UNIVERSITÉ DE PARIS
Laboratoire MAP5 UMR CNRS 8145
École doctorale 386 : Sciences Mathématiques de Paris Centre

THÈSE
Pour obtenir le grade de
DOCTEUR EN MATHÉMATIQUES

Spécialité : Mathématiques appliquées

Présentée par

VIVIEN GOEPP

AN ITERATIVE REGULARIZED METHOD FOR SEGMENTATION WITH APPLICATIONS TO STATISTICS

Sous la direction d'OLIVIER BOUAZIZ et GRÉGORY NUEL.

Soutenue publiquement le 27 septembre 2019 devant un jury composé de

Olivier BOUAZIZ	Université Paris Descartes	Directeur de thèse
Julien CHIQUET	Agro ParisTech	Rapporteur
Chantal GUIHENNEUC	Université Paris Descartes	Examinatrice
Hélène JACQMIN-GADAA	Université de Bordeaux	Examinatrice
Catherine LEGRAND	Université Catholique de Louvain	Rapporteure
Grégory NUEL	Sorbonne Université	Directeur de thèse
Jean-Christophe THALABARD	Université Paris-Descartes	Examinateur
Jean-Philippe VERT	Mines ParisTech	Examinateur

ABSTRACT

This thesis deals with the development of regularized methods using penalized maximum likelihood estimation. More specifically, I use a sparsity-inducing iterative method called adaptive ridge. The latter is competitive compared to other approaches, namely in terms of ease of implementation and computational cost. My work consists in the application of this method to a wide range of problems: survival analysis, spline regression, and spatial segmentation. Applications in several problematics show that the adaptive ridge's good performance in selection, great ease of implementation and low computational cost can make it a good starting point in penalization-base variable selection.

In survival analysis, data are often collected by following a cohort, in which case the events are widely spread through time and the sample is suspected to present heterogeneity. I first focus on developing a method for the inference of the incidence, which allows to detect heterogeneity with respect to the date of birth (or cohort). A closely related problem is the study of the evolution of the inference as a joint function of the age, the date of birth (cohort), and the calendar time (period). Epidemiologists have long resorted to the age-period-cohort model or its submodels. The latter assume linear effects of each variable, which is deemed too simplistic to estimate potentially important features of the incidence. In this framework, I develop a model allowing for the joint estimation of two variables' effects and of their interaction.

Spline regression is known to be a competitive method for non-parametric regression. However the estimated spline depends highly on the initial choice of knots and choosing the best knots is a computationally hard problem. I propose an approach for the estimation of the best knots jointly with the spline function. By initiating a large number of knots and successively removing the least relevant ones, my method makes a slightly restrictive hypothesis to remove much of the computational burden.

In spatial statistics, the spatial domain is often divided into "units" and data are gathered at the unit level. The spatial effect is estimated on each unit and its representation is subject to the arbitrary of the unit division, which makes its interpretation difficult. This can be resolved by regularization, which reduces the variance and increases the interpretability. I present a model for segmentation of spatial data based on the adjacency structure of the units.

RÉSUMÉ

Cette thèse porte sur l'élaboration de méthodes régularisées utilisant l'estimation par maximum de vraisemblance pénalisée. Plus précisément, j'utilise une méthode parsimonieuse itérative, appelée adaptative ridge. Cette dernière est compétitive par rapport à d'autres approches, notamment en termes de facilité de mise en œuvre et de temps de calcul. Mon travail consiste à appliquer cette méthode à un large éventail de problèmes : l'analyse de survie, la régression par *splines* et la segmentation spatiale. Ces applications dans différentes problématiques montrent que la bonne performance de l'*adaptive ridge* en sélection, sa grande facilité de mise en œuvre et son faible coût de calcul peuvent en faire un bon point de départ dans les méthodes de sélection de variable par pénalisation.

En analyse de la survie, les données sont souvent recueillies en suivant une cohorte, auquel cas les événements sont largement répartis dans le temps et l'échantillon peut présenter une hétérogénéité. Je me concentre d'abord sur le développement d'une méthode d'estimation de l'incidence qui permet de détecter l'hétérogénéité par rapport à la date de naissance (ou cohorte). Un problème proche est l'étude de l'évolution de l'inférence en fonction de l'âge, de la date de naissance (*cohort*) et de la date calendaire (*period*). Les épidémiologistes ont longtemps eu recours au modèle age-period-cohort ou à ses sous-modèles. Ces dernières supposent des effets linéaires de chaque variable, ce qui est jugé trop simpliste pour estimer des caractéristiques potentiellement importantes de l'incidence. Dans ce cadre, j'élabore un modèle estimant cojointement l'effet de deux variables et de leur interaction.

La régression par splines est connue pour être une méthode performante de régression non paramétrique. Cependant, la spline estimée dépend fortement du choix initial des nœuds et le choix des meilleurs nœuds est un problème difficile en pratique. Je propose une approche permettant l'estimation des meilleurs nœuds conjointement avec la fonction spline. En initiant un grand nombre de nœuds et en supprimant successivement les moins pertinents, ma méthode fait une hypothèse légèrement restrictive pour diminuer grandement le temps de calcul.

En statistiques spatiales, le domaine spatial est souvent divisé en "unités" et les données sont recueillies au niveau des unités. L'effet spatial est estimé sur chaque unité et sa représentation est soumise à l'arbitraire de la division de l'unité, ce qui rend son interprétation difficile. Ceci peut être résolu par la régularisation, ce qui réduit la variance et augmente l'interprétabilité. Je présente un modèle de segmentation des données spatiales basé sur la structure d'adjacence des unités.

Plan of this thesis

This thesis is organised as follows:

Chapter 1 introduces, compares, and discusses the main statistical approaches to model selection. A great emphasis is laid on penalized likelihood methods. First, I detail the most famous methods of penalized maximum likelihood estimate: the lasso, the elastic-net, and further refinements. I develop on the use of these penalties in the linear model. Secondly, I touch on two of the methods of model selection which are not based on penalization: best subset selection and stepwise selection. In the third section, I introduce the Majorize-Minimization (MM) optimization scheme, which, applied to penalized likelihoods, yields two important iterative penalized methods: the Local Linear Approximation (also termed *adaptive lasso*) and the Local Quadratic Approximation. I then introduce the iteratively defined penalized method used throughout this work: the adaptive ridge. Finally, I develop on the statistical methods which enforce an *a priori* structure on a parameter.

Chapter 2 deals with the application of the adaptive ridge to the context of hazard estimation in survival analysis. After an introduction on the topic, and an illustration of the method to a one-dimensional case, we detail how the fused adaptive ridge allows for a new method of regularized estimation of the bi-dimensional hazard rate, with detection of breakpoints. Even though this methods applies to a wide array of problems, we illustrate it through the angle of age-period-cohort analysis.

Chapter 3 deals with the same problem than Chapter 1, but through the context age-period-cohort analysis. We first introduce the topic of age-period-cohort analysis, as well as the use and drawbacks of the age-period-cohort model. Then, we develop on the "Age-Cohort-Interaction " model, which builds on the works of the previous chapter. This model can be viewed as a generalization of the age-period-cohort analysis, which does not suffer from its defects, at the added cost of computation time.

Chapter 4 deals with a different application of the adaptive ridge. In this part, we apply this method to the problem of finding the best knots to support a regression spline. This problem has long been deemed computationally intractable. We show that provided some simplifying assumptions, our new spline regression method can select the best knots as well as the regression spline in a fast fashion. Before developing on our method, we introduce the topic of spline regression and present the main tools and issues in this topic.

Chapter 5 deals with the applications of the adaptive ridge to regularization of spatially-correlated data. When the statistical problem has a spatial structure that is given in already chosen zones, the problem of regularization becomes challenging. We introduce a method for regularization along a graph, with an application to inference for medical data.

The works of Chapters 2 and 4 has lead to two preprints currently under revision. These two chapters consist of these preprints, reproduced as is and preceded by introductory talks on their respective matters. These two papers are given in Sections 2.2 ad 4.2, with respective supplementary materials given in Sections 2.3 and 4.3 respectively. These papers define their own notations, which

are for the most part consistent with the rest of the manuscript. Except for these sections, the present document forms a consistent manuscript, even if the Chapters were divided so as to be able to be read separately.

Scientific communications

Scientific papers

- Bidimensional estimation of the hazard rate [Submitted][1]
- Age-Cohort-Interaction model [In progress]
- Spline regression [Submitted][2]
- Spatial regularization [In progress]

Conferences

Contributed Talks

- L0 Regularization for the estimation of piecewise constant hazard, *SAM Conference 2017*.
- Regularized Hazard Estimation for age-period-cohort Analysis, *International Workshop on Applied Probability 2018*.
- Estimating interactions effects in the age-cohort model, *Research group "Statistiques et santé" 2018*.

Seminars

- Some applications of the *adaptive ridge* to survival analysis, *Meeting of the LPSM Biology Group*, April 2019.
- Estimation régularisée du risque pour l'analyse age-period-cohort, *MAP5 Seminar of Statistics*, 2017.
- Heterogeneity in Survival Analysis, *Rencontre des Jeunes Statisticiens*, 2019.

Posters

- Regularized Hazard Estimation for age-period-cohort analysis, *Statistical Methods for Post-Genomic Data 2018*.
- Regularized Hazard Estimation for age-period-cohort Analysis, *International Biometric Conference 2018*.
- Interaction Effects in Age-Period-Cohort Analysis, *Statistical Methods for Post-Genomic Data 2019*.

[1] `https://hal.archives-ouvertes.fr/hal-01662197v3`
[2] `https://hal.archives-ouvertes.fr/hal-01853459`

Software

- Package `hazreg`, available on github[3].
- Package `aspline`, available on github[4].
- Package `graphseg`, available on github[5].

[3] `https://github.com/goepp/hazreg`
[4] `https://github.com/goepp/aspline`
[5] `https://github.com/goepp/graphseg`

Notations and Definitions

Definitions

- The L_q norm of a vector $\boldsymbol{u}$ is noted $\|\boldsymbol{u}\|_q$. The L_2 norm is also abbreviated $\|\boldsymbol{u}\|$.
- $\mathbb{E}[X]$ and $\mathbb{V}[X]$ denote the respective expectation and variance of the random variable X.
- Tr denotes the trace.
- $\rightarrow_d$ denotes the convergence in distribution.
- $\boldsymbol{A}^T$ denotes the transpose of the matrix $\boldsymbol{A}$.
- $\min^+\{\boldsymbol{u}\}$ denotes the minimum taken only over the positive values of the vector $\boldsymbol{u}$
- $(x)_+$ denotes the positive part of x, that is $\max(x, 0)$.
- $\mathrm{diag}\{u_i\}_i$ is the diagonal matrix whose non-zero entries are $u_1, \ldots, u_n$.
- $\mathcal{N}(.,.)$ denotes the (possibly multivariate) normal distribution. The two arguments are the expectancy and variance.
- $\#A$ denotes the cardinal of the set A.
- When f is a function and x an element from its domain, $f(x^-)$ and $f\left(x^+\right)$ denote the limits $\lim_{t \to x} f(t)$ in the cases $t < x$ and $t > x$, respectively.

Notations

- Vectors and matrices are noted in bold. When necessary, vectors are identified with column matrices.
- n is the sample size
- $i \in \{1, \ldots, n\}$ is the index of the individuals
- p is the number of covariates
- $j \in \left\{1, \ldots, p\right\}$ is the index of the covariates
- In the context of penalized likelihood methods, $\lambda \in \mathbb{R}$ is the penalty constant; in the context of survival analysis $\lambda : \mathbb{R} \mapsto \mathbb{R}^+$ is the hazard rate.
- β is the parameter to be estimated
- $\boldsymbol{I}$ refers to the identity matrix, whose dimension depend on the context.
- $\ell = -\log L$ is the negative log-likelihood and L is the likelihood.
- I use the symbol "$\triangleq$" when the equality serves as a definition.

Abbreviations

- MM: Majorize-Minimization optimization
- NLL: negative log-likelihood
- LQA: Local Quadratic Approximation
- LLA: Local Linear Approximation
- AIC: Akaike Information Criterion
- BIC: Bayesian Information Criterion
- OLS: Ordinary Least Squares (estimate)
- MLE: Maximum Likelihood Estimate
- PCH: Piecewise Constant Hazard (model)
- LARS: Least Angle Regression

Conflicts in notation between chapters

We have tried to use coherent and non-conflicting notations for the mathematical objects defined in this thesis. However, for the sake of consistency with the conventions of the field, we made the choice to keep conventional notations for known quantities. The instantaneous hazard rate for instance, is noted $\lambda(t)$ (as a function of t) as is standard in survival analysis and in the study of stochastic processes. In other parts of the manuscript, we also used the variable λ to denote the penalty constant in penalized maximum likelihood methods.

These notational conflicts have been kept to ease the understanding of the manuscript. They occur between different chapters but not inside each chapter. We stress that the potential uncertainty is removed when the context is taken into consideration.

Contents

Abstract i

Plan of the thesis v

Scientific communications vii

Definitions and Notations ix

1 Introduction 1
- 1.1 Regularized estimation . . . 4
  - 1.1.1 Ridge regression . . . 5
  - 1.1.2 Lasso estimation . . . 7
  - 1.1.3 Elastic-net . . . 12
  - 1.1.4 Bridge regression . . . 12
  - 1.1.5 Berhu . . . 13
  - 1.1.6 Penalized likelihood methods as a Bayesian prior . . . 13
  - 1.1.7 Non-concave penalties . . . 15
    - 1.1.7.1 Hard-thresholding . . . 17
    - 1.1.7.2 SCAD . . . 18
    - 1.1.7.3 Logarithmic penalty . . . 18
    - 1.1.7.4 Available implementations . . . 19
- 1.2 Subset selection . . . 19
  - 1.2.1 Best subset selection . . . 19
  - 1.2.2 Stepwise Selection . . . 20
- 1.3 Iterative penalized methods . . . 21
  - 1.3.1 MM optimization . . . 21
  - 1.3.2 MM optimization applied to non-concave penalties . . . 22
    - 1.3.2.1 Local Quadratic Approximation . . . 23
    - 1.3.2.2 Local Linear Approximation . . . 24
  - 1.3.3 Adaptive Ridge . . . 25
    - 1.3.3.1 Relation to similar procedures . . . 27
    - 1.3.3.2 Numerical performance . . . 28
- 1.4 Structured variable selection . . . 28
  - 1.4.1 Group lasso . . . 28
  - 1.4.2 Overlapping groups . . . 30
  - 1.4.3 Hierarchical structured sparsity . . . 31
  - 1.4.4 Fused lasso and total variation . . . 32
  - 1.4.5 Fused Adaptive Ridge . . . 34
- 1.5 Conclusion . . . 35

2 Regularized estimation of the hazard rate **37**
2.1 Introduction . . . 38
2.1.1 Survival Analysis . . . 38
2.1.2 The piecewise constant hazard estimation . . . 40
2.1.2.1 Proportional hazard model with piecewise constant hazard . . . 41
2.1.3 Cohort data . . . 42
2.2 Regularized estimation of the hazard rate . . . 43
2.3 Application to the evolution of breast cancer mortality . . . 66

3 Estimating interactions in the age-cohort model **73**
3.1 Introduction to age-period-cohort analysis . . . 74
3.2 The age-cohort-interaction model . . . 75
3.3 The estimating procedure . . . 76
3.4 Choice of the penalty constant . . . 77
3.5 Simulation results . . . 77
3.5.1 Simulation setting . . . 77
3.5.2 Predictive performance . . . 79
3.5.3 Perspective plots . . . 81
3.6 Conclusion . . . 84

4 Spline regression with automatic knot selection **89**
4.1 Introduction to spline regression . . . 91
4.1.1 Spline regression . . . 91
4.1.1.1 Definition of splines . . . 91
4.1.1.2 The truncated power basis . . . 92
4.1.1.3 B-spline basis . . . 93
4.1.1.4 Regression using splines . . . 95
4.1.2 Penalized approaches . . . 97
4.1.2.1 Smoothing splines . . . 97
4.1.2.2 O'Sullivan Penalized Splines . . . 98
4.1.2.3 P-splines . . . 98
4.2 Spline regression with automatic knot selection . . . 99
4.3 Comparison of A-spline with P-spline . . . 129

5 Segmentation of spatial data **135**
5.1 Introduction . . . 136
5.2 A model for spatial segmentation . . . 137
5.2.1 Using graphical data for spatial segmentation . . . 137
5.2.2 Segmentation on a graph . . . 138
5.2.3 The adaptive ridge algorithm on a graph . . . 139
5.3 Simulation . . . 142
5.4 Real data application: Overweight prevalence in the Netherlands . . . 146
5.5 Conclusion . . . 148

6 Conclusion **149**

Conclusion **149**

List of Figures **154**

List of Tables **155**

Bibliography **162**

Chapter 1

Introduction

This thesis deals with the application of penalized methods to different statistical problems. These works all have in common the use of a penalized maximum likelihood method– called the "adaptive ridge" – which performs model selection. The latter comes from a long line of historically, practically, and theoretically important regularization methods, which date back to the very beginning of computational statistics. These methods are at the intersection of different fields: statistics, optimization, and computer science.

It is therefore necessary to introduce the most important of these methods in order to *(i)* highlight the principle of penalized estimation methods, *(ii)* put the adaptive ridge in the context of the other model selection methods, and *(iii)* compare the adaptive ridge with the competing methods. This introduction serves this purpose.

Variable selection methods have been developed since the 1970's, with the use of the stepwise selection and the ridge regression in the context of linear regression. But their use and fame have exploded only in the 1990's and early 2000's, with the development of penalized regression method, and in the first place of the lasso, introduced in the field of statistics by Tibshirani (1996) and in signal processing by Chen et al. (2001) under the name "basis pursuit". These methods have been introduced for linear regression, but their principle applies to any statistical model whose likelihood is computable and practical to maximize. Penalized regression methods consist in adding a term in the negative log-likelihood (NLL) to minimize. This term enforces the estimate to be close to an *a priori* shape or distribution. Certain penalty terms have been found to induce the desired properties of the estimate: the resulting estimate is both better quantitatively, i.e. it has better estimation performance, and qualitatively, i.e. it infers models which are more relevant and easier to interpret, than standard estimates. These methods sparked a revolution in the field of computational statistics: the added penalty term increases the complexity of the computation by a low margin and the benefits are huge in many practical applications.

A whole array of these penalized methods have then been developed and improved upon, with penalties ever more refined and adapted to the problem at hand. However, the initial application of penalized estimation with variable selection is the linear model, and its extension to non-Gaussian errors, the generalized linear model. It seemed that in this case, the goal to find easy- and fast-to-compute methods that are performant both theoretically and practically, has been met. Two other major fields of applications of penalized methods have emerged around this period: high throughput data and wavelet analysis. The former became widely popular in the last decade due to the onset of next-generation sequencing technologies. Since many genes are studied at once, we are in a case where $p \gg n$, usually by a factor ~ 100. The problem posed by this *high-dimensional* setting was seldom met in usual regression settings, and led to the development of penalized regression methods for what is called *high-dimensional statistics*. The latter comes from the development of wavelets bases and their applications to all fields of signal processing. Wavelets are families of functions that form an orthogonal family of $L^2(\mathbb{R})$, are located in both the time and frequency domains, and are all the scaled and shifted versions of one another. These properties make it the tool of choice for

representing signals (i.e. audio signals, images, videos, etc.) efficiently. Indeed, the wavelet bases allow for sparse representation of signals, with major applications to denoising, compression, and compressed sensing. In these problems, the signal sparsity is enforced using a penalized approach. This has sparked the development of penalized likelihood approaches that are specifically fit to the topology of the signal. Note that the leap forward made in that domain around the year 2010 was also enabled by the progress in convex optimization (Boyd and Vandenberghe, 2004). We refer to Mallat (2009) for more details on the wavelet analysis and to Bach (2011) for more details on the application of penalized methods to wavelet analysis.

This introduction is organized as follows. The first part draws a panorama of the main penalized methods for model selection, with comparisons, explanations, and insights. The second part deals with subset selection, the main model selection method that does not used a penalized approach. The third part puts the iterative penalized methods in the theoretical framework of Majorize-Minimization (MM) optimization and compares the main iterative penalized methods amongst which is the adaptive ridge. The last section presents modifications of the penalty terms which enforces the estimate to have a certain sparsity structure, which is more general than being sparse. We arrive at the fused adaptive ridge, which is used in Chapters 2 and 3.

Contents

1.1 Regularized estimation . . . 4
1.1.1 Ridge regression . . . 5
1.1.2 Lasso estimation . . . 7
1.1.3 Elastic-net . . . 12
1.1.4 Bridge regression . . . 12
1.1.5 Berhu . . . 13
1.1.6 Penalized likelihood methods as a Bayesian prior . . . 13
1.1.7 Non-concave penalties . . . 15
1.1.7.1 Hard-thresholding . . . 17
1.1.7.2 SCAD . . . 18
1.1.7.3 Logarithmic penalty . . . 18
1.1.7.4 Available implementations . . . 19
1.2 Subset selection . . . 19
1.2.1 Best subset selection . . . 19
1.2.2 Stepwise Selection . . . 20
1.3 Iterative penalized methods . . . 21
1.3.1 MM optimization . . . 21
1.3.2 MM optimization applied to non-concave penalties . . . 22
1.3.2.1 Local Quadratic Approximation . . . 23
1.3.2.2 Local Linear Approximation . . . 24
1.3.3 Adaptive Ridge . . . 25
1.3.3.1 Relation to similar procedures . . . 27
1.3.3.2 Numerical performance . . . 28
1.4 Structured variable selection . . . 28
1.4.1 Group lasso . . . 28
1.4.2 Overlapping groups . . . 30
1.4.3 Hierarchical structured sparsity . . . 31
1.4.4 Fused lasso and total variation . . . 32
1.4.5 Fused Adaptive Ridge . . . 34
1.5 Conclusion . . . 35

1.1 Regularized estimation

Linear regression. Consider the linear regression setting

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{y}$ is the $n \times 1$-sized response variable, $\boldsymbol{X} = \left(x_{i,j}\right)$ is the $n \times p$-sized design matrix, $\boldsymbol{\beta}$ is the $p \times 1$ parameter vector of linear effects, and $\boldsymbol{\varepsilon}$ is the $p \times 1$ vector of random errors. The columns of $\boldsymbol{X}$ are called the covariates and are noted $\boldsymbol{x}_j$. We consider the case of deterministic design, that is, $\boldsymbol{X}$ is deterministic.

Linear regression is sometimes written

$$\hat{\boldsymbol{y}} = \hat{\beta}_0 + \boldsymbol{X}_1\hat{\beta}_1 + \cdots + \boldsymbol{X}_p\hat{\beta}_p,$$

which includes the estimation of an intercept β_0, adding a row of 1s in the design matrix $\boldsymbol{X}$. Since the maximum likelihood estimate of β_0 is $\bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^n y_i$, we will instead consider the model (1.1) and always assume that the response variable is centered: $\bar{\boldsymbol{y}} = 0$.

It is also necessary that the covariates be scaled, both for the numerical stability of the computations and to be able to compare the effects of each covariate. For instance if the unit of $\boldsymbol{x}_j$ is changed from meters to millimeters, $\hat{\beta}_j$ is multiplied by 1000. Hence the effect of that covariate will be artificially inflated and it will always be selected by a model selection method. Thus, the covariates have to be scaled for their effects to be comparable. Note that there is an equivalence between fitting the parameters on the unscaled covariates and fitting them on the scaled covariates and applying the corresponding scaling to each parameter. Therefore, without loss of generality, we will assume that

$$\sum_{i=1}^n x_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^n x_{ij}^2 = 1, \quad 1 \leq j \leq p.$$

The ordinary least squares (OLS) is

$$\hat{\boldsymbol{\beta}}^{\text{ols}} \triangleq \underset{\boldsymbol{\beta}}{\arg\min} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{1.2}$$

If not stated otherwise, we assume the two classical following assumptions to hold.

Assumption 1. *$y_i = \boldsymbol{x}_i\boldsymbol{\beta}^* + \varepsilon_i$, where ε_i are independent and identically distributed random variables of mean 0 and variance σ^2, and $\boldsymbol{\beta}^*$ is the real value of $\boldsymbol{\beta}$ that we estimate. Note that unless stated otherwise, the errors are not assumed to be normally distributed.*

Assumption 2. *The matrix $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \to \boldsymbol{C}$, where $\boldsymbol{C}$ is positive definite.*

Penalized likelihood. In the forthcoming sections we will discuss penalized regression methods, in which the estimator is defined as the minimizer of the least squares residuals with an added penalty term. In the linear model, when the residuals are normally distributed, the least squares residuals $\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|_2^2$ is (proportional to) the negative log-likelihood $\ell\left(\boldsymbol{\beta}\right)$ of the model (this condition is in fact an equivalence: the negative log-likelihood writes as a sum of squares only when $\varepsilon \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$). The regularization – and variable selection – methods present in this introduction have been developed in the setting of linear regression, with the case of normal residuals as an important specific case. In all generality, they can apply to any parametric model that we want to regularize. In this introduction, the notation $\ell\left(\boldsymbol{\beta}\right)$ refers to the negative log-likelihood in the linear model or in any unspecified parametric model. In applications and illustrations however, we will mostly use the linear model to illustrate the effect of penalized estimation.

We give two examples of such models, which are generalizations of the linear model:

- Robust regression, in which the negative log-likelihood writes

$$\ell\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_i\boldsymbol{\beta}) \tag{1.3}$$

  where ρ is a function giving little importance to large values of the residuals. For instance, when $\rho(x) = |x|$, the model becomes the least absolute deviation, a robust alternative to the linear regression. When ρ is a ρ-function (as defined in Huber et al., 1964), the model is a robust version of the linear model. See Maronna et al. (2006) for a thorough explanation of robust regression.

- The generalized linear model the response variable has a distribution in the exponential family, where the density of $\boldsymbol{y_i}$ (in its canonical form, see McCullagh, 1984, Section 2.2.2) writes

$$f(y) = \exp\left(\theta_i^* T(y) - A(\theta_i^*)) + B(y)\right),$$

  where A and B are known deterministic functions and θ_i^* is the true unknown parameter to be estimated. In the generalized linear model, we estimate the density's parameter vector $\boldsymbol{\theta}^*$ by a function of the linear effect $\boldsymbol{x}_i\boldsymbol{\beta}$, and we have $\mathbb{E}[\boldsymbol{Y}] = g^{-1}\left(\boldsymbol{X}\boldsymbol{\beta}\right)$ where g is a link function. Usually we take the canonical link function given by the distribution of $\boldsymbol{y}$: $g^{-1} = A$. The negative log-likelihood (nll) writes

$$\ell\left(\boldsymbol{\beta}\right) = -\sum_{i=1}^{n} \log f\left(g\left(\boldsymbol{x}_i\boldsymbol{\beta}\right), y_i\right), \tag{1.4}$$

  where f is the distribution of $\boldsymbol{y}$. For example, if $\boldsymbol{y}$ is Poisson distributed, the canonical link function is the log function and the nll writes

$$-\sum_{i=1}^{n} \left\{y_i\boldsymbol{x}_i\boldsymbol{\beta} - \exp(\boldsymbol{x}_i\boldsymbol{\beta}) - \log(y_i!)\right\}.$$

1.1.1 Ridge regression

The OLS estimate is unbiased and has minimal variance amongst all unbiased estimates. However, when the covariates $\boldsymbol{x}_j$ are highly correlated, $\boldsymbol{X}^T\boldsymbol{X}$ becomes close to singular, i.e. when its largest eigenvalue gets close to zero, the variance of the OLS diverges to infinity. In this case, a biased modification of the OLS estimate is necessary to control the variance of the estimate. The ridge regression, defined below, is the most famous of such modifications – owing first to its simplicity. It was introduced in the context of linear regression and we use this specific case to illustrate its principle and effect, but it extends naturally to any other model. It is obtained by adding a regularizing term to the nll, in the way defined below.

Hoerl and Kennard (1970) introduced the ridge estimator. It is defined as the solution to

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2, \tag{1.5}$$

where $\lambda > 0$ is a trade-off hyper-parameter to be chosen.

Problem (1.5) has the explicit solution

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{1.6}$$

Equation 1.6 can be seen as a relaxation of the matrix $\boldsymbol{X}^T\boldsymbol{X}$. When the covariates are too correlated, the columns of $\boldsymbol{X}$ are close to being linearly dependent, and $\boldsymbol{X}^T\boldsymbol{X}$ tends to be singular. Equivalently, this means that the lowest eigenvalue of $\boldsymbol{X}^T\boldsymbol{X}$ tends to zero. If λ_1 is the lowest eigenvalue of $\boldsymbol{X}^T\boldsymbol{X}$, the lowest eigenvalue of $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}$ is $\lambda_1 + \lambda$. Equation 1.6 can be seen as a modification of the OLS that ensures the design matrix does not become singular.

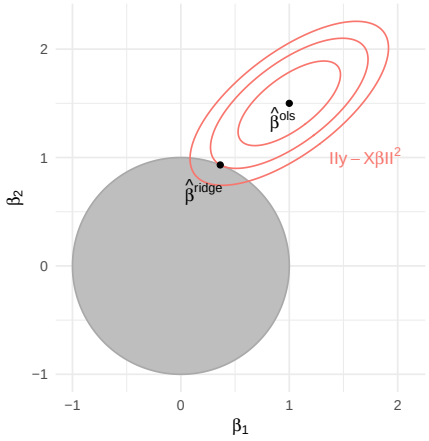


Figure 1.1: Visualization of the ridge estimate as the projection of the OLS onto an L_2 norm ball. The ellipses are level curves of the quadratic form $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2$. The projection onto the circle of radius $t = 1$ has the effect of shrinking the estimate's coordinates together.

Projection on an L_2 ball. Problem (1.5) is the Lagrangian dual of the problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_2^2 \leq t \tag{1.7}$$

for $t > 0$. The two problems have been shown (Boyd and Vandenberghe, 2004, B.1) to have strong duality, that is, for every $\lambda > 0$, there exists a $t > 0$ such that the solution of (1.7) is also the solution of (1.5), and conversely. The decreasing one-to-one relation between λ and t depends on the data.

The equivalent constrained optimization problem helps understand the effect of the penalty included in ridge regression. Figure 1.1 illustrates the shrinkage of $\hat{\boldsymbol{\beta}}$ due to the projection onto the L_2 norm ball. We recall that in a normed vector space, for a chosen norm, the ball of radius $r > 0$ and of center c is the set of elements at a distance r or less to c, i.e. the set $\{x | \|x-c\| \leq r\}$. In this manuscript, we will only consider centered balls, that is, balls of center 0. Balls of radius 1 are called unit balls.

Ridge regression in orthogonal design. Consider the case of *orthogonal design*, that is, when the design matrix is orthogonal: $\boldsymbol{X}^T\boldsymbol{X} = I$. Notice that in this case, the quadratic terms $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{\beta}$ and $\|\boldsymbol{X}^T\boldsymbol{y}-\boldsymbol{\beta}\|^2 = \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{\beta}$ differ only by a constant. Then the linear regression boils down to minimizing $\|\boldsymbol{z}-\boldsymbol{\beta}\|^2$, where $\boldsymbol{z} = \boldsymbol{X}^T\boldsymbol{y}$ is the transformed data. The ridge problem (1.5) rewrites

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{z}-\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|.$$

This particular case is important because it is the case in which the covariates are uncorrelated, and the estimation is done component by component. Also, the last equation can be interpreted as finding the closest vector to the data $\boldsymbol{z}$ with a small L_2 norm. This result can also be seen directly through the simplifications in (1.2) and (1.6).

In this case, then, we have $\hat{\boldsymbol{\beta}}^{\text{ols}} = \boldsymbol{X}^T\boldsymbol{y}$ and

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \frac{\hat{\boldsymbol{\beta}}^{\text{ols}}}{1+\lambda}.$$

Consequently, in orthogonal design, the ridge estimate simply shrinks every coordinate of the OLS towards $\mathbf{0}$ by the same factor.

Minimization of the MSE. Define the mean square error as

$$\text{MSE}_{\text{ols}} = \mathbb{E}\left[\|\hat{\boldsymbol{\beta}}^{\text{ols}} - \boldsymbol{\beta}^*\|_2^2\right]$$

Define $\lambda_1 \geq \cdots \geq \lambda_p > 0$ as the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$. Recall that the OLS is unbiased and from the equality $\hat{\boldsymbol{\beta}}^{\text{ols}} - \boldsymbol{\beta}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}$, we obtain

$$\text{MSE}_{\text{ols}} = \sigma^2 \text{Tr}\{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\} = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}. \tag{1.8}$$

This MSE takes problematically large values when λ_1 is small.

The ridge estimate aims at reducing the MSE to a value lower than (1.8). The ridge estimate reduces the variance of the estimate, at the price of an non-zero bias. More precisely, the MSE of the ridge estimate decomposes as follows. Consider the *singular value decomposition* of $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T, \tag{1.9}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are $n \times p$ and $p \times p$ orthogonal matrices respectively, such that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_p$, and $\boldsymbol{D}$ is a diagonal matrix whose entries $d_1 \geq \cdots \geq d_p \geq 0$ are the *singular values* of $\boldsymbol{X}$. From the equality $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T$, we can see that the d_j^2s are the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$. By plugging (1.9) into (1.6), we get

$$\begin{aligned}
\text{MSE}(\lambda) &= \mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2] \\
&= \lambda^2 \boldsymbol{\beta}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-2}\boldsymbol{\beta}^* + \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-2}\boldsymbol{X}^T\boldsymbol{\varepsilon}\right] \\
&= \lambda^2 \boldsymbol{\beta}^{*T}\boldsymbol{V}\boldsymbol{D}^{-4}\boldsymbol{V}^T\boldsymbol{\beta}^* + \sigma^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2 + \lambda)^2}.
\end{aligned} \tag{1.10}$$

In the last equation, the first term is the bias of the ridge estimate, the second term is its variance. Since

$$\frac{d\text{MSE}}{d\lambda}(\lambda)\bigg|_{\lambda=0^+} < 0,$$

there exists a value of λ such that the ridge estimate has a smaller MSE than the OLS.

1.1.2 Lasso estimation

Why select variables. In many practical situations, the statistician does not know what covariates has an influence on the response variable or does not want to make such an assumption *a priori*. A solution is to add *all* possible variables and find a regression procedure which simultaneously *selects* the relevant covariates and estimates their effect. This is the framework of regression with variable selection.

Lasso as a relaxation of the L_0 norm. Define the L_0 "norm" $\|x\|_0 = \#\{i | x_i \neq 0\}$ as the number of non-zero elements of a vector. This quantity is not a norm, because for $\alpha \neq 0$, $\|\alpha\boldsymbol{x}\|_0 \neq |\alpha|\|\boldsymbol{x}\|$. We still refer to it as a norm since it can be seen as the limit of the L_q norm when $q \to 0$.

The L_0 penalized likelihood approach offers to minimize

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_0. \tag{1.11}$$

The minimizer of (1.11) represents a *trade-off*: it has both to be close to the minimum of the squared residuals and to have few non-zero coordinates. Consequently this estimate will only select the most relevant covariates in the model and estimate their effect.

Unfortunately, to minimize the L_0 norm is NP-hard (Natarajan, 1995), which implies that there is no known algorithm that can solve an L_0 norm-constrained problem in a polynomial time complexity with respect to p. Consequently, this problem has to be solved by trying all the possibilities, which has exponential time complexity in p. This is unfeasible in practice for $p \simeq 50$ or larger.

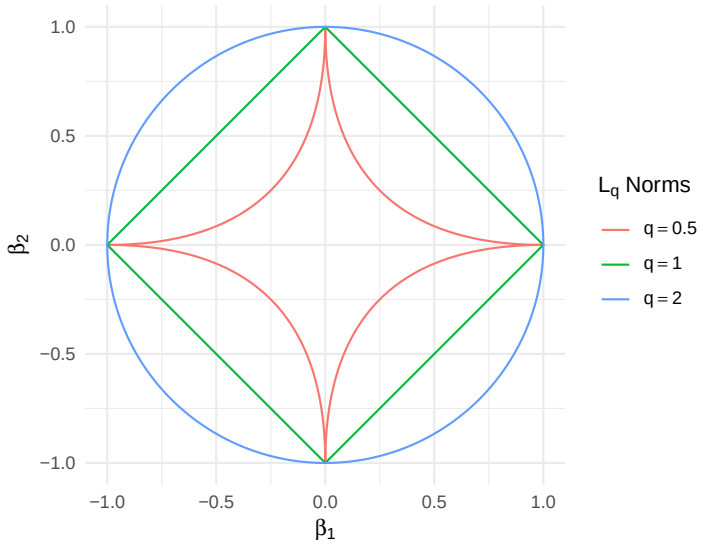


Figure 1.2: Illustration of L_q unit balls for different values of q.

Tibshirani (1996) introduced the lasso estimator defined by

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} \triangleq \underset{\boldsymbol{\beta}}{\arg\min} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \tag{1.12}$$

where $\lambda > 0$ is a hyper-parameter which tunes the trade-off between goodness of fit and regularization. The choice of λ is not touched on in this chapter. Lasso uses the L_1 norm which can be seen as a relaxation of the L_0 norm.

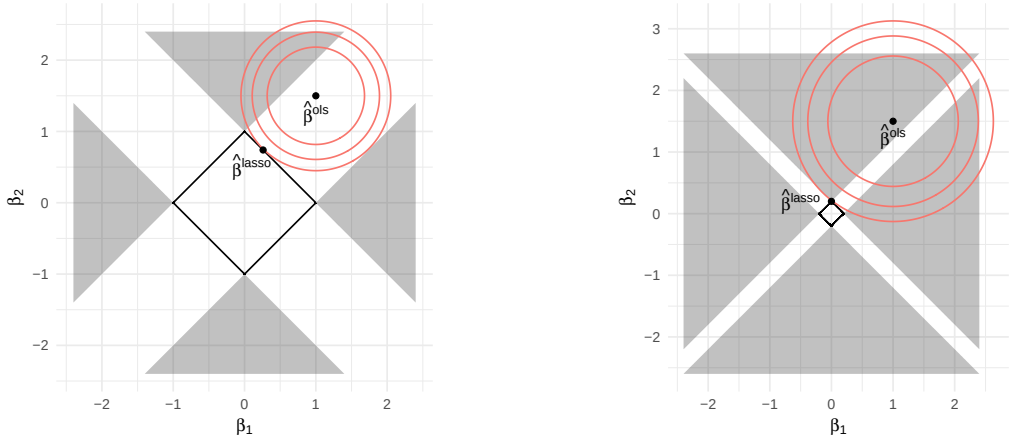
The L_1 norm has been used for many decades to recover sparse vectors. In signal processing, a signal $\boldsymbol{y}$ is expressed as a combination of functions forming a dictionary $\boldsymbol{\phi}$. Since $\boldsymbol{\phi}$ is over-complete, it has more columns than rows, and the coefficient vector $\boldsymbol{x}$ of the decomposition of the signal in the dictionary is not uniquely determined by the problem $\boldsymbol{y} = \boldsymbol{\phi}\boldsymbol{x}$. However one assumes that $\boldsymbol{y}$ has a sparse representation in the dictionary. By enforcing a sparsity constraint over $\boldsymbol{x}$, a compressed representation of the signal $\boldsymbol{y}$ is obtained. Using the L_1 penalized problem to recover a signal is called *basis pursuit*. This method was developed by Chen et al. (2001); it was further analyzed by Donoho and Huo (2001) and Elad and Bruckstein (2001).

Using the Lagrange multipliers, the lasso estimator can be defined as the projection of the OLS onto the L_1 ball of radius t:

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\boldsymbol{\beta}}{\arg\min} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq t, \tag{1.13}$$

where $t > 0$ has a decreasing one-to-one relation with λ. Figure 1.2, which represents L_q unit balls for different values of q, illustrates why the L_1 norm induces sparsity. When $q > 1$, the L_q ball is smooth and the projection onto the ball does not set any coordinate to zero. When $q \leq 1$, the L_q ball has singularities at the axes and the projection onto the ball can set some coordinates to zero. Moreover for $q \geq 1$, L_q norms are convex, which makes computational minimization way easier. The L_1 norm naturally comes out as the only *easily* minimizable penalty which enables variable selection, amongst all L_q norm penalties.

L_1 norm and sparsity. As previously explained, the projection onto an L_1 ball can set some coordinates to zero. This phenomenon is illustrated in Figure 1.3 with $p = 2$. The grey areas represent the half-cones from which a projection on the L_1 ball sets one coordinate to zero. Consequently, when $\hat{\boldsymbol{\beta}}^{\text{ols}}$ has a coordinate of small value, its projection maps it to zero. The lasso has the effect of selecting the covariates with a non-negligible effect on $\boldsymbol{y}$. As illustrated in the figure, as λ increases the



(a) Small penalty (b) Large penalty

Figure 1.3: Illustration of the lasso estimation under orthogonal design. Projection onto two L_1 norm balls of radius t. As t decreases, the OLS will most probably end up in a greyed area and the lasso estimate will become sparse.

radius of the ball decreases and the grey area takes up a larger proportion of the parameter space. Consequently, the lasso will tend to set many covariates to zero as the penalty increases.

