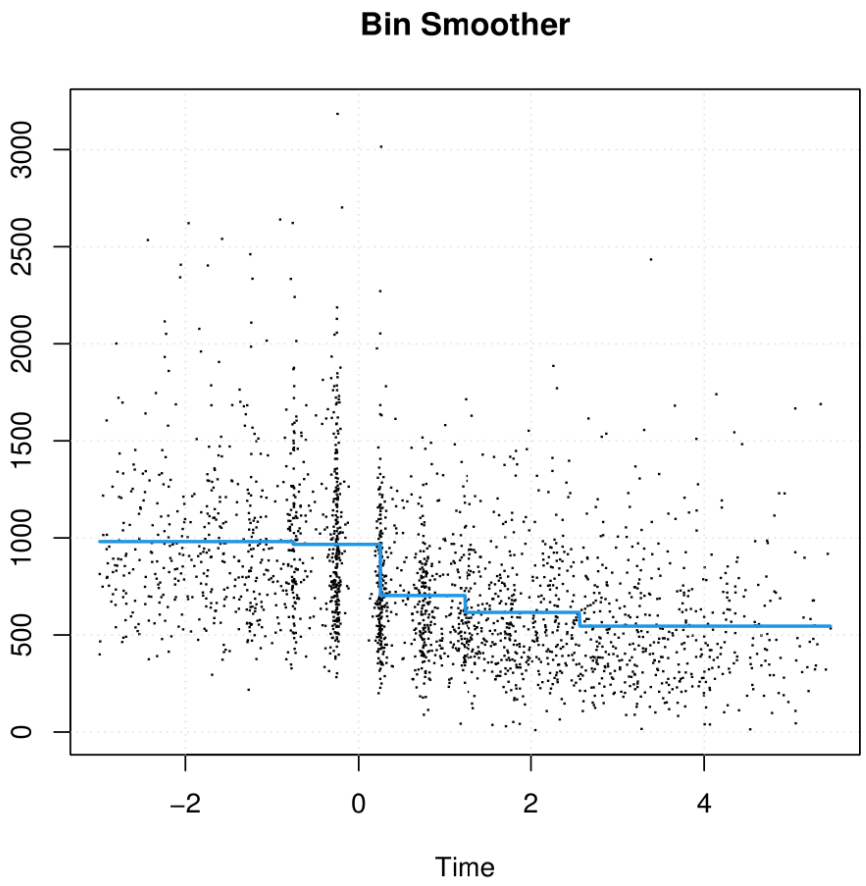


Figure 14: The bin smoother is a better-behaving smoother at the extremes of the covariate space, but shows a discontinuous behaviour in the interior.

7.2.3 Moving average

Since in general we expect the regression function $\widehat{m}(x)$ to be smooth, we can define a smoother with respect to a neighbourhood

$$N_\delta(x) = \{x_i : \|x - x_i\| \leq \delta\},$$

and estimate the δ-neighbour estimator as

$$\widehat{m}(x) = \frac{\sum_{i=1}^n y_i \mathbb{1}_{N_\delta(x)}(x_i)}{\sum_{j=1}^n \mathbb{1}_{N_\delta(x)(x_j)}}.$$

Problem. A δ-neighbour might be empty, since we might have that all x are at distance $\delta + \varepsilon$ from x_i. Therefore, we prefer a neighbourhood which groups a fixed number of observations, the k-neighbourhood.

Def. (Moving average estimator)
If $d_i = \|x - x_i\|$ and $N_k(x) = \{x_i : d_i \leq d_{(2k+1)}\}$, then the **moving average estimator** of $m(x)$ is

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}_{N_k(x)}(x_i)}{2k+1}.$$

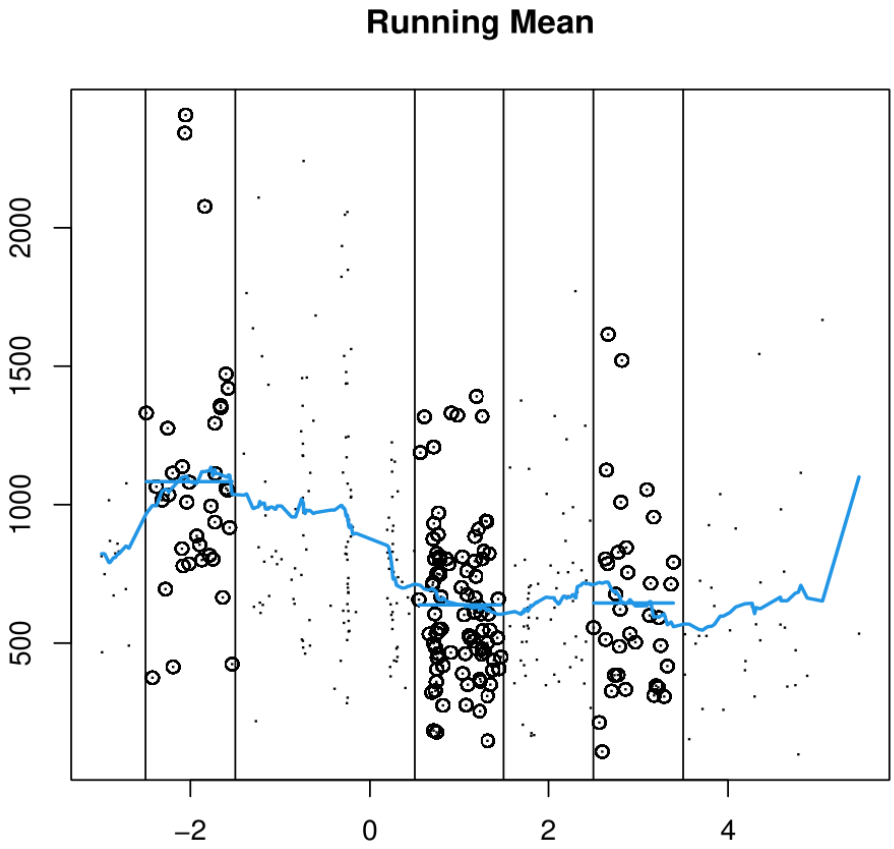


Figure 15: Moving average estimator for a particular choice of k. Note that the number k controls the smoothness of the resulting estimate.

Remark. The estimate such as in Figure 15 is usually too wiggly to be considered useful. Notice we can also fit a line instead of a constant, and the procedure is called **running-line**.

7.3 Kernel smoothers

One of the reasons why the previous smoother is wiggly is because when we move from x_i to x_{i+1}, two points are usually exchanged in the group we average about. If the exchanged points are highly dissimilar, then $\widehat{m}(x_i)$ and $\widehat{m}(x_{i+1})$ may be quite far from each other.

Idea. One way to try and fix this is by making the transition $x_i \to x_{i+1}$ smoother, for instance by using a **kernel smoother**.

7.3.1 Random design

Def. (Nadaraya-Watson estimator)
The **Nadaraya-Watson estimator** is defined as

$$\widehat{m}(x) = \sum_{i=1}^{n} W_i(x) Y_i, \tag{14}$$

where

$$W_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)},$$

and K is a positive integrable function called **kernel** and $h > 0$ is a scale parameter.

Remark. $\sum_{i=1}^{n} W_i(x_i) = 1$, hence $\widehat{m}(x)$ is a weighted average.

Remark. We do not require $\int_{\mathbb{R}} K(u)du = 1$, therefore the kernels might not be normalized.

Remark. This strategy makes sense for the estimation of the regression function,

$$m(x) = \mathbb{E}[Y|X = x] = \frac{\int_{\mathbb{R}} y f_{X,Y}(x, y)dxdy}{f_X(x)}, \tag{15}$$

and the problem is that we do not know neither $f_{X,Y}$ nor f_X. However, assuming that the kernel function is such that

$$\int K_a(u)du = 1, \quad \int_{\mathbb{R}} uK_a(y)du = 0,$$

then we estimate $\widehat{f}_{X,Y}(x, y)$ in (15) using two separate kernels (**product kernel**) for X and Y,

$$\widehat{f}_{X,Y}(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K_x(\frac{x - x_i}{h_x}) K_y(\frac{y - y_i}{h_y}),$$

from which we obtain

$$\widehat{m}(x) = \frac{\int_{\mathbb{R}} y\widehat{f}_{X,Y}(x, y)dxdy}{\widehat{f}_X(x)},$$

$$= \frac{\sum_{i=1}^{n} K_X\left(\frac{x-x_i}{h_x}\right) Y_i}{\sum_{i=1}^{n} K_X\left(\frac{x-x_i}{h_x}\right)}$$

Remark. The Gaussian kernel is not computationally efficient, since it requires all observations (x_i, y_i) for estimating the function at any point x. Since the kernel has a low impact on the resulting regression function, we usually prefer a bounded kernel for complexity reasons.

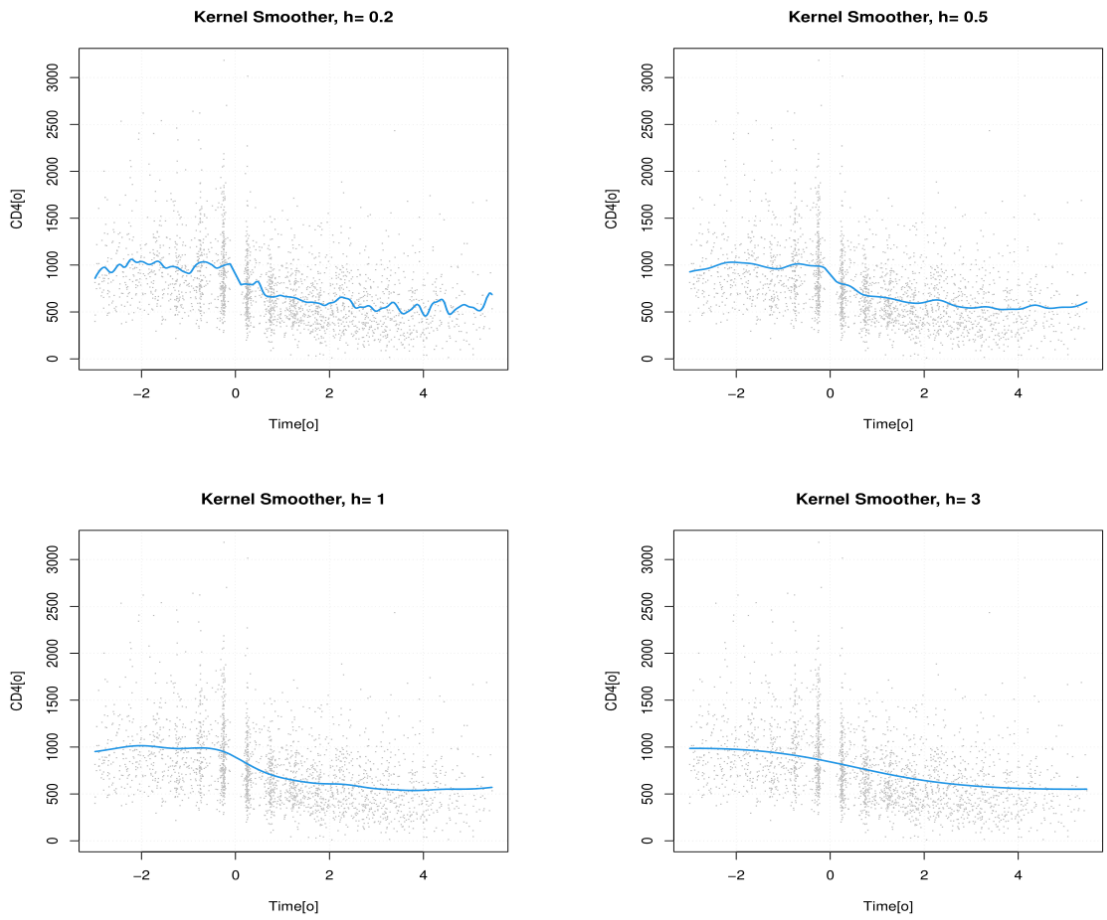


Figure 16: Example of kernel smoothed regression functions, for various choices of bandwidth h.

7.3.2 Fixed design

Suppose now that X is known, e.g. when we have a fixed design procedure and there is no distribution for X. In that case,

$$W_i(x) = \frac{1}{nh} \frac{K\left(\frac{x-x_i}{h}\right)}{f_X(x)},$$

1. *Fixed design*: $x_1 < \ldots < x_n$ and $x_i \in [a, b]$, then the "density" at the chosen points is equal to

$$\widehat{f}_X(x_i) = \frac{1}{n(x_i - x_{i-1})} \tag{16}$$

Def. (Priestley-Chao kernel estimator)

The **Priestley-Chao kernel estimator** for a fixed design uses the "fixed density" in (16),

$$\widehat{m}_{\mathrm{PC}}(x) = \sum_{i=1}^{n} (x_i - x_{i-1}) \frac{1}{h} K\left(\frac{x - x_i}{h}\right) Y_i.$$

2. On the other hand, we can use the weights given by

$$W_i^{\mathrm{GM}}(x) = \int_{s_{i-1}}^{s_i} \frac{1}{h} K\left(\frac{x-u}{h}\right) du, \tag{17}$$

for a choice of $x_{i-1} \leq s_{i-1} \leq x_i$. Usually, the default choice is to set $s_i = (x_i + x_{i+1})/2$, $s_0 = a$, $s_{n+1} = b$.

Def. (Gasser-Müller estimator)

The **Gasser-Müller estimator** uses the weights in (17) to write

$$\widehat{m}_{\mathrm{GM}}(x) = \sum_{i=1}^{n} \left[\int_{s_{i-1}}^{s_i} \frac{1}{h} K\left(\frac{x-u}{h}\right) du \right] Y_i.$$

Problem. A common problem when using nonparametric regression estimators is the **boundary bias**, due to the asymmetric contribution of observations near the boundary (Figure 13).

Solution. One possible idea is to constrain the estimator to be locally linear, such as the LOESS regression that we will discuss later.

7.4 Consistency of the kernel regression estimator

Assume that we have a single covariate X, with a response variable Y that is distributed according to

$$Y_i = m(X_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2),$$

and we apply a kernel density estimator with kernel K and bandwidth h.

Random design model. We want to characterize the behaviour of $\widehat{m}(x)$ as $n \to \infty$, and in order to do that we have to assume that:

1. $\int |K(u)| du < \infty$;
2. $\lim_{|u| \to \infty} uK(u) = 0$;
3. $\mathbb{E}[Y^2] < \infty$;
4. $h \to 0$ and $nh \to \infty$.

Prop. 6 (Consistency of the KDE)

With the above assumptions, at every point of continuity of $m(x)$, we have that

$$\frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)} \xrightarrow{P} m(x).$$

Remark. Under these assumptions, the Nadaraya-Watson estimator in (14) has asymptotic mean-squared error given by

$$\mathrm{AMSE}(\widehat{m}_n) = \frac{h^4}{4} \mu_2^2(K) \left\{ m''(x) + 2 \frac{m'(x) f_X'(x)}{f_X(x)} \right\}^2 + \frac{1}{nh} \frac{\sigma^2}{f_X(x)} \|K\|_2^2. \tag{18}$$

Fixed design model. Assume instead the univariate fixed design model and the following assumptions:

- K has support $[-1,1]$ and $K(-1) = K(1) = 0$;
- m is twice continuously differentiable;
- $\max_i |x_i - x_{i-1}| = O(n^{-1})$;
- $\mathbb{V}[\varepsilon_i] = \sigma^2 < \infty$.

Then, for $n \to \infty$ and $h \to 0$ with $nh \to \infty$ we have that

$$\text{Bias}^2(\widehat{m}_n) = \frac{h^4}{4}\mu_2^2(K)m''(x)^2 + o(h^4),$$

$$\mathbb{V}[\widehat{m}_n] = \frac{1}{nh}\sigma^2\|K\|_2^2 + o\big((nh)^{-1}\big).$$

Remark. The resulting variance is not influenced in the first order by the shape of the regression function $m(x)$, whereas the bias is affected. The asymptotic mean-squared error is therefore

$$\text{AMSE}(\widehat{m}_n) = \frac{h^4}{4}\mu_2^2(K)\underbrace{m''(x)^2}_{\text{unknown}} + \frac{1}{nh}\sigma^2\|K\|_2^2. \tag{19}$$

Remark. The asymptotic MSE of the random design (18) and of the fixed design (19) can be written in both cases as

$$\text{AMSE}(n,h) = \frac{C_1}{nh} + h^4 C_2,$$

and minimizing with respect to h gives the asymptotic behaviour of the optimal bandwidth,

$$h_{\text{opt}} = O(n^{-1/5}), \tag{20}$$

with asymptotic mean-squared error of order $O(n^{-4/5})$, which is less efficient than the $O(n^{-1})$ that we obtain using OLS. This loss of performance is the drawback that we face for using a flexible mean function estimator.

7.5 Local linear regression

Starting from Taylor's theorem, we can say that any smooth function can be approximated with a sufficiently high-degree polynomial. Assume

$$Y_i = m(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} WN(0,\sigma^2)$$

Theorem 10 (Taylor's theorem — original)
Suppose f is a real function on $[a,b]$, $f^{(K-1)}$ is continuous on $[a,b]$, $f^{(K)}(x)$ is bounded for $x \in (a,b)$, then for any distinct points $x_0 < x_1$ in $[a,b]$ there exists a point $x \in (x_0,x_1)$ such that

$$f(x_1) = f(x_0) + \sum_{k=1}^{K-1}\frac{f^{(k)}(x_0)}{k!}(x_1-x_0)^k + \frac{f^{(K)}(x)}{(K)!}(x_1-x_0)^K.$$

Theorem 11 (Taylor's theorem — Young's version)
Let f be such that $f^{(K)}(x_0)$ is bounded for x_0, then

$$f(x) = f(x_0) + \sum_{k=1}^{K} \frac{f^{(k)(x_0)}}{k!}(x - x_0)^k + o(|x - x_0|^K),$$

as $|x - x_0| \to 0$.

Remark. Another refinement of Taylor's theorem called Jackson's Inequality. Suppose f is a real function on $[a, b]$ with K is continuous derivatives, then if $\mathcal{P}_k$ is the space of polynomials of degree k we have that

$$\min_{g \in \mathcal{P}_k} \sup_{x \in [a,b]} |f(x) - g(x)| \leq C \left(\frac{b-a}{2k} \right)^K,$$

where $\mathcal{P}_k$ is the linear space of polynomials of degree k,

$$\mathcal{P}_k = \{a_0 + a_1 x + \ldots + a_k x^k, \ (a_0, \ldots, a_k) \in \mathbb{R}^{k+1}\}.$$

7.5.1 Local linear regression (LOESS)

We will now define the recipe to obtain a **loess** (local regression) smoother for a target covariate x_0. For computational and theoretical purposes we will define this weight function so that only values within a smoothing window $[x_0 - h(x_0), x_0 + h(x_0)]$ will be considered in the estimate of $m(x_0)$.

We define $h(x_0)$ so that we include a fixed $\alpha \times 100\%$ of the data, so that we have approximately regular variance at all points of the estimates for both fixed as well as random designs. Within the smoothing window $[x_0 - h(x_0), x_0 + h(x_0)]$, $m(x)$ is approximated by a polynomial, typically linear or quadratic,

$$m(x) \approx \beta_0 + \beta_1(x - x_0) + \frac{1}{2}\beta_2(x - x_0)^2,$$

Def. (General local linear regression)
The **general local linear regression** estimator of degree p for $m(x)$ at $x = x_0$ is the solution to

$$\sum_{i=1}^{n} \left(Y_i - \beta_0 + \beta_1(x_i - x_0) + \beta_2(x_i - x_0)^2 - \ldots - \beta_p(x_i - x)^p\right)^2 w_i(x_0), \tag{21}$$

where $w_i(x_0) = K\left(\frac{x_i - x_0}{h(x_0)}\right)$ is the weight for the i^{th} observation.

Remarks.

1. The kernel smoother is local linear regression when $p = 0$.
2. This estimator varies with x (in contrast to parametric least squares);
3. We need a careful choice of h in random design framework, since the AMSE in (18) is strongly influenced by $f(x)$.

Remark. The local linear regression allows the estimation of the derivative by using $\widehat{m}$ and calculating

$$\widehat{m}_p^{(\nu)}(x) = \nu!\widehat{\beta}_\nu(x),$$

and we usually have the order of the polynomial about $p = \nu + 1$ or $p = \nu + 3$.

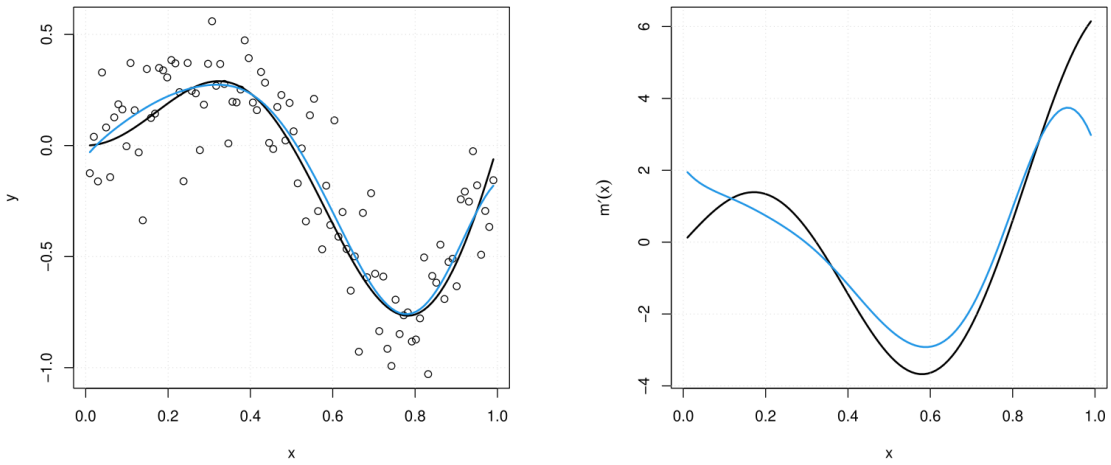


Figure 17: Estimator of the underlying function (*left*) and first derivative (*right*) for a local linear regression. We still observe a problem in the boundary of the support of x.

7.5.2 Estimator for the derivative of $m(x)$

Assume that K has support $[-1, 1]$ and $x \in [a, b]$ and that we want to estimate the k^{the} derivative $m^{(k)}(x)$ using the derivative of the local linear regression estimator in (21). Given some particular functions v and B_1, B_2, B_3 which depend on p, k, m, K (Heckman's notes), we have that

a) *Variance of the estimator*:

$$\mathbb{V}\left[\widehat{m}^{(k)}(x)\right] \sim \frac{v(x)}{nh^{2k+1}},$$

hence we need $nh^{2k+1} \to 0$.

b) *Bias of the estimator*:

$$p - k \text{ EVEN}: \quad \text{Bias}\big(\widehat{m}^{(k)}(x)\big) \sim \begin{cases} B_1(x)h^{p-k+2} & x \in (a, b) \\ B_2(x)h^{p-k+1} & x \in \{a, b\} \end{cases}$$

$$p - k \text{ ODD}: \quad \text{Bias}\big(\widehat{m}^{(k)}(x)\big) \sim B_3(x)h^{p-k+1}$$

p	$x \in [a+h, b-h]$	$x \notin [a+h, b-h]$
0	h^2	h
1	h^2	h^2
2	h^4	h^3
3	h^4	h^4
p odd	h^{p+1}	h^{p+1}
p even	h^{p+2}	h^{p+1}

Figure 18: Asymptotic order of the bias for the derivative estimator.

Remark. Seeing the results above, many researchers usually choose $p-k$ odd, for instance $p = k+1$, so that the bias for the k^{th} derivative is on the same order for both the interior and the boundary of the covariate space.

Remark. When we want to estimate m, i.e. $k = 0$, then using a local linear estimate with $p = 1$ yields asymptotically similar results to a local quadratic estimator but saves lots of computational time.

7.5.3 Robust fitting

If the errors have a symmetric distribution (heavy tails), or if there appears to be outliers we can use a robust extension of the loess. Consider the residuals from the application of the local linear regression estimator,

$$\widehat{\varepsilon}_i = Y_i - \widehat{m}(x_i),$$

then we can consider a bisquare weight function,

$$B(u; b) = \begin{cases} \left[1 - (u/b)^2\right]^2 & \text{if } |u| < b \\ 0 & \text{if } |u| > b \end{cases}$$

and apply a second smoothing on the Y_i's which takes into account the first smoothing iteration. Specifically, we can define the **robust weights** as

$$r_i = B(\widehat{\varepsilon}_i; 6m), \quad m = \text{median}(|\widehat{\varepsilon}_i|),$$

and the local linear regression is repeated by replacing the weights $w_i(x)$ with the new weights $r_i w_i(x)$.

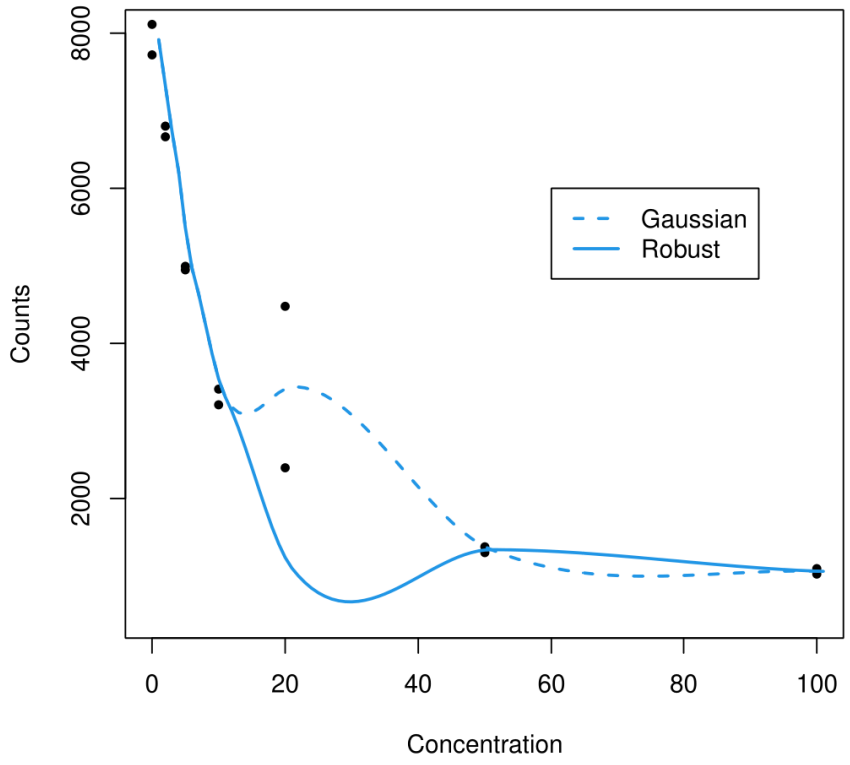


Figure 19: Comparison between robust local linear regression and LOESS.

Remark. The robust estimate is the result of repeating the above procedure several times (e.g. three times), which is how the function `lowess` is implemented in `R`.

7.5.4 Autocorrelated data

Suppose that we observe a times series with equi-spaced times $t_1, t_2, \ldots$, i.e.

$$Y_i = m(t_i) + \varepsilon_i,$$

where ε is a stationary zero-mean process (White Noise process),

$$\mathbb{E}[\varepsilon_i] = 0$$

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_{i+k}) = \gamma(k)$$

with a mixing condition $\sum_{k=1}^{\infty} k|\gamma(k)| < \infty$. Then, the asymptotic mean-squared error of the smoothed estimate takes into account the autocovariance of the process,

$$\mathrm{AMISE}(\widehat{m}_n) = \underbrace{\frac{h^4 \mu_2^2(K) \int m''(u)^2 du}{4}}_{\mathrm{Bias}^2} + \underbrace{\frac{\sigma^2 + 2\sum_{k=1}^{\infty} |\gamma(k)|}{nh} \|K\|_2^2}_{\mathbb{V}},$$

which inflates the variance of the estimator.

7.5.5 Local likelihood model

The local regression framework can be generalized to a general local likelihood regression model, by assuming that the response variable is distributed according to a **dispersion family** (Pace and Salvan, 1997),

$$Y_i \sim f(y, \vartheta(x_i)),$$

where $\vartheta_i = \vartheta(x_i)$ is the regression parameter. Then, the usual log-likelihood for the parameter ϑ is simply

$$\ell(\vartheta) = \sum_{i=1}^{n} \ell\big(Y_i, \vartheta(x_i)\big).$$

Suppose that $\vartheta(x) = \beta_0 + \beta_1 x$, then we can generalized the above quantity to a **local likelihood** for ϑ by weighting each contribution with a kernel, analogously to the weighted sum of squares,

$$\mathcal{L}_x(\vartheta) = \sum_{i=1}^{n} \ell\big(Y_i, \beta_0 + \beta_1(x_i - x)\big) K\left(\frac{x_i - x}{h(x)}\right).$$

Def. (Local likelihood estimator)
Maximizing $\mathcal{L}_x$ over the unknown β's defines the **local likelihood estimator**,

$$\widehat{\vartheta}(x) = \widehat{\beta}_0(x).$$

Remark. Extending to the GLM case, if $f(y, \vartheta)$ is a parametric family of distributions with mean $\mathbb{E}_\vartheta[Y_i] = \mu(\vartheta)$, then $g = \mu^{-1}$ is the **link function** such that $\vartheta = g(\mu)$.

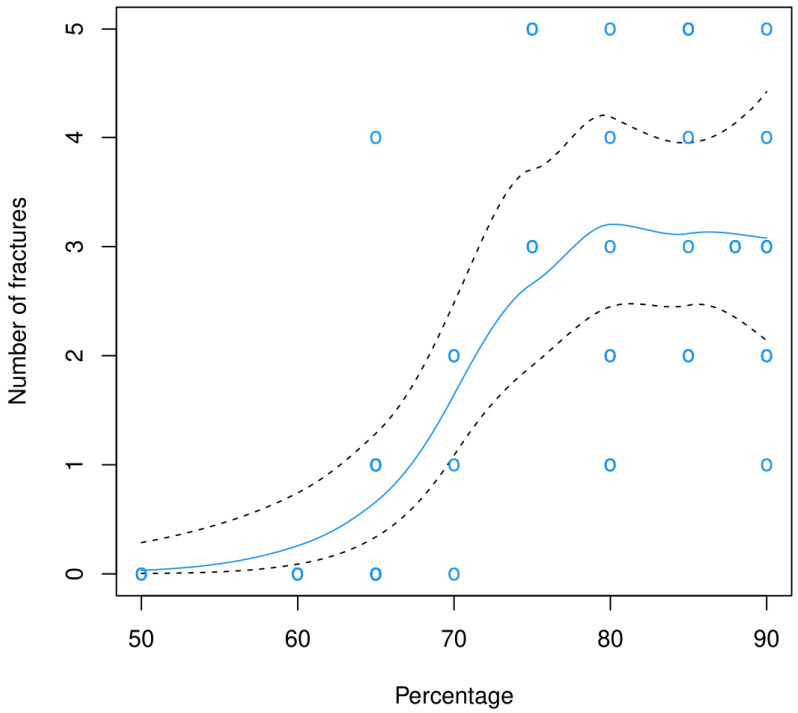


Figure 20: Local log-linear model with $Y \sim \text{Pois}(\mu)$ and $\vartheta = \log \mu$ for the number of fractures as a function of percentage of extraction.

7.6 Orthogonal series estimator

Suppose m is supported on a compact interval, i.e. $[0,1]$, and that we are in a **fixed design** case,

$$Y_i = m(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \text{WN}(0, \sigma^2).$$

where $x_i = i/n$.

Remark. The fixed design assumption is critical, since the estimators are not consistent in the random design case. Consider $L_2[0,1] = \{f : \int_0^1 f(x)^2 dx < \infty\}$, then we have multiple choices for defining a complete orthonormal basis $f_1, f_2, \ldots$, of functions in $L_2[0,1]$ such that

$$\int_0^1 f_i(x) f_j(x) dx = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{22}$$

For instance, standard results from functional analysis can be used to show that that any function $m \in L_2[0,1]$ can be represented as

$$m(x) = \sum_{k=1}^{\infty} m_k f_k(x) dx, \quad m_k = \int_0^1 m(x) f_k(x) dx,$$

where $f_1, f_2, \ldots$ are such that (22) is satisfied.

Example (Fourier series)
Any function $m \in L_2[0,1]$ can be represented as

$$m(x) = m_1 + \sum_{k=1}^{\infty} m_{2k} \sqrt{2} \cos(2\pi k x) + \sum_{k=1}^{\infty} m_{2k+1} \sqrt{2} \sin(2\pi k x),$$

where

$$m_1 = \int_0^1 m(x) dx$$

$$m_{2k} = \sqrt{2} \int_0^1 m(x) \cos(2^k \pi x) dx$$

$$m_{2k+1} = \sqrt{2} \int_0^1 m(x) \sin(2^k \pi x) dx$$

Example (Legendre polynomials)
Wikipedia.

Example (Wavelets)
Wikipedia.

Problem. The basis functions f_k are known (you choose your favourite basis of $L_2[0,1]$), but the m_k's are unknown, since we do not know the true underlying function m.

Solution. A good estimator of the unknown weights m_k's can be shown to be

$$\widehat{m}_k = \frac{1}{n}\sum_{i=1}^{n} Y_i f_k(x_i),$$

and the evaluation is especially fast using the Fast Fourier Transform (FFT).

Remark. Can we really obtain an infinite number of coefficients m_k? If we take a look at

$$\mathbb{E}[\widehat{m}_k] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i]f_k(x_i) = \frac{1}{n}\sum_{i=1}^{n} m(x_i)f_k(x_i) = \frac{1}{n}\sum_{i=1}^{n} m(i/n)f_k(i/n) \approx \int_0^1 m(x)f_k(x)dx,$$

then we see that $\widehat{m}_k$ is simply an asymptotic unbiased estimator of m_k, hence if we use too many coefficients $\widehat{m}_k$ it will be too variable.

Def. (Truncated orthogonal series)
We can define the **truncated orthogonal series** estimator as

$$\widehat{m}_K(x) = \sum_{k=1}^{K} \widehat{m}_k f_k(x).$$

Remark. K plays the role of a smoothing parameter, i.e.

- K small $\implies$ $\widehat{m}_K$ is biased
- K large $\implies$ $\widehat{m}_K$ is too variable

In practice, we choose K using cross-validation and if m is smooth then m_k will only be large for a few initial elements. However, we can have some bias issues if the largest coefficients arise late in the series and we truncate the series too early.

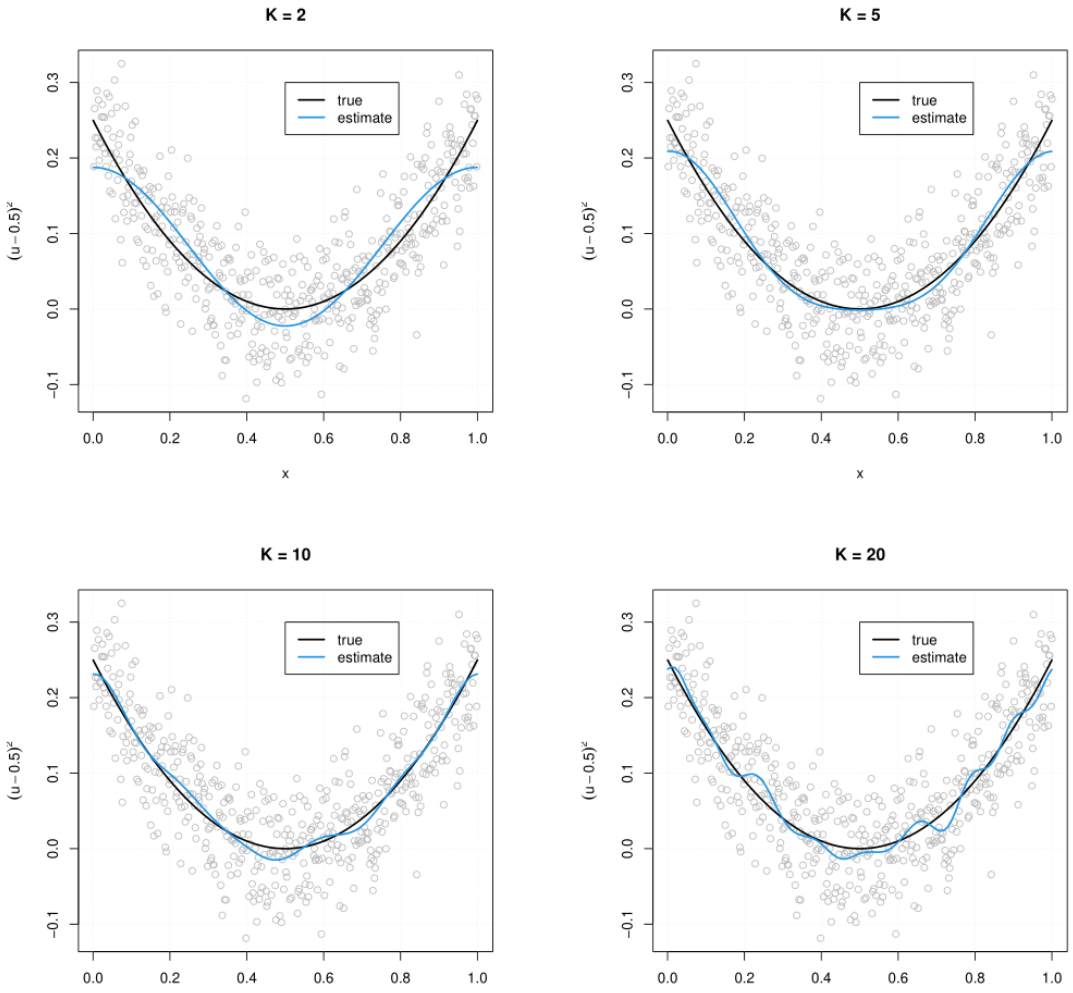


Figure 21: Simulated example of an orthogonal estimate using a Fourier basis.

Lecture 8: Spline regression

2022-04-04

Spline interpolation/regression is a form of nonparametric regression where the interpolant is a special type of piecewise polynomial called a spline. That is, instead of fitting a single, high-degree polynomial to all of the values at once, spline interpolation fits low-degree polynomials to small subsets of the values. In this lecture, we will introduce spline regression and interpolation from a theoretical point of view, whereas in the following lecture we will discuss more statistical considerations.

8.1 Introduction to splines

Before stating the regression problem, we define a suitable space in order to establish the class of regression functions over which we want to pick the best approximation for the regression function in a model of the form

$$Y_i = m(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

Def. (Sobolev space)
We define the **Sobolev space** of order d as the set functions that are absolutely continuous and with square-integrable d^{th} derivative,

$$W_2^d[a,b] = \left\{ m : m^{(j)} \text{ abs. cont. for } j = 0, \ldots, d-1 \text{ and } \int_a^b m^{(d)}(x)^2 dx < \infty \right\}. \tag{23}$$

Our goal is to select the most appropriate function $\widehat{m} \in W_2^d[a,b]$ to approximate m by working on two competing quantities:

1. a **goodness-of-fit measure** with respect to the observed data such as the mean-squared error,

$$\frac{1}{n}\sum_{i=1}^{n} \left(Y_i - m(x_i)\right)^2; \tag{24}$$

2. a measure of **smoothness** of the estimated function,

$$J_d(m) = \int_a^b m^{(d)}(x)^2 dx. \tag{25}$$

As an overall measure of performance we can combine (24) and (25) to write the objective function

$$(1-\vartheta)\sum_{i=1}^{n} \left(Y_i - m(x_i)\right)^2 + \vartheta J_d(m), \quad 0 < \vartheta < 1,$$

or more commonly if $\lambda = \vartheta/(1-\vartheta)$,

$$\frac{1}{n}\sum_{i=1}^{n} (Y_i - m(x_i))^2 + \lambda J_d(m), \quad \lambda > 0. \tag{26}$$

Remark. The parameter λ controls the strength of penalization and points towards a smoother value of the regression function when $\lambda \to \infty$. When $\lambda = 0$, no penalization is enforced and the

function tends to perfectly interpolate the observed values.

In general, any sufficiently smooth function can be well-approximated by a high polynomial such that the remainder term is

$$\text{Rem}(x)^2 \leq \frac{(b-a)^{d-1}}{(2d-1)\left\{(d-1)!\right\}^2} J_d(m),$$

and we can also prove a stronger result, namely that

$$|\text{Rem}(x_j)| \leq C \cdot J_d(m)^{1/2}.$$

Therefore, by controlling the smoothness penalization $J_d(m)$ we can put a bound on the remainder term. Therefore, we minimize the penalized goodness-of-fit criterion (26) with respect to all the function that belong to the Sobolev space $W_2^d[a,b]$ using the **Lagrange multipliers** approach.

Note that $W_2^d[a,b]$ as defined in (23) is an infinite-dimensional space, and so performing optimization of (26) over all $m \in W_2^d[a,b]$ is quite complicated. Therefore, it's important to reduce the infinite-dimensional Sobolev space optimization to a finite-dimensional optimization problem. In order to do so, we introduce the concept of **splines**, which will be the class of function over which we restrict our attention.

Def. (Spline)
A **spline** of order r with knots at $\xi_1, \ldots, \xi_k$ is any function of the form

$$s(x) = \sum_{i=1}^{r-1} \vartheta_i x^i + \sum_{j=1}^{k} \delta_j (x - \xi_j)_+^{r-1}. \tag{27}$$

Remark. We can see that the spline is piecewise polynomial function. Specifically, the previous definition yields $s(x)$ such that

1. $s(x)$ is a piecewise polynomial of order $r-1$ on any subinterval $[\xi_i, \xi_{i+1})$;
2. $s(x)$ has $r-2$ continuous derivatives;
3. $s(x)$ has an $r-1^{\text{st}}$ derivative that is a step function with jumps at $\xi_1, \ldots, \xi_k$.

Let $\mathcal{S}^r(\xi_1, \ldots, \xi_k)$ denote the set of all functions of the form of a spline (27). Then, $\mathcal{S}^r(\xi_1, \ldots, \xi_k)$ is a vector space of dimension $k + r$.

Def. (Natural spline)
A **natural spline** is a spline function of order $r = 2d$ with knots at the observed x_j's.

Remark. In particular, $s(x)$ is a polynomial of order $d-1$ outside the observed interval $[x_{(1)}, x_{(n)}]$.

Remark. The vector space $\mathcal{NS}^{2d}(x_1, \ldots, x_n)$ of natural splines is a subset of the vector space of splines $\mathcal{S}^{2d}(x_1, \ldots, x_n)$ which is obtained by placing $2d$ linear restrictions.

In particular, we have that for $s(x)$ to be a natural spline it must satisfy

$$s(x) = \sum_{i=0}^{2d-1} \vartheta_i x^i + \sum_{j=1}^{k} \delta_j (x - \xi_j)_+^{2d-1},$$

with a constraint given by

$$\vartheta_d = \ldots = \vartheta_{2d-1} = 0.$$

Lemma 4 (Properties of $\mathcal{NS}$)
If $s(x) \in \mathcal{NS}^{2d}(x_1, \ldots, x_n)$, then

1. *$\mathcal{NS}$ has dimension n;*
2. *$s(x)$ is piecewise of order $r-1$ on any subinterval $[x_i, x_{i+1}) \implies r \cdot (n-1)$ coefficients;*
3. *$s(x)$ has $r-2$ continuous derivatives $\implies (r-1) \cdot n$ constraints;*
4. *$s(x)$ is piecewise polynomial of order $r/2$ on any subinterval $[a, x_{(1)})$ and $[x_{(n)}, b] \implies 2(r/2)$ coefficients;*

The total number of degrees of freedom is

$$\textit{Coefficients} - \textit{constraints} = r(n-1) + r - (r-1)n = n.$$

Lemma 5
Let $f_1, \ldots, f_n$ be a basis for $\mathcal{NS}^{2d}(x_1, \ldots, x_n)$, then there are coefficients such that

$$f_j(x) = \sum_{i=0}^{d-1} \vartheta_{ij} x^i + \sum_{i=1}^{n} \delta_{ij} (x - x_i)_+^{2d-1},$$

and if $s(x) = \sum_{j=1}^{n} \beta_j f_j(x)$ and $m \in W_2^d[a, b]$ then

$$\int_a^b m^{(d)}(x) s^{(d)}(x) dx = (-1)^d (2d-1)! \sum_{i=1}^{n} m(x_i) \sum_{j=1}^{n} \beta_j \delta_{ij}.$$

Remark. The last equivalence is important in order to write (25) as a sum instead of as an integral.

Theorem 12 (Minimization of $J_d(m)$)
Let $f_1, \dots, f_n$ be a basis for $\mathcal{NS}^{2d}(x_1, \dots, x_n)$ with associated design matrix $F = \big(f_j(x_i)\big)_{ij}$ and let $\boldsymbol{a} = (a_1, \dots, a_n)^\top$ be a specified vector of constants. Then, if $n > d$, the unique minimizer of $J_d(m)$ in (25) over all $f \in W_2^d[0,1]$ that satisfies $s(x_i) = a_i$ for $i = 1, \dots, n$ is

$$s(x) = \sum_{j=1}^{n} c_j f_j(x),$$

where $\boldsymbol{c} = (c_1, \dots, c_n)^\top$ is the unique solution to $F\boldsymbol{c} = \boldsymbol{a}$. In particular, the matrix F has full rank n.

Consequences. The minimizing criterion is reduced if we replace m by the natural spline $s(x)$ satisfying

$$s(x_i) = m(x_i), \quad i = 1, \dots, n,$$

thus we may minimize the overall loss function over functions of the form

$$s(x) = \sum_{j=1}^{n} c_j f_j(x),$$

where $f_1, \dots, f_n$ is a basis of $\mathcal{NS}^{2d}(x_1, \dots, x_n)$.

Example (Natural cubic splines ($d = 2, r = 4$))
Natural cubic splines are splines such that $\xi_i = x_i$, $s(x)$ is a cubic polynomial on any subinterval $[x_i, x_{i+1})$, and is linear outside $[x_{(1)}, x_{(n)}]$, therefore

$$s(x) = \sum_{i=0}^{3} \vartheta_i x^i + \sum_{j=1}^{n} \delta_j (x - x_j)_+^3, \tag{28}$$

and we have a total of $n + 3$ parameters.

Basis functions. We can define a basis for all spline functions $s(x) \in \mathcal{S}^r(\xi_1, \dots, \xi_k)$, si that

$$s(x) = \sum_{i=1}^{r+k} \alpha_i f_i(x).$$

The truncated power basis for splines is numerically unstable, since there are no orthogonal splines and thus the normal equations are highly ill-conditioned. We instead prefer to use an orthogonal basis set so that computations are more stable

Def. (B-spline basis)
We define the **B-spline basis of order r** with knots at $\xi_1, \dots, \xi_k$ as the set of functions

$$\{B_{i-1,r}(x) : i = -(r-1), \dots, k\}, \tag{29}$$

where $B_{i,1}(x) = \mathbb{1}_{[\xi_i, \xi_{i+1})}(x)$ for $i = 0, \dots, k$ and

$$B_{i,r}(x) = \frac{x - \xi_i}{\xi_{i+r-1} - \xi_i} B_{i,r-1}(x) + \frac{\xi i + r - x}{\xi_{i+r} - \xi_{i+1}} B_{i+1,r-1}(x), \quad i = -(r-1), \dots, k, \tag{30}$$

where as a convention we understand $0/0 = 0$ and $\texttt{NaN}/0 = 0$.

Remark. If $r = 2$ we have the **linear B-splines** and for $r = 4$ we have the **cubic B-splines**.

Remark. Splines that are defined as (30) are linearly independent, hence they are an orthogonal basis. The regression based on B-splines is therefore as stable as orthogonal polynomial regression.