Lasso in orthogonal design: the soft-thresholding operator. In orthogonal design (see previous Section), the lasso problem reduces to the component-wise problem $\hat{\beta}_j^{\text{lasso}} = \text{argmin}_{\beta_j} f(\beta_j)$ where $f(\beta_j) = \frac{1}{2}(\hat{\beta}_j^{\text{ols}} - \beta_j)^2 + \lambda|\beta_j|$ and where $\hat{\boldsymbol{\beta}}^{\text{ols}} = \boldsymbol{X}^T\boldsymbol{y}$ is the OLS estimate in orthogonal design. Since f is the sum of two functions, who minima are in $\hat{\beta}_j^{\text{ols}}$ and 0 respectively, its minimum is located either at 0 or at β_j^{ols}. By considering the two cases $|\beta_j^{\text{ols}}| \leq \lambda$ and $|\beta_j^{\text{ols}}| \leq \lambda$ separately, we get that the lasso solution has the explicit solution given by

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = S(\hat{\boldsymbol{\beta}}^{\text{ols}}) \triangleq \text{sgn}(\hat{\boldsymbol{\beta}}^{\text{ols}}) \left(\left|\hat{\boldsymbol{\beta}}^{\text{ols}}\right| - \lambda\right)_+, \tag{1.14}$$

where the function S is called the *soft-thresholding* function, and where the latter formula is to be considered component by component. This function is equal to zero over the interval $[-\lambda, \lambda]$, which means that in orthogonal design, the lasso sets small value of the data to zero.

We refer to Section 1.1.7 for a graphical representation of the soft-thresholding operator and a discussion of its properties.

Computation of the lasso solution. Nevertheless, since the L_1 norm is not differentiable at the minimum, the computation of the lasso estimation is not straightforward. Two algorithms for solving the lasso have gained popularity since its development.

The Least Angle Regression (LARS) algorithm has been proposed by Efron et al. (2004) and offers to compute the whole *regularization path*[1]. It initiates with no covariates in the model and sequentially adds the covariate that has the maximum correlation with the residuals. The estimate is increased linearly in the direction of equal correlation between all active covariates (hence the names *least angle*). Once a new covariate has the same correlation with the residuals as the current covariate, it is added to the set of active covariates and becomes the current covariate. This method is explained in more details in the Appendix.

[1]The regularization path of a regularized model is the path of solutions $\hat{\boldsymbol{\beta}}(\lambda)$ represented as a function of λ. This path illustrates the progressive effect of λ on the fitted parameters and is often obtained over a grid of values of λ

Another approach is the coordinate descent, a simple optimization method that is particularly fit to solving the lasso. This method is explained here.

The coordinate descent was initially proposed for the lasso by Fu (1998) under the name of *shooting algorithm*. It was generalized to other related penalties by Friedman et al. (2007). The method solves the lasso problem for a fixed λ.

The algorithm starts with an initial guess $\hat{\boldsymbol{\beta}}^{(0)}$ and minimizes the penalized residuals with respect to each component in a fixed cyclical order until convergence. For $1 \leq k \leq p$, we denote by β_k the k-th component of $\boldsymbol{\beta}$. The minimization of (1.12) with respect to β_k writes

$$\min_{\beta_k} \sum_{i=1}^{n} \left(\tilde{y}_i^{(k)} - x_{i,k}\beta_k \right)^2 + \lambda |\beta_k| \tag{1.15}$$

where $\tilde{y}_i^{(k)} \triangleq y_i - \sum_{j \neq k} x_{ij}\beta_j$. The solution to this problem is simply the soft-thresholding operator applied to the modified output:

$$\beta_k^* = \left(\left| \sum_{i=1}^{n} \tilde{y}_i^{(k)} x_{ik} \right| - \lambda \right)_+ \operatorname{sgn}\left(\sum_{i=1}^{n} \tilde{y}_i^{(k)} x_{ik} \right) \tag{1.16}$$

Hence the coordinate descent algorithm for the lasso is simple, easy to implement and fast to compute. However, contrarily to LARS, it does not compute the solution path for all penalties; we fix a grid of penalties and compute the coordinate descent for each penalty. To speed up the computation, the initial value in the coordinate descent is taken as the minimum found for the previous penalty. This computational trick is often referred to as *warm start*. The coordinate descent procedure with one penalty is given in Algorithm 1.

Algorithm 1 Coordinate descent with one penalty

function COORDINATE-DESCENT($\boldsymbol{y}, \boldsymbol{X}, \lambda$)
  $\hat{\boldsymbol{\beta}} \leftarrow \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}\boldsymbol{y}$
  $k \leftarrow 1$
  while not converge **do**
    $\tilde{y}_i^{(k)} \leftarrow y_i - \sum_{j \neq k} x_{ij}\hat{\beta}_j$
    $\hat{\beta}_k \leftarrow S\left(\sum_{i=1}^{n} \tilde{y}_i^{(k)} x_{ik}, \lambda\right)$
    $k \leftarrow \left(k \bmod p\right) + 1$
  end while
end function

In the lasso regression, the function to minimize is the sum of the residuals, which is convex differentiable, and the L_1 norm, which writes as the sum of univariate convex functions. The rational behind the use of coordinate descent to solve the lasso problem is the following property.

Property 1. *Consider a function $f : \mathbb{R}^p \to \mathbb{R}$ which writes $f(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + \sum_{j=1}^{p} h_j(\beta_j)$ where g is convex differentiable and the $h_i, 1 \leq i \leq p$ are convex. Assume that $\boldsymbol{\beta}^*$ is a minimum of f along all coordinates, that is: $\forall i \in \{1, \ldots, p\}, \forall a \in \mathbb{R}, f(\boldsymbol{\beta}^* + a\boldsymbol{e}_i) \geq f(\boldsymbol{\beta}^*)$ where $\boldsymbol{e}_j$ is the j^{th} basis vector. Then $\boldsymbol{\beta}^*$ is a global minimum.*

Proof. Let $\boldsymbol{\beta}^*$ be a minimum of f along all coordinates. For any $\boldsymbol{\beta}$:

$$\begin{aligned} f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*) &= g(\boldsymbol{\beta}) - g(\boldsymbol{\beta}^*) + \sum_{j=1}^{p} \left(h_j(\beta_j) - h_j\left(\beta_j^*\right) \right) \\ &\geq \sum_{j=1}^{p} \left(\frac{\partial g}{\partial \beta_j}(\boldsymbol{\beta}^*) \left(\beta_j - \beta_j^*\right) + h_j(\beta_j) - h_j\left(\beta_j^*\right) \right) \end{aligned}$$

The inequality comes from the convexity and differentiability of g. Since f is maximal along all coordinates, every term in the sum is positive. This completes the proof. □

This property gives an incentive to use the coordinate descent when the function to minimize has the aforementioned form. This is the case for the lasso penalized regression as well as for any lasso problem where the likelihood of the model is log-concave. This property alone does not guarantee the coordinate descent converges to the global optimum. When g is strictly convex, the coordinate descent has been proven to converge (Tseng, 1988).

Theoretical properties of the lasso. In this section we consider the asymptotic performances of the lasso. Since this selection method performs selection and estimation, we need to define asymptotic properties for both. The definitions and properties present in this section are found in Zou (2006).

Let $\mathscr{A}_n = \{j : \hat{\beta}_j \neq 0\}$ be the selected covariates and $\mathscr{A}^* = \{j : \beta_j^* \neq 0\}$ the true non-zero covariates. The cardinal of $\mathscr{A}^*$, i.e. the number of true non-zero covariates, is noted p_0.

Definition 1. *An estimate $\hat{\boldsymbol{\beta}}$ is said to be consistent in selection if* $\lim_n \mathbb{P}(\mathscr{A}_n = \mathscr{A}^*) = 1$.

For simplicity, we assume without loss of generality that $\mathscr{A}^* = \{1, \dots, p_0\}$. Let

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix}, \tag{1.17}$$

where $\boldsymbol{C}_{11}$ is a $p_0 \times p_0$ matrix. Let $\hat{\boldsymbol{\beta}}_{\mathscr{A}^*}$ (resp. $\boldsymbol{\beta}^*_{\mathscr{A}^*}$) be the first p_0 elements of the estimate $\hat{\boldsymbol{\beta}}$ (resp. $\boldsymbol{\beta}^*$). If an estimate is consistent in selection, the support $\mathscr{A}^*$ will be known with large probability for n large enough. The question arises of the estimation quality of $\hat{\boldsymbol{\beta}}_{\mathscr{A}^*}$.

Definition 2. *The estimate $\hat{\boldsymbol{\beta}}$ is said to be v_n-consistent in estimation if* $v_n(\hat{\boldsymbol{\beta}}_{\mathscr{A}^*} - \boldsymbol{\beta}^*_{\mathscr{A}^*}) \to_d \mathscr{N}(\mathbf{0}, \boldsymbol{C}_{11})$.

Note that $\boldsymbol{C}_{11}$ is the variance matrix of the estimate *knowing the right model*, i.e. knowing $\mathscr{A}^*$. An estimate which is both consistent in selection and $\sqrt{n}$-consistent in estimation is said to have the *oracle properties*: it performs as well as the maximum likelihood estimate would if we knew the true covariate indices. Consequently, the oracle properties is the best we can ask for in a variable selection method, in terms of prediction accuracy. This means that when an estimate has the oracle properties, it is always more advantageous to use it instead of the OLS.

These considerations have to be moderated by the fact that *(i)* in practice n is not always very large *(ii)* the conditions about λ are related to its asymptotic speed.

Zou (2006) proved the following property.

Property 2. *The lasso estimate is not necessarily consistent in selection. More precisely, if $\lambda_n = \mathscr{O}(\sqrt{n})$, $\limsup_n \mathbb{P}(\mathscr{A}_n = \mathscr{A}) \leq c < 1$, where the constant c depends on the true model.*

A small modification of the lasso has been proven to enjoy the oracle properties. This estimate is presented hereafter.

The adaptive lasso. Zou (2006) introduced a slight modification of the lasso with a weighted L_1 norm:

$$\hat{\boldsymbol{\beta}}^{\mathrm{al}} \triangleq \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^{p} w_j |\beta_j|. \tag{1.18}$$

The weights are defined as $w_j = 1/|\hat{\beta}_j|^\gamma$, where $\hat{\boldsymbol{\beta}}$ is any consistent estimate of $\boldsymbol{\beta}$ (for instance the OLS) and where $\gamma > 0$ is a hyper-parameter to be chosen. The weights have the effect to penalize more coordinates which have a small estimate. Due to this rescaling, the adaptive lasso is proven to have the oracle properties when $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$. Moreover, the adaptive lasso estimate has the same computational cost as the lasso.

1.1.3 Elastic-net

The elastic-net penalty was introduced by Zou and Hastie (2005):

$$\lambda\left(\alpha\|\boldsymbol{\beta}\|_2^2+(1-\alpha)\|\boldsymbol{\beta}\|_1\right) \tag{1.19}$$

where $\alpha\in[0,1]$ is an hyper-parameter to be fixed. The elastic-net penalty defines a norm that is a compromise between the L_1 and the L_2 norm: as α varies, the elastic-net estimate varies continuously between the lasso and the ridge estimates. For $\alpha=0$ it becomes the lasso penalty and for $\alpha=1$ it becomes the ridge penalty.

Contrarily to the L_q norm ($1\leq q\leq 2$), the elastic-net penalty is non-differentiable, as illustrated in Figure 1.4. Consequently, the elastic-net always performs selection (for $\alpha\neq 1$) like the lasso and shrinkage like the ridge (which reduces the MSE). Comparatively, the L_q norm ($1<q<2$) does not perform model selection and is of little interest. As discussed hereafter, the elastic-net enjoys the desirable properties from both estimates.

The thresholding function of the elastic-net under orthogonal design is

$$\hat{\beta}_i^{\mathrm{en}}=\frac{(|\hat{\beta}_i^{\mathrm{ols}}|-\lambda_1/2)_+}{1+\lambda_2}\mathrm{sgn}\{\hat{\beta}_i^{\mathrm{ols}}\}. \tag{1.20}$$

This thresholding function has the same behavior as the soft thresholding for small values of $\hat{\beta}_i^{\mathrm{ols}}$ and the same behavior as the ridge thresholding for large values of $\hat{\beta}_i^{\mathrm{ols}}$. An illustration is given in Figure 1.7 with $\lambda_1=2$ and $\lambda_2=1$.

The elastic net is also more stable in selection than the lasso. More precisely, if several covariates are very correlated, the lasso will arbitrarily select one covariate and discard the others. This is proven (Zou and Hastie, 2005) not to be the case for the elastic-net: the difference $\left|\beta_i-\beta_j\right|$ is bounded by $1-\mathrm{corr}\{\boldsymbol{x}_i,\boldsymbol{x}_j\}$, so that highly correlated covariate effects have close to equal estimates. This property is known as the *grouping effect*: the elastic net selects important covariates in group.

Moreover, in the $p\gg n$ scenario, lasso does not perform well in selection because it cannot select more than n covariates. The elastic-net overcomes this obstacles and, even with small α, can select any number of covariates.

Finally, the elastic-net estimate writes also as the minimizer of

$$\boldsymbol{\beta}^T\left(\frac{\boldsymbol{X}^T\boldsymbol{X}+\lambda_2\boldsymbol{I}}{1+\lambda_2}\right)\boldsymbol{\beta}-2\boldsymbol{X}^T\boldsymbol{y}+\lambda_1\|\boldsymbol{\beta}\|_1, \tag{1.21}$$

hence it is computed as a lasso estimate with a modified design matrix.

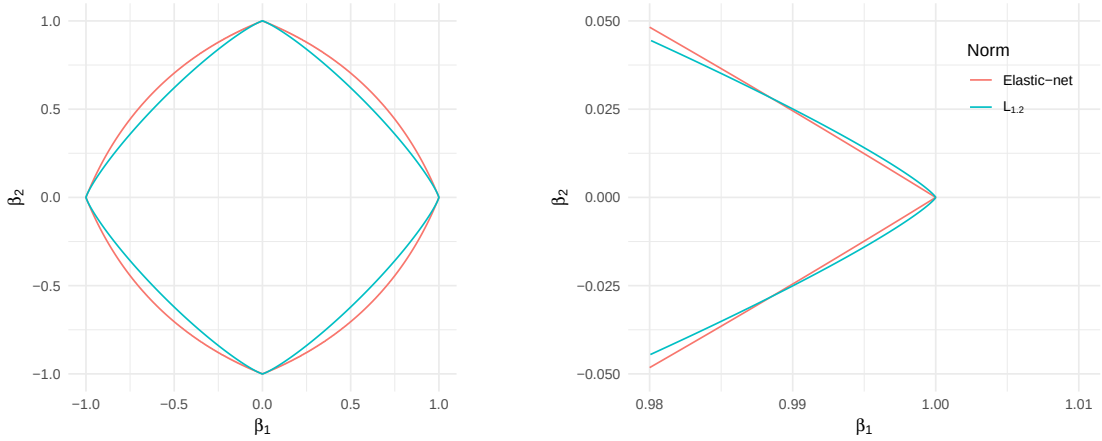
The LARS algorithm can be adapted to estimate the elastic-net. Recall that LARS computes the whole regularization path. Thus the LARS algorithm is called on a grid of values of λ_2 and the selected value is the one which minimizes the (ten-fold) cross-validation error. The computational cost of this procedure is K times more than that of a lasso fit, where K is the length of the grid of λ_2.

1.1.4 Bridge regression

Bridge regression is based on the L_q penalty:

$$\hat{\boldsymbol{\beta}}^{\mathrm{bridge}}\triangleq\arg\min_{\boldsymbol{\beta}}\ell\left(\boldsymbol{\beta}\right)+\lambda\sum_{j=1}^{p}\left|\beta_j\right|^{\gamma} \tag{1.22}$$

where $\gamma>0$ is a parameter to be fixed. Bridge regression was first introduced by Frank and Friedman (1993) as a generalization of the ridge ($\gamma=2$) and the lasso ($\gamma=1$) and was later studied by Fu (1998), who compared it with the lasso. For $0<q\leq 1$ the L_q penalty is singular around zero and the bridge regression performs variable selection. This property is illustrated in Figure 1.2 which represents unit balls of L_q norms for different values of q.



(a) Unit radius balls

(b) Zoom around $(1,0)$

Figure 1.4: Illustration of why the elastic net induces sparsity and not the L_q norm ($1 < q < 2$). Unit radius balls of the elastic-net norm ($\alpha = 0.2$) and the $L_{1.2}$ norm.

Asymptotic properties of the bridge regression has been studied by Knight and Fu (2000) in both cases $\gamma \leq 1$ and $\gamma > 1$.

Fu (1998) proposed an algorithm based on the Newton-Raphson procedure to compute the bridge estimate when $\gamma > 1$. However, when $0 < \gamma < 1$, the penalty is not convex and its minimization is computationally challenging.

The bridge penalty is part of a general family of sparsity inducing penalties called *non-concave* penalties These are defined in the next section. The oracle properties of the bridge estimate and the algorithms available to compute it are given in this section.

1.1.5 Berhu

The Berhu penalty was introduced by Owen (2006):

$$p_\lambda(|\theta|) = \begin{cases} |\theta| & |\theta| \leq \lambda \\ \frac{\theta^2+\lambda^2}{2\lambda} & |\theta| > \lambda \end{cases} \tag{1.23}$$

It is equal to the absolute value around zero and is quadratic away from zero. Consequently it is sometimes referred to as *inverse Huber* function (Huber et al., 1964).

The presence of the singularity of the penalty at zero makes it a variable selection method. The Berhu penalty replaces the absolute value penalty by a quadratic penalty for large values in order to benefit from the properties of the ridge regression: the grouping effect, and an increased accuracy in estimation. The Berhu penalty can be seen as mixing the L_1 and L_2 penalties; in this regards it is very similar to the elastic-net.

1.1.6 Penalized likelihood methods as a Bayesian prior

Let the variable $\mathcal{M}$ denote the model. Recall that in the Bayesian framework, the data are generated from a model which is also assumed to be random. The distribution of the model is an *a priori* information that we must provide. The goal of Bayesian inference is to infer the posterior likelihood $\mathbb{P}(\mathcal{M}|\text{data})$. From the Bayes formula:

$$\mathbb{P}(\mathcal{M}|\text{data}) = \frac{\mathbb{P}(\text{data}|\mathcal{M})\,\boldsymbol{\pi}(\mathcal{M})}{\mathbb{P}(\text{data})}, \tag{1.24}$$

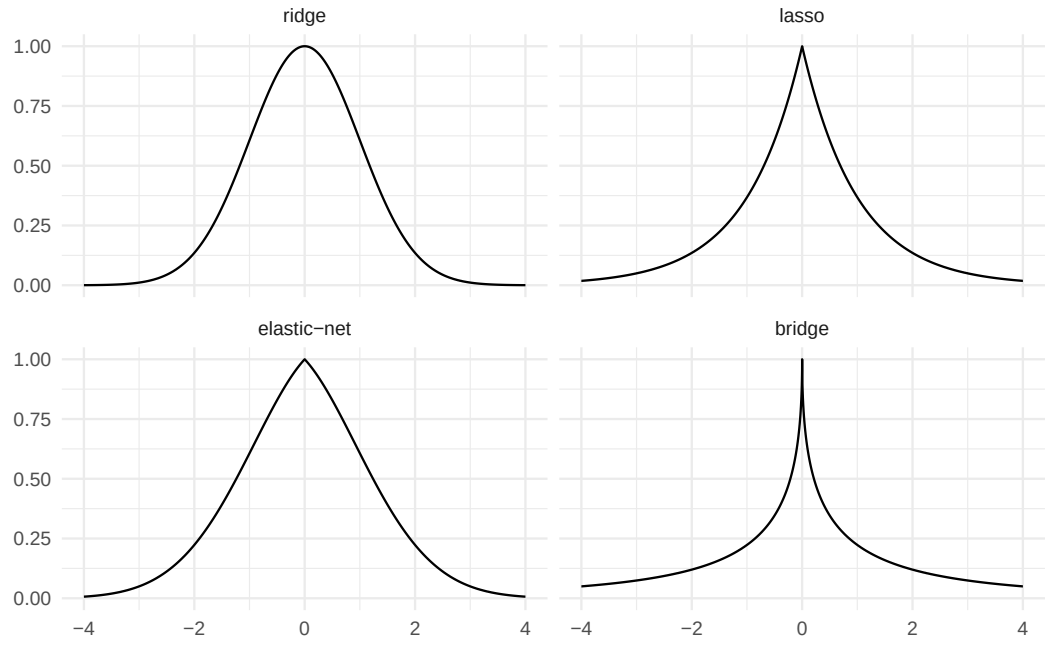


Figure 1.5: Bayesian priors on the parameter with the Ridge, Lasso, Elastic-net ($\alpha = 0.5$), and Bridge ($q = 0.5$) penalties.

where $\boldsymbol{\pi}(\mathscr{M})$ is the prior distribution of the model, which includes *a priori* information about the model distribution.

Here, the symbol for probability must be understood in the sense of density. The prior on the data does not intervene in the minimization of the integrated negative log-likelihood. The first probability in the right-hand side of (1.24) is the likelihood (of the data). In this case, the model $\mathscr{M}$ is the parameter $\boldsymbol{\beta}$ from which the data is generated. Then, Bayesian inference is made by minimizing:

$$-2\log\mathbb{P}\left(\boldsymbol{\beta}|\text{data}\right) = 2\ell\left(\boldsymbol{\beta}\right) - 2\log\boldsymbol{\pi}(\boldsymbol{\beta}), \tag{1.25}$$

where we recall that $\ell(\boldsymbol{\beta})$ is the negative log-likelihood (of the data conditionally on the model). Equation 1.25 is very similar to penalized maximum likelihood estimation. Notice that Equations (1.25) and (1.5) are the same on the condition that

$$-2\log\boldsymbol{\pi}\left(\boldsymbol{\beta}\right) = \lambda\|\boldsymbol{\beta}\|^2.$$

Consequently, the ridge penalty can be seen as a Bayesian inference with a normal prior distribution on the parameter: $\boldsymbol{\beta} \sim \mathscr{N}\left(\mathbf{0}, \lambda^2\right)$. The hyper-parameter λ is the standard deviation of the centered normal prior on the parameter. More generally, the L_q norm penalty corresponds to the prior

$$\boldsymbol{\pi}(\boldsymbol{\beta}) = \exp\left(-\lambda\frac{\|\boldsymbol{\beta}\|_q^q}{2}\right).$$

The lasso regression corresponds to a prior on the parameter with a Laplace distribution. These priors are represented in Figure 1.5 in the univariate case. Likewise, the elastic-net corresponds to the prior

$$\boldsymbol{\pi}\left(\boldsymbol{\beta}\right) \propto \exp\left(-\lambda\frac{\alpha\|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\|\boldsymbol{\beta}_1\|}{2}\right),$$

which is also represented in Figure 1.5.

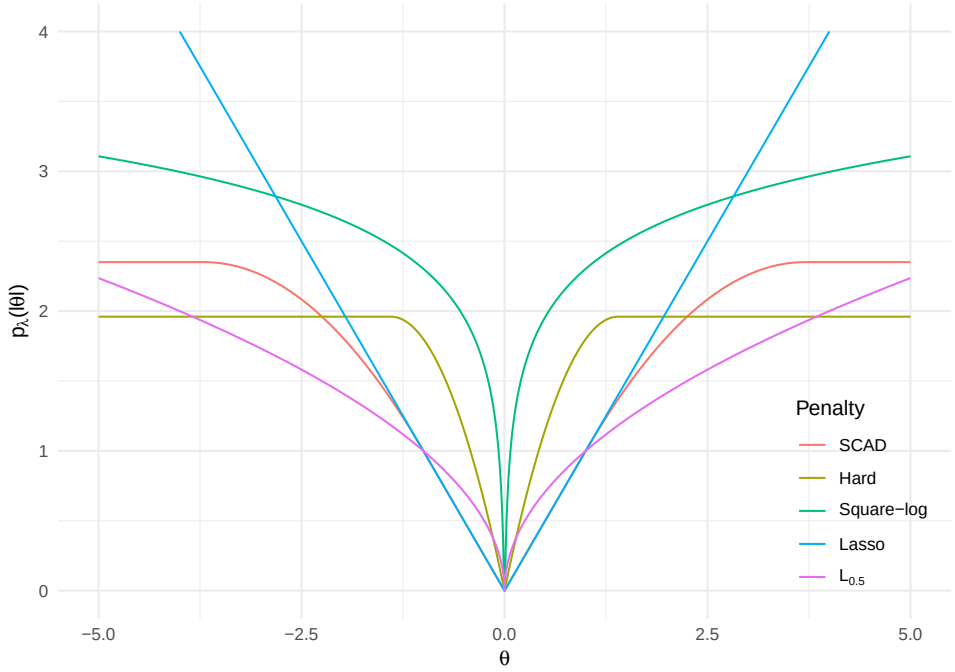


Figure 1.6: Several non-concave penalties with different values of λ.

1.1.7 Non-concave penalties

Non-concave penalized estimates are defined as the minimizers of

$$\ell\left(\boldsymbol{\beta}\right) + \sum_{j=1}^{p} p_\lambda\left(\left|\beta_j\right|\right) \tag{1.26}$$

where p is a *non-concave* function, or *non-concave* penalty, in the following sense.

Definition 3 (Non-concave function)**.** *A real-valued function $p_\lambda(|\theta|)$ defined on $\mathbb{R}$ is said to be non-concave if it is* (i) *even on $\mathbb{R}$ with $p_\lambda(0) = 0$* (ii) *non-decreasing and concave on $(0, \infty)$* (iii) *differentiable over $\mathbb{R}^*$.*

The differentiability condition can be relaxed to piecewise differentiability without without loss of generality. Since it is not a restrictive assumption in practice, we will assume differentiability for the sake of the presentation. We emphasize the fact that in this work, the term "non-concave" is not used to say that a function is not concave.

In the cases where $p_\lambda(|\beta_j|)$ is proportional to λ, we will denote $p_\lambda(|.|)$ by $\lambda p(|.|)$ and drop the index λ. One may also want to penalize some variables more than others, because of some *a priori* knowledge about their importance. This is done by setting a different penalty for each coefficient. For the sake of the presentation, we will assume that all the components of $\boldsymbol{\beta}$ are penalized the same.

The estimates defined by (1.26), where p_λ is a non-concave penalty, will be called *non-concave* estimates. Many important sparse estimates write as non-concave estimates, which makes (1.26) a good framework to analyze and compare sparsity-inducing penalized likelihood methods. Using the definition of non-concave penalty, we can *(i)* give conditions on the penalty for the corresponding estimate to be sparse, *(ii)* give theoretical results shared by non-concave sparse estimates, *(iii)* compare the behavior and performance of sparsity-inducing estimates by comparing their corresponding penalty functions.

The lasso and the bridge ($0 < q < 1$) are both non-concave estimates ; this follows directly from (1.12) and (1.22) respectively. In the following sections we define two important non-concave penalties: the hard-thresholding and the smoothly clipped absolute deviation (SCAD). Figure 1.6 represents several non-concave penalties. Since non-concave penalties are concave and decreasing

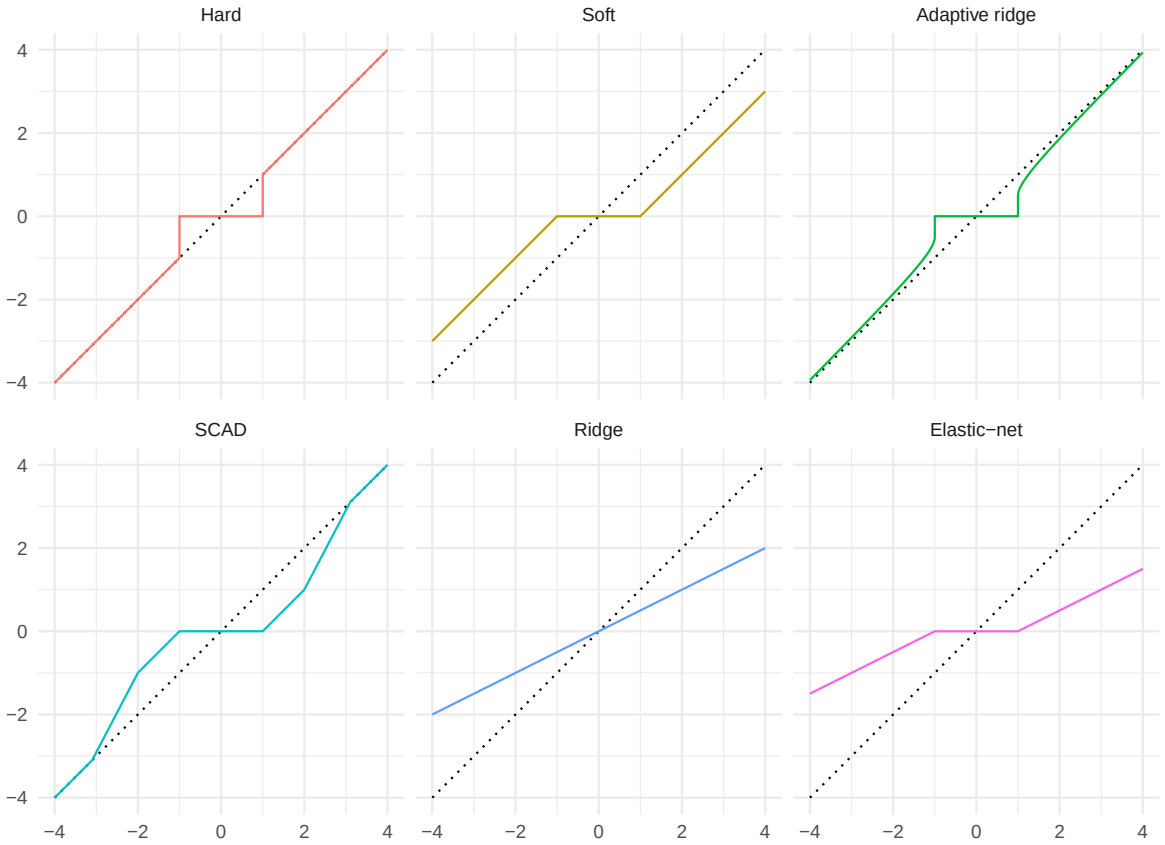


Figure 1.7: Thresholding functions in orthogonal design.

on $(0,\infty)$, they verify $p'(0^+) > 0$. Since they are even functions, we also have $p'(0^-) < 0$. Thus by definition, they are not differentiable at 0. This property is proven (Fan and Li, 2001) to be a sufficient condition for a penalty to perform variable selection. Thus all non-concave penalties such that $p'_\lambda(0^+) > 0$ perform variable selection.

Figure 1.6 illustrates the shapes of several non-concave penalties. The sharper the penalty, the better it approximates the L_0 norm, which is the desired penalty. However, sharp penalties are harder to minimize, which makes the estimation procedure more computationally intensive. First, penalties like the SCAD and hard-thresholding, which are constant away from zero, are hard to minimize. The lasso is an outstanding non-concave penalty: it is *minimally* non-concave, since it is linear on $(0,\infty)$, which is the *least concave* function. Moreover, the derivative of the L_1 penalty is constant on $(-\infty,0)$ and $(0,\infty)$ and thus its minimization cannot be performed using classic optimization methods. Finally, note that empirically, the sharper the spike around zero, the "stronger" the estimates will be enforced to be sparse.

Figure 1.7 represents the thresholding functions of some non-concave penalties under the case of orthogonal design: $\hat{\beta}_j = f(\hat{\beta}_j^{\text{ols}}) = f(\boldsymbol{X}^T\boldsymbol{y})$. Asymptotically the variables with a non-zero effect on $\boldsymbol{y}$ will have an OLS estimate far away from zero and the variables with no effect on $\boldsymbol{y}$ will be around zero. Consequently, the penalty

(i) Performs variable selection only if the thresholding function is identically equal to zero around a neighborhood of zero;

(ii) Estimates the non-zero effect without bias only if the thresholding function is close to the identity function (dotted lines on the figure) for large values of the data.

(iii) Is *stable in selection* if the thresholding function is continuous. (By *stable* we mean that the selected model is not highly sensitive to small variations in the data.)

The convergence properties of the non-concave penalties was studied by Fan and Li (2001). Under mild regularity conditions on λ_n and p_λ the non-concave penalties have the oracle properties. The only restrictive condition is that

$$\limsup_{n\to\infty} \limsup_{\theta\to 0} p'_{\lambda_n}(\theta)/\lambda_n > 0. \tag{1.27}$$

The minimization of non-concave functions is a computationally difficult task. Derivative-based optimization methods are not efficient in this context, because *(i)* the derivative of the penalty is not always continuous and *(ii)* the second order derivative $p''_\lambda(|\theta|)$ takes infinitely small values. Specific optimization methods have to be used to derive numerical estimation procedures that are stable and fast to compute. These methods are introduce in the framework of MM optimization, in Section 1.3.1.

1.1.7.1 Hard-thresholding

Consider the L_0 penalty

$$p_\lambda(|\theta|) = \lambda^2 \mathbb{1}_{\theta\neq 0}, \tag{1.28}$$

also called the entropy penalty.

Note that in orthogonal design, the penalized least squares is $\left(y_i - \beta_i\right)^2 + \lambda^2 \mathbb{1}_{\beta_i\neq 0}$, and since $\|x\|_0$ is constant on $(-\infty,0)$ and $(0,\infty)$, it takes its minimal value either at y_i or at 0. The minimal value is taken at 0 when $\lambda^2 < y_i^2$. Thus, the thresholding function of the L_0 penalty is the so-called *hard thresholding* rule (Figure 1.7):

$$\hat{\beta}^{\text{hard}} = \beta \mathbb{1}_{|\beta|<\lambda}. \tag{1.29}$$

This thresholding only clips the values below the threshold λ to zero, and leaves the other values unmodified. As discussed, this penalty makes the computation of the estimation NP-hard.

Let us now consider the penalty function

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 \mathbb{1}_{|\theta|<\lambda}. \tag{1.30}$$

Surprisingly, under orthogonal design, this penalty also gives the hard thresholding rule (see Antoniadis, 1997; Fan, 1997). Since it is a much simpler quantity to minimize, this penalty is used when we want to have an unbiased variable selection method. This penalty is also referred to as hard thresholding penalty; it is represented in Figure 1.6.

The discontinuity in the hard thresholding makes the model selection unstable. The next penalty offers to remedy this problem, while still having an unbiased estimate of the parameters.

1.1.7.2 SCAD

The smoothly clipped absolute deviation (SCAD) penalty was introduced by Fan (1997) and analyzed by Fan and Li (2001). It is defined by

$$p'_\lambda(|\beta|_j) = \lambda I(\beta_j \leq \lambda) + \lambda \frac{(a\lambda - \beta_j)_+}{(a-1)\lambda} I(\beta_j > \lambda), \tag{1.31}$$

for some $a > 2$ and for $\beta_j > 0$. This penalty is linear over $[0, \lambda]$, parabolic over $[\lambda, a\lambda]$, and constant over $[a\lambda, \infty)$. Therefore it behaves like the lasso for small values of the data and like the hard-thresholding penalty for large values of the data. Under orthogonal design, the SCAD has the following thresholding function (see Figure 1.7):

$$\hat{\beta}_j = \begin{cases} \operatorname{sgn}(\beta_j)\left(|\beta|_j - \lambda\right)_+ & \text{when } |\beta_j| < 2\lambda \\ \{(a-1)\beta_j - \operatorname{sgn}(\beta_j)\, a\lambda\} / (a-2) & \text{when } 2\lambda < |\beta_j| \leq a\lambda \\ \beta_j & \text{when } |\beta_j| > a\lambda \end{cases} \tag{1.32}$$

This thresholding rule equals that of the lasso for $|\beta_j| < \lambda$ and equals the hard thresholding for $|\beta_j| > a\lambda$.

In addition to λ, the parameter a also needs to be determined. We can use cross-validation over a two-dimensional grid, but the computational cost can be deterrent. Fan and Li (2001) performed simulations under orthonormal design with a prior distribution $\beta_j \sim_{\text{iid}} \mathcal{N}(0, a\lambda)$ and estimated the L_2 risk $\mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2]$. They proposed to take the value $a = 3.7$ and noted that the risk does not vary a lot for different values of a.

Note that the SCAD is the only non-concave penalty which is not proportional to λ. This has no influence on the minimization of (1.26) for a fixed value of λ. However, this can make it more complicated to minimize (1.26) for a sequence of values of λ. This difference can make the added computational cost of the SCAD deterrent. We refer to Zou and Li (2008) for more details.

1.1.7.3 Logarithmic penalty

In order to recover highly sparse estimates, non-concave penalties are required to have a "sharp" spike at zero, or in other words, to have a derivative that vanishes away from zero. We have discussed that penalties which are constant away from zero are harder to minimize. An obvious and simple choice is then to use the logarithmic function as a penalty: $\theta \mapsto \log(|\theta|)$. Since the logarithm equals $-\infty$ at zero, this function does not comply as a penalty function. To remedy this issue, we cap the log function away from $-\infty$ in the following manner: define the "log penalty"

$$p(|\theta|) = \log(|\theta| + \varepsilon), \tag{1.33}$$

where $\varepsilon > 0$ is a numerical constant bigger than computer precision, but smaller than the order of magnitude of the estimates. Typically, we chose a value of ε such that $\varepsilon \ll \min\{|\beta_j^{\text{ols}}|, 1 \leq j \leq p\}$.

Name	Penalty function	Additional parameter	Package	Reference
LASSO	$p_\lambda(\|\theta\|) = \|\theta\|$	no	`glmnet`	Friedman et al. (2010b)
Elastic-net	$p_\lambda(\|\theta\|) = (1-\alpha)\|\theta\| + \alpha\theta^2$	$0 < \alpha < 1$	`glmnet`	Friedman et al. (2010b)
Bridge	$p_\lambda(\|\theta\|) = \|\theta\|^q$	$0 < q < 1$	`ncpen`	Huang et al. (2008) Lee et al. (2016)
SCAD	$p'_\lambda(\|\theta\|) = \lambda\mathbb{1}_{\theta\leq\lambda} + \lambda\frac{(a\lambda-\theta)_+}{(a-1)\lambda}\mathbb{1}_{\theta>\lambda}$	$a > 2$	`ncpen` `ncvreg`	Kim et al. (2018) Breheny and Huang (2011)
Adaptive ridge	$p_\lambda(\|\theta\|) = \log\left(1 + \frac{\theta^2}{\varepsilon}\right)$	$\varepsilon \sim 10^{-12}$	`l0ara`	Dai et al. (2018) Frommlet and Nuel (2016)

Table 1.1: Available implementations in `R` of the penalization-based variable selection methods for the linear model.

Surprisingly, this penalty has never been formally introduced or even used in penalized mle estimation, to the best of our knowledge. It was however used in the field of compressed sensing by Candès et al. (2008), which also uses a small constant ε to cap the values of the penalty away from $-\infty$.

Note that Zou and Li (2008) make reference to the penalty $p(|\theta|) = \log(|\theta|)$. More precisely, they define a numerical procedure whose corresponding non-concave penalty would be the logarithm function.

1.1.7.4 Available implementations

We gather in this section a short summary table of the aforementioned penalties with references to their available implementations in `R`: see Table 1.1. No known implementations of the Berhu penalty and the log-penalty have been found.

1.2 Subset selection

In this section we present the methods of variables selection which are not based on penalized estimation. Instead, the methods select the best set of variables by choosing a set of variables and fitting the model without penalty. These methods are simple to implement and are historically important, but since the set of subsets of variables has a cardinal growing exponentially with the number of variables, they become impractical in the high dimensional case $p \gg n$. They have also been criticised for being unstable in selection (see Breiman, 1996).

1.2.1 Best subset selection

The naive approach to model selection is to fit all possible models without penalization and to compare them with a predefined metric. The metric can be a cross-validation procedure or a model selection criterion, as the AIC (Akaike, 1974) or the BIC (Schwarz, 1978). Since each model is defined by a subset of the variables $\{1, \cdots, p\}$, this approach is called *best subset selection*. This method is computationally feasible only for small number of variables. The total number of subsets of $\{1, \cdots, p\}$ is 2^p. When $p = 30$, there are approximately one billion models to fit. In the case of model regression, where model fitting is computationally fast, this is the limiting order of magnitude: when $p \simeq 40$, best subset selection if no longer viable.

Branch and Bound. Furnival and Wilson (1974) introduced the "branch and bound" approach (also called *leaps and bound*) to variable selection (see also Hand, 1981). This algorithm uses a trick to reduce the computational cost of best subset selection. Consider the subsets of $\{1, \cdots, p\}$ as a binary tree, with the full model at the root and all the possible submodels at the leaves; one moves from a

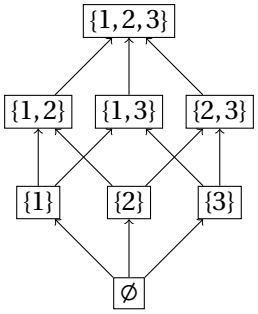


Figure 1.8: All possible submodels with three variables, forming a Hasse diagram of the power-set of $\{1,2,3\}$.

vertex to its two children by either keeping or removing one variable. The naive implementation of the best subset method fits all submodels in this tree. The branch and bound does the same, but uses bounds over the model selection criterion to prune some parts of the tree that are suboptimal compared to the rest of the tree. For the sake of the presentation, assume that the model criterion is the AIC: for model m with p_m parameters and with MLE estimate $\hat{\theta}_m$, we select the model with the lower value of $\text{AIC}(m) = -2\ell(\hat{\theta}_m) + 2p_m$. Consider a set of models M and define the model m^{u} which includes all variables present in any model $m \in M$ and the model m^{i} which includes only the variables present in all models $m \in M$. From the inequalities

$$\ell(\hat{\theta}_{m^{\text{i}}}) \leq \ell(\hat{\theta}_m) \leq \ell(\hat{\theta}_{m^{\text{u}}}) \tag{1.34}$$

$$p_{m^{\text{i}}} \leq p_m \leq p_{m^{\text{u}}}, \tag{1.35}$$

we know that for any model $m \in M$, the AIC has the following lower bound: $\text{AIC}(m) \geq -2\ell(\theta_{m^{\text{u}}}) + 2p_{m^{\text{i}}}$. (We can derive similar lower bounds for other selection criteria.) The branch and bound explores the binary tree from the root and keeps track of the lowest AIC encountered so far. If the current vertex M has a lower bound greater than the lowest AIC in memory, then M cannot include the model with the lowest AIC. It is useless to explore M, and the branch and bound removes it from the tree search. This method is implemented in `R` in the package `leaps` (Miller, 2017).

The branch and bound makes the number of explored models way smaller in practice, but it is still exponential with p. For $p > 40$, the best subset selection becomes unviable. We then have to compare a smaller number of submodels; this is the approach developed in the following section.

1.2.2 Stepwise Selection

Stepwise selection consists in exploring a small subset of all the possible submodels. Consider the Hasse diagram of the submodels partially ordered by the inclusion between models (see Figure 1.8). On one extremity of the diagram is the model with no variables and on the other extremity is the model with all variables. The stepwise selection methods explores one path between the null model and the full model, using a greedy approach. We differentiate between three strategies of exploration. The *forward-stepwise* method – or forward selection – starts at the null model and incrementally adds variables to the model, drawing a path in the direction of the arrows in Figure 1.8. The *backward-stepwise* method – or backward elimination – starts at the full model and incrementally removes variables from the model, drawing a path in the opposite direction than the arrows in Figure 1.8. The *forward-backward* method is a combination of the two methods: at each step, we choose to either remove or add a variable.

The greedy decision rule for the removal or addition of a new variable $\boldsymbol{x}_j$ is based on testing whether $\hat{\beta}_j = 0$. When comparing the potential variables to add (or remove), it is not necessary to fit the whole model. The new estimate is computed using the current estimate and an update of the $\boldsymbol{QR}$ decomposition of $\boldsymbol{X}_c^T\boldsymbol{X}_c$, where $\boldsymbol{X}_c$ is the current design matrix. This considerably improves the

computational burden of this method. For more information on stepwise selection procedures, see Miller (2002, Section 3). These methods are implemented in the `R` package `leaps` (Miller, 2017).

1.3 Iterative penalized methods

1.3.1 MM optimization

MM optimization (see Lange, 2004, Section 6) is a family of optimization procedures. The MM algorithm generalizes the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). The acronym MM stands for both *majorization-minimization* (if we want to minimize a function) and *minorization-maximization* (if we want to maximize it). For the sake of the presentation, we focus on the *majorization-minimization* version.

Let $f\left(\boldsymbol{\beta}\right)$ be the function to minimize. Let $\boldsymbol{\beta}^{(k)}$ be the current point. We say that a function $g\left(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}\right)$ majorizes f if it satisfies:

$$\begin{cases} g\left(\boldsymbol{\beta}^{(k)}|\boldsymbol{\beta}^{(k)}\right) = f\left(\boldsymbol{\beta}^{(k)}\right) \\ g\left(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}\right) \geq f\left(\boldsymbol{\beta}\right), \forall \boldsymbol{\beta} \in \mathscr{C} \end{cases} \tag{1.36}$$

where $\mathscr{C}$ is the convex set on which both f and g are defined. The idea of the MM algorithm is to minimize $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$ instead of $f(\boldsymbol{\beta})$ directly.

In MM optimization, the parameter at the next step $\boldsymbol{\beta}^{(k+1)}$ is defined as any vector verifying $g(\boldsymbol{\beta}^{\boldsymbol{(k+1)}}|\boldsymbol{\beta}^{(k)}) < g(\boldsymbol{\beta}^{\boldsymbol{(k)}}|\boldsymbol{\beta}^{(k)})$. We could set

$$\boldsymbol{\beta}^{(k+1)} = \arg\min_{\boldsymbol{\beta}} g\left(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}\right),$$

but this is not required – and is not considered optimal for most choices of g and of the minimization algorithm for $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$. The MM optimization comes from the fact that for any g satisfying (1.36):

$$f\left(\boldsymbol{\beta}^{(k+1)}\right) \leq g\left(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}\right) \leq g\left(\boldsymbol{\beta}^{(k)}|\boldsymbol{\beta}^{(k)}\right) = f\left(\boldsymbol{\beta}^{(k)}\right), \tag{1.37}$$

so that the MM iterations always reduces the value of f.

An optimization procedure with this *descend* property is praised for being stable. Notice that (1.37) does not imply convergence of the algorithm to the global minimum. Under mild regularity conditions (Lange, 2004), the MM converges to a local minimum. Except when f is convex, we cannot ensure that MM converges to the global minimum.

The effectiveness of the MM optimization depends highly on the choice of the majorizing function. Its expression must be simple, so that its minimization is not too computationally intensive. Yet it must be "close" enough to f, in the sense that $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{\boldsymbol{(k)}}) - f(\boldsymbol{\beta}^{(k)})$ does not increase too much in the neighborhood of $\boldsymbol{\beta}^{(k)}$. Each choice of a majoring function g yields a different MM algorithm. The function g is usually a local (linear or quadratic) approximation of the function f around the current point $\boldsymbol{\beta}^{(k)}$.

In the following section, we will show that many iteratively weighted penalized estimation arise as the MM algorithm of some non-concave penalty.

Remark: EM is a special case of MM. We show here that the Expectation-Maximization (EM) algorithm is a specific case of the MM optimization. The EM algorithm allows to perform inference in the presence of missing, or unobserved data, and was introduced by Dempster et al. (1977). We refer to McLachlan and Krishnan (2007) for an exhaustive reference to the EM.

We give a definition in a simple setting. Assume that you observe data from a random variable $\boldsymbol{Y}$. Assume also that $\boldsymbol{Y}$ is incomplete and that it comes from a random variable $\boldsymbol{X}$ that is not observed. Define $\boldsymbol{\theta}$ the parameter of the distribution of $\boldsymbol{X}$ and note $f(x|\boldsymbol{\theta})$ and $g(y|\boldsymbol{\theta})$ the respective densities of

$\boldsymbol{X}$ and $\boldsymbol{Y}$ conditionally on $\boldsymbol{\theta}$. The goal is to make inference on $g(y|\boldsymbol{\theta})$. In the case of missing data, we have $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{Z})$, where $\boldsymbol{Z}$ is the missing data. But the EM works in a more general setting, where we assume that there is a function t that maps $\boldsymbol{X}$ to $\boldsymbol{Y}$. The trick for the statistician is to find a complete r.v. $\boldsymbol{X}$ whose distribution (and likelihood) is simple.

We can now define the EM algorithm. It consists of iterating between two steps: the E steps defines the conditional expectancy of the missing data conditionally on the observed data (and of the parameter). The M step maximizes this density. By sequentially maximizing a marginalized density, we can infer the parameter of the missing data.

Algorithm 2 Expectation-Maximization

1: **function** EM($\boldsymbol{\theta}^{(0)}$)
2: $k = 1$
3: **while** not converge **do**
4: $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) \leftarrow \mathbb{E}\left[\log f(\boldsymbol{X}|\boldsymbol{\theta})|\boldsymbol{Y}, \boldsymbol{\theta}^{(k)}\right]$
5: $\boldsymbol{\theta}^{(k)} \leftarrow \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$
6: $k \leftarrow k+1$
7: **end while**
8: **end function**

It is clear from the algorithm that the E-step plays the role of a local approximation of $g(y|\theta)$. The next proposition shows that the E step consists in fact of a function minorizing $\log g(y|\boldsymbol{\theta})$.

Property 3. *The function* $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) - Q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)}) + \log(g(y|\boldsymbol{\theta}^{(k)}))$ *minorizes* $\log(g(y|\boldsymbol{\theta}^{(k)}))$ *in the sense of* (1.36)*:*

$$\begin{cases} q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) \leq & \log(g(y|\boldsymbol{\theta})), \forall \boldsymbol{\theta} \\ q(\boldsymbol{\theta}^k|\boldsymbol{\theta}^{(k)}) = & \log(g(y|\boldsymbol{\theta}^{(k)})). \end{cases}$$

Thus E-step is the local approximation using $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ *and the M-step is the maximization of* $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$.

Proof. The equality case is trivial. It suffices to prove that $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) - \log(g(y|\boldsymbol{\theta})) \leq Q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)}) - \log(g(y|\boldsymbol{\theta}^{(k)}))$. We need to apply the Lemma 1. The detailed proof of the applicability of this lemma is somewhat technical, and is skipped here (we refer to Lange, 2004, p. 139). We have

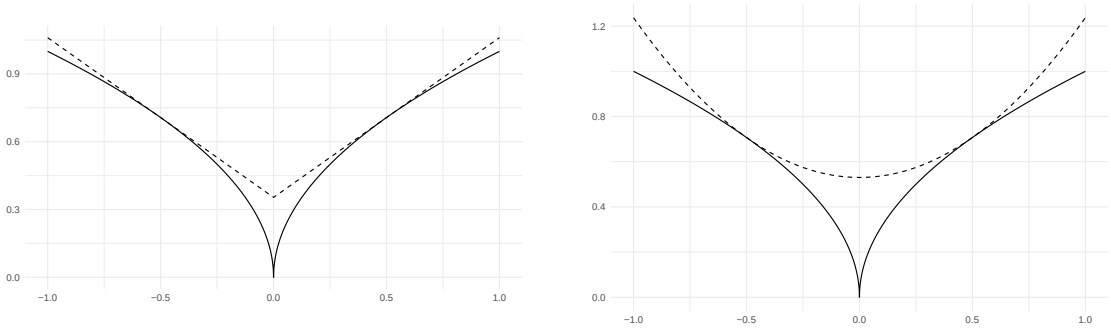
$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) - \log(g(y|\boldsymbol{\theta})) =& \mathbb{E}\left[\log(f(X|\boldsymbol{\theta}))|Y = y, \boldsymbol{\theta}^{(k)}\right] - \mathbb{E}\left[\log(g(Y|\boldsymbol{\theta}))|Y = y, \boldsymbol{\theta}^{(k)}\right] \\ =& \mathbb{E}\left[\log\left(\frac{f(X|\boldsymbol{\theta})}{g(Y|\boldsymbol{\theta})}\right)|Y = y, \boldsymbol{\theta}^{(k)}\right] \\ \leq& \mathbb{E}\left[\log\left(\frac{f(X|\boldsymbol{\theta}^{(k)})}{g(Y|\boldsymbol{\theta}^{(k)})}\right)|Y = y, \boldsymbol{\theta}^{(k)}\right] \\ =& Q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)}) - \log(g(y|\boldsymbol{\theta}^{(k)})) \end{aligned}$$

This completes the proof. □

Lemma 1. *Given two almost surely positive densities* u *and* v, $\mathbb{E}_h\left[\log(h)\right] \geq \mathbb{E}_h\left[\log(k)\right]$.

1.3.2 MM optimization applied to non-concave penalties

MM optimization can be used to minimize (1.26). Define $f(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|)$ as the function to minimize, where $p_\lambda(|\theta|)$ is a non-concave function. The optimization algorithm will depend on the choice of the dominating function g. In this section, we present two possibilities for g, which yield to different numerical procedures. Even though they aim at minimizing the same quantity, theses two algorithms have different properties in practice.



(a) LLA (b) LQA

Figure 1.9: Local linear approximation (a, in dashed line) and local quadratic approximation (b, in dashed line) of the bridge penalty (solid line) $p(|\theta|) = |\theta|^{0.5}$ around the current point $\theta^{(k)} = 0.5$.

1.3.2.1 Local Quadratic Approximation

Fan and Li (2001) proposed a dominating function based on the Local Quadratic Approximation (LQA) of the penalty function. The original paper introduced this approximation as a mere numerical trick. It was formalized as an MM procedure by Zou and Li (2008).

The first order approximation of $p'_\lambda(|\theta|)$ around $\theta^{(k)}$ writes

$$\big(p_\lambda(|\theta|)\big)' = p'_\lambda(|\theta|)\mathrm{sgn}(\theta) \simeq \frac{p'_\lambda(|\theta^k|)}{|\theta^k|}\theta \tag{1.38}$$

Taking the antiderivative of the both terms in the previous equation yields the following approximation of the penalty (see Figure 1.9b):

$$p_\lambda(|\theta|) \simeq q_\lambda(\theta|\theta^{(k)}) \triangleq p_\lambda(|\theta^{(k)}|) + (\theta^2 - {\theta^{(k)}}^2)\frac{p'_\lambda(|\theta^{(k)}|)}{2|\theta^{(k)}|} \tag{1.39}$$

We will establish that $q_\lambda(\theta|\theta^k)$ majorizes $p_\lambda(|\theta|)$. We prove the property for $\theta \in (0,\infty)$ and since the same reasoning applies to $\theta \in (-\infty,0)$ (and to $\theta = 0$ by continuity), the property will be proven over $\mathbb{R}$. For $\theta \geq 0$, we have

$$\begin{aligned}\Big[q_\lambda(\theta|\theta^{(k)}) - p_\lambda(|\theta|)\Big]' &= \theta\frac{p'_\lambda(|\theta^{(k)}|)}{|\theta^{(k)}|} - \mathrm{sgn}(\theta)p'_\lambda(|\theta|) &(1.40)\\ &= \theta\left(\frac{p'_\lambda(|\theta^{(k)}|)}{|\theta^{(k)}|} - \frac{p'_\lambda(|\theta|)}{\theta}\right). &(1.41)\end{aligned}$$

Since $\theta \mapsto p'_\lambda(|\theta|)/\theta$ is non-increasing over $(0,\infty)$, the right hand-size term of the last equation is strictly positive over $(0,\theta^{(k)})$ and strictly negative over $(\theta^{(k)},\infty)$. Consequently $q_\lambda(\theta|\theta^{(k)}) - p_\lambda(|\theta|)$ is decreasing over $(0,\theta^{(k)})$ and increasing over $(\theta^{(k)},\infty)$. With $q_\lambda(\theta^{(k)}|\theta^{(k)}) - p_\lambda(|\theta^{(k)}|) = 0$, this completes the proof. The function $p(|\theta|)$ and its LQA $q(\theta|\theta^{(k)})$ around $\theta^{(k)} = 0.5$ are represented in Figure 1.9a.

When minimizing $q(\theta|\theta^{(k)})$, we can neglect the constant terms. We derive the following dominating function for the penalized nll:

$$g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}) = \ell(\boldsymbol{\beta}) + \sum_{j=1}^{p}\frac{p'_\lambda(|\beta_j^{(k)}|)}{2|\beta_j^{(k)}|}\beta_j^2. \tag{1.42}$$

The minimization of $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$ is a weighted ridge problem whose solution is explicit (from Equation 1.6). The MM algorithm iterates over

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \left(\nabla^2 \ell(\boldsymbol{\beta})^{(k)} + \Omega_\lambda(\boldsymbol{\beta}^{(k)})\right)^{-1} \left(\nabla \ell(\boldsymbol{\beta}^{(k)}) + \Omega_\lambda(\boldsymbol{\beta}^{(k)})\boldsymbol{\beta}^{(k)}\right), \tag{1.43}$$

until convergence, where $\boldsymbol{\Omega}_\lambda(\boldsymbol{\beta}^{(k)})$ is a diagonal matrix with diagonal entries $\left\{p'_\lambda(|\beta_j^{(k)}|)/2|\beta_j^{(k)}|\right\}$.

A notable special case is the linear model, for which (1.43) simplifies to:

$$\boldsymbol{\beta}^{(k+1)} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Omega}_\lambda(\boldsymbol{\beta}^{(k)}))^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{1.44}$$

The denominator in (1.39) is source of numerical instabilities when β_j gets close to zero. To overcome this problem, we can either

(i) Set β_j to zero if its absolute value becomes smaller than a fixed threshold, and continue the iterations without the jth component This solution is easy to implement, but once a component of $\boldsymbol{\beta}$ is set to zero, it will never be selected again. Thereby, the iterative estimate has same disadvantage as stepwise backward elimination: once a variable is removed from the model, it is never considered again. This restriction can be very limiting in practice.

(ii) Bound the denominator away from zero in (1.42) by fixing a small $\varepsilon > 0$ and solving

$$\boldsymbol{\beta}^{(k)} = \arg\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) + \sum_{j=1}^{p} \frac{p'_\lambda(|\beta_j^{(k)}|)}{2(|\beta_j^{(k)}| + \varepsilon)}\beta_j^2, \tag{1.45}$$

as suggested by Hunter and Li (2005). This perturbed version is not an MM algorithm any more, and the descent property is not verified. Hunter and Li (2005) showed that (1.45) is the dominating function of a perturbed version of the negative nll, and the LQA estimator converges towards the argmax of this quantity. Under a mild regularity condition (Hunter and Li, 2005) on $f(\boldsymbol{\beta})$, the limit estimator obtained by (1.45) is close to that of (1.44) when ε tends to zero.

1.3.2.2 Local Linear Approximation

Zou and Li (2008) introduced a dominating function based on the local linear approximation of the penalty function. A first order approximation of $p_\lambda(|\theta|)$ around the current point $\theta^{(k)}$ gives the Local Linear Approximation (LLA):

$$p_\lambda(|\theta|) \simeq l_\lambda(\theta|\theta^{(k)}) \triangleq p_\lambda(|\theta^{(k)}|) + (|\theta| - |\theta^{(k)}|)p'_\lambda(|\theta^{(k)}|) \tag{1.46}$$

Lets us prove that this approximating function majorizes $p_\lambda(|\theta|)$. Since both terms of (1.46) are even with respect to θ, one can assume $\theta > 0$ in the following proof. The local linear approximation $l(\theta|\theta^{(k)})$ of the bridge penalty $p(|\theta|) = |\theta|^{0.5}$ around the current point $\theta^{(k)} = 0.5$ is represented in Figure 1.9a. From

$$\left[l_\lambda(\theta|\theta^{(k)}) - p_\lambda(|\theta|)\right]' \quad = \quad p_\lambda(|\theta^{(k)}|) - p_\lambda(|\theta|), \tag{1.47}$$

and the concavity of p_λ, it follows that $l_\lambda(\theta|\theta^{(k)}) - p_\lambda(|\theta|)$ is minimal at $\theta^{(k)}$, where it equals zero. This completes the proof.

Derive the following dominating function for the penalized nll:

$$g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}) = \ell(\boldsymbol{\beta}) + \sum_{j=1}^{p} p'_\lambda(|\beta_j^{(k)}|)|\beta_j|. \tag{1.48}$$

Minimizing this quantity is a weighted lasso problem. When the model at stake is the linear model or the generalized linear model, $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$ can be minimized efficiently using the LARS algorithm.

The LLA iteration step is more computationally intensive than that of the LQA, which is explicit. However the fact that each iteration step of the LLA is a lasso problem has deep consequences. The LLA converges towards a sparse model using a sequence of sparse models. Consequently, the LLA need not be very close to converge to provide a sparse model. Comparatively, the LQA's converging sequence is only asymptotically sparse, and many iterations might be necessary to obtain a good approximation of the limiting sparse model. In practice, much fewer iteration steps are carried out for the LLA than for the LQA.

The MM converges to a local minimum, so the choice of the initial guess $\boldsymbol{\beta}^{(0)}$ is important. For both LLA and LQA, the standard choice is to set $\boldsymbol{\beta}^{(0)}$ as the unpenalized maximum likelihood. When the model is not overparametrized, this initial guess is considered to be not too far from the global minimum.

Zou and Li (2008) go as far as introducing the *one-step* estimator, which is obtained by performing only one step of the MM algorithm:

$$\boldsymbol{\beta}^{(1)} = \underset{\boldsymbol{\beta}}{\arg\min} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} p'_\lambda(|\beta_j^{(0)}|)|\beta_j|, \tag{1.49}$$

where the initial guess is the OLS: $\boldsymbol{\beta}^{(0)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. This estimator is a weighted lasso estimator, similarly to Zou (2006)'s adaptive lasso. In fact, for $0 < q < 1$, the one-step estimate with the L_q penalty is the adaptive lasso with parameter $\gamma = 1 - q > 0$.

1.3.3 Adaptive Ridge

Definition of the adaptive ridge. The Adaptive Ridge is an iterative penalized regression method. It shares some similarities with the adaptive lasso, with the notable difference that each iteration is less computationally expensive. As its name suggests, it consists in iterating over a weighted ridge penalty, with the value of the weights adapted at each iteration.

The adaptive ridge method was used by Rippe et al. (2012) as a numerical trick for approaching an L_0 norm penalty. It was analyzed in a more general setting by Frommlet and Nuel (2016) as an iterative procedure to numerically approach any L_q penalty, for $q \geq 0$. This method iterates over a weighted L_2 penalty problem while adaptively changing the weights at each iteration. It is defined as the following procedure:

(i) Set $\boldsymbol{w}^{(0)} = \mathbf{1}$

(ii)
$$\hat{\boldsymbol{\beta}}^{(k)} = \underset{\boldsymbol{\beta}}{\arg\min}\, \ell\left(\boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{p} w_j^{(k-1)} \beta_j^2 \tag{1.50}$$

(iii)
$$w_j^{(k)} = (|\beta_j^{(k)}|^\gamma + \varepsilon^\gamma)^{(q-2)/\gamma} \tag{1.51}$$

(iv) Iterate between the last two steps until convergence.

Equations (1.50) and (1.51) are justified empirically by the following reasoning. Assume $\boldsymbol{\beta}^{(k)}$ is reasonably close to the limit estimate $\boldsymbol{\beta}^{(\infty)}$. Then $\boldsymbol{\beta}^{(k+1)}$ will be not too far from $\boldsymbol{\beta}^{(k)}$ and the quantity can be well approximated by

$$w_j^{(k-1)}\beta_j^2 = \frac{\beta_j^2}{(|\beta_j^{(k)}|^\gamma + \varepsilon^\gamma)^{(2-q)/\gamma}} \simeq \frac{\beta_j^2}{(|\beta_j|^\gamma + \varepsilon^\gamma)^{(2-q)/\gamma}}.$$

This function is differentiable and approximates well the L_q penalty. It is represented in Figure 1.10 for different values of q. The solid lines represent the penalties and the dotted lines represent their approximations. The parameter q is usually chosen *a priori* ; it is rarely cross-validated in

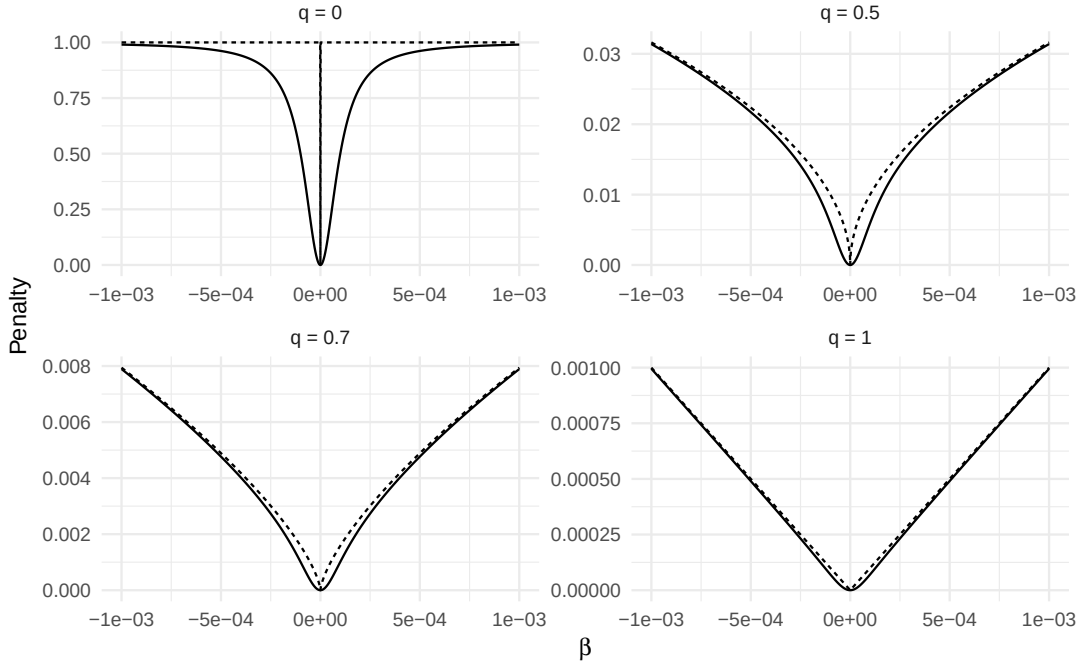


Figure 1.10: L_q penalties (solid lines) and their approximations by the Adaptive Ridge (dotted lines) with $\gamma = 2$ and $\varepsilon = 10^{-4}$.

practice. Namely, for $q = 1$, the Adaptive Ridge gives another algorithm for computing the lasso. The estimating function is continuous even in the case where $q = 0$ and the penalty is not continuous.

The approximation is very close to the L_q penalty on the sets $(-\infty, \varepsilon)$ and (ε, ∞), thus choosing a very small value for ε will increase the quality of the approximation. However ε has to be large enough to avoid the numerical instabilities in (1.51). Using a theoretic argument, Hunter and Li (2005) proposed a value that depends on the smallest valued OLS component. Using a simulation study, Frommlet and Nuel (2016) proposed the value 10^{-8}.

The parameter $\gamma > 0$ determines how well the approximating function tries to fit to the penalty around zero. Simulations carried by Frommlet and Nuel (2016) highlight that setting a large value of γ is however not necessary in practice and show that $\gamma = 2$ seems a reasonable choice.

The L_0 adaptive ridge. The adaptive ridge is numerically efficient to approximate any L_q penalty, $q \geq 0$. Setting $q > 0$ yields a procedure to numerically estimate the bridge estimator. Since, to our knowledge, the L_0 norm has never been efficiently approximated in the contexted of penalized maximum likelihood estimation, the case $q = 0$ is of particular interest. We refer to this case as the "L_0 adaptive ridge". Note that this term does not signify that the penalty used is the L_0 norm. The reweighing step of the L_0 adaptive ridge with $\gamma = 2$ writes:

$$w_j^{(k)} = \frac{1}{{\beta_j^{(k)}}^2 + \delta^2}$$

In orthogonal design, the limit estimate ($k \to \infty$) of the L_0 adaptive ridge is:

$$f(\beta_j^{\text{ols}}) = \begin{cases} 0 & |\beta_j^{\text{ols}}| \leq 2\sqrt{\lambda} \\ \frac{1}{2}\left(\beta_j^{\text{ols}} + \text{sgn}(\beta_j^{\text{ols}})\sqrt{{\beta_j^{\text{ols}}}^2 - 4\lambda}\right) & |\beta_j^{\text{ols}}| > 2\sqrt{\lambda} \end{cases} \tag{1.52}$$

which is represented in Figure 1.7. This thresholding function is equal to zero on the interval $[-2\sqrt{\lambda}, 2\sqrt{\lambda}]$, which proves that the limit estimate of the L_0 adaptive ridge is sparse.

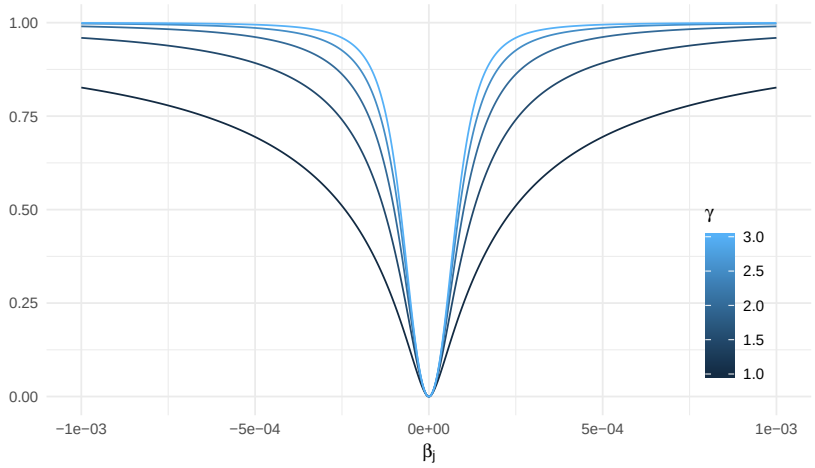


Figure 1.11: Approximations of L_q penalties by the Adaptive Ridge for different values of γ, and with $q = 0$ and $\varepsilon = 10^{-4}$.

The adaptive ridge iteration (with $\gamma = 2$) has already been used by Chartrand (2008) in the context of compressed sensing. In compressed sensing, the signal $\boldsymbol{y}$ can be recovered completely from its encoding $\boldsymbol{\beta}$ since the matrix $\boldsymbol{X}$ has more columns than rows. The author iteratively solves

$$\min_{\boldsymbol{\beta}} \sum_{j=1}^{p} w_j \beta_j^2 \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta},$$

instead of (1.50), but the weights w_i are adaptive as in (1.51).

In their implementation, the authors adaptively reduce the value of ε as the absolute error between successive values of $\hat{\boldsymbol{\beta}}^{(k)}$ decreases. To ensure numerical accuracy, the value of ε is capped at 10^{-8}. The authors compare the square log penalty to Candès et al. (2008)'s log penalty. The square log penalty yields higher recovery rates of $\boldsymbol{y}$ with very sparse reconstructed signal $\hat{\boldsymbol{\beta}}$.

The L_1 adaptive ridge. The adaptive ridge could be used to approximate the L_1 norm, by setting $q = 1$ in (1.50). The L_1 adaptive ridge bypasses the numerical difficulty posed by the minimization of the L_1 norm. However, different tools have been developed to solve the lasso problem efficiently (see Section 1.1.2). Consequently, this version of the adaptive ridge is not particularly superior to the lasso as far as computational speed is concerned. We note that it was already introduce by Vogel and Oman (1996), who define the iterative procedure as a "lagged diffusivity fixed points iteration". The authors developed this method in the framework of image denoising and the minimization of the penalized likelihood (Equation 1.50) was conducted using the Newton-Raphson method.

1.3.3.1 Relation to similar procedures

Weighted L_1 regression. Candès et al. (2008) introduced a weighted L_1 regression problem in the context of compressed sensing. The algorithm uses a weighted L_1 penalty (see Algorithm 3). The constant $\varepsilon > 0$ ensures that the procedure stays stable when a coefficient is set to zero. It also helps against numerical instability issues arising when dividing by small numbers. Candès et al. (2008) offer to adapt the value of ε at each iteration, setting it to the smallest absolute value of the coefficients of the current estimate (but no smaller than 10^{-3}).

The formula for the weights is very close to that of the non-iterative procedure adaptive lasso (see Equation 1.18) with $\gamma = 1$. In compressed sensing, the matrix $\boldsymbol{X}$ has more columns than rows and the problem $\min \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|$ is ill-posed. Consequently there is no OLS estimate to plug in the formula of the weights. This justifies the use of an adaptive procedure where the OLS estimate in Equation 1.18 is replaced by the current estimate.

Algorithm 3 Candes's weighted L_1 minimization

function ITERATIVE $L_1(\boldsymbol{y}, \boldsymbol{X}, \lambda, \varepsilon)$
  $\boldsymbol{w}^{(0)} \leftarrow \mathbf{1}$
  $k \leftarrow 0$
  while not converge **do**
    $\boldsymbol{\beta}^{(k)} \leftarrow \arg\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} w_j^{(k)} |\beta_j|$
    $w_j^{(k)} \leftarrow \frac{1}{|\beta_j^{(k)}| + \varepsilon}$
  end while
end function

Consider now the non-concave penalized problem

$$\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \log\left(|\beta_j| + \varepsilon\right). \tag{1.53}$$

1.3.3.2 Numerical performance

This procedure is strikingly easy to implement and fast to compute. Indeed, the ridge problem is easier to solve than the lasso problem and the reweighing step has complexity $\mathcal{O}(p)$. Comparatively, the lasso estimate is more computationally intensive. This makes the adaptive ridge quicker to compute than Zou and Li (2008)'s one step estimates an Candès et al. (2008)'s adaptive procedure for the log penalty. The downside of the adaptive ridge is that the k-th step estimate is not sparse: only the limit estimate $\hat{\beta}^{(\infty)}$ is sparse. In practice, the adaptive ridge procedure must be stopped when it is estimated to have converged, and the estimate's coordinates that are close to zero must be set at zero.

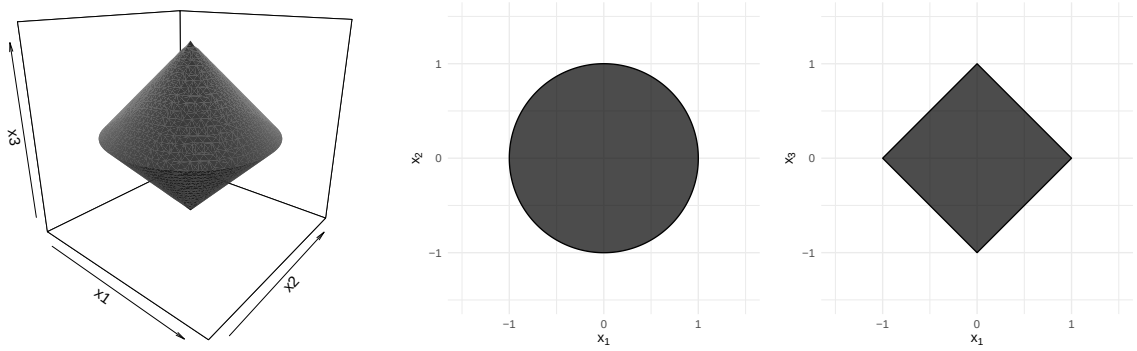
1.4 Structured variable selection

The methods introduced previously are useful in cases with the prior assumption that $\boldsymbol{\beta}^*$ is sparse, i.e., many of its components are zero. In some applications, the true parameter is assumed to have a simple structure other than having many components to zero. This specific *sparsity structure* can vary greatly and depends on the relationship between parameters. For instance if $\boldsymbol{\beta}^*$ is assumed piecewise constant, this prior is expressed as the sparsity of $(\beta_j - \beta_{j-1})$. In general, many prior structures of the parameters are expressed as the sparsity of a transformation of the data. This section presents different possible transformations available to ensure different sparsity structures.

1.4.1 Group lasso

In some practical situations in statistical regression, the variables belong to a group. An important example of such situation is the regression against factor variables that are represented as dummy variables. Another instance of *grouped variable* occurs when we know *a priori* that some variables have a common effect on the response variable. In these cases, it makes little sense to perform variable selection that would select some variables inside a group and not the others. Variable selection techniques have been developed to take into account the presence of groups, and to select groups of variables together.

Yuan and Lin (2006) introduced the group lasso, building on ideas by Bakin (1999) and Lin and Zhang (2006). Consider a partition $\mathcal{G} = (\mathcal{G}_k)_{1 \leq k \leq K}$ of the set of indices of variables $\{1, \cdots, p\}$. Assume that the variables are partitioned into K groups which are given by $\mathcal{G}$. For $k \in \{1, \cdots, K\}$, p_k denotes the number of covariates in the k-th group (we have $\sum_k p_k = p$) and $\boldsymbol{\beta}_k$ denotes the vector containing



(a) In 3d (b) In the plane $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ (c) In the plane $(\boldsymbol{x}_1, \boldsymbol{x}_3)$

Figure 1.12: Unit ball of the group lasso norm with three variables and with $\mathcal{G} = \{\{1,2\},\{3\}\}$. We represent the 3d ball (a) and its restriction to the variables $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ (b) and $\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$ (c).

only the variables in $\mathcal{G}_k$. Without loss of generality, we can write $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_K^T)^T$. The group lasso is defined as the mimimizer of

$$\ell(\boldsymbol{\beta}) + \lambda \sum_{k=1}^{J} w_k \left\| \beta_k \right\|_q, \tag{1.54}$$

with positive weights $w_{p_j} > 0$.

This *group norm*, also called L_1/L_q norm or mixed norm (Bach, 2011) is the norm obtained by a L_1 averaging of the L_q norm defined on each group. It behaves as an L_q norm between variables of each group and as an L_1 norm between groups. Consequently, for $q > 1$, the $\mathrm{L}_1/\mathrm{L}_q$ norm performs selection between groups and performs shrinkage (without selection) within each group. It is easy to see that the $\mathrm{L}_p/\mathrm{L}_q$ penalty is a norm for every $p \geq 1$, $q \geq 1$. Notice that when the partition of $\{1, \cdots, p\}$ is the trivial partition composed of singletons, (1.54) becomes the lasso penalty and when there is only one group, (1.54) becomes the ridge penalty. The group lasso can therefore be seen as a compromise between the ridge and the lasso penalty.

There are several possible choices for q. Some authors (Turlach et al., 2005), (Zhao et al., 2009) consider the $\mathrm{L}_1/\mathrm{L}_\infty$ mixed norm, mainly because optimization with this penalty is numerically efficient (see Mairal et al., 2011; Zhao et al., 2006). However the preferred choice in the literature is the Euclidean norm, $q = 2$, which shrinks equally in all directions.

We now illustrate the group penalty on a simple example. Consider the case of $p = 4$ variables divided into $K = 2$ groups: $\mathcal{G}_1 = \{1,2\}$ and $\mathcal{G}_2 = \{3,4\}$. The group lasso penalty writes $\sqrt{\beta_1^2 + \beta_2^2} + \sqrt{\beta_3^2 + \beta_4^2}$. This mixed norm is a L_1 norm with respect to the groups p_1 and p_2, and is a L_2 norm inside each group. Consequently, the L_1 norm will enforce selection of $(\boldsymbol{\beta}_{p_1}, \boldsymbol{\beta}_{p_2})$ and will not enforce selection of variables inside each group. The variables $1,2$ can be selected together but not alone – and the same goes for $3,4$. The effect of the variables $1,2$ inside the group p_1 is given by the L_2 norm. The selection effect of the mixed norm is illustrated in Figure 1.12, which represents the unit ball in the case where the $p = 3$ variables are partitioned by $\mathcal{G} = \{\{1,2\},\{3\}\}$. This ball is a double cone with singularities at its top and bottom and at its central circle. As is the case for any penalty norm, these singularities give the sparsity structure encoded by the mixed norm: either β_1 and β_2 are jointly set to zero (in which case $\boldsymbol{x}_3$ is estimated with shrinkage), or only β_3 is set to zero (in which case $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are estimated with ridge shrinkage).

The weights w_{p_j} are included to adjust for the relative importance of the groups. If one group has twice as many components as another, it will be regularized $\sqrt{2}$ times less – assuming all variables have equal importance. Consequently, the choice $w_{p_j} = \sqrt{p_j}$ is often used to renormalize the importance of each group.

Yuan and Lin (2006) introduced a group LARS algorithm, a group version of the LARS algorithm which successively adds groups of variable to the active set. Contrarily to the LARS, which can be

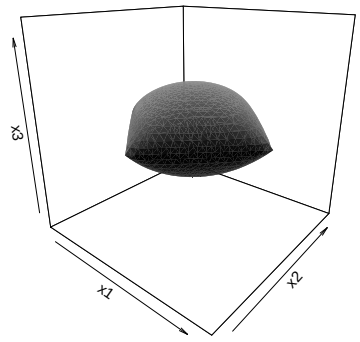


Figure 1.13: Unit ball for the L_1/L_2 mixed norm with overlapping groups $\{\{1,3\},\{2,3\}\}$.

easily modified to compute the lasso solution (Efron et al., 2004), the group LARS generally differs from the solution of (1.54). However, since the L_1/L_2 penalty is a norm, the penalized likelihood (1.54) is convex and the group lasso solution can be computed in reasonable time using the coordinate descent algorithm (or any convex optimization method). Roth and Fischer (2008) presented a comp utationally efficient algorithm to compute the group lasso.

The group lasso can be naturally extended to any mixed norm. One could for instance consider the L_p/L_q mixed norm, where $p < 1$ and $q > 1$, such that the norm *between* groups induces sparsity and the norm *inside* does not induce sparsity. Turlach et al. (2005) introduced a version of the group lasso using the L_1/L_∞ norm.

We emphasize that the group lasso is to be used when clear *a priori* knowledge is known about the variable group. Theoretically, we could set the partition as a Bayesian variable and try to find the best partition. We could explore the set of partitions of $\{1,\cdots,d\}$ by following the vertices of the Hasse diagram defined by the partial order on partitions (see Figure 1.8). We could explore this set using a discrete MCMC algorithm. However, in practice, this Hasse diagram is highly connected so exploring this space with an MCMC approach is computationally unfeasible.

1.4.2 Overlapping groups

The previous section considers a bagging of the covariates based on a partitioning of $\{1,\cdots,p\}$. This group sparsity structure corresponds mainly to the case of factor covariates represented using dummy variables. In this section we consider generalizations to subsets of $\{1,\cdots,p\}$ which do not form a partition. It turns out that the sparsity obtained from such overlapping groups are of great use to include interesting *a priori* structure between the variables. This extension of the group lasso was introduced by Zhao et al. (2009) and Jacob et al. (2009).

We keep the same notation as in the previous section, except that here $\mathcal{G} = (\mathcal{G}_1,\cdots,\mathcal{G}_K)$ is an element of the power set of $\{1,\cdots,p\}$. Without loss of generality, we can assume that each variable is present in at least one group $\mathcal{G}_k$. We consider the same penalty as before:

$$\ell(\boldsymbol{\beta}) + \lambda \sum_{k=1}^{K} \left\| \boldsymbol{\beta}_k \right\|_{q_k}, \tag{1.55}$$

where $q_k > 1$ is the order of the penalty norm inside $\mathcal{G}_k$. As already discussed in the previous section, setting $q_k = 2$ for all k will give a ridge penalty inside each group.