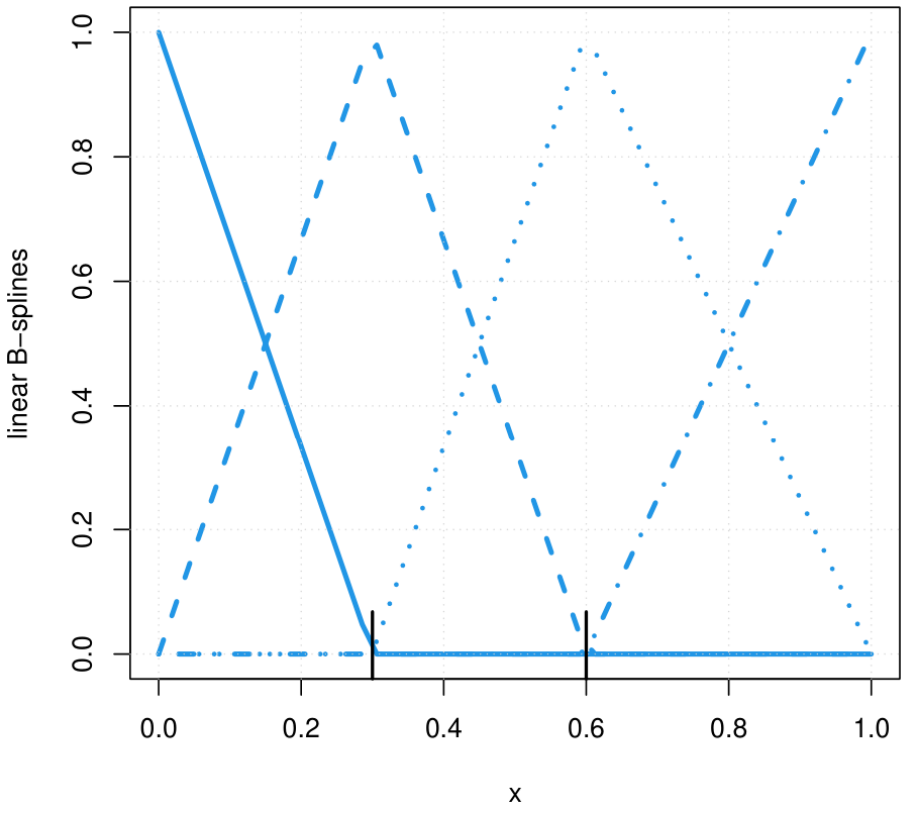


Figure 22: Example of the first linear B-splines ($r = 2$).

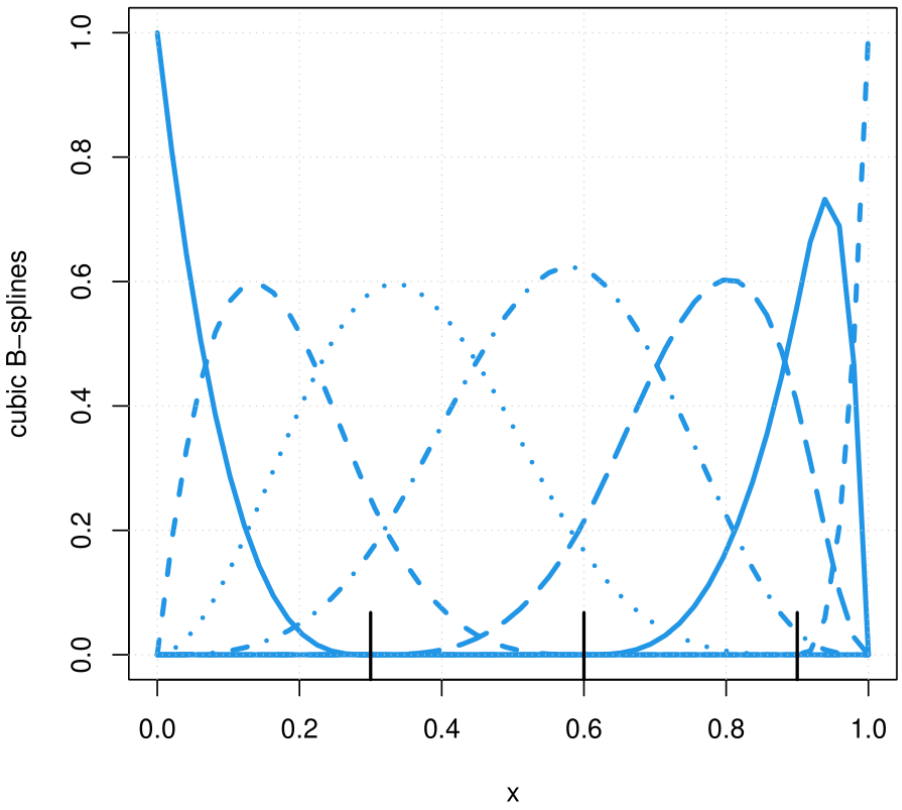


Figure 23: Example of the first cubic B-splines ($r = 2$).

Key properties. Let $a = x_1 < \ldots < x_n = b$, then we have that for the natural cubic splines are such that $s(x_i) = y_i$ and $g''(x_1) = g''(x_n) = 0$.

Prop. 7 (NS minimizes the penalty term)
A natural cubic splines minimizes

$$J_2(g) = \int_{x_1}^{x_n} g''(x)^2 dx,$$

for all $g \in W_2^2[a, b]$ that interpolates (x_i, y_i) for $i = 1, \ldots, n$.

Remark. Since $m(x) = a + bx$ has null second derivative, it minimizes the penalty. Therefore, as $\lambda \to \infty$ we have that $m(x) \to a + bx$ (Figure 24).

Prop. 8 (NS minimizes the penalized loss function)
A cubic spline minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - m(x_i)\big)^2 + \lambda \int_{x_1}^{x_n} m''(x)^2 dx, \quad \lambda > 0$$

for all $m \in W_2^2[x_1, x_n]$.

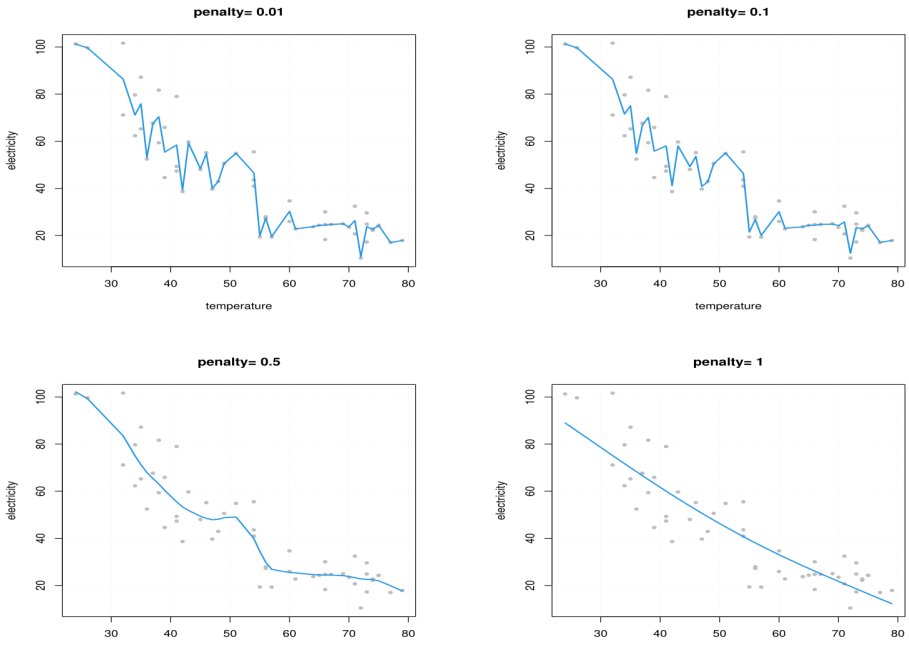


Figure 24: As the penalty parameter increases, a higher smoothness of the resulting spline is enforced. For $\lambda \to \infty$, the spline tends to a linear regression function.

Final remarks.

› Using splines we can also get an estimate of the derivative, similarly to kernel density regression.

The coefficients $\beta_\lambda = (\beta_{\lambda 1}, \ldots, (\beta_{\lambda n})^\top$ are the solution to

$$(F + n\lambda G)\beta_\lambda = \boldsymbol{y}, \tag{31}$$

where $F = \big(f_j(x_i)\big)_{ij}$, $G = (-1)^d(2d-1)!\delta_{ij}$, and thus

$$\widehat{\beta}_\kappa = (F + n\lambda G)^{-1}\boldsymbol{y}.$$

From this, it's immediate to see that the estimated regression curve is a linear combination of the observed $\boldsymbol{y}$,

$$\widehat{m} = F(F + n\lambda G)^{-1}\boldsymbol{y} = P_\lambda \boldsymbol{y}.$$

Alternatively, we can premultiply (31) by $F^\top$ to write

$$(F^\top F + n\lambda F^\top G)\beta_\lambda = F^\top \boldsymbol{y},$$

and the solution is

$$\beta_\lambda = (F^\top F + n\lambda \Omega)^{-1} F^\top \boldsymbol{y}. \tag{32}$$

where

$$\Omega = \left(\int_a^b f_i^{(d)}(x) f_j^{(d)}(x) dx \right)_{ij}.$$

From (32), we can see that the fitted values are

$$\widehat{m} = F(F^\top F + n\lambda\Omega)^{-1}F^\top \boldsymbol{y} = S_\lambda \boldsymbol{y},$$

where S_λ is called **hat matrix**.

Example (Demmler and Reinsch basis)
The Demmler and Reinsch basis is defined as the natural linear splines that interpolate the constant and the set of functions

$$g_j(x) = \sqrt{2}\cos(j\pi x), \quad j = 1, \ldots, n-1.$$

We find that we can write the fitted values using the D-R basis is (slides)

$$\widehat{m}_\lambda(x) = \frac{1}{n}\sum_{i=1}^{n} K_n(x, x_i; \lambda) y_i,$$

where

$$K_n(x, z; \lambda) = 1 + \sqrt{2}\sum_{j=1}^{n-1} \frac{\cos(j\pi z) f_{j+1}(x)}{1 + \lambda\gamma_j},$$

and this kernel is well-approximated by

$$K\left(\frac{x-z}{\sqrt{\lambda}}\right) \cdot \frac{1}{\sqrt{\lambda}},$$

where $K(u)$ is the Laplace density.

Remark. Splines have a lot in common with kernel density estimators and, in general, yield similar results.

8.2 Least squares regression splines

Suppose that we want to find the least squares minimizer when

$$\mathbb{E}[Y|X = x] = m(x) \approx \sum_{k=1}^{K} \beta_k f_k(x),$$

then we simply have that if $F = \big(f_k(x_j)\big)_{k,j}$ then

$$\widehat{\beta} = (F^\top F)^{-1} F^\top \boldsymbol{y},$$

and

$$\widehat{m}(x) = F(F^\top F)^{-1} F^\top \boldsymbol{y}.$$

- **Truncated power basis**: $F^\top F$ is ill-conditioned and the solution is unstable.
- **B-spline basis**: computationally fast and stable, $F^\top F$ can be inverted using $\mathcal{O}(n)$ calculations.

Extensions. We can consider straightforward extensions of spline regression, for instance by applying it to:

› **Dependent data**: $\mathbb{V}[\boldsymbol{Y}] = \Sigma$, then

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} (\boldsymbol{y} - F\beta)\Sigma^{-1}(\boldsymbol{y} - F\beta) = (F^\top \Sigma^{-1} F)^{-1} F^\top \Sigma^{-1} \boldsymbol{y}$$

› **General likelihoods**: for instance, in logistic regression we can assume $Y_i \sim \text{Ber}\big(\vartheta(x_i)\big)$, where

$$m(x_i) = \text{logit}\left(\vartheta(x_i)\right) = \log \frac{\vartheta(x_i)}{1 - \vartheta(x_i)} = \sum_{k=1}^{K} \beta_k f_k(x_i),$$

which is easy to maximize using the `glm` function in R after constructing the regression matrix F.

› **Monotone data**: suppose that $\beta_k \geq 0$, then the linear combination $s(x) = \sum_{k=1}^{K} \beta_k f_k(x_i)$ of the B-splines basis $f_1, \ldots, f_K$ is **non-decreasing** (Ramsay, 1988). Hence, we want to find the solution to

$$\begin{aligned} \underset{\beta}{\text{argmin}} \quad & \|\boldsymbol{y} - F\beta\|^2, \\ \text{s.t.} \quad & \beta_k \geq 0 \end{aligned}$$

We can do so by using the `cobs` package.

8.3 Choosing the number and location of knots

We have many ways of doing so, which might be more or less technical:

1. Stepwise variable selection (addition/deletion) based on linear models;
2. Penalized regression splines (Eilers and Marx, 1996);
3. Zhou Chen 2001 Jasa using "spatially adaptive" splines;
4. Bayesian regression splines with a prior on the number of knots and on their location (Dimatteo et al., 2001).

8.3.1 Stepwise variable selection

Using the `polspline` package, we can apply methods based on:

› **Forward selection**: start from `y ~ 1` and iteratively add each term separately whenever the new variable is significant.

› **Backward selection**: start from `y ~ .` and iteratively remove each term separately whenever the chosen variable is not significant.

› **Mixed selection**: we can either add, delete, or swap a term in the model for one that is not in the model.

› **Information criteria**: we can use criteria such as

– Corrected AIC for regression

$$\mathrm{AIC}_c = n \log \mathrm{RSS} + n \frac{1 + p/N}{1 - (p+2)/N}.$$

– BIC criterion

$$\mathrm{BIC} = n \log \mathrm{RSS} + p \log n.$$

– Mallow's criterion

$$C_p = \frac{\mathrm{RSS}_p}{\mathrm{RSS}_K/(n-K)} - n + 2p.$$

8.3.2 Penalized regression splines

Models with different K are not nested, therefore we cannot use stepwise selection. On the other hand, we can penalized the regression coefficients $\beta = (\beta_1, \dots, \beta_K)^\top$ by solving the regression problem

$$\|\boldsymbol{y} - F\beta\|^2 + \lambda \beta^\top P \beta, \tag{33}$$

where $P \succeq 0$ and $\lambda \geq 0$. We can chose P by observing that

$$\begin{aligned}
\int_a^b m^{(2)}(x)^2 dx &\approx \int_a^b \left(\sum_{k=1}^K \beta_k f_k''(x) \right)^2 dx \\
&= \int_a^b \sum_{k=1}^K \sum_{j=1}^K \beta_k f_k''(x) \beta_j f_j''(x) dx \\
&= \sum_{k=1}^K \sum_{j=1}^K \beta_k \left(\int_a^b f_k''(x) \beta_j f_j''(x) \right) \beta_j dx \\
&= \beta^\top P \beta,
\end{aligned}$$

by setting $P = (p_{ij})_{ij} = \left(\int_a^b f_i''(x) f_j''(x) dx \right)_{ij}$. With this choice, the solution to (33) is a simple ridge estimator.

Other choices.

› We can use the penalized B-splines (Eilers and Marx, 1996)

– Penalizing $\sum_{j=1}^{k+r+1} \beta_j^2$, the function $f \equiv 0$ is reproduced but not constants.

– Penalizing $\sum_{j=2}^{k+r+1} (\beta_j^2 - \beta_{j-1})^2$, constant functions are reproduced but not lines.

– Penalizing $\sum_{j=3}^{k+r+1} (\beta_j^2 - 2\beta_{j-1} + \beta_{j-2})^2$, lines are reproduced but not quadratic. ...

8.3.3 Knot selection

In general, we should use a lot of knots but less than n. Ruppert et al. (??) recommend 4-5 points between knots but no more than 20-40, although we can choose them by cross-validation. The choice of location can either be chosen using the quantiles of the x values, although the computational burden is higher. Using equi-spaced knots allows a computationally-easier solution.

Remark. Calculating a ridge linear regression can be performed by observing that

$$\operatorname*{argmin}_{\beta} \|\boldsymbol{y} - F\beta\|^2 + \lambda\beta^\top P\beta,$$

where $P = D^\top D$, then we can write the solution as

$$\widehat{\beta} = \operatorname*{argmin}_{\beta} \left\| \begin{pmatrix} \boldsymbol{y} \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} F \\ \sqrt{\lambda} D \end{pmatrix} \right\|^2$$

Lecture 9: Statistical issues

2022-04-06

9.1 Degrees of freedom

Alternatively, sometimes we want the estimate of m to predict well. Specifically, let Y_i^* be an unobserved "future" observation independent of the (x_i, Y_i)'s but with the same distribution as Y_i.

Def. (Predictive squared error)
The **predictive squared error** of $Y^*1, \ldots, Y_n^*$ is defined as

$$\text{PSE} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left(\widehat{m}(x_i) - Y_i^*\right)^2\right]. \tag{34}$$

Remark. Note that we can write

$$\begin{aligned}\text{PSE} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left(\widehat{m}(x_i) - m(x_i) + m(x_i) - Y_i^*\right)^2\right] \\ &= \text{MSE}(\widehat{m}) + \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}[Y_i],\end{aligned}$$

hence

$$\operatorname{argmin} \text{PSE} = \operatorname{argmin} \text{MSE}.$$

There are quite a few methods in the literature for choosing a smoothing parameter:

1. subjectively, by looking at plots or by specifying the effective number of parameters;
2. via a plug-in method, an automatic method that requires asymptotic formulae to "plug into";
3. via cross-validation, an automatic method that doesn't require asymptotic formulae;
4. via a bias estimation technique called EBBS (empirical bias bandwidth selection) introduced by Ruppert (1997);
5. by recasting the problem as a random effects problem or Bayesian problem.

Methods 1–3 are well-known and have been studied for a long time. On the other hand, method 5 hasn't been extensively studied but is most likely appropriate only for estimating m.

Remark. Most of the smoothers presented here are linear smoothers of the form

$$s(x) = \sum_{j=1}^{n} S_j(x) Y_j.$$

Def. (Linear smoother)
we say that $\widehat{m}$ is a **linear smoother** if there exists an $n \times n$ matrix S independent of $\boldsymbol{Y}$ such that

$$\widehat{m} = S\boldsymbol{Y}.$$

Remark. This representation allows us to easily write the variance of $\widehat{m}$ as

$$\mathbb{V}[S\boldsymbol{Y}] = S\,\mathbb{V}[\boldsymbol{Y}]S^\top \overset{iid}{=} \sigma^2 SS^\top.$$

As they are linear smoothers (possibly depending on some parameter λ), we can study the behaviour of the MSE since we know that

$$\begin{aligned}\mathrm{MSE}(\lambda) &= \mathrm{Bias}(\widehat{m})^2 + \mathbb{V}[\widehat{m}] \\ &= \frac{1}{n} e_\lambda^\top e_\lambda + \frac{1}{n}\operatorname{tr}(S_\lambda S_\lambda^\top)\sigma^2,\end{aligned}$$

and thus the predictive squared error is

$$\mathrm{PSE}(\lambda) = \frac{1}{n} e_\lambda^\top e_\lambda + \left\{1 + \frac{1}{n}\operatorname{tr}(S_\lambda S_\lambda^\top)\right\}\sigma^2.$$

How can we characterize the amount of smoothing being performed?

- The smoothing parameters provide a characterization, but they do not permit us to compare between different smoothers.
- Using the connections between smoothing and multivariate linear regression we can compare the amount of smoothing both locally and globally, using the matrix S_λ
- We will define **variance reduction**, **effective number of parameters** and **influence** for linear smoothers.

Def. (Variance reduction)
We define the **variance reduction** as the proportion of explained variance

$$\frac{\mathbb{V}[\widehat{m}(x)]}{\mathbb{V}[Y]} \overset{iid}{=} \frac{\sigma^2 \sum_{i=1}^n S_i^2(x)}{\sigma^2} = \sum_{i=1}^n S_i^2(x).$$

Remark. Under mild conditions, one can show that the explained variance is less than one.

Remark. Recall that

$$\sum_{i=1}^n \mathbb{V}[\widehat{m}(x_i)] = \operatorname{tr}(SS^\top)\sigma^2,$$

and so the total variance reduction is

$$\frac{\sum_{i=1}^n \mathbb{V}[\widehat{m}(x_i)]}{\sum_{i=1}^n \mathbb{V}[Y_i]} = \frac{\operatorname{tr}(SS^\top)}{n},$$

and we note that in linear regression we have that

$$\sum_{i=1}^n \mathbb{V}[\widehat{m}(x_i)] = p\sigma^2.$$

Def. (Degrees of freedom (effective number of parameters))
For a linear smoother, the **degrees of freedom (effective number of parameters)** is

$$\mathrm{df} = \mathrm{tr}(SS^\top). \tag{35}$$

Alternative (i). In linear regression, since $S = X(X^\top X)^{-1}X^\top$, then $SS^\top = S$ and $\mathrm{df} = p$. Hence, $\mathrm{tr}(S)$ can also be a definition for the **effective number of parameters** in linear smoothers.

Alternative (ii). We notice that for a linear smoother,

$$\mathbb{E}[(\boldsymbol{Y} - \widehat{\boldsymbol{m}})^\top(\boldsymbol{Y} - \widehat{\boldsymbol{m}})] = \big\{n - 2\,\mathrm{tr}(S) + \mathrm{tr}(SS^\top)\big\}\sigma^2,$$

and in linear regression this is exactly $(n-p)\sigma^2$. From this, we also have that

$$p = 2\,\mathrm{tr}(S) - \mathrm{tr}(SS^\top),$$

therefore we can also use $2\,\mathrm{tr}(S) - \mathrm{tr}(SS^\top)$ as a third definition for effective number of parameters.

In summary, we can use as the definition of **degrees of freedom** for a smoother, the quantities

1. $\mathrm{tr}(S)$;
2. $\mathrm{tr}(SS^\top)$;
3. $\mathrm{tr}(2S - SS^\top)$.

If $SS^\top = S$, then all quantities $1. - 3.$ are the same and, if S is not symmetric, we can show that under relatively mild assumptions,

$$1 \le \mathrm{tr}(SS^\top) \le \mathrm{tr}(S) \le 2\,\mathrm{tr}(S) - tr(SS^\top) \le n.$$

Figure 25 shows this relationship for a `loess` regression function.

Influence. The sensitivity of $\widehat{m}(x_i)$ to the i^{th} data point can be measured by the element s_{ii}, much like the influence analysis applied to the i^{th} element in linear regression.

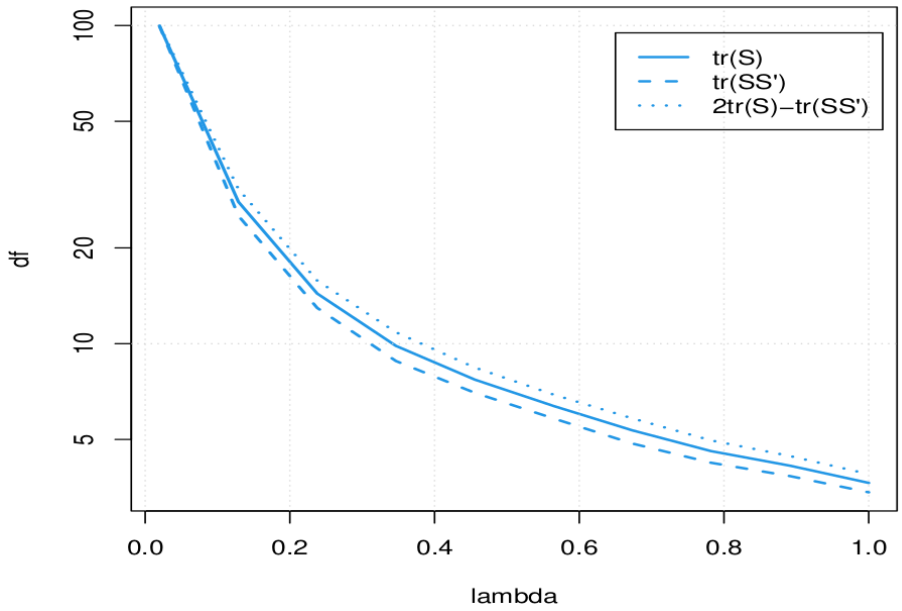


Figure 25: Comparison of three definitions of effective number of parameters for a local linear linear regression, `loess` (p = 2).

9.2 Eigendecomposition analysis

For a linear smoother $\widehat{m} = S\boldsymbol{y}$ with symmetric smoothing matrix $S_{n\times n}$, the eigendecomposition of S can be used to gain insight about its behavior.

9.2.1 Symmetric smoothers

We study the behaviour of its eigendecomposition

$$S = UDU^\top = \sum_{j=1}^{n} \vartheta_j \boldsymbol{u}_j \boldsymbol{u}_j^\top,$$

where $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ is an orthonormal basis of eigenvectors of S with eigenvalues $\vartheta_1 \geq \vartheta_2 \geq \cdots \geq \vartheta_n$, and $D = \text{diag}(\vartheta_1, \ldots, \vartheta_n)$.

Example (Linear regression)
If we consider the simple linear regression $Y_i = a + bx_i + \varepsilon_i$, the resulting matrix

$$H = X(X^\top X)^{-1} X^\top \boldsymbol{y},$$

has only two nonzero eigenvalues, where $X = (\mathbf{1}_n \;\; \boldsymbol{x})$.

Example (Cubic splines)
Cubic splines are an example of a symmetric smoother, and its eigenvectors resemble poly-

nomials of increasing degree. On the other hand, `loess` and "nearest-neighbor" methods are not symmetric smoothers.

Prop. 9 (Eigenvalues for cubic splines)
It is possible to show that the first two eigenvalues of S are $\vartheta_1 = \vartheta_2 = 1$ and they correspond to linear functions of the predictor such that $J_d(m) = 0$. Moreover, we have that $0 < \vartheta_n < \ldots < \vartheta_3 < 1$.

Proof.
No.

□

Remark. The action of the smoother can be understood as follows: if $\boldsymbol{y} = \boldsymbol{u}_j$, then it is shrunk by an amount ϑ_j such that

- $\vartheta_j = 1 \implies S\boldsymbol{u}_j = \boldsymbol{u}_j$ and there is no smoothing.
- $\vartheta_j = 0 \implies S\boldsymbol{u}_j = \boldsymbol{0}$ and there is a lot of smoothing.

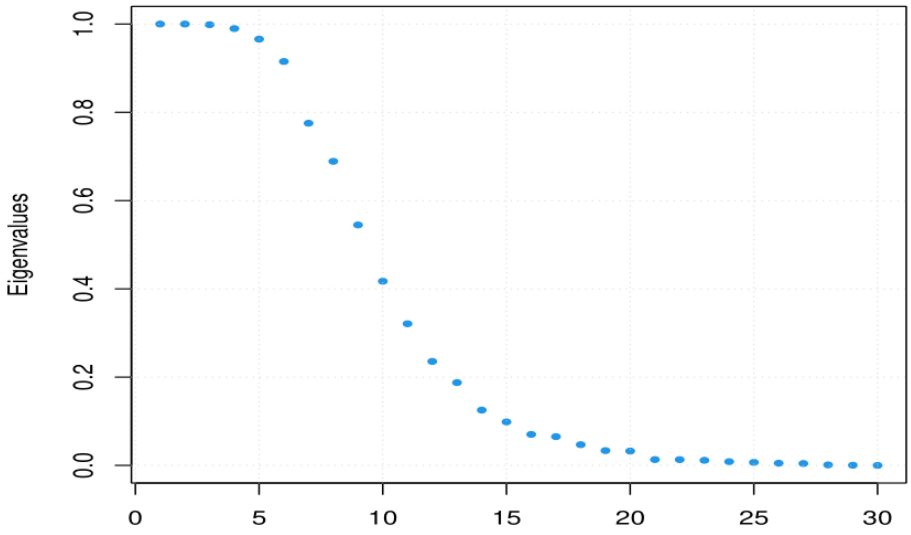


Figure 26: Example of behaviour of eigenvalues 1 through 30 for S in a smoothing spline. Eigenvectors later in the sequence correspond to more localized components.

9.2.2 Filtering

Cubic smoothing splines, regression splines, linear regression, polynomial regression are all symmetric smoothers. If S is not symmetric we have complex eigenvalues and the above decomposition is not as easy to interpret. However, we can still gain insight by using the singular value decomposition

$$S = UDV^\top,$$

from which we can write the smoothed function as

$$\widehat{m} = S\boldsymbol{Y} = UDV^\top \boldsymbol{Y}.$$

This corresponds to a sequence of transformations on $\boldsymbol{Y}$:

1. A basis transformation $Z = V^\top \boldsymbol{Y}$;
2. Shrinking using $D\boldsymbol{Z}$ the components that are related to "unsmooth components";
3. Transforming back to $\widehat{m} = U\widehat{\boldsymbol{Z}}$.

The change of basis idea described above has been explored in the Wavelet theory.

Example (Signal processing)
In the signal processing theory, we "filter" a signal using a smoothing operation, for instance by considering

$$Y_t = \frac{1}{3}(Y_{t-1} + Y_t + Y_{t+1}),$$

and the **transfer function** describes how the power of certain frequency components are reduced.

Low-pass filter. The power of the higher frequency components are reduced.

High-pass filter. The power of the lower frequency components are reduced.

Idea. We can view the eigenvalues of smoother matrices S as transfer functions, and under this framework the smoothing spline can be considered a low-pass filter. If we look at the eigenvectors of the smoothing spline we notice they are similar to sinusoidal components of increasing frequency.

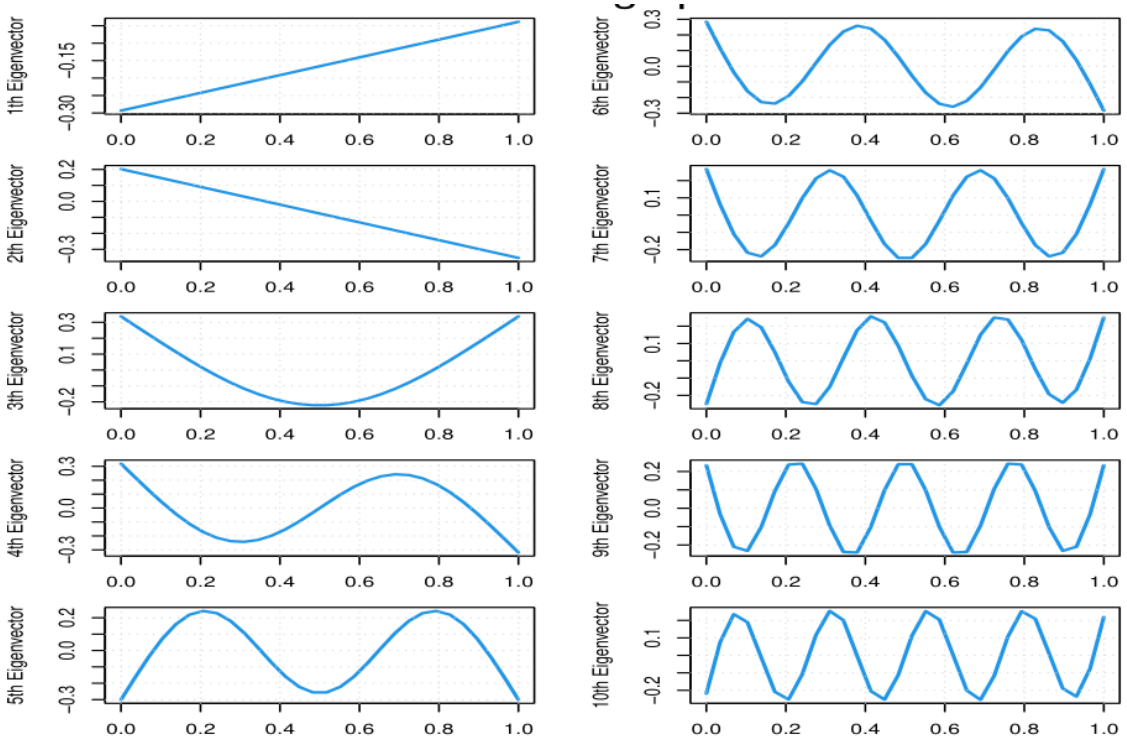


Figure 27: First 10 eigenvectors of smoothing splines, for a regularly-spaced grid of observations. We observe that the higher eigenvectors represent higher-frequency components of the spectrum.

9.3 Bandwidth selection

9.3.1 Plug-in methods

For pointwise evaluation we have that

$$\mathbb{E}[\widehat{m}_h^{(j)}(x) - m^{(j)}(x)]^2 = \frac{V(x)}{n^\alpha h^\beta} + h^\gamma B^2(x), \tag{36}$$

whereas for global evaluation we have

$$\int_a^b \mathbb{E}[\widehat{m}_h^{(j)}(x) - m^{(j)}(x)]^2 w(x)dx = \frac{1}{n^\alpha h^\beta} \int_a^b V(x)w(x)dx + h^\gamma \int_a^b B^2(x)w(x)dx. \tag{37}$$

Optimal bandwidth. For a local smoothing parameter choice we choose h_{opt} that minimizes the pointwise asymptotic MSE (36). If we are interested in a global smoothing choice, we can minimize the expression (37). in both cases, we estimate the unknowns σ^2, $f_X(x)$ and the derivatives of m. Then, we plug the estimates into the formula for h_{opt}.

Example (Useful trick)

A trick is to replace $w(x)$ with $f(x)$ in (37), in order to minimize instead

$$\int_a^b \mathbb{E}[\widehat{m}_h^{(j)}(x) - m^{(j)}(x)]^2 f(x)dx,$$

so that the component of $f(x)$ vanishes from the variance component. The remaining unknowns are σ^2 and $m^{(p+1)}(x)$.

Example ($p = 1, j = 0$)
Differentiating ... with respect to h and setting it equal to zero, we obtain

$$h_{\text{opt}} = Hn^{-1/5}, \tag{38}$$

where

$$H = \left(\frac{(b-a)\int_a^b K^2(u)du}{\left(\int_a^b u^2 K(u)du\right)^2} \right)^{1/5} \cdot \left(\frac{\sigma^2}{\int_a^b m''(x)^2 f_X(x)dx} \right)^{1/5}, \tag{39}$$

hence the only unknowns are σ^2 and $m''(x)$.

Estimating σ^2. The simplest method is to use

$$\widehat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (Y_i - Y_{i-1})^2,$$

and it can be shown (*exercise*) that $\widehat{\sigma}^2 \xrightarrow{P} \sigma^2$.

Estimating $m''(x)$. The first idea could be to estimate using the *Rule Of Thumb I* (ROT):

1. Use a parametric estimate $\widehat{m}_{\text{par}}$ to get $\widehat{m}''_{\text{par}}$.
2. Estimate $\int_a^b m''(x)^2 f(x)dx$ with $\sum_{i=1}^n \widehat{m}''_{\text{par}}(x_i)^2$ and plug into (39) to get h_{ROT}.

Instead of using a parametric method, we can apply a more involved algorithm which yields better results (Ruppert et al., 1995) and is currently implemented in the `KernSmooth` package (Algorithm 3).

Algorithm 3 `KernSmooth` estimation of $m''(x)$ (Ruppert, 1997)

1: Estimate m with $\widehat{m}_{\text{par}}$ parametrically.

2: Differentiate $\widehat{m}_{\text{par}}$ to get $\widehat{m}''_{\text{par}}$ and $\widehat{m}^{(iv)}_{\text{par}}$, and plug into

$$\int_a^b m''(x)m^{(iv)}f_X(x)dx \approx \frac{\sum_{i=1}^n m''(x_i)m^{(iv)}(x_i)}{n}.$$

3: Estimate the optimal bandwidth for $m''(x)$ using

$$g_{\text{opt}} = \mathcal{G} \cdot \left(\frac{\sigma^2(b-a)}{\int m''(x)m^{(iv)}(x)f_X(x)dx}\right)^{1/7} n^{-1/7}.$$

4: Estimate $\widehat{m}''_{g_{\text{ROT}}}(x)$ by knowing that

$$\widehat{m}^{(\nu)}(x) = \nu!\widehat{\beta}_\nu(x).$$

5: Plug the estimated quantities into H to get

$$h_{\text{opt}} = \widehat{H}n^{-1/5}.$$

9.3.2 Cross-validation

Remember that one of our goals is trying to find the optimizer

$$h_{\text{opt}} = \operatorname*{argmin}_h \text{MSE}(h) = \operatorname*{argmin}_h \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\widehat{m}_h(x_i) - m(x_i)]^2,$$

and the idea of **cross-validation** is to consider an independent set of data (x_i^*, y_i^*) from the same data-generating process of $Y_1, \ldots, Y_n$. The idea is that

$$\operatorname*{argmin}_h \text{PSE}(h) = \operatorname*{argmin}_h \text{MSE}(h) + \sigma^2, \tag{40}$$

and so finding an optimal $\widehat{h}_{\text{opt}}$ using (40) can provide a decent estimate of h_{opt}.

Cross-validation tries to imitate this by leaving out points $\{(x_i, y_i) : i \in n_k\}$ one subset $k = 1, \ldots, K$ at a time and estimating the smooth function at x_i based on the remaining $n - n_k$ points. The cross-validation estimate of h_{opt} is

$$\widehat{h}_{\text{opt}} = \operatorname*{argmin}_h \text{CV}(h) = \operatorname*{argmin}_h \frac{1}{n}\sum_{k=1}^K \left[Y_i - \widehat{m}_h^{(-k)}(x_i)\right]^2, \tag{41}$$

and under the assumption that $\widehat{m}_h^{(-n_k)}(x_i) \approx \widehat{m}_h(x_i)$ we have that

$$\mathbb{E}[\text{CV(h)}] \approx \text{PSE}(h).$$

By the law of large numbers, we also have that

$$\operatorname*{argmin}_h \mathbb{E}[\text{CV}(h)] \approx \operatorname*{argmin}_h \text{PSE}(h).$$

In practice. We optimize (41) using a grid of values h and choosing the h that minimizes it.

LOOCV for linear smoothers. In linear smoothers $\widehat{m} = S\boldsymbol{Y}$ such that constants are smoothed into constants,

$$S\mathbf{1}_n = \mathbf{1}_n \iff \sum_{j=1}^{n} s_{ij} = 1, \quad \text{for all } i.$$

we have that the quantity $\widehat{m}_h^{(-i)}(x_i)$ can be written as

$$\widehat{m}_h^{(-i)}(x_i) \propto \sum_{j \neq i} s_{ij} y_j = \frac{1}{1 - s_{ii}} \sum_{j \neq i} s_{ij} y_j, \tag{42}$$

since the smoother preserves constants also for $\widehat{m}^{(-i)}$ and thus must integrate to one. We can write (42) as

$$\widehat{m}_h^{(-i)}(x_i) = \sum_{j \neq i} s_{ij} y_j + s_{ii} \widehat{m}_h^{(-i)}(x_i),$$

from which

$$\begin{aligned} Y_i - \widehat{m}_h^{(-i)}(x_i) &= y_i - \sum_{j \neq i} s_{ij} y_j + s_{ii} \widehat{m}_h^{(-i)}(x_i) \\ &= y_i - \widehat{m}_h(x_i) + s_{ii}\big(Y_i - \widehat{m}_h^{(-i)}(x_i)\big) \\ &= \frac{Y_i - \widehat{m}_h(x_i)}{1 - s_{ii}}. \end{aligned}$$

Finally, we can write the CV criterion for a linear smoother as

$$\mathrm{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i - \widehat{m}_h(x_i)}{1 - s_{ii}} \right)^2.$$

Remark. Sometimes it's faster to compute $\mathrm{tr}(S)$ instead of computing the whole S and then finding the elements. In the linear model, for instance,

$$\mathrm{tr}\left(X(X^\top X)^{-1} X^\top\right) = \mathrm{tr}\left(X^\top X (X^\top X)^{-1}\right) = p.$$

Therefore, we prefer using the *Generalized Cross Validation* (GCV) criterion

$$\mathrm{GCV}(h) = \sum_{i=1}^{n} \left(\frac{Y_i - \widehat{m}_h(x_i)}{n - \mathrm{tr}(S)} \right)^2. \tag{43}$$

Remark. Note that the quantity

$$(n - \mathrm{tr}(S)) \cdot \mathrm{GCV}(h) = \sum_{i=1}^{n} \frac{\overbrace{\big(Y_i - \widehat{m}_h(x_i)\big)^2}^{\mathrm{RSS}}}{\underbrace{n - \mathrm{tr}(S)}_{\text{res. df}}}, \tag{44}$$

and so the above equation (44) looks like an estimate of the variance of the ε_i's, with the numerator being the residual sum of squares from a regression and the denominator being the appropriate degrees of freedom.

9.3.3 Pointwise bootstrap

Assuming a regression model

$$Y_i = m(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

then we can draw a sample of size n with replacement from (x_i, y_i) to get a bootstrap-based confidence interval for the regression function $\widehat{m}(x)$ using the percentile CI's.

Remark. The resulting envelope can't be interpreted as a $1 - \alpha$ confidence interval for the true curve, but rather a representation of the variability in the process yielding $\widehat{m}(x)$.

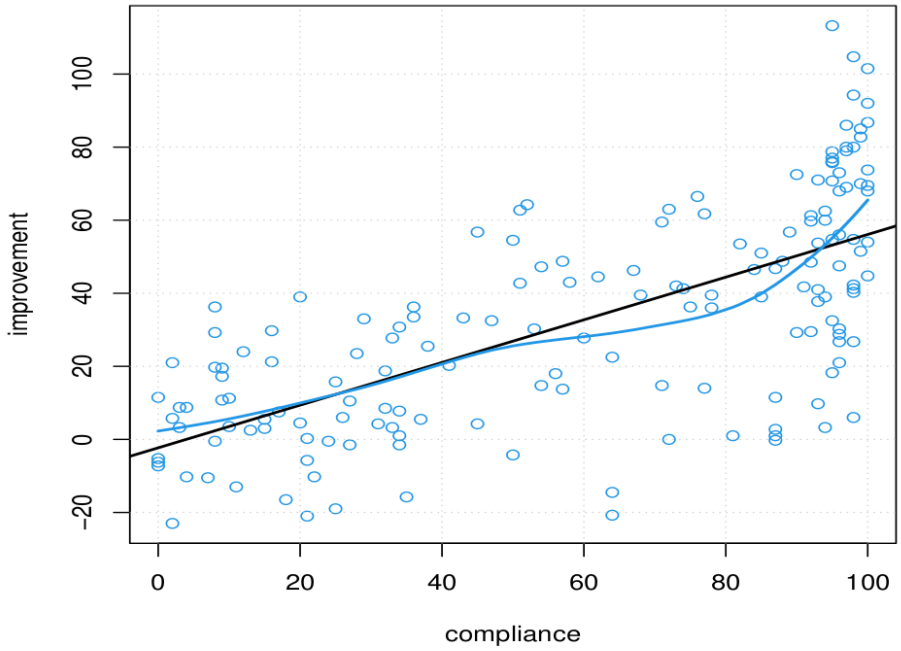


Figure 28: Example of a fitted linear model and a `loess` fit.

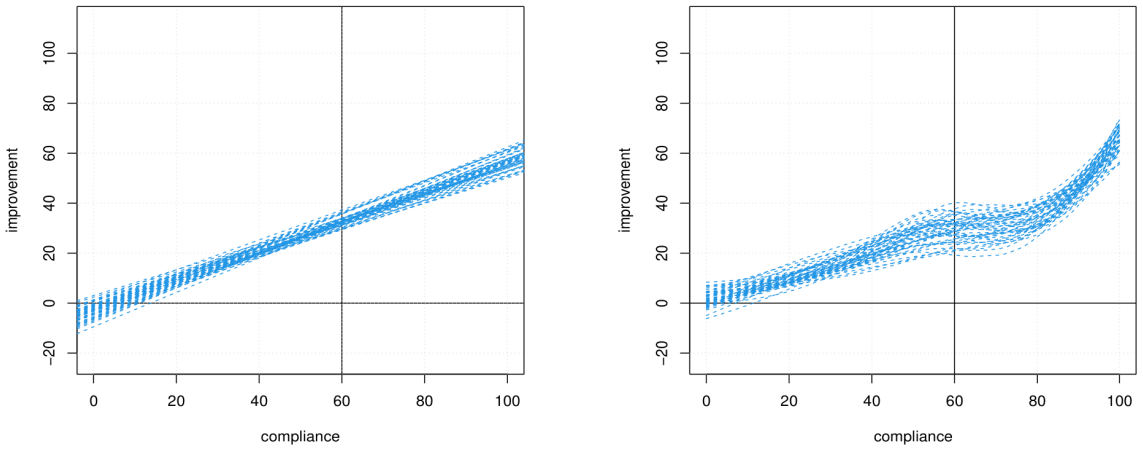


Figure 29: Bootstrap-based confidence sets for the linear model (*left*) and the `loess` fit (*right*) using $B = 50$.

We can show that if $\widehat{\boldsymbol{m}} = S\boldsymbol{Y}$, then the variance-covariance matrix of $\widehat{\boldsymbol{m}}$ is

$$\mathbb{V}[\widehat{\boldsymbol{m}}] = \sigma^2 SS^\top,$$

and pointwise standard errors are $\sigma^2 \operatorname{diag}(SS^\top)$. Since our estimate is usually biased,

$$\mathbb{E}[\widehat{\boldsymbol{m}}] = S\boldsymbol{m} \neq \boldsymbol{m},$$

then the confidence intervals are much more convenient to calculate for $S\boldsymbol{m}$ than for $\boldsymbol{m}$. Think of $\tilde{\boldsymbol{m}} = S\boldsymbol{m}$ as the best possible approximation to the "truth" $\boldsymbol{m}$, then we would like to construct global confidence bands.

9.3.4 Global confidence bands

We now discuss how to build a confidence band which contains the whole function with a fixed level of confidence. Suppose for instance that σ^2 is known, then from the properties of Gaussian random variables it's easy to see that

$$(\widehat{\boldsymbol{m}} - \tilde{\boldsymbol{m}})^\top (\sigma^2 SS^\top)^{-1} (\widehat{\boldsymbol{m}} - \tilde{\boldsymbol{m}}) \sim \chi^2_n, \tag{45}$$

and thus the confidence set for $\tilde{\boldsymbol{m}}$ of probability $1 - \alpha$ is

$$C_\alpha = \left\{ \boldsymbol{g} \in \mathbb{R}_n : (\widehat{\boldsymbol{m}} - \boldsymbol{g})^\top (\sigma^2 SS^\top)^{-1} (\widehat{\boldsymbol{m}} - \boldsymbol{g}) \sim \chi^2_n \right\}.$$

Since in practice σ^2 has to be estimated, we can use the approximation

$$\widehat{\sigma}^2 = \frac{(\boldsymbol{y} - \widehat{\boldsymbol{m}})^\top (\boldsymbol{y} - \widehat{\boldsymbol{m}})}{n - \operatorname{tr}(2S - SS^\top)},$$

and plug-in the estimator $\widehat{\sigma}^2$ into (45) to get

$$(\widehat{\boldsymbol{m}} - \tilde{\boldsymbol{m}})^\top (\widehat{\sigma}^2 SS^\top)^{-1} (\widehat{\boldsymbol{m}} - \tilde{\boldsymbol{m}}) \sim ??,$$

whose distribution however depends on the underlying true distribution of the ε_i's. We have two alternatives:

1. If the Gaussian assumption holds, then we can argue that the distribution is

$$\left\{n - \text{tr}(2S - SS^\top)\right\} + \text{tr}(SS^\top \sim F_{\text{tr}(SS^\top), n-\text{tr}(2S-SS^\top)}.$$

2. If we are unsure, we can use the bootstrap to construct an approximate distribution $\widehat{G}$ of G. In this case, we prefer the **semiparametric bootstrap** (Algorithm 4).

Algorithm 4 Semiparametric bootstrap

1: Estimate $\widehat{m}$ using a smoother.
2: Calculate residuals $\widehat{\varepsilon} = \boldsymbol{y} - \widehat{\boldsymbol{m}}$.
3: Obtain the bootstrap sample $\boldsymbol{Y}^* = \widehat{\boldsymbol{m}} + \widehat{\varepsilon}^*$, where $\widehat{\varepsilon}^*$ are resampled values from $\widehat{\varepsilon}$.
4: Calculate the statistic

$$(\widehat{\boldsymbol{m}}^* - \widehat{\boldsymbol{m}})^\top \left(\widehat{\sigma}^{*2} SS^\top\right)^{-1} (\widehat{\boldsymbol{m}}^* - \widehat{\boldsymbol{m}})$$

Lecture 10: Further extensions

2022-04-11

10.1 Multivariate splines methods

The problem with the multidimensional case is that there is no unique way of seeing the problem, and different approaches might lead to comparable results.