First, note that the mixed norm (1.55) remains a norm when $\mathcal{G}$ is no longer a partition. Consequently the penalty is convex and the group lasso has the same computational cost when the groups have overlapping.

We now develop on the type of sparsity structure that overlapping groups can encode. Overlapping groups allow to encode *a priori* information of the type "if one variable is removed, other variables have to be selected". For the sake of the illustration, consider the following example, with

4 covariates $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_4$. The *a priori* knowledge here is that the variable 1 and 3 are closely related, as well as 1 and 2. Define $\mathcal{G} = (\{1,2\},\{2,3\})$. The group lasso estimate is the minimizer of

$$\ell(\boldsymbol{\beta}) + \lambda\left(\sqrt{\boldsymbol{\beta}_1^2 + \boldsymbol{\beta}_2^2} + \sqrt{\boldsymbol{\beta}_2^2 + \boldsymbol{\beta}_3^2}\right).$$

This penalty sets *all* the variables inside one of the groups to zero. This penalty is represented in Figure 1.13. As illustrated, the singularities enforce that either $\beta_1 = \beta_3 = 0$ or $\beta_2 = \beta_3 = 0$. Consequently the estimated model will either only select the variable 3 or only the variable 2.

There are a large number of possibilities to use overlapping groups to include prior knowledge on the link between variables. However, defining groups that encode the desired sparsity structure is a difficult task, and is somewhat counter-intuitive. Indeed if a variable is selected in the model, all the variables which have one group in common with this variable will be kept in the model. Consequently the selected variables will be formed by the complementary of a union of the groups $\mathcal{G}_k$. It is often easier to express the relationship between variables through a measure of their proximity and we want to use the groups defined using this proximity measure. The relevant penalty corresponds to selecting variables whose support is formed by a union of the groups $\mathcal{G}_k$. Jacob et al. (2009) introduced this group norm. For instance, with the group $\mathcal{G} = \{\{1,3\},\{2,3\}\}$, this penalty removes either β_2 or β_3 from the model, and does not select β_3. This method is of great use in signal processing, where it is easy to specify the *prior* structure of an object using a union of basic structures. For examples of applications in signal processing and image processing, see Jenatton et al. (2011); Huang et al. (2009).

Sparse Group lasso. Friedman et al. (2010a) introduced an extension of the group lasso which enforces the estimate to be both sparse and sparse with respect to its group partition. We use the notations of Section 1.4.1, i.e. $\mathcal{G} = (\mathcal{G}_k)_{1\leq k\leq K}$ denotes a partition of $\{1,\cdots,p\}$. The penalty of the sparse group lasso is obtained by adding an L_1 penalty to (1.54) and choosing $q = 2$. The sparse group lasso estimate minimizes

$$\ell(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\sum_{k=1}^{K}\sqrt{p_k}\|\boldsymbol{\beta}_k\|_2.$$

Note that this penalty is a mere particular case of the overlapping group penalty. Indeed, the sparse group penalty rewrites as (1.55) with the overlapping group $\tilde{\mathcal{G}} = \mathcal{G} \cup \{\{1\},\cdots,\{p\}\}$. In its generalized version (Simon et al., 2013), the sparse group lasso uses a weighted mean between the lasso and the group lasso.

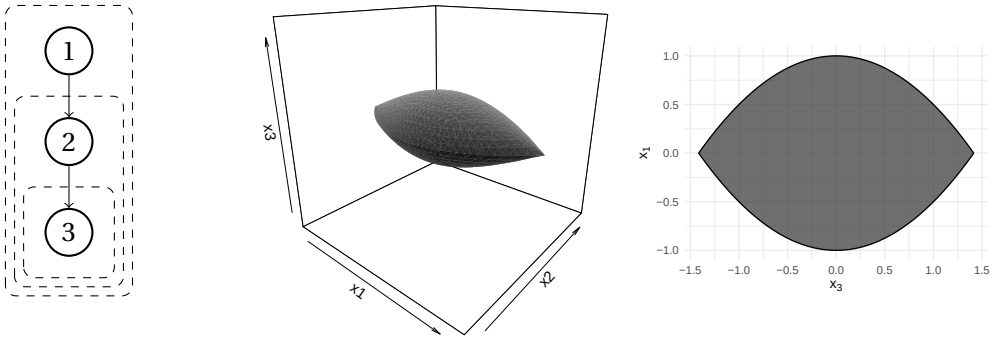
1.4.3 Hierarchical structured sparsity

A particular case of the overlapping group lasso is of great interest in many applications. Suppose that we dispose of a hierarchy between the variables, in the sense that one variable can be more important or more primary than another one. We may want to include this *a priori* knowledge as a constraint on $\boldsymbol{\beta}$. If $\boldsymbol{x}_1$ is "above" $\boldsymbol{x}_2$ in this hierarchy, we say that $\boldsymbol{x}_1$ precedes $\boldsymbol{x}_2$, and it is noted $\boldsymbol{x}_1 \prec \boldsymbol{x}_2$ The hierarchy corresponds to a directed acyclic graph (DAG) between the variables, with the property that the parent variables precede their children. In the following, we assume that the hierarchy includes all the variables. We identify the variables to the vertices of the DAG.

We can build a grouping of the variables such that each variable will only be considered for removal from the model if all of its children have already been removed. We define $\mathcal{G}$ as follows: it contains all the singletons of child-less nodes, then all the groups of size 2 composed of the child-less nodes and their respective parents, then all the groups of size 3 composed of the child-less vertices and their grand-parents, etc until all the variables are included. Note that the groups are nested: if a group contains a variable, it overlaps with all groups containing this variable. There is a one-to-one correspondence between a DAG and its nested group $\mathcal{G}$.

We define the hierarchical group lasso in the same way as before:

$$\ell(\boldsymbol{\beta}) + \lambda\sum_{k=1}^{K}\left\|\boldsymbol{\beta}_k\right\|_2. \tag{1.56}$$



(a) Hierarchy (b) Unit ball in 3d (c) Unit ball in the plane $(\boldsymbol{x}_1, \boldsymbol{x}_3)$

Figure 1.14: (a) Hierarchical structure of the overlapping group $\mathcal{G} = \{\{1,2,3\},\{2,3\},\{3\}\}$, (b) Unit ball of the corresponding mixed norm, (c) Projection of the unit ball on the plane $\boldsymbol{x}_2 = 0$.

The following argument illustrates why the overlapping group penalty encodes the desired hierarchical structure. From the way $\mathcal{G}$ was constructed, we know that any variable $\boldsymbol{x}_k$ appears in (1.56) through an L_1 norm of β_k and an L_2 norm of all the groups containing $\boldsymbol{x}_k$ and its dominating variables. Consequently, $\boldsymbol{x}_k$ will always be removed from the model before any of its ancestors.

We provide a simple example to illustrate hierarchical sparsity structures. Consider the three variables $\boldsymbol{x}_1 \prec \boldsymbol{x}_2 \prec \boldsymbol{x}_3$. The group for this hierarchy is $\mathcal{G} = \{\{1,2,3\},\{2,3\},\{3\}\}$ (see Figure 1.14a). The corresponding penalty is $\sqrt{\beta_1^2+\beta_2^2+\beta_3^2}+\sqrt{\beta_2^2+\beta_3^2}+\sqrt{\beta_3^2}$. The unit ball is represented in Figure 1.14b. The two "spikes" in the unit ball encode the sparsity $\beta_2 = 0$. But since they are located in the plane $\beta_3 = 0$, the sparsity induced by these singularities necessarily induce that $\beta_3 = 0$. The shape of the unit ball correctly encodes the sparsity structure of Figure 1.14a. Finally, Figure 1.14c represents the projection of the unit ball on the variables β_1 and β_3. It illustrate that the third variable will be removed from the model before the first variable.

1.4.4 Fused lasso and total variation

In this section, the parameter vector is assumed to have a natural ordering. This is the case in signal processing, when the response variable $y_i = y(t_i)$ is indexed by time and each parametrization reflects this time ordering. Tibshirani et al. (2005) was among the first to introduces a variable selection method on a transformation of the data. In their method, called *fused lasso*, one assumes that the vector is sparse and that it deviates from zero only on intervals over which it is constant. The fused lasso minimizes

$$\ell(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|, \tag{1.57}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are trade-off constants. When taking $\lambda_2 = 0$, this estimate becomes the lasso estimate. When taking $\lambda_1 = 0$, the penalty becomes the total variation of the parameter.

The total variation, defined by $\mathrm{TV}(\boldsymbol{x}) = \sum_{j=2}^{n} |\boldsymbol{x}_j - \boldsymbol{x}_{j-1}|$, quantifies how much a sequence varies. Thus, the total variation penalty enforces the parameter to be piecewise constant. In the linear model with orthogonal design $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$, the total variation regularized problem writes

$$\underset{\boldsymbol{\beta}}{\arg\min} \|\boldsymbol{y} - \boldsymbol{\beta}\|_2^2 + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|, \tag{1.58}$$

which is referred to as total variation denoising in the signal processing community (Rudin et al., 1992a). It is used to recover noisy signals under the assumption that the original signal is piecewise constant.

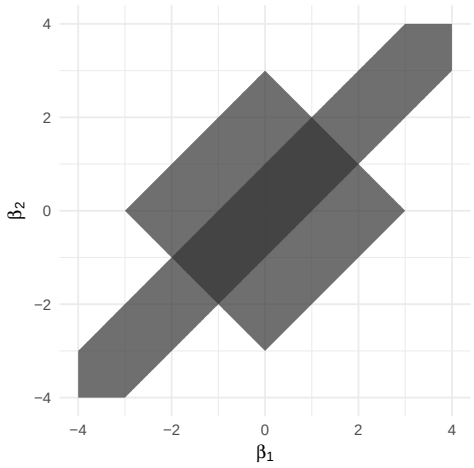


Figure 1.15: Admissible solution set for the fused lasso penalty (dark grey) and for the lasso and total variation penalties (light grey).

The joint use of lasso and total variation penalty enforces the signal to be both sparse and piecewise constant. The admissible solution set for the fused lasso penalty is illustrated in Figure 1.15 (dark grey). In light grey, the diagonal band represents the admissible set of $|\beta_2 - \beta_1| \leq c_2$ and the rotated square represents the L_1 ball $|\beta_1| + |\beta_2| \leq c_1$, with $c_1 = 3$ and $c_2 = 1$.

The fused lasso inherits the following properties from the lasso:

(i) It is as computationally efficient, since Friedman et al. (2007) showed that a small modification of the coordinate descent is viable to solve the fused lasso.

(ii) It enforces the same level of sparsity on the parameter than the lasso. Recall that if $p > n$, the lasso selects at most n parameters (see Rosset et al., 2004, Appendix A). For the fused lasso, this property remains true when replacing n by the number of constant plateaux: $\sum_{j=1}^{p} \mathbb{1}_{\beta_j \neq \beta_{j-1}}$ (with $\beta_0^* = 0$).

Two-dimensional fused lasso. Consider now the case where the parameter $\boldsymbol{\beta}$ is parametrized by two indices: $\boldsymbol{\beta} = (\beta_{j,j'})$, where $1 \leq j \leq q$ and , $1 \leq j' \leq q'$, with $q \times q' = p$. In some applications, this parametrization appears naturally in the model, which then has a bi-dimensional structure. Then the fused regularization introduced in the previous section generalizes naturally to the two-dimensional case. Notice that in (1.58), we penalize over the difference operator $\Delta\beta_j \triangleq \beta_j - \beta_{j-1}$ of the parameter. This operator generalizes in two dimensions to the penalty

$$\sum_{j=1}^{q} \sum_{j'=1}^{q'} |\beta_{j,j'} - \beta_{j-1,j'}| + |\beta_{j,j'} - \beta_{j,j'-1}|. \tag{1.59}$$

which defines the two-dimensional total variation penalty As for the one-dimensional case, we can add regularization over β_j directly, so as to form the bi-dimensional equivalent of (1.57), called bi-dimensional fused lasso. Note that there is not a unique difference operator in two dimensions.

This two-dimensional setting arises in image processing: each parameter corresponds to the grey level of one pixel in a numerically-stored image. See Friedman et al. (2007) for an illustration of the bi-dimensional fused lasso, applied to image denoising. Many statistical models dealing with spatial data also falls within the scope of this bi-dimensional parameter structure.

Minimization of (1.59) does not enter in the framework of the coordinate descent, as presented in Property 1. However, the optimization method can be slightly modified so as to solve (1.59), as was shown by Friedman et al. (2007). Optimization methods for the bi-dimensional fused regularization

is not given here in all generality. Instead, we present the computational approach used for the fused adaptive ridge later in this section.

Many extensions of the bi-dimensional fused lasso are possible. First, the L_1 penalty can be generalized to any penalty presented in this Chapter. In particular, this thesis focuses on the adaptive ridge applied on bi-dimensional regularization, which we will call *fused adaptive ridge*. Second, (1.59) makes use of the one-dimensional finite difference operator of the first order. To enforce $\boldsymbol{\beta}$ to be close to linear instead of constant, we would use the penalty

$$4\beta_{j,j'} - \beta_{j-1,j'} - \beta_{j+1,j'} - \beta_{j,j'+1}, \beta_{j,j'-1}$$

or

$$8\beta_{j,j'} - \beta_{j-1,j'} - \beta_{j+1,j'} - \beta_{j,j'+1}, \beta_{j,j'-1} - \beta_{j-1,j'-1} - \beta_{j-1,j'+1} - \beta_{j+1,j'+1} + \beta_{j+1,j'-1}.$$

The two serve the same role, the difference between them being out of the scope of this work. Extensions to any order are available, although the first order given in (1.59) is the most used in many applications.

1.4.5 Fused Adaptive Ridge

The work presented in this thesis makes extensive use of the (L_0) adaptive ridge, and mostly of the fused adaptive ridge, used in Chapters 2 and 3. We recall the method here and discuss some computational considerations that will be developed further in the corresponding Chapters.

Generalizing (1.59) to the adaptive ridge penalty is straightforward. The weighted penalty of the adaptive ridge becomes a weighted difference along each dimension. Thus, we use two arrays of weights: $\boldsymbol{v}$ and $\boldsymbol{w}$.

Algorithm 4 Fused Adaptive Ridge

1: **function** FUSED ADAPTIVE RIDGE(λ)
2: $\quad \boldsymbol{w}^{(0)} \leftarrow \mathbf{1}, \quad \boldsymbol{v}^{(0)} \leftarrow \mathbf{1}$
3: $\quad k \leftarrow 1$
4: $\quad$ **while** not converge **do**
5: $\quad\quad \boldsymbol{\beta}^{(k)} \leftarrow \operatorname{argmin} \ell(\boldsymbol{\beta}) + \lambda \sum_{j=2}^{q} \sum_{j'=2}^{q'} v_{j+1,j'+1}(\beta_{j,j'} - \beta_{j-1,j'})^2 + w_{j+1,j'+1}(\beta_{j,j'} - \beta_{j,j'-1})^2$
6: $\quad\quad \boldsymbol{v}^{(k)}_{j+1,j'+1} \leftarrow \left((\beta_{j,j'} - \beta_{j-1,j'})^2 + \varepsilon^2\right)^{-1}$
7: $\quad\quad \boldsymbol{w}^{(k)}_{j+1,j'+1} \leftarrow \left((\beta_{j,j'} - \beta_{j,j'-1})^2 + \varepsilon^2\right)^{-1}$
8: $\quad\quad k \leftarrow k+1$
9: $\quad$ **end while**
10: **end function**

The method is given in Algorithm 4. Notice that line 6 and 7 are done in complexity $O(q \times q')$. The computational bottleneck is the minimization of the penalized likelihood in line 5. This is typically done using the Newton-Raphson method. This method requires a number of successive inversions of the Hessian (i.e. the second order derivative) of the penalized likelihood. This matrix equals the Hessian of $\ell(\boldsymbol{\beta})$ to which is added the second order derivative of the penalty term (given in line 6). Note by $H(\boldsymbol{\beta})$ this function, whose values are $(qq') \times (qq')$ matrices. Note that the form of the fused penalty induces that H is a (symmetric) band matrix. More precisely, assuming that $\boldsymbol{\beta}$ is ordered $\boldsymbol{\beta} = (\beta_{1,1}, \ldots, \beta_{1,q'}, \beta_{2,1}, \ldots, \beta_{q,1}, \ldots, \beta_{q,q'})$ it has 5 non-zeros diagonals: the main three diagonals, and two diagonals located at distance $\pm q'$ of the main diagonal. Inverting H can be done cleverly by making use of its symmetry and bandedness, using QR decomposition for banded matrices. We have found no way of taking advantage of the fact that most diagonals are zero to speed up this inversion; hence the QR decomposition method inverts H in $O(q \times q \times q')$, which is usually of the same order as $p^{3/2}$.

This remark is only relevant is $\nabla^2 \ell(\boldsymbol{\beta})$ is also banded, with a band no bigger than q'. This turns out to be the case in our applications. More details of computational considerations are given in the corresponding chapters.

1.5 Conclusion

In this introducing chapter I have developed on the many and diverse methods of penalized likelihood estimation that perform selection of variables.

In a first part, I explain in details the simplest and mostly used methods of penalized likelihood: the lasso and its extensions: elastic-net and bridge regression. These methods' have historically been developed from the generalization of the norm penalty: the lasso penalty (L_1 norm), which is often introduced as a relaxation of the L_0 norm, was considered an extension of the ridge penalty (L_2 norm) to enable both shrinkage and sparsity. Most other methods are built as extensions of the L_1 norm, as is the case for the elastic-net and the bridge. A breakthrough was made when Fan and Li (2001) introduced the idea of "non-concave" penalty, generalizing the previous methods to a wide class of sparsity inducing penalties. These non-concave penalties benefit from satisfying asymptotic properties of consistence in selection and asymptotic normality in estimation, with the lowest possible variance. These penalties are hard to minimize, making the computation of their corresponding estimate a difficult problem. Fan and Li (2001) and Zou and Li (2008) introduced two different iterative procedures to minimize these penalties: the LQA and the LLA.

The next Section is dedicated to briefly introducing the other important approach to variable selection: stepwise selection. Instead of enforcing a prior on the distribution of the estimated parameter, like the penalized methods do, the stepwise selection methods (including forward-backward selection) performs variable selection by successively fitting the model with different subset of variables.

In Section 1.3, I resume to penalized methods and I develop on the iterative methods for solving penalized likelihood problems. I also introduce the iterative penalized method that is central to this thesis: the adaptive ridge. I discuss its strengths and particularities and I develop on a particular case of interest: the L_0 adaptive ridge.

Finally in the last section I introduce an important application of sparsity-inducing estimation: structured sparsity. In many applications, the parameter is assumed to have a certain characteristic. Two examples have a historical importance and have played a large role in the development of structured sparsity: *(i)* in image processing, images are assumed to be mainly composed of large regions of equal brightness and *(ii)* in many models, the variables are assumed to belong to groups, and we want to select the variables by groups. Structured sparsity focuses on finding ways to apply the penalized likelihood variable selection methods to a specific function of the parameter.

The work of this thesis makes an extensive use of structure sparsity in different applications. We will consistently use the (L_0) adaptive ridge with a structured sparsity adapted to the problem at hand.

Why the adaptive ridge? Many efficient penalized likelihood methods have been successfully brought to light and the class of non-concave penalties seem to enjoy both optimal theoretical properties and numerical procedures for fast computation. We see the adaptive ridge as a new and interesting approach to iterative model selection methods. Its simplicity of implementation makes it easily applicable to new statistical problems. Indeed, the adaptive ridge procedure is a weighted ridge problem; for which readily available numerical solvers can be used.

The adaptive ridge was introduced by Rippe et al. (2012) and Frommlet and Nuel (2016) with an empirical justification for its formulation. In Chapter 1.3.3, we make an important connection between the adaptive ridge and non-concave penalty. The adaptive ridge is the numerical procedure obtained when applying Fan and Li (2001)'s LQA to a particular penalty, which is a relaxation of the logarithmic function. This provides the pertinent framework for a study of the properties of the adaptive ridge. It can be shown that although the relaxation used to generalize the non-concave penalty to the adaptive ridge procedure makes its computation easier, the theoretical properties of the latter do not apply for the former. We lay the ground for a study of these properties.

Appendix: Least angle regression

The least angle regression (LARS) is an iterative method for penalized regression introduced by Efron et al. (2004). A modified version of the LARS, called lasso-modified LARS or abusively LARS, allows to compute the solution path of the lasso. Indeed, the lasso estimate, as a function of the parameter λ, is piecewise linear. As long as the active variables are the same, the estimates are a linear function of λ. Each time a variable is included in or removed from the model, the slope of the other variables' effects shift. The LARS – and the LARS-lasso – make use of this property: the estimates are only computed at the values of λ were a variable is added to or removed from the model. The entire solution path is then drawn by connecting the dots. In this respect, the LARS-lasso compares favorably with the coordinate descent, which computes the lasso estimate only on a grid of values of λ.

The LARS-lasso is initiated with no "active" variables in the model and it iteratively adds variables one by one in the model. The current estimate of $\boldsymbol{y}$ is noted $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and is initiated at zero. The LARS-lasso operates by selecting the variable with the greatest correlation with the current residuals $\hat{\boldsymbol{c}} = \boldsymbol{X}^T(\boldsymbol{y} - \hat{\boldsymbol{y}})$ and adding is to the active set of variables. The current estimate is then updated in the direction $\boldsymbol{u}$ such that the current residuals have equal correlation with the newly added variable than with the other active variables, that is: $\hat{\boldsymbol{y}} \leftarrow \hat{\boldsymbol{y}} + \hat{\alpha}\boldsymbol{u}$. The value of $\hat{\alpha}$ is chosen such that the correlation between the current residuals and the new variable becomes equal to the correlation between the current residuals and any other non-active variable. The first non-active variable to satisfy this condition is in turn added as an active variable. Then, the current estimate $\hat{\boldsymbol{y}}$ is updated in the direction with equal correlation between the two last variables to be activated. By iteration, the current estimate follows the path of equal correlation – i.e. of equal angle – between all covariates, until they are all added to the model. Hence the name "least angle regression".

The LARS-lasso is the same as the LARS algorithm, except for the following small modification. When moving the current estimate along the least angle direction, if an active variable j makes its estimate cross zero: *(i)* update the current estimate at the corresponding value of $\hat{\alpha}$, *(i)* remove j from the set of active variables, and *(iii)* carry on the LARS procedure. The LARS-lasso was proven by Efron et al. (2004) to make the algorithm compute the lasso solution path, under the condition that no two variables have equal correlation with the current estimate. This condition is always met in practice. The proof is technical and there does not seem to be an intuitive explanation to the connexion between least-angle regression and L_1 norm.

Chapter 2

Regularized estimation of the hazard rate

In this section we introduce a statistical method for hazard rate estimation with automatic detection of breakpoints. This methods uses the adaptive ridge penalization to effectively produce a piecewise constant estimate of the hazard rate, where the cuts are chosen from the data. This method is general and could be applied to a wide range of statistical problems, like density estimation. We only present it in a more specific setting, where instead of estimating a density, we estimate a typical quantity of interest in survival analysis: the hazard rate, which can be seen as a conditional density. This choice is motivated by two reasons. First, this problem was motivated by the study of heterogeneity in survival analysis. Secondly, we present another model in Chapter 3 which answers the problems of age-period-cohort analysis and is built on the method presented in this Chapter. Thus, we choose to present our method in the specific context of survival analysis, and more specifically age-period-cohort analysis.

More precisely, we focus on the estimation of the hazard rate $\lambda(t,u)$ taken as a function of two variables t and u, were t is the time dependent variable and u is another variable. In particular, we present our method in the context of age-period-cohort analysis. We will show that we can consider only the case of two variables: one is time-dependent (either the age or the period) and the other, the cohort (i.e. date of birth), is not time-dependent. For the sake of simplicity, our method is illustrated with u as the cohort variable.

This chapter is organized as follows. In a first part, we give a general introduction to time-to-event data and the statistical study of censored variables. We go on to introduce a simple model for the estimation of the hazard rate in the case of right-censored data. This model is the piecewise constant hazard model, with automatic detection of the breakpoints. This first model was developed by Bouaziz and Nuel (2017). We present it because it provides a good illustration of application of the fused adaptive ridge to hazard estimation. In fact, the model developed in this Chapter is built as an extension of the work from Bouaziz and Nuel (2017). We then present the setting of age-period-cohort analysis, and more specifically, the estimation of the hazard rate in presence of another continuous variable. Finally, Sections 2.2 and 2.3 present our contribution.

Contents

2.1 Introduction 38
2.1.1 Survival Analysis 38
2.1.2 The piecewise constant hazard estimation 40
2.1.2.1 Proportional hazard model with piecewise constant hazard 41
2.1.3 Cohort data 42
2.2 Regularized estimation of the hazard rate 43
2.3 Application to the evolution of breast cancer mortality 66

2.1 Introduction

2.1.1 Survival Analysis

Survival analysis is the statistical study of duration variables, sometimes called *time-to-event* variables. This duration variable is the time between an origin and an *event of interest*. The origin is the moment where the individual starts being at risk of having the event. (In the example where death is the event of interest, the origin is the birth of the individual.) In order to collect time-to-event data we have to wait until the event of interest has occurred. In cases where the duration time is long (as in demographics and epidemiology), studies are stopped before all events have occurred.

Thus, not all the duration times are observed. We can only observe the event of interest (for instance death) if it happened before all events that would make its observation impossible (for instance that the individual exit the epidemiological study). This problem can occur for a large proportion of the sample, for example with demographic studies where the event of interest is death. Removing the corrupted individual from the study and carrying the statistical inference with the remaining sample is a mistake because it would introduce a bias in the observed sample.

This problem is called *censoring*. When the variable cannot be observed after the occurrence of another incompatible event, it is called *right-censoring*. In this section, right-censoring is statistically defined and the assumptions under which inference on the event of interest can be recovered are discussed.

Time-to-event data. Let $T^* \geq 0$ be the time-to-event random variable of interest. In the case of right-censoring, T^* is not observed. Instead we observe (T_i, Δ_i), an i.i.d sample of the observed time T and the right-censoring time C:

$$T \triangleq \min\left(T^*, C\right) \tag{2.1}$$

$$\Delta \triangleq \mathbb{1}_{T^* \leq C} = \mathbb{1}_{T = T^*}. \tag{2.2}$$

The term *right censoring* comes from the fact it is the higher values of T^* that are censored by C – and hence the *right* part of the distribution. To observe Δ_i is to know whether we observe T_i^* or C_i.

Terminology. We provide and recall a few terms used in time-to-event data. The *event of interest* is the event which marks the end point of the duration T^* at stake. It is typically death of the onset of a disease. A duration T^i is said to be *observed* if $\Delta_i = 1$, otherwise, it is said to be *censored* (and $\Delta_i = 0$). An individual is said to be *at risk* at a time t if $T_i > t$, that is, if he is at risk of having the event of interest.

Survival function. The survival function is defined as

$$S(t) \triangleq \mathbb{P}(T > t) = 1 - F(t^-), \tag{2.3}$$

where F is the cumulative probability distribution function of the variable of interest. The survival function at *time t* gives the probability of survival at time t – i.e. the probability that the individual has not had the event of interest at time t.

Hazard Rate. The hazard rate (or instantaneous hazard) rate is defined as the infinitesimal probability of the event to occur now, conditionally on the fact that it has not occurred yet:

$$\lambda^*(t) \triangleq \lim_{\delta t \to 0} \frac{\mathbb{P}(t \leq T^* < t + \delta t \mid T^* \geq t)}{\delta t}. \tag{2.4}$$

This function is important in survival analysis for (i) theoretical reasons as is illustrated in the next paragraph (ii) practical reasons because it has an important meaning in demographics and epidemiology.

Many simple statistical models are formulated using the hazard rate rather than the survival function. Denote by f the density of T^*; then simple calculations yield $\lambda(t) = f(t)/S(t)$. This equation shows that there is a one-to-one correspondence between a distribution and its hazard rate function (if it exists).

Assumptions on the censoring variable. Define the following assumptions.

Assumption 3 (Independent censoring)**.** *The random variables C and T^* are independent*

Assumption 4 (Non-informative censoring)**.** *The censoring variable C does not depend on the model parameters*

Inference under right-censoring. Define the *crude* hazard rate as:

$$\lambda(t) \triangleq \lim_{\delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \delta t \mid T \geq t, C \geq t)}{\delta t}. \tag{2.5}$$

We recall the following property (Fleming and Harrington, 2011, p. 25).

Property 4. *Under Assumption 3, $\lambda(t) = \lambda^*(t)$.*

This property is essential for inference with censored data. Indeed, the crude hazard depends on T and C and not on T^*. Consequently, under Assumption 3, inference on T^* is straightforward and we can perform inference on T^* (through its hazard rate) without directly observing it.

This assumption is not equivalent to Property 4. It is stronger than necessary and we can construct examples of dependent censoring where $\lambda = \lambda^*$ (Fleming and Harrington, 2011, Exercice 1.8). The precise assumption that is equivalent to Property 4 is

$$\lim_{\delta t \to 0} \frac{1}{\delta t} \mathbb{P}\left(t \leq T^* < t + \delta t | T^* \geq s\right) = \lim_{\delta t \to 0} \frac{1}{\delta t} \mathbb{P}\left(t \leq T^* < t + \delta t | T^* \geq t, C \geq t\right), \tag{2.6}$$

for all t such that $\mathbb{P}(T^* > u) > 0$ (see Fleming and Harrington, 2011, , Theorem 1.3.1). This equality signifies that C does not convey any information supplementary on T^*. However, for the sake of simplicity, we assume independent censoring and not the latter, weaker assumption. This is a simplification often made in the literature.

Moreover, wherever necessary, it is implicitly assumed that T^* and C are absolutely continuous with respect to the Lebesgue measure.

Remark on censoring and competing risk. As explained previously, the censoring is due to the individual leaving the study. We make remark here to note that not all events preventing the event of interest from being observed are censoring events. Take the example where the event of interest is the birth of the first child. If the individual leaves the study before his first child is born, this individual is censored. But the death of the individual is not a censoring variable, because after dying, the individual is no longer at risk. Censoring requires that the individual still be at risk of having the event of interest. The question "is the individual still at risk ?" allows to understand whether an alternative event is a censoring event or a competing risk. This simple question is often overlooked in applications.

2.1.2 The piecewise constant hazard estimation

The piecewise constant hazard model. In this section we introduce a simple model upon which this chapter's work is built. Let $\boldsymbol{c} = (c_1, \dots, c_J)$ be a vector of positive-valued *cuts* sorted in increasing order: $c_1 < \cdots < c_J$. In the piecewise constant hazard model the hazard is constant on the intervals defined by $\boldsymbol{c}$:

$$\lambda(t) = \sum_{j=1}^{J} \exp(\alpha_j) \mathbb{1}_{c_{j-1} \le t < c_j}, \tag{2.7}$$

where α_j are the values of the log-hazard rate over each interval and where $c_0 = 0$ and $c_{J+1} = \infty$ by convention. The logarithmic transformation ensures that the parameter $\boldsymbol{\alpha}$ can take any real value, and thus, no constraints need to be added to its estimation. The parameter of the model is $\boldsymbol{\alpha} \in \mathbb{R}^J$.

The interest of the piecewise constant hazard model comes from its simplicity. Indeed the hazard rate is directly interpretable and in many applications (e.g. medicine or epidemiology), the practitioner is interested in a simple hazard function in order to draw conclusions. On the other hand, with sufficiently enough cuts, this model is very flexible. One of the aim of this chapter is to choose the number and values of the cuts from the data. In this manuscript we use the acronym "PCH" to denote the piecewise constant hazard function.

In this section we will discuss the estimation in the PCH model with an automatic detection of the knots. This gives a good introduction to the contribution of this chapter, which is estimation of bi-dimensional hazard rate with automatic selection of the knots. This work was developed by Bouaziz and Nuel (2017) who also introduced a procedure for inferring $\boldsymbol{\alpha}$.

Maximum likelihood estimation. Under Assumptions 3 and 4, the likelihood writes

$$L(\boldsymbol{\alpha}) = \prod_{i=1}^{n} f(T_i)^{\Delta_i} S(T_i)^{1-\Delta_i}. \tag{2.8}$$

This product has n terms: the information provided by individual i is $f(T_i)$ if the event is observed (i.e. $\Delta_i = 1$) and $S(T_i) = P(T^* > T_i)$ if the event is censored (i.e. $\Delta_i = 0$). The NLL $\ell = -\log L$ writes

$$\ell(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \left\{ \int_0^{T_i} \lambda(t)\, dt - \Delta_i \log(\lambda(T_i)) \right\}. \tag{2.9}$$

Combining (2.7) into (2.9), we get

$$\begin{aligned} \ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n} \int_0^{T_i} \lambda(t) dt - \sum_{i=1}^{n} \Delta_i \log(\lambda(T_i)) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{J} e^{\alpha_j} \int_0^{T_i} \mathbb{1}_{c_{j-1} \le t < c_j} dt - \sum_{i=1}^{n} \sum_{j=1}^{J} \alpha_j \Delta_i \mathbb{1}_{c_{j-1} \le T_i < c_j}. \end{aligned}$$

Define

$$R_{i,j} \triangleq \int_0^{T_i} \mathbb{1}_{c_{j-1} \le t < c_j} dt = \mathbb{1}_{T_i \ge c_{j-1}} (\max(T_i, c_j) - c_{j-1}) \tag{2.10}$$

and

$$O_{i,j} \triangleq \Delta_i \mathbb{1}_{c_{j-1} \leq T_i < c_j}. \tag{2.11}$$

The former quantity is equal to the time spent by individual i in the interval $[c_{j-1}, c_j)$. It is called *time at risk*. The latter quantity is equal to the number of observed events in the interval $[c_{j-1}, c_j)$ – note that the censored events are not taken into account here. Then

$$\ell(\boldsymbol{\alpha}) = \sum_{j=1}^{J} \exp(\alpha_j) R_j - \alpha_j O_j, \tag{2.12}$$

where $R_j = \sum_{i=1}^{n} R_{i,j}$ is the *total time at risk* and $O_j = \sum_{i=1}^{n} O_{i,j}$ is the *total number of events*, both in the interval $[c_{j-1}, c_j)$.

A few remarks need to be made at this point. The two quantities $(O_j)_j$ and $(R_j)_j$ are exhaustive statistics. They gather all the information from the data $(T_i, \Delta_i)_i$ with respect to the PCH model. They have a crucial role in any survival model with a piecewise constant hazard. When there is no ambiguity, they will be referred to as the "exhaustive statistics".

Moreover, $\ell(\boldsymbol{\alpha})$ writes as a very simple function of the exhaustive statistics. In fact Equation 2.12 is also the likelihood of the Poisson regression model whose response variable is $O_j \sim \mathscr{P}(\mu_j)$ with mean $\mu_j = \log(\alpha_j) R_j$. For this reason, the PCH model is often called “Poisson model”, even though it does not assume that the O_js be Poisson distributed or that the R_js be fixed quantities. Notice that the MLE is explicit here, since the minimum of $\ell(\boldsymbol{\alpha})$ is

$$\alpha_j^{\text{mle}} = \log\left(\frac{O_j}{R_j}\right).$$

Fused penalization using the adaptive ridge. The former considerations yield an explicit maximum likelihood estimate of the hazard in the PCH model. We want to regularize over the successive differences of α_j. Define the penalized negative log-likelihood corresponding to the fused adaptive ridge regularization:

$$\ell^{\text{pen}}(\boldsymbol{\alpha}, \boldsymbol{w}) \triangleq \ell(\boldsymbol{\alpha}) + \lambda \sum_{j=1}^{J-1} w_j (\alpha_{j+1} - \alpha_j)^2, \tag{2.13}$$

where $\lambda > 0$ is a trade-off constant and $(w_j)_{1 \leq j \leq J-1}$ is the vector of weights used in the adaptive ridge.

The penalized estimate is obtained by iterating between (2.13) and an update of the weights $\boldsymbol{w}$, as in Algorithm 4. The principle of the estimation follows the iterative procedure of the adaptive ridge, which is given in Section 1.3.3.

2.1.2.1 Proportional hazard model with piecewise constant hazard

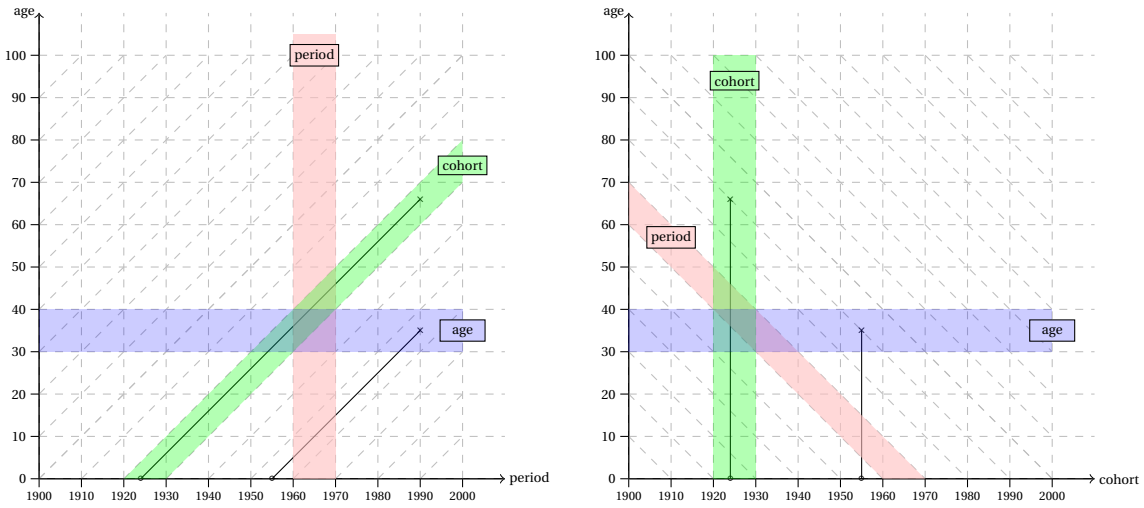
The previous model generalizes easily to the setting of right censored data with covariate, that is, the data consists of $(T_i, \Delta_i, X_i)_{1 \leq i \leq n}$, where $\boldsymbol{x}_i$ is a $p \times 1$ vector of covariates. We assume that the covariate are not time-dependent. Consider now the PCH model with proportional effect of the covariates, that is:

$$\lambda(t|\boldsymbol{x}_i) = \lambda_0(t) \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}), \tag{2.14}$$

where $\lambda_0(t)$ is the piecewise constant baseline hazard (see Equation 2.7) and $\boldsymbol{\beta} = (\beta_j)_{1 \leq j \leq p}$ is the p-length parameter of the covariates' effect. Define $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ the $(J+p)$-length parameter vector of this model: $\boldsymbol{\alpha}$ represents the values of the hazard rate across time and $\boldsymbol{\beta}$ represents the proportional effect due to the covariates.

Equation 2.14 is called proportional hazard model. In this model, the parameter $\boldsymbol{\beta}$ has an important interpretation. Assume for the sake of the illustration that the j-th variable is binary. Then under this model, two individuals whose covariates are equal except the j-th have proportional hazard rates:

$$\lambda(t|x_{i,j} = 1) = \lambda(t|x_{i',j} = 0) \exp(\beta_j).$$



(a) Lexis Diagram: age-period

(b) Lexis Diagram: age-cohort

Figure 2.1: Lexis Diagrams in the Age-Period plane (a) and in the Age-Cohort plane (b).

In this model, $\log \beta_j$ can be interpreted as the multiplicative effect of the j-th covariate on the baseline hazard rate $\lambda_0(t)$. We choose to model a proportional effect of the covariates because it offers a simple model with highly interpretable parameter $\boldsymbol{\beta}$. Unlike Cox's model (Cox, 1972) – which is semiparametric – the baseline hazard is specified here, and the model is fully parametric.

2.1.3 Cohort data

In demographics and epidemiology, an important problem is the study of the evolution of the hazard rate of death (or disease onset) with respect to time. But there are different time scales that we can consider.

- the *cohort* refers to the date of the time origin.
- the *period* refers to the calendar time (i.e., the date).
- The *age* refers to the time spent since the time origin.

When the event of interest is the death from a specific disease, the time origin is the onset of the disease and the *age* variable is the time since the disease onset.

When the event of interest is death or the onset of a disease, the time origin is the birth of each individual. Then, the *age* variable is the age of the individual and the cohort variable is the date of birth of the individual.

The two time-dependent variables are *period* and *age*. In all cases, we have the linear relation:

$$\text{period} = \text{cohort} + \text{age}. \tag{2.15}$$

This section introduces inference of the hazard rate, when one considers more than one of these three variables.

The Lexis Diagram. Consider Figure 2.1a, called the Lexis diagram. It represents the plane with the age in the y-axis and the period in the x-axis. The life of 2 individuals is represented by the black solid lines. These lines are oblique at a 45-degree angle: at the increment of each year, the age of every person increments by one year. In the Lexis diagram, the time origin is the birth of the individual and the lines represent the time spent at risk. The event of interest is represented by the dot at the end

of each line. There is another variable that is hidden in these two variables, the date of birth of each individual, or *cohort*, which is the x-intercept of the lines.

Figure 2.1b represents the age-cohort plane: the only time-dependent variable is age, on the y-axis. The cohort is on the x-axis, so that the life of the individuals are vertical lines, growing upwards as time passes. The same two individuals represented in Figure 2.1a are also represented here.

The age-cohort plane may be easier to represent the evolution of the time at risk, because they are represented by vertical lines. But the Age-period plane is a more natural presentation of the problem, because the period appears directly in most time-to-event data bases. Indeed, cohort studies have a starting date and an end date, and consequently the times for the event are comprised between two limiting periods. Between these two periods, events are registered for individuals with any age, and any date of birth (i.e. cohort). Consequently, cohort study data are represented as events comprised in a rectangle in the age-period plane.

Inference of the hazard rate. We present here the different approaches taken to infer the hazard rate over the Lexis diagram.

Age-period-cohort analysis can be interested in estimating the hazard as a bivariate function, defined in the following way:

$$\lambda(t|u) = \lim_{\delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \delta t | T \geq t, U = u)}{\delta t}, \tag{2.16}$$

where U is the cohort variable.

An important approach is that of the factor models. The factor models assume a discretization of the variables: suppose the age and cohort variables are discretized into J and K intervals respectively. Then the period variable is naturally divided into $J+K$ intervals (see Figure 2.1b, where $J+K=10$ and we count 20 period intervals). Consider the discretized hazard rate $(\lambda_{j,k})_{j,k}$ obtained by discretizing (2.16). The factors model assume that the age, cohort, and period variables have an additive effect on $\log(\lambda_{j,k})$:

$$\log(\lambda_{j,k}) = \alpha_j + \beta_k + \gamma_{j+k}, \tag{2.17}$$

where the parameter vectors $\boldsymbol{\alpha}$, β_k, and $\boldsymbol{\gamma}_{j+k}$ are the effects of each variable. Equation (2.17) is called the *age-period-cohort* model. Because of the linear dependency between the three variables (Equation 2.15), this model is not identifiable (Clayton and Schifflers, 1987) and arbitrary constraints need to be added. We also define three submodels of (2.17): the age-period ($\log\lambda_{j,k} = \alpha_j + \gamma_{j+k}$), age-cohort ($\log\lambda_{j,k} = \alpha_j + \beta_k$), and period-cohort ($\log\lambda_{j,k} = \beta_k + \gamma_{j+k}$) models, which are identifiable. The factors models denote the age-period-cohort model and its submodels. They play an important role in epidemiology (see Section 3.1) but the additive effect that they assume on the variables can be too restrictive.

In this chapter, we focus on developing a regularized estimation of the discretized hazard. In the next chapter, we develop an extension of the factor models that uses regularization to infer the effect of two variables and the interaction between the two effects. Both chapters rely on the application of the adaptive ridge to the discretized hazard rate.

2.2 Regularized estimation of the hazard rate

In this section, I include the preprint entitled “Regularized Bidimensional Estimation of the Hazard Rate”. This work was conducted with my PhD supervisors and with the help of Jean-Christophe Thalabard.

Regularized Bidimensional Estimation of the Hazard Rate

Vivien Goepp[1], Jean-Christophe Thalabard[1], Grégory Nuel[2], and Olivier Bouaziz[1]

[1]*MAP5 (CNRS UMR 8145, 45, rue des Saints-Pères, 75006 Paris)*
[2]*LPSM (CNRS UMR 8001, 4, Place Jussieu, 75005 Paris)*

May 2018

Abstract

In epidemiological or demographic studies, with variable age at onset, a typical quantity of interest is the incidence of a disease (for example the cancer incidence). In these studies, the individuals are usually highly heterogeneous in terms of dates of birth (the cohort) and with respect to the calendar time (the period) and appropriate estimation methods are needed. In this article a new estimation method is presented which extends classical age-period-cohort analysis by allowing interactions between age, period and cohort effects. In order to take into account possible overfitting issues, a penalty is introduced on the likelihood of the model. This penalty can be designed either to smooth the hazard rate or to enforce consecutive values of the hazards to be equal, leading to a parsimonious representation of the hazard rate. The method is evaluated on simulated data and applied on breast cancer survival data from the SEER program.

Keywords Survival Analysis, Penalized Likelihood, Piecewise Constant Hazard, Age-Period-Cohort Analysis, Adaptive Ridge Procedure

Introduction

In epidemiological or demographic studies, with variable age at onset, a typical quantity of interest is the incidence or the hazard rate of a disease (for example the cancer incidence). In these studies, individuals are recruited and followed-up during a long period of time,

usually from birth. The data are then reported either in the form of registers, which contain the number of observed cases and the number of individuals at risk to contract the disease, or in the form of the observed time for each individual. These types of studies are of great interest for the statistician, especially when the event of interest will tend to occur at late ages, such as in cancer studies. However, these data are usually highly heterogeneous in terms of dates of birth and with respect to the calendar time. In such cases, it is therefore very important to take into account the variability of the age, the cohort (date of birth) and the period (the calendar time) in the hazard rate estimation. This is usually done using age-period-cohort estimation methods (see Yang and Land, 2013, and citations therein).

In age-period-cohort analysis, the effects of age, period and cohort are fit as factor variables in a regression model where the output is the logarithm of the hazard rate. However, this induces an identifiability problem due to the relationship: period = age + cohort. There have been several solutions proposed to this problem. Osmond and Gardner (1982) proposed to compute each submodel (age-cohort, age-period, and period-cohort) and use a weighting procedure to combine the three models. Different constraints have also been proposed to make the age-period-cohort model identifiable. However, as noticed by Heuer (1997, p 162), the obtained estimates highly depend on the choice of the constraints. Holford (1983) proposed to directly estimate the linear trends of each effect. This procedure leads to results that are difficult to interpret. See Carstensen (2007) for a detailed discussion of the identifiability problem of the age-period-cohort model. More recently, Kuang et al. (2008) proposed to estimate the second order derivatives of the three effects. This model is implemented in the package `apc` Nielsen (2015). Finally, Carstensen (2007) proposed to first fit one submodel (say age-cohort) and then to fit the period effect over the residuals of the first model. This model is implemented in the `R` package `Epi` (Carstensen et al., 2017), Plummer and Carstensen (2011).

All these approaches can be viewed as parametric models, where the parameters are the age, period, and cohort vector parameters. As such they are also restrictive because they do not allow for interactions between the three effects, that is they assume that one effect does not depend on the other effect's value. A different approach consists in considering the hazard rate as a function of age and either period or cohort and to estimate this bi-dimensional function in a non-parametric setting. No specific structure of the hazard rate is assumed. However, for moderate sample sizes, non-parametric approaches are prone to overparametrization. As a consequence, regularized methods have been proposed in order to avoid overfitting in this non-parametric context. A kernel-type estimator was proposed by Beran (1981) and McKeague and Utikal (1990) where the cumulative hazard is smoothed using a kernel function. See Keiding (1990) for a thorough discussion of methods for hazard inference in age-period-cohort analysis. More recently, Currie and Kirkby (2009) proposed a spline estimation procedure to infer the hazard rate as a function of two variables. The authors use a generalized linear model using B-splines and overfitting is dealt with using a penalization over the differences of adjacent splines' coefficients.

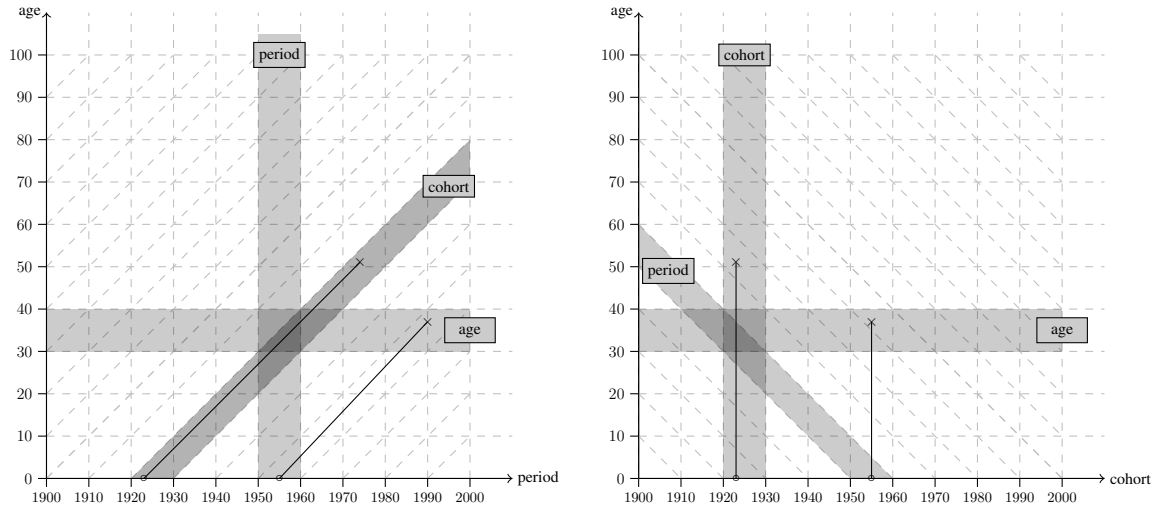
In this article, we propose a new non-parametric method for bi-dimensional hazard rate estimation. As the previous non-parametric approaches, this model considers the estimation of the hazard rate with respect to two variables, i.e. either age-cohort, age-period, or period-cohort, without assuming any specific structure on the hazard rate. Inference is made in two dimensions, but through the linear relationship period = age + cohort, the hazard rate can be represented as a function of any two of the three variables. Finally, in order to take into account the issue of overfitting, we use the L_0 penalization procedure introduced by Rippe et al. (2012), Frommlet and Nuel (2016), and Bouaziz and Nuel (2017). This penalty offers a segmentation of the hazard rate into constant areas. It makes use of an approximation of the L_0 norm which is computationally tractable. The novelty of this method lies in the parsimonious representation of the bi-dimensional hazard rate into segmented areas. In particular, the method can efficiently exhibit cohort, age or period effects, that is, specific changes of the hazard rate due to the date of birth, the age or the calendar time. Our approach also allows L_2 norm penalization, which will induce a smoothed estimate of the hazard in a similar way as the aforementioned non-parametric methods.

Our model is introduced in Section 1. The regularization method is then presented in Section 2. In Section 3, the penalty term selection problem is discussed. Finally, the performance of our model is assessed through a simulation study in Section 4 and illustrations on the SEER cancer dataset is provided in Section 5.

1 Modeling strategy

In the age-period-cohort setting, the date of birth (the cohort) U of each individual is available and the variable of interest is a time-to-event variable of this individual denoted T. The data are subject to right-censoring and they are represented as tabulated data over the J cohort intervals and the K age intervals $[c_0, c_1), [c_1, c_2), \ldots, [c_{J-1}, c_J)$ and $[d_0, d_1), [d_1, d_2), \ldots, [d_{K-1}, d_K)$ respectively, with the convention $c_0 = d_0 = 0$ and $c_J = d_K = \infty$. On a sample of n individuals, the available data can then be rewritten in terms of the exhaustive statistics $\boldsymbol{O} = (O_{1,1}, \ldots, O_{J,K})$, $\boldsymbol{R} = (R_{1,1}, \ldots, R_{J,K})$, where for $j = 1, \ldots, J$, $k = 1, \ldots, K$, $O_{j,k}$ represents the number of observed events that occurred in the j-th cohort interval $[c_{j-1}, c_j)$ and k-th age interval $[d_{k-1}, d_k)$ and $R_{j,k}$ represents the total times individuals were at risk in this j-th cohort and k-th age interval. In the case of register data, the discretization $(c_j), (d_k)$ is imposed by the data and the available data is directly $\boldsymbol{R}$ and $\boldsymbol{O}$, which are often called the *cases* and *person-years*, respectively. See for instance Carstensen (2007) for an example of such data. The aim is to use the available data to provide an estimator of the hazard rate, defined in the age-cohort setting as:

$$\lambda(t|u) = \lim_{dt \to 0} \frac{1}{dt} \mathrm{P}(t < T < t + dt | T > t, U = u),$$



(a) Lexis diagram: Age-Period diagram

(b) Age-Cohort diagram

Figure 1: Diagrams representing the lives of individuals: in the age-period plane (a) – called Lexis diagram – and in the age-cohort plane (b). Solid lines represent lives of individuals until occurrence of the event of interest. The same age, cohort, and period intervals are displayed in light gray. The intersection of two intervals forms a parallelogram and the intersection of three intervals forms a triangle.

in the situation where $\lambda(t,u)$ is assumed to be piecewise constant. That is, we assume that

$$\lambda(t|u) = \sum_{j=1}^{J}\sum_{k=1}^{K} \lambda_{j,k}\mathbb{1}_{[c_{j-1},c_j)\times[d_{k-1},d_k)}(t,u),$$

and inference is made over the $J \times K$ dimension parameter $\boldsymbol{\lambda} = (\lambda_{1,1}, \ldots, \lambda_{J,K})$. Note that the hazard can be equivalently defined as a function of age and period or as a function of period and cohort where the period is defined as the calendar time, that is: period = cohort + age. For illustration, the change of coordinates between the age-period and age-cohort diagrams is represented in Figure 1. In our models, the hazard will be considered as a function of solely age and cohort since the influence of any of the two elements of age, period or cohort can be retrieved using this reparametrization.

Following Aalen et al. (2008, p. 224) the negative log-likelihood takes the form

$$\ell_n(\boldsymbol{\lambda}) = \sum_{j=1}^{J}\sum_{k=1}^{K}\{\lambda_{j,k}R_{j,k} - O_{j,k}\log(\lambda_{j,k})\}. \tag{1}$$

The authors also noticed that this log-likelihood is equivalent to a log-likelihood arising from a Poisson model. However, note that no distribution assumptions are made on the data and in particular the $O_{j,k}$ are not assumed to be Poisson distributed (see Carstensen, 2007, for a discussion on the "Poisson" model). Minimizing ℓ_n yields an explicit maximum likelihood estimator $\widehat{\lambda}^{\text{mle}}_{j,k} = O_{j,k}/R_{j,k}$. However, for moderate sample sizes this estimator is overfitted, especially in places of the age-cohort plane where few events are recorded. To remedy this problem we propose in the following to penalize the differences between adjacent values of the hazard in the log-likelihood.

For computation convenience, we first reparametrize the model: $\eta_{j,k} = \log\lambda_{j,k}$, for $1 \le j \le J$ and $1 \le k \le K$. The estimate is obtained by minimizing the penalized function

$$\ell_n^{\kappa}(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w}) = \ell_n(\boldsymbol{\eta}) + \frac{\kappa}{2}\sum_{j=1}^{J-1}\sum_{k=1}^{K} v_{j,k}\left(\eta_{j+1,k} - \eta_{j,k}\right)^2 + \frac{\kappa}{2}\sum_{j=1}^{J}\sum_{k=1}^{K-1} w_{j,k}\left(\eta_{j,k+1} - \eta_{j,k}\right)^2, \tag{2}$$

where $\ell_n(\boldsymbol{\eta})$ was defined in (1), κ is a penalty constant used as a tuning parameter, and $\boldsymbol{v} = (v_{1,1}, \ldots, v_{J-1,K})$, $\boldsymbol{w} = (w_{1,1}, \ldots, w_{J,K-1})$ are constant positive weights of respective dimensions $(J-1)K$ and $J(K-1)$. Note that the case $\kappa = 0$ corresponds to the maximum likelihood estimation and the case $\kappa = \infty$ corresponds to a hazard uniformly constant over the age and cohort intervals. The parameter κ needs to be chosen in an appropriate way in order to obtain a compromise between these two extreme situations.

This model does not attempt to estimate the age, period and cohort effect as parameter vectors. Instead, it performs a regularized estimation of $\boldsymbol{\lambda}$ that has no age-period-cohort-type structure. Two choices for the weights $\boldsymbol{v}$ and $\boldsymbol{w}$ can be made: one will lead to a

smooth hazard rate and the other to a segmented hazard rate. This will be discussed in the next section. The choice of the optimal value for κ is addressed in Section 3.

Minimization of ℓ_n^κ is performed using the Newton-Raphson algorithm (see Algorithm 1). Let $U_n^\kappa(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w}) = \partial \ell_n^\kappa / \partial \boldsymbol{\eta}$ be the gradient vector of the negative log-likelihood and $I_n^\kappa(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w}) = \partial U_n^\kappa(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w}) / \partial \boldsymbol{\eta}^T$ be its Hessian matrix.

For $1 \leq j, j' \leq J$ and $1 \leq k, k' \leq K$, we have

$$\frac{\partial \ell_n}{\partial \eta_{j,k}}(\boldsymbol{\eta}) = \exp(\eta_{j,k}) R_{j,k} - O_{j,k}, \qquad \frac{\partial^2 \ell_n(\boldsymbol{\eta})}{\partial \eta_{j',k'} \partial \eta_{j,k}} = \mathbb{1}_{j=j',k=k'} \exp(\eta_{j,k}) R_{j,k}, \quad \text{and}$$

$$\begin{aligned}\frac{\partial \ell_n^\kappa}{\partial \eta_{j,k}}(\boldsymbol{\eta}) = \frac{\partial \ell_n(\boldsymbol{\eta})}{\partial \eta_{j,k}} &+ \kappa \left[-v_{j,k} (\eta_{j+1,k} - \eta_{j,k}) + v_{j-1,k} (\eta_{j,k} - \eta_{j-1,k}) \right] \\ &+ \kappa \left[-w_{j,k} (\eta_{j,k+1} - \eta_{j,k}) + w_{j,k-1} (\eta_{j,k} - \eta_{j,k-1}) \right],\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ell_n^\kappa(\boldsymbol{\eta})}{\partial \eta_{j',k'} \partial \eta_{j,k}} = \frac{\partial^2 \ell_n}{\partial \eta_{j',k'} \partial \eta_{j,k}}(\boldsymbol{\eta}) &+ \kappa \left[\mathbb{1}_{j=j',k=k'} \left(v_{j',k'} + v_{j'-1,k'} + w_{j',k'} + w_{j',k'-1} \right) \right. \\ &- v_{j',k'} \mathbb{1}_{j=j'+1,k=k'} - v_{j'-1,k'} \mathbb{1}_{j=j'-1,k=k'} \\ &\left. - w_{j',k'} \mathbb{1}_{j=j',k=k'+1} - w_{j',k'-1} \mathbb{1}_{j=j',k=k'-1} \right].\end{aligned}$$

As a consequence, the Hessian matrix can be written

$$I_n^\kappa(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w}) = \frac{\partial^2 \ell_n(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} + \kappa B(\boldsymbol{\eta}),$$

where $B(\boldsymbol{\eta})$ is a band matrix of bandwidth equal to $\min(J, K) - 1$. Thus the Hessian matrix has the same structure as $B(\boldsymbol{\eta})$ and the calculation of $I_n^\kappa(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w})^{-1} U_n^\kappa(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w})$ has a $\mathcal{O}(\min(J, K) JK)$ complexity instead of $\mathcal{O}(J^3 K^3)$. Fast inversion of the Hessian matrix is done using Cholesky decomposition as implemented in `Rcpp` in the package `bandsolve` [1].

2 Choice of the regularization parameters v and w

In this section, two different expressions of the weights $\boldsymbol{v}$ and $\boldsymbol{w}$ are proposed which correspond to two different types of regularization of the hazard rate. The first one yields a smooth estimate. The second one uses an iterated adaptation of the weights to approximate an L_0 norm penalization of the first order differences.

[1] http://github.com/Monneret/bandsolve

Algorithm 1 Newton-Raphson Procedure with constant weights

function NEWTON-RAPHSON($\boldsymbol{O}, \boldsymbol{R}, \kappa, \boldsymbol{v}, \boldsymbol{w}$)
  $\boldsymbol{\eta} \leftarrow \mathbf{0}$
  while not converge **do**
    $\boldsymbol{\eta}^{\text{new}} \leftarrow \boldsymbol{\eta} - I_n^{\kappa}(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w})^{-1} U_n^{\kappa}(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w})$
    $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta}^{\text{new}}$
  end while
  return $\boldsymbol{\eta}$
end function

2.1 L_2 Norm Regularization

A ridge-type penalization is performed when setting $\boldsymbol{v} = \boldsymbol{w} = \mathbf{1}$. In this case the penalization corresponds to the square of the first-order differences of $\boldsymbol{\delta}$. In the penalized estimation model, this choice of weights yields a globally smooth estimator of the hazard rate. Note that our penalized maximum likelihood model will yield similar results as the spline method of Ogata and Katsura (1988) presented in Section 1. In our method the penalization is performed over the first order differences of the parameter while in the spline method it is performed over the second order differences. This means that for arbitrarily large values of the penalty constant, the regularized hazard will be a constant function instead of a linear function. This model will be referred to as L_2 regularized estimation or smooth estimation.

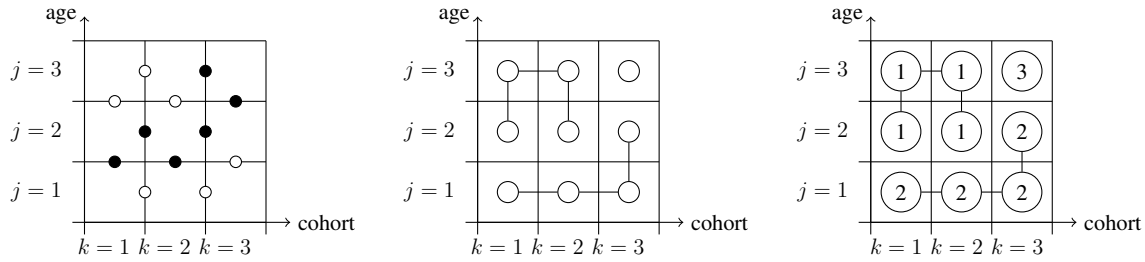
Finally, one notes that Equation 2 allows for some flexibility in the regularization. Indeed, manually setting the weights $\boldsymbol{v}$ and $\boldsymbol{w}$ will allow to tune the importance of the regularization between different regions of the plane and between the two variables.

2.2 Approximate L_0 Norm Regularization

Following the work from Rippe et al. (2012), Frommlet and Nuel (2016), and Bouaziz and Nuel (2017) an adaptive ridge procedure is performed when the weights are updated at each iteration of the Newton-Raphson algorithm. At the m-th iteration of the Newton-Raphson algorithm the weights are computed from the following formulas:

$$\begin{cases} v_{j,k}^{(m)} = \left(\left(\eta_{j+1,k}^{(m)} - \eta_{j,k}^{(m)} \right)^2 + \varepsilon_v^2 \right)^{-1}, \\ w_{j,k}^{(m)} = \left(\left(\eta_{j,k}^{(m)} - \eta_{j,k-1}^{(m)} \right)^2 + \varepsilon_w^2 \right)^{-1}, \end{cases}$$

where ε_v and ε_w are constants negligible compared to 1 (in practice one typically chooses $\varepsilon_v = \varepsilon_w = 10^{-5}$). We iterate between minimizing ℓ_n^{κ} for fixed weights and reevaluat-



(a) Weighted differences (b) Corresponding graph (c) Connected components

Figure 2: Representation of the method used to select the constant areas for the adaptive ridge procedure. In this example, $J = K = 3$. In Panel (a), the circles represent the values of the differences $v_{j,k} \left(\eta_{j+1,k} - \eta_{j,k}\right)^2$ and $w_{j,k} \left(\eta_{j,k+1} - \eta_{j,k}\right)^2$: empty circles correspond to the value 0 and filled circles correspond to the value 1. Panel (b) represents the graph that is generated from these values. Adjacent nodes whose difference is null are connected by a vertice. Panel (c) represents the last step, where the connected components of the graph are extracted. Each connected component corresponds to one constant area. The numbering is arbitrary.

ing the weights such that at the m-th step, $v_{j,k}^{(m)}(\eta_{j+1,k}^{(m)} - \eta_{j,k}^{(m)})^2 \simeq \|\eta_{j+1,k}^{(m)} - \eta_{j,k}^{(m)}\|_0$ and $w_{j,k}^{(m)}(\eta_{j,k+1}^{(m)} - \eta_{j,k}^{(m)})^2 \simeq \|\eta_{j,k+1}^{(m)} - \eta_{j,k}^{(m)}\|_0$, where $\|.\|_0$ denotes the L_0 norm – i.e. $\|u\|_0 = 0$ if $u = 0$ and $\|u\|_0 = 1$ otherwise. In other words, this adaptive ridge procedure approximates the L_0 norm regularization over the differences of $\eta_{j,k}$ and yields a segmentation of $\eta_{j,k}$ into piecewise constant areas. As with other classical penalized methods (e.g. LASSO, ridge) and as pointed out in Frommlet and Nuel (2016), the adaptive ridge penalization scheme induces a shrinkage bias. Therefore, after segmentation of the $\eta_{i,j}$s, the hazard rate is estimated on each constant area using the unpenalized maximum likelihood estimator. More precisely, at convergence of the adaptive ridge algorithm, $v_{j,k} \left(\eta_{j+1,k} - \eta_{j,k}\right)^2$ will be approximately equal to 0 if $|\eta_{j+1,k} - \eta_{j,k}|$ is smaller than ε_v and approximately equal to 1 if $|\eta_{j+1,k} - \eta_{j,k}|$ is greater than ε_v – and similarly for $w_{j,k} \left(\eta_{j,k+1} - \eta_{j,k}\right)^2$. Then one creates the graph whose vertices are the JK discretization cells and whose edges are the connexions between adjacent cells that have differences close to 0. Each connected component of this graph is a different area over which the hazard has been estimated to be constant. The extraction of connected components from the graph is done using the package `igraph` (Csardi and Nepusz, 2006). The log-hazard $\eta^{(r)}$ of the r-th constant area is such that $\forall [c_{j-1}, c_j) \times [d_{k-1}, d_k) \in r, \eta_{j,k} = \eta^{(r)}$. Finally, the values of $\eta^{(r)}$ are not estimated using the results of the adaptive ridge algorithm, but by unpenalized maximum likelihood estimation: $\widehat{\eta}^{(r)} = \log\left(O^{(r)}/R^{(r)}\right)$ where $O^{(r)}$ is the number of events in the r-th constant area and $R^{(r)}$ is the time at risk in the r-th constant area.

This estimation method will be called L_0 regularized estimation or segmented estima-

tion. This method is illustrated through the toy-example of Figure 2 and the adaptive ridge procedure is summarized in Algorithm 2. In practice, the stopping criterion for the adaptive ridge algorithm is when the absolute difference between successive values of the weighted differences is smaller than a predefined value – we use 10^{-8} in our implementation.

Algorithm 2 Adaptive Ridge Procedure

function ADAPTIVE-RIDGE($\boldsymbol{O}, \boldsymbol{R}, \kappa$)
  $\boldsymbol{\eta} \leftarrow \mathbf{0}$
  $\boldsymbol{v} \leftarrow \mathbf{1}$
  $\boldsymbol{w} \leftarrow \mathbf{1}$
  while not converge **do**
    $\boldsymbol{\eta}^{\text{new}} \leftarrow$ NEWTON-RAPHSON($\boldsymbol{O}, \boldsymbol{R}, \kappa, \boldsymbol{v}, \boldsymbol{w}$)
    $v_{j,k}^{\text{new}} \leftarrow \left(\left(\eta_{j+1,k}^{\text{new}} - \eta_{j,k}^{\text{new}} \right)^2 + \varepsilon_v^2 \right)^{-1}$
    $w_{j,k}^{\text{new}} \leftarrow \left(\left(\eta_{j,k}^{\text{new}} - \eta_{j,k-1}^{\text{new}} \right)^2 + \varepsilon_w^2 \right)^{-1}$
    $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta}^{\text{new}}$
  end while
  Compute $(O^{\text{new}}, R^{\text{new}})$ for selected $(\boldsymbol{\eta}, \boldsymbol{v}^{\text{new}}, \boldsymbol{w}^{\text{new}})$
  $\boldsymbol{\eta}^{\text{new}} \leftarrow \log\left(\boldsymbol{O}^{\text{new}} / \boldsymbol{R}^{\text{new}}\right)$
  return $\boldsymbol{\eta}^{\text{new}}$
end function

3 Choice of the penalty constant κ

In practice, the hazard rate needs to be estimated for a set of penalty constants and the choice of κ is determined as the penalty that provides the best compromise between model fit and reduced variability of the hazard rate estimate. For the L_0 regularization model, different values of the penalty constant lead to different segmentations of the $\eta_{j,k}$. As a consequence, the problem of choosing the optimal penalty constant can be rephrased as the problem of choosing the optimal model among a set of models $\mathcal{M}_1, \ldots, \mathcal{M}_M$, where each of these models corresponds to a different segmentation of the $\eta_{j,k}$ and M is the maximum number of different models. In this section we propose different methods to select the optimal model. Comparison of the efficiency of the different methods will be analyzed in Section 4 on simulated data.

We recall that $\boldsymbol{R}$ and $\boldsymbol{O}$ are the exhaustive statistics and $\boldsymbol{\eta}$ is the parameter to be estimated in our two models. Bayesian criteria attempt to maximize the posterior probability $\mathrm{P}(\mathcal{M}_m|\boldsymbol{R}, \boldsymbol{O}) \propto \mathrm{P}(\boldsymbol{R}, \boldsymbol{O}|\mathcal{M}_m)\pi(\mathcal{M}_m)$, where $\mathrm{P}(\boldsymbol{R}, \boldsymbol{O}|\mathcal{M}_m)$ is the integrated likelihood

and $\pi(\mathcal{M}_m)$ is the prior distribution on the model. This problem is equivalent to minimizing $-2\log \mathrm{P}(\mathcal{M}_m|\boldsymbol{R},\boldsymbol{O})$. By integration

$$\mathrm{P}(\boldsymbol{R},\boldsymbol{O}|\mathcal{M}_m) = \int_{\boldsymbol{\eta}} \mathrm{P}(\boldsymbol{R},\boldsymbol{O}|\mathcal{M}_m,\boldsymbol{\eta})\pi(\boldsymbol{\eta})d\boldsymbol{\eta},$$

where $\mathrm{P}(\boldsymbol{R},\boldsymbol{O}|\mathcal{M}_m,\boldsymbol{\eta})$ is the likelihood and $\pi(\boldsymbol{\eta})$ is the prior distribution of the parameter, which is taken constant in the following. Thus Bayesian criteria are defined as

$$-2\log\left(\mathrm{P}(\mathcal{M}_m|\boldsymbol{R},\boldsymbol{O}\right) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m\log n - 2\log\pi(\mathcal{M}_m) + \mathcal{O}_{\mathrm{P}}(1),$$

where q_m is the dimension of the model $\mathcal{M}_m$ i.e., the number of constant areas selected by the adaptive ridge algorithm.

The BIC (Schwarz, 1978) corresponds to the Bayesian criterion obtained when one neglects the term $\pi(\mathcal{M}_m)$, which is equivalent to having a uniform prior on the model:

$$\mathrm{BIC}(m) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m\log n. \tag{3}$$

As explained by Żak-Szatkowska and Bogdan (2011), a uniform prior on the model is equivalent to a binomial prior on the model dimension $\mathcal{B}(JK, 1/2)$. When the true model's dimension is much smaller than the maximum possible dimension JK, the BIC tends to give too much importance to models of dimensions around $JK/2$, which will result in underpenalized estimators. To this effect, Chen and Chen (2008) have developed an extended Bayesian information criterion called EBIC_0 (or EBIC for short). One can write $\pi(\mathcal{M}_m) = \mathrm{P}(\mathcal{M}_m|\mathcal{M}_m \in \mathcal{M}_{[q_m]})\mathrm{P}(\mathcal{M}_m \in \mathcal{M}_{[q_m]})$ where $\mathcal{M}_{[q_m]}$ is the set of models of dimension q_m. The EBIC_0 criterion is defined by setting $\mathrm{P}(\mathcal{M}_m|\mathcal{M}_m \in \mathcal{M}_{[q_m]}) = 1/\binom{JK}{q_m}$ and $\mathrm{P}(\mathcal{M}_m \in \mathcal{M}_{[q_m]}) = 1$. Thus

$$\pi(M_m) = \binom{JK}{q_m}$$

and

$$\mathrm{EBIC}_0\left(m\right) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m\log n + 2\log\binom{JK}{q_m}. \tag{4}$$

Note that the EBIC_0 assigns the same *a priori* probability to all models of same dimension. Therefore, when the true model's dimension is not close to $JK/2$ the EBIC_0 will be able to select this model more easily. Namely, when the true model's dimension is very small the EBIC_0 will tend to choose very sparse models.

The last criterion that will be used is the Akaike Information Criterion (Akaike, 1998), or AIC, defined as $\mathrm{AIC}(m) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + 2q_m$. This criterion is known for performing better than the BIC in terms of mean squared error, however the BIC will tend to select sparser models than the AIC.

Note that Bayesian criteria and the AIC can only be used for the L_0 regularized estimation only, since the L_2 model does not perform a model selection. An alternative to performing model selection is to use the K-fold cross validation. With this method, the data are split at random into L parts. The estimated parameter obtained when the l-th part is left out is noted $\widehat{\eta}^{-l}(\kappa)$ and the cross-validated score is defined as

$$\mathrm{CV}(\kappa) = \sum_{l=1}^{L} \ell_n^{\kappa,l}(\widehat{\boldsymbol{\eta}}^{-l}),$$

where $\ell_n^{\kappa,l}$ is the negative log-likelihood evaluated on the l-th part of the data. The optimal penalty constant is obtained by minimizing $\mathrm{CV}(\kappa)$ with respect to κ. The L-fold cross validation method can be used for both the L_0 regularized estimation and the L_2 regularized estimation. However, this method is numerically time consuming as the estimator has to be computed L times while Bayesian criteria or the AIC provide direct methods to perform model selection from the original estimator. In the simulation studies and data analysis, we set $L = 10$.

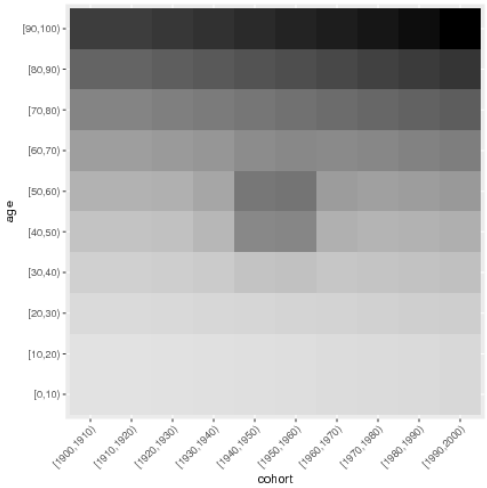
4 Simulation study

4.1 Simulation designs

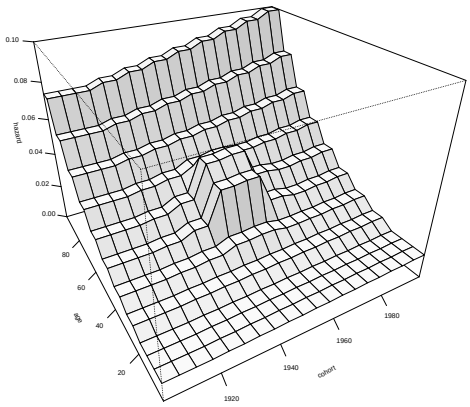
In this section, our segmented estimation method with L_0 norm is compared with the AGE-COHORT model and with the smoothed hazard estimate with the L_2 norm. The different criteria for model selection are also compared with each other. We present two simulation designs. In the first one, the true hazard rate is generated from a smooth age-cohort model which includes an interaction term on a small region of the age-cohort plane. In the second case, the true hazard rate is a piecewise constant function with four heterogeneous areas. The two true hazards are displayed in Figure 3, both in greyscale and in perspective plot.

The simulation design is as follows. We set $J = 10$ equally spaced age intervals and $K = 10$ equally spaced cohort intervals. The age intervals are defined as $[0, 10)$, ..., $[90, 100]$ and the cohorts intervals are defined as $[1900, 1910)$,..., $[1990, 2000]$. In order to simulate a dataset, the cohorts are first sampled on $K = 10$ cohort group intervals of 10 years length ranging from 1900 to 2000. Censoring is then simulated as a uniform distribution over the age interval $[75, 100]$ for all cohorts such that all observed events are comprised in the age interval $[0, 100]$. Since in practice one does not know the appropriate discretization in advance, a different discretization was used for the estimation procedure : the age and cohort intervals were defined as 5-year length intervals instead of 10 for the true hazard. As a result, a total of 20×20 parameters need to be estimated.

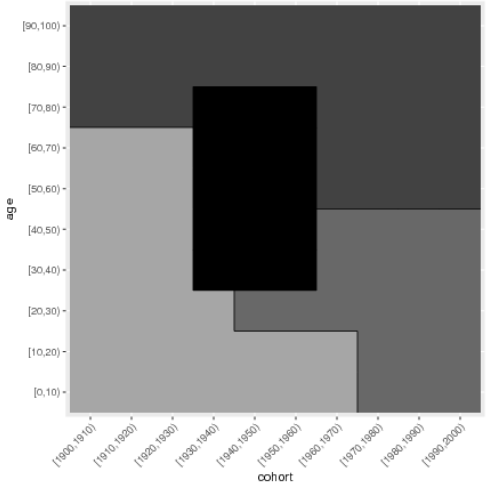
For each of the two designs, we simulated data of sample sizes 100, 400, 1000, 4000, and 10000. For each sample size, the simulation and estimation were replicated 500 times.



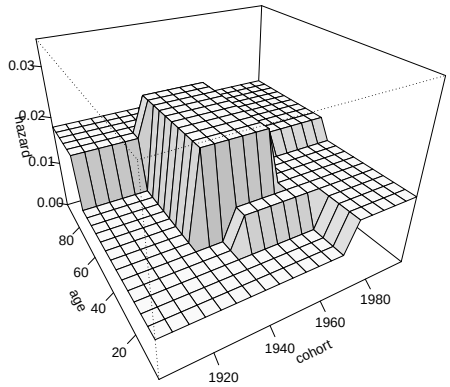
(a) Smooth true hazard – heatmap



(b) Smooth true hazard – perspective



(c) Piecewise constant true hazard – heatmap



(d) Piecewise constant true hazard – perspective

Figure 3: True hazard of the two simulation designs: smooth hazard in heatmap (a) and perspective plot (b) and piecewise constant hazard in heatmap (c) and perspective plot (d).

	L_0 method				L_2 method
Sample size	**AIC**	**BIC**	**EBIC**	**CV**	**CV**
100	412.90	401.50	4.60	4.50	1.00
400	269.60	225.50	37.90	7.10	1.00
1000	168.00	111.50	4.30	3.60	1.00
4000	79.40	24.90	5.00	3.40	1.00
10000	26.10	5.90	4.60	2.40	1.00

(a) Smooth true hazard

	L_0 method				L_2 method
Sample size	**AIC**	**BIC**	**EBIC**	**CV**	**CV**
100	429.90	428.90	1.30	1.30	1.00
400	81.70	64.30	3.00	2.40	1.00
1000	34.20	16.80	3.80	3.50	1.00
4000	12.20	2.20	1.50	1.90	1.00
10000	6.70	0.80	0.60	0.80	1.00

(b) Piecewise constant true hazard

Table 1: Relative mean squared errors with respect to the cross-validated L_2 estimator, for different sample sizes and different estimation methods. Panel (a): smooth true hazard. Panel (b): piecewise constant true hazard.

Smooth true hazard The smooth true hazard (Figures 3a and 3b) is generated using the age-cohort model $\log \lambda_{j,k} = \mu + \alpha_j + \beta_k$ with an intercept $\mu = \log(10^{-2})$. The age effect vector $\boldsymbol{\alpha}$ and cohort effect vector $\boldsymbol{\beta}$ are arithmetic sequences such that $\alpha_2 = 0$, $\alpha_J = 2.5$, $\beta_2 = 0$, and $\beta_K = 0.3$. An interaction term is added to the hazard. It corresponds to a bump in the hazard located in the neighbourhood of the region of the age-cohort plane (45,1945). The bump is defined as 10 times the Gaussian density function with mean $(1945, 45)$ and with a diagonal variance-covariance matrix with diagonal equal to $(50, 50)$. This true hazard displays a sharp increase for high values of the age, which implies that few events will be recorded in this region. On average, 91 % of the events are observed in this simulation design.

Piecewise constant true hazard The piecewise constant true hazard (Figures 3c and 3d) has four constant areas over the age-cohort square $[0, 100] \times [1900, 2000]$. On average, 71 % of the events are observed in this simulation design.

4.2 Performance of the estimation methods in terms of MSE

Our two estimation methods (L_0 and L_2 norm) are compared in terms of the Mean Squared Errors in each simulation scenario. The different selection methods for the penalty (AIC, BIC, EBIC and cross-validation) are also compared. The results are presented in Table 1. On the overall, the EBIC and cross-validated criteria outperform the AIC and the BIC for the two simulations scenarios. This is particularly true for small sizes where the AIC and the BIC behave very poorly. As expected, the L_2 norm estimator is the most performant of all estimators in the smooth true hazard scenario (Table 1a) and the L_0 method performs better in the piecewise constant hazard scenario (Table 1b) than in the smooth true hazard scenario. The L_2 norm estimator is also the most performant of all estimators in the piecewise constant hazard scenario except for very large sample sizes ($n = 10000$) where the BIC, EBIC and cross-validated criterion provide slightly better performances. Finally, in both scenarios, the EBIC always outperforms the AIC, the BIC and the cross-validated criterion. Different censoring rates were also studied which showed a degradation of the performances of the overall estimators as the percentage of censoring increased. The performance in terms of number of selected areas was also investigated. It showed that the EBIC and CV criterion perform better at selecting sparse models with few areas, while the AIC and BIC tend to overestimate the true number of areas. Indeed, for sample size 4000, the 80% inter-quantile range of the selected number of areas is $[3, 5]$ for the EBIC and $[1, 5]$ for the CV, whereas it is $[3, 13]$ and $[36, 72]$ for the BIC and AIC respectively. These experiments are not reported here.