Our goal is to characterize the conditional expectation,

$$\mathbb{E}[Y|\boldsymbol{x}] = m(\boldsymbol{x}) = m(x_1, \ldots, x_p), \tag{46}$$

for instance by using a regression function of the form

$$m(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{47}$$

Equation (47) is very useful since it summarizes the contribution of each predictor with a single coefficient and provides an easy way to predict Y for a new set of covariates $x_1, \ldots, x_p$.

However, in many cases we are interested in finding a nonlinear estimation procedure for (46). Because Taylor's theorems also applies to multidimensional functions it is relatively straightforward to extend local regression to cases where we have more than one covariate. For instance, by expanding $m(x_1, x_2)$ around $\boldsymbol{x}_0 = (x_{01}, x_{02})$ we can write

$$m(x_1, x_2) \approx \beta_0 + \beta_1(x_1 - x_{01}) + \beta_2(x_2 - x_{02}) + \beta_3(x_1 - x_{01})(x_2 - x_{02}) + \frac{1}{2}\beta^4(x_1 - x_{01})^2 + \frac{1}{2}\beta_5(x_2 - x_{02})^2, \tag{48}$$

and we can find the local quadratic regression $\widehat{m}(\boldsymbol{x}_0)$ by finding the $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_5)^\top$ that minimizes

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i(\boldsymbol{x}_0)\big\{Y_i - m(x_1, x_2)\big\},$$

where $m(x_1, x_2)$ is approximated by (48). Here, $w(\boldsymbol{x}_0)$ is a **multivariate kernel** and there are two straightforward ways to construct it:

Def. (Product kernel)
We define a **product kernel** as the function defined by the product

$$w_i(\boldsymbol{x}_0) = \prod_{j=1}^{p} w_{ij}(x_{0j}),$$

where each w_{ij} is a marginal kernel for the j^{th} variable,

$$w_{ij}(x_{0j}) = w\left(\frac{x_{ij} - x_{0j}}{h_j}\right).$$

Def. (Global kernel)
We define a **global kernel** as

$$w_i(\boldsymbol{x}_0) = w\left(\frac{D(\boldsymbol{x}, \boldsymbol{x}_0)}{h}\right),$$

where $D(\boldsymbol{x}, \boldsymbol{x}_0)$ is a distance function between $\boldsymbol{x}$ and $\boldsymbol{x}_0$ and $h > 0$ is a bandwidth parameter.

Remark. Defining a distance depends on the particular problem, in particular

› For spatial data with covariates that are measured in the same unit of measurement, we can use the **Euclidean** distance,

$$D_E(\boldsymbol{x}, \boldsymbol{x}_0) = \sqrt{\sum_{j=1}^{p}(x_j - x_{0j})^2}.$$

› For different units of measurement, we can rescale them by $v_j = \mathrm{sd}(x_j)$ and use

$$D_E(\boldsymbol{x}, \boldsymbol{x}_0) = \sqrt{\sum_{j=1}^{p}\left(\frac{x_j - x_{0j}}{v_j}\right)^2}.$$

› Finally, for correlated covariates we can use a more generalized distance,

$$D_G(\boldsymbol{x}, \boldsymbol{x}_0) = (\boldsymbol{x} - \boldsymbol{x}_0)^\top V^{-1}(\boldsymbol{x} - \boldsymbol{x}_0),$$

where V is the variance-covariance matrix of the x's.

Remark. Methods of bandwidth selection and statistical inference for local polynomial multiple regression are essentially identical to the methods discussed previously for nonparametric simple regression.

10.1.1 Curse of dimensionality

The curse of dimensionality is both a computational and geometrical effect that affects our ability of estimating nonparametrically a regression function in the multivariate case.

› **Computationally**, this refers to the fact that the computational burden of some methods can increase exponentially with dimension.

› **Statistically**, we have that to maintain a given degree of accuracy of a nonparametric estimator, the sample size must increase exponentially with the dimension p. Specifically, from (??) we have that for a p-dimensional nonparametric regression,

$$\mathrm{AMSE}(\widehat{m}_{h_n}) = \mathcal{O}\big(n^{-4/(4+p)}\big),$$

and so by fixing a target value $\mathrm{AMSE} = C$ we need a sample size of order

$$n = \frac{1}{C}^{(4+p)/p}.$$

10.1.2 Thin plate splines

References: Wood (2017, §5.5)

We want to find

$$\underset{m}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m(\boldsymbol{x}_i)\}^2 + \lambda J_d(m), \tag{49}$$

where $J_d(m)$ is a smoothness penalty function,

$$J_d(m) = \int_{\mathbb{R}^p} \sum_{\boldsymbol{\alpha} \in \Lambda_d} \frac{d!}{\alpha_1! \cdots \alpha_p!} \left(\frac{\partial^d}{\partial x_1^{\alpha_1} \cdots \partial x_p^{\alpha_p}} m(\boldsymbol{x}) \right)^2 d\boldsymbol{x},$$

where $\Lambda_d = \{\boldsymbol{\alpha} \in \mathbb{R}^p : \alpha_1 + \ldots + \alpha_p = d\}$ and $2d > p$.

Remark. The null space of J_d, $\ker J_d = \{\phi : J_d(\phi) = 0\}$, is a set of polynomial functions. Therefore, increasing λ in (49) has the effect of increasing the penalization towards a linear model.

One can show that in the two-dimensional case the solution has the form

$$m(\boldsymbol{x}) = \sum_{j=1}^{D} \underbrace{\alpha_j \phi_j(\boldsymbol{x})}_{\text{polyn. basis}} + \sum_{i=1}^{n} \underbrace{\delta_i \eta_{dp}\big(\|\boldsymbol{x} - \boldsymbol{x}_i\|\big)}_{\text{isotropic penalization}}, \tag{50}$$

where ϕ_j are linearly independent polynomials of degree $p < d$.

Remark 1. From (50), we can see that the thin-plate splines are isotropic, that is, curvature in all directions is penalized equally and $J_d(m)$ is invariant under rotation of the coordinates. This makes sense when $m(\boldsymbol{x})$ is a function of covariates that are measured with the same measurement units.

Remark 2. Moreover, (50) allows us to obtain the thin-plate splines solution by using a penalized linear model approach with some constraints on the coefficients (Wood, 2017). In general, it has an $\mathcal{O}(n^3)$ computational cost which can be lowered to $\mathcal{O}(n^2 k)$ by using a rank-k truncated polynomial basis (Wood, 2003) and by applying a Lanczos iteration (Wood, 2017, §B.11). Implementations can be found in the R packages `gss`, `fields`, and `mgcv`.

Remark 3. We still obtain a linear smoother $\widehat{m}(\boldsymbol{x}) = S\boldsymbol{y}$ and therefore all discussions about effective degrees of freedom (35), bandwidth selection (43) etc...

10.1.3 Tensor product splines

Isotropic smooths assume that a unit change in one variable is equivalent to a unit change in another variable, in terms of function variability; when this is not the case, isotropic smooths can lead to poor models. A different approach to constructing multidimensional splines is using the tensor product basis.

Def. (Tensor product basis)
Suppose that $\{\phi_{1j},\ j = 1, \dots, K_1\}$ and $\{\phi_{2k},\ k = 1, \dots, K_2\}$ are a set of basis functions for x_1 and x_2, respectively. Then, the **tensor product basis** for $\boldsymbol{x} = (x_1, x_2)$ is defined as

$$\{\psi_{jk}(\boldsymbol{x}) = \phi_{1j}(x_1) \cdot \phi_{2k}(x_2),\ j = 1, \dots, K_1,\ k = 1, \dots, K_2\}.$$

Remark 1. A general function of $\boldsymbol{x}$ can be written in terms of the tensor product basis as

$$m(\boldsymbol{x}) = \sum_{j=1}^{K_1} \sum_{k=1}^{K_2} \beta_{jk} \phi_{1j}(x_1) \phi_{2k}(x_2).$$

Remark 2. By appropriately ordering the β_{jk} into a vector $\boldsymbol{\beta}$, we have that the model matrix X can be written in terms of the matrices X_1 and X_2 for the univariate smooth as

$$X = X_1 \odot X_2,$$

where $\odot$ is the row-wise Kronecker product.

Def. (Kronecker product)
The **Kronecker product** of $A_{m \times n}$ and $B_{p \times q}$ is the $pm \times qn$ block matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & \dots & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{pmatrix}$$

Def. (Row-wise Kronecker product)
The **row-wise Kronecker product** of $A_{m \times n}$ and $B_{m \times q}$ is a $m \times np$ matrix. If B_i denotes the i^{th} row of B, it is defined as

$$A \odot B = \begin{pmatrix} a_{11}B_1 & \dots & a_{1n}B_1 \\ \vdots & \dots & \vdots \\ a_{m1}B_m & \dots & a_{mn}B_m \end{pmatrix}$$

10.1.4 L-splines

(Miscellanea slides)

We want to define a particular null space, $\ker J_d(m)$, such that the minimizer of the penalty is a more complicated function. Specifically, we define a **functional** $\mathcal{L}$ that determines the

$$J(m) = \int \big(\mathcal{L}m(x)\big)^2 dx.$$

For instance, if we define

$$L = D^m + \sum_{j=0}^{m-1} w_j D^j,$$

where D^j is the j^{th} order differential operator, then we have that

$$(\mathcal{L}m)(x) = m^{(m)}(x) + \sum_{j=0}^{m-1} w_j(x) m^{(j)}(x),$$

where w_j are real-valued and continuous, and this operator defines a **model-based penalization**. This generates $\ker J_d$ which is
The null space corresponds to the eigenvectors of S (in $\widehat{m} = S\boldsymbol{y}$) which correspond to $\lambda_i = 1$. These eigenvectors are the functions that can be **reproduced exactly**.

Operator $\mathcal{L}$	Parametric family for $\mathcal{N}_0$
D^2	$\{1, x\}$
D^4	$\{1, x, x^2, x^3\}$
$D^2 + \gamma D, \quad \gamma \neq 0$	$\{1, \exp(-\gamma x)\}$
$D^4 + \omega^2 D, \quad \omega \neq 0$	$\{1, x, \cos(\omega x), \sin(\omega x)\}$
$(D^2 - \gamma D)(D^2 + \omega^2 D), \quad \gamma, \omega \neq 0$	$\{1, \exp(\gamma x), \cos(\omega x), \sin(\omega x)\}$
$D - w(\cdot)I, \quad w(x) \neq 0$	$\{\exp\left[\int_0^x w(u)du\right]\}$
$D^2 - w(\cdot)D, \quad w(x) \neq 0$	$\{1, \int_0^x \exp\left[\int_0^u w(v)dv\right] du\}$

Figure 30: Operators $\mathcal{L}$ and the corresponding parametric family for $\ker \mathcal{L}$.

Remark. From Figure 30, we observe that by using $\mathcal{L} = D^4 + \omega^2 D$, we can apply L-splines to periodic time series in order to preserve seasonal components.

Remark. This trick allows the definition of multivariate splines such as

$$\psi_{jkl}(x_1, x_2, x_3) = \phi_{1j}(x_1)\psi_{kl}(x_2, x_3),$$

where $\psi_{kl}(x_2, x_3)$ is a thin-plate spline which has some properties in terms of reproduced functions (Figure 30).

10.2 Additive models

In additive models we assume that the response is linear in the predictors effects and that there is an additive error, so that we can study the effect of each predictor separately. The model is

$$Y_i = \sum_{j=1}^{p} m_j(x_j) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \tag{51}$$

where each m_j is a smooth function of the j^{th} covariate.

Remark. No matter the dimension of the covariates, the regression surface $m(\boldsymbol{x})$ can be interpreted by inspecting each function marginally (Figure 31).

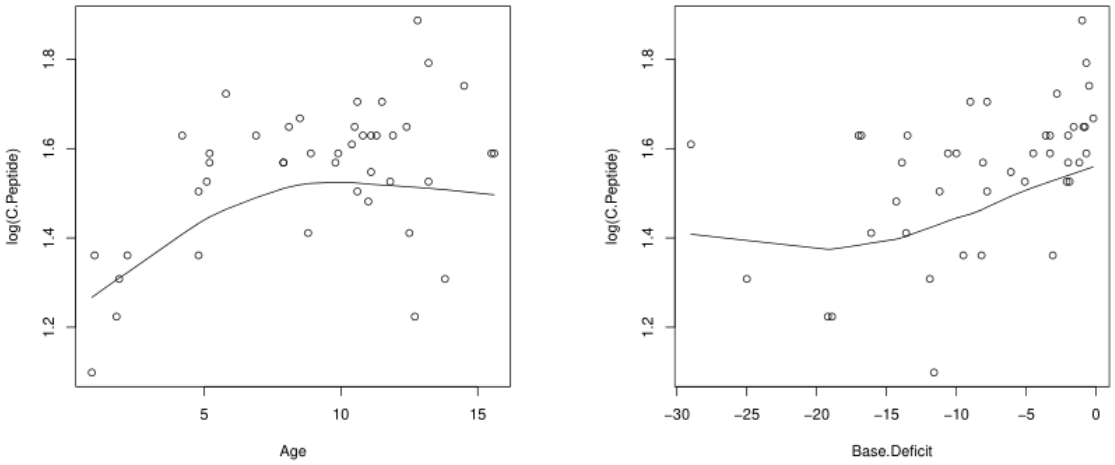
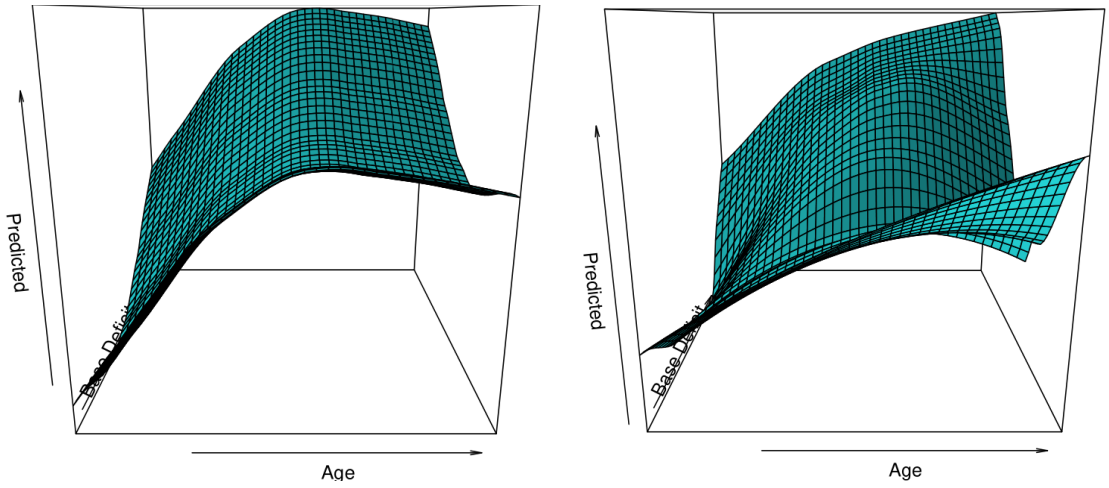


Figure 31: marginalEffectAdditiveModel



(a) Additive model.

(b) Fully nonparametric model.

Figure 32: Comparison between an additive model and a complete nonparametric model on the same dataset.

10.2.1 Estimation

The estimation procedure is based on the **backfitting algorithm** of Hastie et al. (2013), which is summarized in Algorithm 5. The idea behind backfitting is that

$$\mathbb{E}[Y - \widehat{\alpha} - \sum_{j=1}^{p-1} \widehat{m}_j(x_j)|x_p] \approx m_p(x_p),$$

and therefore the partial residuals with respect to all $p-1$ covariates are approximately

$$\widehat{\varepsilon}_i \approx m_p(x_p) + \delta_i, \quad \delta_i \overset{\text{iid}}{\sim} \text{WN}(0, \sigma_\delta^2).$$

The marginal regression function can be estimated by any smoothing technique that we have reviewed until now.

Algorithm 5 Backfitting

Input: $\varepsilon > 0$ convergence criterion.

Init:
  $\hat{\alpha} = \sum_{i=1}^{n} y_i/n$
  $\hat{f}_j = 0$ per ogni $j = 1, \dots, p$
  $\text{tol} = \varepsilon + 1$
while $\text{tol} > \varepsilon$ **do**
  for $j = 1, \dots, p$ **do**
    $\hat{f}_j = \texttt{smooth}\big(y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})\big)$
    $\hat{f}_j = \hat{f}_j - n^{-1} \sum_{i=1}^{n} \hat{f}_j(x_{ij})$ ▷ $\hat{f}$ has zero mean, but this stabilizes rounding errors
  end for
  $\text{tol} = \texttt{distance}(\widehat{f}^{(new)}, \widehat{f}^{(old)})$.
end while

Remark 1. Multiple theoretical justifications can be raised for the backfitting algorithm, for instance by considering the univariate regression

$$\sum_{i=1}^{n} \Big\{ Y_i - \sum_{j=1}^{p} m_j(x_{ij}) \Big\}^2 + \sum_{j=1}^{p} g l_j \int \big\{ m_j''(t) \big\}^2 dt.$$

The solution to the above problem is given by

$$\widehat{m}_j = (I + \lambda_j P_j)^{-1} \big(\boldsymbol{y} - \sum_{k \neq j} \widehat{m}_k \big), \quad j = 1, \dots, p.$$

Therefore, by stacking the operator $S_j = (I + \lambda_j P_j)^{-1}$ in matrix notation, we can write

$$\begin{pmatrix} I & S_1 & \dots & S_1 \\ S_2 & I & \dots & S_2 \\ \dots & \dots & \ddots & \dots \\ S_p & S_p & \dots & I \end{pmatrix} \begin{pmatrix} m_1 \\ \vdots \\ m_p \end{pmatrix} = \begin{pmatrix} S_1 \boldsymbol{y} \\ \vdots \\ S_p \boldsymbol{y} \end{pmatrix},$$

which can be solved by applying the **Gauss-Seidel algorithm**. This has been shown by Buja et al. (1989) to be equivalent to solving the back-fitting algorithm, hence providing a theoretical justification.

Remark 2. The use of the backfitting algorithm allows fitting models which have mixed type of smoothing procedures. Moreover, some components might be parametric, nonparametric, and also include multivariate smooth functions to model nonparametric interactions of terms.

10.2.2 Projection pursuit

The projection pursuit idea (Friedman and Tukey, 1974) came out of wanting to do nonparametrics in higher dimensions, while still keeping the complexity at an acceptable level. The main idea is that

we form an additive model in $\boldsymbol{x}$, but using univariate nonparametric smoothers of linear combinations (**projections**) of the covariates,

$$Y_i = \sum_{j=1}^{K} f_j(\boldsymbol{\alpha}_j^\top \boldsymbol{x}) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2). \tag{52}$$

Remark 1. The additive model (51) is a special case of (52) when $\boldsymbol{\alpha}_j = (0\ 0\ \cdots\ 1\ 0\ \cdots\ 0)$.

Remark 2. The f_j's can be estimated via backfitting, although the weights α_j are generally only fit once.

References

Buja, A. et al. (1989). «Linear Smoothers and Additive Models». In: *The Annals of Statistics* 17.2, 453–510.

Dimatteo, I. et al. (2001). «Bayesian Curve-fitting with Free-knot Splines». In: *Biometrika* 88.4, 1055–1071.

Eilers, P. H. C. and Marx, B. D. (1996). «Flexible Smoothing with B-splines and Penalties». In: *Statistical Science* 11.2, 89–121.

Friedman, J. and Tukey, J. (1974). «A Projection Pursuit Algorithm for Exploratory Data Analysis». In: *IEEE Transactions on Computers* C-23.9, 881–890.

Hall, P. (1987). «On Kullback-Leibler Loss and Density Estimation». In: *The Annals of Statistics* 15.4, 1491–1519.

Hastie, T. et al. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer Nature.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective.* Singapore: World Scientific Pub.

Ramsay, J. O. (1988). «Monotone Regression Splines in Action». In: *Statistical Science* 3.4, 425–441.

Ruppert, D. et al. (1995). «An Effective Bandwidth Selector for Local Least Squares Regression». In: *Journal of the American Statistical Association* 90.432, 1257–1270.

Ruppert, D. (1997). «Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation». In: *Journal of the American Statistical Association* 92.439, 1049–1062.

Wasserman, L. (2005). *All of Nonparametric Statistics.* New York: Springer.

Wood, S. N. (2003). «Thin Plate Regression Splines». In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.1, 95–114.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition.* 2° edizione. Boca Raton: Chapman and Hall/CRC.

Part II

Multilevel models

Instructor: Stefano Mazzuco

The literature on hierarchical models is quite wide and it can be difficult to present a unified view of the topic. For the most part, we will follow the exposition of
We will discuss the motivation of introducing a "random" coefficient, how these model work and the advantages and disadvantages—if any—with respect to more classical approaches. Specifically, we will talk about different aspects of these models:

- data structure;
- inference, from both a frequentist and Bayesian point of view;
- connection between nonparametric models (smoothing) and mixed models.

Lecture 11: Hierarchical modelling

2022-03-30

11.1 Hierarchical structures

We will start from the idea of hierarchical data, which is the classical justification for applying multilevel models. Consider the `sleepstudy` dataset, in which we are given repeated measurements related to the same individual (Table 1).

	Reaction	Days	Subject
1	249.5600	0	308
2	258.7047	1	308
3	250.8006	2	308
4	321.4398	3	308
5	356.8519	4	308
6	414.6901	5	308

Table 1: sleepStudyDatasetTable

Idea. We might think that multiple observations within subject are correlated with each other, and would like to model this relationship.

From Figure 33 we can observe that there is a complex structure of dependence between each individual subject when compared to the overall pooled model.

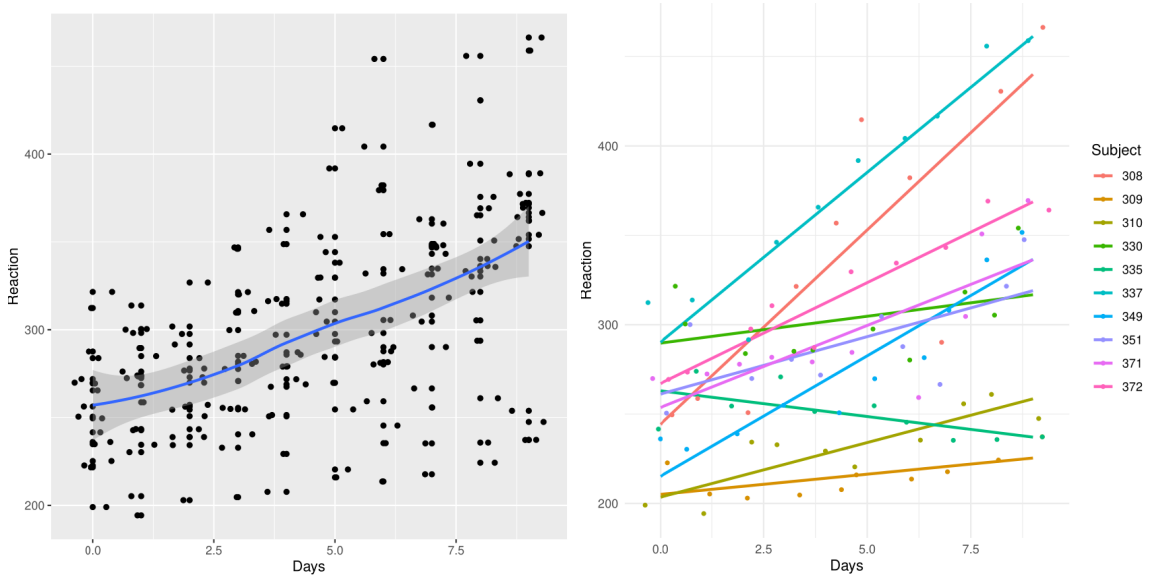


Figure 33: hierarchicalModelNoHierarchyExample

Hierarchical structure might be observed in different ways:

a) Longitudinal studies, **repeated observations** hierarchy.

b) **Multistage** structure, such as country → region → city.

c) **Imperfect hierarchies** which are linked together but not embedded (e.g. multiple observations for the same combination of student and teacher)

The second key idea in multilevel models is to **borrow information** between groups to improve estimation when the sample size within a specific group is not large enough to provide accurate information (Figure 34).

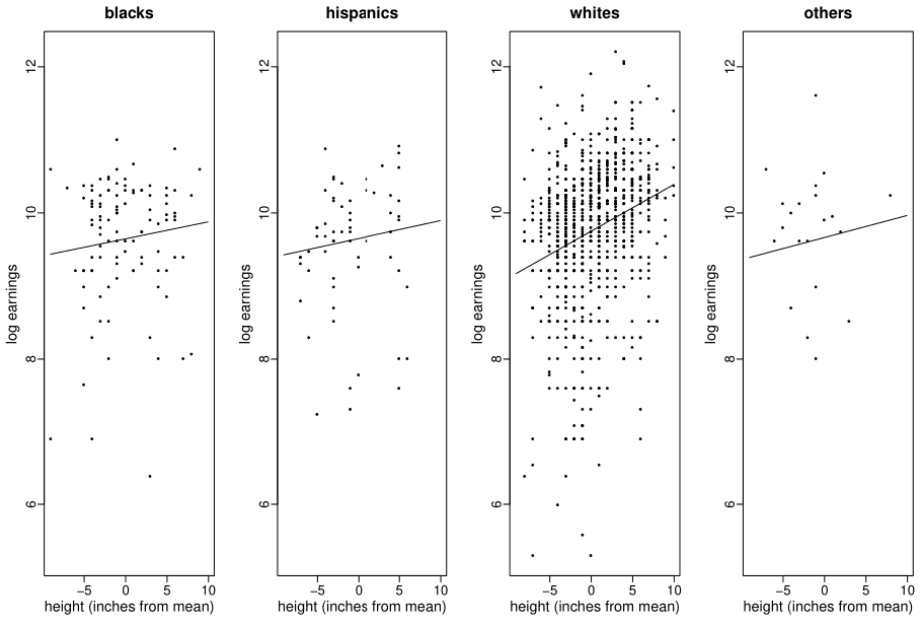


Figure 34: Example of the need of borrowing information between groups with high sample sizes and groups with low sample sizes.

Remark. In the case represented in Figure 34, there is no hierarchical structure within the groups. We are interested in the "random effect" specification only for improving estimation between each group (Gelman and Hill, 2007).

11.2 Hierarchical linear model

We define the basic objects of the regression analysis, by specifying a conditional regression

$$\boldsymbol{Y}|\boldsymbol{B} = \boldsymbol{b} \sim \mathcal{N}(Z\boldsymbol{b} + X\boldsymbol{\beta}, \sigma^2 I_n). \tag{53}$$

In the above equation, the $\boldsymbol{B} \in \mathbb{R}^q$'s are called **random effects** and are given a prior distribution that is usually centered in zero,

$$\boldsymbol{B} \sim \mathcal{N}(0, \Sigma_\vartheta),$$

where Σ_ϑ is a positive-definite matrix which can be decomposed as

$$\Sigma_\vartheta = \Lambda_\vartheta \Lambda_\vartheta^\top.$$

Using the above decomposition, we can rewrite the random effect term as

$$\boldsymbol{B} = \Lambda_\vartheta \boldsymbol{U}, \quad \boldsymbol{U} \sim \mathcal{N}(0, I_q).$$

The model in (53) can be estimated via **penalized least squares** (PLS), which amounts to minimizing the following quantity,

$$(\boldsymbol{\beta}, \boldsymbol{u}) = \underset{\boldsymbol{\beta},\boldsymbol{u}}{\operatorname{argmin}} \underbrace{\|\boldsymbol{y} - X\boldsymbol{\beta} - Z\Lambda_\vartheta \boldsymbol{u}\|_2^2}_{\text{sum sq. residuals}} + \underbrace{\|\boldsymbol{u}\|_2^2}_{\text{penalty}}$$

which is equivalent to calculating the solution to

$$(\boldsymbol{\beta}, \boldsymbol{u}) = \underset{\boldsymbol{u}}{\operatorname{argmin}} \left\| \begin{pmatrix} \boldsymbol{y} - X\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} Z\Lambda_\vartheta \\ I_q \end{pmatrix} \boldsymbol{u} \right\|_2^2.$$

Since $U_\vartheta = Z\Lambda_\vartheta$, the conditional mean satisfies the condition

$$[U_\vartheta U_\vartheta^\top + I_q]\mu_{\boldsymbol{U}|\boldsymbol{Y}} = U_\vartheta^\top (\boldsymbol{Y} - X\boldsymbol{\beta}).$$

Both Z and U are usually **sparse matrices**, and therefore calculations are usually sped up by applying a Cholesky decomposition to the original matrix.

Example (Linear model)
Consider a linear model where we are interested in minimizing

$$\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2,$$

then if we can apply a Cholesky decomposition $X^\top X = R^\top R$ with R upper triangular, then we have that

$$(X^\top X)^{-1} X^\top y = R^{-1}(R^{-1})^\top R^\top y,$$

and the quantity R^{-1} is straightforward to calculate since it is upper triangular. This might speed up computations when the data is abundant ($n \gg p$), since the Cholesky decomposition has complexity $\mathcal{O}\big(p^3 + (np^2)/2\big)$.

11.2.1 Likelihood

The model in (53) is defined from the **conditional distribution** of $\boldsymbol{Y}|\boldsymbol{B}$. However, in practice $\boldsymbol{y}$ is observed while the random effect $\boldsymbol{b}$ is unknown, and we want to estimate the conditional distribution $\boldsymbol{U}|\boldsymbol{Y} = \boldsymbol{y}$. From straightforward calculations, we can write the joint distribution of $(\boldsymbol{Y}, \boldsymbol{U})$ as

$$f_{\boldsymbol{Y},\boldsymbol{U}}(\boldsymbol{y}, \boldsymbol{u}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\|\boldsymbol{y} - X\boldsymbol{\beta} - U\boldsymbol{u}\|^2} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\|\boldsymbol{u}\|^2},$$

hence the conditional distribution is Gaussian and the conditional mean is

$$\mu_{\boldsymbol{U}|\boldsymbol{Y}=\boldsymbol{y}} = \underset{u}{\operatorname{argmin}} \left\{ \|\boldsymbol{y} - X\boldsymbol{\beta} - U(\theta)\boldsymbol{u}\|^2 + \|u\|^2 \right\}.$$

We can derive the profile deviance

$$-2\ell(\boldsymbol{\vartheta}, \boldsymbol{\beta}|\boldsymbol{y}) = \log L(\boldsymbol{\vartheta})^2 + n\left(1 + \log \frac{2\pi r^2(\boldsymbol{\vartheta}, \boldsymbol{\beta})}{n}\right).$$

Unfortunately, the above deviance depends on σ^2, and thus we need to minimize it in order to evaluate the conditional estimate

$$\widehat{\sigma}^2 = \frac{r^2(\theta, \beta)}{n},$$

so that we can derive the profile deviance

$$-2\tilde{\ell}(\theta, \beta|\boldsymbol{y}) = \log|L(\theta)|^2 + n\left\{1 + \log\left(\frac{2\pi r^2(\theta, \beta)}{n}\right)\right\}.$$

The estimate of σ^2 might be seriously biased, and we need to take into account the loss of degrees of freedom when estimating it. The trick is reducing the likelihood function by using the **restricted maximum likelihood** (REML) estimate of

$$\widehat{\boldsymbol{\vartheta}} = \underset{\boldsymbol{\vartheta}}{\operatorname{argmax}}\, L_R(\boldsymbol{\vartheta}, \sigma^2|\boldsymbol{y}) = \underset{\boldsymbol{\vartheta}}{\operatorname{argmax}} \int L(\boldsymbol{\vartheta}, \boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) d\boldsymbol{\beta},$$

The profiled REML deviance is

$$-2\tilde{\ell}_R(\theta) = \log|L|^2 + \log|R_x|^2 + (n-p)\left\{1 + \log\left(\frac{2\pi r^2(\theta)}{n-p}\right)\right\},$$

where R_x is the $p \times p$ matrix used to extend the Cholesky decomposition.

Problem. When using REML estimates, we cannot apply likelihood ratio tests and AIC/BIC procedures to compare models.

11.2.2 Borrowing of information

Consider the simplest linear mixed model,

$$y_{ij} = \beta_{0j} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_\varepsilon), \quad \beta_{0j} \sim \mathcal{N}(0, \sigma^2_\beta),$$

then we can explicitly calculate the estimates based on an **Empirical Bayes** (EB) procedure,

$$\widehat{\beta}_{0j} = \frac{\sigma^2_\beta}{\sigma^2_\beta + \sigma^2_\varepsilon/n_j}\bar{y}_{\cdot j} + \frac{\sigma^2_\varepsilon/n_j}{\sigma^2_\beta + \sigma^2_\varepsilon/n_j}\bar{y}_{\cdot\cdot} = \frac{\sigma^2_\beta \bar{y}_{\cdot j} + \frac{\sigma^2_\varepsilon}{n_j}\bar{y}_{\cdot\cdot}}{\sigma^2_\beta + \sigma^2_\varepsilon/n_j}. \tag{54}$$

We can see that (54) is a weighted average between the total mean $\bar{y}_{\cdot\cdot}$ and the group average $\bar{y}_{\cdot j}$, weighted by the strength of the group variance σ^2_β with respect to the noise variance σ^2_ε. In particular, the higher the sample size n_j, the higher importance is given to the specific group mean $\bar{y}_{\cdot j}$.

Remark. Using (54) prevents the unrepresented groups from having wildly different means, which could be unrealistic and a possible consequence of the high uncertainty on the estimate.

Bayes. The empirical Bayes approach in (54) is simply a frequentist approximation of the full Bayesian model

$$Y_i|\mu \sim \mathcal{N}(\mu, \sigma^2_0)$$

$$\mu \sim \mathcal{N}(\mu_0, \tau^2_0)$$

which results in the posterior mean

$$\mathbb{E}[\mu|\boldsymbol{y}] = \frac{\tau_0^2 \bar{y} + \frac{\sigma_0^2}{n}\mu_0}{\tau_0^2 + \frac{\sigma_0^2}{n}}$$

Pooling. The model can be described as an intermediate model between:

- the **complete-pooling** model, where $y_{ij} = \mu + \varepsilon_{ij}$;
- the **no-pooling** model, $y_{ij} = \mu + \beta_j + \varepsilon_{ij}$ with β_j "fixed".

The strength of the pooling depends on how informative the j^{th} group is with respect to the other groups.

Overfitting. The above distinction can also be phrased in terms of finding a balance between **overfitting** with a no-pooling model and **underfitting** with a complete-pooling model.

11.3 Generalized linear mixed models

References: Berger et al. (1999)

We can extend the LMM framework in (53) to the **generalized linear mixed model** (GLMM) framework for a dispersion family by specifying

$$g\big(\mathbb{E}[Y|U=u]\big) = X\boldsymbol{\beta} + Z\Lambda\boldsymbol{u},$$

where g is a **link function** that maps the linear predictor onto the support of $\mathbb{E}[Y]$ (Pace and Salvan, 1997).

Attention. Unlike standard linear models, there are some differences when using generalized linear models. In this case, the expected value of Y is not β.

Inference can be carried out by numerically maximizing the **marginal likelihood**,

$$L(\vartheta, \beta|\boldsymbol{y}) = \int_{\mathbb{R}^q} f(\boldsymbol{y}|\boldsymbol{u}) f(\boldsymbol{u}) d\boldsymbol{u}, \tag{55}$$

which can be carried out using a two-step procedure:

1. First, by determining the conditional mode

$$\tilde{\boldsymbol{u}}(\boldsymbol{y}|\vartheta, \beta) = \operatorname*{argmax}_{\boldsymbol{u}} f(\boldsymbol{y}|\boldsymbol{u}) f(\boldsymbol{u}). \tag{56}$$

2. Conditionally on the value of $\tilde{\boldsymbol{u}}$, the deviance can be calculated as

$$D(\beta, \vartheta|\boldsymbol{y}) = D_{\text{GLM}}(\boldsymbol{y}, \mu(\tilde{\boldsymbol{u}})) + \|\tilde{\boldsymbol{u}}\|_2^2 + \log|\boldsymbol{L}|^2, \tag{57}$$

where $\boldsymbol{L}$ is the Cholesky factor. Then, (57) has to be maximized in order to obtain the estimate for the fixed effects β and the variance component θ.

Calculation of the mode In order to compute (56), we can minimize the weighted residual sum of squares,

$$\tilde{\boldsymbol{u}}(\vartheta, \beta) = \underset{\boldsymbol{u}}{\operatorname{argmin}} \left\| \begin{pmatrix} W^{1/2}(\mu)(\boldsymbol{y} - \mu(\boldsymbol{u})) \\ -\boldsymbol{u} \end{pmatrix} \right\|_2^2,$$

where $W(\mu)$ is a diagonal matrix with weights given by the inverse of the variance of Y. This estimation method is called **penalized iteratively reweighted least squares** (PIRLS). Note that estimates for generalized linear models (without random effects) are usually IRLS (without penalization).

Maximization of the deviance The deviance in (57) can be maximized via **Laplace approximation** (Pace and Salvan, 1997), which is faster but not accurate, or via **Gauss-Hermite quadrature**, which is slower but more accurate. Since numerical quadrature scales poorly with the dimensionality of the parameter space, the Laplace approximation performs better when the number of coefficients and random effects is large.

Another possibility would be to use penalized likelihood, which yields biased estimates for large values of standard deviation of the parameters. In general, this choice is nowadays deprecated.

Example (Covid-19 mortality)

We consider the Covid-19 outbreak and its impact on the "excess mortality", i.e. the difference between the observed number of deaths and the expected one, in the case there had not been any outbreak.

Naive approach. But how do we define the expected number of deaths? The naive method would be by taking the past five-years average (2015–2019), but the problem is that the trend in mortality is not taken into account when applying this procedure.

Mixed model approach. We can try using a longitudinal model to estimate both the baseline mortality, and the possible observed excess mortality.

Lecture 12: Multilevel models (ii)

2022-04-01

12.1 Residual analysis

We now consider a linear mixed model of the form

$$\boldsymbol{Y} = X\boldsymbol{\beta} + Z\boldsymbol{b} + \varepsilon,$$

and we observe that there are multiple sources of variability, $(\boldsymbol{b}, \varepsilon)$ that contribute to the overall variance of the observed data. Therefore, we have different types of residuals that can be inspected:

1. Level 1 residuals: $\varepsilon_i = y_i - x_i^\top \widehat{\boldsymbol{\beta}} - z_i^\top \widehat{\boldsymbol{b}}$
2. Level 2 residuals: $z_i^\top \widehat{\boldsymbol{b}}$
3. Composite residuals: $y_i - x_i^\top \widehat{\boldsymbol{\beta}} = Z\widehat{\boldsymbol{b}_i} + \varepsilon_i$

Problem. These residuals are interrelated among them, and ignoring this correlation may lead to some **confounding** issues in diagnosing model deficiencies. Specifically, it is possible that issues in one type of residuals might affect also other residuals.

Solution. The recommendation in this case is to perform a upward residual analysis by checking first level–1 residuals—which are not confounded by the others—then level–2 residuals, and finally the composite residuals.. Going the other way round may provide misleading evidence of model deficiencies.

12.1.1 Level-1 residuals

References: Christensen et al. (1992)

In order to check for **homogeneity of residual variance** across groups, we can use a test statistic based on

$$d_i = \frac{\log s_i^2 - \sum_{i=1}^n (n_i - r_i) \log s_i^2 / \sum_{i=1}^n (n_i - r_i)}{\sqrt{2/(n_i - r_i)}},$$

where s_i^2 is the residual variance within each group based on separate ordinary least squares regressions and r_i is the rank of the corresponding model matrix. Therefore, the statistic $H = \sum_{i=1}^n d_i^2$ has an approximate $\chi^2_{g^*-1}$ distribution, where g^* is the number of groups after having disregarded the small ones (n < 10).

12.1.2 Level-2 residulas

Through the use of level-2 residuals, we can check

a) whether **additional explanatory variables** can contribute significantly to the model;

b) the assumption of **linearity** of the level–2 explanatory variables;

c) the assumption of **normality** of the whether the level–2 residuals.

12.1.3 Influence analysis

References: Demidenko and Stukel (2005)

Another standard tool in the analysis of residuals is **influence analysis**, that is, the systematic investigation of whether some particular observations are given disproportionate importance in model estimation.

Deletion. The easiest way in linear models to assess this influence is by removing a subset of observations and observing the resulting change in the estimates.

In a standard linear model we assume independents units, whereas in a hierarchical model we assume another type of dependence. In these cases, we might assess the influence (Cook's distance) for both level-1 and level-2 units to the resulting estimates, although we cannot rely on large sample asymptotic results.

We can calculate how influential can be a (level 1 or level 2) unit with respect to the precision of the fixed effects estimates. For distance, we can use the following individual statistics,

$$\text{CovTrace}(\beta)_i = \left|\text{tr}\left\{\widehat{\mathbb{V}}(\widehat{\beta})^{-1} - \widehat{\mathbb{V}}(\widehat{\beta}_{-i}) - p\right\}\right|$$

$$\text{CovRatio}(\beta)_i = \det \widehat{\mathbb{V}}(\widehat{\beta}_{-i}) \cdot \det \widehat{\mathbb{V}}(\widehat{\beta})^{-1}$$

where $\widehat{\mathbb{V}}[\widehat{\beta}_{-i}]$ is the covariance matrix of $\widehat{\beta}$ estimated without the i^{th} unit.

$$i^{\text{th}} \text{ unit not influential } \implies \text{CovTrace} \approx 0, \text{CovRatio} \approx 1.$$

These estimates are computationally not easy to to obtain, although parametric bootstraps might reduce the computational cost. Another less demanding statistic is the **relative variance change** given by

$$\text{RVC}_i = \frac{\widehat{\vartheta}_l^{(-i)}}{\widehat{\vartheta}_l} - 1$$

where $\widehat{\vartheta}_l^{(-i)}$ is the estimate of the l^{th} variance component without the i^{th} unit.

12.1.4 Leverage

Leverage points are points which fitted values that are unusually far away from the observed ones, and in a linear model they are usually measured by the rate of change

$$h_i = \frac{\partial \widehat{y}_i}{\partial y_i} = x_i^\top (X^\top X)^{-1} x_i = H_{ii},$$

where $H = X(X^\top X^{-1})X$. In a linear mixed model, the **leverage score** becomes

$$H_i = \underbrace{X_i (X_i^\top V_i^{-1} X_i)^{-1} X_i V_i^{-1}}_{H_{1i}} + \underbrace{Z_i D Z_i^\top V_i^{-1} (I - H_{1i})}_{H_{2i}},$$

which can be simplified in $H_i = H_{1i} + H_{2i}$.

Remark. H_{2i} is confounded by H_{1i}, and the two components might not be separated.

Reference. See R package `HLMdiag` for diagnostics in linear mixed models.

12.1.5 Randomized quantile residuals

References: Dunn and Smyth (1996)

Although residuals for GLMs are harder than in LMs, reweighted and Pearson residuals still carry useful information about the model. Dunn and Smyth (1996) suggested applying a bootstrap-based approach for constructing standardized residuals (`DHARMa` package). For each y_i, a fitted value is simulated conditionally to X_i in order to obtain an empirical cumulative distribution function. The bootstrap can either be **parametric** or **nonparametric**.

Idea. If the model is correctly specified, we expect that all values of the cumulative distribution function should appear with equal probability. That is, the distribution of the observed quantiles should be approximately Unif$(0, 1)$. This approach is similar to Bayesian p-values, and using a nonparametric bootstrap might be useful in specific situations such as biased estimators (ridge regression).

Remark. Dunn and Smyth (1996) also provide some examples in which randomized quantile residuals give visual evidence of a misspecified model, whereas ordinary (deviance or Pearson) residuals fail to do so.

Conditionality. Using this approach, we can also decide whether to simulate all stochastic levels (**unconditional simulation**) or simulate only the unit level (**conditional simulation**). Unconditional simulation allows testing the model as a whole, at the cost of adding more variability and thus reducing the power of the test.

Lecture 13: Generalised Estimating Equations

2022-04-13

Generalised Estimating Equations (GEE) is a method which is usually employed for longitudinal data, especially when the distribution of Y is not normal (e.g. binary or count data). Theoretically, mixed models can be used for longitudinal data: individuals i are observed j times, leading to

$$Y_{ij} = \alpha_i + \beta^\top x_{ij} + \varepsilon_{ij}, \tag{58}$$

with

$$\alpha_i \sim \mathcal{N}(0, \sigma^2_\alpha), \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_\varepsilon), \quad \alpha_i \perp\!\!\!\perp \varepsilon_{ij}.$$

Again, when using a GLM, we have a link between the predictors and the mean of the distribution.

Remark. In the linear model (58), we have that for each unit i,

$$\operatorname{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\varepsilon},$$

which is the **intraclass correlation coefficient**. When the response variable is binary, such correlation is more complicated to define, although the important aspect is that we assume a constant correlation between intra-individual observations.

Instead of using random effects to define the correlation structure, we can separately model the dependence of response with covariates and the within–unit correlation structure. We do so by defining a **marginal model** for the response Y_{ij}.

13.1 Marginal models

A marginal model focuses on the population average after integrating out the random effects, that is, we assume that

$$g\big(\mathbb{E}[Y_{ij}]\big) = x_i^\top \beta.$$

On the other hand, we explicitly model the marginal variance using the intraclass correlation as a function of μ_{ij},

$$\mathbb{V}[Y_{ij}] = v(\mu_{ij}) \cdot \phi,$$

by introducing an additional parameter α to be estimated.

Remark. In a marginal model, we do not estimate the random intercepts and instead try to estimate the intraclass correlation parameter α so that the model specification is equivalent to some random effect model. We can also obtain a population level model from a mixed effect model, although it is an average of the subject-specific quantities.

Several structure of covariance can be employed in order to complicate the structure:

- **Independence:** $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

- **Exchangeability**: $\begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$
- **Order-1 autoregressive**: $\begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$
- **Unstructured**: $\begin{pmatrix} 1 & \rho_{12} & \rho_{23} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{pmatrix}$

When the dependence structure becomes more complex, such as the unstructured case, maximizing the likelihood is extremely expensive. However, GEE procedures avoid this bottleneck: **structural parameters** (the ρ_{ij}'s in the above cases) are estimated separately from regression ones (β).

The idea of GEE comes from the **quasi-likelihood** approach, specifically by estimating a parameter ϑ when solving a score equation of the type

$$\sum_{i=1}^{n} \phi(\vartheta|y_i) = 0,$$

and the resulting estimate $\widehat{\vartheta}$ is termed **M-estimate**.

- This approach has a main advantage: it avoids to deal with intractable likelihood functions.
- Moreover, might be also more robust, since different models can have the same score function. Therefore, M-estimates might hold for a larger set of models than with maximum likelihood approaches.
- At the same time, we have some disadvantages: AIC/BIC and likelihood ratio tests are unavailable, and parameter estimates are less efficient.

Example (Quasi-likelihood in GLMs)
In a GLM, we directly model the mean and variance of the response variable,

$$\mathbb{E}[Y_i] = \mu_i(\beta)$$
$$\mathbb{V}[Y_i] = \phi v(\mu_i).$$

Since the estimating equation is

$$\ell'(\vartheta) = 0 \iff \frac{1}{\phi} \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} \frac{1}{v(\mu_i)} (y_i - \mu_i) = 0,$$

we observe that the estimating equation only depends through the mean μ_i and variance $v(\mu_i)$ of the model.

Remark. Likelihood theory also hold for quasi-likelihood models, specifically we can prove an asymptotic distribution of the form

$$(\widehat{\beta} - \beta)(X^\top W^{-1} X)^{-1}(\widehat{\beta} - \beta) \dot{\sim} \chi_p,$$

under standard regularity assumptions.

Example (Logistic regression)
In a logistic regression , Liang and Zeger (1986) suggest a moment estimator for α,

$$\widehat{\alpha} = \frac{1}{\tilde{N}} \sum_{i=1}^{m} \sum_{j<k} r_{ij} r_{jk},$$

where r_{ij} are the Pearson residuals and $\tilde{N} = \ldots$.

Remark. Estimate is performed through iteratively reweighted least squares, where the weight $V(Y_i)^{-1}$ needs to be estimated from the values of β.

Remark. In a GEE we assume a specific correlation structure, and there are two solutions for calculating the standard error:

1. **Naive**: assuming that the correlation structure is correct, the standard error are correct and the estimates of β are consistent.
2. **Huber-White**: sandwich estimators of the standard error, which are more robust to misspecification.

In general, for a linear mixed model we have that

$$\mu_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_i,$$

and therefore

$$\mathbb{E}[\mu_{ij}] = \beta_0 + \beta_1 x_{ij} + \int \alpha_i dF(\alpha_i) = \beta_0 + \beta_1 x_{ij}.$$

Hence, the fixed effects are the population averages in the marginal model whereas μ_{ij} is the expected value in the subject-specific model. For a generalized linear mixed model, such interpretation is not straightforward since

Diagnostics. Agresti (2002) suggests that GEE estimates are valid even if the correlation structured is misspecified. In general, we can start by using an exchangeable structure and checking how the coefficient estimates change with other correlation structures.

Lecture 14: Bayesian hierarchical models

2022-04-20

Although the Bayesian perspective is different from the frequentist approach, we can discard these differences to focus on the practical advantages that we have in the hierarchical modelling framework. Indeed, as the hierarchical model becomes more complex we have seen that inference becomes difficult, leading us to use different solutions to **regularize** the likelihood function (penalized likelihoods, GEE, EM, REML).

14.1 Bayesian inference

The Bayesian approach to inference provides numerous advantage since "regularization" is naturally provided by the prior distribution. The starting point is associating to each parameter some initial **prior distribution** $\pi(\vartheta)$, whose choice depends on

a) the prior information that we have on the parameters before running the model, or;

b) the required strength of regularization for estimating the model.

The Bayesian inference is based on the posterior distribution of the parameter,

$$\pi(\boldsymbol{\vartheta}|\boldsymbol{y}) = \frac{L(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})}{\int_\Theta L(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})d\boldsymbol{\vartheta}} = \frac{L(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})}{L(\boldsymbol{y})}, \tag{59}$$

where $L(\boldsymbol{y}|\vartheta)$ is the likelihood of the data conditional on ϑ and $L(\boldsymbol{y})$ is the marginal likelihood. The main issue for several year has been that the integral at denominator of (59) usually does not have an analytical solution (except for particular conjugate cases).

However, in GLMs no conjugate prior is available. We could potentially rely on asymptotic normal approximation but the result might be too inaccurate for actual inference. This is especially the case when the posterior distribution is highly skewed, such as in logistic regression with high number of successes or failures.

Other approximation methods are based on ABC and INLA.

14.2 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) provides an approach for generating samples from the posterior distribution. With this procedure, in the limit as the samples $\to \infty$ we are exactly sampling from the posterior distribution of the parameters. This way, we obtain an "empirical" summary of the distribution: posterior mean, variance, percentiles, empirical pdf and empirical cdf.

Moreover, by simply applying a transformation g to the sampled values $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_B$, we obtain the posterior distribution of any functional $g(\boldsymbol{\vartheta})$.

Let $\boldsymbol{\vartheta}^t$ be the vector of parameters at the t^{th} iteration, with a fixed starting value for ϑ^0. MCMC generates a new value of $\boldsymbol{\vartheta}^t$ that depends on the data and $\boldsymbol{\vartheta}^{t-1}$ via a Markov Chain,

$$\mathbb{P}(\boldsymbol{\vartheta}^{t+1} \in A|\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_t) = \mathbb{P}(\boldsymbol{\vartheta}^{t+1} \in A|\boldsymbol{\vartheta}^t).$$

The goal of the constructed stochastic process is for the chain to be **ergodic** with the target invariant distribution given by the posterior distribution $\pi(\boldsymbol{\vartheta}|\boldsymbol{y})$. For a discrete chain X_t with support on $S = \{1, 2, \ldots, n\}$, this can be represented by saying that

$$\lim_{t\to\infty} P_j^t \pi_j = \pi(j|\boldsymbol{y}), \quad \text{for all } j \in S.$$

A more general extension can be found for continuous state spaces, albeit it necessitates a more technical discussion.

Remark. Note that the starting value $\boldsymbol{\vartheta}^0$ does not affect the limit distribution of the Markov Chain. However, randomizing the starting point might help with convergence issues.

14.2.1 Gibbs sampling

Gibbs sampling (GS) is a MCMC algorithm that proposes new values of $\boldsymbol{\vartheta}^t$ by sampling from each marginal conditional distribution. That is, for simulating a new value $\boldsymbol{\vartheta}^t$ we sample

$$\begin{aligned}
&\pi(\boldsymbol{\vartheta}_1^t|\boldsymbol{\vartheta}_2^{t-1}, \ldots, \boldsymbol{\vartheta}_p^{t-1}, \boldsymbol{y}) \\
&\pi(\boldsymbol{\vartheta}_2^t|\boldsymbol{\vartheta}_1^t, \boldsymbol{\vartheta}_3^{t-1}, \ldots, \boldsymbol{\vartheta}_p^{t-1}, \boldsymbol{y}) \\
&\quad \vdots \\
&\pi(\boldsymbol{\vartheta}_p^t|\boldsymbol{\vartheta}_1^t, \boldsymbol{\vartheta}_2^t, \ldots, \boldsymbol{\vartheta}_{p-1}^t, \boldsymbol{y})
\end{aligned}$$

Under some regularity conditions, after some iterations, the chain converges to $\pi(\boldsymbol{\vartheta}|\boldsymbol{y})$. However, the values of the Markov Chain before convergence should be discarded. We do so by defining a reasonable value of **burn-in** observations to be discarded. For simple GLM models, a burn-in of 100 iterations could be enough, whereas for more complicated models a much higher value could be necessary.

Example (Beta-Binomial)
We consider the joint distribution for (X, Y) given by

$$f(x, y) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1},$$

and we are interested in the marginal distribution of X. Analytically, we can recover this distribution as the beta-binomial

$$f(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)\ldots}{\ldots}$$

Conditioning on x yields a Beta distribution for $Y|X$, whereas by conditioning on Y yields a Binomial distribution for X. Alternating the simulation between $X|Y$ and $Y|X$ yields the posterior distribution for (X, Y) using the Gibbs sampler.

Example (Probit model)
In the probit model, we have

$$\pi(\beta) = \mathcal{N}(\beta_0, \Sigma_\beta),$$

with likelihood

$$\pi(\boldsymbol{y}|\beta, X) = \prod_{i=1}^{n} \Phi(\boldsymbol{x}_i^\top \beta)^{y_i} \big(1 - \Phi(\boldsymbol{x}_i^\top \beta)\big)^{1-y_i}.$$

However, Albert and Chib (1993) have shown that by introducing latent variables $Z_1, \dots, Z_n$ such that $Y_i = \mathbb{1}_{Z_i > 0}$, then the posterior distribution can be calculated explicitly. By assigning a truncate multivariate normal prior to $\boldsymbol{Z}$, we obtain the full conditional distribution of β (normal), Z_i (truncated normal), and scale parameters λ_i (Gamma) of Z_i.

Remark. Sometimes, Gibbs samplers are hard to define due to the fact that we need the explicit conditional distributions. Sometimes we have tricks to do so, but other times we may need different approaches.

14.2.2 Metropolis-Hastings

In several cases, deriving the full conditional distribution of parameters is not easy, or even possible. A popular alternative to GS is the **Metropolis–Hastings** (MH) algorithm, which can be seen as a generalisation of the Gibbs sampler. Formally, we the algorithm works as follows:

1. we sample $\tilde{\vartheta}_j^t$ from a **proposal distribution** $q_j(\cdot|\vartheta_j^{t-1})$;
2. we accept the proposal with probability

$$p = \min\left\{1, \frac{\pi(\tilde{\vartheta}_j^t) L(\boldsymbol{y}|\tilde{\vartheta}_j^t) q_j(\vartheta_j^{t-1}|\vartheta_j^t)}{\pi(\tilde{\vartheta}_j^{t-1}) L(\boldsymbol{y}|\tilde{\vartheta}_j^{t-1}) q_j(\vartheta_j^t|\vartheta_j^{t-1})}\right\};$$

3. If we reject the proposal, then $\vartheta_t^j = \vartheta_j^{t-1}$.

Remark. MH is tempting since we do not need to derive any conditional distribution. The problem is that the convergence might be extremely slow, and the trick is to find a proposal distribution that guarantees a high proportion of acceptance rate while still exploring the parameter space.

Combinations. GS and MH methods can be combined, and this might be handy for hierarchical models. Specifically, we can apply GS whenever we have conjugacy, and employ a MH step whenever the conditional distribution is unfeasible to calculate.

14.2.3 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) generalizes the classical MCMC scheme by introducing Hamiltonian dynamics in the evolution of the proposal. Specifically, the proposed new value for the chain is simulated using a time-reversible and volume-preserving numerical integrator (typically the leapfrog integrator) to propose a move to a new point in the state space. Specifically, a simple HMC algorithm uses the following scheme:

1. A momentum variable is initialized $\boldsymbol{p}_0 \sim \mathcal{N}(\boldsymbol{0}, m \cdot I_p)$.

2. Generate a new proposed value $(\theta^*, p^*) = \big(\theta(\tau), -p(\tau)\big)$ by approximating the solution $\big\{\big(\theta(t), p(t)\big), t \in [0, \tau]\big\}$ of Hamilton's equations,

$$\frac{\partial \theta}{\partial t} = m^{-1} \cdot \boldsymbol{p}, \quad \frac{\partial p}{\partial t} = -\nabla_\theta U(\theta),$$

   where $U(\theta) = -\log \pi(\theta)$ is the potential energy.

3. Accept the proposed value with probability

$$\alpha = \min\left\{1, \frac{\pi(\theta^*, p^*)}{\pi(\theta, p)}\right\}.$$

Lecture 15: Hierarchical GAMs

2022-04-27

GAM are often considered for flexible regression functions, whereas hierarchical linear models (HLM) are useful due to their ability to "borrow information" when the sparsity is high. These two class of models are much closer than what we might think, since both borrow information:

- GAM: borrowing is performed between different regions of predictor X. We penalize squared deviations from a completely smooth function.
- HLM: borrowing is done between different groups. Group-level effects are pulled towards global effects.

Given this strong connection between HLMs and GAMs, we can think about extending GAMs by allowing smooth functional relationships between predictor and response to vary between groups. These functional relationships across groups are shrunk towards a global functional relation due to the hierarchical effect.

Idea: Just like in HLM, we would like to penalize functions that are far away from the common (mean) shape of functional relationship, and the penalization is higher when the sample size of the group is low.

15.1 Review of GAMs

Given a response variable y and a set of J predictors $x_1, \ldots, x_J$, we specify the model as

$$\mathbb{E}[Y] = g^{-1}\Big(\alpha + \sum_{j=1}^{J} f_j(x_j)\Big),$$

where $g(\cdot)$ is the **link function** and each f_j is a smoother specific to the j^{th} predictor. Each function f_j is represented by a weighted sum of K basis functions,

$$f_j(x_j) = \sum_{k=1}^{K} \beta_{jk} f_{jk}(x_j),$$

and the number of basis functions K determines the maximum complexity of the function f_j.

Note. This does not mean that the actual complexity of f_j becomes automatically higher when we increase K, because increasing the penalization to each β_{jk} can reduce the total complexity of the model.

15.1.1 Penalization

A penalty matrix S is defined to ensure the appropriate complexity of f_j, so that the maximization is performed on

$$\underset{\beta}{\operatorname{argmax}}\, L - \lambda \beta^\top S \beta,$$

where λ is a **penalty parameter** that controls the trade-off between bias and variance. A higher value of λ corresponds to a less complex f_j.

15.1.2 Basis functions

The choice of basis functions usually amounts to three choices:

a) **Thin plate regression splines** are useful when several predictors should be included but we assume that the amount of smoothing is the same for each covariate.

b) **Cyclic cubic smoothers** are useful with cyclic components (seasonal effects), since the start and end of the smoother are constrained to match in value and first derivative.

c) **Random effects** can act as a sort of basis functions.

Consider the model

$$y = X\beta + \varepsilon$$
$$\beta \sim \mathcal{N}\left(0, S^{-1}\frac{\sigma^2}{\lambda}\right)$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2 \cdot I),$$

then we can apply a eigendecomposition $S = U\tilde{\Lambda}U$ to get the rotated values of β, $\tilde{\beta} = U^\top \beta$. Then, we get that

$$\beta^\top S \beta = \tilde{\beta}^\top \tilde{\Lambda} \tilde{\beta},$$

and if we partition the data into fixed and random effects we obtain

$$Y = W\gamma + Zb + \varepsilon,$$

where

$$b \sim \mathcal{N}\left(0, \tilde{\Lambda}^{-1}\frac{\sigma^2}{\lambda}\right).$$

Hence, $\lambda = \sigma_\varepsilon^2/\sigma_b^2$ is the ratio between residual and random effects variances, which acts as a smoothing parameter, and can be estimated using REML.

Example (Nonlinear relation with groups)
When modelling nonlinear relation with different groups, we have multiple choices:

› A single smoother for all observations (pooling smoothers, pooling smoothness)

$$y_{ij} = f(x_{ij}) + \alpha_i + \varepsilon_{ij}.$$

› A global smoother with group–level smoothers that have the same smoothness (partial pooling smoothers, pooling wiggliness),

$$y_{ij} = f(x_{ij}) + f_i(x_{ij}) + \varepsilon_{ij}.$$

› A global smoother with group–level smoothers that have varying smoothness (partial pooling of smoothers, no pooling smoothness)

› Group–specific smoothers without a global smoother, that have the same smoothness (no pooling smoothers, pooling smoothness)

$$y_{ij} = f_i(x_{ij}) + \varepsilon_{ij}$$

› Group–specific smoothers without a global smoother that have varying smoothness (no pooling smoothers, no pooling smoothness).

Remark. If the smoothness varies across groups and we set a common penalty, then we will have high variance in groups with high smoothness and overly-smoothed estimates (high bias) in groups with low smoothness. This is the usual **bias-variance trade-off**.

Concurvity. Concurvity is the analogue to collinearity, and measures how well a smoother can be approximated by a combination of other smoothers. If we consider model 2 and 3 in the example, the global term is entirely concurved with the groupwise smoothers.

References

Albert, J. H. and Chib, S. (1993). «Bayesian Analysis of Binary and Polychotomous Response Data». In: *Journal of the American Statistical Association* 88.422, 669–679.

Berger, J. O. et al. (1999). «Integrated Likelihood Methods for Eliminating Nuisance Parameters». In: *Statistical Science* 14.1, 1–28.

Christensen, R. et al. (1992). «Case-Deletion Diagnostics for Mixed Models». In: *Technometrics* 34.1, 38–45.

Demidenko, E. and Stukel, T. A. (2005). «Influence Analysis for Linear Mixed-Effects Models». In: *Statistics in Medicine* 24.6, 893–909.

Dunn, P. K. and Smyth, G. K. (1996). «Randomized Quantile Residuals». In: *Journal of Computational and Graphical Statistics* 5.3, 236–244.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. Singapore: World Scientific Pub.

Part III

High-dimensional data

Instructor: Davide Risso

References: Box et al. (2005) — Classic experimental design

Burtini et al. (2015), Lai (2001) — Multi-armed bandits

Efron and Hastie (2016) — Statistical learning

Hastie et al. (2015), Buhlmann and Geer (2011) — Sparse & High-dimensional models

Even today, statisticians can provide useful insights in the data-collecting methodology when it is required to carry out an experiment. In general, experiments might be use to both assess the adequacy of a theory or to develop a predictive model. Since the properties of the experiments determine the quality of the resulting estimators, it is of primary importance to know the basics of experimental design. Knowing the statistical properties of the data is a valuable skill to have even today.

> *"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of".*
>
> (R. A Fisher)

After a review of experimental design, we will discuss some methods which may find wide applicability in modern high-dimensional settings, such as penalized regression and graphical models.

Lecture 16: Experimental design

2022-05-04

Experimental design is one the earliest field of statistics, which deals with how to collect data so that the results will best be able to answer the question of interest. Although it is an old topic, it still finds a lot of relevance today both on static (Meng, 2018) as well as in in sequential design and bandit problems.

16.1 Introduction

In experimental design, we want to compare/manipulate different scenarios and see the effect on a certain response variable. The term **experiment** is much broader than the usual meaning which implies labs, test tubes, etc. Examples of applications are

- **Agricolture**: where it all started. Classic design and terminology largely come from Fisher's work at Rothamsted Experimental Station, where he worked on crop yields.
- **Medicine**: perhaps the most widely known examples of experimental design are clinical trials, in which patients are randomly assigned to treatments.
- **Industry**: collect information so as to control a production process. Standard applications are quality control procedures.
- **Advertising**: Often called "AB Testing", this is usually another fancy name for hypothesis testing. What advertising is most effective? What is the best website layout?
- **Epidemiology**: (More generally, observational studies to compare populations). Usually can't manipulate conditions, but can use ideas of experimental design to design how to pick the populations wisely.

16.1.1 Confounding

Confounding describes the situation when two variables both influence the response and the effects cannot be isolated from each other. A simple example is when

$$Y_{ij} = \alpha x_1 + \beta x_2 + (\alpha\beta) x_1 x_2 + \varepsilon_{ij},$$

and our design only observes $\boldsymbol{x} = \{(1,0),(0,1)\}$. In this case, we cannot estimate all the parameters and the effect of x_1 and x_2 is confounded. This is a property of **the experimental design** and not of the variables, and indicate a flaw of the study. Obvious confounders are unobserved variables, non-probabilistic samples, and interactions between effectiveness and drug administration—for which we use double-blinds and placebo.

Example (Expensive drug)
An expensive drug might be administered only to high-end hospitals, and therefore there can be a confounding effect due to the fact that rural hospitals are not included.

Remark. In general, it's easier to think about what when wrong on an experiment when analyzing data, while it's harder to think about the confounding upfront.

In order to design a sensible experiment, one must know how they will analyze the data. Even if there are currently no data, the design of an experiment is based on statistical properties of the analysis (and data).

Flipside. In order to correctly analyze the data, we must understand the experimental design with which data has been collected.

Steps of an experiment:

1. **Design** – choices you make before collecting the data
   - What measurement to make (the response)?
   - What conditions to compare (the treatment)?
   - What is a unit that will get a treatment (individual person? groups of people, e.g. a hospital?)
   - How many samples do I need?
   - Which units get which (combination) of treatments?
2. **Running the experiment**
3. **Analysis** – how you analyze the data created from a particular design to answer the scientific question.

Example (Shoe manufacturing)
A certain manufacturer of kids shoes wants to test two different materials for soles, to decide which one leads to less wear.

- Response: shoe wear, hence discuss how to measure rubber consumption.
- Treatment: material A vs material B.
- Units: sample kids from the interest population.
- Allocation: how do we give treatments? Easiest would be randomizing.
- Sample size: do a power calculation for a t test.

However, a more intelligent sample would give both shoes to every child in order to perform a paired t test. An even more intelligent design would give left and right shoes of the two competing materials to each child in order to test the wear of the soles under comparable conditions.

Remark. Both designs are valid, although conditioning on the specific child is better than simply

Remark. Even in this simple example, we spent a good 20 minutes talking about it. We can imagine that a large scale study might involve lots of planning and studying beforehand.

An experimental design such as the example above is called **randomized block design**:

› `boy` is the **blocking variable** upon which we condition.

› `material` is the treatment variable, of which we **randomize** the assignment.

Blocking and randomization are the two main devices of experimental design.

16.2 Type of designs

16.2.1 Completely randomized

The simplest design is the **completely randomized (CR) factorial design**, also known as randomized basic factorial design. The design refers to both our choice of treatments:

› **complete**: all combinations of the levels are observed

and the mechanism for assignment:

› **randomized**: units are randomly assigned across treatments.

If we have r replications of each treatment, we say that the design is balanced.

Remark. Balanced designs result in better properties, since the precision of the estimates is equal and thus the power is greater.

More treatments. Note that we can have one or more treatments. If we have k treatments we write $CR[k]$.

16.2.2 Complete block design

If we decide to control for a covariate by blocking, it means that we will have our units already divided into groups based on a blocking factor. Then in each of our blocks, we can do a single replication of all the treatments (i.e. if T treatments, we have T units in each block). Each treatment is assigned randomly within a block to the T units.

Example (Boys' shoes)
If we have two materials, we can use the boy as a block because they have two feet. If we have more materials, we cannot do that.

Remark. Again, depending on whether the treatments come from a single variable or the combinations of multiple variables we can write $CB[k]$.

Replication. Of course, if we wanted to, we could have cT units in each block, i.e. repeat each treatment c times within each value of the blocking factor.

This is often called a **Generalized Complete Block** design. This is a choice of trade-off, since replicating within a block usually limits how many different blocks you can afford to have.

What are some advantages of replicating within a block versus getting more blocks?

Example (Shoes)
Given that we can afford a total of 40 measurements, do we want to run four measurements for 10 children or run two measurements for 20 children?

In general, it's better to maximize the number of blocks although for budget reasons it's usually less costly to repeat a measurement in an already-sampled block.

16.3 Latin squares design

If we have two or more blocking factors, we can think of crossing the blocking factors and use the combinations as blocks and then create a complete block design with the combination as the individual blocks.

Problem. It is not always possible (or easy) to assign all treatments to all block combinations, think for instance the fact that the effect of a drug might depend on the order of administration.

An alternative is to relax the requirement of observing all treatments in all block combinations, by using a **latin square design**.

Idea. We cross the blocking factors, but the resulting block combination can only have one treatment assigned to the combination.

	Treatment Order			
	1	2	3	4
Participant 1	A	B	D	C
Participant 2	B	C	A	D
Participant 3	C	D	B	A
Participant 4	D	A	C	B

Figure 35: Latin square design, observe that we have one treatment each for every row and every column.

Example (Milk yield)
The goal of this study is to compare three diets (full grain, partial grain, and roughage) to see what effect diet has on how much milk a cow gives.

Three cows were given each of the three diets for six weeks. The pounds of milk produced are recorded for each six-week period.

The six week period was chosen because it was long enough to effect milk production and not too long because the cow produces less milk the longer the time since pregnancy.

Cow/Weeks	1-6	7-12	13-18
I	608 (roughage)	716 (partial)	845 (full)
II	885 (partial)	1086 (full)	711 (roughage)
III	940 (full)	766 (roughage)	832 (partial)

Figure 36: Latin square design for the cow milk study.

The response is the milk yield, the treatment is the diet, and the blocking factor are the cows.

Randomization. In latin squares, randomization comes from the fact that there is a fixed number of latin squares to choose from, and we randomly choose one of them.

Generalizations. There are generalizations to rectangular problems, such as $n > p$ or vice versa.

16.4 Analysis

We will not talk much about analysis, because you already know how to analyze these situations: GLMs are used in virtually all cases.

Remark. One thing to notice is that usually we are dealing with categorical variables: if we assign treatments, it means we are assigning a specific value of a variable.

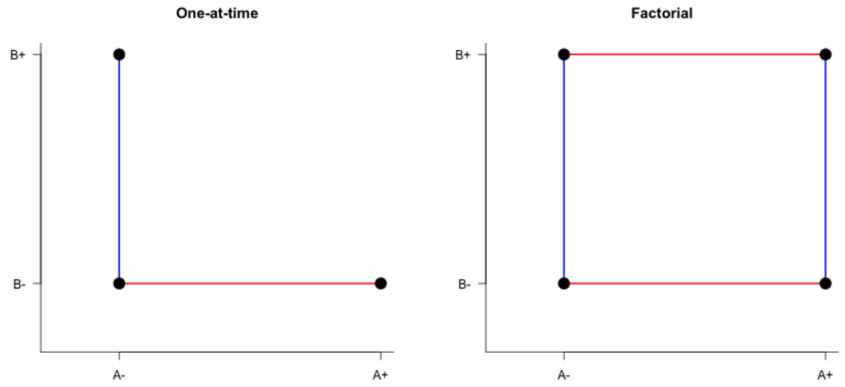


Figure 37: factorial-design-vs-one-at-time

Factorial designs are superior in the sense that

a) We can observe interactions between covariates.

b) We increase statistical efficiency even if there are no interactions.

Example (Efficiency)

Suppose we want to test A^- vs A^+, then we need to calculate averages and study the differences between them. In one-at-a-time design we would compute

$$\bar{y}_{(A^+,B^-)} - \bar{y}_{(A^-,B^-)},$$

and assuming that we know σ^2, the variance of the effect for $N = 48$ will be

$$\frac{\sigma^2}{8} \quad \text{or} \quad \frac{\sigma^2}{6},$$

depending on how you assign the observations. In the case of a factorial design, the estimate of the effect is

$$\frac{1}{2}\left\{\bar{y}_{(A^+,B^-)} - \bar{y}_{(A^-,B^-)} + \bar{y}_{(A^+,B^+)} - \bar{y}_{(A^-,B^+)}\right\},$$

whose variance is $\sigma^2/16$ for $N = 48$.

Example (Psychology experiment)

Subjects are told to perform as many repetitions of a given clerical task as they can in a 1-hour period. The response is the number of tasks correctly performed. All subjects performed a small set of similar clerical tasks as practice before the main study.

16 subjects participated in the experiment for credit toward a requirement of their introductory psychology course (credit group); 16 others were recruited from other classes and paid $10 for the hour (paid group).

In each group (credit or money) half the subjects (selected randomly) were told they had performed unusually well on the practice trials (positive feedback), and half were told they had performed poorly (negative feedback).

Finally, within each of the four groups created by the manipulations just described, half of the subjects (at random) were told that performing the tasks quickly and accurately was correlated with other important job skills (self motivation), whereas the other half were told that good performance would help the experiment (other motivation).

Factor	Levels	Experimental?
Group	Credit or Paid	Not randomized, given
Class	Psychology or Other	No, confounded with Group
Feedback	Positive or Negative	Yes
Motivation	Self or Other	Yes

Figure 38: Factors and their experimental status for the psychology experiment.

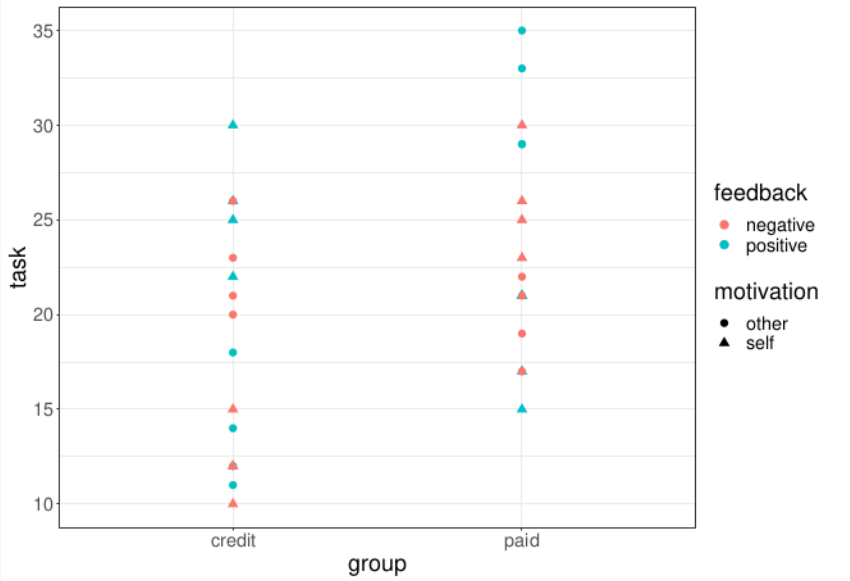


Figure 39: psychology-experiment-plot

We can study a three-way interaction by using three-way ANOVA. Let

$$y_{ijkl} = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\beta\gamma)_{kl} + (\alpha\gamma)_{jl} + (\alpha\beta\gamma)_{jkl} + \varepsilon_{ijkl},$$

where $(\alpha\beta)_{jk}$ is understood as a parameter and not a product. Moreover, we need zero-sum constraints on the parameters,

$$\sum_j \alpha_j = 0, \quad \sum_k \beta_k = 0, \quad \sum_l \gamma_l = 0,$$

plus those in two and three-way interactions. The ANOVA with zero-sum constraints in R can be obtained by using the command

```
options(contrasts=c('contr.sum', 'contr.sum')
```

16.4.1 2^k factorial designs

When we have k dichotomous variables, a common scenario is when we have a total of $T = 2^k$ combinations. There are many settings in which we can't apply a full factorial experiments:

- Too few experiments: We may not be able to run the full $r2^k$ experiments that are needed for a full factorial. We may also be limited by the design points (the number of changes you implement jointly), rather than (or in addition to) the total number of observations (e.g. A/B testing)
- Too small blocks: We might be limited by the number of treatments we can run in a block (e.g. kids' shoes example, only two feet). For example, a complete block design requires running all treatments once in each block. But if each block can only run $<$ 2 k experiments, then you have to have less than the full set of treatments in each block.

In such cases, we can use a design called **fractional factorial design**, in which not all the treatment combinations are observed.

Remark. There is a huge literature on which combinations to choose so that we are only confounding higher order interactions (Box et al., 2005, §6).

16.5 Consequences of design on analysis

A first consequence is the choice of fixed or random effects when analyzing a dataset. It is generally easier to reject the null hypothesis under the fixed model, so it is more conservative (safer?) to make effects random.

Practical considerations. We usually consider our treatments as fixed effects and blocks/nuisance parameters as random effects. This is because usually treatments are what we want our statistical power to be focused on.

Sometimes the nature of the experiment naturally leads to multilevel/hierarchical models, which we refer as **nested designs**.

Example (Nested design)
A typical example is with education: we are interested in measuring effects about children, but children are taught in classrooms. Hence we often cannot apply treatments at the child level, but rather at the school or teacher level.

Example (Nested design (ii))
An example of nested design approach is the following strategy for detecting diabetes in dogs:

1. Each dog gets randomly assigned diabetes or not (kind of completely random), there are ways to induce diabetes in dogs.
2. Apply a measurement to each dog with both technologies (kind of complete-block)

This approach is also called **split-plot design**, since it was first developed for agricultural studies.

Remark. Nested designs are applied, because we cannot perform a complete-block design due to real-world constraints.

Lecture 17: Sequential experimental design

2022-05-04

Sequential decision making is much harder than usual experimental design, and still offers lots of theoretical challenges for modern applications.

17.1 Introduction

Sequential testing or decision making refers to making decisions in contexts where the data is coming in over time. The goal of sequential testing is to monitor the incoming data over time and also make decisions over time.

A simple way of framing this problem is to think about sequential hypothesis testing.

Example (Sequential treatment)
For simplicity, assume that at each time t, a pair of participants join and are randomly assigned to two treatments and their outcomes measured,

$$(X_1, Y_1), (X_2, Y_2), \dots$$

Idea. Rather than waiting until collecting all N samples, we continually monitor the results.

For instance, we can calculate

$$Z_t = \frac{\overline{Y}_t - \overline{X}_t}{\widehat{\sigma}_t/\sqrt{t}},$$

and we can use the series of statistics $Z_1, Z_2, \dots, Z_t, \dots$ to sequentially test the hypothesis

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

Remark. The basic ideas were developed by Wald during WWII for weapons designs, but today they are most commonly seen in two contexts:

1. **Clinical trials**: we can consider that patients may be recruited over time, and assigned to one of two drug therapies. We often have what is called grouped sequential testing, which means that the monitoring is done at intervals, e.g. after r pairs of patients have been recruited. Another common alternative in clinical trials is that the full set of N patients is recruited at the beginning of the trial, but the effects of the drug have to be monitored over a long period of time.
2. **A/B testing**: data is continuously collected over time, where treatments are random decisions between two different ads/layouts for a subset of the users. Stopping before the maximum length of the experiment allows increasing the revenue by switching earlier to the more successful option.

17.1.1 Multiple testing

The multiple testing problem inherent in this set up is probably pretty obvious. This is not a problem in non-sequential clinical trials, because we have the theoretical guarantees given by standard statistical theory when everything is set up *a priori*. In sequential testing, even if there is no difference we have a strong likelihood of observing a difference by chance due to the repeated examinations of the data.

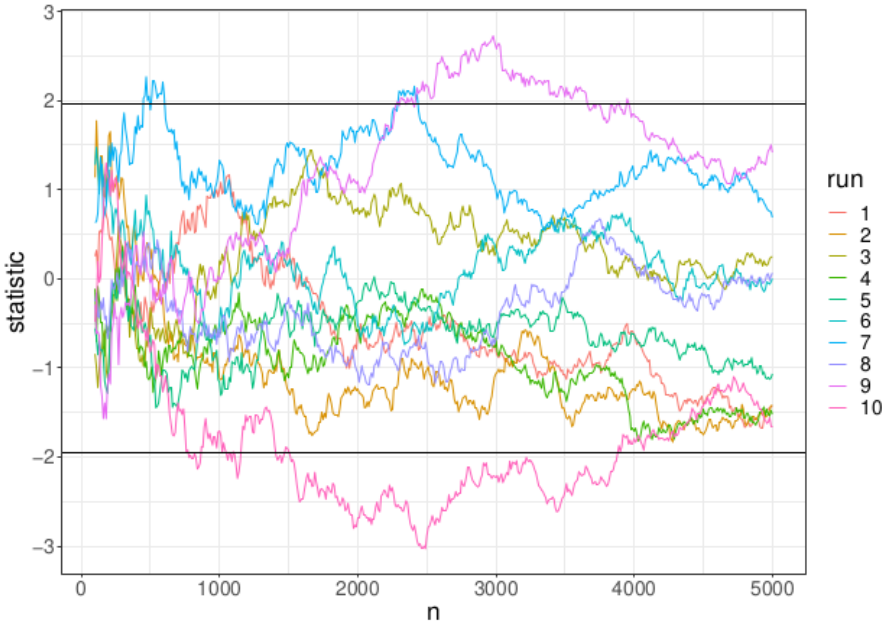


Figure 40: Examples of false positives in multiple sequential tests.

17.1.2 Selection bias

Another statistical problem is one of performing inference after stopping the trial.

Example (Post-stopping inferencec)
In case of a clinical trial we need at a minimum to report confidence intervals of the effect size after stopping. Moreover, we are likely to also do inference on other endpoints, for example by looking for significant side-effects from the drug or sub-populations that do not perform well on the drug.

Remark. All of these inferential tasks will be biased by the fact that the data was used to determine when to stop the trial. This is an example of the **winner's curse**, that is, the fact that the sample is in favour of the hypothesis of finding a positive result. Other sources of bias are possible, depending on the selection rule (Shin et al., 2019)

17.2 Multi-armed bandits

Suppose that we have a set of J possible treatments/conditions, and a response Y to the individual treatment that will be collected over time. Rather than always assigning participants equally across

the groups, we assign them with probabilities that are based on the previous results. Formally, we observe

$$Y_t\big(\tau_{j(t)}\big),$$

where $\tau_{j(t)}$ is the treatment chosen at time t. The data might look like

$$Y_1(\tau_2), Y_2(\tau_1), Y_3(\tau_3), Y_4(\tau_2), \ldots,$$

with the idea that

- at the beginning of the trial, we assign **uniformly**;
- over time, we focus on the **most promising results**.

Differences between bandits and sequential testing is that

- **Sequential testing**: mimic the conclusions we could have made from a simple randomized trial if we just waited for the trial to end.
- **Bandits**: try to find the optimal treatment by maximizing $\mathbb{E}\left[\sum_{j<t} Y_j\right]$.

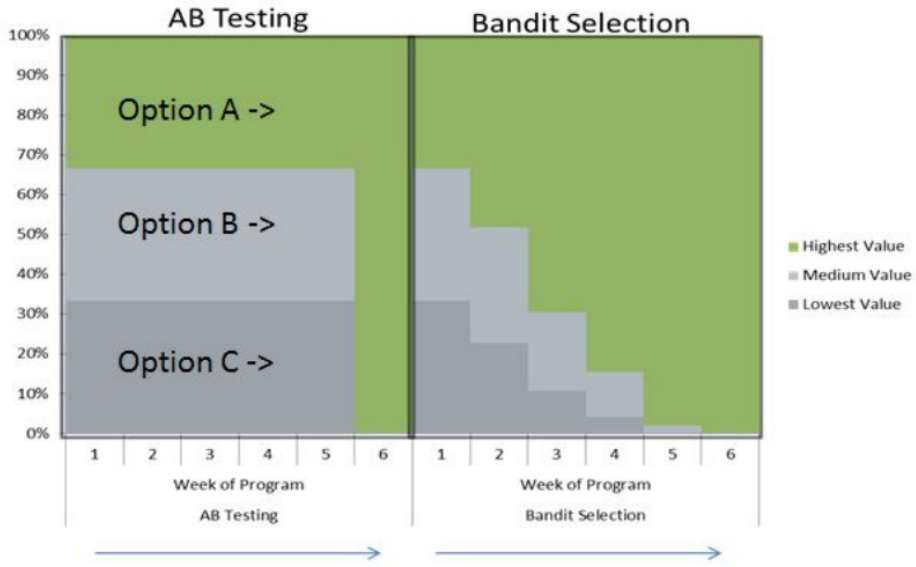


Figure 41: As soon as A starts being better, the probability of assigning it will increase over time.

Advantages. Bandit algorithms have an advantage over standard testing procedures since

a) We can adjust to new options because there is no "testing phase", e.g. we can seamlessly add new treatments over time.

b) The bandit will switch from exploration to exploitation of already-known promising treatments, so that we do not waste potential resources on treatments that do not work.

Choice. Choice between bandit algorithms and sequential hypothesis testing is linked to **different goals**. In drug studies we want to be confident that the best treatment is actually better, and so we try to be more conservative about our results. On the other hand, in A/B testing we simply want to improve revenue and this does not require being conservative with our inferential conclusions.

17.3 Sequential testing

Suppose that we want to repeatedly test the hypotheses

$$\begin{cases} H_0 : \vartheta = \vartheta_0 \\ H_1 : \vartheta = \vartheta_1 > \vartheta_0 \end{cases}$$

and we do so based on large values of a test statistic S_t based on $X_1, \dots, X_t$. The main idea of classical **sequential testing** is that at each time t we can either

1. If $S_t \geq U_t$, reject H_0.
2. If $S_t \leq L_t$, reject H_1 (difference from classic Hyp. test).
3. If $L_t < S_t < U_t$, continue provided that $t < N$, the maximum predetermined size.

Statistically, the question is to determine the best test statistic S_t and the optimal thresholds L_t, U_t so that the overall Type I error of the entire procedure is controlled,

$$\mathbb{P}(\text{reject } H_0 | H_0) \leq \alpha.$$

The most commonly known sequential testing method is the **Sequential Probability Ratio Test** (SPRT) developed for the simple case of testing a simple-vs-simple hypothesis (Wald, 1945).

Def. (Optimality)
The **optimality** of a sequential hypothesis test is defined as minimizing how long we expect to run the trial before we can stop, $\mathbb{E}[T]$, where

$$T = \inf_{t \geq 1} \{S_t > U \vee S_t < L\}$$

Remark. Given $\alpha = \mathbb{P}(S_T > U | H_0)$ and $\beta = \mathbb{P}(S_T < L | H_1)$, we want to use the least amount of observations needed for reaching a conclusion.

Theorem 13 (Wald (1945))
In the case of simple hypotheses, the optimality criteria on $\mathbb{E}[T]$ is satisfied by the likelihood ratio statistic,

$$S_t = \prod_{i=1}^{t} \frac{f_1(x_i)}{f_0(x_i)},$$

and the optimal cutoff values L^* and U^* are such that

$$L^* \geq \frac{\beta}{1-\alpha},$$

$$U^* \leq \frac{1-\beta}{\alpha}.$$

Remark. In practice, we can use the right-hand side of the two equations to set the cutoff values.

Proof.
Let $S_t = \prod_{i=1}^{t} f_1(x_i)/f_0(x_i)$, then we have that if $\mathcal{R}_1$ is the rejection region,

$$\begin{aligned} 1-\beta &= \int_{\mathcal{R}_1} f_1(x)dx = \int_{\mathcal{R}_1} \frac{f_1(x)}{f_0(x)} f_0(x)dx \\ &= \int_{\mathcal{R}_1} S_t \cdot f_0(x)dx \geq U_t \int_{\mathcal{R}_1} f_0(x)dx \\ &= U_t\alpha, \end{aligned}$$

and therefore $U_t \leq (1-\beta)/\alpha$.

With the same reasoning, we can write

$$\begin{aligned} 1-\alpha &= 1 - \int_{\mathcal{R}_1} f_0(x)dx = 1 - \int_{\mathcal{R}_1} \frac{f_0(x)}{f_1(x)} f_1(x)dx \\ &= 1 - \int_{\mathcal{R}_1} S_t^{-1} \cdot f_1(x)dx \\ &= \int_{\mathcal{R}_0} S_t^{-1} \cdot f_1(x)dx \\ &\geq L_t \int_{\mathcal{R}_1} f_1(x)dx \\ &= L_t\beta, \end{aligned}$$

and so $L_t \geq \beta/(1-\alpha)$.

□

Remark. This is a very powerful result but it relies on the fact that we know both ϑ_0 and ϑ_1. However, we have an advantage in that we do not require a specific parametric assumption.

Example (Gaussian with known variance)
Consider

$$X_t - Y_t \sim \mathcal{N}(\vartheta, 1),$$

with the hypothesis

$$\begin{cases} H_0 : \vartheta = 0 \\ H_1 : \vartheta = 0.1 \end{cases}$$

and that we fix $\alpha = 0.05$ and $\beta = 0.1$, then the cutoff values will be $L = 0.105$ and $U = 18$.

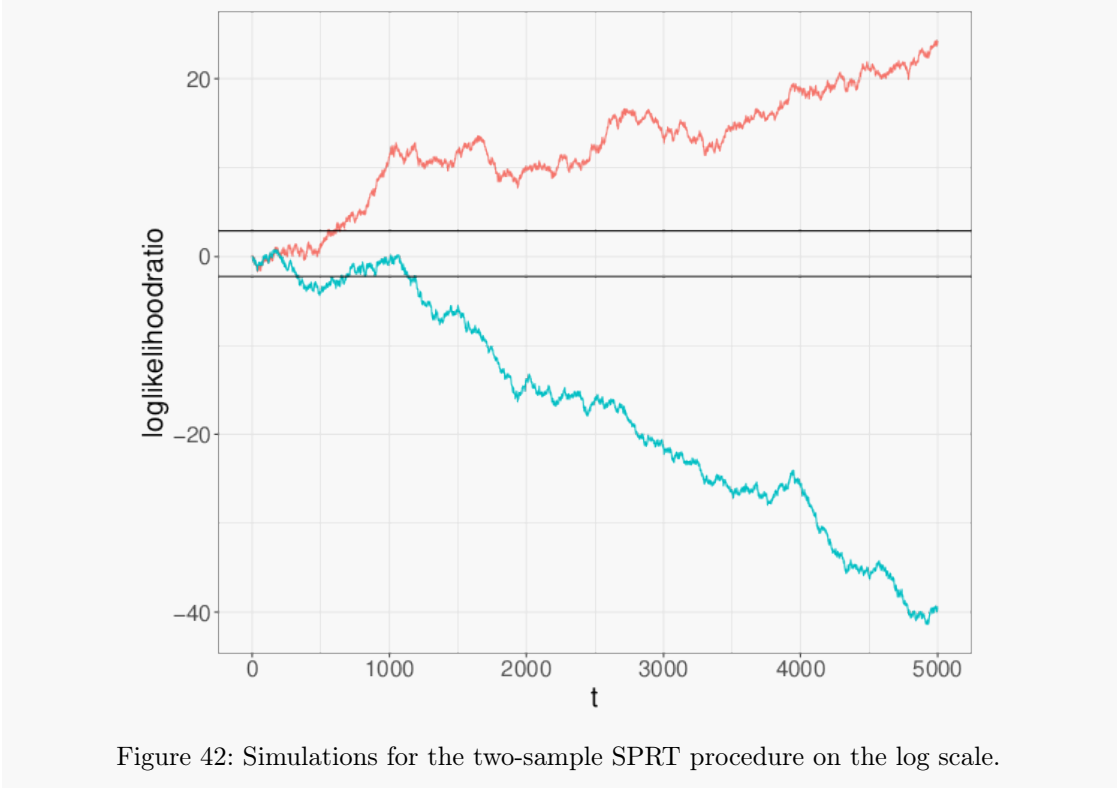


Figure 42: Simulations for the two-sample SPRT procedure on the log scale.

Remark. It is quite odd to keep the same cutoffs even though we increase the sample size. Moreover, the classic SPRT is valid only for simply hypotheses such as normal distributions with known variance.

When the random variables are normal, it seems more logical to use a t-test with unknown variances.

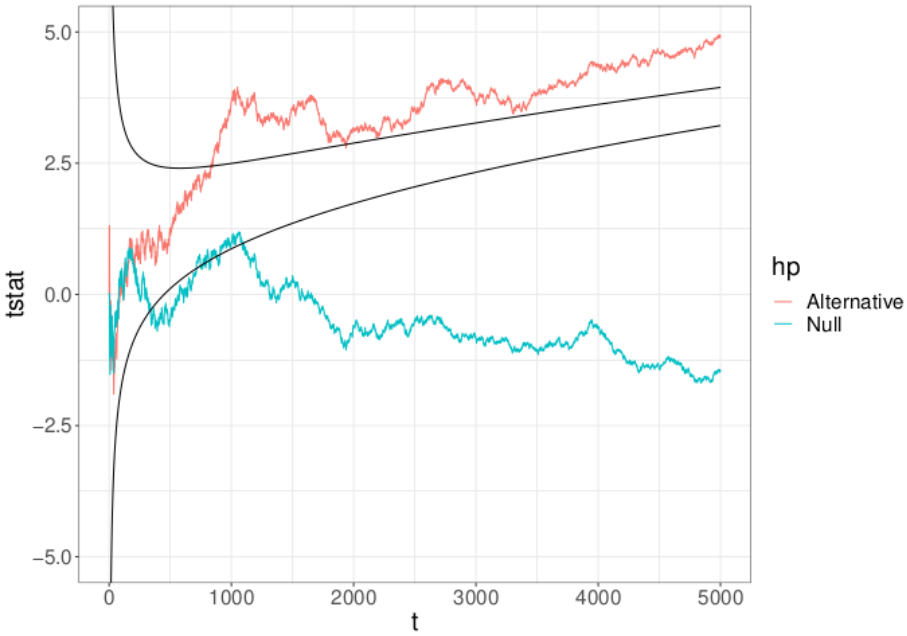


Figure 43: SPRT procedure for a two-sample t-test with unknown variance.

Def. (Confidence sequence)
A $(1-\alpha)$ **confidence sequence** for a parameter ϑ is a sequence of confidence intervals $\text{CI}_t = (L_t, U_t) \subseteq \mathbb{R}$ that satisfies

$$\mathbb{P}(\vartheta \ni \text{CI}_t \text{ for all } t \geq 1) \geq 1 - \alpha. \tag{60}$$

Remark. Property (60) is sometimes referred to as a **uniform coverage guarantee** and it is equivalent to imposing a control on the FWER.

***p*-values** We can also define its dual, often called **always valid** p-values as $p = 1/S_t$ to show that

$$\mathbb{P}(\text{exists a } t \text{ s.t. } p_t \leq \alpha | H_0) \leq \alpha,$$

and to prove it we can use the fact that

$$\mathbb{P}(\text{exists a } t \text{ s.t. } S_t \geq 1/\alpha | H_0) \leq \alpha.$$

17.4 Multi-armed bandits

A multi-armed bandit problem is a process where at time t an agent must choose between J options, called **arms**. In the case of only two treatments there is not a large gain in using bandit strategies, although for multiple arms the improvement might be substantial.

Once an arm j is chosen at time t, the responses

$$Y_t = X_t(A_t)$$

are observed, while all other potential outcomes $\{X_t(j), j \neq A_t\}$ are unobserved.

The basic bandit problem (**stochastic bandit**) assumes stationarity for $X_t(j)$, that is, $X_t(j)$ are i.i.d with mean μ_j, independently of time t. Moreover, we assume a Bernoulli bandit where $X_t(j) \in \{0, 1\}$ so that μ_j is the probability of success. Finally, a bandit will have a **finite time horizon**, meaning that there is a non-random time T at which the process will be stopped.