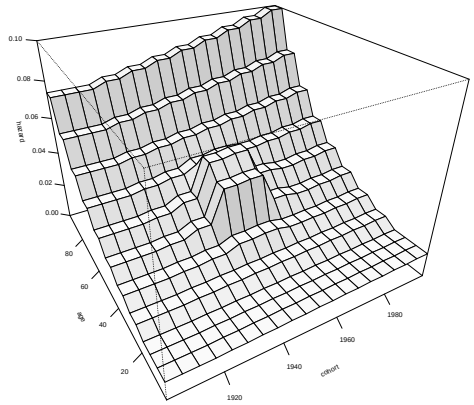
In conclusion, the simulation experiments suggest to use the EBIC among all different criteria for the L_0 norm estimator as it provides the best tradeoff between computation time and estimation performance.

4.3 Perspective plots of the estimation methods

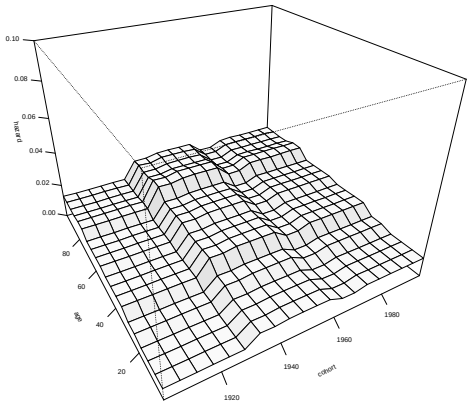
In this section the performance of our adaptive ridge (L_0 norm) and ridge (L_2 norm) estimates is assessed visually by comparison of the true hazard. The standard age-cohort model (Holford, 1983) has also been implemented. This model assumes that the hazard has the following expression:

$$\log \lambda_{j,k} = \mu + \alpha_j + \beta_k,$$

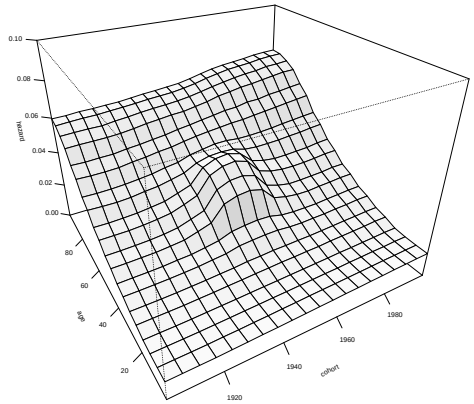
where μ is the intercept, $\boldsymbol{\alpha}$ is the age effect and $\boldsymbol{\beta}$ is the cohort effect. It should be noted that this model does not allow for interactions between age and cohort effects. Perspective plots of the median hazard estimations over 500 replications are presented in Figures 4 and 5 for the smooth and piecewise constant true hazard respectively. For the L_0 regularized estimate, the penalty constant is chosen using the EBIC.



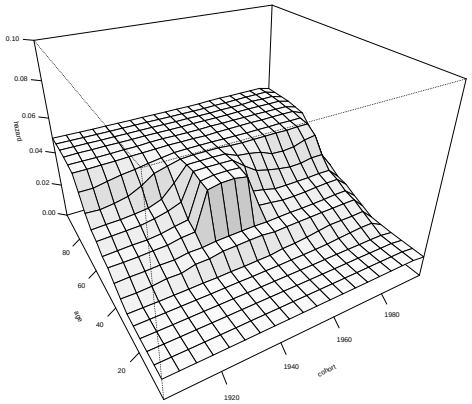
(a) True hazard



(b) Median of age-cohort estimates

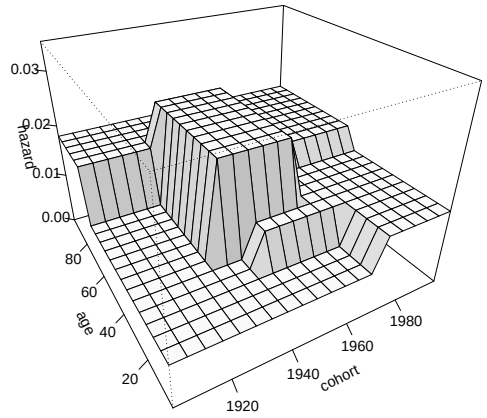


(c) Median of smooth estimates

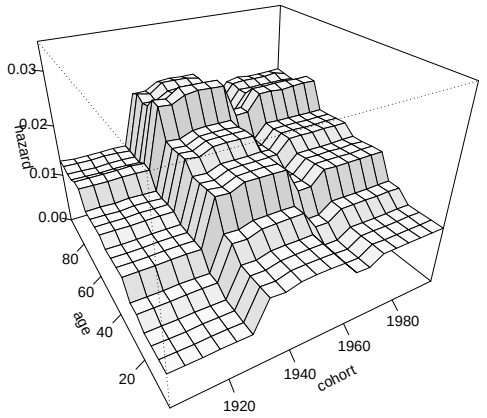


(d) Median of segmented estimates

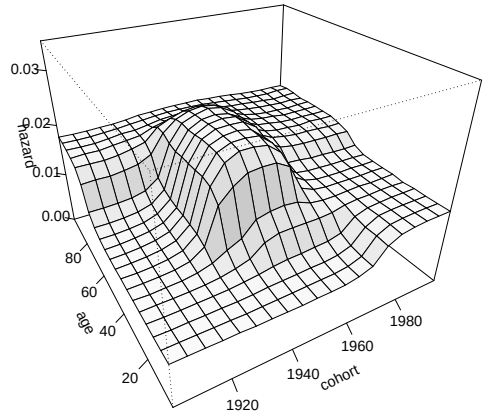
Figure 4: Smooth true hazard and corresponding estimates. The sample size is 4000 and the hazard estimates are medians taken over 500 simulations. The estimations are performed in the age-cohort plane and with different methods. Panel (a) represents the true hazard used to generate the data, Panel (b) represents the hazard estimated using the age-cohort model, Panel (c) represents the smoothed estimate, and Panel (d) represents the segmented estimate with the EBIC criterion.



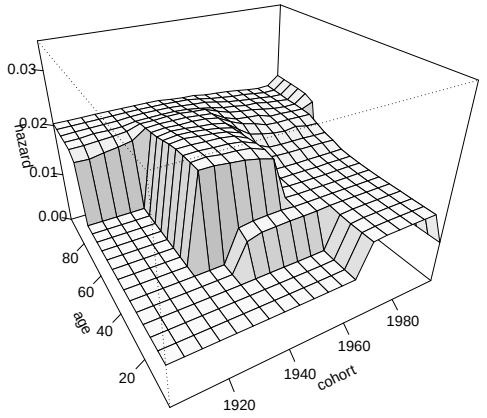
(a) True hazard



(b) Median of age-cohort estimates



(c) Median of smooth estimates



(d) Median of segmented estimates

Figure 5: Piecewise constant true hazard and corresponding estimates. The sample size is 4000 and the hazard estimates are medians taken over 500 simulations. The estimations are performed in the age-cohort plane and with different methods. Panel (a) represents the true hazard used to generate the data, Panel (b) represents the hazard estimated using the age-cohort model, Panel (c) represents the smoothed estimate, and Panel (d) represents the segmented estimate with the EBIC criterion.

In Figure 4, it is seen that the age-cohort model is not able to estimate the central bump in the hazard. On the contrary, the smoothed estimate accurately recovers the shape of the true hazard except for the high values of age where few events are observed. Interestingly, it is seen that our segmentation method provides similar results as the smoothing technique even though the true hazard is not piecewise constant.

The results in Figure 5 yield similar conclusions. The age-cohort model behaves very poorly due to its constrained structure while the ridge and adaptive estimates provide satisfactory results. In particular the shape of the true hazard is correctly captured by the adaptive ridge on the majority of replicated samples.

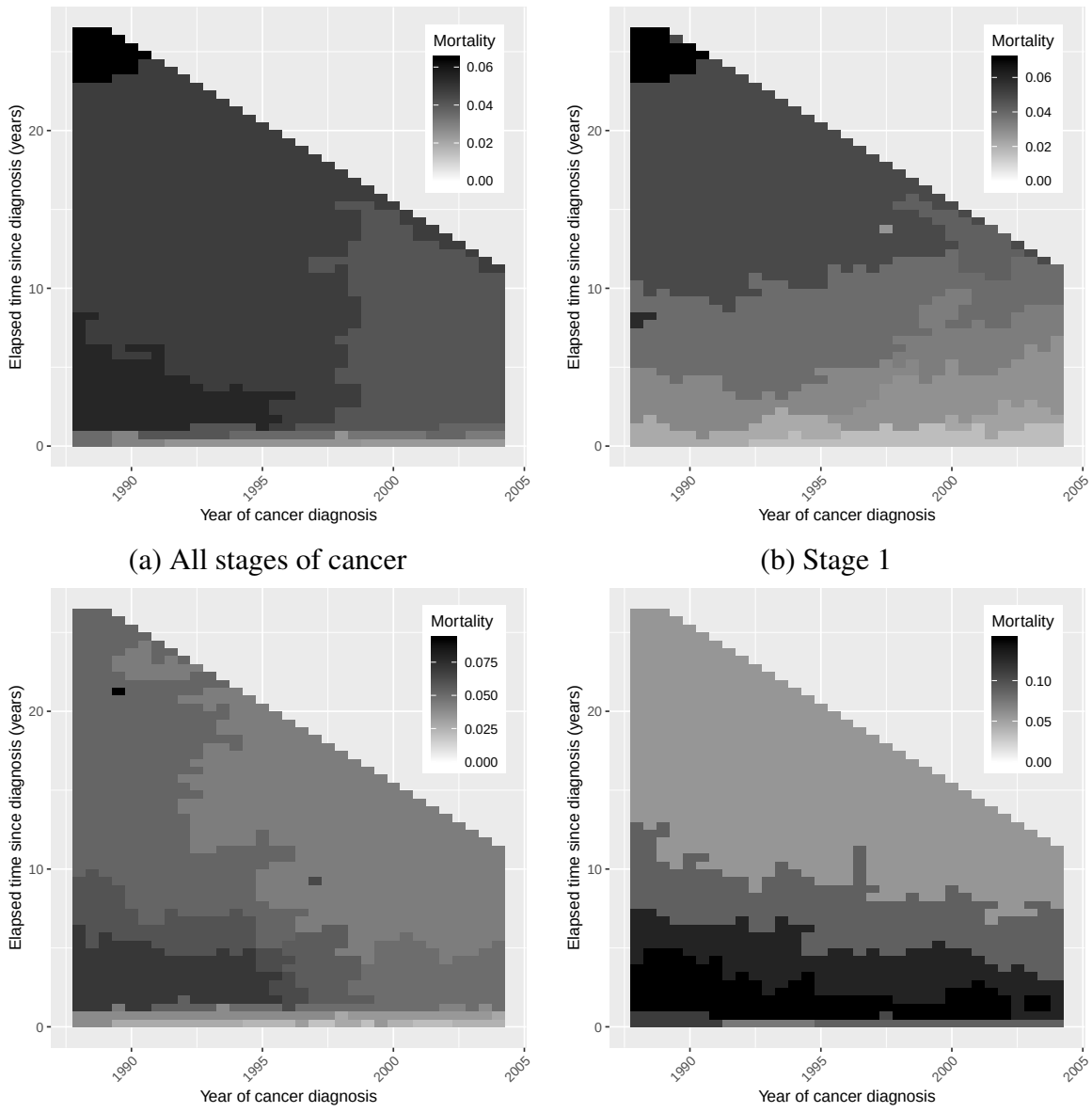
5 Real data application

Our method is applied to data of survival times after diagnosis of breast cancer. The dataset is provided by the Surveillance, Epidemiology, and End Results (SEER) Program from the US National Cancer Institute (NCI). SEER collects medical data of cancers (including stage of cancer at diagnosis and the type of tumor) and follow-up data of patients in the form of a registry. Around 28 percent of the US population is covered by the program. The registry started in February 1973 and the available current dataset includes follow-up data until January 2015. We refer to the website `https://seer.cancer.gov/` for information about the SEER Program and its publicly available cancer data.

In this study the duration of interest T is the time from breast cancer diagnosis to death in years, the variable U is the date of diagnosis (in years) and the period is the calendar time (in years). Patients continuously entered the study between 1973 and 2015 and right-censoring occurred for patients that were still alive at the end of follow-up or for those that were lost to follow-up.

The breast cancer data was extracted using the package `SEERaBomb`. For the sake of comparison, the subsample of malignant, non-bilateral breast tumor cancers was extracted from the dataset, such that the data comprises $1,265,277$ women with 60 percent of censored individuals. Times from diagnosis to last day of follow-up vary between 0 and 41 years, and the dates of cancer diagnosis U_i vary between 1973 and 2015. Death from another cause than cancer is available in the dataset and is accounted for as right-censoring.

The implementation of our adaptive ridge method aims at two goals. Firstly we aim at simultaneously detecting a cohort effect and an age effect, that is the evolution of the mortality with respect to the time elapsed since cancer diagnosis (age effect) and with respect to the date of diagnosis (cohort effect). Secondly, our method will provide estimation of the hazard rates on the resulting heterogeneous areas. The method is first applied on the whole sample of 1265277 individuals. In order to take into account the fact that mortality from cancer highly depends on the cancer stage, we also perform a stratified analysis with respect to the stage of cancer at diagnosis. For this purpose, we use the cancer stage classification



(a) All stages of cancer

(b) Stage 1

(c) Stage 2

(d) Stage 3

Figure 6: Estimated hazard of death after diagnosis of breast cancer for different stages of cancer. The estimate is obtained with the L_0 regularization. The upper right corner of every graph corresponds to the region where no data are available. Note that the scales are different between panels.

provided by the SEER data: we keep the patients with cancer stages 1, 2, and 3 at the time of diagnosis. This classification closely follows that of the American Joint Comitee on Cancer (AJCC), 3^{rd} Edition; the details are given at page 86 of the manual entitled Comparative Staging Guide for Cancer, available at `https://seer.cancer.gov`. The main difference between the two classifications is that the SEER Program classifies the cases where lymph node status cannot be assessed as if there was no regional lymph node metastasis.

The L_0 estimates for the whole sample and for each cancer stage are displayed in Figure 6. We see that the different stages of cancer at diagnosis have a great impact on the survival times. For Stage 1 cancers, the mortality is low between 0 and 4–5 years after diagnosis, and steadily increases afterwards. The date of diagnosis seems to have no impact on the mortality of Stage 1 cancers. On the other hand, Stage 2 cancers exhibit a strong effect of the date of diagnosis: around $1995 - 1997$, the mortality significantly decreases. This can correspond to an improvement of the treatment of breast cancer around that period in the United States. Finally, Stage 3 cancers display a very high hazard rate across all dates of diagnosis. This seems to indicate that the evolution in treatments of breast cancer had a significant impact on the survival times after diagnosis, but almost exclusively when cancers were diagnosed at Stage 2. Two additional analyses of the hazard rate with stratification with respect to age at diagnosis and estrogen receptor status were performed in the Supplementary Materials. The results suggest that the shift in mortality around year 1996 could correspond to the introduction of hormone-blocking therapy (Fisher et al., 1999).

Conclusion

In this article, we have introduced a new estimation method to deal with age-period-cohort analysis. This model assumes no specific structure of the effects of age and cohort and the hazard rate is directly estimated without estimating the effects. In order to take into account possible overfitting issues, a penalty is used on the likelihood to enforce similar consecutive values of the hazard to be equal. Two different types of penalty terms were introduced. One leads to a ridge type regularization while the other leads to a L_0 regularization. Different selection methods of the penalty parameter were also introduced. To our knowledge, a segmented estimation model of this kind has never been introduced in this context.

Using simulated data, it has been shown that the cross validated ridge estimator and the $EBIC_0$ adaptive ridge estimator perform the best in terms of mean squared error. The cross validation criterion was shown to provide the best fit of the hazard rate, but its very high computationally cost makes it non-competitive. In this context, this modified BIC criterion comes out as a powerful tool to select the *best* bias-variance tradeoff.

The method was successfully applied to data of survival after breast cancer provided by the SEER program. The segmented estimate of the hazard rate displays important information about the shift in mortality after being diagnosed of breast cancer in the United States

in the mid-1990s.

Our method could be directly extended to a different discretization of the age-period-cohort plane, such as $1 \times 1 \times 1$-year triangles that are represented in dark gray in Figure 1 (see Section 3 of Carstensen, 2007, for an example of this discretization). Another extension would be to consider other types of penalizations. Instead of estimating a piecewise constant hazard, one could estimate a piecewise linear hazard by penalizing over second order differences of the hazard.

Acknowledgement The authors are thankful to the National Cancer Institute for providing U.S. mortality data on cancer.

Conflict of Interest The authors have declared no conflict of interest.

References

O. O. Aalen, Ø. Borgan, and S. Gjessing. *Survival and Event History Enalysis: A Process Point of View*. Statistics for Biology and Health. Springer, New York, NY, 2008. OCLC: ocn213855657.

H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

R. Beran. Nonparametric Regression with Randomly Censored Survival Data. Technical report, Technical Report, University of California, Berkeley, 1981.

O. Bouaziz and G. Nuel. L0 Regularization for the Estimation of Piecewise Constant Hazard Rates in Survival Analysis. *Applied Mathematics*, 08(03):377–394, 2017.

B. Carstensen. Age–Period–Cohort Models for the Lexis Diagram. *Statistics in Medicine*, 26(15):3018–3045, 2007.

B. Carstensen, M. Plummer, E. Laara, and M. Hills. *Epi: A Package for Statistical Analysis in Epidemiology*. 2017.

J. Chen and Z. Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

G. Csardi and T. Nepusz. The igraph Software Package for Complex Network Research, 2006.

I. D. Currie and J. G. Kirkby. Smoothing Age-Period-Cohort Models with P -splines: A Mixed Model Approach. 2009.

B. Fisher, J. Dignam, N. Wolmark, D. L. Wickerham, E. R. Fisher, E. Mamounas, R. Smith, M. Begovic, N. V. Dimitrov, R. G. Margolese, C. G. Kardinal, M. T. Kavanah, L. Fehrenbacher, and R. H. Oishi. Tamoxifen in treatment of intraductal breast cancer: National Surgical Adjuvant Breast and Bowel Project B-24 randomised controlled trial. *THE LANCET*, 353:8, 1999.

F. Frommlet and G. Nuel. An Adaptive Ridge Procedure for L0 Regularization. *PLoS ONE*, 11(2):e0148620, 2016.

C. Heuer. Modeling of Time Trends and Interactions in Vital Rates Using Restricted Regression Splines. *Biometrics*, 53(1):161–177, 1997.

T. R. Holford. The Estimation of Age, Period and Cohort Effects for Vital Rates. *Biometrics*, 39(2):311–324, 1983.

N. Keiding. Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 332(1627):487–509, 1990.

D. Kuang, B. Nielsen, and J. P. Nielsen. Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika*, 95(4):979–986, 2008.

I. W. McKeague and K. J. Utikal. Identifying Nonlinear Covariate Effects in Semimartingale Regression Models. *Probability Theory and Related Fields*, 87(1):1–25, 1990.

B. Nielsen. Apc: An R Package for Age-Period-Cohort Analysis. *The R Journal*, 7(2), 2015.

Y. Ogata and K. Katsura. Likelihood Analysis of Spatial in Homogeneity for Marked Point Patterns. *Annals of the Institute of Statistical Mathematics*, 40(1):29–39, 1988.

C. Osmond and M. J. Gardner. Age, Period and Cohort Models Applied to Cancer Mortality Rates. *Statistics in Medicine*, 1(3):245–259, 1982.

M. Plummer and B. Carstensen. Lexis: An R Class for Epidemiological Studies with Long-Term Follow-Up. *Journal of Statistical Software*, 38(5):1–12, 2011.

R. C. A. Rippe, J. J. Meulman, and P. H. C. Eilers. Visualization of Genomic Changes by Segmented Smoothing Using an L0 Penalty. *PLoS ONE*, 7(6):e38230, 2012.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

Y. Yang and K. C. Land. *Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications*. Chapman & Hall/CRC Interdisciplinary Statistics, 2013.

M. Żak-Szatkowska and M. Bogdan. Modified Versions of the Bayesian Information Criterion for Sparse Generalized Linear Models. *Computational Statistics & Data Analysis*, 55(11):2908–2924, 2011.

2.3 Application to the evolution of breast cancer mortality

In this section, I include the supplementary material to the paper “Regularized Bidimensional Estimation of the Hazard Rate”.

Supplementary Material to: Regularized Bidimensional Estimation of the Hazard Rate

Vivien Goepp[1], Jean-Christophe Thalabard[1], Grégory Nuel[2], and Olivier Bouaziz[1]

[1]*MAP5 (Department of mathematics, 45, rue des Saints-Pères, 75006 Paris)*
[2]*LPMA (Department of mathematics, 4, Place Jussieu, 75005 Paris)*

August 2018

1 Application to Breast Cancer Mortality: Stratification with respect to the Age at Diagnosis

The mortality of breast cancer is known to greatly vary on whether the cancer is pre or post-menopausal (Consensus, 1985). Consequently, a thorough analysis of the mortality from breast cancer would require to stratify with respect to the menopausal status at diagnosis. Since this covariate is not present in the data, we decided to stratify the sample with respect to the age of the patient at diagnosis, which is a proxy of menopausal status. Most women are known to have their menopause between 45 and 55 years old (Hill, 1996; Henderson et al., 2008; Gold, 2011), with 25th, 50th, 75th percentiles ranging from years 47-49, 50-51, 52-54, respectively, according to countries and surveys (Mishra et al., 2017). Consequently, based on the available information in SEER, for each cancer stage, the patients were divided into three classes of age at diagnosis: $(.,45]$, $(45,55]$, and $(55,.)$ as a proxy for pre- menopausal, peri- menopausal and post- menopausal ages, respectively. The resulting estimated hazards are represented in Figure 1.

The stage 1 cancer patients younger than 45 and the stage 3 cancer patients older than 55 display the same mortality across all dates of diagnosis, i.e. with no cohort effect.

Moreover, the mortality of stage 1 cancer patients aged 45 and older at diagnosis has a slight cohort effect corresponding to a progressive decrease in the mortality across all survival times (Peto et al., 2000). This could suggest a trend of slow and steady improvement of the treatment of breast cancer in the United States over the period $1887-2005$.

Finally, we observe a clear decrease of the mortality for stage 2 cancers for all three age classes. This shift is located at the year 1995 for middle-aged patients and around the years $1997-1998$ for patients younger than 45 and older than 55. The same drop in mortality is observed for stage 3 cancers with patients younger than 45 at diagnosis, around year 1995. This could correspond to the introduction of improvements in the treatments of breast cancer in the United States (Consensus, 1985). Among the three main medical innovations, which can be considered in this period, the improvement of the surgical procedures for the loco- regional control of the disease and the assessment

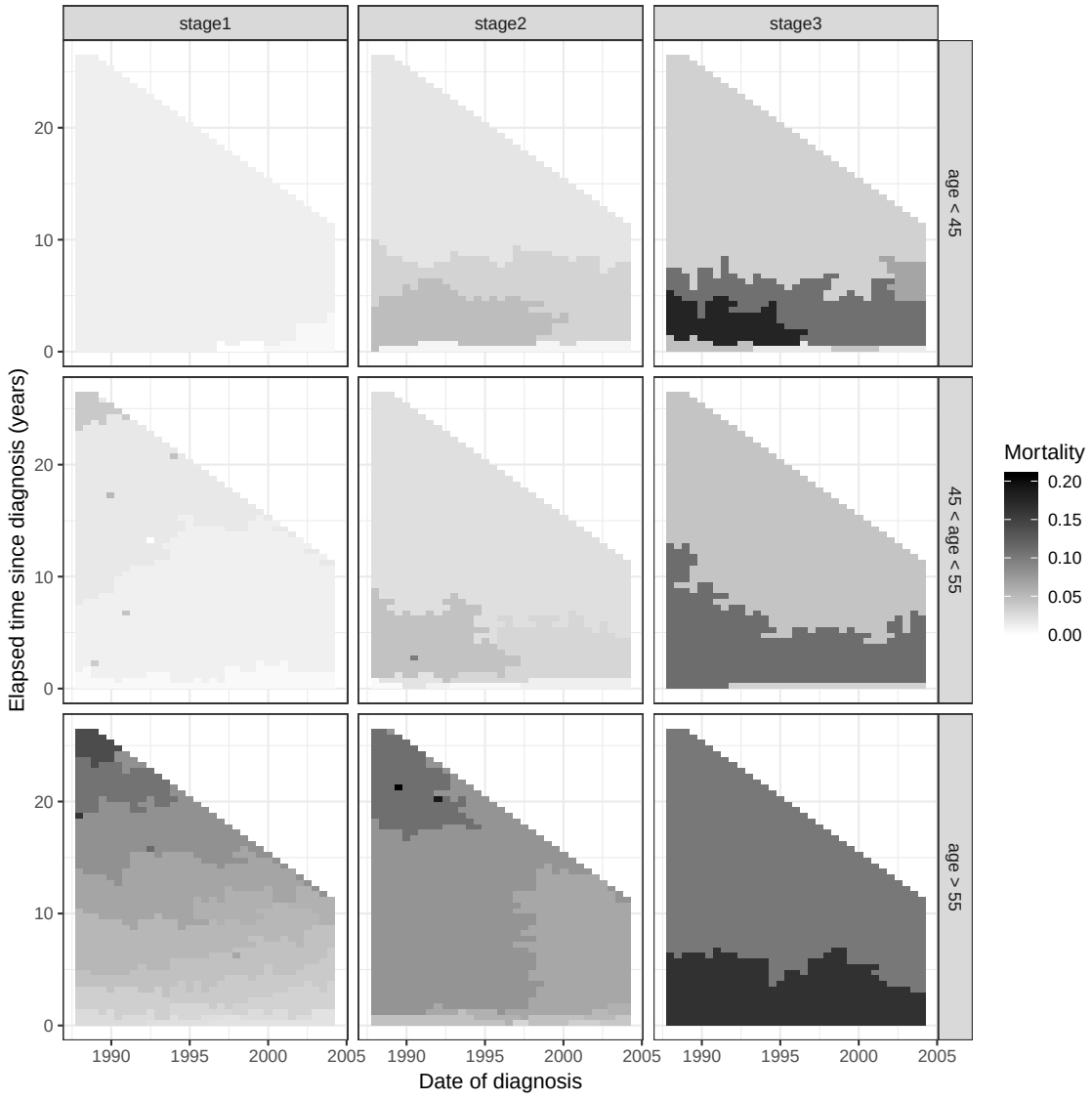


Figure 1: Estimated hazard of death since diagnosis of breast cancer for different cancer stages and for different ages at diagnosis. The estimate is obtained with the L_0 regularization. The upper right corner of every graph corresponds to the region where no data are available. All graphs share the same scale.

of the beneficial effect of hormone-receptor therapies could be reflected in the observed survival in stages 1-2, whereas the later emergence during this period of new classs of chemotherapeutic agents like taxoids (Rowinsky et al., 1992; Crown et al., 2004) or herceptin-based therapies targeted on new class of tumor markers (Pegram et al., 1998; Emens and Davidson, 2004) would be related with the changes in survival observed in stage 3. In the next section, we will use a stratified analysis to understand the effect of hormone-receptor therapies on the mortality shift in the mid-1990s.

2 Application to Breast Cancer Mortality: Stratification with respect to the Estrogen Receptor Status

The cohort effect highlighted in the previous section could correspond to the introduction of Selective Estrogen Receptor Modulator (SERM) treatments and in particular the use of Tamoxifen as a treatment for breast cancer, showing improved survival in women with estrogen receptor positive tumor, initially in post- menopausal women (Fisher et al., 1989), later in both post - and premenopausal women (Early Breast Cancer Trialists' Collaborative Group, 1988; Fisher et al., 1998; Pritchard, 2005; Cochrane, 2008). Indeed, Tamoxifen was gradually used in the early years of 1990's (Gail et al., 1999; Harlan et al., 2002; Mariotto et al., 2006) to decrease the mortality of breast cancer patients. This treatment is only efficient on estrogen receptor-sensitive cancers. To validate our hypothesis, we conducted the estimation of mortality separately for patients with estrogen receptor sensitive and non-sensitive cancers. Since stage 2 cancers displayed a strong cohort effect across all ages at diagnosis, we only kept stage 2 cancers in this study. The estimated mortality is given in Figure 2. Note that the spikes in the mortality are an artifact of the segmentation procedure when the sample sizes tend to be too small in some regions of the age-cohort plane and are not to be taken into account in the interpretation of the mortality.

There is a clear difference in the evolution of mortality with respect to time at diagnosis between sensitive and non-sensitive estrogen cancers. For estrogen sensitive cases, the mortality displays the same sudden decrease around years $1997 - 1998$ as in Figure 1, across all age classes. In particular for individuals aged 55 or more at the time of diagnosis, the mortality has gradually decreased for estrogen sensitive patients, whereas it did not evolve with time for estrogen non-sensitive patients. On the other hand, the mortality for non-estrogen sensitive cancers displays almost no cohort effect for all ages at diagnosis (Knight et al., 1977).

The same analysis was run with stratification with respect to progesterone receptor status, with very similar morality estimates (results not shown here). Further analyses could be carried out to better understand the effect of the introduction of hormone-blocking therapies on mortality. However, the segmentation of the hazard rate, even with this simple stratified analysis, highlighted that the adoption of SERM therapies in the United States is a potential reason for the sharp decrease of mortality in the middle of the 1990s (Peto et al., 2000).

References

Cochrane. Tamoxifen for early breast cancer. *The Cochrane database of systematic reviews*, (4):

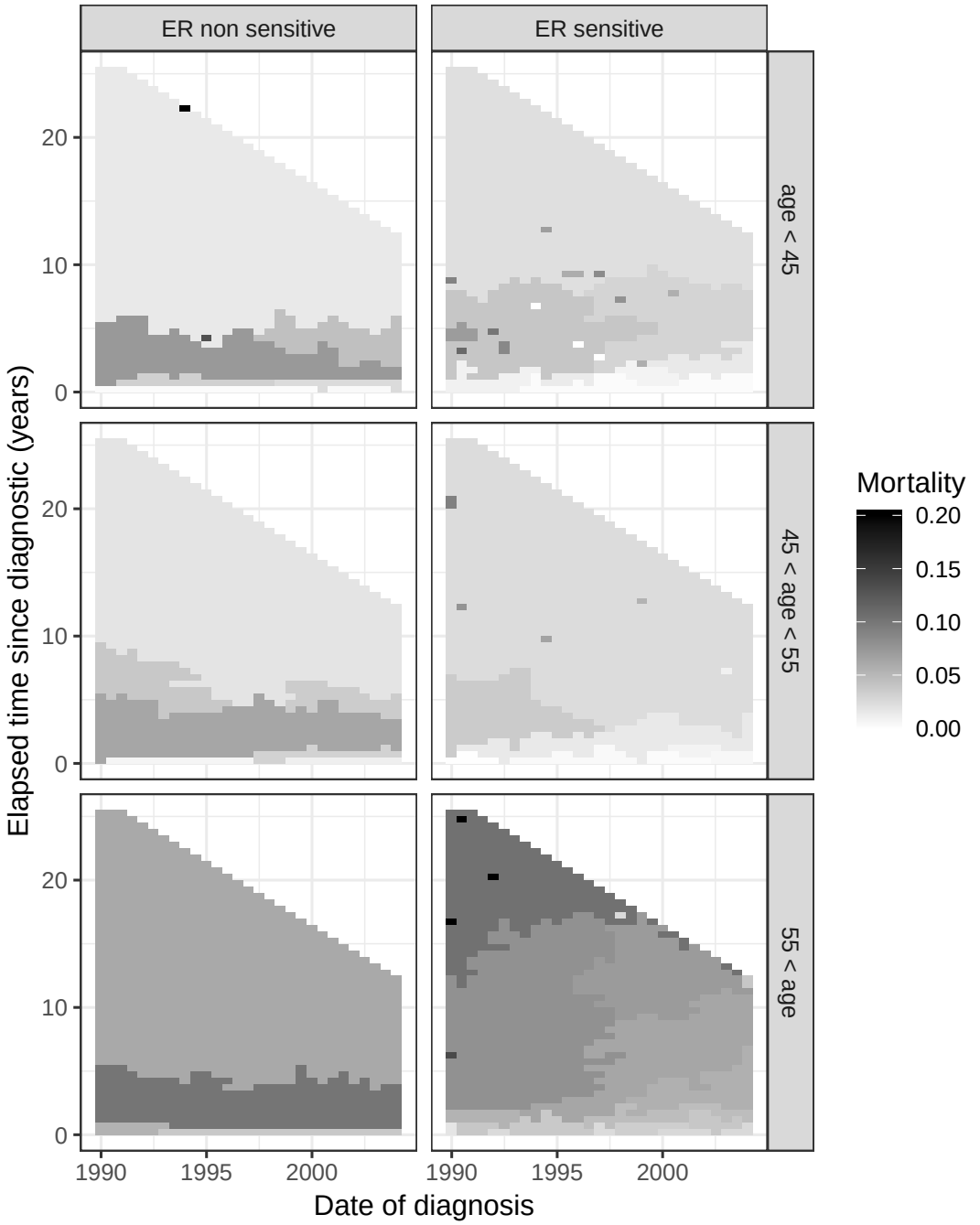


Figure 2: Estimated hazard of death since diagnosis of breast cancer for Stage 2 cancers. The estimation is carried separately for three classes of age at diagnosis: $(., 45]$, $(45, 55]$, and $(55, .)$ and for sensitive and non-sensitive estrogen receptor cancers. Inference is made with the L_0 regularization. All graphs share the same scale.

CD000486, 2008.

Consensus. Consensus conference. Adjuvant chemotherapy for breast cancer. *JAMA*, 254(24): 3461–3463, 1985.

J. Crown, M. O'Leary, and W.-S. Ooi. Docetaxel and paclitaxel in the treatment of breast cancer: A review of clinical experience. *The oncologist*, 9 Suppl 2:24–32, 2004.

Early Breast Cancer Trialists' Collaborative Group. Effects of adjuvant tamoxifen and of cytotoxic therapy on mortality in early breast cancer. An overview of 61 randomized trials among 28,896 women. *The New England journal of medicine*, 319(26):1681–1692, 1988.

L. A. Emens and N. E. Davidson. Trastuzumab in breast cancer. *Oncology*, 18(9):1117–28; discussion 1131–2, 1137–8, 2004.

B. Fisher, J. Costantino, C. Redmond, R. Poisson, D. Bowman, J. Couture, N. V. Dimitrov, N. Wolmark, D. L. Wickerham, and E. R. Fisher. A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors. *The New England journal of medicine*, 320(8):479–484, 1989.

B. Fisher, J. P. Costantino, D. L. Wickerham, C. K. Redmond, M. Kavanah, W. M. Cronin, V. Vogel, A. Robidoux, N. Dimitrov, J. Atkins, M. Daly, S. Wieand, E. Tan-Chiu, L. Ford, and N. Wolmark. Tamoxifen for Prevention of Breast Cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *JNCI: Journal of the National Cancer Institute*, 90(18):1371–1388, 1998.

M. H. Gail, J. P. Costantino, J. Bryant, R. Croyle, L. Freedman, K. Helzlsouer, and V. Vogel. Weighing the Risks and Benefits of Tamoxifen Treatment for Preventing Breast Cancer. *Journal of the National Cancer Institute*, 91(21):18, 1999.

E. B. Gold. The timing of the age at which natural menopause occurs. *Obstetrics and Gynecology Clinics of North America*, 38(3):425–440, 2011.

L. C. Harlan, J. Abrams, J. L. Warren, L. Clegg, J. Stevens, and R. Ballard-Barbash. Adjuvant therapy for breast cancer: Practice patterns of community physicians. *Journal of Clinical Oncology*, 20(7):1809–1817, 2002.

K. D. Henderson, L. Bernstein, B. Henderson, L. Kolonel, and M. C. Pike. Predictors of the timing of natural menopause in the Multiethnic Cohort Study. *American journal of epidemiology*, 167 (11):1287–1294, 2008.

K. Hill. The demography of menopause. *Maturitas*, 23(2):113–127, 1996.

W. A. Knight, R. B. Livingston, E. J. Gregory, and W. L. McGuire. Estrogen receptor as an independent prognostic factor for early recurrence in breast cancer. *Cancer Research*, 37(12):4669–4671, 1977.

A. B. Mariotto, E. J. Feuer, L. C. Harlan, and J. Abrams. Dissemination of adjuvant multiagent chemotherapy and tamoxifen for breast cancer in the United States using estrogen receptor information: 1975-1999. *Journal of the National Cancer Institute. Monographs*, (36):7–15, 2006.

G. D. Mishra, N. Pandeya, A. J. Dobson, H.-F. Chung, D. Anderson, D. Kuh, S. Sandin, G. G. Giles, F. Bruinsma, K. Hayashi, J. S. Lee, H. Mizunuma, J. E. Cade, V. Burley, D. C. Greenwood, A. Goodman, M. K. Simonsen, H.-O. Adami, P. Demakakos, and E. Weiderpass. Early menarche, nulliparity and the risk for premature and early natural menopause. *Human Reproduction*, 32(3): 679–686, 2017.

M. D. Pegram, G. Pauletti, and D. J. Slamon. HER-2/neu as a predictive marker of response to breast cancer therapy. *Breast Cancer Research and Treatment*, 52(1-3):65–77, 1998.

R. Peto, J. Boreham, M. Clarke, C. Davies, and V. Beral. UK and USA breast cancer deaths down 25% in year 2000 at ages 20-69 years. *The Lancet*, 355(9217):1822, 2000.

K. Pritchard. Endocrinology and hormone therapy in breast cancer: Endocrine therapy in premenopausal women. *Breast Cancer Research*, 7(2):70–76, 2005.

E. K. Rowinsky, N. Onetto, R. M. Canetta, and S. G. Arbuck. Taxol: The first of the taxanes, an important new class of antitumor agents. *Seminars in Oncology*, 19(6):646–662, 1992.

Chapter 3

Estimating interactions in the age-cohort model

Contents

3.1 Introduction to age-period-cohort analysis 74
3.2 The age-cohort-interaction model 75
3.3 The estimating procedure 76
3.4 Choice of the penalty constant 77
3.5 Simulation results 77
3.5.1 Simulation setting 77
3.5.2 Predictive performance 79
3.5.3 Perspective plots 81
3.6 Conclusion 84

3.1 Introduction to age-period-cohort analysis

In the study of diseases with age at onset, the incidence – or hazard rate – is an important quantity to estimate. In these studies, data are either present in the form of registers or collected from a cohort that is being followed. In both cases, the individuals are usually very heterogeneous in terms of date of birth – also called *cohort*. Equivalently, the event of interest is observed at widely different calendar times – also called *period*. It is therefore necessary to adjust for this heterogeneity. Namely, many models have been developed to infer the hazard rate as a function of age, cohort, and period, called age, cohort, and period effects. This is the goal of age-period-cohort analysis.

The simplest models in age-period-cohort analysis are the age-cohort and age-period models (Clayton and Schifflers, 1987). These models assume that each variable has an additive effect on the logarithm of the hazard. In the age-cohort model,

$$\log \lambda_{j,k} = \alpha_j + \beta_k, \tag{3.1}$$

where α_j is the age effect at age j, β_k is the cohort effect at calendar time c, and λ is the hazard rate. In this model, the hazard rate is assumed to be the product of the age effect and the cohort effect. Since there are fewer parameters than values of the hazard, this model yields a regularized estimation of λ, and thus of the age and cohort effect. But this model is very restrictive because it assumes that the age and cohort have an additive effect only, that is, the log-hazard difference between two ages is the same for all cohorts and that the log-hazard difference between two cohorts is the same for all ages. The full model in age-period-cohort analysis is called the age-period-cohort model (Clayton and Schifflers, 1987). It jointly estimates the three additive effects:

$$\log \lambda_{j,k} = \alpha_j + \beta_k + \gamma_{j+k}, \tag{3.2}$$

where γ_l is the period effect at calendar time $l = j + k$. Because of the linear dependency (age + cohort = period) between the variables, these effects are only identifiable up to a linear trend. This makes the estimated effects hard to interpret and many attempts have been made to overcome this difficulty. Carstensen (2007) offered to choose a cohort of reference and to infer the relative hazard with respect to this cohort. Kuang et al. (2008) introduced a reparametrization based on the second order derivatives of the effects which allows estimation of the three effects when the index set $\{(i, j)\}$ is a trapezoid.

Both the two-factor models and the age-period-cohort model are useful to infer the influence of the corresponding variables over the hazard rate. Indeed the age, cohort, and period effects are important information to understand the evolution of the incidence. However they can represent a simplistic view of the evolution of the hazard along time. We study the age-cohort model where an interaction term is added. This term is introduced in order to capture non-linear relations between age and cohort effects. The interaction between age and cohort can be interpreted as regions of the age-cohort plane where the hazard is either unexpectedly high or unexpectedly low. Thus, the interaction between effects is an important tool to detect peculiar features of the event of interest.

In this chapter, we introduce a generalization of the age-cohort model which includes the estimation of the interaction between the age and cohort effects. This model is called age-cohort-interaction model, or "ACI" model. For the sake of simplicity, we only consider the age-cohort model in this work, but the period-cohort and age-period models can be generalized in the same way to estimate interactions between effects. We introduce a penalized likelihood approach to jointly infer the age and cohort effects as well as the interaction effect. For the model to be interpretable, we enforce a regularization of the interaction term. We offer a choice between two types of regularization: one provides a smooth estimate of the interaction and one provides a segmented estimate of the interaction. We illustrate these two regularizations and their respective advantages.