Goal. Choosing the arms is done so that we maximize the **total reward**,

$$\mathbb{E}\left[\sum_t Y_t\right],$$

or, equivalently, minimize the **expected total regret** against an oracle strategy that knows the optimal action A_t at all times,

$$\tilde{R}(T) = \max_j \sum_{t=1}^{T} X_t(j) - \sum_{t=1}^{T} Y_t,$$

and the goal is to minimize $\mathbb{E}[\tilde{R}(T)]$. In reality we do not know the oracle choices and therefore we use a **pseudo-regret**,

$$R(T) = \sum_{t=1}^{T} \left(\max_j \mathbb{E}\left[X_t(j)\right] - Y_t \right)$$

$$= T\mu^* - \sum_{t=1}^{T} Y_t,$$

and we minimize $\mathbb{E}[R(T)]$.

Remark. In general $\mathbb{E}[R(T)] \leq \mathbb{E}[\tilde{R}(T)]$, hence $R(T)$ is a weaker notion of regret.

Example (A/B testing)
Suppose that we need to decide which "suggested article" to show. We can model the response as a Bernoulli variable: success if the user clicks on the suggestion, failure otherwise.

Assume that we have $T = 100$, politics is effective $1/4$ of times, sports $1/2$ of times, fashion $1/8$ of times and technology $1/6$ of times. Finally, suppose that each choice was presented 25 times, then the average regret would be

$$\mathbb{E}[R(T)] = 100 \cdot 1 - (1/4 + 1/2 + 1/8 + 1/6) \cdot 25 = 73.95.$$

There are several methods proposed to solve this problem, and they include:

- ε-greedy
- Thompson sampling
- Upper Confidence Bound (UCB) algorithms.

17.4.1 ε-greedy

The obvious approach is to use a **greedy approach**, i.e. always selecting the arm j with highest mean $\widehat{\mu}_j$. Each arm starts with equal probability ("uninformative prior") and we update the probability estimate over time.

Problem. The algorithm can be stuck in a non-optimal arm.

In an ε-first algorithm we spend a fixed number of rounds exploring and then goes with the best arm:

1. Exploration phase: try each arm n times.
2. Select the arm $\widehat{A}$ with highest average reward and play it in all remaining rounds.

The expected regret for J arms is

$$\mathbb{E}[R(T)] \leq T^{2/3} \cdot O(J \cdot \log T)^{1/3}$$

ε-greedy approaches uses a different strategy,

1. Toss a coin with time-dependent success probability ε_t
2. If successful we choose an arm uniformly at random
3. Otherwise we choose the arm with the highest average value so far.

If $\varepsilon_t = t^{-1/3}(T \log t)^{-\frac{1}{3}}$, then the ε-greedy method achieves regret bound for each t

$$\mathbb{E}[R(t)] \leq t^{2/3} \cdot O(J \cdot \log t)^{1/3}.$$

We have a regret bound for all values of t, which is a strong improvement over the ε-first algorithm.

17.4.2 Softmax

Other algorithms try to avoid the explicit exploration phase by including information on the posterior probabilities. The **softmax** method uses information on the arms to select them with a probability proportional to a function of $\widehat{\mu}_t(j)$ (Boltzmann exploration),

$$\mathbb{P}(\text{choose } j|\tau) = \frac{e^{\widehat{\mu}_t(j)/\tau}}{\sum_j e^{\widehat{\mu}_t(j)/\tau}}.$$

17.4.3 Thompson sampling

Thompson Sampling (TS) makes this idea more formal under a Bayesian setting. Assuming that we have a posterior distribution for each μ_j, for instance

$$\mu_j|Y_1, \ldots, Y_{t-1} \sim \text{Beta}\big(S_t(j) + \alpha_j, N_t(j) - S_t(j) + \beta_j\big),$$

then we apply the following procedure,

1. Sample values $M_1, \ldots, MJ$ from the posterior distributions of $\mu_1|\boldsymbol{Y}, \ldots, \mu_J|\boldsymbol{Y}$.
2. Choose the arm with the largest M_j.

Remark. For more general distributions, we can choose the next arm based on the expected regret of Y_t under each posterior distribution,

$$\underset{j}{\text{argmax}}\, \mathbb{E}_{\widehat{\mu}_t(j)}[r(X_t(j))|\boldsymbol{Y}].$$

17.4.4 Upper Confidence Bound (UCB)

For each arm, at each t we can create a confidence interval around the estimate of the mean with upper limit

$$\widehat{\mu}_t(j) + U_t(j),$$

where $U_t(j)$ is called the **upper confidence index**. At each t, we pick the arm so that

$$A_t = \underset{j}{\text{argmax}}\, \widehat{\mu}_t(j) + U_t(j),$$

in order to balance the fact that we have an high reward with the variability of the estimate. With the proper choice of $U_t(j)$ we have that (Lai and Robbins, 1985),

$$\mathbb{E}[R(T)] = O(\log T),$$

and the upper confidence bound is derived from the confidence sequences of an equivalent sequential test based on the GLR statistic. Other variants of the algorithm are introduced to make it easier to calculate the $U_t(j)$'s.

UCB1. If $N_t(j)$ is the number of times arm j has been played, we can use a straightforward solution given by

$$U_t(j) = \sqrt{\frac{2\log t}{N_t(j)}}.$$

This algorithm achieves a multiple of the optimal bound but it has been empirically shown to converge very slowly in practice.

UCB2. We can break the time points into epochs $1, 2, \ldots, \nu, \ldots$ and we calculate the arm j chosen based on $U_\nu(j)$. Once that arm is chosen, it is played $m_\nu(\alpha)$ times before picking a new j for the next epoch $\nu + 1$, based on

$$U_\nu(j) = \sqrt{\frac{(1-\alpha)\log(eN/\tau(M_\nu(j)))}{2\tau(M_\nu(j))}}.$$

Once an arm j^* is selected, it is played for

$$m_\nu = \tau(M_\nu(j^*) + 1) - \tau(M_\nu(j^*))$$

times, giving an exponential increase in the number of times the arm is played. This algorithm is particularly useful if there are many arms so that we have adequate time to explore all of them.

17.5 Generalizations

The main assumption so far is to have a stationary process with i.i.d rewards. Other generalizations include **adversarial bandits**, which assume nothing about the distribution of the $X_t(j)$'s.

17.5.1 Adversarial bandits

The name adversarial bandit stems from the fact that we can assume that there is an adversary that chooses $X_t(j)$ to be the worst case at every t, in the sense that we cannot learn what to do next. The following assumptions are made:

- The adversary may pick the rewards according to a deterministic function $f(t)$, which is assumed to be bounded.
  1. If the rewards are chosen at the beginning, it is an **oblivious adversary**
  2. Otherwise, it is an **adaptive adversary** and f can be a function of past choices of the player/agent.
- The player picks an arm based on a (probabilistic) strategy without awareness of the adversary's selections of rewards.
- The rewards are assigned.

Problem. We cannot trust the exploitation phase as safely as before, otherwise the player can always be outmaneuvered by the adversary that learns the strategy. Moreover, even though the function f defining the strategy of the adversary can be assumed to be deterministic, the observed rewards are based on the player's actions. Therefore, the adversary's strategy is also random.

Exp3. The **Exp3 algorithm** calculates a vector of probabilities,

$$p_t(j) = \mathbb{P}(A_t = j | A_1, Y_1, \ldots, A_{t-1}, Y_{t-1}),$$

using a estimation that is based on a forgetting factor $\gamma \in (0, 1)$,

$$w_t(j) = e^{\gamma \widehat{S}_{t-1}(j)},$$

to calculate

$$p_t(j) = (1 - \gamma) \frac{w_t(j)}{\sum_j w_t(j)} + \gamma \frac{1}{J}.$$

High values of γ sends the algorithm towards exploring, whereas lower values of γ are more inclined to exploiting. After picking j^*, we update the weights in j^* using

$$w_{k+1}(j^*) = w_k(j^*) \cdot e^{(\gamma Y_t)/(J \cdot p_t(j^*))},$$

and Exp3 will not converge to a single arm unless $\gamma = 0$.

Remark. This will make the algorithm more robust to non-stationarity, but evidently is not optimal if there is a single clear winner.

Other generalizations of the bandit problem are:

- Contextual bandits: introduce covariates in the reward function.
- Dueling bandits: only two choices are picked and "duel".
- Combinatorial bandits: find the best of $K < J$ arms at each time.

Lecture 18: Models for high-dimensional data

2022-05-11

18.1 Introduction

High-dimensional statistics refers to statistical inference when the number of parameters p is larger than the number of observations n, either in the supervised or unsupervised settings. We will see two examples:

- **High-dimensional regression**, where $p > n$.
- **Multivariate models**, in which we want to learn to learn the correlation structure of a multi-dimensional phenomenon.

Random effect models are a possible way of dealing with situations in which we would have too many fixed effects. Here, we will stress the similarity and differences between mixed effect models and penalized regression.

Example (Gene expression)
In genomics, most experiments rely on high-throughput technologies (such as microarrays and DNA/RNA sequencing).

Genomic datasets are characterized by thousands of variables, such as the abundance (expression) of genes or proteins, which can be measured for each individual. However, it's expensive to perform a full genomic analysis on an individual, and thus the number of statistical units is usually small.

Research. Usually, we have two goals:

a) predict the status of an individual (healthy/diseased) based on their gene expression;

b) identify the differentially expressed genes, i.e. the genes that are significantly different between two (or more) groups.

Whereas for *a*) a machine learning method or lasso/ridge may be useful, for *b*) a common approach is to specify a model for each gene and use empirical Bayes techniques to borrow strength among them.

18.2 Empirical Bayes

Assume that the number of claims for policy holder k follows a Poisson distribution with parameter ϑ_k, with ϑ_k varying randomly,

$$X_k \sim \text{Pois}(\vartheta_k).$$

A Bayesian approach would place a prior distribution on ϑ and use the posterior expected value

$$\mathbb{E}[\vartheta | X = x] = \frac{\int_0^{+\infty} \vartheta p_\vartheta(x) g(\vartheta) d\vartheta}{\int_0^{+\infty} p_\vartheta(x) g(\vartheta) d\vartheta}.$$

Since $\mathbb{E}[X] = \vartheta$, this quantity can be use to predict the expected number of claims for the customer in the next year.

Problem. We might not know the prior distribution of the data, but we can leverage the following result.

Prop. 10 (Robbins' formula)
Let $X|\theta$ be a Poisson random variable and let θ have any prior distribution $g(\theta)$. Then, we have that

$$\mathbb{E}[\vartheta|X=x] = \frac{(x+1)f(x+1)}{f(x)}. \tag{61}$$

Proof.

$$\begin{aligned}\mathbb{E}[\theta|X=x] &= \frac{\int_0^\infty \theta \cdot e^{-\theta}\theta^x/x!g(\theta)d\theta}{\int_0^\infty e^{-\theta}\theta^x/x!g(\theta)d\theta} \\ &= \frac{(x+1)\int_0^\infty e^{-\theta}\theta^{x+1}/(x+1)!g(\theta)d\theta}{\int_0^\infty e^{-\theta}\theta^x/x!g(\theta)d\theta} \\ &= \frac{(x+1)f(x+1)}{f(x)}.\end{aligned}$$

□

Using (61), we obtain the **nonparametric Empirical Bayes** estimate of the posterior mean by plugging in the observed counts,

$$\widehat{\mathbb{E}}[\vartheta|X=x] = \frac{(x+1)\widehat{f}(x+1)}{\widehat{f}(x)} = \frac{(x+1)y_{x+1}}{y_x}. \tag{62}$$

Remark. We can get an analytical solution because we use a Bayesian approach, although we are "forgetting" about the prior distribution $g(\vartheta)$.

Trick. We are using the *concept* of a prior distribution to borrow strength between observations in a frequentist setting. This methodology is particularly useful when we have a lot of parallel problems, for which it is beneficial to borrow strengths between them to obtain a common solution.

Note that (62) uses a nonparametric formulation, but choosing a parametric formulation may be more robust for small values of $\widehat{f}(x)$,

$$\vartheta \sim \text{Gamma}(\alpha, \beta).$$

This amounts to using the observed counts in order to estimate the values of α and β using the marginal density of X. This is convenient since by conjugacy the marginal density is the negative binomial distribution,

$$f_{\alpha,\gamma}(x) \propto \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)}\gamma^x(1-\gamma)^\alpha,$$

where we used the reparameterization $\gamma = \beta/(1+\beta)$. Then, we can maximize over α and γ and use $\widehat{f}(x) = f_{\widehat{\alpha},\widehat{\gamma}}(x)$ in (61).

	nonparam	gamma
0	0.168	0.164
1	0.363	0.398
2	0.527	0.632
3	1.333	0.866
4	1.429	1.100
5	6.000	1.334
6	1.750	1.569
7	NA	NA

Figure 44: Comparison of Robbins' formula using the nonparametric and Gamma approaches.

18.3 Shrinkage estimation

While most of the classic statistical literature focus on unbiased estimators, in certain situations introducing a deliberate bias can benefit the overall performance of the estimator. This is the case of shrinkage, where extreme values from few observations are regularized towards a common value and, in particular, of the James-Stein estimator.

18.3.1 James-Stein estimator

Suppose that we observe $x_1, \ldots, x_n$ where

$$X_i|\mu_i \sim \mathcal{N}(\mu_i, 1),$$

and that the prior distribution is chosen as

$$\mu_i \sim \mathcal{N}(M, A).$$

It's easy to show that the posterior distribution is

$$\mu_i|x_i \sim \mathcal{N}(M + B(x_i - M), B), \quad B = A/(A+1).$$

Using a frequentist approach, we can compare the posterior mean

$$\widehat{\mu}_i^B = M + B(x_i - M),$$

with the MLE $\widehat{\mu}_i^{\text{MLE}} = x_i$. We have that

$$\mathbb{E}\left[\sum_{i=1}^n (\widehat{\mu}_i^B - \mu_i)^2\right] = n \cdot B,$$

$$\mathbb{E}\left[\sum_{i=1}^n (\widehat{\mu}_i^{\text{MLE}} - \mu_i)^2\right] = n.$$

and thus $\widehat{\mu}_i^B$ has B times the risk of the MLE. For instance, it has half the MSE of the MLE when choosing $A = 1$.

Paradox. We can expect the JS estimator to do better since we assume $\mu_i \sim \mathcal{N}(M, A)$. The remarkable fact is the following theorem, that is, the JS estimator works better than the MLE even though μ_i do not come from the same distribution. It turns out to be a "paradox", in the sense that there would be no *a priori* reason to shrink towards a common estimate.

Theorem 14 (James-Stein)
Suppose that $X_i|\mu_i \sim \mathcal{N}(\mu_i, 1)$ *independently for* $i = 1, \ldots, n$ *with* $n \geq 3$. *Then,*

$$MSE(\widehat{\mu}^{JS}) < MSE(\widehat{\mu}^{MLE}),$$

for all choices of $\mu \in \mathbb{R}^n$.

Remark. Specifically, the MSE for the James-Stein estimator is

$$\mathrm{MSE}(\widehat{\mu}^{\mathrm{JS}}) = nB + 3(1 - B).$$

Remark. While in low-dimensional settings it may be preferable to have an unbiased estimator like the MLE, in high dimensional settings, shrinkage estimation has become modern practice. Shrinkage estimation tends to produce better results **in general**, at the possible expense of extreme cases.

18.4 Ridge regression

Consider the classic linear regression model

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I),$$

whose maximum likelihood estimator is given by

$$\widehat{\beta} = (X^\top X)^{-1} X^\top Y.$$

We further assume to have standardized the columns of X, so that $(X^\top X)_{ii} = 1$ for all i's and that $\bar{y} = 0$ so that we do not need an intercept.

Ridge regression is a shrinkage method designed to improve the estimation of β, and is particularly important when $p > n$. The ridge regression estimator is defined for $\lambda \geq 0$ as

$$\widehat{\beta}(\lambda) = (X^\top X + \lambda I)^{-1} X^\top y,$$

or equivalently

$$\widehat{\beta}(\lambda) = (X^\top X + \lambda I)^{-1} X^\top X \widehat{\beta},$$

from which it is clear to see how $\widehat{\beta}(\lambda)$ is a shrunken version of $\widehat{\beta}$.

Remark. Ridge regression can be seen as an *ad hoc* procedure to fix the singularity in order to invert the matrix $X^\top X$, e.g. when we have collinearity or/and $p > n$.

Scale of X. Although linear regression is equivariant under scale transformation of the variables x_j, this is not the case for ridge regression since the penalization $\|\beta\|_2^2$ treats all coefficients equally. This is the reason why we usually scale the covariates prior to applying the ridge regression procedure (Figure 45).

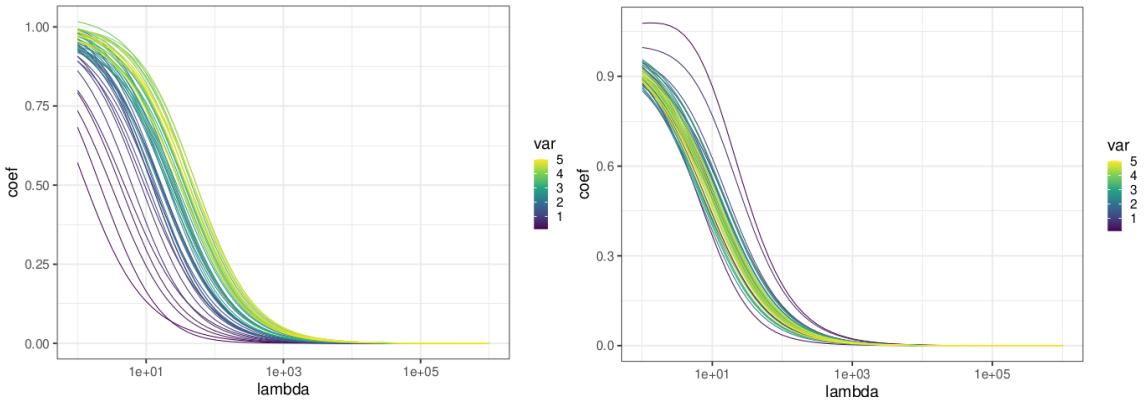


Figure 45: Behaviour of the ridge coefficients for 50 covariates and $\beta_0 = (1, 1, \ldots, 1)$ and $\Sigma_{jj} = j/10$, when predictors are either not standardized (*left*) or standardized (*right*).

Intercept. We usually don't want to penalize the intercept, and so it is left out from the estimation procedure.

18.4.1 Bayesian interpretation

Ridge regression can be interpreted as the maximum a posteriori under a prior distribution

$$\beta|\sigma^2 \sim \mathcal{N}_p\left(0, \frac{\sigma^2}{\lambda} I\right), \tag{63}$$

for which

$$\mathbb{E}[\beta|y] = (X^\top X + \lambda I)^{-1} X^\top X \widehat{\beta}.$$

Intuitively, the larger λ the more strongly we assume that β lies near zero.

Bias/Variance. As often is the case in statistics, we have a bias-variance trade-off. With the ridge estimator, we are (deliberately) introducing a bias in order to reduce the variance when compared to the least squares estimator.

Under (63), we have the log-posterior density

$$\log f(\beta|y, \sigma^2) \propto -\frac{1}{2\sigma^2}\left\{\|y - X\beta\|^2 + \lambda\|\beta\|_2^2\right\}, \tag{64}$$

and the ridge estimator is the maximum of the penalized regression (64).

18.4.2 Comparison with James-Stein

Both ridge and James-Stein estimators are a form of shrinkage. However, if we apply the James-Stein shrinkage to the normal linear model, we get a different shrinkage estimator,

$$\widehat{\beta}^{\mathrm{JS}} = \left(1 - \frac{(p-2)\sigma^2}{\widehat{\beta}^\top X^\top X \widehat{\beta}}\right) \widehat{\beta},$$

for which we know from 14 that

$$\mathbb{E}\left[\|\widehat{\mu}^{\mathrm{JS}} - \mu\|^2\right] < p\sigma^2.$$

Although there is no such guarantee for ridge regression (Efron and Hastie, 2016), and although it is not easy to choose the λ parameter without resorting to cross-validation, the ridge estimator shows good empirical performance.

18.4.3 Link to random effect models

Turning the model to a random effects model,

$$Y = Z\gamma + \varepsilon, \quad \gamma \sim \mathcal{N}(0, \sigma_\gamma^2 I), \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I),$$

then one way of estimating the random effect γ is given by

$$\widehat{\gamma} = \operatorname*{argmin}_\gamma \left\{ \|y - Z\gamma\|^2 + \frac{1}{\sigma_\gamma^2}\|\gamma\|^2 \right\},$$

which, up to a reparametrization of the penalty parameter, is equivalent to the ridge estimate.

Lecture 19: Inference using the lasso

2022-05-11

19.1 Lasso estimator

Other penalizations can be used instead of the L^2 norm in (64), for instance we might choose the L^1 penalty on the coefficients,

$$\tilde{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left\{ \|y - X\beta\|^2 + \lambda \|\beta\|_1 \right\}. \tag{65}$$

Note that (65) can be rewritten as a constrained optimization problem,

$$\underset{\beta_0, \beta}{\operatorname{argmin}} \left\{ \|Y - X\beta\|_2^2 \right\}, \quad \text{s.t.} \|\beta\|_1 \leq t,$$

whose result is compared to the ridge constrained optimization in Figure 46. The most important feature of lasso is to obtain both **shrinkage** and **variable selection**, whereas ridge regression only shrinks the estimates.

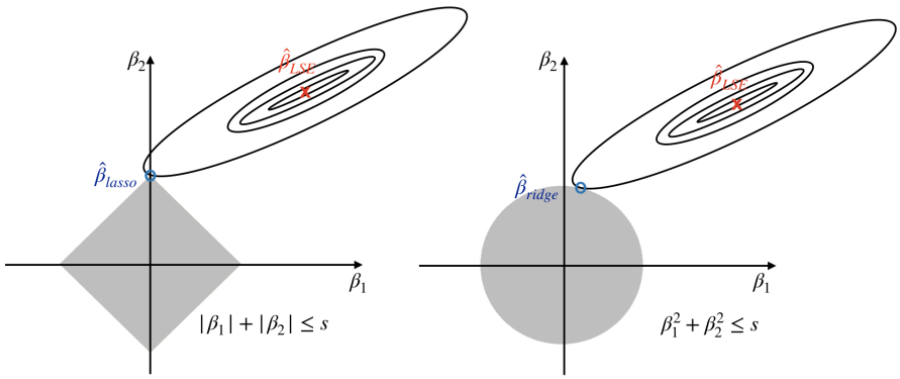


Figure 46: Comparison between the lasso (*left*) and ridge (*right*) penalization.

In addition to shrinking the estimates, we also get some coefficients to be set exactly to zero.

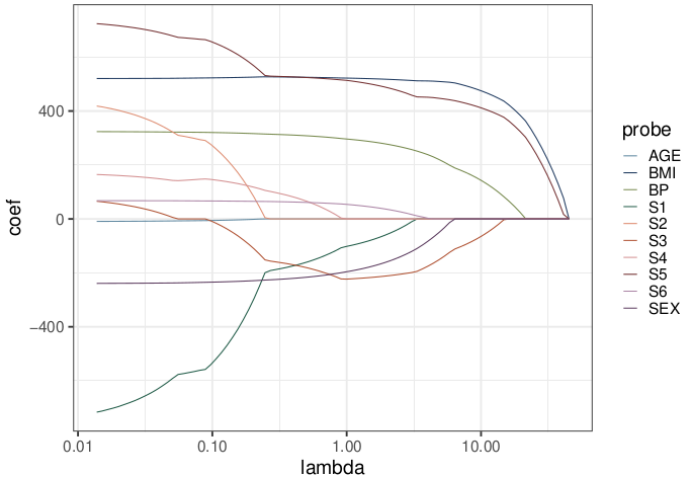


Figure 47: Coefficient path for the lasso regression.

Nonconvex optimization. More generally, one could write the problem using an ℓ_q norm so that $q < 1$ results in a nonconvex optimization problem, whereas $q \geq 1$ yields a convex problem. We observe that $q = 1$ (lasso) is the smallest value that yields a convex problem. Convexity greatly simplifies the computation, and sparsity is convenient both for computation and interpretability of the results.

Sparsity. If the true model is sparse, the lasso can somewhat estimate the parameters efficiently even without knowing which k parameters are nonzero.

19.1.1 Penalty

The bound t, and hence the parameter λ, controls the complexity of the model and has to be estimated differently from the model coefficients. To estimate the best value of λ, the most common strategy is to randomly split the data into training and test sets, then fitting the model on the training set and estimating its performance in the test set using a grid of values of λ.

Repeating this procedure for all combinations of K subsets, using $K-1$ for training and one for testing, is called **cross-validation**. Typical values of K are between 5 and 10 (***K*-fold cross-validation**), depending on the sample size, although if $K = n$ we call it **leave-one-out cross-validation**.

Remark. A natural choice of λ is the value that minimizes the cross-validation mean-squared error. If we want to emphasize sparsity, we can use the "one-standard-error rule", choosing the largest value of λ that yields an error within one standard error above its minimum value.

19.1.2 Computational issues

The lasso can be seen as a convex minimization problem, since it is as a quadratic program with a convex constraint. There are many sophisticated quadratic programming methods to solve the lasso, one of which is the **coordinate descent** optimization routine.

Note that the minimization problem for a single predictor z is equivalent to

$$\operatorname*{argmin}_{\theta} \mathcal{L}(\beta) = \operatorname*{argmin}_{\beta} \left\{ \frac{1}{2}(\beta - \widehat{\beta})^2 + \lambda|\beta| \right\}, \tag{66}$$

where $\widehat{\beta} = (z^\top z)^{-1} z^\top y$ is the maximum likelihood estimator of β. Assuming that z is standardized, we have that $\widehat{\beta} = n^{-1} \sum_{i=1}^n z_i y_i$. Then, if $\beta > 0$ the loss function in (66) has derivative

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta} = \begin{cases} \beta - \widehat{\beta} + \lambda, & \text{if } \beta > 0 \\ 0 & \text{if } \beta = 0 \\ \beta - \widehat{\beta} - \lambda, & \text{if } \beta < 0 \end{cases}$$

which has solution

$$\widehat{\beta}(\lambda) = \begin{cases} \widehat{\beta} - \lambda & \text{if } \widehat{\beta} > \lambda \\ 0 & \text{if } |\widehat{\beta}| \leq \lambda \\ \widehat{\beta} + \lambda & \text{if } \widehat{\beta} < -\lambda \end{cases}$$

The above estimator can be written as

$$\widehat{\beta}(\lambda) = \operatorname{sgn}(\widehat{\beta})\big(|\widehat{\beta}| - \lambda\big)_+ = S_\lambda\left(\frac{1}{n}\sum_{i=1}^n z_i y_i\right),$$

where S_λ is the **soft-thresholding operator**. In the multivariate case we can apply a **cyclic coordinate descent** by minimizing the objective function for predictor j while holding fixed all other $\widehat{\beta}_k$ to their current values for each $k \neq j$.

$$\tilde{\beta}_j(\lambda) = S_\lambda\left(\frac{1}{n}\sum_{i=1}^n x_{ij}(y_i - \tilde{y}_i^{(j)})\right),$$

where

$$\tilde{y}_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\widehat{\beta}_k.$$

An equivalent form of the update is

$$\widehat{\beta}_j(\lambda) = S_\lambda\left(\widehat{\beta}_j + \frac{1}{n}\sum_{i=1}^n x_{ij} r_i\right),$$

where $r_i = y_i - \widehat{y}_i$ are the full residuals.

Convergence. Under mild conditions, it's provable that the pathwise coordinate descent algorithm will converge to the lasso solution.

Orthogonality. If the predictors are orthogonal, a single iteration over the partial residuals gives the lasso solution. This is particularly important in the signal processing setting, since orthogonal bases such as wavelets are particularly useful.

Warm start. Solving for multiple values of λ can be done by starting with a value

$$\lambda_{\max} = \max_j \left|\frac{1}{n}\sum_{i=1}^n x_{ij} y_i\right|,$$

so that $\widehat{\beta}(\lambda_{\max}) = 0_p$. Then, we proceed by slowly decreasing λ and using each previous solution as the starting point ("warm start") for the coordinate descent algorithm.

19.1.3 Degrees of freedom

If we consider the full procedure using the data to select λ, we usually overestimate the true degrees of freedom of the model. However, one can show that for a fixed penalty λ, the number of nonzero coefficients k_λ is an unbiased estimate of the degrees of freedom of the model (Zou et al., 2007).

Remark. This is due to the shrinkage properties of the lasso, which is the price that we need to pay as a trade-off. This result is a basis for the **covariance test** for testing the significance of the predictors.

19.1.4 Further topics

There is a huge body of work related to theoretical aspects of the lasso (Buhlmann and Geer, 2011). Here, we only list some of the theoretical guarantees, without going into details.

- **Uniqueness of the solution**: the lasso solution is unique if the column of X are not linearly dependent.
- **Consistency for prediction**: if the true parameter is sparse relative to $n/\log p$, then the lasso is consistent for prediction.
- **Recovery of the nonzero support set** ("sparcistency"): assuming rather restrictive conditions, with probability tending to one the true set of nonzero variables is included in the set of all possible lasso sub-models.

19.2 Inference

If our goal is prediction, then we have no further things to discuss: the lasso is an extremely simple method that can be used as an alternative to any machine learning method in the regression or classification setting.

However, often our goal is not prediction: we could be interested in variable selection and/or statistical inference. Unlike for prediction, using the lasso for statistical inference is not straightforward.

19.2.1 Bayesian lasso

One possibility is to take a Bayesian approach: this has the advantage of automatically providing an inferential framework. We can specify the model

$$Y|\beta, \lambda, \sigma \sim \mathcal{N}(X\beta, \sigma^2 I),$$

with conditional priors

$$\beta|\lambda, \sigma \sim \text{Laplace}(0, \sigma/\lambda)$$

$$\sigma \sim 1\sigma^2$$

and under this model the negative log posterior density is

$$-\log f(\beta|y, \lambda, \sigma^2) \propto \frac{1}{2\sigma^2}\|y - X\beta\|_2^2 + \frac{\lambda}{\sigma}\|\beta\|_1.$$

Placing a conjugate prior for λ,

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}e^{-\delta\lambda^2},$$

has the advantages of a) performing model selection without choosing λ and b) taking into account the posterior variability in λ.

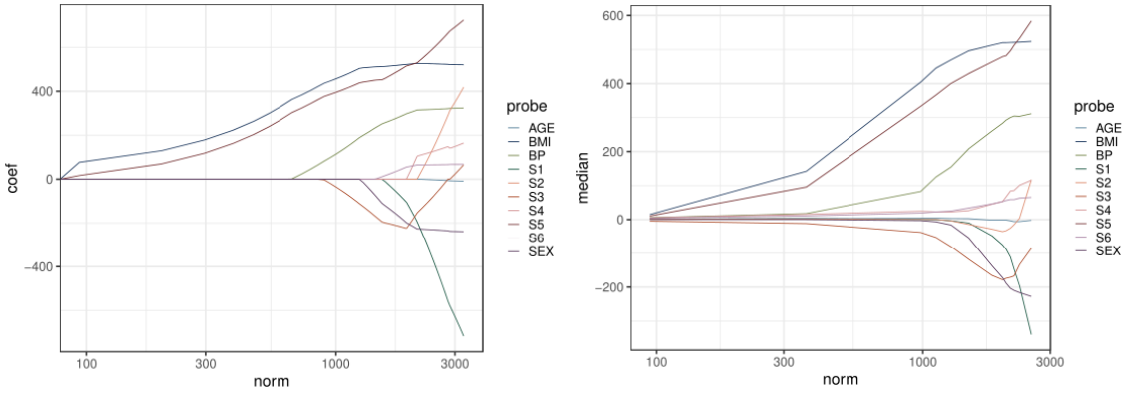


Figure 48: Comparison between the lasso solution paths using the frequentist (*left*) and Bayesian approaches (*right*).

19.2.2 Bootstrap

Assume instead that we do not want to be Bayesian, then we can apply a bootstrap procedure to perform statistical inference on the coefficients. After having chosen $\widehat{\lambda}_{\text{CV}}$, we can use a nonparametric bootstrap procedure to obtain the sampling distribution of $\widehat{\beta}(\lambda)$, although this requires B replications of the CV procedure.

Remark. Although we are evaluating the uncertainty of the estimates, we are doing so by assessing the uncertainty based on the selection λ. However, many other times we are interested in the specific distribution of $\widehat{\beta}(\widehat{\lambda}_{\text{CV}})$. The question of "fixed λ" inference is the object of the so-called **post selection inference**.

19.2.3 Hypothesis testing

Things are complicated by the fact that the lasso performs variable selection as part of the procedure. It is not immediate, for instance, to devise a test stepwise regression that accounts for the adaptive nature of the procedure. Surprisingly, such a test is available for the lasso and is based on the **least-angle regression algorithm**.

Consider the sequence of **knots** $\lambda_1 > \lambda_2 > \ldots > \lambda_K$ returned by the LAR, then we want to test the significance of the predictor entered in the active set at step k. Denote with $\mathcal{A}_{k-1}$ the set of active predictors before the tested predictor entered the set and with $\widehat{\beta}(\lambda + 1)$ the estimate at the end of the step. We refit the lasso with $\lambda = \lambda_{k+1}$ but using only the variables in $\mathcal{A}_{k-1}$.

The **covariance test statistic** (Lockhart et al., 2014) is defined as

$$T_k = \frac{1}{\sigma^2}\big(y^\top X\widehat{\beta}(\lambda_{k+1}) - y^\top X\widehat{\beta}_{\mathcal{A}_{k-1}}(\lambda_{k+1})\big),$$

and measures how much of the covariance between the outcome and the fitted model can be attributed to the predictor that has just entered the model.

Remark. Under H_0 that all $k-1$ true signals already entered the model, $T_k \xrightarrow{d} \text{Exp}(1)$ as $n, p \to \infty$. When σ^2 is unknown, it can be estimated using the full model and the null distribution becomes $F_{2,n-p}$.

Problem. The covariance test uses a set of **conditional hypotheses**: at each stage of LAR, we are testing whether the coefficients of all other predictors not yet in the model are zero. The fundamental problem is that the variables that enter the data at time k depend on the data, and so we are testing the order of the inclusion of the variables. Bühlmann et al. (2014) more or less argue that these hypotheses are nonsensical and of small practical interest.

19.2.4 Debiased lasso

A different approach to post-selection inference does not attempt at making inference on the coefficients estimated by the lasso, but aims at estimating the confidence intervals for the full set of population regression coefficients, assuming a linear model.

A debiased version of the lasso estimator can be defined as

$$\widehat{\beta}^d = \widehat{\beta}(\lambda) + \frac{1}{n}\Theta X^\top (y - X\widehat{\beta}(\lambda)),$$

where Θ is an appropriate inverse of $\widehat{\Sigma} = \frac{1}{n}X^\top X$.

Remark. If $n > p$, then Θ is the proper inverse and the estimator is unbiased, and several authors provide estimates of Θ so that

$$\widehat{\beta}^d \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{n}\Theta\widehat{\Sigma}\Theta^\top\right),$$

so that confidence intervals may be formed for the β_j's.

Nodewise regression Consider the "nodewise regression" that uses the lasso to regress each variable in X to the other $p-1$ variables,

$$\widehat{\gamma}_j = \underset{\gamma \in \mathbb{R}^{p-1}}{\text{argmin}} \left\{ \frac{1}{2n}\|X_j - X_{-j}\gamma\|_2^2 + \lambda_j \|\gamma\|_1 \right\},$$

then we can define

$$\widehat{\Theta} = \widehat{T}\widehat{C},$$

where $\widehat{T}^2 = \text{diag}(\widehat{\tau}_1^2, \ldots, \widehat{\tau}_p^2)$ with

$$\widehat{\tau}^2 = \frac{1}{n}\|X_j - X_{-j}\widehat{\gamma}_j\|_2^2 + \lambda_j \|\widehat{\gamma}_j\|_1,$$

and $\widehat{C}$ is defined as

$$\widehat{C}_{ij} = \begin{cases} 1 & \text{if } i = j, \\ -\widehat{\gamma}_{ij} & \text{if } i \neq j. \end{cases}$$

19.3 Feature screening

Often, the lasso is also used for the goal of variable selection: since some coefficients will be estimated as exactly zero, we can use the lasso as a **screening procedure** to identify what are the variables that influence our response **without any test of hypotheses**.

A general solution is to use **penalized likelihood** approaches. Due to the fact that the likelihood will keep increasing as we increase p, a typical solution is to penalize for the increase in dimension using AIC or BIC.

Under some assumptions and with $\lambda \approx \sqrt{\log(p)/n}$, it can be shown (Buhlmann and Geer, 2011, §6) that

$$\|\widehat{\beta}(\lambda) - \beta\|_q \xrightarrow{P} 0, \quad \text{as } n \to \infty.$$

Therefore, if $S_0 = \{j \in \{1, \dots, p\} : \beta_j \neq 0\}$, we can consider the lasso as producing a screening estimator

$$\widehat{S}(\lambda) = \{j \in \{1, \dots, p\} : \widehat{\beta}_j(\lambda) \neq 0\},$$

hoping that $\widehat{S}(\lambda)$ may allow to infer S_0. A simpler and more relevant result can be obtained by considering the set of **relevant** covariates as

$$S_0^R(C) = \{j \in \{1, \dots, p\} : |\beta_j| \geq C\}, \quad C \in \mathbb{R}.$$

One can show that, for any fixed $C \in (0, +\infty)$ we have that the following **beta-min condition** holds,

$$\mathbb{P}(\widehat{S}(\lambda) \supset S_0^R(C)) \xrightarrow{n \to \infty} 1. \tag{67}$$

The **feature-screening** property described by Equation (67) means that with high probability we include the relevant covariates.

However, an important corollary is that the every Lasso estimated model has $\min\{n, p\}$ estimated coefficients. Then, if $p \gg n$ we typically obtain a massive dimensionality reduction of the problem even with the smallest choice of penalty.

Remark. This means that choosing λ to achieve optimal screening will be particularly hard when $p \gg n$.

Lecture 20: Extensions of the Lasso

2022-05-18

Up until now we have discussed a penalized version of the linear model. In practice, generalized linear models are needed when our response is not continuous and extending the lasso to those settings is important. Moreover, the lasso does not work well in the case of highly correlated predictors and several solutions have been proposed to extend the model to address this issue. In general, the idea is to leverage some natural grouping of the predictors.

20.1 Lasso for GLMs

Instead of using the least squares, we minimize the **penalized negative log-likelihood**

$$\widehat{\beta}(\lambda) = \underset{\beta_0,\beta}{\operatorname{argmin}} \left\{ -\frac{1}{n} L(\beta_0, \beta; y, X) + \lambda \|\beta\|_1 \right\}. \tag{68}$$

From the computational point of view, it may not be as easy to optimize as the least squares likelihood.

Remark. We can still use the coordinate descent, although there may not be a closed-form solution to the optimization.

20.1.1 Logistic regression

Logistic regression is the prime example of lasso for GLM, which is widely used in a variety of settings. Because of its importance in classification and machine learning (document classification, genome disease association, ...), a lot of development has been devoted to solving the problem of penalized logistic regression.

Let $Y_i | X_i = x \sim \operatorname{Bin}\big(1, \pi(x)\big)$, with

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta^\top x,$$

then the negative log-likelihood (68) takes the form

$$-\frac{1}{n}\sum_{i=1}^{n} \left\{ y_i(\beta_0 + \beta^\top x_i) - \log\left(1 + e^{\beta_0 + \beta^\top x_i}\right) \right\} + \lambda\|\beta\|_1. \tag{69}$$

Note how this is a convex optimization problem, since

$$-yf(x) + \log\left(1 + e^{f(x)}\right)$$

is a sum of convex functions, and therefore is a convex optimization problem.

Glmnet. The R package `glmnet` uses a **proximal-Newton** optimization by approximating the negative log-likelihood using a quadratic function,

$$Q(\beta_0, \beta) = -\frac{1}{2n}\sum_{i=1}^{n} w_i (z_i - \beta_0 - x_i^\top \beta)^2 + C(\tilde{\beta}_0, \tilde{\beta})^2,$$

where

$$z_i = \tilde{\beta}_0 + x_i^\top \tilde{\beta} + \frac{y_i - \tilde{\pi}(x_i)}{\tilde{\pi}(x_i)(1 - \tilde{\pi}(x_i))}.$$

We thus iteratively apply the following optimization scheme:

a) newton update to minimize Q, which is a simple weighted least squares problem;

b) apply the coordinate descent to the quadratic approximation.

20.1.2 Signed variables

In the machine learning community, it is more common to encode the response as $\{-1, 1\}$ rather than $\{0, 1\}$. When using sign variables, the penalized negative log-likelihood becomes

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i(\beta_0 + \beta^\top x_i)}\right) + \lambda\|\beta\|_1, \tag{70}$$

where the product $y_i(\beta_0 + \beta^\top x_i)$ is usually referred to as the **margin** in the support-vector machine literature.

› **Positive margin**: correct classification.

› **Negative margin**: incorrect classification.

Since minimizing (70) corresponds to maximizing the margin, we would like all of them to be positive.

Problem. It is well known that a logistic regression without penalization will fail on **separable data**, i.e. on two classes that are linearly separable. This means that the data could in principle be perfectly separated into two classes by a single hyperplane. Logistic regression fails in this setting because maximum likelihood finds

$$\log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \log\frac{1}{0} = +\infty = \beta_0 + \beta^\top x.$$

Solution. When adding a penalization term, the problem goes away by choosing a sufficiently large value of λ. This is important because when $n \ll p$ the classes are almost always linearly separable, unless there are exact ties in covariate space for the two classes.

Remark. This implies that we should be careful with small values of λ in logistic regression, as they can lead to very unstable models.

Remark 2. Very small values of λ have an important meaning in the machine learning community since they reveal a link between **penalized logistic regression** and **support vector machines**.

Given a boundary $\mathcal{B} = \{x \in \mathbb{R}^p : f(x) = 0\}$ associated to $f(x; \beta_0, \beta) = \beta_0 + \beta^\top x$, the euclidean distance between a point x_0 and the boundary is

$$d_2(x_0, \mathcal{B}) = \inf_{z \in \mathcal{B}} \|z - x_0\|_2 = \frac{|f(x_0)|}{\|\beta\|_2}.$$

Then, the quantity

$$\frac{yf(x)}{\|\beta\|_2}$$

is the signed Euclidean distance from the boundary; negative if the sign of y disagrees with that of the prediction $f(x)$. Hence, the optimal separating hyperplane $f^*(x)$ for separable data solves the optimization problem

$$M_2^* = \max_{\beta_0,\beta} \left\{ \min_{i\in\{1,\dots,n\}} \frac{y_i f(x_i)}{\|\beta\|_2} \right\}.$$

Considering the ridge logistic regression, we can prove that the solution for a small λ's are such that

$$\lim_{\lambda\to 0} \left\{ \min_{i\in\{1,\dots,n\}} \frac{y_i f(x_i, \widehat{\beta}(\lambda))}{\|\widehat{\beta}(\lambda)\|_2} \right\} = M_2^*.$$

Analogously, the lasso solution converges to the ℓ_1 version of the SVM, M_∞^*.

20.1.3 Conclusion

- For small values of λ, the logistic regression solution coincides with the SVM solution.
- The SVM approach leads to more stable numerical estimates in this region.
- Logistic regression is more useful in the sparser part of the solution path.

20.2 Elastic-Net

The main problem of the lasso is that it does not handle highly correlated predictors very well. The **elastic-net** solution is a compromise between the ridge and lasso approaches, which is obtained by solving the convex problem

$$\widehat{\beta}(\alpha,\lambda) = \underset{\beta_0,\beta}{\operatorname{argmin}} \left\{ \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - x_i^\top\beta)^2 + \lambda\Big(\frac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1\Big) \right\}.$$

Here, $\alpha \in [0,1]$ is a tuning parameter which balances between the lasso solution ($\alpha = 1$) and ridge solution ($\alpha = 0$).

Remark. For any $\alpha < 1$ and $\lambda > 0$, the elastic net problem is strictly convex, i.e. a unique solution exists irrespective of the **correlations** or **duplications** of the predictors. A coordinate descent algorithm similar to that of the lasso is used to solve the optimization.

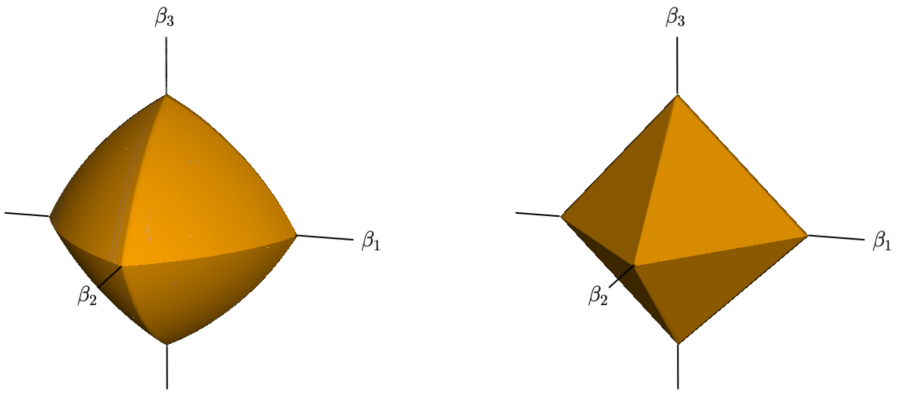


Figure 49: The elastic net constraint (*left*) compared to the lasso constraint (*right*) for three predictors, β_1, β_2 and β_3.

20.3 Group Lasso

Many settings exist for which we have a natural grouping of our predictors:

- features may be **structurally grouped**, e.g. the dummy variables of a categorical predictor with multiple categories.
- another example is in genomics, in which genes are organized in biological pathways. Genes in each pathway tend to be correlated, as they are involved in the same biological process.

In these cases, it may be desirable to have all coefficients within a group become nonzero (or zero) simultaneously. The **group lasso** solves the convex problem

$$\widehat{\vartheta}(\lambda) = \underset{(\vartheta_0,\vartheta)}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \vartheta_0 - \sum_{j=1}^{J} Z_j \vartheta)^2 + \lambda \sum_{j=1}^{J} \|\vartheta_j\|_2 \right\}. \tag{71}$$

Remark. We apply the ℓ_2 norm *within* the group and the ℓ_1 norm *between* groups by summing them.

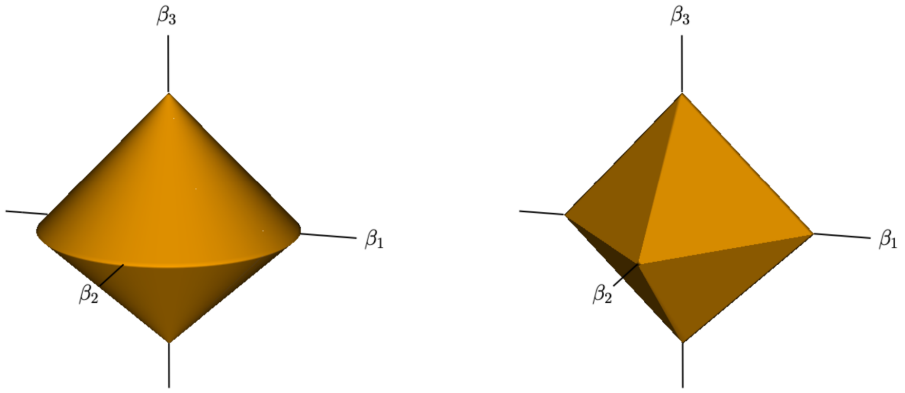


Figure 50: The group lasso constraint (*left*) compared to the lasso constraint (*right*) with $\vartheta_1 = (\beta_1, \beta_2)$ and $\vartheta_2 = \beta_3$.

Remark. Depending on λ, either the entire vector $\widehat{\vartheta}_j$ will be zero or not. When $p_j = 1$, we have $\|\vartheta_j\|_2 = |\vartheta_j|$ and if all groups are singletons, the problem reduces to the lasso. Moreover, since all groups are equally penalized the method tends to select larger groups; versions that use weighted penalties such as $\lambda_j = \sqrt{p_j}\lambda$ exist.

20.3.1 Computational aspects

Minimization of (71) can be done blockwise by finding the solution to

$$\widehat{\vartheta}_j = \left(Z_j^\top Z_j + \frac{\lambda}{\|\widehat{\vartheta}_j\|_2} I \right)^{-1} Z_j^\top r_j, \tag{72}$$

where $r_j = y - \sum_{k \neq j} Z_k \widehat{\vartheta}_k$ is the j^{th} partial residual. Since (72) contains $\widehat{\vartheta}_j$ in the right-hand side, the update is not explicit and has to be approximated.

› **Orthonormal variables**: we can immediately apply the solution

$$\widehat{\vartheta}_j = \left(1 - \frac{\lambda}{\|Z_j^\top r_j\|_2} \right)_+ Z_j^\top r_j.$$

› **General case**: we can use a gradient-descent-like numerical update rule.

Sparsity. Group lasso is atypical since a group includes all coefficients within a specific group. In certain applications we would like sparsity both *between* groups (grouped lasso) and *within groups* (standard lasso). The **sparse group lasso** (Simon et al., 2013) achieves this by balancing between the two approaches,

$$\widehat{\vartheta}(\lambda) = \underset{(\vartheta_0, \vartheta)}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \vartheta_0 - \sum_{j=1}^{J} Z_j \vartheta)^2 + \lambda \sum_{j=1}^{J} (1-\alpha) \|\vartheta_j\|_2 + \alpha \|\vartheta_j\|_1 \right\}. \tag{73}$$

Remark. The logic is the same to the elastic-net, but with a ℓ_1 penalization at the group level.

Algorithm. Since the problem is still convex, a similar algorithm based on block coordinate descent can be used to optimize the function. In the case of orthonormal Z_j, we have

$$\widehat{\vartheta}_j = \left(1 - \frac{\lambda(1-\alpha)}{\|S_{\lambda\alpha}(Z_j^\top r_j)\|_2} \right)_+ S_{\lambda\alpha}(Z_j^\top r_j),$$

where $S_\mu(z) = \operatorname{sgn}(z)(z - \mu)_+$ is the soft-thresholding operator.

20.3.2 Overlap group lasso

Sometimes variables can belong to more than one group, for instance genes can belong to more than one biological pathway. The overlap group lasso (Jacob et al., 2009) is a modification of the group lasso that allows variables to contribute to more than one group.

Idea. We replicate the variable for each group in which it appears and we fit the regular group lasso. For a variable X_j that is replicated twice, X_{j1} and X_{j2}, the coefficient will be given by

$$\widehat{\beta}_j = \widehat{\vartheta}_{j1} + \widehat{\vartheta}_{j2},$$

hence it will be zero only if both estimates are zero. As a consequence variable X_j has a higher chance of being included in the model because it belongs to two groups.

Remark. This is better than considering the same parameter in both groups, since

a) $\widehat{\vartheta}_j = 0$ in a group imposes $\widehat{\vartheta}_j = 0$ in the other group, hence we either select both groups or neither group is selected.

b) from the optimization point of view, we would not be able to apply the block coordinate optimization anymore.

Recasting. Instead of duplincating the predictors, we introduce a new variable

$$\nu_{jk} = \begin{cases} \beta_k & \text{if } X_k \text{ is in group } j \\ 0 & \text{otherwise} \end{cases}$$

and leverage the fact that $\beta_k = \sum_{j=1}^{J} v_{jk}$. Then, we can recast the overlap group lasso into the problem

$$\underset{\nu_j \in \mathcal{V}_j}{\operatorname{argmin}} \left\{ \frac{1}{2} \|y - X\Big(\sum_{j=1}^{J} \nu_j\Big)\|_2^2 + \lambda \sum_{j=1}^{J} \|\nu_j\|_2 \right\},$$

where $\mathcal{V}_j \subseteq \mathbb{R}^p$ is the subspace of all possible vectors ν_j. The main result of (Jacob et al., 2009) is to rewrite the above optimization problem in terms of the original β's as

...

Application. One important application of the overlap group lasso is to enforce the hierarchy in models that contain interactions. Recall that it may be favorable to always include all the main effects when including interactions

20.3.3 Fused lasso

When there is spatial or temporal ordering of the observations, we might expect that contiguous observations behave similarly. For instance in copy number variation studies, whole chromosomal regions are duplicated or deleted, hence if we have multiple genes in the affected region we will see a coordinated response.

The **fused lasso** (Tibshirani et al., 2005) exploits the structure of the signal to solve the following optimization problem,

$$\underset{\vartheta \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \vartheta_i)^2 + \lambda_1 \sum_{i=1}^{n} |\vartheta_i| + \lambda_2 \sum_{i=2}^{n} |\vartheta_i - \vartheta_{i-1}| \right\}. \tag{74}$$

The logic behind (74) is to both shrink the coefficients using the ℓ_1 norm and to encourage neighbouring coefficients to be similar to each other.

Lecture 21: Graphical models

2022-05-25

Probabilistic Graphical Models provide a framework for building parsimonious models for high dimensional data. Here, we do not have a regression setting, but rather a situation in which we want to characterize the multivariate distribution of our data. Instead of using the graph to represent the statistical units, we will use them to describe the multivariate relationships between the variables in our model.

21.1 Introduction

Idea. The main idea is that there is a one-to-one relation between the structure of a graph and the resulting conditional independence of a set of random variables.

There are many application of graphs in statistics, mainly divided in two classes:

1. **Structure learning** or **graphical model selection**: the structure of the graph is unknown and needs to be estimated from the data. This is the main focus of this class.
2. **Prior information** can be added to the model by imposing the graph structure using domain knowledge.
3. **Group testing**: in yet other cases the interest lies in estimating the difference between two graphs in two conditions.

There is a huge literature on graphical models, with the two main classes of models being

› graphical models for Gaussian random variables;

› graphical models for discrete random variables.

We will focus only on Gaussian Graphical Models here. See e.g. Lauritzen (1996) for a more general treatment of the subject.

21.2 Basics of graphical models

Consider the collection of random variables $X = (X_1, \dots, X_p)$. These random variables can always be associated to the vertices (nodes) of some underlying graph. The essential idea is to leverage the properties of the graph (i.e., its structure) to constrain the distribution of the random vector X.

Def. (graph)
A **graph** $G = (V, E)$ is the pair given by the **vertex set** $V = \{1, \dots, p\}$ and the **edge set** $E = \{(s, t) : s \neq t, s, t \in V\}$.

Def. (Directed edge)
A **directed edge** is a pair $(s, t) \in E$ such that $(t, s) \notin E$.

Remark. Conversely, if $(s, t) \in E \implies (t, s) \in E$ then the edges are undirected.

Remark. A graph might contain a mixture of directed and undirected edges, although we will focus only on undirected graphs.

Causality. Directed graphs can be used to represent causality relationships between between random variables. A popular class of directed graphs is the **Directed Acyclic Graphs** (DAGs).

By denoting with P the probability distribution of X,

$$(X_1, \dots, X_p) \sim P,$$

then the pair (G, P) is referred to as a **graphical model**.

A subset of vertices a defines an **induced subgraph**

$$G_A = (A, E_n A \times A),$$

and a subgraph is **complete** if if all pairs of its vertices are connected in G. A **clique** $C \subseteq V$ is a **fully connected** subgraph, i.e.

$$(s, t) \in E \quad \forall s, t \in C.$$

A clique is **maximal** if it is not strictly contained within another clique, and we denote with $\mathcal{C}$ the set o all cliques.

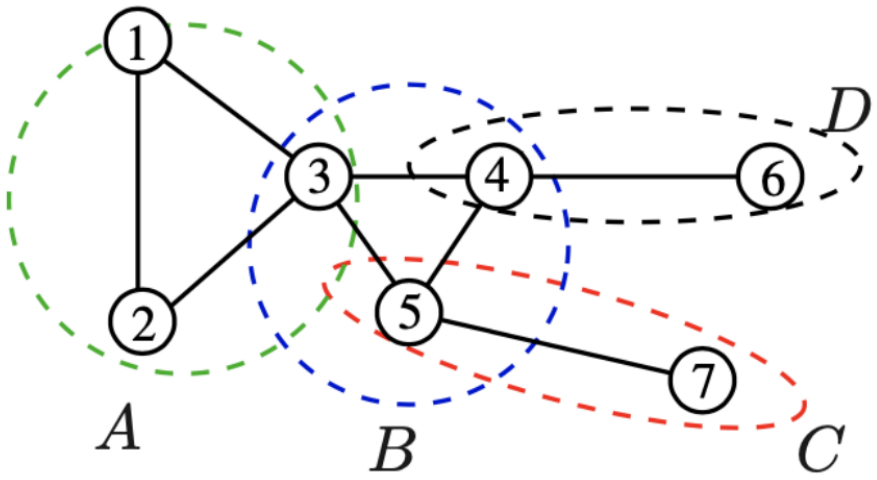


Figure 51: Example of a graph with four maximal cliques.

Def. (Separated subsets)
Two disjoint subsets $A, B \subset V$ are **separated** by a subset S (disjoint from A and B) if all paths from A to B contain vertices from S.

Remark. In other words, a set S is a **separator** (or **cut set**) if it separates the graph into disconnected components.

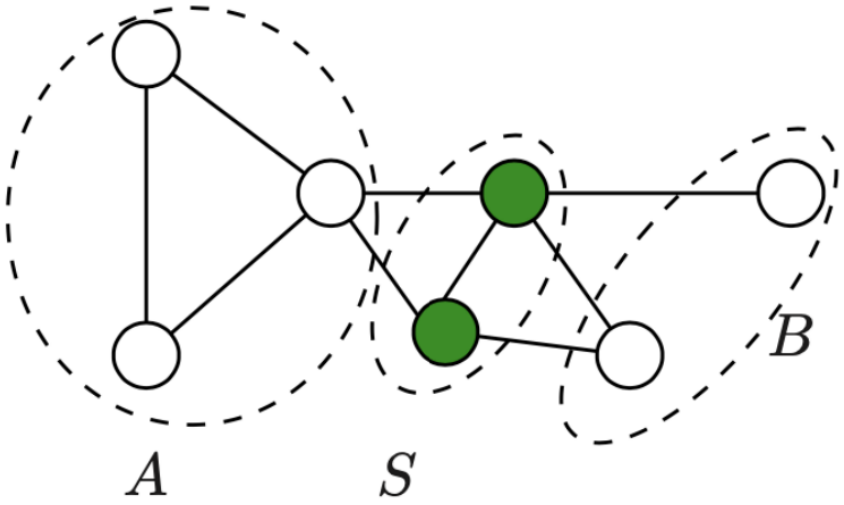


Figure 52: Example of a separator in a graph.

Graphical models are based on the notion of conditional independence.

Def. (Conditional independence)
Random variables X_1 and X_3 are **conditionally independent** given random variable X_2 if the conditional distribution

$$\mathbb{P}(X_1|X_2, X_3) = \mathbb{P}(X_1|X_2),$$

and we usually indicate this with $X_1 \perp\!\!\!\perp X_3|X_2$.

Remark. Given X_2, the knowledge of X_3 does not contribute to explain X_1.

Cliques. The concepts of cliques and separators allow us to define the factorization property of a graph. In fact, conditional independence allows to factorize the joint distribution of the variables of a graph into its cliques.

$$\begin{aligned} p(x_1, x_2, x_3) &= \psi_1(x_1, x_2)\psi_2(x_2, x_3) \\ &= p(x_1, x_2)p(x_3|x_2) \\ &= p(x_1|x_2)p(x_2, x_3) \end{aligned}$$

Hence we have an equivalence between conditional independence and clique factorization of the graph.

Example
Consider a graph of the form

$$1 \longrightarrow 2 \longrightarrow 3,$$

then we have two cliques,

$$\mathcal{C} = \{C_1, C_2\}, C_1 = \{1, 2\}, C_2 = \{2, 3\},$$

and one separator,

$$\mathcal{S} = \{S\}, S = \{2\}.$$

Hence, we have the following conditional independence,

$$X_1 \perp\!\!\!\perp X_3 | X_2.$$

Since the edge $1 \longrightarrow 3$ is missing, we can interpret it as imposing conditional independence given the separator.

Def. (Decomposable graph)
A graph G is **decomposable** $\iff$ the set of cliques of G can be ordered so as to satisfy the running intersection property. That is, for every $i = 2, \ldots, k$,

$$S_i = C_i \cap \bigcup_{j=1}^{i-1} C_j \implies S_i \in C_l \text{ for some } l < i - 1.$$

Remark. Although the ordering may not be unique, the graph uniquely determines the set of cliques and separators.

Def. (Pairwise Markov property)
We say that P satisfies the **pairwise Markov property** with respect to the undirected graph $G = (V, E)$ if for any pair of unconnected vertices $(s, t) \notin E$,

$$X_s \perp\!\!\!\perp X_t | X_{V \setminus \{s,t\}}.$$

Remark. In other words, if an edge between two nodes is absent then the two random variables are conditionally independent given all the other variables.

Def. (Local markov property)
We say that P satisfies the **local Markov property** with respect to the undirected graph $G = (V, E)$ if for any vertex $s \in V$,

$$X_s \perp\!\!\!\perp X_{V \setminus \mathcal{N}^+(s)} | X_{\mathcal{N}(s)},$$

where $\mathcal{N}(s) = \{t \in V : (s, t \in E)\}$ is the **neighborhood set** of s and $\mathcal{N}^+(s) = \mathcal{N}(s) \cup \{s\}$ is the **closure** of s.

Remark. In other words, the random variable X_s is conditionally independent of the rest of the nodes given its neighbors.

Def. (Global Markov property)
We say that P satisfies the **global Markov property** with respect to the undirected graph $G = (V, E)$ if for any triple of disjoint sets A, B, C such that C separates A and B,

$$X_A \perp\!\!\!\perp X_B | X_C.$$

Remark. Note that if P has a positive and continuous density with respect to Lebesgue measure, i.e. P does not assign zero probability to any assignment of the variables, the three Markov properties are equivalent.

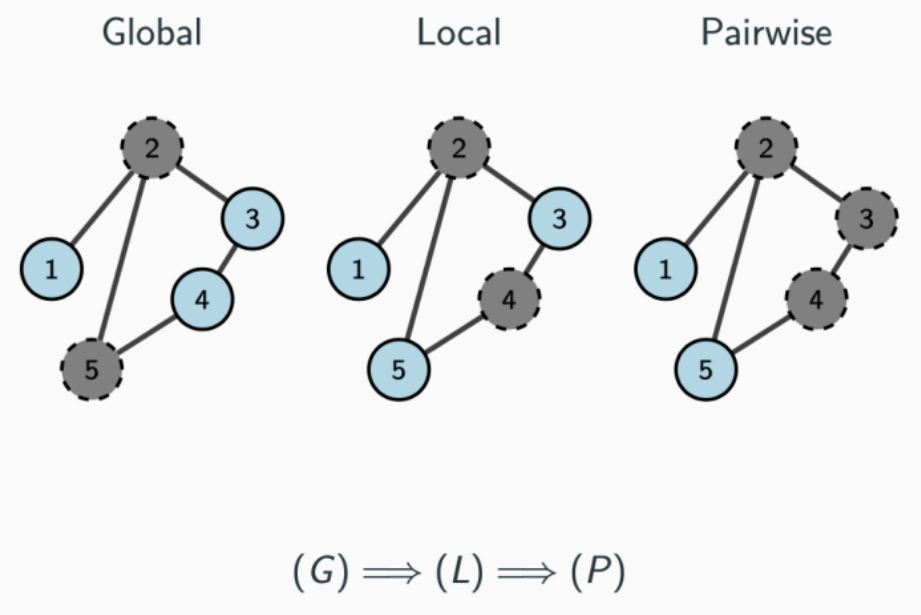


Figure 53: Comparison between the three Markov properties in a graph.

Markov chains. Note that a Markov chain is a special case of a graph in which the global Markov property holds. Markov chains are a chain-structured graph with edges

$$E = \{(1, 2), (2, 3), \ldots, (p - 1, p)\}.$$

In this case, each node $s \in \{2, \ldots, p - 1\}$ is a separator, which makes the "past" conditionally independent from the "future".

Def. (Conditional independence graph)
We say that a graphical model (G, P) is a **conditional independence graph** if G is undirected and the pairwise Markov property holds.

If P is Markov relative to G, the joint distribution of the graph can be decomposed as

$$p(x_V) = p(x_{C_1}) \prod_{j=1}^{k} p(x_{R_j} | x_{S_j}),$$

where $R_j = C_j \setminus S_j$. The main advantage is that now we have reduced the dimensionality of the problem, from p to the cardinality of the largest clique; this is typically a big advantage if the graph is very sparse.

21.3 Gaussian graphical models

Let us consider a special case of the model seen so far, in which

$$(X_1, \dots, X_p) \sim \mathcal{N}_p(\mu, \Sigma).$$

A Gaussian Graphical Model is a conditional independence graph in which P is a multivariate Gaussian. In this case, the equivalence between pairwise, local, and global Markov properties holds since the Gaussian distribution is absolutely continuous.
Since

$$p_{\mu,\Sigma}(x) = \dots$$

we can reparametrize it in terms of the canonical parameters

$$p_{\gamma,\Theta} = \exp\left\{\sum_{k=1}^{p} \gamma_s x_s - \frac{1}{2}\sum_{s,t=1}^{p} \theta_{st} x_s x_t - A(\Theta)\right\},$$

where $\Theta = (\theta_{ij})_{i,j} = \Sigma^{-1}$ is the precision matrix, $\gamma = \Theta\mu$, and $A(\Theta) = \dots$. Since conditional independence means that $\theta_{st} = 0$ if $(s,t) \notin E$, the conditional independence graph has a nice interpretation in terms of placing elements $\theta_{st} = 0$ in the **precision matrix** of the distribution.

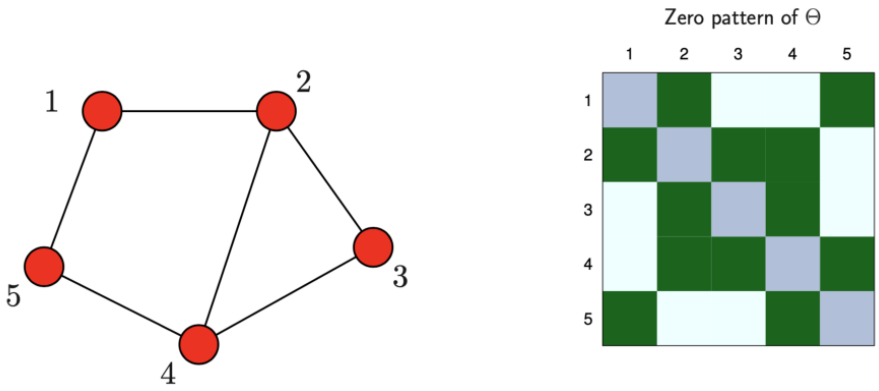


Figure 54: Relationship between the conditional independence and graph of the distribution.

Remark. Note that this result translates into an "if and only if" interpretation of the pairwise Markov property. Hence, in Gaussian Graphical model we have a stronger result than the Markov property.

Partial correlation The partial correlation of X_S and X_t given $X_{V\setminus\{s,t\}}$ as

$$\rho_{st|V\setminus\{s,t\}} = -\frac{\theta_{st}}{\sqrt{\theta_{ss}\theta_{tt}}}.$$

This is important because partial correlation is directly related to regression.

Consider the following regression model

$$X_s = \beta_t X_t + \sum_{r \in V \setminus \{s,t\}} \beta_r X_r + \varepsilon,$$

then we can show that

$$\beta_t = -\frac{\theta_{st}}{\theta_{ss}},$$

hence, the edge between s and t is present if and only if the corresponding coefficient in the regression model is different from zero. This links the structure learning problem with variable selection and lasso-based procedures.

21.4 Graphical lasso

As said, one of the main problem is that of learning the structure of the graph from the data. Here, we discuss some approaches based on ℓ_1 penalization, but the problem is more general and can be tackled, e.g., by using a hypothesis testing approach (Nguyen and Chiogna 2021).

For simplicity, and since we are only interested in the structure of the graph, we assume that $\mu = 0$. Assuming $X \sim \mathcal{N}_p(0, \Theta^{-1})$, then the rescaled log-likelihood can be written as

$$L(\Theta; X) = \log\det\Theta - \operatorname{tr}(S\Theta),$$

where $S = X^\top X/n$ is the empirical covariance matrix. The maximum likelihood estimator of Θ is

$$\widehat{\Theta} = S^{-1},$$

which has two problems:

1. in general, S^{-1} will have no elements equal to zero;
2. for $p > n$, S will be singular and the MLE cannot be computed.

One solution is to add an ℓ_1 penalty, which makes the problem become

$$\underset{\Theta \succeq 0}{\operatorname{argmax}} \left\{\log\det\Theta - \operatorname{tr}(S\Theta) - \lambda\rho_1(\Theta)\right\},$$

where we use a ℓ_1 norm penalization on the off-diagonal elements of Θ,

$$\rho_1(\Theta) = \sum_{s \neq t} |\theta_{st}|.$$

This problem is termed the **graphical lasso** by Friedman et al. (2008). A necessary and sufficient condition for Θ to maximize the penalized log-likelihood

$$\Theta^{-1} - S - \lambda\Gamma(\Theta) = 0,$$

where $\Gamma(\Theta)$ is a $p \times p$ matrix whose (i, j) element is the subgradient of $|\theta_{ij}|$,

$$\Gamma(\theta_{ij}) = \begin{cases} 1 & \text{if } \theta_{ij} > 0 \\ -1 & \text{if } \theta_{ij} < 0 \\ a \in [-1, 1] & \text{if } \theta_{ij} = 0. \end{cases}$$

21.4.1 Estimation

The graphical lasso is as convex optimization problem and a unique solution can be found via a blockwise coordinate descent algorithm. The main idea is to partition all matrices into one column versus the rest, and for convenience we choose the last:

$$\Theta = \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^\top & \theta_{22} \end{pmatrix}, \quad S = \begin{pmatrix} S_{11} & s_{12} \\ s_{12}^\top & s_{22} \end{pmatrix}.$$

Now we can rewrite this problem as

$$W_{11}\beta - s_{12} + \lambda\Gamma(\theta_{12}) = 0, \tag{75}$$

where

- $\beta = -\theta_{12}/\theta_{22}$.
- W_{11} is the $(p-1) \times (p-1)$ block of Θ^{-1}.
- s_{12} and θ_{12} are $p-1$ non-diagonal elements of the p^{th} row and column of S and Θ.
- s_{22} and θ_{22} are the p^{th} diagonal element of S and Θ.

Hence, this is equivalent to a modified version of the estimating equations for a lasso regression, since the subgradient equations are

$$\frac{1}{n} Z^\top Z\beta - \frac{1}{n} Z^\top y + \lambda \operatorname{sgn}(\beta) = 0.$$

Hence if we set

$$s_{12} = \frac{1}{n} Z^\top y$$

$$W_{11} = \frac{1}{n} Z^\top Z,$$

then we can solve each blockwise step by using a modified algorithm for the lasso, and treating each variable as the response and the other $p-1$ as the predictors.

A better solution can be obtained by noticing that if the solution takes the form

$$\Theta = \begin{pmatrix} \Theta_{11} & 0 & \dots & 0 \\ 0 & \Theta_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Theta_{kk} \end{pmatrix},$$

for some ordering of the variables, then the graphical lasso problem can be solved separately for each block, and the solution is constructed from the individual solutions (Witten et al., 2011).

From that paper, a necessary and sufficient condition for the solution to the graphical lasso problem to be block diagonal with blocks $C_1, \ldots, C_k$ is that $|S_{ij}| < \lambda$ for all $i \in C_k$, $j \in C_l$, $k \neq l$. This implies that one can simply screen the off-diagonal elements in a given column of S to determine if the corresponding node is unconnected from all other nodes.

Faster algorithms. The two statements above can be exploited to define two faster algorithms, compared to the "regular" graphical lasso, which requires $O(p^3)$ operations:

1. Identify all the q fully unconnected nodes, define an ordering for which each unconnected node is a group and the rest of the $p - q$ nodes is the last group: this leads to $O(p^2 + (p-q)^3)$ operations.
2. Identify the k connected components of a graph and solve the k graphical lasso problems: this leads to $O(p^2 + \sum_k |C_k|^3)$ operations. This is a big advantage if k is large or if the $|C_k|'s$ are small compared to p.

While algorithm 2 is generally faster, algorithm 1 could be better if there are many unconnected nodes.

21.5 Neighborhood-based methods

An alternative approach to structure learning is to leverage the local Markov property. Instead of using W_{11} in (75), we could use S. This turns out to be equivalent to performing a multivariate linear regression on all the variables together.

If we regress X_s on $X_{V\setminus\{s\}}$, thanks to the local Markov property it suffices to regress X_s on $X_{\mathcal{N}(s)}$. This approach was proposed by Meinshausen and Bühlmann (2006) and considers the p **nodewise regression** problems,

$$X_s = \sum_{t \neq s} \beta_t^{(s)} X_t + \varepsilon^{(s)}, \quad s = 1, \ldots, p,$$

and we can use the lasso to obtain an estimate of the neighborhood of s,

$$\widehat{S}^{(s)} = \{t : \widehat{\beta}^{(s)}(\lambda) \neq 0\}.$$

Problem. A slight issue is that we will have two estimated coefficients $\widehat{\beta}_t^{(s)}$ and $\widehat{\beta}_s^{(t)}$ for each edge (s, t), and we need to decide what to do when the two regression do not agree.

1. `OR` rule: include edge if either one of them are $\neq 0$.
2. `AND` rule: include edge if both of them are $\neq 0$.

Remark. `AND` rule of course produces sparser solutions.

Remark. Meinshausen and Bühlmann (2006) shows that the conditions for consistent estimation are weaker in the case of nodewise regression than the case of graphical lasso. Therefore, nodewise regression is more general, although at first it may seem less powerful than the simultaneous approach.

21.5.1 Faithfulness assumption

Instead of using a penalized approach, we might introduce a new assumption based on the graph in order to infer the properties using a hypothesis test.

Recall that by the Markov property

$$C \text{ separates } A \text{ and } B \implies X_A \perp\!\!\!\perp X_B | X_C,$$

although the reverse is not true in general. This means that we can infer from the graph some conditional independence properties of the distribution P. However, P may include other conditional independence relations.

Def. (Faithfulness)
The distribution P is **faithful** to the graph G if the following equivalence holds,

$$C \text{ separates } A \text{ and } B \implies X_A \perp\!\!\!\perp X_B | X_C, \tag{76}$$

Remark. $\implies$ follows from the Markov property, whereas the faithfulness assumption requires $\impliedby$. In other words, all conditional independences can be read from the graphical separation of the nodes.

Example (Failure of faithfulness)
As an esample of failure of faithfulness we can consider the fully-connected graph ...
and consider the following relationships

$$X_1 = \varepsilon$$

$$X_2 = \alpha X_1 + \varepsilon_2$$

$$X_3 = \beta X_1 + \gamma X2 + \varepsilon_3,$$

where $\varepsilon_1, \varepsilon_2, \varepsilon_3 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. For instance,

$$\text{Cov}(X_1, X_3) = \beta + \alpha\gamma,$$

and we can enforce marginal independence by choosing the coefficients so that

$$\beta + \alpha\gamma = 0,$$

e.g. $\alpha = \beta = 1$ and $\gamma = -1$.

Assuming faithfulness, we have that

$$(s,t) \notin E \iff \rho_{st|C} = 0 \quad \text{for all } C \subseteq V \setminus \{s,t\}.$$

This means that we can hierarchically screen marginal correlations $\rho_{st|C} = 0$ with $|C|$ small. If one of them is zero, we know that there is no edge between s and t.

This proposition is the basis for the so-called "Peter-Clarke" (PC) algorithm (Spirtes et al., 2001), which is an iterative multiple testing procedure for inferring zero partial correlations, typically by testing that the coefficients of a local regression model are equal to 0.

Algorithm 6 PC algorithm

Start with a sample $X_1, \ldots, X_p$ with a fully connected graph.
$d \leftarrow 0$
while $d < m$ **or** $E = \emptyset$ **do**
    Select an edge (s, t) in G.
    Choose a set $C \subseteq \mathcal{N}(s) \setminus \{t\}$ with $|C| = d$.
    Test the conditional independence $X_s \perp\!\!\!\perp X_t | C$
    If the variables are conditionally independent remove the edge (s, t).
    Repeat for all $C \subseteq \mathcal{N}(s) \setminus \{t\}$ with $|C| = d$ or until (s, t) is removed.
    Repeat for all pairs (s, t) adjacent in G.
    $d \leftarrow d + 1$.
end while

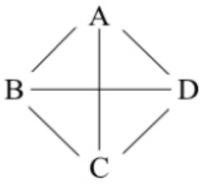
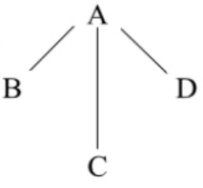
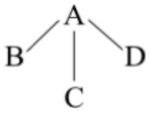
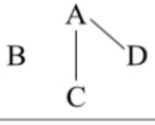
Level	Graph	#CI tests	Test	Result	Updated Graph
1		1	I(A, B)?	No	
		2	I(A, C)?	No	
		3	I(A, D)?	No	
		4	I(B, A)?	No	
		5	I(B, C)?	Yes	
		6	I(B, D)?	Yes	
		7	I(C, A)?	No	
		8	I(C, D)?	Yes	
		9	I(D, A)?	No	
2		10	I(A, B | C)?	Yes	
		11	I(A, C | D)?	No	
		12	I(A, D | C)?	No	

Figure 55: Example of PC algorithm applied to a small graph.

Remark. The `AND` and `OR` rules still find usefulness here, as well as type-I error and multiple comparison error control.

References

Box, G. E. P. et al. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. 2nd edition. Hoboken, N.J: Wiley-Interscience.

Buhlmann, P. and Geer, S. V. D. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. 2011° edizione. Heidelberg ; New York: Springer-Nature New York Inc.

Bühlmann, P. et al. (2014). «Discussion: "A Significance Test for the Lasso"». In: *The Annals of Statistics* 42.2. arXiv: 1405.6792 [math, stat].

Burtini, G. et al. (2015). «A Survey of Online Experiment Design with the Stochastic Multi-Armed Bandit». In: *arXiv:1510.00757 [cs, stat]*. arXiv: 1510.00757 [cs, stat].

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. New York: Cambridge University Press.

Friedman, J. et al. (2008). «Sparse Inverse Covariance Estimation with the Graphical Lasso». In: *Biostatistics* 9.3, 432–441.

Hastie, T. et al. (2015). *Statistical Learning with Sparsity*. 1st edition. Boca Raton: Routledge.

Jacob, L. et al. (2009). «Group Lasso with Overlap and Graph Lasso». In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. New York, NY, USA: Association for Computing Machinery, 433–440.

Lai, T. L. and Robbins, H. (1985). «Asymptotically Efficient Adaptive Allocation Rules». In: *Advances in Applied Mathematics* 6.1, 4–22.

Lai, T. L. (2001). «Sequential Analysis: Some Classical Problems and New Challenges». In: *Statistica Sinica* 11.2, 303–351.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press.

Lockhart, R. et al. (2014). «A Significance Test for the Lasso». In: *The Annals of Statistics* 42.2, 413–468.

Meinshausen, N. and Bühlmann, P. (2006). «High-Dimensional Graphs and Variable Selection with the Lasso». In: *The Annals of Statistics* 34.3, 1436–1462.

Meng, X.-L. (2018). «Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election». In: *The Annals of Applied Statistics* 12.2, 685–726.

Shin, J. et al. (2019). *Are Sample Means in Multi-Armed Bandits Positively or Negatively Biased?* arXiv: 1905.11397 [math, stat].

Simon, N. et al. (2013). «A Sparse-Group Lasso». In: *Journal of Computational and Graphical Statistics* 22.2, 231–245.

Spirtes, P. et al. (2001). *Causation, Prediction, and Search*. Ed. by F. Bach. Second. Adaptive Computation and Machine Learning Series. Cambridge, MA, USA: A Bradford Book.

Tibshirani, R. et al. (2005). «Sparsity and Smoothness via the Fused Lasso». In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.1, 91–108.

Wald, A. (1945). «Sequential Tests of Statistical Hypotheses». In: *The Annals of Mathematical Statistics* 16.2, 117–186.

Witten, D. M. et al. (2011). «New Insights and Faster Computations for the Graphical Lasso». In: *Journal of Computational and Graphical Statistics* 20.4, 892–900.

Zou, H. et al. (2007). «On the "Degrees of Freedom" of the Lasso». In: *The Annals of Statistics* 35.5, 2173–2192.

Part IV

Models for complex and dependent data

Instructor: Mauro Bernardi

Lecture 22: Bayesian linear regression

2022-05-27

22.1 Bayesian regression

We start from the specification of Bayesian linear regression and how this is used in practice. An easy framework to understand calculations, since it is possible to generalize and extend those computations to other type of models, such as probit regression, GLMs, and semiparametric spline/basis regression. We consider the Gaussian linear model

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2),$$

for which we need to postulate a prior distribution on $(\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^{p+1}$. Several alternatives of prior distributions might be noninformative, conjugate, independent, informative, Zellner's g-prior, hierarchical, ...

Prop. 11 (Full conditionals linear model)
Under $\sigma^2 \sim IGa(\nu, \lambda)$ and $\boldsymbol{\beta}|\sigma_\varepsilon^2 \sim N(\boldsymbol{\beta}_0, \sigma_\varepsilon^2 D_0)$, the full conditional distributions of $(\boldsymbol{\beta}, \sigma_\varepsilon^2)$ are

$$\pi(\boldsymbol{\beta}|y, X, \sigma_\varepsilon^2) \propto N_p(\widehat{\Sigma}_n(D_0^{-1}\boldsymbol{\beta}_0 + X^\top y), \sigma_\varepsilon^2 \widehat{\Sigma}_n) \tag{77}$$

$$\pi(\sigma_\varepsilon^2|y, X, \boldsymbol{\beta}) \propto IGa\left(\nu + \frac{n+p}{2}, \lambda + \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{2}\right), \tag{78}$$

where $\tilde{X} = \begin{pmatrix} X \\ D_0^{-1/2} \end{pmatrix}$, $\tilde{y} = \begin{pmatrix} y \\ 0_p \end{pmatrix}$, $\varepsilon = y - X\boldsymbol{\beta}$ and $\widehat{\Sigma}_n = (X^\top X + D_0^{-1})^{-1}$.

Remark. Full conditionals are needed in order to be able to efficiently sample using the Gibbs sampler algorithm.

Bias. Prior information introduces bias, which in this case is given by $(X^\top X + D_0^{-1})^{-1}(D_0^{-1}\boldsymbol{\beta}_0)$. All Bayesian solutions are biased but we allow it in order to improve stability and efficiency of the estimation procedures

Conjugacy. The choice of a conjugate prior is such that

$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta}|\sigma^2) \times \pi(\sigma^2),$$

where $\pi(\sigma^2) = \text{IGa}(\cdot|\nu, \lambda)$ is the conjugate choice, although it is not a good idea in high dimension.

Computation. Most of the statistical learning problems face the problem of a quantity which has a structure of

$$\widehat{\Sigma}_n = (X^\top X + D_0^{-1})^{-1},$$

which has usually a $O(p^3)$ computational cost.

Prop. 12 (Joint distribution)
For the linear regression model, then the joint posterior distribution in the conjugate case is

$$\pi(\boldsymbol{\beta}, \sigma_\varepsilon^2 | y, X) = \left(\sigma_\varepsilon^2\right)^{-\frac{n+p}{2}+\nu+1} \exp\left\{-\frac{\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{2\sigma_\varepsilon^2} - \frac{\lambda}{\sigma_\varepsilon^2}\right\} \exp\left\{-\frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top D_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{2\sigma_\varepsilon^2}\right\}, \quad (79)$$

where $\varepsilon = y - X\boldsymbol{\beta}$.

Bayes' theorem. The most important result is that

$$\pi(\theta, y) = L(y|\theta)\pi(\theta),$$

and once we can write the joint distribution we can run simulation algorithms. Note that L and π are independent, hence we should either not observe the data or observe a **small subset of data** before calibrating the prior.

In some special cases, the **marginal likelihood** (or **model evidence**) $m(y)$ of the model is available and fully analytically tractable, from which we can calculate the posterior $\pi(\theta|y)$ as

$$\pi(\theta|y) = \frac{\pi(\theta, y)}{m(y)} = \frac{L(y|\theta)\pi(\theta)}{\int_\Theta l(y|\theta) \cdot \pi(\theta) d\theta}. \quad (80)$$

In the linear model, we do so by integrating out the parameters $(\boldsymbol{\beta}, \sigma_\varepsilon^2)$ from the joint posterior.

Model evidence. The name model evidence stems from the fact that θ has been canceled out, and therefore this quantity describes how well the model fits the data when averaged over the prior distribution. The BIC is an asymptotic approximation to $m(y)$, which is used as a model selection tool. Hence, $\pi(\theta|y)$ is used to make inference on the parameter, whereas $m(y)$ can be used to assess how well the model fits the data.

Prop. 13 (Marginal likelihood under the conjugate prior)
Assuming without loss of generality $\boldsymbol{\beta}_0 = 0$*, then the marginal likelihood for the Bayesian linear model under the conjugate prior is*

$$m(y|X) \propto |\tilde{X}^\top \tilde{X}|^{-1/2} |D_0|^{-1/2} \left(\lambda + \frac{S^2}{2}\right)^{-(\nu+n/2)}, \quad (81)$$

where $S^2 = y^\top H y$, $H = I_n - X\widehat{\Sigma}_n X^\top$ *and* $\widehat{\Sigma}_n = (X^\top X + D_0^{-1})^{-1}$.

Remark. The model is encapsulated both in the functional form of (81) and in the value of the hyperparameters.

Remark. This is proportional to a multivariate t distribution and can be obtained by considering the joint posterior and factorizing it in terms of the full conditionals of $\boldsymbol{\beta}$ and σ_ε^2.

Proof.
Given the joint posterior distribution for $(\boldsymbol{\beta}, \sigma_\varepsilon^2)$ and assuming $\boldsymbol{\beta}_0 = \mathbf{0}$,

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \sigma_\varepsilon^2 | y, X) &= \left(\sigma_\varepsilon^2\right)^{-\frac{n+p}{2}+\nu+1} \exp\left\{-\frac{\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{2\sigma_\varepsilon^2} - \frac{\lambda}{\sigma_\varepsilon^2}\right\} \exp\left\{-\frac{\boldsymbol{\beta}^\top D_0^{-1} \boldsymbol{\beta}}{2\sigma_\varepsilon^2}\right\} \\
&= \left(\sigma_\varepsilon^2\right)^{-p/2} \exp\left\{-\frac{y^\top y}{2\sigma_\varepsilon^2}\right\} |D_0|^{-1/2} |\widehat{\Sigma}_n|^{1/2} |\widehat{\Sigma}_n|^{-1/2} \exp\left\{-\frac{\lambda}{\sigma_\varepsilon^2}\right\} \\
&\quad \times \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(\beta - \widehat{\Sigma}_n X^\top y)^\top \widehat{\Sigma}_n^{-1} (\beta - \widehat{\Sigma}_n X^\top y)\right\} \\
&\quad \times \exp\left\{\frac{1}{2\sigma_\varepsilon^2} y^\top X \widehat{\Sigma}_n X^\top y\right\} \left(\sigma_\varepsilon^2\right)^{-(n/2+\nu+1)},
\end{aligned}$$

integration with respect to $\boldsymbol{\beta}$ can be done in a straightforward way by noticing that only the second row of the last equality contains $\boldsymbol{\beta}$, and it is the kernel of a $N_p(\widehat{\Sigma}_n X^\top y, \sigma_\varepsilon^2 \widehat{\Sigma}_n)$. Thus the integral with respect to $\boldsymbol{\beta}$ amounts to the normalizing the density,

$$\frac{\sigma_\varepsilon^2}{(2\pi)^{p/2} \det \widehat{\Sigma}_n^{1/2}},$$

and we obtain

$$\pi(\sigma_\varepsilon^2, y | X) \propto \exp\left\{-\frac{2\lambda + S^2}{2\sigma_\varepsilon^2}\right\} |D_0|^{-1/2} |\widehat{\Sigma}_n|^{1/2} \left(\sigma_\varepsilon^2\right)^{-(n/2)+\nu+1}, \tag{82}$$

and $S^2 = y^\top y - y^\top X (X^\top X + D_0^{-1})^{-1} X^\top y$. Therefore the full conditional distribution of σ_ε^2 where the regression parameters are marginalised out is

$$\sigma_\varepsilon^2 | y, X \sim \mathrm{IGa}(\nu + n/2, \lambda + S^2/2).$$

Again, integrating out σ_ε^2 from equation (82) yields the marginal likelihood which is proportional to (81).

□

We provide an alternative formulation for the full conditional distribution of $\boldsymbol{\beta}$, which will find wide application in later discussions about dynamic linear models.

Prop. 14 (Full conditional of $\boldsymbol{\beta}$)
For the linear regression model with conjugate priors the full conditional distribution of $\boldsymbol{\beta}|\sigma_\varepsilon^2$ can be rewritten as

$$\boldsymbol{\beta}|y, X, \sigma_\varepsilon^2 \sim N(\widehat{\boldsymbol{\beta}}_n, \sigma_\varepsilon^2 \widehat{\Sigma}_n), \tag{83}$$

where

$$F = I_n + XD_0X^\top \tag{84}$$

$$K = D_0X^\top F^{-1} \tag{85}$$

$$\widehat{\Sigma}_n = (I_p - KX)D_0, \tag{86}$$

$$\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + K(y - X\boldsymbol{\beta}_0) \tag{87}$$

Remark. The quantities in (87) and (86) are the same quantities that we obtained before,

$$\widehat{\boldsymbol{\beta}}_n = \widehat{\Sigma}_n(D_0^{-1}\boldsymbol{\beta}_0 + X^\top y)$$
$$\widehat{\Sigma}_n = (X^\top X + D_0^{-1})^{-1}.$$

Remark. In (87), the quantity $y - X\boldsymbol{\beta}_0$ represents the prior error term by which we correct the prior belief $\boldsymbol{\beta}_0$ using the Kalman gain K.

Proof.
For the linear regression under conjugate prior, we can write the joint distribution of $(y, \boldsymbol{\beta})|\sigma_\varepsilon^2$ as a Gaussian random variable. The marginal moments of y are

$$\mathbb{E}[y] = \mathbb{E}_{\boldsymbol{\beta}}[\mathbb{E}_{y|\boldsymbol{\beta}}[y]] = \mathbb{E}_{\boldsymbol{\beta}}[X\boldsymbol{\beta}] = X\boldsymbol{\beta}_0.$$
$$\mathbb{V}[y] = \mathbb{E}_{\boldsymbol{\beta}}[\mathbb{V}_{y|\boldsymbol{\beta}}[y]] + \mathbb{V}_{\boldsymbol{\beta}}[\mathbb{E}_{y|\boldsymbol{\beta}}[y]] = \sigma_\varepsilon^2(I_n + XD_0X^\top) = \sigma_\varepsilon^2 F,$$

and so $y \sim N(X\boldsymbol{\beta}_0, \sigma_\varepsilon^2 F)$. As for the covariance, we can write

$$\begin{aligned}\mathrm{Cov}(y, \boldsymbol{\beta}) &= \mathbb{E}[(y - \mathbb{E}[y])(\boldsymbol{\beta} - \mathbb{E}[\boldsymbol{\beta}])^\top] = \mathbb{E}[(X(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \varepsilon)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top] \\ &= X\,\mathbb{E}[(\boldsymbol{\beta} - \boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top] = \sigma_\varepsilon^2 XD_0.\end{aligned}$$

Jointly, we have that

$$\begin{pmatrix} y \\ \boldsymbol{\beta} \end{pmatrix} \Big| \sigma_\varepsilon^2 \sim N_{p+n}\left(\begin{pmatrix} X\boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_0 \end{pmatrix}, \sigma_\varepsilon^2 \begin{pmatrix} F & XD_0 \\ D_0X^\top & D_0 \end{pmatrix}\right),$$

and by leveraging standard results for the Gaussian distribution (see Appendix), we have that

$$\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + D_0X^\top F^{-1}(y - X\boldsymbol{\beta}_0) = \boldsymbol{\beta}_0 + K(y - X\boldsymbol{\beta}_0),$$
$$\widehat{\Sigma}_n = D_0 - D_0X^\top F^{-1}XD_0 = (I_p - KX)D_0.$$

□

Lecture 23: Bayesian linear regression (ii)

2022-06-01

From the previous proof, we observe that the Kalman gain can be written as

$$K = D_0 X^\top F^{-1} = \operatorname{Cov}(\boldsymbol{\beta}, Y)\, \mathbb{V}[Y]^{-1},$$

whereas $\varepsilon_0 = y - X\boldsymbol{\beta}_0$ are interpreted as prior residuals.

Prop. 15 (Alternative formula for $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{n}}$)
The update equation for $\widehat{\Sigma}_n$ can also be written as

$$\widehat{\Sigma}_n = (I_n - KX)D_0 = (I_p - KX)D_0(I_p - KX)^\top + KK^\top.$$

Proof.

$$\begin{aligned}
\widehat{\Sigma}_n &= (I_n - KX)D_0 = (I_p - KX)D_0(I_p - KX)^\top + KK^\top. \\
&= D_0 - 2KXD_0 + K(I_n + XD_0X^\top)K^\top \\
&= D_0 - 2KXD_0 + KFK^\top \\
&= D_0 - 2KXD_0 + KXD_0 \\
&= (I_n - KX)D_0 && \text{(by def. of } F \text{ and } K\text{)}
\end{aligned}$$

□

Remark. This form of the variance update is more computationally expensive, but it also leads to more stable results. The subtraction $I_n - KX$ does not guarantee that $\widehat{\Sigma}_n$ is symmetric because of floating-point rounding errors, which are frequent enough to make it a relevant issue.

Prop. 16 (Optimality)
The Bayesian updating is optimal, in the sense that the estimated variance has the smallest value of $\operatorname{tr}\widehat{\Sigma}_n$*,*

$$K = \operatorname*{argmin}_K \operatorname{tr}\widehat{\Sigma}_n$$

Proof.
Differentiating with respect to K yields

$$\begin{aligned}
\frac{\partial}{\partial K}\operatorname{tr}\widehat{\Sigma}_n &= \frac{\partial}{\partial K}\operatorname{tr}\left(D_0 - 2KXD_0 + KFK^\top\right) \\
&= \frac{\partial}{\partial K}\left(\operatorname{tr} D_0 - 2\operatorname{tr}(KXD_0) + \operatorname{tr}(KFK^\top)\right) \\
&= -2(XD_0)^\top + 2FK^\top,
\end{aligned}$$

and equating it to zero yields $KF = XD_0 \iff K = XD_0F^{-1}$.

□

Prop. 17 (Marginal posterior of $\boldsymbol{\beta}$)
For the linear regression model under $\sigma_\varepsilon^2 \sim IGa(\nu, \lambda)$ and $\boldsymbol{\beta}|\sigma_\varepsilon^2 \sim N(\boldsymbol{\beta}_0, \sigma_\varepsilon^2 D_0)$ the marginal posterior distribution of the regression parameters $\boldsymbol{\beta}$ is

$$\pi(\boldsymbol{\beta}|y, X) \propto St_{n+2\nu}\left(\widehat{\boldsymbol{\beta}}_n, \frac{2\lambda + S^2}{n + 2\nu}\widehat{\Sigma}_n\right).$$

Prop. 18 (Full conditional of $\boldsymbol{\beta}$)
For the linear regression model under $\sigma_\varepsilon^2 \sim IGa(\nu, \lambda)$ and $\boldsymbol{\beta}|\sigma_\varepsilon^2 \sim N(\boldsymbol{\beta}_0, \sigma_\varepsilon^2 D_0)$ the full conditional distribution of the regression parameters $\boldsymbol{\beta}$ after further n_0 observations is

$$\boldsymbol{\beta}|y, y_0, X, X_0, \sigma_\varepsilon^2 \sim N(\widehat{\boldsymbol{\beta}}_{n,0}, \sigma_\varepsilon^2\widehat{\Sigma}_{n,0}),$$

where

$$\begin{aligned}
K_{n,0} &= \widehat{\Sigma}_n X_0^\top F_{n,0}^{-1} \\
F_{n,0} &= I_r + X_0\widehat{\Sigma}_n X_0^\top \\
\widehat{\Sigma}_{n,0} &= (I_p - K_{n,0}X_0)\widehat{\Sigma}_n \\
\widehat{\boldsymbol{\beta}}_{n,0} &= \widehat{\boldsymbol{\beta}}_n + K_{n,0}(y_0 - X_0\widehat{\boldsymbol{\beta}}_n),
\end{aligned}$$

and $\widehat{\boldsymbol{\beta}}_n = \widehat{\Sigma}_n X^\top y$.