The rest of this chapter is constructed as follows. Section 3.2 introduces the age-cohort-interaction model. Section 3.3 develops on a procedure for the estimation of the model. In Section 3.4, we explain how to choose the tuning parameter of the regularization. Finally, our method is illustrated on simulated data in Section 3.5.

3.2 The age-cohort-interaction model

We assume that the time of interest is subject to independent right-censoring (Fleming and Harrington, 2011, p. 27), such that the available data is of the form (U_i, T_i, Δ_i), $i = 1, \ldots, n$, where T_i is the censored time-to-event of individual i and Δ_i equals 0 if individual i is censored and 1 otherwise. The hazard rate as a function of the cohort u and age t is defined by

$$\lambda(t|u) = \lim_{\delta t \to 0} \frac{1}{\delta t} \mathrm{P}(t < T < t + \delta t | T > t, U = u),$$

where U is the random variable of the cohort. The cohort variable is discretized into J intervals $[c_0, c_1), \ldots, [c_{J-1}, c_J)$ and the age variable is discretized into K intervals $[d_0, d_1), \ldots, [d_{K-1}, d_K)$. The data is rewritten using the exhaustive statistics $O = \left(O_{1,1}, \ldots, O_{J,K}\right)$ and $R = \left(R_{1,1}, \ldots, R_{J,K}\right)$, where $O_{j,k}$ and $R_{j,k}$ are respectively the number of observed events and total time at risk in the j-th cohort interval and in the k-th age interval. We assume that $\lambda(t|u)$ is constant over each set $[c_{j-1}, c_j) \times [d_{k-1}, d_k)$ and infer the discretized hazard rate $\left(\lambda_{1,1}, \ldots, \lambda_{J,K}\right)$. The negative log-likelihood takes a simple expression:

$$\ell_n(\boldsymbol{\lambda}) = \sum_j \sum_k \lambda_{j,k} R_{j,k} - O_{j,k} \log\left(\lambda_{j,k}\right). \tag{3.3}$$

We introduce the age-cohort-interaction ("ACI") model:

$$\log \lambda_{j,k} = \mu + \alpha_j + \beta_k + \delta_{j,k}, \tag{3.4}$$

where α_j is the cohort effect for the j-th cohort interval, β_k is the age effect for the k-th age interval, and $\delta_{j,k}$ is the interaction term for the (j,k)-th rectangle. The parameter μ plays the role of the intercept (that is $\lambda_{1,1} = \mu$), and we impose $\alpha_1 = \beta_1 = \delta_{1,k} = \delta_{j,1} = 0$ so that the model has JK freely varying parameters and is identifiable. The parameter of the model is $\boldsymbol{\theta} = (\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta})$, where $\boldsymbol{\alpha} = (\alpha_j)_{2 \le j \le J}$, $\boldsymbol{\beta} = (\beta_k)_{2 \le k \le K}$, and $\boldsymbol{\delta} = (\delta_{j,k})_{2 \le j \le J, 2 \le k \le K}$.

Note that the interaction term $\boldsymbol{\delta}$ between the age and cohort is different from the period effect $\boldsymbol{\gamma}$ in the age-period-cohort model (3.2). The period effect can be seen as an interaction effect between age and cohort, which is a function of the period only. Contrarily to the age-period-cohort model, our model can estimate an interaction between age and cohorts which is any function of these two variables. Consequently, the ACI model can be seen as an extension of the age-cohort model which is also more general than the age-period-cohort model.

The negative log-likelihood takes the form

$$\ell_n(\boldsymbol{\theta}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \exp\left(\mu + \alpha_j + \beta_k + \delta_{j,k}\right) R_j^k - \left(\mu + \alpha_j + \beta_k + \delta_{j,k}\right) O_j^k. \tag{3.5}$$

For the parameter $\boldsymbol{\delta}$ to correctly estimate the interaction between age and cohort – that is, the part of the hazard which diverges from the age-cohort model – one needs to impose a constraint on $\boldsymbol{\delta}$. We want the δs to be zero-valued everywhere except for values of (j,k) for which the hazard is too different from the age-cohort model. To this end, the parameters are estimated using the penalized negative log-likelihood:

$$\ell_n(\boldsymbol{\theta}) - \frac{\kappa}{2} \sum_{j,k} \{\|\delta_{j,k} - \delta_{j-1,k}\|_0 + \|\delta_{j,k} - \delta_{j,k-1}\|_0\},$$

where $\|x\|_0$ is the L_0 norm, equal to 0 if $x = 0$ and to 1 otherwise and $\kappa > 0$ is a trade-off parameter. This penalty is proportional to the number of non-zero differences of $\delta_{j,k}$. Consequently, the penalized estimation enforces the differences of δ to be equal to zeros except in a small amount of times. Since $\delta_{1,k} = \delta_{j,1} = 0$, this penalty also enforces the interaction term to be equal to zero except where necessary.

This penalty term ensures that $\boldsymbol{\delta}$'s second order total variation is small. Thus the parameters μ, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ estimate the age-cohort model that is the closest to the data, and $\boldsymbol{\delta}$ estimates the divergence of the data from the age-cohort model. The penalty constant tunes the trade-off between goodness-of-fit and regularization. When $\kappa = 0$, the ACI model is subject to no regularization and the corresponding estimate is the maximum likelihood estimate. When pen $\to \infty$, the interaction term is set to be uniformly equal to zero and the ACI model is the age-cohort model. In this sense, the ACI model is an extension of the age-cohort model which allows for interaction between age and cohort effects.

This penalty is however not continuous, which makes the maximization of the penalized log-likelihood intractable in practice. The next section deals with a numerical method for approximating the latter penalized likelihood.

3.3 The estimating procedure

Following the works from Rippe et al. (2012) and Frommlet and Nuel (2016), we use an iterative algorithm to approximate the L_0 norm penalty. The estimate is the minimizer of:

$$\ell_n^{\kappa}(\theta) \triangleq \ell_n(\theta) + \frac{\kappa}{2} \sum_{j,k} v_{j,k} \left(\delta_{j,k} - \delta_{j-1,k}\right)^2 + w_{j,k} \left(\delta_{j,k} - \delta_{j,k-1}\right)^2, \tag{3.6}$$

where $v_{j,k}$ and $w_{j,k}$ are positive weights. The procedure iterates between minimizing (3.6) with fixed weights and updating the weights, using the formula

$$\begin{cases} v_{j,k}^{(m)} = \left(\left(\delta_{j+1,k}^{(m)} - \delta_{j,k}^{(m)}\right)^2 + \varepsilon_v^2 \right)^{-1}, & 1 \le j \le J-1, \quad 2 \le k \le K, \\ w_{j,k}^{(m)} = \left(\left(\delta_{j,k}^{(m)} - \delta_{j,k-1}^{(m)}\right)^2 + \varepsilon_w^2 \right)^{-1}, & 2 \le j \le J, \quad 1 \le k \le K-1, \end{cases} \tag{3.7}$$

where (m) is the iteration step, $\delta_{j,k}^{(m)}$ is the parameter from the last iteration and ε_v and ε_w are positive constants negligible compared to 1 (in practice we choose $\varepsilon_v = \varepsilon_w = 10^{-6}$ (Frommlet and Nuel, 2016)).

At convergence, $v_{j,k}\left(\delta_{j+1,k} - \delta_{j,k}\right)^2$ are very close to 1 if the adjacent values $\delta_{j+1,k}$ and $\delta_{j,k}$ have been estimated to have different values and to 0 if they have been estimated to have the same value – and similarly for $w_{j,k}\left(\delta_{j,k+1} - \delta_{j,k}\right)^2$. We typically use a threshold of 10^{-8}, so that values smaller than 10^{-8} are set to 0 and values larger than $1 - 10^{-8}$ are set to 1. Then one creates the graph whose vertices are the JK discretization rectangles and whose edges are the connexion between adjacent rectangles that have a difference close to 0. As with other classical penalized methods (e.g. lasso, ridge) and as pointed out in Frommlet and Nuel (2016), the adaptive ridge penalization scheme induces a shrinkage bias. Therefore, the unpenalized maximum likelihood estimate is used to infer the value of the hazard over the constant areas estimated using the adaptive ridge procedure. The values of $\delta^{(r)}$ are not estimated using the results of the adaptive ridge algorithm but by unpenalized maximum likelihood estimation: $\widehat{\delta}^{(r)} = O^{(r)} / R^{(r)}$ where $O^{(r)}$ is the number of events in the r-th constant area and $R^{(r)}$ is the time at risk in the r-th constant area.

The estimating procedure is summarized in Algorithm 5. See also (Bouaziz and Nuel, 2017) and (Goepp et al., 2018) for implementations of the adaptive ridge procedure in similar contexts. The penalized likelihood in Equation (3.6) is minimized using the Newton-Raphson algorithm. Expressions of first and second order derivatives of ℓ_n^{κ} are given in Appendix A. The Hessian matrix has JK rows and columns, so its inversion is computationally intensive. A fast inversion method for its inversion is detailed in Appendix B.

Algorithm 5 Adaptive Ridge Procedure

function ADAPTIVE-RIDGE($\boldsymbol{O}, \boldsymbol{R}, \kappa$)
  $\boldsymbol{\theta} \leftarrow \mathbf{0}$
  $\boldsymbol{v} \leftarrow \mathbf{1}$
  $\boldsymbol{w} \leftarrow \mathbf{1}$
  while not converge **do**
    $\boldsymbol{\theta}^{\text{new}} \leftarrow$ NEWTON-RAPHSON($\boldsymbol{O}, \boldsymbol{R}, \kappa, \boldsymbol{v}, \boldsymbol{w}$)
    $v_{j,k}^{\text{new}} \leftarrow \left(\left(\delta_{j+1,k}^{\text{new}} - \delta_{j,k}^{\text{new}}\right)^2 + \varepsilon_v^2\right)^{-1}$
    $w_{j,k}^{\text{new}} \leftarrow \left(\left(\delta_{j,k}^{\text{new}} - \delta_{j,k-1}^{\text{new}}\right)^2 + \varepsilon_w^2\right)^{-1}$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{\text{new}}$
  end while
  Compute $(O^{\text{new}}, R^{\text{new}})$ selected $(\boldsymbol{\delta}, \boldsymbol{v}^{\text{new}}, \boldsymbol{w}^{\text{new}})$
  $\widehat{\boldsymbol{\theta}} \leftarrow \boldsymbol{O}^{\text{new}} / \boldsymbol{R}^{\text{new}}$
  return $\widehat{\boldsymbol{\theta}}$
end function

3.4 Choice of the penalty constant

In practice, the hazard rate needs to be estimated for a set of penalty constants and the optimal κ is chosen as the penalty that provides the best compromise between model fit and reduced variability of the hazard rate estimate. In the piecewise constant estimation, different values of the penalty constant yield to different segmentations of $\delta_{j,k}$. As a consequence, to choose the optimal penalty constant is to choose the optimal model, where each model corresponds to a different segmentation of the $\delta_{j,k}$s.

The computational burden of cross validation dissuades us from using it for the ACI model. We offer to use a criterion for model selection. The AIC (Akaike, 1974), the BIC (Schwarz, 1978), and, following a similar work in Goepp et al. (2018), a modified version of the BIC called "EBIC" (for "Extended BIC", see Chen and Chen, 2012). The latter is defined as a BIC criterion with a specific prior on the model. It is defined as

$$\text{EBIC}(\kappa) = \text{BIC}(\kappa) + 2\log\binom{JK}{m(\kappa)}, \tag{3.8}$$

where $m(\kappa)$ is the model dimension. Notice that the term on the right hand side of the latter equation takes small values when the model dimension is close to 1 or to JK, the maximal dimension. Thus the prior of the EBIC gives more weight to models of small and large dimensions, compared to $JK/2$.

Comparison of the efficiency of the different methods will be analyzed in Section 3.5 on simulated data.

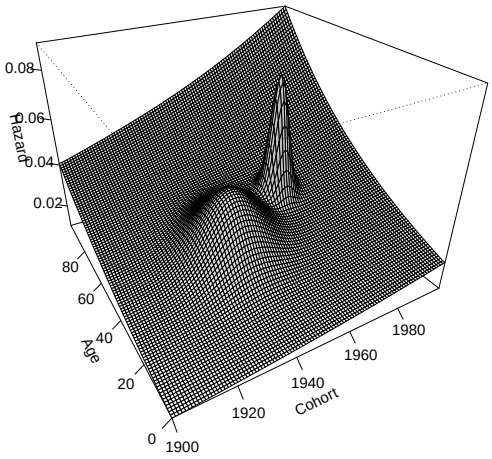
3.5 Simulation results

3.5.1 Simulation setting

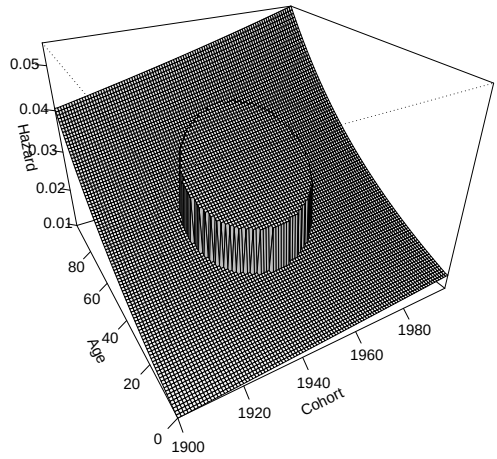
The simulation setting is as follows. The data is simulated using a "true" (conditional) hazard rate $\lambda(t|u)$, given as piecewise constance function:

$$\lambda(t|u) = \sum_{j=1}^{J}\sum_{k=1}^{K} \lambda_{j,k} \mathbb{1}_{[c_{j-1},c_j)\times[d_{k-1},d_k)}(t,u), \tag{3.9}$$

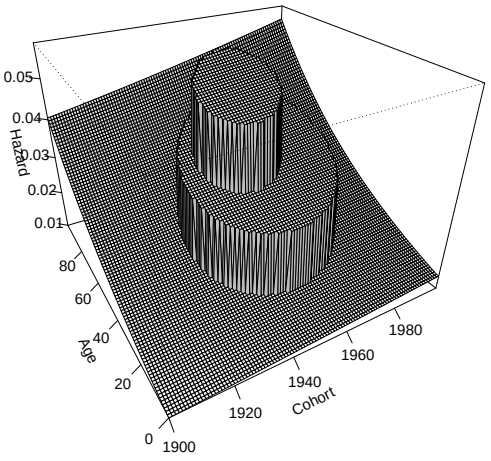
where $(\lambda_{j,k})$ is the discrete hazard rate. We set $J = 20$ equally spaced age intervals $[0,5), \ldots, [95,100]$ and $K = 20$ equally spaced cohort intervals $[1900,1905), \ldots, [1995,2000]$. The censoring distribution



(a) Design 1



(b) Design 2



(c) Design 3

Figure 3.1: True hazard ($\lambda^*_{j,k}$) for simulation design.

is uniform over $[75,100] \times [1900,2000]$. For inference, we use the same discretization into 5-year age and cohort intervals. Consequently, there are $J \times K = 400$ parameters in the full model. We use three simulation designs, with three different hazard rates; see Figure 3.1.

For each design, we simulate data of sample sizes 400, 100, 4000, and 10000. For each sample size, the simulation and estimation were replicated 100 times. In all simulation designs, the log-hazard rate is the sum of and age effect, a cohort effect, and an interaction effect:

$$\log \lambda^*_{j,k} = \mu^* + \alpha^*_j + \beta^*_k + \delta^*_{j,k},$$

where $\mu^* = \log(10^{-2})$ and (α^*_j) and (β^*_k) are sequences in arithmetic progression ranging from $\alpha^*_2 = 0$ to $\alpha^*_J = 1.4$ and from $\beta^*_2 = 0$ to $\beta^*_K = 0.3$ respectively. We defined three simulation designs, with different choices for the interaction effect $\delta^*_{j,k}$.

Simulation Design 1**.** In this simulation design $\delta^*_{j,k}$ is the sampling of the mixture of two Gaussian densities:

$$\delta^*_{j,k} = 15 f_{\boldsymbol{\mu}_1,\boldsymbol{\sigma}_1}(c_j, d_k) + 3 f_{\boldsymbol{\mu}_2,\boldsymbol{\sigma}_2}(c_j, d_k),$$

where $f_{\boldsymbol{\mu},\boldsymbol{\sigma}}(t,u)$ denotes the bivariate Gaussian density at age t and cohort u. We set $\boldsymbol{\mu}_1 = (40,1940)^T$, $\boldsymbol{\mu}_2 = (50,1965)$, $\boldsymbol{\sigma}_1 = 100\boldsymbol{I}_2$, and $\boldsymbol{\sigma}_2 = 10\boldsymbol{I}_2$. The resulting hazard is represented in Figure 3.1a: the general (increasing) trend is given by the age and cohort effects, and the two "bumps" are given by the interaction effect. The first bump is spread out and has a large amplitude; the second bump however is smaller and more spiky. The latter can be difficult to estimate.

Contrarily to Design 2, the interaction term has a low amplitude compared to the combined (additive) effect of age and cohort. Consequently, the interaction effect is purposefully hard to estimate in this design, as it is hard to separate it from the age and cohort effects.

Simulation Design 2**.** In this simulation design, the interaction term is a piecewise function whose support is a circle at the center of the age-cohort plane:

$$\delta^*_{j,k} = 0.01 * \mathbb{1}((j,k) \in \mathscr{C}(1950,50,500)).$$

where $\mathscr{C}\left(m_k, m_j, R\right)$ is the circle of center $\left(m_k, m_j\right)$ and of radius R.

The interaction term is a very simple piecewise constant function. The resulting hazard is represented in Figure 3.1b. In this design, it is required that the ACI model infers correctly the support of $\delta^*_{j,k}$.

Simulation Design 3**.** This simulation design is piecewise constant with two levels:

$$\delta^*_{j,k} = 0.015 * \mathbb{1}\left((j,k) \in \mathscr{C}(1950,50,700)\right) + 0.015 * \mathbb{1}\left((j,k) \in \mathscr{C}(1955,45,200)\right).$$

The resulting hazard is represented in Figure 3.1c. The two terms in this equations are functions whose supports are circles and one circle is inside the other. Consequently the interaction effect has a large support and a large amplitude (compared to the age and cohort effects). Hence it is easy to estimate the presence of the interaction term. However it can be harder to estimate correctly the presence and shape of the higher level.

3.5.2 Predictive performance

We represent the estimation performance of the ACI model compared to two other estimates: the AC model and the MLE. Note that in all simulation designs, the interaction term takes values close to zero and consequently the true hazard is purposefully close to a age-cohort model. This simulation setting is aimed at quantifying how well the interaction term can be inferred by the ACI model. Another simulation setting should be used to quantify how well the ACI model performs when the true hazard

Sample size	MLE	AC model	ACI model		
			AIC	BIC	EBIC
400	2.905×10^{-2}	6.320×10^{-4}	6.666×10^{-4}	$\mathbf{6.235 \times 10^{-4}}$	$\mathbf{6.235 \times 10^{-4}}$
1000	8.813×10^{-3}	$\mathbf{2.951 \times 10^{-4}}$	3.027×10^{-4}	2.954×10^{-4}	2.954×10^{-4}
4000	3.167×10^{-3}	$\mathbf{7.417 \times 10^{-5}}$	8.109×10^{-5}	7.671×10^{-5}	7.435×10^{-5}
10000	1.303×10^{-3}	$\mathbf{4.133 \times 10^{-5}}$	5.700×10^{-5}	4.136×10^{-5}	4.136×10^{-5}

(a) Simulation Design 1

Sample size	MLE	AC model	ACI model		
			AIC	BIC	EBIC
1000	4.332×10^{-3}	9.263×10^{-5}	9.911×10^{-5}	9.299×10^{-5}	$\mathbf{9.209 \times 10^{-5}}$
4000	6.459×10^{-4}	$\mathbf{2.701 \times 10^{-5}}$	3.059×10^{-5}	2.702×10^{-5}	2.702×10^{-5}
10000	2.195×10^{-4}	$\mathbf{1.463 \times 10^{-5}}$	1.662×10^{-5}	1.465×10^{-5}	1.465×10^{-5}
40000	4.976×10^{-5}	7.992×10^{-6}	$\mathbf{7.808 \times 10^{-6}}$	8.002×10^{-6}	8.002×10^{-6}

(b) Simulation Design 2

Sample size	MLE	AC model	ACI model		
			AIC	BIC	EBIC
400	1.578×10^{-2}	$\mathbf{4.305 \times 10^{-4}}$	4.377×10^{-4}	4.368×10^{-4}	4.368×10^{-4}
1000	7.168×10^{-3}	$\mathbf{2.068 \times 10^{-4}}$	2.325×10^{-4}	2.074×10^{-4}	2.074×10^{-4}
4000	1.051×10^{-3}	$\mathbf{7.273 \times 10^{-5}}$	9.689×10^{-5}	7.302×10^{-5}	7.302×10^{-5}
10000	3.242×10^{-4}	5.386×10^{-5}	5.585×10^{-5}	$\mathbf{5.381 \times 10^{-5}}$	5.409×10^{-5}

(c) Simulation Design 3

Table 3.1: Mean squared errors of the AC model, the ACI model, and the maximum likelihood estimate. For each sample size, the mean squared error was computed over 100 repetitions. The smallest mean squared error in each row is highlighted in bold.

is very dissimilar from an additive effect between age and cohort. This is beyond the scope of this study.

In this section, we compare the predictive performance of the ACI model with the AIC, BIC, and EBIC criteria, to the AC model and to the MLE. We simulate the data a repeated number of $L = 100$ times. For each repetition, we compute the ACI, AC, and MLE estimates. The MLE estimate is explicit: $\hat{\lambda}_{j,k} = O_{j,k}/R_{j,k}$. The AC estimate is not explicit but is easy to compute – we provide a Newton-Raphson-based estimation in the package `hazreg`. For each estimate, we then compute the mean squared error

$$\text{MSE} = \frac{1}{L}\sum_{l=1}^{L} \left\| \hat{\boldsymbol{\lambda}}_l - \boldsymbol{\lambda}^* \right\|^2,$$

where $\hat{\boldsymbol{\lambda}}_l$ is the estimated hazard for the l-th sample. This procedure is completed for different sample sizes: 400, 1000, 4000, and 10000. A sample size of 10000 is considered small for applications in epidemiology. With $J = K = 20$, this sample size is 25 times larger than the number of parameters.

The results are gathered in Table 3.1. The ACI model performs always (almost) as well as the AC model, which performs way better than the MLE. This is coherent with the fact that in three simulation designs, the true hazard is close (in L_2 norm) to the age-cohort model. For larger sample sizes, the ACI model outperforms the AC model (results not shown here). No criteria stands out as better for the task of model selection. These remarks hold likewise for all three simulation designs.

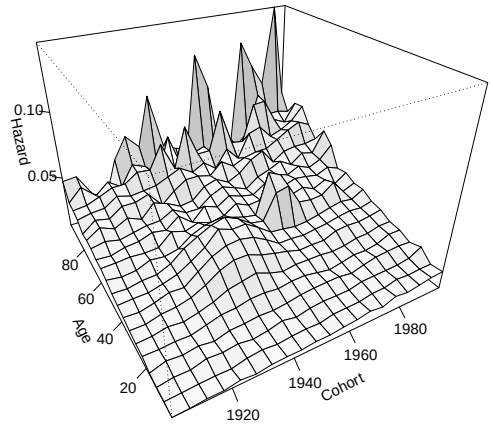
Consequently, the ACI model is shown to perform as well as the AC model, even when the true hazard is close to that of the AC model. In addition, the ACI infers the interaction effect. In the next section, we will qualify, how well the interaction term is estimated.

3.5.3 Perspective plots

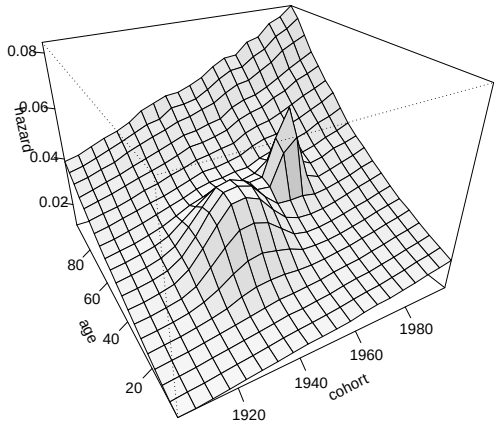
In this section, we represent the hazard rate estimated by the ACI model. We also represent the interaction effect. We simulate datasets of sample size 40000 a repeated number of 100 times. We represent the pointwise medians of the estimates over the 100 replications. The estimates are represented in Figures 3.2, 3.3, and 3.4 for the three simulation designs, respectively.

Figure 3.2 represents the results for Simulation Design 1. We compare the hazard estimated with the ACI model (Figure 3.2b) and with the MLE (Figure 3.2a). The MLE infers accurately the two "bumps" forming the interaction effect, as well as the general trend forming the age and cohort effects. But the regularizing effect of the ACI model improves the quality of the estimation, especially for large values of the age, where there are fewer data points and thus a lower signal-to-noise ratio. We also represent the decomposition of the estimated hazard (Figure 3.2b) in the age and cohort effects and the interaction effect: Figure 3.2c represents $\left(\exp(\mu + \alpha_j + \beta_k)\right)_{j,k}$ and Figure 3.2d represents $\left(\exp(\delta_{j,k})\right)_{j,k}$. The interaction effect is accurately estimated, which leads to think the ACI model is useful to detect the presence of interactions between the age and cohort effects. (We stress that the functions represented here are medians over 100 repetitions. The estimation of interaction effect with such precision as that of Figure 3.2 requires a greater sample size than 40000). The estimated age and cohort effects presented in Figure 3.2c are very close to their true values. Figures 3.2e and 3.2f represent the age $(\alpha_2, \cdots, \alpha_J)$ and cohort $(\beta_2, \cdots, \beta_K)$ effects respectively The estimates of each of the 100 replications are in light grey, their median is in red and the true value is in blue. The age effect is well estimated: the estimate is close to the true value and the replicated estimate highlight a symmetrical variance thereof. The cohort effect displays the same behavior, but with a decreased quality of estimation: the empirical variance is large and the median estimate presents two bumps, whereas the true cohort effect has none. This is explained by the small values taken by $\boldsymbol{\beta}^*$ compared to that of μ^*, $\boldsymbol{\alpha}^*$, and $\boldsymbol{\delta}^*$. We used the AIC selection criterion in this simulation. The same simulations were done with the BIC and EBIC criterion (results not shown here) and yield similar, albeit visually slightly worst, results. This remark holds for Figures 3.3-3.4.

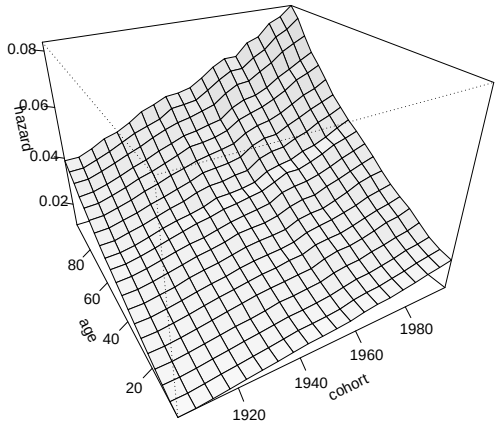
Figure 3.3 represents the estimated hazard with the ACI model for Simulation Design 2. Figure 3.3a represents the true hazard. The hazard estimated by the AC model and ACI model are repre-



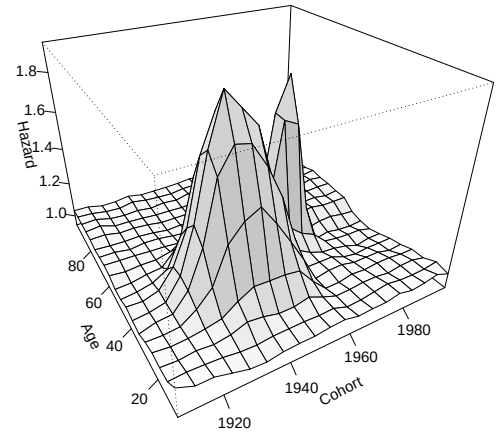
(a) MLE hazard



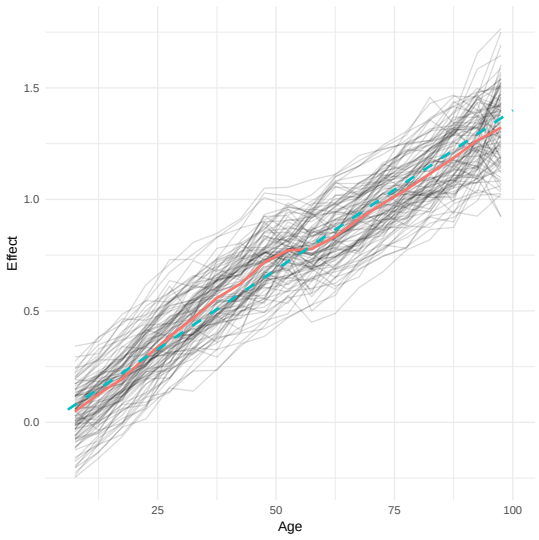
(b) Estimated hazard (ACI model)



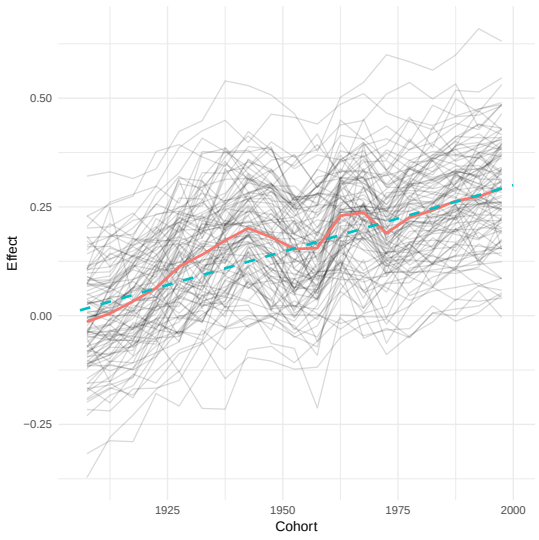
(c) Age and cohort effects (ACI model)



(d) Interaction effect (ACI model)

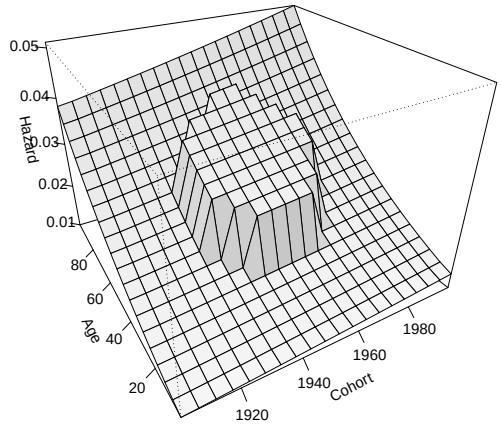


(e) Age effect (ACI model)

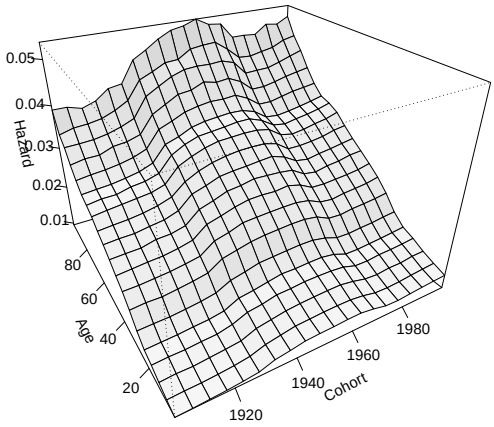


(f) Cohort effect (ACI model)

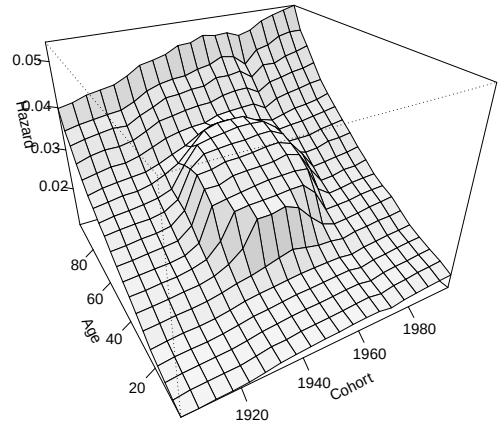
Figure 3.2: Estimates of the ACI model in Simulation Design 1. (a) Maximum likelihood estimate (b) Estimated hazard ($\log \lambda_{j,k}$) in the ACI model (c) Age and cohort effects ($\mu + \alpha_j + \beta_k$) in the ACI model (d) Interaction effect ($\delta_{j,k}$) from the ACI model (e) Age effects (α_j) (f) Cohort effects (β_k).



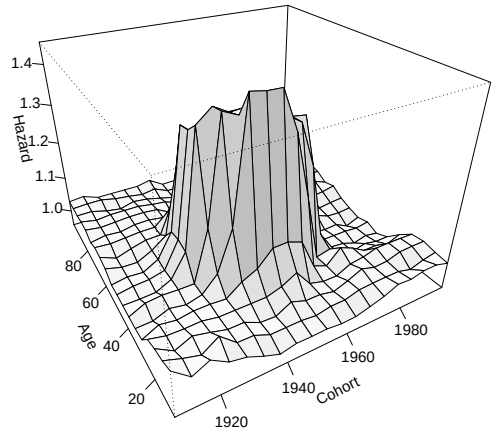
(a) True hazard



(b) AC Model



(c) Estimated hazard (ACI model)



(d) Estimated interaction effect (ACI model)

Figure 3.3: Estimates in Simulation Design 2 with the ACI model: (a) True hazard, (b) AC model ($\log \lambda_{j,k} = \mu + \alpha_j + \beta_k$), (c) Estimated hazard in the ACI model, and (d) Interaction effect ($\delta_{j,k}$) in the ACI model.

sented in Figures 3.3b and 3.3c. The AC model fails to estimate the shape of the hazard. In particular, the interaction term cannot – by definition – be estimated by the AC model, and the estimated age and cohort effects are skewed to fit the interaction term. Figure 3.3 represents the interaction term. The estimated interaction is not visually close to a piecewise constant function. Its circular support is however well estimated.

Figure 3.4 represents the same results as Figure 3.3 for Simulation Design 3. The remarks are the same: the ACI model infers both the age-cohort trend and the interaction term satisfyingly. The added difficulty of this simulation design is the presence of two levels instead of one. Figure 3.4d shows that in this case, the interaction effect is still well estimated.

3.6 Conclusion

We have introduced a new model to estimate the interaction between effects in the age-period-cohort setting. This model was developed to generalize the age-cohort model and is called age-cohort-interaction model. (The age-period and period-cohort models can be generalized in the same fashion.)

This model uses penalized maximum likelihood estimation. Indeed, a fused adaptive ridge method is used to regularize the interaction term. This regularization forces the interaction term to be piecewise constant. We propose different model selection criteria for the choice of the penalty constant: the AIC, the BIC, and the EBIC. We compare the criteria using simulations. The AIC is underpenalizing and consequently, the estimated interaction effect is piecewise constant with many different values (simulation results not shown here). The BIC and EBIC tend to select similar penalties and consequently to provide the same estimate, the EBIC being slightly more penalizing than the BIC (results not shown here).

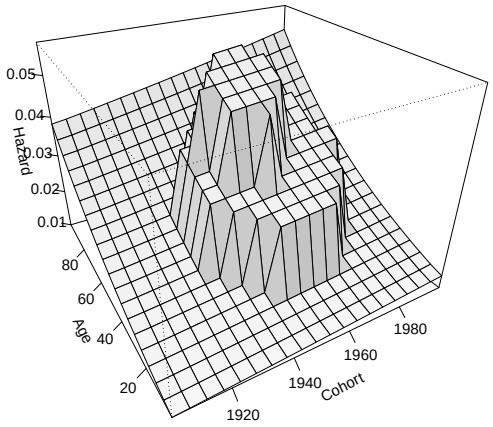
The estimation performance of our model is evaluated on simulations and compared with that of the age-cohort model and the MLE. With sample size 400, 1000, 4000, and 10000, our method performs as well as the age-cohort model and outperform the MLE. For higher sample sizes, our model slightly outperforms the age-cohort model (results not shown here). Finally, perspective plots of the interaction term illustrate that the age-cohort-interaction model performs well at estimating the presence and support of interaction between age and cohort effects. The shape of the interaction is inferred correctly in the case where it is piecewise constant.

Perspectives include using a bootstrapping of the data to provide empirical confidence intervals of the estimate. Averaging the collection of estimates provides an estimate of the interaction term that is not piecewise constant. I have performed exploratory analysis in this direction, with promising results. Moreover, I have started to apply this method to the study of breast cancer incidence among Norwegian women. I plan to apply the ACI model to the data provided by the NOWAC cohort (Lund et al., 2008). Finally, the ACI model could be used to test for the presence of a (non-linear) interaction effect. Indeed, if we had an asymptotic distribution for the estimated interaction effect, one could build a test with null hypothesis "$\forall\, j,k,\delta_{j,k}=0$". This would require to establish asymptotic properties of the adaptive ridge. We refer to the conclusion of this manuscript for perspectives on this matter.

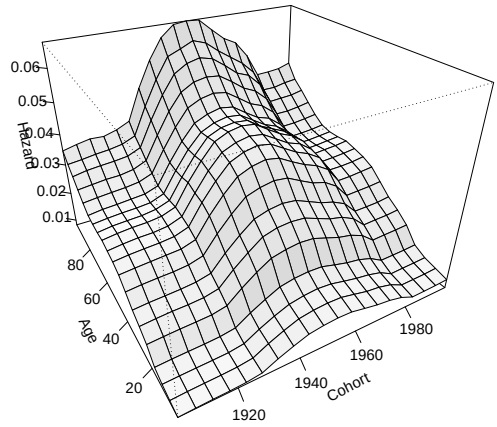
Appendix A: Expressions of the score and Hessian matrix

Unpenalized likelihood in the ACI model. We give the expressions of the (unpenalized) negative log-likelihood given in Equation 3.5. Its first order derivative is given by

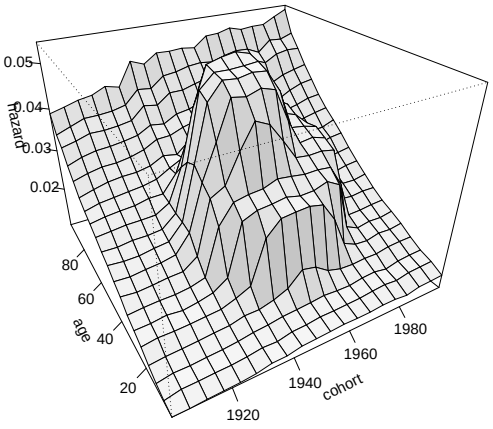
$$\frac{\partial \ell}{\partial \mu}(\boldsymbol{\theta}) = \sum_{j=1}^{J}\sum_{k=1}^{K} \exp\left(\mu+\alpha_j+\beta_k+\delta_{jk}\right) R_j^k - O_j^k,$$



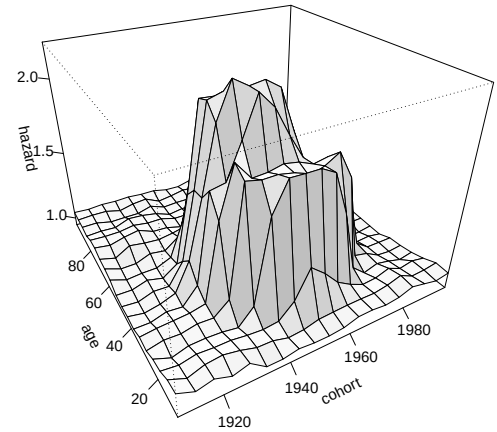
(a) True hazard



(b) AC model



(c) Estimated hazard (ACI model)



(d) Estimated interaction effect (ACI model)

Figure 3.4: Estimates in Simulation Design 2 with the ACI model: (a) True hazard, (b) AC model ($\log \lambda_{j,k} = \mu + \alpha_j + \beta_k$), (c) Estimated hazard in the ACI model, and (d) Interaction effect ($\delta_{j,k}$) in the ACI model.

$$\frac{\partial \ell}{\partial \alpha_{j'}}(\boldsymbol{\theta}) = \sum_{k=1}^{K} \exp\left(\mu + \alpha_{j'} + \beta_k + \delta_{j'k}\right) R_{j'}^{k} - O_{j'}^{k},$$

$$\frac{\partial \ell}{\partial \beta_{k'}}(\boldsymbol{\theta}) = \sum_{j=1}^{J} \exp\left(\mu + \alpha_j + \beta_{k'} + \delta_{jk'}\right) R_{j}^{k'} - O_{j}^{k'},$$

and

$$\frac{\partial \ell}{\partial \delta_{j'k'}}(\boldsymbol{\theta}) = \exp\left(\mu + \alpha_{j'} + \beta_{k'} + \delta_{j'k'}\right) R_{j'}^{k'} - O_{j'}^{k'},$$

with $1 \leq j' \leq J$ and $1 \leq k' \leq K$. Its second order derivative is given by

$$\frac{\partial^2 \ell}{\partial \mu^2}(\boldsymbol{\theta}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \exp\left(\mu + \alpha_j + \beta_k + \delta_{jk}\right) R_{j}^{k},$$

$$\frac{\partial^2 \ell}{\partial \alpha_{j''} \partial \mu}(\boldsymbol{\theta}) = \sum_{k=1}^{K} \exp\left(\mu + \alpha_{j''} + \beta_k + \delta_{j''k}\right) R_{j''}^{k},$$

$$\frac{\partial^2 \ell}{\partial \beta_{k''} \partial \mu}(\boldsymbol{\theta}) = \sum_{j=1}^{J} \exp\left(\mu + \alpha_j + \beta_{k''} + \delta_{jk''}\right) R_{j}^{k''},$$

$$\frac{\partial^2 \ell}{\partial \delta_{j''k''} \mu}(\boldsymbol{\theta}) = \exp\left(\mu + \alpha_{j''} + \beta_{k''} + \delta_{j''k''}\right) R_{j''}^{k''},$$

$$\frac{\partial^2 \ell}{\partial \alpha_{j''} \partial \alpha_{j'}}(\boldsymbol{\theta}) = \mathbb{1}_{j'=j''} \sum_{k=1}^{K} \exp\left(\mu + \alpha_{j''} + \beta_k + \delta_{j''k}\right) R_{j''}^{k},$$

$$\frac{\partial^2 \ell}{\partial \beta_{k''} \partial \alpha_{j'}}(\boldsymbol{\theta}) = \exp\left(\mu + \alpha_{j'} + \beta_{k''} + \delta_{j'k''}\right) R_{j'}^{k''},$$

$$\frac{\partial^2 \ell}{\partial \delta_{j''k''} \partial \alpha_{j'}}(\boldsymbol{\theta}) = \mathbb{1}_{j'=j''} \exp\left(\mu + \alpha_{j''} + \beta_{k''} + \delta_{j''k''}\right) R_{j''}^{k''},$$

$$\frac{\partial^2 \ell}{\partial \delta_{j''k''} \partial \beta_{k'}}(\boldsymbol{\theta}) = \mathbb{1}_{k'=k''} \exp\left(\mu + \alpha_{j''} + \beta_{k''} + \delta_{j''k''}\right) R_{j''}^{k''},$$

$$\frac{\partial^2 \ell}{\partial \beta_{k''} \partial \beta_{k'}}(\boldsymbol{\theta}) = \mathbb{1}_{k'=k''} \sum_{j=1}^{J} \exp\left(\mu + \alpha_j + \beta_{k''} + \delta_{jk''}\right) R_{j}^{k''},$$

$$\frac{\partial^2 \ell}{\partial \delta_{j''k''} \partial \delta_{j'k'}}(\boldsymbol{\theta}) = \mathbb{1}_{j'=j'',k'=k''} \exp\left(\mu + \alpha_{j''} + \beta_{k''} + \delta_{j''k''}\right) R_{j''}^{k''},$$

with $1 \leq j', j'' \leq J$ and $1 \leq k', k'' \leq K$.