Remark. These equations are the same as before, and are obtained simply by replacing $\boldsymbol{\beta}_0$ with $\widehat{\boldsymbol{\beta}}_n$ and D_0 with $\widehat{\Sigma}_n$, that is, using the posterior distribution as the new prior.

Remark. The usefulness of these relations is that we can update the posterior distribution of our model without having to recompute the full posterior distribution every time a new observation is available. This is especially relevant if

a) data is collected over time (Kalman filter);

b) we are not able to compute the posterior on the whole dataset in one sweep.

Prop. 19 (Posterior predictive distribution)
For the linear regression model under $\sigma_\varepsilon^2 \sim IGa(\nu, \lambda)$ and $\boldsymbol{\beta}|\sigma_\varepsilon^2 \sim N(\boldsymbol{\beta}_0, \sigma_\varepsilon^2 D_0)$ the posterior predictive distribution for $y_0 \in \mathbb{R}^r$ associated to the design matrix X_0 is

$$\pi(y_0|y, X, X_0) \propto St_{n+2\nu}\left(X_0\widehat{\boldsymbol{\beta}}_n, \frac{2\lambda + S^2}{n + 2\nu}F\right),$$

where $S^2 = y^\top Hy$, $H = I_n - X\widehat{\Sigma}_n X^\top$, $F = I_r + X_0\widehat{\Sigma}_n X_0^\top$.

Proof.

□

Remark. Recall that the marginal likelihood (81) provides information about the **in-sample model fit** when we integrate out the prior parameters. Here, the predictive distribution is a "marginal likelihood" that provides information on the new observations y_0.

Proof.
We obtain the posterior predictive distribution analogously to the marginal likelihood in (81), by replacing $\boldsymbol{\beta}_0$ with $\widehat{\boldsymbol{\beta}}_n$ and D_0 with $\widehat{\Sigma}_n$.

□

Example (Jeffreys' prior)
For the linear regression mode, the Jeffreys' prior assumes

$$\pi(\boldsymbol{\beta}, \sigma_\varepsilon^2) \propto \frac{1}{\sigma_\varepsilon^2}.$$

This prior distribution is such that the marginal likelihood $m(y|X)$ is still finite and thus the posterior distribution is still proper.

Remark. If $p > n$, this prior is not useful since $D_0 = 0_{p\times p}$ and $\widehat{\Sigma}_n = (X^\top X)^{-1}$ is not invertible.

Exercise. Find the posterior distribution, full conditionals, posterior predictive and the marginal likelihood under the Jeffreys' prior.

Proof.
The normal-inverse-gamma conjugate prior distribution can be written as

$$\pi(\beta, \sigma_\varepsilon^2) = \left(\sigma_\varepsilon^2\right)^{-\frac{n+p}{2}+\nu+1} \exp\left\{-\frac{\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}}{2\sigma_\varepsilon^2} - \frac{\lambda}{\sigma_\varepsilon^2}\right\} \exp\left\{-\frac{\boldsymbol{\beta}^\top D_0^{-1}\boldsymbol{\beta}}{2\sigma_\varepsilon^2}\right\},$$

and therefore the Jeffreys prior can be written as the limit of the above distribution for $D_0^{-1} \to \mathbf{0}_{p\times p}$, $\nu \to -\frac{p}{2}$, and $\lambda \to 0$. In this case, we have that

$$\beta|\sigma_\varepsilon^2, y \sim N_p\big(X(X^\top X)^{-1}X^\top y, \sigma_\varepsilon^2(X^\top X)^{-1}\big)$$

$$\sigma_\varepsilon^2|y \sim IGa\left(\frac{n}{2}, \frac{\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}}{2}\right).$$

The posterior predictive then becomes

$$\pi(y_0|y, X, X_0) \propto St_{n-p}\left(X_0\widehat{\beta}_{\text{ols}}, \frac{\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}}{n-p}F\right).$$

□

Example (Conditionally conjugate prior)
The conditionally conjugate prior postulates a prior distribution of the form

$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta})\pi(\sigma^2),$$

where $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, D_0)$ and $\sigma_\varepsilon^2 \sim \text{IGa}(\nu, \lambda)$. Notice that in this case the prior distribution for $\boldsymbol{\beta}$ does not depend on σ_ε^2.

Remark. The problem with this prior distribution is that we do not have a closed-form expression for the model evidence.

Remark. However, since we can compute $\pi(\boldsymbol{\beta}|\sigma^2)$ and $\pi(\sigma^2|\boldsymbol{\beta})$ we can apply a Gibbs sampler in order to obtain a sample from the joint posterior distribution.

Exercise. Find the posterior, the full conditionals of $\boldsymbol{\beta}$ and σ_ε^2, the posterior predictive, and the marginal likelihood conditionally on σ_ε^2.

Proof.
For a fixed value of σ_ε^2, the posterior distribution of the regression coefficients is

$$\pi(\boldsymbol{\beta}|y, X, \sigma_\varepsilon^2) \propto N_p(\widehat{\Sigma}_n(D_0^{-1}\boldsymbol{\beta}_0 + X^\top y/\sigma_\varepsilon^2), \widehat{\Sigma}_n),$$

where $\widehat{\Sigma}_n = (X^\top X/\sigma_\varepsilon^2 + D_0^{-1})^{-1}$. Moreover, the posterior distribution of σ^2 is simply

$$\sigma^2|Y \sim IGa\left(\nu + \frac{n}{2}, \lambda + \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{2}\right).$$

□

23.1 Choosing the hyperparameters

23.1.1 Data-dependent prior

Data-dependent priors are a useful way of creating default priors, even though, strictly speaking, they violate the principle of not using the data twice. For instance, for $\boldsymbol{\beta}|\sigma^2$ we can use

$$\boldsymbol{\beta}_0 = \widehat{\boldsymbol{\beta}}_{\text{MLE}}, \quad D_0 = (X^\top X)^{-1}/\sigma^2,$$

whereas for σ^2 we can use

$$\nu = \frac{1}{2}, \quad \lambda = \frac{1}{2}\widehat{\sigma}^2_{\text{MLE}}.$$

Remark. These are both unit information priors, that is, the strength of the prior is equivalent to one sample.

Problem. As with the Jeffreys' prior before, the matrix $X^\top X$ is not invertible if $p > n$.

Data-dependence. The problem of using the information twice is on $\boldsymbol{\beta}_{\text{MLE}}$ and $\widehat{\sigma}^2_{\text{MLE}}$, since we use the observed data y. Instead, the choice of D_0 does not violate Bayes' theorem since we always

assume X to be fixed. A possible solution is to compute the prior on a small portion of data (X_0, y_0) and then calculate the posterior on the rest of the data.

23.1.2 Zellner's g-prior

A way to partially avoid the choice of hyperparameter is to use Zellner's g-prior, which also does not make use of the data twice.

Def. (Zellner's g-prior)
We define **Zellner's g-prior** as the following conditionally-conjugate prior distribution,

$$\boldsymbol{\beta}|\sigma_\varepsilon^2, g \sim N\big(\mathbf{0}, g \cdot \sigma_\varepsilon^2 (X^\top X)^{-1}\big)$$

$$\sigma_\varepsilon \sim \text{IGa}(\nu, \lambda),$$

where $g > 0$ is a free parameter that controls the strength of the prior over the posterior distribution.

Prop. 20 (Posterior under Zellner's prior)
Under Zellner's g-prior, the posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$ *is*

$$\boldsymbol{\beta}|\sigma^2 y, X, g \sim N(\widehat{\boldsymbol{\beta}}_n, \widehat{\Sigma}_n)$$

$$\sigma^2|y, X, g \sim IGa(\nu, +n, \lambda, +S_g^2),$$

where the updated hyperparameters are

$$H = I_n - X(X^\top X)^{-1} X^\top$$

$$S_g^2 = y^\top H y$$

$$\widehat{\boldsymbol{\beta}}_n = \frac{g}{1+g}(X^\top X)^{-1} X^\top y$$

$$\widehat{\Sigma}_n = \frac{\sigma^2 g}{1+g}(X^\top X)^{-1}.$$

Remark. The prior has covariance matrix proportional to the inverse Fisher information matrix for $\boldsymbol{\beta}$, similarly to a Jeffreys' prior.

Invariance. Zellner's g-prior is the only one that is invariant to change of scales in the predictors.

Proportionality. We observe that the posterior hyperparameters are such that $\widehat{\boldsymbol{\beta}}_n \propto \widehat{\boldsymbol{\beta}}_{\text{MLE}}$ and $\widehat{\Sigma}_n \propto (X^\top X)^{-1}$. Moreover, these solutions are invariant to the scale of the predictors, since both $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$ and the Fisher information are invariant as well.

Default choice. If the data is scaled to have a diagonal covariance matrix, we usually choose the default value of $g = 100$ so that we have a small amount of regularization with minimal consequences on the MLE.

Gibbs sampler. We can marginalize out σ^2 and obtain the joint distribution using

$$\pi(\boldsymbol{\beta}, \sigma^2 | y, X, g) = \pi(\boldsymbol{\beta} | y, X, g, \sigma^2) \pi(\sigma^2 | y, X, g),$$

and thus we can use simple Monte Carlo simulation to sample from the posterior distribution, which has obvious advantages over MCMC.

23.2 Model selection and sparsity

Sparse model regression are usually formulated in terms of solution to penalized loss functions,

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n} \|y - X\boldsymbol{\beta}\|_2^2 + \mathcal{P}_\lambda(\boldsymbol{\beta}),$$

where different penalty functions lead to the usual solutions such as lasso, ridge, bridge, etc...
All these model have a Bayesian interpretation in terms of prior distribution on the model parameters, although the resulting estimates do not have a sparse interpretation. Indeed, they only provide a solution in terms of posterior distributions but none of the parameters are exactly zero.

23.2.1 Bayesian lasso

As we have seen in the previous lectures, the lasso solution has the interpretation of being the maximum a posteriori of the regression coefficients a Laplace conditional prior distribution. However, we can consider the fully-Bayesian approach to perform inference on the model coefficients.

Def. (BLASSO prior)
The **BLASSO prior** is defined by the Laplace prior,

$$y | \boldsymbol{\beta}, \sigma_\varepsilon^2 \sim N(X\boldsymbol{\beta}, \sigma_\varepsilon^2 I_n)$$

$$\boldsymbol{\beta}_j | \lambda, \sigma_\varepsilon^2 \overset{\text{iid}}{\sim} \text{Laplace}(0, \sqrt{\sigma_\varepsilon^2}/\lambda)$$

Stochastic representation. Stochastic (or **hierarchical**) representation is a standard way of retrieving conjugacy when we have a non-conjugate setting such as the above. This amounts to introducing auxiliary conjugate variables with an assigned prior distribution, which are then marginalized out in order to retrieve the original model.

Since the prior is not conjugate, we can consider the following hierarchical representation that is equivalent to the BLASSO prior,

$$\begin{aligned}
y | \boldsymbol{\beta}, \sigma_\varepsilon^2 &\sim N(X\boldsymbol{\beta}, \sigma_\varepsilon^2 I_n) \\
\boldsymbol{\beta} | \sigma_\varepsilon^2, \boldsymbol{\tau} &\sim N(0, \sigma_\varepsilon^2 D_{\boldsymbol{\tau}}), \quad D_{\boldsymbol{\tau}} = \text{diag}(\tau_1^2, \dots, \tau_p^2) \\
\tau_j^2 &\overset{\text{iid}}{\sim} \text{Exp}(\lambda^2/2), \quad j = 1, \dots, p.
\end{aligned}$$

Proof.
Since they are independent, the marginal prior distribution of each $\beta_j|\sigma_\varepsilon^2$ can be obtained by integrating out with respect to the auxiliary parameter τ_j,

$$\begin{aligned}
\int_0^\infty \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2\tau}} e^{-\frac{z^2}{2\sigma_\varepsilon^2\tau}} \frac{\lambda^2}{2} e^{-\frac{\lambda^2\tau}{2}} d\tau &= \frac{\lambda^2}{2} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \int_0^\infty \tau^{-\frac{1}{2}} e^{-\frac{1}{2}\left(\frac{z^2}{\sigma_\varepsilon^2\tau + \lambda^2\tau}\right)} d\tau \\
&= \frac{\lambda^2}{2} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} 2\sqrt{\frac{z}{\lambda\sigma_\varepsilon}} K_{\frac{1}{2}}\left(\sqrt{\frac{\lambda^2 z^2}{\sigma_\varepsilon^2}}\right) \\
&= \frac{\lambda^2}{2} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} 2\sqrt{\frac{z}{\lambda\sigma_\varepsilon}} \sqrt{\frac{\pi}{2}} \frac{\sigma_\varepsilon}{\lambda z} e^{-\frac{a|z|}{\sigma_\varepsilon}} \\
&= \frac{\lambda}{2\sigma_\varepsilon} e^{-\frac{\lambda|z|}{\sigma_\varepsilon}}.
\end{aligned}$$

□

From the hierarchical representation, we can obtain the full conditional distributions

$$\begin{aligned}
\boldsymbol{\beta}|- &\sim N(\widehat{\boldsymbol{\beta}}_n, \sigma_\varepsilon^2 \widehat{\Sigma}_n) \\
(\tau_j^2)^{-1}|- &\sim \mathrm{IN}(\mu_{\tau_j}, \lambda^2),
\end{aligned}$$

where $\mu_{\tau_j} = \sqrt{\lambda^2\sigma_\varepsilon^2/\boldsymbol{\beta}_j^2}$ and

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_n &= \widehat{\Sigma}_n X^\top y \\
\widehat{\Sigma}_n &= (X^\top X + D_\tau^{-1})^{-1}.
\end{aligned}$$

Remark. Remember that the lasso prior with a single λ only works for standardized variables

Alternatively, we can write the posterior distribution of the Bayesian lasso regression model as the following orthant-wise Normal distribution.

Prop. 21 (Joint posterior for β under BLASSO)
Applying Bayes' theorem to the lasso regression model, we can prove that the posterior distribution of $\boldsymbol{\beta}$ is a mixture of orthant-wise truncated Normal distributions,

$$\pi(\boldsymbol{\beta}|y, X, \sigma_\varepsilon, \lambda) = \sum_{z \in \mathcal{Z}} \omega_z^*(y, X, \lambda, \sigma_\varepsilon) \frac{\phi_p(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}^z, \Sigma)}{\Phi_p^z(\widehat{\boldsymbol{\beta}}^z, \Sigma)} \mathbb{1}_{O_z}(\boldsymbol{\beta}),$$

where

$$\Phi_p^z(\widehat{\boldsymbol{\beta}}^z, \Sigma) = \int_{O_p^z} \mathcal{N}(t|\widehat{\boldsymbol{\beta}}^z, \Sigma) dt$$

$$\widehat{\boldsymbol{\beta}}^z = \widehat{\boldsymbol{\beta}}_{ols} - \lambda \sigma_\varepsilon^{-1} \cdot \Sigma z$$

$$\Sigma = \sigma_\varepsilon^2 (X^\top X)^{-1}$$

$$\omega_z(y, X, \lambda, \sigma_\varepsilon) = \int \pi(y|X, \boldsymbol{\beta}, \sigma_\varepsilon^2) \pi(\boldsymbol{\beta}|\lambda, \sigma_\varepsilon) \mathbb{1}_{O_p^z}(\boldsymbol{\beta}) d\boldsymbol{\beta} = \frac{\Phi_p^z(\widehat{\boldsymbol{\beta}}^z, \Sigma)}{\phi_p(0|\widehat{\boldsymbol{\beta}}^z, \Sigma)}$$

$$\omega^* = \frac{\omega_z}{\sum_{z \in \mathcal{Z}} \omega_z}.$$

Proof.
Consider a univariate Gaussian linear regression model of the form

$$y = x\boldsymbol{\beta} + \varepsilon,$$

with prior distribution

$$\pi(\boldsymbol{\beta}) \propto e^{\frac{\lambda}{\sigma_\varepsilon}|\boldsymbol{\beta}|}.$$

Then, the posterior is proportional to

$$\pi(\boldsymbol{\beta}|y, x) \propto \prod_{i=1}^n \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(y_i - x_i\boldsymbol{\beta})^2\right\} \exp\left\{-\frac{\lambda}{\sigma_\varepsilon}|\boldsymbol{\beta}|\right\}.$$

Assume that we know that the posterior is a normal, then the variance is unaffected by the prior and is equal to $\sigma_\varepsilon^2(x^\top x)^{-1}$. Moreover, the absolute value can be written as

$$|\boldsymbol{\beta}| = \begin{cases} \boldsymbol{\beta} & \text{if } \boldsymbol{\beta} \geq 0 \\ -\boldsymbol{\beta} & \text{if } \boldsymbol{\beta} < 0, \end{cases}$$

and therefore if $\boldsymbol{\beta}$ is positive we have a penalization $\exp\{-\lambda/\sigma_\varepsilon\boldsymbol{\beta}\}$ on the likelihood whereas if $\boldsymbol{\beta}$ is negative we apply a sign. Combining the two and completing the square gives the formula for the posterior mean.

□

Remark. The variance—or curvature—of the resulting posterior distribution for $\boldsymbol{\beta}$ is unaffected by the prior, since the kernel is proportional to

$$\pi(\boldsymbol{\beta}) \propto \prod_{j=1}^{p} e^{\frac{\lambda}{\sigma_\varepsilon}|\boldsymbol{\beta}_j|},$$

and thus it does not contain any square terms that can influence Σ.

Gibbs sampler We can setup a Gibbs sampler for the model parameters by writing the full conditional distribution of $\boldsymbol{\beta}_j|\boldsymbol{\beta}_{-j}, -$ as the mixture distribution

$$\pi(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{-j}, -) = \omega_j \cdot \phi(\boldsymbol{\beta}_j|\widehat{\boldsymbol{\beta}}_j^-, \sigma_j^2) \cdot \mathbb{1}_{(-\infty,0)} + (1-\omega_j) \cdot \phi(\boldsymbol{\beta}_j|\widehat{\boldsymbol{\beta}}^+, \sigma_j^2) \cdot \mathbb{1}_{(0,+\infty)},$$

where

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_j^+ &= (x_j^\top x_j)^{-1}\big(x_j^\top(y - X_{-j}\boldsymbol{\beta}_{-j}) - \sigma_\varepsilon\lambda\big) \\
\widehat{\boldsymbol{\beta}}_j^- &= (x_j^\top x_j)^{-1}\big(x_j^\top(y - X_{-j}\boldsymbol{\beta}_{-j}) + \sigma_\varepsilon\lambda\big) \\
\sigma_j^2 &= \sigma_\varepsilon^2(x_j^\top x_j)^{-1},
\end{aligned}$$

and ω_j is a mixing weight. Proof of this result is given in the appendix of the slides.

23.2.2 Spike and slab prior

In order to actually enforce sparsity on the model coefficients, Mitchell and Beauchamp (1988) and George and McCulloch (1997) postulated the **continuous spike-and-slab** prior distribution,

$$\pi(\boldsymbol{\beta}|\sigma_\varepsilon^2, v_0, v_1, \theta) = \prod_{j=1}^{p} \big\{(1-\theta)\phi(\boldsymbol{\beta}_j|0, \sigma_\varepsilon^2 v_{0,j}) + \theta\phi(\boldsymbol{\beta}_j|0, \sigma_\varepsilon^2 v_{1,j})\big\}, \tag{88}$$

where $v_{0,j} \ll v_{1,j}$ in order to force a solution which is very close to zero. Usually, $v_0/v_1 \approx 10^5$ in order to avoid label switching in the posterior sampling algorithm.

If $v_{0,j} \to 0$ and $v_{1,j} = v_1$ then the **spike and slab** prior becomes

$$\pi(\boldsymbol{\beta}|\sigma_\varepsilon^2, v_1, \theta) = \prod_{j=1}^{p} \big\{(1-\theta)\delta_0(\boldsymbol{\beta}_j) + \theta\phi(\boldsymbol{\beta}_j|0, \sigma_\varepsilon^2 v_1)\big\}. \tag{89}$$

Remark. Using the spike and slab solution, we have a probabilistic evaluation of whether the model coefficient is zero or not. This model is the "gold standard" for Bayesian variable selection in linear regression.

Remark. Moreover, if v_1 is large then the width of the slab does not excessively add bias to the value of the nonzero coefficients.

For the spike and slab prior in (89), we can write the following hierarchical representation by introducing a vector of indicators $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$,

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_{|\boldsymbol{\gamma}|}(0, \sigma_\varepsilon^2 v_1 I_{|\boldsymbol{\gamma}|})$$

$$\gamma_i \overset{\text{iid}}{\sim} \text{Ber}(\theta), \quad i = 1, \ldots, p.$$

Prop. 22 (Full spike-and-slab conditionals)
For the linear regression model under the spike and slab prior distributions, with $\sigma_\varepsilon^2 \sim IGa(\nu, \lambda)$, $\boldsymbol{\beta}|\sigma_\varepsilon^2 \sim N(0, \sigma_\varepsilon^2 D_{0,\gamma})$ *and* $\theta \sim Beta(\xi, \varphi)$, *for a* ***fixed*** *value of* γ *the full conditionals of the model parameters are*

$$\boldsymbol{\beta}_\gamma|- \sim N_{p_\gamma}\left(\widehat{\Sigma}_\gamma X_\gamma^\top y, \sigma_\varepsilon^2 \widehat{\Sigma}_\gamma\right)$$

$$\sigma_\varepsilon^2|- \sim IGa\left(\nu + \frac{n + p_\gamma}{2}, \lambda + \frac{\varepsilon_\gamma^\top \varepsilon_\gamma}{2}\right)$$

$$\theta|- \sim Beta(\xi + p_\gamma, \varphi + p - p_\gamma),$$

where $\widehat{\Sigma}_\gamma = (X_\gamma^\top X_\gamma + D_{0,\gamma}^{-1})^{-1}$.

Remark. From these solutions we can sample from the posterior distribution of the parameters only if we know the true value of γ. In order to simulate from $\gamma|-$ we can leverage a marginal likelihood approach, since in the conditionally conjugate case we know this quantity.

Prop. 23 (Marginal likelihood under spike and slab)
Integrating out $(\boldsymbol{\beta}_\gamma, \sigma_\varepsilon^2)$ *from the unnormalized joint posterior yields the posterior distribution of the model indicator* γ, *which is given by*

$$m(\gamma|y, X_\gamma) \propto \ell(\gamma|y, X_\gamma)\pi(\gamma), \tag{90}$$

where if $S_\gamma^2 = y^\top y - y^\top X_\gamma \widehat{\Sigma}_\gamma y$, *then*

$$\ell(\gamma|y, X_\gamma) \propto |\widehat{\Sigma}_\gamma|^{-1/2} |D_{0,\gamma}|^{-1/2} \left(\lambda + \frac{S_\gamma^2}{2}\right)^{-(\nu + n/2)}$$

$$\pi(\gamma) \propto \binom{p}{p_\gamma} \theta^{p_\gamma} (1 - \theta)^{p - p_\gamma}.$$

Remark. Exact sampling from (90) is feasible only under small values of p. For larger values, we sample according to a reversible-jump MCMC algorithm (Green, 1995):

1. Sample an indicator $\gamma^*|\gamma \sim q(\gamma^*|\gamma)$ from

$$q(\gamma^*|\gamma) = \frac{1}{\binom{p}{d}}, \quad \text{where } \sum_{j=1}^{p} |\gamma_j^* - \gamma_j| = d.$$

2. The sampled indicator is accepted with the usual Metropolis-Hastings acceptance probability,

$$\alpha = \min\left\{1, \frac{m(\gamma|y, X_{\gamma^*})}{m(\gamma|y, X_\gamma)}\right\}.$$

Problems. Two possible problems are a) the dimension of the model space, which requires a large number of iterations to fully explore and b) computing the variance update $\widehat{\Sigma}_\gamma = \left(X_\gamma^\top X_\gamma + D_{0,\gamma}^{-1}\right)^{-1}$. This problem can be solved by computing the spectral decomposition of $X_\gamma^\top X_\gamma$, since

$$\begin{aligned}
\widehat{\Sigma}_\gamma &= \left(X_\gamma^\top X_\gamma + \frac{1}{v_1} I_{p_\gamma}\right)^{-1} \\
&= \left(U_\gamma V_\gamma U_\gamma^\top + \frac{1}{v_1} I_{p_\gamma}\right)^{-1} \\
&= U_\gamma^\top \Big(\underbrace{V_\gamma + \frac{1}{v_1} I_{p_\gamma}}_{\text{diagonal}}\Big)^{-1} U_\gamma \qquad\qquad (\text{since } U_\gamma^\top U_\gamma = I)
\end{aligned}$$

and this is easy to compute because V is assumed to be a diagonal matrix. Our problem is that the design matrix X_γ always changes whenever γ changes, and therefore we cannot compute this inverse once and for all.

Instead of comparing all the $2^{|\gamma|} - 1$ possible models, we might only be interested in finding the regression coefficients which are nonzero. In order to do so, we can leverage a Gibbs sampling approach on the weights θ in Equation (89) rather than sampling from the model indicator γ.

Prop. 24 (Stochastic Search Variable Selection)
Under the spike and slab, the full conditional distribution of the regression parameters $\boldsymbol{\beta}_j$ is

$$\boldsymbol{\beta}_j | \boldsymbol{\beta}_{-j}, - \sim \tilde{\omega}_j \delta_0 + (1 - \tilde{\omega}_j) N(\widehat{\boldsymbol{\beta}}_j, \widehat{\sigma}_j^2),$$

where

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_j &= (x_j^\top x_j + v_0^{-1})^{-1} x_j^\top (y - X_{-j} \boldsymbol{\beta}_{-j}) \\
\widehat{\sigma}_j^2 &= \sigma_\varepsilon^2 (x_j^\top x_j + v_0^{-1})^{-1} \\
\tilde{\omega}_j &= \left\{1 + \frac{\omega\tau}{(1-\omega)\sigma_\varepsilon} \cdot \frac{\Phi(\widehat{\boldsymbol{\beta}}_j / \sigma_j^2)}{\phi(0 | \widehat{\boldsymbol{\beta}}_j, \sigma_j^2)}\right\}^{-1}
\end{aligned}$$

Remark. This gives us the posterior distribution for $\boldsymbol{\beta}$ and is extremely computationally efficient, since we obtain marginal distributions for $\boldsymbol{\beta}_j$ and γ_j.

Covariate selection. Keep in mind that this approach selects the covariates but not the model, as we would obtain using the full spike-and-slab conditionals.

Complexity. Although computationally simple, we need $B \cdot p$ simulations in order to obtain a posterior distribution since we need to sweep the whole p-dimensional vector for each iteration of the Gibbs sampler. Hence, the rjMCMC using the full conditionals can be faster if p is very large and the set p^* of relevant regressors is small, e.g. $p \approx 10^4$ and $p^* \approx 100$.

Lecture 24: Bayesian linear regression (iii)

2022-06-06

24.1 Generalized linear models

In this section we consider several extension to non-Gaussian regression models: the probit, logit and the Poisson regression. For all cases we consider the solution based on data augmentation techniques which make things Gaussian. From this, since we can use all previously known results to perform statistical inference.

24.1.1 Probit regression

Def. (Probit regression model)
The **probit regression model** is defined as

$$y_i|\boldsymbol{\beta} \sim \text{Ber}(\psi_i)$$
$$\psi_i = \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})$$
$$\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta}),$$

where Φ is the cumulative distribution function of the standard Normal distribution.

Assuming for the vector of regression parameters the following prior distribution,

$$\boldsymbol{\beta}|v_1 \sim N(\beta|0, v_1 I_p),$$

we follow the approach of Albert and Chib (1993) in order to obtain a straightforward Gibbs sampler algorithm. Since the posterior distribution of the GLM,

$$L(y, X, \boldsymbol{\beta}) \times \pi(\boldsymbol{\beta}), \tag{91}$$

is not available we augment the data by introducing latent variables y_i^* so that

$$L(y, y^*, X, \boldsymbol{\beta}) \times \pi(\boldsymbol{\beta}) \tag{92}$$

has known form. If we can then show that by averaging over the latent variables we obtain the same likelihood from which we started,

$$\int L(y, y^*, X, \boldsymbol{\beta}) dy^* = L(y, X, \boldsymbol{\beta}), \tag{93}$$

we can use (92) in place of (91) in order to perform inference on the posterior parameters. In practice we numerically approximate the integral (93) by using a Gibbs sampler algorithm that samples from the full conditional distributions, and then by averaging over y^*.

Remark. The price that we pay using data augmentation is the fact that the augmented space is of **larger dimension** due to y^*. Therefore, we introduce further uncertainty in the estimation procedure.

Probit regression. For the probit regression model, we can write the following data-augmentation scheme by introducing a Gaussian latent variable,

$$y_i^* = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0,1)$$

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* < 0, \end{cases}$$

and our goal is now to obtain the full conditional distribution of the latent variables y_i^*. The full conditional can be written as

$$y_i^* | y_i, \boldsymbol{x}_i, \beta \sim N(\boldsymbol{x}_i^\top \boldsymbol{\beta}, 1) \cdot \mathbb{1}_{(0,+\infty)} \cdot \mathbb{1}_{\{1\}}(y_i) + N(\boldsymbol{x}_i^\top \beta, 1) \cdot \mathbb{1}_{(-\infty,0)} \cdot \mathbb{1}_{\{0\}} y_i.$$

On the other hand, the joint full conditional distribution of the regression parameters $\boldsymbol{\beta}$ becomes

$$\boldsymbol{\beta} | y^*, y, X \overset{\text{d}}{=} \boldsymbol{\beta} | y^*, X \sim N\big(\widehat{\beta}_n, \widehat{\Sigma}_n\big),$$

where

$$\widehat{\beta}_n = \widehat{\Sigma}_n X^\top y^*$$

$$\widehat{\Sigma}_n = (X^\top X + D_0^{-1})^{-1}.$$

The following Proposition provides the marginal likelihood for the binary probit regression model.

Prop. 25 (Marginal likelihood probit model)
For the probit regression model and prior distribution $\boldsymbol{\beta} | v_1 \sim N(\beta | 0, D_0)$ with $D_0 = v_1 I_p$, the marginal likelihood is proportional to

$$L(X, y | y^*) \propto \exp\left\{-\frac{S^2}{2}\right\} |D_0|^{-1/2} |\widehat{\Sigma}_n|^{1/2},$$

where $S^2 = {y^*}^\top H y^*$, $H = I_n - X \widehat{\Sigma}_n X^\top$, and $\widehat{\Sigma}_n = (X^\top X + D_0^{-1})^{-1}$.

Remark. It is relevant that S^2 contains y^*, since it tells us the conditional model fit by conditioning on y^*. After convergence of the Gibbs sampler, we can average the sampled values over y^* in order to obtain the unconditional marginal likelihood $L(X, y)$.

24.1.2 Logistic regression

We now consider the classical logistic regression and provide a Gibbs sampler for its Bayesian version.

Def. (Logistic regression model)
The **logistic regression model** is defined as

$$y_i|\beta \sim \text{Ber}(\psi_i)$$

$$\psi_i = \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\top \beta}}$$

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}_0, D_0).$$

The problem of finding a data-augmentation scheme for the logit regression model is very recent and has been solved by Polson et al. (2013).

Prop. 26 (Pólya-Gamma representation)
Let $p(\omega)$ denote the density of the random variable $\omega_i \sim PG(1,0)$. Then, the following integral identity holds for all $a \in \mathbb{R}$ and $b > 0$,

$$\frac{(e^\psi)^a}{(1+e^\psi)^b} = \frac{1}{2^b} e^{\kappa\psi} \int_0^\infty \underbrace{e^{-\omega\psi^2/2}}_{\propto \text{Gaussian}} p(\omega) d\omega, \tag{94}$$

where $\kappa = a - b/2$

Remark. This tells us that the logistic function can be written as a mixture between a Gaussian distribution over a latent factor which is distributed as a Pólya-Gamma random variable.

Application. Using the latent factor, if $\tilde{y}_i = y_i - \frac{1}{2}$ the complete-data likelihood can be written as

$$\begin{aligned}
\pi(y, \omega | X, \boldsymbol{\beta}) &= \prod_{i=1}^n \mathbb{P}(Y_i = y_i | \boldsymbol{x}_i, \omega_i, \alpha, \boldsymbol{\beta}) \pi(\omega_i) \\
&= \prod_{i=1}^n \frac{\left(e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\right)^{y_i}}{1 + e^{\boldsymbol{x}_i^\top \beta}} \pi(\omega_i) \\
&= \prod_{i=1}^n \frac{1}{2} \exp\left\{ y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \frac{\boldsymbol{x}_i^\top \beta}{2} \right\} \exp\left\{ -\frac{\omega_i (\boldsymbol{x}_i^\top \boldsymbol{\beta})^2}{2} \right\} \\
&= \frac{1}{2^n} \exp\left\{ \sum_{i=1}^n \tilde{y}_i (\boldsymbol{x}_i^\top \boldsymbol{\beta}) \right\} \exp\left\{ -\sum_{i=1}^n \frac{\omega_i (\boldsymbol{x}_i^\top \boldsymbol{\beta})^2}{2} \right\}.
\end{aligned}$$

Under the Gaussian prior, the posterior distribution of $\boldsymbol{\beta}$ becomes

$$\begin{aligned}\pi(\beta|y,X,\omega) &\propto \exp\left\{\sum_{i=1}^n \tilde{y}_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right\} \exp\left\{-\sum_{i=1}^n \frac{\omega_i(\boldsymbol{x}_i^\top\boldsymbol{\beta})^2}{2}\right\}\pi(\boldsymbol{\beta}) \\ &\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^n \omega_i\left(\boldsymbol{x}_i^\top\boldsymbol{\beta} - \frac{\tilde{y}_i}{\omega_i}\right)^2\right\}\pi(\boldsymbol{\beta}) \\ &\propto \exp\left\{-\frac{1}{2}(\bar{y}-X\beta)^\top W(\bar{y}-X\boldsymbol{\beta})\right\}\pi(\boldsymbol{\beta}),\end{aligned}$$

where $\bar{y} = (\frac{\tilde{y}_1}{\omega_1},\dots,\frac{\tilde{y}_n}{\omega_n})$, $W = \mathrm{diag}(\omega_1,\dots,\omega_n)$. In the above equation, we have the posterior distribution that is normally-distributed with mean and variance

$$\begin{aligned}\widehat{\mu}_n &= \widehat{\Sigma}_n(X^\top\tilde{y} + D_0^{-1}\boldsymbol{\beta}_0) \\ \widehat{\Sigma}_n &= (X^\top W X + D_0^{-1})^{-1}.\end{aligned}$$

Therefore, we only need to sample from a $\mathrm{PG}(1,0)$ in order to obtain a sample from the posterior distribution of $\boldsymbol{\beta}$. Finally, the conditional distribution of $\omega_i|y,\beta$ is $\mathrm{PG}(1,\boldsymbol{x}_i^\top\boldsymbol{\beta})$.

Remark. In the logistic regression, the posterior distribution of $\boldsymbol{\beta}$ has the matrix W in the covariance matrix, whereas we do not have it when we use the probit regression. Simulating from the probit regression posterior is much simpler from a computational point of view.

24.1.3 Poisson model

Def. (Log-Gamma distribution)
Let $X \sim \mathrm{Gamma}(\alpha,\kappa)$, then the **Log-Gamma distribution** is defined as the distribution of $Y = \log X \sim \mathrm{LGamma}(\alpha,\kappa)$ with probability density function

$$f_Y(y) = \frac{\kappa^\alpha}{\Gamma(\alpha)} e^{\alpha y - \kappa e^y}\mathbb{1}_{(-\infty,\infty)}(y). \tag{95}$$

Def. (Multivariate Log-Gamma distribution)
Let $\boldsymbol{y} = \boldsymbol{c} + V\boldsymbol{w}$ with $\boldsymbol{w} = (w_1,\dots,w_p)$, $w_i \sim \mathrm{LGamma}(\alpha_i,\kappa_i)$ $\boldsymbol{c} \in \mathbb{R}^p$ be the location parameter and $V \in \mathbb{R}^{p\times p}$ the scale matrix. Then, the **multivariate Log-Gamma distribution** is the density function of $\boldsymbol{Y} \sim \mathrm{MLGamma}(\boldsymbol{\alpha},\boldsymbol{\kappa},\boldsymbol{c},\boldsymbol{V})$.

Prop. 27 (Conjugacy of Log-Gamma distribution)
Let $Z|Q \sim Pois(e^Q)$ and assume that $Q \sim LGamma(\alpha,\kappa)$, then

$$Q|Z \sim LGamma(Z+\alpha,\kappa+1).$$

Proof.
The density of $Z|Q$ is given by

$$f_{Z|Q}(z) = e^{-e^q} \frac{(e^q)^z}{z!} \propto e^{zq-e^q},$$

which has the same density as (95) with hyperparameters $\alpha = z$ and $\kappa = 1$.

□

Def. (Poisson Log-Gamma regression model)
The **Poisson Log-Gamma regression model** is defined as

$$\begin{aligned} y_i|\lambda_i &\sim \text{Pois}(e^{\lambda_i}) \\ \lambda_i &= \boldsymbol{x}_i^\top \boldsymbol{\beta} \\ \boldsymbol{\beta} &\sim \text{MLG}(0, v_1 \cdot I_p, \alpha_\beta \cdot \mathbf{1}_p, \kappa_\beta \cdot \mathbf{1}_p). \end{aligned}$$

Posterior distribution. Under the previous model, the joint distribution of $(y, \boldsymbol{\beta})$ can be written as

$$\dots,$$

and the posterior distribution of the regression parameters is

$$\dots$$

See slides 51–53 for calculations.

24.2 Gaussian process

Here we derive a generalization of Bayesian linear regression, with possibly infinitely many basis functions, using Gaussian process regression. In general, they define a distribution directly over functions by using the kernel trick, which is one of the most important ideas in machine learning. This lets us easily incorporate assumptions like smoothness, periodicity time-space dependence, etc., which are usually hard to encode as priors over regression weights.

24.2.1 Gaussian process regression

Gaussian processes are so that two fields of research stem, where in both a multivariate Gaussian is introduced as a prior distribution.

1. **Gaussian process prior**, where the generic multivariate Gaussian is the prior distribution.
2. **Gauss-Markov random field**, where we also have the Markov property on the precision matrix of the multivariate Gaussian, similarly to a Jordan canonical form.

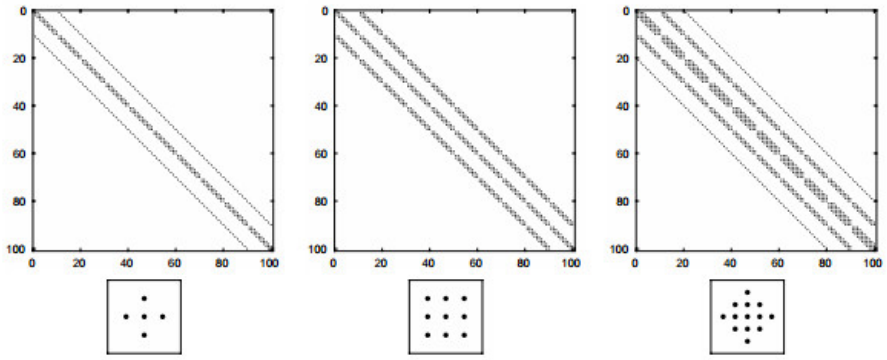


Figure 56: Nonzero values of the precision matrix $\Omega = \Sigma^{-1}$ in a Gauss-Markov random field.

Under this framework, the matrix inversion has order $O(k^2)$ instead of $O(p^3)$, where k is the number of bands that are different from zero. From the conditional independence, after lag k we do not have dependence between observations. This is particularly useful for autoregressive models, for instance the AR(1),

$$y_t = \phi y_{t-1} + \varepsilon_t.$$

has diagonals with nonzero values at $k = \pm 1$.

Def. (Gaussian process regression)

A **Gaussian process regression** (GPR) assumes the following representation,

$$y_i | f, \boldsymbol{x} \sim \mathcal{N}\big(f(\boldsymbol{x}), \sigma^2\big) \tag{96}$$

$$f | \boldsymbol{x} \sim \mathcal{GP}\big(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\big) \tag{97}$$

Sampling. The $\mathcal{GP}$ prior is defined so that for any finite collection of points $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell)$, $\boldsymbol{x}_j \in \mathbb{R}^p$, the joint density of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell) \sim N\big((m(\boldsymbol{x}_1), \ldots, m(\boldsymbol{x}_\ell))^\top, [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j}\big)$ Sampling from the GP on a grid of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ simply amounts to sampling from the following multivariate normal distribution,

$$\begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} m(\boldsymbol{x}_1) \\ \vdots \\ m(\boldsymbol{x}_n) \end{pmatrix}, \Big(k(\boldsymbol{x}_i, \boldsymbol{x}_j)\Big)_{i,j=1,\ldots,n} \right)$$

Parameters. The Gaussian process can be defined by choosing a mean function $m : \mathcal{X} \longrightarrow \mathbb{R}$ and a covariance function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ so that

$$\mathbb{E}[f(\boldsymbol{x})] = m(\boldsymbol{x}),$$

$$\text{Cov}[f(\boldsymbol{x}), f(\boldsymbol{x}')] = k(\boldsymbol{x}, \boldsymbol{x}').$$

Although $m(\boldsymbol{x})$ is usually chosen to be zero, the choice of $k(\boldsymbol{x}, \boldsymbol{x}')$ is crucial as it essentially determines the behavior of the GP prior. When k is a function that depends only on the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$,

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = k(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|),$$

the GP prior is **stationary** (or **isotropic**).

Remark. The GP prior is useful to approximate other prior distributions by leveraging the zeros in the precision matrix to write it as a Gauss-Markov random field. Then sampling and updating from the posterior is faster, and this is especially convenient for high-dimensional models.

Example (Bayesian optimization)
Bayesian optimization is a process which tries to minimize a function with measurement errors

$$\min f(x) + \varepsilon$$

where we know that there is an actual underlying function f but cannot observe it directly unless with measurement errors. A common solution is to employ a Gaussian process to obtain the posterior for f and optimize it numerically.

24.2.2 Posterior computation

As for any regression problem, the main use of GPR models is for predicting $f(\boldsymbol{x}_1^*), \dots, f(\boldsymbol{x}_{n^*}^*)$ over a set of new predictors by using the posterior distribution conditionally on the observed data. The joint density function for $\boldsymbol{y}$ and $\boldsymbol{f}_{\boldsymbol{x}^*}$ is given by

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{f}_{\boldsymbol{x}^*} \end{pmatrix} \sim MVN\left(\begin{pmatrix} \boldsymbol{m}_{\boldsymbol{x}} \\ \boldsymbol{m}_{\boldsymbol{x}^*} \end{pmatrix}, \begin{pmatrix} K_{\boldsymbol{x},\boldsymbol{x}} + \sigma^2 I_n & K_{\boldsymbol{x},\boldsymbol{x}^*} \\ K_{\boldsymbol{x},\boldsymbol{x}^*}^\top & K_{\boldsymbol{x}^*,\boldsymbol{x}^*} \end{pmatrix} \right),$$

which is a consequence of the fact that $\boldsymbol{\varepsilon}^* \perp\!\!\!\perp f(\boldsymbol{x})$. Indeed,

$$\mathrm{Cov}(\boldsymbol{y}, \boldsymbol{f}_{\boldsymbol{x}^*}) = \mathrm{Cov}(\boldsymbol{f}_{\boldsymbol{x}} + \boldsymbol{\varepsilon}, \boldsymbol{f}_{\boldsymbol{x}^*}) = \mathrm{Cov}(\boldsymbol{f}_{\boldsymbol{x}}, \boldsymbol{f}_{\boldsymbol{x}^*}) = K_{\boldsymbol{x},\boldsymbol{x}^*}.$$

Corollary 2 (Posterior distribution of the Gaussian process)
Using the usual conditioning formulas for the multivariate Gaussian distributions, we have that

$$\boldsymbol{f}_{\boldsymbol{x}^*} | \boldsymbol{y} \sim MVN(\mathbb{E}[\boldsymbol{f}_{\boldsymbol{x}^*} | \boldsymbol{y}], \mathbb{V}[\boldsymbol{f}_{\boldsymbol{x}^*} | \boldsymbol{y}]), \tag{98}$$

$$\boldsymbol{y}_{\boldsymbol{x}^*} | \boldsymbol{y} \sim MVN\big(\mathbb{E}[\boldsymbol{f}_{\boldsymbol{x}^*} | \boldsymbol{y}], \mathbb{V}[\boldsymbol{f}_{\boldsymbol{x}^*} | \boldsymbol{y}] + \sigma^2 I_{n^*}\big), \tag{99}$$

where

$$\mathbb{E}[\boldsymbol{f}_{\boldsymbol{x}^*} | \boldsymbol{y}] = \boldsymbol{m}_{\boldsymbol{x}^*} + K_{\boldsymbol{x},\boldsymbol{x}^*}^\top (K_{\boldsymbol{x},\boldsymbol{x}} + \sigma^2 I_{n^*})^{-1} (\boldsymbol{y} - \boldsymbol{m}_{\boldsymbol{x}}),$$

$$\mathbb{V}[\boldsymbol{f}_{\boldsymbol{x}^*} | \boldsymbol{y}] = K_{\boldsymbol{x}^*,\boldsymbol{x}^*} - K_{\boldsymbol{x},\boldsymbol{x}^*}^\top (K_{\boldsymbol{x},\boldsymbol{x}} + \sigma^2 I_{n^*})^{-1} K_{\boldsymbol{x},\boldsymbol{x}^*}.$$

Remark. The only difference between this process and ordinary Bayesian linear regression is that $X^\top X$ is substituted by $K_{\boldsymbol{x},\boldsymbol{x}}$ in the calculation of the mean and variance. This is another instance of the **kernel trick** that is commonly used in statistical learning theory.

Computational Issues. Without using a Gauss-Markov random field structure, both expressions require the inversion of a matrix of size $n \times n$, which is an $O(n^3)$ Cholesky operation unless $K_{\boldsymbol{x},\boldsymbol{x}}$ has a special structure that can be exploited. Once the inverse matrix has been computed, evaluation of $\mathbb{E}[\boldsymbol{f}_{\boldsymbol{x}^*}|\boldsymbol{y}]$ is an $O(n)$ operation and evaluation of $\mathbb{V}[\boldsymbol{f}_{\boldsymbol{x}^*}|\boldsymbol{y}]$ is $O(n^2)$. The same holds for $\mathbb{E}[\boldsymbol{y}_{\boldsymbol{x}^*}|\boldsymbol{y}]$ and $\mathbb{V}[\boldsymbol{y}_{\boldsymbol{x}^*}|\boldsymbol{y}]$.

Type-II MLE. Regardless of the mean and covariance functions adopted for the GPR, they are both usually defined in terms of low-dimensional parameters, which need to be estimated appropriately for making predictions with either (98) or (99). The usual approach for jointly estimating the generic parameter vector θ and σ^2 is to perform type-II maximum likelihood (to highlight that we are doing maximum likelihood within a nonparametric Bayesian model) over the model marginal log-likelihood

$$\log p(\boldsymbol{y}; \sigma^2, \boldsymbol{\theta}) = -\frac{n}{2} \log |K_{\boldsymbol{x},\boldsymbol{x}} + \sigma^2 I| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{m}_{\boldsymbol{x}})^\top (K_{\boldsymbol{x},\boldsymbol{x}} + \sigma^2)^{-1} (\boldsymbol{y} - \boldsymbol{m}_{\boldsymbol{x}})$$

Full Bayesian approach. Taking a fully-Bayesian inferential approach to GPR requires specifying a prior distribution for $\boldsymbol{\theta}$ and σ^2 instead of treating them as unknown quantities to be estimated via maximum likelihood. Given $p(\sigma^2, \boldsymbol{\theta})$, the generic joint prior density function for $\boldsymbol{\theta}$ and σ^2, the posterior density function for the GPR model parameter is

$$p(\sigma^2, \boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\sigma^2, \boldsymbol{\theta}) p(\sigma^2, \boldsymbol{\theta}).$$

In general this does not admit an explicit expression, and instead requires efficient MCMC sampling schemes or variational approximation procedures in order to be implemented. Also, the predictive distributions for $\boldsymbol{f}_{\boldsymbol{x}^*}$ now becomes

$$p(\boldsymbol{f}_{\boldsymbol{x}^*}|\boldsymbol{y}) = \int p(\boldsymbol{f}_{\boldsymbol{x}^*}|\boldsymbol{y}, \sigma^2, \boldsymbol{\theta}) p(\sigma^2, \boldsymbol{\theta}|\boldsymbol{y}) d\sigma^2 d\boldsymbol{\theta},$$

and equivalently for the posterior predictive for $\boldsymbol{y}_{\boldsymbol{x}^*}$. In both cases the GPR posterior predictive distribution can be interpreted as a mixture of multivariate Gaussian distributions, with $p(\sigma^2, \boldsymbol{\theta}|\boldsymbol{y})$ acting as mixing density.

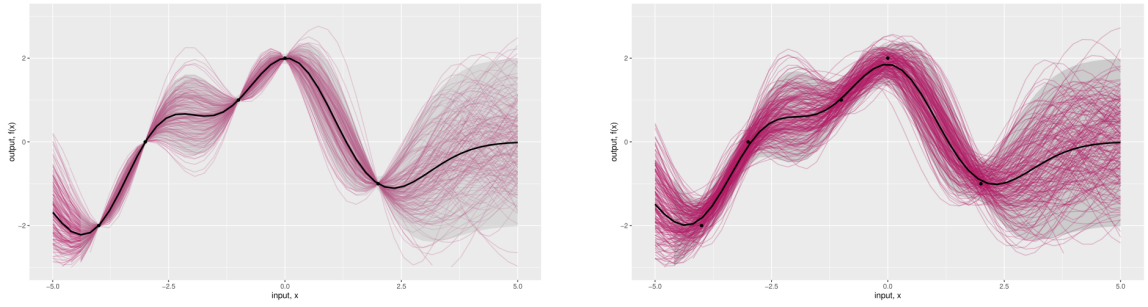


Figure 57: Posterior predictive distribution for $\boldsymbol{f}_{\boldsymbol{x}^*}$ (*left*) and for $\boldsymbol{y}_{\boldsymbol{x}^*}$ (*right*) using Gaussian process regression. Note that the posterior variance of $\boldsymbol{f}_{\boldsymbol{x}^*}$ is zero when $\boldsymbol{x}^* = \boldsymbol{x}$.

Lecture 25: Multivariate regression models

2022-06-08

We now consider the class of models where the response variable $\boldsymbol{Y}$ is itself a multivariate random vector. This may be useful to jointly predict the values of the random vector by including information on the covariance structure of the distribution.

25.1 Gaussian multivariate regression

To define the basic Gaussian multivariate regression (GMR) model, consider the q separate regression models for the j^{th} response variable $\boldsymbol{y}_j \in \mathbb{R}^N$ using the same design matrix $X \in \mathbb{R}^{N \times p}$,

$$\begin{aligned}
\boldsymbol{y}_1 &= X\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1 \\
\boldsymbol{y}_2 &= X\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2 \\
&\vdots \\
\boldsymbol{y}_q &= X\boldsymbol{\beta}_q + \boldsymbol{\varepsilon}_q
\end{aligned}$$

where the $\boldsymbol{\beta}_j$'s are regression parameters and $\boldsymbol{\varepsilon}_j$ are Gaussian error terms such that $\mathbb{E}[\boldsymbol{\varepsilon}_j] = \mathbf{0}_N$ and

$$\begin{aligned}
\text{Cov}(\boldsymbol{\varepsilon}_j, \boldsymbol{\varepsilon}_j) &= \Sigma, \\
\text{Cov}(\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_s) &= \mathbf{0}_{N \times N}, \quad t \neq s.
\end{aligned}$$

Then, we can stack the observations on the response variables, the vector of error terms and of regression parameters as

$$\begin{aligned}
Y &= (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q) \\
E &= (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_q) \\
B &= (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q),
\end{aligned}$$

and the multivariate model can be rewritten as the matrix-variate model

$$Y = XB + E, \tag{100}$$

where $E \sim N(\mathbb{O}, \Sigma, I)$ is the matrix-variate Gaussian distribution with mean equal to the zero matrix and variance-covariance matrices

$$\begin{aligned}
\Sigma = \left(\sigma_{ij}\right)_{i,j} \quad & \text{for the columns of } Y \\
I_N \quad & \text{for the rows of } Y
\end{aligned}$$

Prop. 28 (Posterior distribution under non-informative prior)
For the matrix-variate Gaussian regression model defined in Equation (100) *under the non-informative (Jeffreys' prior) on* (B, Σ) *defined as*

$$\phi(B, \Sigma) = |\Sigma|^{-(k+1)/2}, \tag{101}$$

the joint posterior distribution distribution of $(B, \Sigma)|Y$ *is*

$$B|\Sigma, Y \sim MN(\widehat{B}, (X^\top X)^{-1}, \Sigma)$$

$$\Sigma|Y \sim IW(S, N),$$

where

$$\widehat{B} = (X^\top X)^{-1} X^\top Y, \quad S = Y^\top (I_N - P_X) Y,$$

and $P_X = X(X^\top X)^{-1} X^\top$. *Here,* $MN(\boldsymbol{a}, B, C)$ *indicates the matrix-variate normal distribution with mean* $\boldsymbol{a}$, *column-wise covariance* B *and row-wise covariance* C.

Prop. 29 (Marginal distributions under non-informative prior)
Furthermore, under the prior (101), *the marginal distributions of* (B, Σ) *are*

$$B|Y \sim MT_{p\times q}(\widehat{B}, X^\top X, S, N - p)$$

$$\Sigma|Y \sim IW_q(S, N),$$

where $MT_{p\times q}(\boldsymbol{a}, B, C, \nu)$ *is the matrix-variate* t *distribution with non centrality parameter* $\boldsymbol{a}$ *and* ν *degrees of freedom.*

Proof.

Proof 1 (Posterior distribution under non-informative prior)

To prove Propositions 1 and 2, note that the joint posterior distribution of $(\mathbf{B}, \mathbf{\Sigma})$ *is:*

$$\begin{aligned}
\varphi(\mathbf{B}, \mathbf{\Sigma}) &\propto |\mathbf{\Sigma}|^{-(N+k+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{XB})^\intercal(\mathbf{Y} - \mathbf{XB})\right]\right\} \\
&\propto |\mathbf{\Sigma}|^{-p/2} |\mathbf{\Sigma}|^{-(N-p+k+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{\Sigma}^{-1}\left(\mathbf{S} + (\mathbf{B} - \widehat{\mathbf{B}})^\intercal \mathbf{X}^\intercal \mathbf{X}(\mathbf{B} - \widehat{\mathbf{B}})\right)\right]\right\} \\
&\propto \varphi_{MN}(\mathbf{B}|\widehat{\mathbf{B}}, (\mathbf{X}^\intercal\mathbf{X})^{-1}, \mathbf{\Sigma}) \times \varphi_{IW}(\mathbf{\Sigma}|\mathbf{S}, N - p),
\end{aligned} \tag{10}$$

where $\mathbf{S} = (\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})^\intercal(\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})$. *Now, let us consider the quadratic form in equation* (10), *we have:*

$$(\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})^\intercal \mathbf{X}(\mathbf{B} - \widehat{\mathbf{B}}) = \left(\mathbf{Y}^\intercal\mathbf{X} - \widehat{\mathbf{B}}^\intercal\mathbf{X}^\intercal\mathbf{X}\right)(\mathbf{B} - \widehat{\mathbf{B}}) = 0. \tag{11}$$

The result follows immediately by applying Corollary 11 in Appendix.

□

Remark. Note that the posterior distribution of (B, Σ) is a proper distribution even under the improper Jeffreys' prior. Moreover, the posterior moments of the parameters (see expected value of multivariate student and inverse-Wishart) are

$$\mathbb{E}[B|Y] = \widehat{B}$$

$$\mathbb{E}[\Sigma|Y] = \frac{1}{N-p-1} \cdot S$$

$$\mathbb{V}[\text{vec}\, B|Y] = \frac{1}{N-p-1} S \otimes (X^\top X)^{-1}$$

Prop. 30 (Posterior distribution under the conjugate prior)
For the matrix-variate Gaussian regression model defined in equation (100) *under the conjugate prior distribution for* (B, Σ)*, defined as* $p(B, \Sigma) = p(B|\Sigma)p(\Sigma)$*, with:*

$$B|\Sigma \sim MN_{p\times q}(\overline{B}_0, \Sigma, P_0), \quad \Sigma \sim IW_q(C_0, v_0),$$

then the joint posterior distribution for (B, Σ) *is*

$$B|\Sigma, Y \sim MN_{p\times q}(\tilde{B}, \Sigma, \tilde{\Sigma})$$

$$\Sigma|Y \sim IW_q(C_0 + S + Q(\widehat{B}), N + v_0),$$

where $Q(\widehat{B}) = (\widehat{B} - \overline{B}_0)\tilde{\Sigma}P_0^{-1}(\widehat{B} - \overline{B}_0)$*, with* $\widehat{B} = (X^\top X)^{-1}X^\top Y$ *and*

$$\tilde{\Sigma} = (X^\top X + P_0^{-1})^{-1}, \quad \tilde{B} = \tilde{\Sigma}(P_0^{-1}\overline{B}_0 + X^\top Y). \tag{102}$$

Proof.
Slides 11–14.

□

Prop. 31 (Posterior variance Kalman form)
The posterior variance-covariance matrix (102) *can be rewritten as follows,*

$$\tilde{\Sigma} = (X^\top X + P_0^{-1})^{-1} = (I_p - KX)P_0, \tag{103}$$

where K *is the Kalman gain,*

$$F = I_N + XP_0X^\top$$

$$K = P_0X^\top F^{-1}$$

Gaussianity. In the non-Gaussian case, we apply formula (103) to a data-augmented quantity that replaces Y. The price that we pay for moving to the non-Gaussian case is an increase in variability in the posterior sampling algorithm.

Proof.
Using the Sherman-Morrison-Woodbury formula the posterior variance-covariance matrix becomes

$$\begin{aligned}\tilde{\Sigma} &= (X^\top X + P_0^{-1})^{-1} \\ &= P_0 - P_0 X^\top (I_N + X P_0 X^\top)^{-1} X P_0 \\ &= P_0 - P_0 X^\top F^{-1} X P_0 \\ &= (I_p - KX) P_0.\end{aligned}$$

□

Prop. 32 (Full conditionals under the conjugate prior)
Under the conjugate prior, the full conditional distributions are

$$B|\Sigma, Y \sim MN_{p\times q}(\tilde{B}, \Sigma, \tilde{\Sigma})$$

$$\Sigma|B, Y \sim IW_q(C_0 + Q^*(B), N + v_0 + p + q),$$

where

$$Q^*(B) = (Y - XB)^\top (Y - XB) + (B - \overline{B_0})^\top P_0^{-1} (B - \overline{B}_0).$$

Proof.
Slide 23.

□

Prop. 33 (Posterior distribution under the conditionally conjugate prior)
Under the conditionally conjugate prior for (B, Σ) defined as

$$B \sim MN_{p\times q}(\overline{B}_0, \Omega_0 \otimes P_0)$$

$$\Sigma \sim IW_n(C_0, v_0),$$

then the full conditional distributions in vectorized form are

$$B|\Sigma, Y \sim N_{pq}(\tilde{B}_{vec}, \tilde{\Sigma}_{vec})$$

$$\Sigma|B, Y \sim IW_q(Q(B), T + v_0),$$

where $Q(B) = C_0 + (Y - XB)^\top (Y - XB)$ and

$$\tilde{\Sigma}_{vec} = (\Sigma^{-1} \otimes X^\top X + \Omega_0^{-1} \otimes P_0^{-1})^{-1}$$

$$\tilde{B}_{vec} = \tilde{\Sigma}_{vec}\big((\Omega_0^{-1} \otimes P_0^{-1})\bar{\boldsymbol{b}}_0 + (\Sigma \otimes X^\top)\boldsymbol{y}\big),$$

where $\boldsymbol{y} = \mathrm{vec}(Y)$, $\boldsymbol{b} = \mathrm{vec}(B)$ and $\bar{\boldsymbol{b}}_0 = \mathrm{vec}(\overline{B}_0)$.

To induce sparse solutions we can assume a Dirac spike-and-slab prior that relies on an auxiliary latent p-dimensional selection vector $\Gamma = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_p)$, where $\boldsymbol{\gamma}_j \in \{\mathbf{1}, \mathbf{0}\}$. Therefore, the prior distribution conditional on the selection vector $\boldsymbol{\gamma}$ is specified as follows

$$B|\Sigma, \boldsymbol{\gamma} \sim \mathrm{MN}_{p\times q}(B_{\gamma,0})$$

25.2 Dynamic autoregressive models

Let $(\boldsymbol{y}_1, \dots, \boldsymbol{y}_T) \in \mathbb{R}^{k\times T}$ be a sequence of k-dimensional observations. We can define a dynamic model to describe the evolution over time of the observations, according to a specific evolution dynamic.

Def. (Dynamic autoregressive model)
We define a **dynamic autoregressive model** of order p, $\boldsymbol{Y}_t \sim \mathrm{VAR}(p)$, as

$$\boldsymbol{y}_t = \boldsymbol{\nu} + \Phi_1 \boldsymbol{y}_{t-1} + \dots + \Phi_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t, \quad t = 0, \pm 1, \pm 2, \dots, \tag{104}$$

with $\boldsymbol{\varepsilon}_t \sim \mathrm{WN}(\mathbf{0}, \Sigma)$.

Def. (Stability)
We say that the VAR model (104) is **stable** and (104) is such that all roots z of the characteristic equation

$$|\Phi(z)| = \det(I - \Phi_1 \boldsymbol{z} - \dots - \Phi_p \boldsymbol{z}^p) = 0 \tag{105}$$

are in modulus greater than one, i.e.,

$$\nexists z : |z| \leq 1 \text{ and } |\Phi(z)| = 0. \tag{106}$$

Def. (Companion form)
Assume that $\boldsymbol{Y}_t \sim \mathrm{VAR}(p)$, then the **companion form** of the autoregressive process is defined as

$$\begin{pmatrix} \boldsymbol{y}_t \\ \boldsymbol{y}_{t-1} \\ \boldsymbol{y}_{t-2} \\ \vdots \\ \boldsymbol{y}_{t-p+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\nu} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_{p-1} & \Phi_p \\ I_k & 0 & \cdots & 0 & 0 \\ 0 & I_k & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_k & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_{t-1} \\ \boldsymbol{y}_{t-2} \\ \boldsymbol{y}_{t-3} \\ \vdots \\ \boldsymbol{y}_{t-p} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{107}$$

which can be represented more compactly as a VAR(1) model on the modified vector

$$\boldsymbol{y}_t^* = \boldsymbol{a}_0 + A_1 \boldsymbol{y}_{t-1}^* + \boldsymbol{u}_t, \quad \boldsymbol{u}_t \sim \mathrm{N}(\mathbf{0}, \Sigma_u), \tag{108}$$

with $\Sigma_u = \mathrm{diag}(1, 0, \dots, 0) \otimes \Sigma$.

Prop. 34 (Stability condition of the companion form)
Let $\boldsymbol{Y}_t \sim$ VAR(p), then we can use the companion form (107) (and the compact version (108)) to write the characteristic equation (105) of $\boldsymbol{Y}_t$ as

$$\det(I_{kp} - A_1 \boldsymbol{z}) = 0.$$

Prop. 35 (Mean of a VAR process)
Assume that $\boldsymbol{Y}_t \sim$ VAR(p) is a stable process, then the unconditional mean of $\boldsymbol{Y}_t$ is

$$\mathbb{E}[\boldsymbol{Y}_t] = \boldsymbol{\nu} + \sum_{j=1}^{p} \Phi_j \, \mathbb{E}[\boldsymbol{y}_{t-j}] = (I_{pk} - A_1)^{-1} \boldsymbol{a}_0.$$

Prop. 36 (Variance-covariance of a VAR process)
Assume that $\boldsymbol{Y}_t \sim$ VAR(p) is a stable process, then cross-covariance function of $\boldsymbol{Y}_t^*$ (defined in ...) satisfies the following recursion,

$$\Gamma_{\boldsymbol{y}^*}(0) = A_1 \Gamma_{\boldsymbol{y}^*}(0) A_1^\top + \Sigma_u$$

$$\Gamma_{\boldsymbol{y}^*}(h) = A_1 \Gamma_{\boldsymbol{y}^*}(h-1),$$

and for the original process $\boldsymbol{Y}_t$ we can select the first covariance block of the model in companion form,

$$\Gamma_{\boldsymbol{y}}(h) = J \Gamma_{\boldsymbol{y}^*(h)}(h),$$

where $J = \begin{pmatrix} I_k & 0_{k\times k} & \dots & 0_{k\times k} \end{pmatrix}$.

Remark. The initial values of the recursion can be determined by exploiting the properties of the vec operator,

$$\begin{aligned} \operatorname{vec} \Gamma_{\boldsymbol{y}^*(0)} &= (A_1 \otimes A_1) \operatorname{vec}\left(\Gamma_{\boldsymbol{y}^*}(0)\right) + \operatorname{vec} \Sigma_{\boldsymbol{u}} \\ &= (I_{(kp)^2} - A_1 \otimes A_1)^{-1} \operatorname{vec} \Sigma_{\boldsymbol{u}}. \end{aligned}$$

25.2.1 Estimation

In order to estimate the VAR(p) model, we can recast the problem as a multivariate linear regression problem. Specifically, define

$$\begin{aligned}
Y &= \begin{pmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_T \end{pmatrix} && \in \mathbb{R}^{k\times T} \\
B &= \begin{pmatrix} \Phi_0 & \Phi_1 & \Phi_2 & \cdots & \Phi_p \end{pmatrix} && \in \mathbb{R}^{k\times(kp+1)} \\
\varepsilon &= \begin{pmatrix} \boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2 & \cdots & \boldsymbol{\varepsilon}_T \end{pmatrix} && \in \mathbb{R}^{k\times T} \\
z_t &= \begin{pmatrix} 1 & \boldsymbol{y}^\top & \boldsymbol{y}_{t-1}^\top & \cdots & \boldsymbol{y}_{t-p+1}^\top \end{pmatrix} && \in \mathbb{R}^{(kp+1)\times 1} \\
Z &= \begin{pmatrix} \boldsymbol{z}_1 & \boldsymbol{z}_2 & \cdots & \boldsymbol{z}_{T-1} \end{pmatrix} && \in \mathbb{R}^{(kp+1)\times T}
\end{aligned}$$

Def. (Matrix and vector formulation of the VAR(p))
We define the **matrix and vector formulation of the VAR(p)** as

$$Y = ZB + \varepsilon, \quad \varepsilon \sim \mathrm{MN}_{k\times t},$$

which can then be rewritten in terms of a standard linear regression by using the vec operator,

$$\mathrm{vec}(Y) = (I_{kp+1} \otimes Z)\,\mathrm{vec}(B) + \mathrm{vec}(\varepsilon)$$

$$\boldsymbol{y} = (I_{kp+1} \otimes Z)\boldsymbol{\beta} + \boldsymbol{u}.$$

Remark. From this representation, ordinary least squares or conjugate Bayesian analysis can provide a simple solution to the estimation of the VAR model.

25.2.2 Prediction

For predicting new observations, in a Bayesian setting we average over the posterior distribution of the parameters in order to obtain a predictive distribution,

$$\boldsymbol{y}_t(h) = \int f(\boldsymbol{y}_{t+h}^*|\Theta)p(\Theta|Y)d\Theta,$$

where we indicate by Θ the whole set of unknown parameters.

25.3 Prior shrinkage

Dynamic autoregressive processes are powerful tools for the analysis of multivariate time series. Flexibility is a positive aspect of these processes that has fostered their application in many areas but from a statistical point of view it often results in over-parameterization and over-fitting, which inevitably leads to inaccurate predictions.

Some solutions have been proposed, such as

› variable selection priors (**george2008**);

› factor models (Stock and Watson, 2006);

› steady-state priors (Villani, 2008);

› Bayesian model averaging (Garratt et al., 2009);

› combination forecasts.

25.3.1 Bayesian factor models

Bayesian factor models are relevant since they reduce the complexity of the problem. Extending the factor model to the dynamic setting can be done via state-space models, which will be the topic of the Kalman filter specialist course.

Def. (Factor model)
A **factor model** is defined as the hierarchical model

$$\begin{aligned} \boldsymbol{Y}_t|\boldsymbol{f}_t &\sim N(B\boldsymbol{f}_t, \Sigma) \\ \boldsymbol{f}_t &\sim N(0, I_k), \quad t = 1, 2, \dots, \end{aligned} \tag{109}$$

where $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_m^2)$ and B is the matrix of the factor loadings with $B_{ii} > 0$ for all i.

Remark. Due to the model structure (109), the following restrictions are placed on $\boldsymbol{Y}_t$,

$$\begin{aligned} \mathbb{V}[\boldsymbol{Y}_t|\boldsymbol{f}_t] &= \Sigma, \\ \mathbb{V}[\boldsymbol{Y}_t] &= BB^\top + \Sigma. \end{aligned}$$

By using the same idea as the matrix-variate form of the model, we can write (109) as

$$\begin{aligned} Y &= F\Theta + \varepsilon, \qquad \varepsilon \sim \text{N}_{m\times T}(0, \Sigma \otimes I_T) \\ F &\sim N_{kT}(0, I_{kT}). \end{aligned}$$

Remark. In order to estimate the model from a frequentist point of view, we can interpret F as having a normal prior distribution and therefore use Bayesian methods to perform inference.

Prop. 37 (Invariance of the factor model)
The basic factor model is invariant with respect to orthogonal transformations,

$$\begin{cases} B^* = BP^\top \\ \boldsymbol{f}_t^* = P\boldsymbol{f}_t \end{cases}, \quad \textit{with } P^\top P = I_k.$$

Remark. The classical solution is to impose a restriction such as $B^\top\Sigma^{-1}B = I_k$, or to use a particular block structure for B.

25.4 Optimization methods

References: Lange (2015)
Lange (2016)

In this last part of the course, we will discuss some optimization methods which may find wide applicability in any statistical problem. Suppose that we want to solve a difficult optimization problem,

$$\underset{x}{\operatorname{argmin}} f(x), \tag{110}$$

and we instead have a function $g(x)$ that is much simpler to optimize with the following properties:

(i) There exists at least one x^* such that

$$f(x^*) = g(x^*).$$

(ii) $f(x)$ is majorized by $g(x)$ in all other points, that is,

$$f(x) < g(x) \quad \text{for all } x \in \mathcal{X} \setminus \{x^*\}.$$

Def. (Surrogate function)
Suppose that (i) and (ii) above hold, then $g(x)$ is called a **surrogate function** of $f(x)$.

Remark. If the optimization problem is convex, we can solve the optimization problem for $g(x)$, and property (i) drives the solution towards the optimal solution of (110). However, the price that we pay is that we have to iterate the minimization of g in order to reach the true minimum (110).

Remark. If f is concave, we only have to replace (ii) with $f(x) > g(x)$. This method is called "MM", that is, either **Minimization of a Maximizer** or **Maximization of a Minorizer**.

Valid majorizing functions can be found by leveraging some well-known inequalities:

- Jensen's inequality - EM algorithm;
- quadratic upper bound principle;
- supporting hyperplane for a convex function - Gradient descent;
- chord above the graph property of a convex function;
- arithmetic-geometric mean inequality;
- Cauchy-Schwartz inequality.

References

Albert, J. H. and Chib, S. (1993). «Bayesian Analysis of Binary and Polychotomous Response Data». In: *Journal of the American Statistical Association* 88.422, 669–679.

Garratt, A. et al. (2009). «Real-Time Prediction With U.K. Monetary Aggregates in the Presence of Model Uncertainty». In: *Journal of Business & Economic Statistics* 27.4, 480–491.

George, E. I. and McCulloch, R. E. (1997). «Approaches for Bayesian Variable Selection». In: *Statistica Sinica* 7.2, 339–373.

Green, P. J. (1995). «Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination». In: *Biometrika* 82.4, 711–732.

Lange, K. (2015). *Optimization*. Second. New York: Springer.

Lange, K. (2016). *MM Optimization Algorithms*. SIAM.

Mitchell, T. J. and Beauchamp, J. J. (1988). «Bayesian Variable Selection in Linear Regression». In: *Journal of the American Statistical Association* 83.404, 1023–1032.

Polson, N. G. et al. (2013). «Bayesian Inference for Logistic Models Using Polya-Gamma Latent Variables». In: *arXiv:1205.0310 [stat]*. arXiv: 1205.0310 [stat].

Stock, J. H. and Watson, M. (2006). *Forecasting with Many Predictors*. Handbook of Economic Forecasting. Elsevier, 515–554.

Villani, C. (2008). *Optimal Transport: Old and New*. 2009th edition. Aalborg: Springer.

Part V

Kalman filter and dynamic linear models

Instructor: S.J. Koopman

References: Durbin and Koopman, 2012

In this part of the course we will discuss models based on dynamic Gaussian linear specification, via the use of the Kalman filter relationships. This formula will be the basis for parameter estimation, signal extraction, filtering, and prediction of future values.

Starting from the local level model, which is a stylized version of the dynamic liner model, we will build the intuition for more complex dynamics which can be used to model general relationships over time.

Lecture 26: Introduction and Local Level Model

2022-06-22

A time series is a set of observations y_t, each one recorded at a specific time t and ordered over time. We assume to have $t = 1, \dots, n$ observations and regard the time series $y_1, y_2, \dots, y_n$ as a realized sample from a **stochastic process**. Our goal is to model the evolution of the time series over time, which finds relevant application in a wide variety of tasks and fields, including economic policy, financial decision making, climate change monitoring, and forecasting.

26.1 Time series

Much of time series analysis and forecasting is about distinguishing the signal from the noise from observed data.

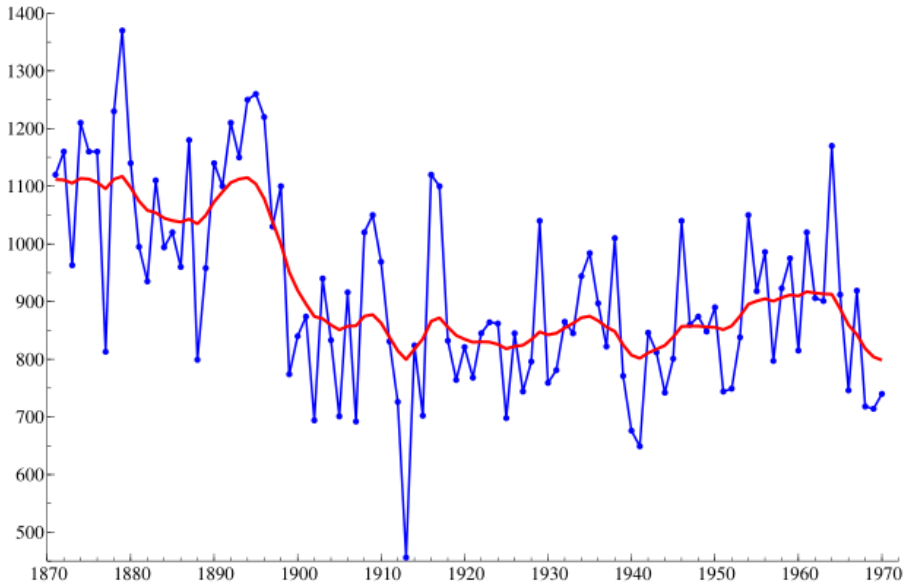


Figure 58: Signal and noise for the `nile` dataset.

Over the years, the literature of time series modelling has developed a host of tools to deal with these types of data, including

- **ARMA-type models**
  - Autoregressive models
  - ARMA models
  - Long-memory models and fractional integration
  - Vector autoregressive models, cointegration, vector error correction models
- **Latent variable models**
  - Regime switching
  - Markov-switching
  - Threshold autoregression

  - Smooth transition models.

› **State-space models**

  - Dynamic regression and error correction models
  - General state space models

› Generalized autoregressive conditional heteroskedasticity (GARCH) models

› Autoregressive conditional duration models and related models

26.2 Local level model

The local level model is defined using the following two-stage model.

Def. (Local level model)
The **local level model** is defined as the two-stage model

$$y_t = \mu + \varepsilon_t, \qquad \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$$

$$\mu_{t+1} = \mu_t + \eta_t, \qquad \eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\eta^2).$$

Remark. The time-varying level is modelled as a random walk process whose updates are independent from ε_s for all s and are updated in $t+1$ instead of t for easier computation later on.

Initial conditions. The model requires an initial specification for μ_1 in order to be estimated.

Stationarity. Note that the processes for both μ_t and for y_t are nonstationary.

The local level model is defined in terms of two unknown parameters, σ_ε^2 and σ_η^2, which are unknown and are used to define the **signal-to-noise ratio**

$$q = \frac{\sigma_\eta^2}{\sigma_\varepsilon^2}.$$

Some trivial special cases are $\sigma_\eta^2 = 0$, which is the i.i.d model and $\sigma_\varepsilon^2 = 0$ which is the random walk model.

Remark. The local level model is a basic illustration of a **state space model**.

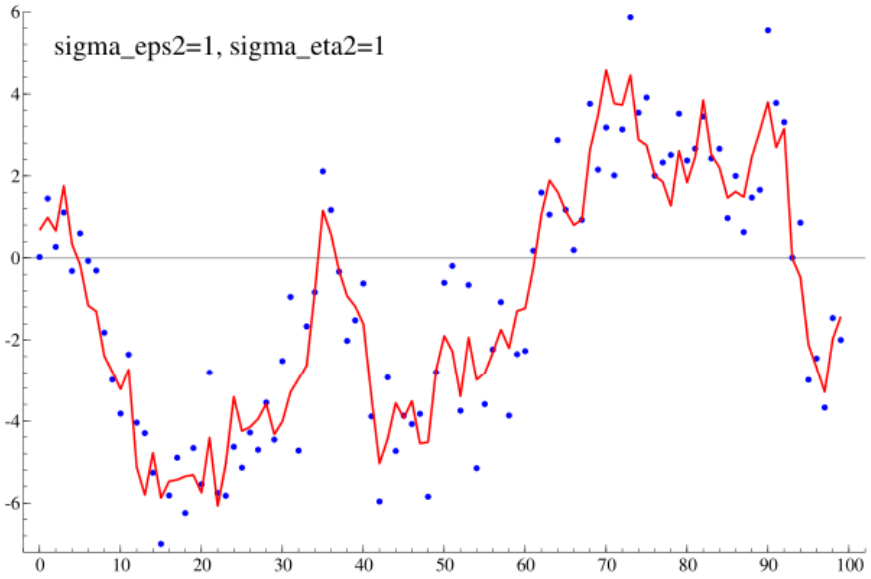


Figure 59: Example of a local level model alongside the true value of the latent variable μ_t in red.

26.2.1 Properties of the LLM

Some basic properties of the local linear model are:

› **Stationarity of Δy_t**,

$$\Delta y_t = \Delta\mu_t + \Delta\varepsilon_t = \eta_{t-1} + \varepsilon_t - \varepsilon_{t-1}.$$

› **Autocovariance of Δy_t** since $\mathbb{E}[\Delta Y_t] = 0$,

$$\gamma_0 = \mathbb{E}[\Delta Y_t \Delta Y_t] = \sigma_\eta^2 + 2\sigma_\varepsilon^2$$

$$\gamma_1 = \mathbb{E}[\Delta Y_t \Delta Y_{t-1}] = -\sigma_\varepsilon^2$$

$$\gamma_\tau = \mathbb{E}[\Delta Y_t \Delta Y_{t-\tau}] = 0 \qquad \text{for } \tau \geq 2.$$

› **Autocorrelation of Δy_t**, since $\rho_\tau = \gamma\tau/\gamma_0$ we can write

$$\rho_0 = 1$$

$$\rho_1 = -\frac{\sigma_\varepsilon^2}{\sigma_\eta^2 + 2\sigma_\varepsilon^2} = -\frac{1}{q+2}$$

$$\rho_\tau = 0 \qquad \text{for } \tau \geq 2.$$

Observe that the autocorrelation is constrained to be

$$-1/2 \leq \rho_1 \leq 0,$$

hence the LLM postulates $\Delta Y_t \sim \text{MA}(1)$ with a constrained acf and $Y_t \sim \text{ARIMA}(0,1,1)$.

Since we can write the LLM as a MA(1) mode, we could specify

$$\Delta Y_t = \xi_t + \theta \xi_{t-1}, \quad \xi_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

whose acf is $\rho_1 = \theta/(1+\theta^2)$. Therefore, we have a restricted parameter space for $\theta \in (-1, 0)$. To express θ as a function of q, we can solve the equality of the acf's to get

$$\theta = \frac{1}{2}(\sqrt{q^2 + 4q} - 2 - q).$$

26.2.2 Generalizations

Some immediate generalizations can take into account a stationary innovation of the level,

$$\mu_{t+1} = \phi \mu_t + \eta_t, \quad |\phi| < 1,$$

or some modifications where $\text{Cov}(\varepsilon_t, \eta_t) = \rho \neq 0$.

26.3 Signal extraction and prediction

We are now interested in the problem of extracting information from $y_1, \ldots, y_t$ in order to estimate the values of the underlying unobservable state.

26.3.1 Signal extraction

Consider a local level model

$$y_t = \mu + \varepsilon_t, \qquad \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$$

$$\mu_{t+1} = \mu_t + \eta_t, \qquad \eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\eta^2),$$

and assume that (i) we have collected observations for $y_1, \ldots, y_{t-1}$ and (ii) the conditional density $f(\mu_t | y_1, \ldots, y_{t-1})$ is normal with **known** mean a_t and **known** variance p_t. Then, we have that

$$\mu_t | y_1, \ldots, y_{t-1} \sim \mathcal{N}(a_t, p_t),$$

and by collecting the observation y_t, the conditional density of interest turns out to be normal as well and

$$\mu_t | y_1, \ldots, y_t \sim \mathcal{N}(a_{t|t}, p_{t|t}).$$

Our goal is to characterize the mean $a_{t|t}$ and variance $p_{t|t}$ of the **filtered states** $\mu_t | y_1, \ldots, y_t$.
Denote the forecast for y_t as

$$\widehat{y}_t = \mathbb{E}[Y_t | y_{1:t-1}] = \mathbb{E}[\mu_t + \varepsilon_t | y_{1:t-1}] = a_t.$$

The corresponding one-step ahead **prediction error** is

$$v_t = y_t - \widehat{y}_t = y_t - a_t,$$

where it holds that

$$\mathbb{E}[v_t] = \mathbb{E}[(\mu_t - a_t + \varepsilon_t)] = 0$$

$$\mathbb{V}[v_t] = \mathbb{V}[(\mu_t - a_t) + \varepsilon_t] = p_t + \sigma_\varepsilon^2,$$

and therefore $Y_t|y_{1:t-1} \sim \mathcal{N}(a_t, p_t + \sigma_\varepsilon^2)$.
To obtain an expression for $a_{t|t}$ and $p_{t|t}$ we observe that

$$\mu_t|y_{1:t} \overset{\mathrm{d}}{=} \mu_t|v_t, y_{1:t-1},$$

since $v_t = y_t - \mathbb{E}[\mu_t|y_{1:t-1}]$. Then, we have that

$$\begin{aligned} f(\mu|v_t, y_{1:t-1}) &= f(\mu_t, v_t|y_{1:t-1})/f(v_t|y_{1:t-1}) \\ &= f(\mu_t|y_{1:t-1})f(v_t|\mu_t, y_{1:t-1})/f(|y_{1:t-1}) \\ &= \mathcal{N}(a_t, p_t) \times \mathcal{N}(\mu_t - a_t, \sigma_\varepsilon^2)/\mathcal{N}(0, p_t + \sigma_\varepsilon^2), \end{aligned}$$

and by completing the square we have that the **filter density** $f(\mu_t|y_{1:t})$ is $\mathcal{N}(a_{t|t}, p_{t|t})$, where

$$k_t = \frac{p_t}{p_t + \sigma_\varepsilon^2}$$

$$a_{t|t} = a_t + k_t v_t$$

$$p_{t|t} = k_t \sigma_\varepsilon^2.$$

Moreover, we have that the **predicted signal** density $f(\mu_{t+1}|y_{1:t})$ is $N(a_{t+1}, p_{t+1})$ where

$$a_{t+1} = \mathbb{E}[\mu_{t+1}|y_{1:t}] = \mathbb{E}[\mu_t + \eta_t|y_{1:t}] = a_{t|t}$$

$$p_{t+1} = \mathbb{V}[\mu_t + \eta_t|y_{1:t}] = p_{t|t} + \eta_\eta^2.$$

Lecture 27: State space methods

2022-06-23

The important key variables in a state space model are the observation vector y_t and the state vector α_t, which contains all the dynamic features in the model. Here, we will discuss the main methods related to such models.

27.1 State space model

The state space model is a convenient representation or formulation for almost all linear Gaussian time series models, and is used mostly for the purpose of using the Kalman filter and its related algorithms.

Def. (State space model)
The linear Gaussian **state space model** is defined by three components:

- **Observation equation**:

$$\boldsymbol{y}_t = Z_t \boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{\text{iid}}{\sim} \mathcal{N}(0, H_t)$$

- **State transition equation**:

$$\boldsymbol{\alpha}_{t+1} = T_t \boldsymbol{\alpha}_t + R_t \boldsymbol{\zeta}_t, \quad \boldsymbol{\zeta}_t \overset{\text{iid}}{\sim} \mathcal{N}(0, Q_t),$$

  with $\boldsymbol{\zeta}_t \perp\!\!\!\perp \boldsymbol{\varepsilon}_s$ for all t, s.

- **Initial condition**: $\boldsymbol{\alpha}_1 \sim N(\boldsymbol{a}_1, P_1)$ independently from all $\boldsymbol{\zeta}_t, \boldsymbol{\varepsilon}_s$.

Remark. The system matrices T_t, Z_t, R_t, Q_t, H_t are fixed at time t, and are known functions of the parameter vector. They determine the dynamic structure of the model.

Example (Local level model)
The local level model is the simplest example of a state space model, which can be obtained by setting the transition matrices $Z_t = T_t = 1$, $R_t = 1$ and the process variances $H_t = \sigma_\varepsilon^2$ and $Q_t = \sigma_\eta^2$.

Example (AR(1) model)
The AR(1) model can be written in state space form by setting $Z_t = 1$, $T_t = \phi$, $R_t = 1$, $H_t = 0$, and $Q_t = \sigma^2$.

Example (ARMA(p,q) model)
A general ARMA(p, q) model can be written in state space model by choosing the transition

matrices

$$Z_t = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}, \quad T_t = \begin{pmatrix} \phi_1 & 1 & 0 & \dots & 0 \\ \phi_2 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{m-1} & 0 & 0 & \dots & 1 \\ \phi_m & 0 & 0 & \dots & 0 \end{pmatrix},$$

and the innovation matrices

$$R_t = \begin{pmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{m-1} \end{pmatrix}, \quad H_t = 0, \quad Q_t = \sigma^2.$$

Generalization to ARIMA(p, d, q) are also possible (Durbin and Koopman, 2012).

Example (Dynamic linear regression)
A dynamic linear regression model can be obtained by defining

$$y_t = X_t\beta_t + \varepsilon_t,$$

$$\beta_{t+1} = \beta_t + \eta_t.$$

Take home. The state space form encompasses many standard models that are usually employed in statistics:

- Unobserved components time series models.
- ARIMA models.
- Dynamic regression models.
- Sums of linear dynamic components (Durbin and Koopman, 2012).
- Any linear Gaussian dynamic process (Durbin and Koopman, 2012).

27.2 Initial conditions

27.2.1 Stationary process

For the AR(1) model, the initial condition $\alpha_1 \sim \mathcal{N}(a_1, p_1)$ is the unconditional distribution of the stationary process of μ_t. Hence, we can choose

$$a_1 = 0, \quad p_1 = \sigma_\eta^2/(1-\phi^2),$$

and this approach leads to a general solution for all stationary elements in the state vector. For a more general model,

$$\begin{aligned}\mathbb{V}[\alpha_{t+1}] = \mathbb{V}[\alpha_t] &\iff \mathbb{V}[\alpha_{t+1}] = T\,\mathbb{V}[\alpha_t]T^\top + RQR^\top \\ &\iff P_1 = TP_1T^\top + RQR^\top \\ &\iff \operatorname{vec}(P_1) = (I - T \otimes T)^{-1}\operatorname{vec}(RQR^\top).\end{aligned}$$

27.2.2 Non-stationary process

When the process is non-stationary, we might have something like the random walk model. In this case, $a_1 = 0$ but

$$p_1 = \frac{\sigma_\eta^2}{(1-\phi^2)} \xrightarrow{\phi \to 1} \infty,$$

and therefore we cannot proceed as we did before to obtain p_1.

27.3 Kalman filter

Given the state space model, the unobserved state α_t can be estimated from the observations through the use of the Kalman filter. We will need the following two lemmas in order to obtain the required equations.

Lemma 6 (Conditioning normal distributions)
Suppose (X, Y) are jointly normal, then we have that

$$\mathbb{E}[X|Y] = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$$

$$\mathbb{V}[X|Y] = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}^\top.$$

Lemma 7 (Conditioning normal distributions (ii))
Suppose (X, Y, Z) are jointly normal with $\mathbb{E}[Z] = 0$ and $\Sigma_{yz} = 0$, then we have that

$$\mathbb{E}[X|Y, Z] = \mathbb{E}[X|Y] + \Sigma_{xz}\Sigma_{zz}^{-1}z$$

$$\mathbb{V}[X|Y, Z] = \mathbb{V}[X|Y] - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{xz}^\top.$$

Theorem 15 (Kalman filter)
If we define the following quantities,

$$\begin{aligned}
\nu_t &= y_t - Z_t a_t \\
F_t &= Z_t P_t Z_t^\top + H_t \\
K_t &= (T_t P_t Z_t^\top) F_t^{-1} \\
a_{t+1} &= T_t a_t + K_t \nu_t \\
P_{t+1} &= T_t P_t T_t^\top + R_t Q_t R_t^\top - K_t F_t K_t^\top,
\end{aligned}$$

then, for the state space model the distribution of the filtered states are given by

$$\begin{aligned}
\alpha_{t+1}|y_{1:t} &\sim \mathcal{N}\big(a_{t+1}, P_{t+1}\big) \\
a_{t+1} &= \mathbb{E}\left[\alpha_{t+1}|y_{1:t}\right], \\
P_{t+1} &= \mathbb{V}\left[\alpha_{t+1}|y_{1:t}\right]
\end{aligned}$$

whereas the predictive distribution for $y_{t+1}|y_{1:t}$ is

$$\begin{aligned}
Y_{t+1}|y_{1:t} &\sim \mathcal{N}\big(\hat{y}_{t+1}, F_{t+1}\big) \\
\hat{y}_{t+1} &= \mathbb{E}\left[Y_{t+1}|y_{1:t}\right], \\
F_{t+1} &= \mathbb{V}\left[Y_{t+1}|y_{1:t}\right],
\end{aligned}$$

Proof.
Consider the set of observations $y_{1:t} = \{y_{1:t-1}, y_t\} \equiv \{y_{1:t-1}, \nu_t\}$, then we have that $\mathbb{E}[\nu_t y_{t-j}] = 0$ for any $j = 1, \ldots, t-1$. If we define $X = \alpha_{t+1}, Y = y_{1:t-1}, Z = \nu_t$, then we can apply 7 to obtain

$$\begin{aligned}
\mathbb{E}[X|Y] &= \mathbb{E}[\alpha_{t+1}|y_{1:t-1}] = T_t\,\mathbb{E}[\alpha_t|y_{1:t-1}] + R_t\,\mathbb{E}[\zeta_t] = T_t a_t. \\
\Sigma_{xz} &= \mathbb{E}[\alpha_{t+1}\nu_t^\top] = T_t\,\mathbb{E}[\alpha_t \nu_t^\top] + R_t\,\mathbb{E}[\zeta_t \nu_t^\top] = T_t P_t Z_t^\top \\
\Sigma_{zz} &= \mathbb{V}[v_t] = F_t.
\end{aligned}$$

Finally, by 7 we have

$$\begin{aligned}
\mathbb{E}[X|Y,Z] &= \mathbb{E}[X|Y] + \Sigma_{xz}\Sigma_{zz}^{-1} z = T_t a_t + T_t P_t Z_t^\top F_t^{-1} \nu_t = T_t a_t + K_t \nu_t \\
P_{t+1} &= T_t P_t T_t^\top + R_t Q_t R_t^\top - K_t F_t K_t^\top.
\end{aligned}$$

□

The derived Kalman filter computes the **predictions** of the states directly,

$$a_{t+1} = \mathbb{E}[\alpha_{t+1}|y_{1:t}], \quad P_{t+1} = \mathbb{V}[\alpha_{t+1}|y_{1:t}].$$

However, we can also work with the **filtered estimates**,

$$a_{t|t} = \mathbb{E}[\alpha_t | y_{1:t}], \quad P_{t|t} = \mathbb{V}[\alpha_t | y_{1:t}],$$

which can be updated to the predicted values for the states, which for clearness we now indicate as $a_{t+1} = a_{t+1|t}$, $P_{t+1} = P_{t+1|t}$. The alternative Kalman filter is obtained by defining

$$\begin{aligned} M &= P_t Z_t^\top F_t^{-1} \\ a_{t|t} &= a_{t|t-1} + M_t \nu_t \\ P_{t|t} &= P_{t|t-1} - M_t F_t M_t^\top \\ a_{t+1|t} &= T_t a_{t|t} \\ P_{t+1|t} &= T_t P_{t|t} T_t^\top + R_t Q_t R_t^\top. \end{aligned}$$

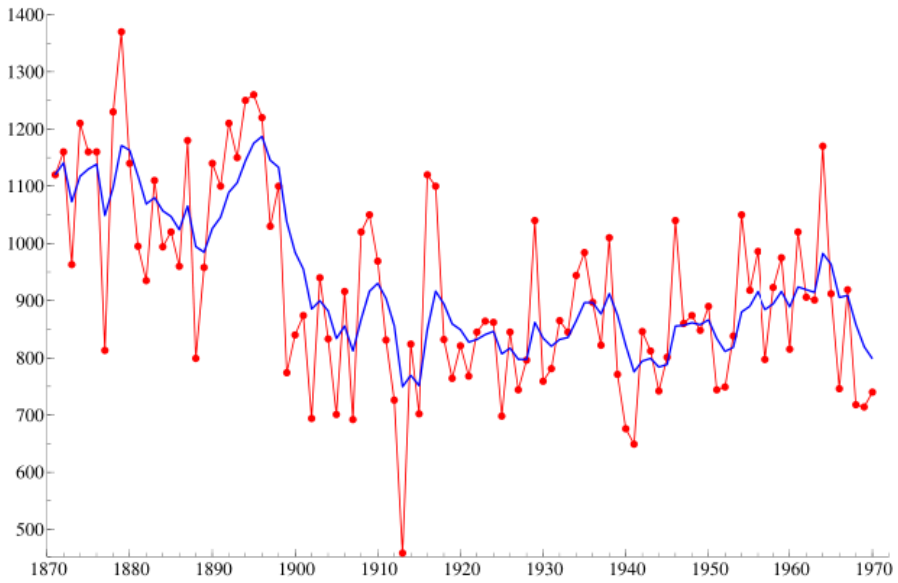


Figure 60: Filtered states (blue) for the `nile` dataset, using a local level model.

Remark. Under correct model specification, the standardised residuals are such that

$$\frac{\nu_t}{\sqrt{F_t}} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1),$$

and thus we can apply standard tests for normality, heteroskedasticity, and serial correlation. A recursive algorithm is also available for calculating smoothed disturbances (auxiliary residuals), which can be used to **detect breaks** and outliers;

Finally, we can perform **model comparison** and parameter restriction via the use of likelihood ratio tests, AIC and BIC.

References

Durbin, T. l. J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods*. 2 edizione. Oxford: OUP Oxford.