Likelihood with piecewise constant regions. Let Z_l define the l-th region, for $1 \leq l \leq L$. Let ℓ^{c} be the constrained negative log-likelihood function, depending on the parameter vector $\boldsymbol{\theta} = \left(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^{(1)}, \ldots, \delta^{(L)}\right)$ of size $1 + J - 1 + K - 1 + L$. Then

$$\lambda_{j,k} = \exp\left(\mu + \alpha_j + \beta_k + \sum_{l=1}^{L} \mathbb{1}_{(j,k) \in Z_l} \delta^{(l)}\right)$$

and

$$\ell^{\mathrm{c}}(\boldsymbol{\theta}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \lambda_{j,k} R_{j,k} - \log\left(\lambda_{j,k}\right) O_{j,k}. \tag{3.10}$$

Then

$$\frac{\partial \ell^{\mathrm{c}}}{\partial \delta^{(l')}}(\boldsymbol{\theta}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \mathbb{1}_{(j,k) \in Z_{l'}} \left(\lambda_{j,k} R_{j,k} - O_{j,k}\right)$$

$$\frac{\partial^2 \ell^{\mathrm{c}}}{\partial \mu \partial \delta^{(l')}}(\boldsymbol{\theta}) = \sum_{j=1}^{J}\sum_{k=1}^{K} \mathbb{1}_{(j,k)\in Z_{l'}} \lambda_{j,k} R_{j,k}$$

$$\frac{\partial^2 \ell^{\mathrm{c}}}{\partial \alpha_{j'} \partial \delta^{(l')}}(\boldsymbol{\theta}) = \sum_{k=1}^{K} \mathbb{1}_{(j',k)\in Z_{l'}} \lambda_{j',k} R_{j',k}$$

$$\frac{\partial^2 \ell^{\mathrm{c}}}{\partial \beta_{k'} \partial \delta^{(l')}}(\boldsymbol{\theta}) = \sum_{k=1}^{K} \mathbb{1}_{(j',k)\in Z_{l'}} \lambda_{j,k'} R_{j,k'}$$

$$\frac{\partial \ell^{\mathrm{c}}}{\partial \delta^{(l')} \partial \delta^{(l'')}}(\boldsymbol{\theta}) = \mathbb{1}_{l'=l''} \sum_{j=1}^{J}\sum_{k=1}^{K} \mathbb{1}_{(j,k)\in Z_{l'}} \lambda_{j,k} R_{j,k}$$

Penalized Likelihood. The penalized likelihood's derivatives with respect to μ, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ are unchanged. For $2 \le j' \le J$ and $2 \le k' \le K$ we have:

$$\begin{aligned}\frac{\partial \ell^{\kappa}}{\partial \delta_{j'k'}}(\boldsymbol{\theta}) = \frac{\partial \ell}{\partial \delta_{j'k'}}(\boldsymbol{\theta}) &+\kappa\left[-v_{j',k'}\left(\delta_{j'+1,k'}-\delta_{j',k'}\right)+v_{j'-1,k'}\left(\delta_{j',k'}-\delta_{j'-1,k'}\right)\right]\\ &+\kappa\left[-w_{j',k'}\left(\delta_{j',k'+1}-\delta_{j',k'}\right)+w_{j',k'-1}\left(\delta_{j',k'}-\delta_{j',k'-1}\right)\right],\end{aligned} \tag{3.11}$$

with by convention $\delta_{J+1,k} = \delta_{j,K+1} = 0$ and $v_{J,k} = w_{j,K} = 0$. For $2 \le j' \le J$, $2 \le k' \le K$, $2 \le j'' \le J$ and $2 \le k'' \le K$ we have:

$$\begin{aligned}\frac{\partial^2 \ell^{\kappa}}{\partial \delta_{j',k'} \partial \delta_{j'',k''}}(\boldsymbol{\theta}) = \frac{\partial^2 \ell}{\partial \delta_{j',k'} \partial \delta_{j'',k''}}(\boldsymbol{\theta}) &+\kappa\left[\mathbb{1}_{j''=j',k''=k'}\left(v_{j',k'}+v_{j'-1,k'}+w_{j',k'}+w_{j',k'-1}\right)\right.\\ &-v_{j',k'}\mathbb{1}_{j''=j'+1,k''=k'}-v_{j'-1,k'}\mathbb{1}_{j''=j'-1,k''=k'}\\ &\left.-w_{j',k'}\mathbb{1}_{j''=j',k''=k'+1}-w_{j',k'-1}\mathbb{1}_{j''=j',k''=k'-1}\right]\end{aligned} \tag{3.12}$$

Appendix B: Fast Inversion of the Hessian matrix

The Newton-Raphson iteration requires the computation of $\left(I_n^{\kappa}\right)^{-1} U_n^{\kappa}$. This linear problem is the computational bottleneck of the algorithm. The Hessian matrix has the form

$$I_n^{\kappa} = \begin{pmatrix} A & B \\ B^T & D \end{pmatrix}$$

where

$$A = \frac{\partial^2 \ell_n^{\kappa}}{\partial(\mu,\boldsymbol{\alpha},\boldsymbol{\beta})\partial(\mu,\boldsymbol{\alpha},\boldsymbol{\beta})^T}, \quad B = \frac{\partial^2 \ell_n^{\kappa}}{\partial(\mu,\boldsymbol{\alpha},\boldsymbol{\beta})\partial\boldsymbol{\delta}^T}, \quad \text{and} \quad D = \frac{\partial^2 \ell_n^{\kappa}}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^T}.$$

The Hessian matrix is inverted using the Schur block matrix inversion formula (Zhang, 2005):

$$\left(I_n^{\kappa}\right)^{-1} = \begin{pmatrix} S^{-1} & -S^{-1}BD^{-1} \\ -D^{-1}CS^{-1} & D^{-1}+D^{-1}CS^{-1}BD^{-1}, \end{pmatrix}$$

where $S = A - BD^{-1}C$ is the Schur complement of the matrix D. Consequently

$$(I_n^{\kappa})^{-1} U_n^{\kappa} = \begin{pmatrix} S^{-1}U_1 - S^{-1}BD^{-1}U_2 \\ -D^{-1}B^T S^{-1}U_1 + D^{-1}U_2 + B^T S^{-1}BD^{-1}U_2 \end{pmatrix}, \tag{3.13}$$

where $U_1 = \partial \ell_n^{\kappa} / \partial\left(\mu,\boldsymbol{\alpha},\boldsymbol{\beta}\right)$ and $U_2 = \partial \ell_n^{\kappa} / \partial\boldsymbol{\delta}$. Since D is a symmetric band diagonal matrix, $D^{-1}B^T$ is computed using Cholesky factorization and the computation complexity of $\left(I_n^{\kappa}\right)^{-1} U_n^{\kappa}$ using Equation (3.13) reduces from $\mathcal{O}\left((JK)^3\right)$ to $\mathcal{O}\left(\max(J,K)^2 JK\right)$.

Chapter 4

Spline regression with automatic knot selection

The problem of finding a function f that links a (potentially multivariate) explanatory variable $\boldsymbol{x}$ to a univariate response variable $\boldsymbol{y}$ is central in statistics. In all generality the link between $\boldsymbol{x}$ and $\boldsymbol{y}$ is not affine and the linear model is often deemed too restrictive. Often, the link between $\boldsymbol{x}$ and $\boldsymbol{y}$ is supposed complex and is unknown, and we do not want to make restrictive assumptions on f. This is the framework of non-parametric regression, where the unknown function f is assumed to belong to a family of functions that cannot be configured by $\mathbb{R}^p$ (or a subset thereof). Instead, we infer the function from the data without additional assumptions on f.

We count four important tools for non-parametric regression: Gaussian process regression, kernel regression, spline regression, and regression trees. Among those, spline regression has been praised for being simple and efficient in many practical cases. This domain has grown in popularity since the development of computational statistics in the years 1980's and has gained mainstream appeal in the years 2000's. Strictly speaking, spline regression does not fall inside the category of non-parametric regression, because splines are families of parametric functions. They are still termed non-parametric because *(i)* they are over-parametrized and each parameter has no interpretability inside the model and *(ii)* splines with a sufficiently large number of parameters can approximate any function inside some infinite-dimensional function space with arbitrary precision (the details are beyond the scope of this work, see Curry and Schoenberg, 1966). Some refer to spline regression as being semiparametric, which is also technically incorrect.

Spline regression is simple in practice: we set the number of knots, their position, and the order of the spline (knots are positions on the x-axis where the spline displays a discontinuity in its derivative and a spline's order accounts for its regularity). Conditionally on this choice, the regression is a parametric method and is easy to compute. However, with sufficiently large order and number of knots, the spline approximates any function inside a very large (i.e. infinite-dimensional) function set. This is the simplifying trick which makes spline regression both simple and powerful.

In practice, performing regression with one choice of knots (and order) is not enough: we need to select the best number and position of knots and order *a posteriori*. This is a complicated problem. The order of the spline is always chosen by the statistician. With sufficiently enough knots, a spline of sufficiently high order can approximate any spine of lower order. Thus the order is chosen as the lowest value such that the fit is deemed satisfactory (rarely more than 5 in practice). The knots are however harder to choose. The regression will change drastically if there are too many knots or if they are not well placed. This is the hidden cost of spline regression. Two main solutions have been brought forward. The first avoids the problem by voluntarily over-parametrizing the splines and taking overfitting into account with penalization. The second tackles the problem of optimizing with respect to the knots, either through a Bayesian framework or through a well-chosen optimization scheme.

In this chapter, we introduce a novel approach to automatic spline selection. It consists in jointly

selecting the number and position of knots as well as the fitted spline. The method relies on the adaptive ridge and we iterate between penalized spline regressions to find the best unpenalized regression spline. We iteratively remove the least relevant knots from a large initial collection of knots. This approach takes a simplifying approach to the problem, but the computational burden is vastly reduced and simulations show that the resulting fit has good prediction performance.

This chapter is organized as follows. In the first part, we introduce splines and spline regression through its different methods, with emphasis on interpretation and computational cost. We elaborate on the different unpenalized and penalized spline regression methods. The paper "Spline Regression with Automatic Knot Selection" is given in the second part.

We refer to Green and Silverman (2000), Härdle (1990), and Wahba (1990) for reviews on spline regression.

Contents

4.1 Introduction to spline regression 91
4.1.1 Spline regression 91
4.1.1.1 Definition of splines 91
4.1.1.2 The truncated power basis 92
4.1.1.3 B-spline basis 93
4.1.1.4 Regression using splines 95
4.1.2 Penalized approaches 97
4.1.2.1 Smoothing splines 97
4.1.2.2 O'Sullivan Penalized Splines 98
4.1.2.3 P-splines 98
4.2 Spline regression with automatic knot selection 99
4.3 Comparison of A-spline with P-spline 129

4.1 Introduction to spline regression

4.1.1 Spline regression

We consider the problem of regressing a real-valued response variable $\boldsymbol{y} = (y_i)_{1\leq i\leq n}$ onto a real-valued explanatory variable $\boldsymbol{x} = (x_i)_{1\leq i\leq n}$, where n is the sample size. We assume that there exists an unknown function f such that

$$y_i = f(x_i) + \varepsilon_i, \quad 1 \leq i \leq n, \tag{4.1}$$

where ε_i are realizations of a centered random variable $\boldsymbol{\varepsilon}$, not necessarily independent. The values of the x_is are assumed to be inside an interval $[a,b]$. Without loss of generality, we order the $\boldsymbol{x}$ variable, so that $x_1 \leq \cdots \leq x_n$. In this setting, the x variable is considered deterministic and the aim of regression is to find an approximation $\hat{f}$ of the f that links $\boldsymbol{x}$ to $\boldsymbol{y}$. This problem being ill-posed, we must assume that f satisfies some additional property. This property is most often a level of smoothness, for instance, an order of differentiability.

4.1.1.1 Definition of splines

Splines are a parametric family of functions that are smooth and general enough to approximate a wide range of functions. Splines are defined using a series of knots, which are base points for the spline located on the x-axis. Let $q \geq 1$ denote the number of knots. Let $\boldsymbol{t} = (t_j)$ define the knots in increasing order, such that $a \leq t_1 < \cdots < t_q \leq b$. Furthermore, we denote by $k \geq 1$ the order of the spline. For $j \in \{1, \cdots, q-1\}$, the splines of order k are the functions defined on $[a,b]$ which are piecewise polynomials of degree $k-1$ on each interval $[t_j, t_{j+1})$ and whose $k-2$-th derivatives are defined and continuous at each knot t_j (we use the convention that a function's derivative of a negative order is the function itself). For illustration, splines of order 1 are piecewise constant functions with possible discontinuities at each t_j; we easily verify that these functions are polynomials of degree 0 on each interval $[t_j, t_{j+1})$ and that they are possibly discontinuous at the knots. Splines of order 2 are broken lines, i.e. piecewise linear functions which are continuous but possibly non-differentiable at each knot t_j. We can easily verify that these are the functions that are polynomials of degree 1 on each knot interval and are continuous everywhere.

The set of splines of order k defined over the same knots are a linear space of dimension $k+q$ (this property is easy to verify and we omit the proof). Moreover, splines are slightly more general functions than polynomials, which allows them to approximate a wider range of functions, in the

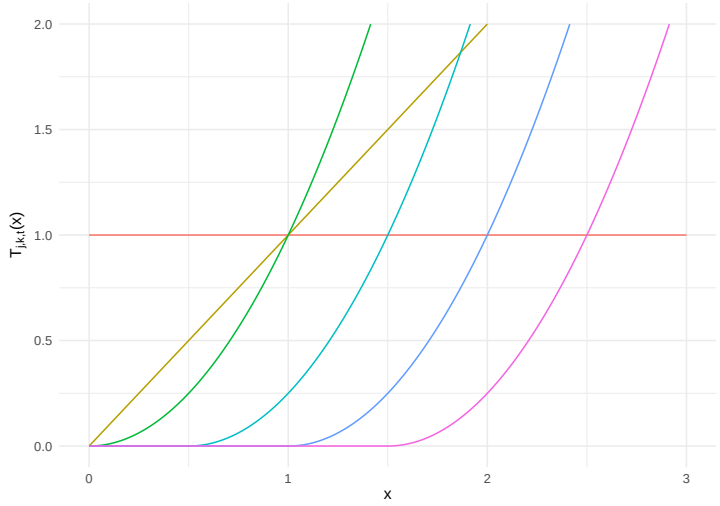


Figure 4.1: Representation of the truncated power basis $(T_{j,k,\boldsymbol{t}}(x))_{1\leq j\leq k+q}$ for $k=3$ and $\boldsymbol{t}=(0.5,1,1.5)$.

sense of interpolation. It turns out that polynomials of degree smaller of equal than a constant $k \geq 0$ are a poor choice of function family for interpolation. For example, polynomial interpolation is globally sensitive to a local perturbation of the data, while spline interpolation will only change locally if the data is perturbed locally. See (De Boor, 1978, Chapter 2) for a detailed discussion of the drawback of polynomial bases for function approximation.

In the following, we will use splines as approximating functions $\hat{f}$ for the regression (4.1). We stress that since few knots are used ($q \leq 40$ typically), and with an order k typically smaller than 5, the spline linear space has a low dimension. This makes the regression problem a parametric estimation problem, in a parameter space of small dimension. We first have to define a well-suited base of the spline linear space.

4.1.1.2 The truncated power basis

The truncated power basis is a basis for the spline linear space. For fixed q, $\boldsymbol{t}$ and k, the j-th truncated power basis spline if defined as

$$T_{j,k,\boldsymbol{t}}(x) = \begin{cases} x^j & \text{if} \quad 0 \leq j \leq q-1 \\ (x-t_{j-k})_+^{q-1} & \text{if} \quad k+1 \leq j \leq k+q. \end{cases} \tag{4.2}$$

We recall that the notation $(x)_+$ denotes the positive part $\max(x,0)$; hence the naming of *truncated power* basis. The indices t, k, and $\boldsymbol{t}$ will be dropped from the notation $T_{j,k,\boldsymbol{t}}$ when they can be inferred from the context.

We will show that these functions form a basis of the spline space. It is easy to see that it has $k+q$ elements who are linearly independent. It remains to show that each element of the basis is a spline. For $j \in \{1,\cdots,k-1\}$, $T_{j,k,\boldsymbol{t}}$ is a spline of order k because is it a polynomial of degree $k-1$. For $j \in \{k,\cdots,k+q\}$, $T_{j,k,\boldsymbol{t}}$ is polynomial on both $[a,t_j]$ and $[t_j,b]$ and it is continuously differentiable on $[a,b]$ up to the order $k-1$. This completes the proof.

Numerical stability issues. An example of truncated power basis is given in Figure 4.1, for $\boldsymbol{t} = (t_1,t_2,t_3),= (0.5,0.1,0.15)$ and $k=3$. The truncated power basis is then composed of the $k+q=6$ elements $x^0=1$, $x^1=x$, x^2, $(x-t_1)_+^2$, $(x-t_2)_+^2$, and $(x-t_3)_+^2$.

The figure also illustrates the major problem with the truncated power basis: it yields poorly conditioned problems. We develop on this issue here. When the number of knots increases, two consecutive knots t_j and t_{j+1} can be so close that the truncated basis functions $(x-t_j)_+^q$ and $(x-$

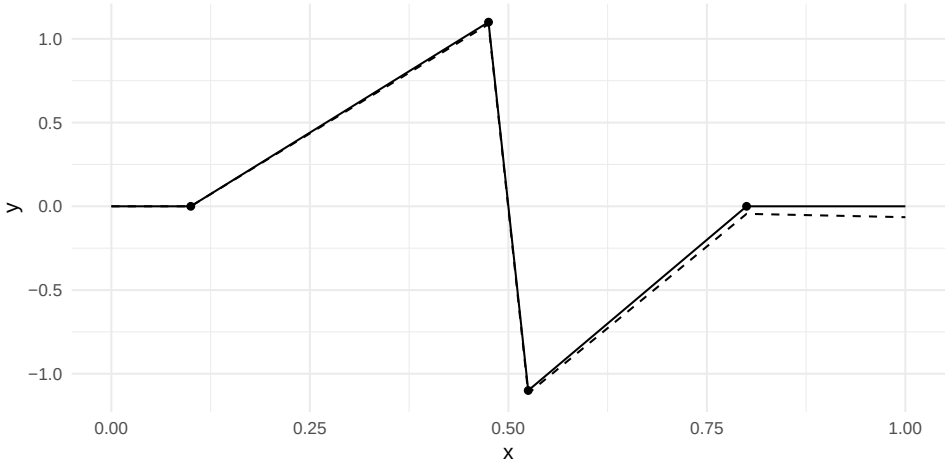


Figure 4.2: Illustration of the numerical instability of the decomposition in truncated power basis, with a spline of order 2 supported by 3 knots (tick line) and its decomposition with 2-digit machine precision coefficients (dotted line).

$t_{j+1})_+^q$ become very highly correlated. As is shown after, the fitting of splines to a data set (x_i, y_i) is done by linear regression. The design matrix of this linear regression is the $n \times k+q$ matrix composed of the basis functions $T_{j,k,\boldsymbol{t}}(x_i)$ evaluated at the data points. When the columns of this matrix are too close to being correlated, numerical precision issues occur. This is due to the subtraction of numbers that are potentially large and have almost opposite values. In the truncated power basis, the base functions span over a large number of knots, so the error made on one coefficient has an effect on whole fitted function (more precisely, on the right side of the corresponding knot). This issue makes the representation of spline in the truncated power basis prone to numerical instabilities, and thus unfit to computation. As an illustrating example, consider the example adapted from (De Boor, 1978, page 85) of a spline of order $q = 2$ defined on $[a, b] = [0, 1]$ and supported over the knots $(t_1, t_2, t_3, t_4) = (0.1, 0.5 - h/2, 0.5 + h/2, 0.8)$, with values $(0, 1.1, -1.1, 0)$ at the knots, and where h is a small constant. This spline allows the following decomposition in the truncated power basis:

$$f(x) = \alpha_1(x-t_1)_+ + \alpha_2(x-t_2)_+ + \alpha_3(x-t_3)_+ + \alpha_4(x-t_4)_+, \quad \text{where} \quad \begin{cases} \alpha_1 = & \frac{1.1}{t_2-t_1} \\ \alpha_2 = & -\left(\frac{2.2}{h} + \frac{1.1}{t_2-t_1}\right) \\ \alpha_3 = & \left(\frac{2.2}{h} + \frac{1.1}{t_4-t_3}\right) \\ \alpha_4 = & -\left(\frac{1.1}{t_4-t_3}\right). \end{cases}$$

When h gets small, we have $\alpha_2 \simeq -\alpha_3$, and both get large in absolute value. With an illustrative machine precision of 2 significant numbers and $h = 0.05$, we have $\alpha_1 \simeq 2.9$, $\alpha_2 \simeq -47$, $\alpha_3 \simeq 48$, and $\alpha_4 \simeq -4$. Figure 4.2 illustrates the loss of quality suffered by this decomposition.

This makes the truncated power basis unfit in practice. In the next section, we introduce the spline basis which solves all the pitfalls that the truncated power basis has.

4.1.1.3 B-spline basis

The B-spline basis is a reparametrization of the truncated power basis. It is defined by recurrence, for fixed order $k \leq 1$ and fixed knots $\boldsymbol{t}$, as

$$\begin{aligned} B_{j,1,\boldsymbol{t}} &= \mathbb{I}\{t_j \leq x < t_{j+1}\} \\ B_{j,k,\boldsymbol{t}}(x) &= \frac{x - t_j}{t_{j+k-1} - t_j} B_{j,k-1,\boldsymbol{t}}(x) + \frac{t_{j+k} - x}{t_{j+k} - t_{j+1}} B_{j+1,k,\boldsymbol{t}}(x), \quad \text{for} \quad k > 1, \end{aligned} \tag{4.3}$$

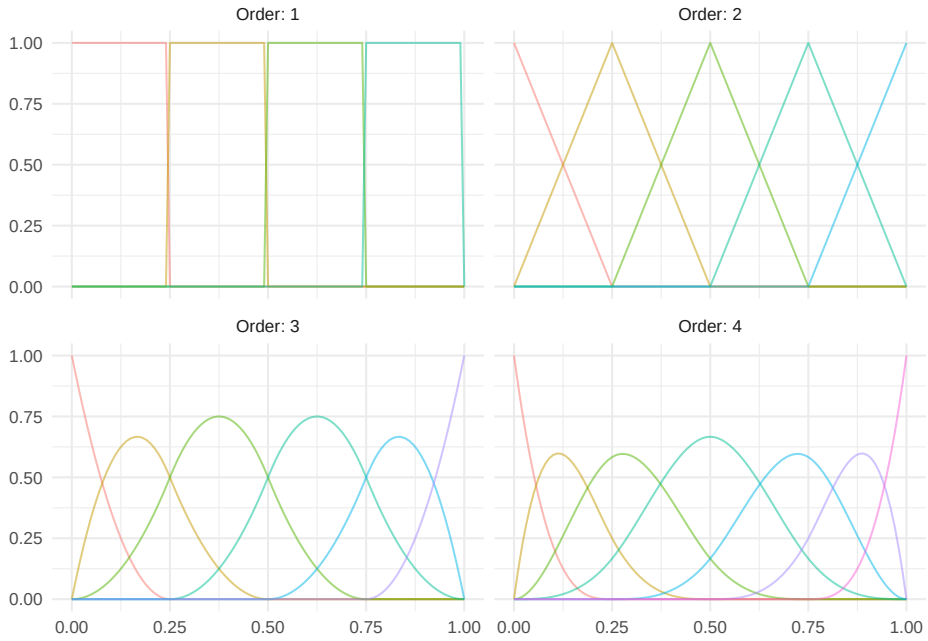


Figure 4.3: B-spline bases of order 1 to 4 with three equally spaced knots $\boldsymbol{t} = c(0.25, 0.5, 0.75)$.

where by convention $t_0 = a$ and $t_{q+1} = b$. We can easily verify by recurrence that this formula defines indeed splines. There are $k+q$ B-splines of order k, which form a base. The suffixes of the splines are dropped when they can be inferred from context.

Consider first the B-spline basis of order 1: they are the characteristic functions of the inter-knot intervals $[t_j, t_{j+1})$. Compare now these functions to the truncated power basis of order 1: $T_{j,1,\boldsymbol{t}} = \mathbb{I}\{t_j \leq x\}$. Here, the B-splines are equal to the consecutive difference of two truncated power splines. This allows them to have a small support and to have have the same order of magnitude. Likewise, the B-splines of order 2 are triangular functions such that $B_{j,2}(t_{j-1}) = 1$, $B_{j,2}(x) = 0$ for $x \notin [t_{j-2}, t_j)$. We can show that the B-splines have the following properties for any k:

- $B_{j,k}(x)$ has exactly the support $[t_{j-k}, t_j) \cap [a, b]$.
- At any point $x \in [a, b]$, there are exactly k B-splines with non-zero values.
- The B-spline base is normalized, i.e. at any point $x \in [a, b]$, $\sum_{j=1}^{k+q} B_{j,k}(x) = 1$.

This makes the representation of splines in this base better fitted to computations. More precisely

- The decomposition coefficients of a spline are of the same order of magnitude as the spline.
- These coefficients represent the local value of the spline, and have no global effect on the spline.

This makes B-splines optimally fitted to spline representation. For instance, in the example given in the last section (see Figure 4.2), B-spline decomposition recovers perfectly the initial spline.

For the sake of completeness, we give an illustration of B-spline bases of different orders and with equally spaces knots in Figure 4.3.

Computation of B-splines. The recurrence formulas defining B-splines are directly used to compute B-splines. De Boor (1978) gave an algorithm to compute B-splines of any order based on B-spline of lower order (recall that splines of order 1 are piecewise constant). This method is efficient for computing B-splines and most B-spline implementations use FORTRAN routines implementing de Boors' algorithm (de Boor, 1977).

4.1.1.4 Regression using splines

Spline regression as linear regression. We want to solve the regression problem (4.1) using a spline to estimate f. Given an order and the sequence of knots, we define

$$\hat{f}(x) = \sum_{j=1}^{k+q} \alpha_j B_j(x), \tag{4.4}$$

where $\boldsymbol{\alpha} = (\alpha_j)$ is the vector parameter which gives the decomposition of $\hat{f}$ in B-splines. We chose the L_2 norm as our loss function, i.e. we select the estimated function which minimizes the sum of square

$$\sum_{i=1}^{n} (y_i - f(x_i))^2.$$

Notice now that from this equation and (4.4), the criterion to minimize writes

$$\left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha} \right\|^2,$$

where the $n \times (k+q)$ design matrix $\boldsymbol{X}$ has (i,j)-th entry $B_j(x_i)$. This is the framework of linear regression, whose estimate is explicit:

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

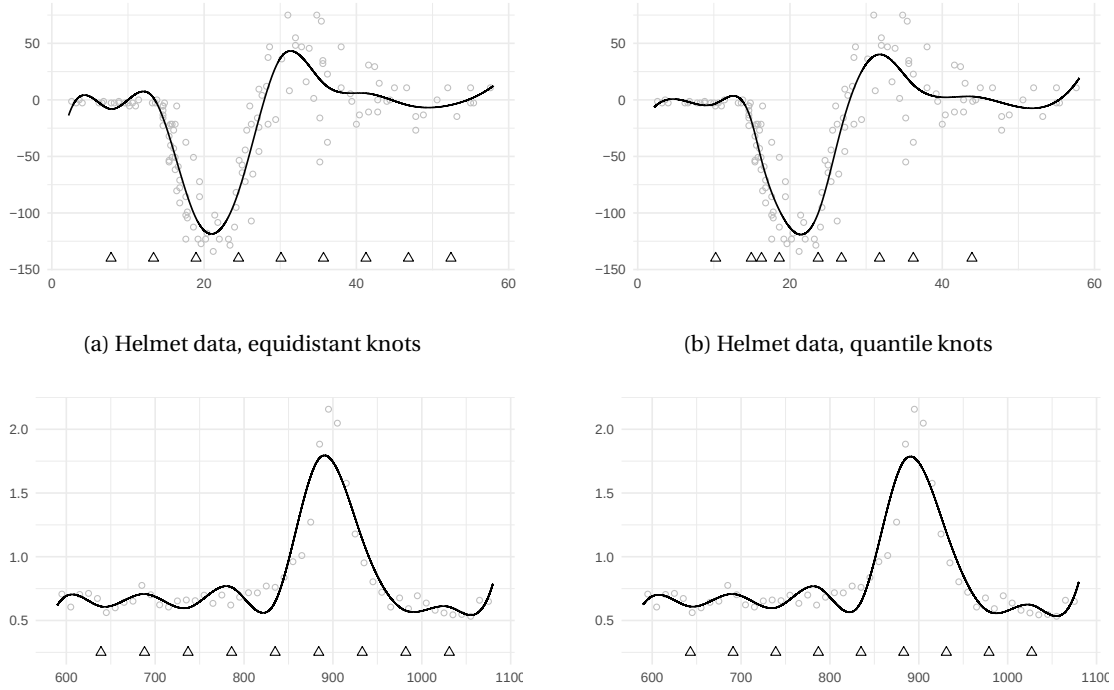
From the aforementioned properties of B-splines, the design matrix has the following form. For $j \in \{0, \cdots, q\}$ let n_j be the number of points x_i in the interval $[t_j, t_{j+1})$ (we have $\sum_j n_j = n$). Then $\boldsymbol{X}$ has non-zero values on blocks of size $n_j \times k$ ordered one above the other, each block being shifted by one entry compared to the previous block. An easy calculation shows that $\boldsymbol{X}^T\boldsymbol{X}$ is banded, of bandwidth k: its (j,j')-th entry is zero if $|j - j'| > k$.

Numerical efficiency. As noted by (Hastie et al., 2001, Chap. 5, Appendix), the numerical cost is vastly reduced in this setting. Its inversion is computed efficiently in the following steps: *(i)* compute the Cholesky decomposition $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{L}^T$, with complexity $O(nk^2)$, *(ii)* solve $\boldsymbol{L}\tilde{\boldsymbol{x}} = \boldsymbol{X}^T\boldsymbol{y}$ for $\tilde{\boldsymbol{x}}$ using forward substitution, in complexity $O(n)$, and *(iii)* solve $\boldsymbol{L}^T\boldsymbol{x} = \tilde{\boldsymbol{x}}$ for $\boldsymbol{x}$ using back substitution, in the same complexity. Overall, smoothing spline regression is computed in $O(nk^2)$ time complexity. In practice, k is small (often set to 4) and so fitting B-splines is fast.

Remark. In the following, we will discuss about the properties of the underlying function f. Indeed, the order of the spline to choose and the number and location of knots all depend on the shape of f, as does the quality of the regression. In this chapter, we say that a function is *regular*, or *smooth*, if is has few non-zero high-order derivatives. This is a local notion, which means that a function can be smoother in one place of the segment $[a,b]$ and less smooth in another. Since no assumption is made on f, we don't give a proper mathematical definition of *regularity*; it is used to discuss the difficulties of spline regression.

Choice of the knots. Spline regression is fast and easy to implement and it provides a flexible enough setting to approximate a large spectrum of functions. But the quality of the fit depends on the choice of the knots. With splines of order 1 and 2, the knots ought to be placed at the cut-points in the data, which can be selected by visual inspecting (although this method is not data-driven and highly arbitrary). For splines of higher order, the selection criterion is more difficult to choose.

It turns out that equally-spaced knots is not always a good choice, because the function f may have more variability in one region than in another. One may choose equally spaced knot, which seems to be a "neutral" setting, and which works well when the true function has enough regularity. However, a spline with too many knots will *(i)* be prone to "wiggliness" and *(ii)* be overfitted, because as the number of knots gets close to the number of data points, the fit becomes an interpolation.



(a) Helmet data, equidistant knots

(b) Helmet data, quantile knots

(c) Titanium data, equidistant knots

(d) Titanium data, quantile knots

Figure 4.4: Spline regression with 9 knots placed at equi-distance (c, d) and at the quantiles (a, b), for two different data sets.

A spline with more knots has more flexible fitting to the data, i.e. it can fit to functions with high variability (in the sense that it has higher order derivatives with important values). Since B-splines are locally adaptive, a region with more knots will accommodate high variability of the function in this region. Thus, we would want to choose the knots in an adaptive, data-driven way that puts more knots when it is necessary to fit the high variability, and less when the function is smoother. This is a difficult task.

A possible solution is to place the knots at the quantiles of (x_i), allowing for higher adaptability where there are many data points. But this is efficient only if highly-varying regions have more data points, which is not always verified. We illustrate this problem on two different datasets (Figure 4.4). The "helmet" data (see de Souza and Heckman, 2013) set is helmet crash test data, representing acceleration against time ($n = 132$). It is repeatedly used to illustrate spline regression methods and compare them (Silverman, 1985; Eilers and Marx, 1996). The "titanium" data represents a heat property of titanium as a function of its temperature ($n = 49$) (used for instance by Dierckx, 1995; de Boor, 1986; Jupp, 1978). Since this data contains close to no noise, we add a Gaussian noise of standard deviation 0.05 to the y values in order to make the inference reasonably challenging. Compare now Figures 4.4b and 4.4b, which display spline regressions of order 4 with 9 knots placed at the quantiles, for two different data sets. In the first one, the x_is are not spread uniformly, and the sharp shift at $x \sim 17$ is where the x_is are the most dense. The knots set at the quantiles allow for many knots in this region, and the fit is of good quality comparatively to equidistant knots (Figure 4.4a). On the other one, the x_is are uniformly spread, and there is a high variability in one region. This lead a poor fit (at least as poor as with equidistant knots, see Figure 4.4c).

The problem of finding the "best" position of the knots is a difficult one. In the next section, we present a few solutions present in the literature.

4.1.2 Penalized approaches

In this part, we will see some methods that circumvent the difficulty of choosing the knot's position. Instead, they set a (too) large number of knots and deal with overfitting through penalized estimation. They all differ in their choice of knot number and placement and type of penalization, but essentially follow this simple principle.

Natural cubic splines. We first introduce a family of spline called *natural cubic spline*.

Spline of order 2 are continuous, splines of order 3 are continuous and have continuous first order derivatives, splines of order 4 are continuous and have continuous first and second order derivatives. For order ≤ 3, the splines' shifts at the knots are visible. For order 4 and larger, it becomes almost impossible to tell where the knots are: the splines are seamless. This is why in practice splines of order 4 are chosen when f is assumed "smooth" without any other assumption. They are called *cubic* splines – they are made of piecewise cubic polynomials.

Around the boundary of $[a,b]$, there are fewer data points then in the center of $[a,b]$. Consequently, the regression has more variance close to the margins a and b. To remedy this issue, one solution is to add constraints at the boundary. For cubic splines, we can add the constraints that the spline be linear on $(-\infty, t_1]$ and $[t_q, \infty)$, i.e. that the second and third order derivatives are zero left of t_1 and right of t_q. This adds 4 constraints on the spline, which can now be parametrized by $k+q-4=q$ freely varying parameters.

Definition 4. *A natural cubic spline is a spline of order* 4 *linear outside of its boundary knots.*

4.1.2.1 Smoothing splines

Consider the problem of finding the function $\hat{f}$ which solves

$$\arg\min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_{\mathbb{R}} \left[f''(t)\right]^2 dt \tag{4.5}$$

where f is any function such that this quantity is defined and $\lambda > 0$ is a smoothing parameter. Consequently, f belongs to the Sobolev space of functions whose derivatives of order 0, 1, and 2 are square-integrable. Equation 4.5 is minimization problem in infinite dimension.

Fortunately, the solution to (4.5) is proven to be a natural cubic spline with n knots placed at the x_is. These splines are called *smoothing splines*. Consider a basis $(N_j(x))_{1\leq j\leq n}$ for the space of cubic splines defined on the knots (x_i) and write $\hat{f}(x) = \boldsymbol{\alpha}^T \boldsymbol{N}(x)$ (we can define such a basis with properties similar to B-splines. We do not give the details here for the sake of the presentation and refer to Hastie et al., 2001, Section 5.2.2). Then problem (4.5) rewrites

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\alpha}^T \boldsymbol{\Omega} \boldsymbol{\alpha} \tag{4.6}$$

where $\boldsymbol{X}$ is the design matrix of the values of $N_j(x_i)$ and

$$\Omega_{i,j} = \int_{\mathbb{R}} N_i''(t) N_j''(t) dt \tag{4.7}$$

is the $n \times n$ symmetric matrix enforcing the smoothing of the spline. Consequently, the smoothing spline is not over-parametrized, because its estimated vector is the solution of a ridge regularized regression:

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{\Omega})^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

Since $\boldsymbol{\Omega}$ is positive definite, it acts as a regularization of the estimate, in a way similar to ridge regression (if $\boldsymbol{\Omega}$ were diagonal, the previous equation would be a weighted ridge regression). Note that $\boldsymbol{\Omega}$ acts as a weighting matrix and it depends on $\boldsymbol{x}$ and λ, but not on $\boldsymbol{y}$. Hence the term *smoothing spline*:

the projection matrix $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ that maps $\boldsymbol{y}$ to $\hat{\boldsymbol{y}}$ is obtained by a smoothing independent from $\boldsymbol{y}$.

It comes that $\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{\Omega}$ is symmetric definite positive and 4-banded. Since smoothing splines are of order 4, we know that $\boldsymbol{X}^T\boldsymbol{X}$ is also 4-banded. Consequently smoothing splines have the same computational cost as B-splines.

4.1.2.2 O'Sullivan Penalized Splines

O'Sullivan (1986) introduced a penalized regression spline that it closely related to smoothing splines. Indeed, the method (called O'Sullivan splines for clarity) uses cubic B-splines (although any order could be used without modifications to the method) with $k+4$ knots located anywhere on $[a,b]$. The estimate is the cubic spline

$$f(x)=\sum_{j=1}^{k+4}\alpha_j B_{j,4}(x)$$

Consider the $n\times(k+4)$ design matrix $\boldsymbol{X}$ with (i,j)-th entry $B_j(x_i)$. O'Sullivan penalized splines use the same penalization as smoothing splines, but in the context of B-splines (instead of natural splines). Likewise (4.5), the estimate is defined as

$$\underset{\boldsymbol{\alpha}}{\arg\min}\left\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\alpha}\right\|+\lambda\int_{x_{\min}}^{x_{\min}}\left\{\sum_{j=1}^{k+4}\alpha_j B''_{j,4}(t)\right\}^2 dt, \tag{4.8}$$

and likewise smoothing splines, the estimate reduces to

$$\hat{\boldsymbol{a}}=(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}\boldsymbol{y}, \tag{4.9}$$

where the penalty matrix $\boldsymbol{\Omega}$ is defined by (4.7), where N_j is replaced by $B_{j,4}$. The rationale behind (4.8) is similar to smoothing splines. Using many knots creates wiggliness in the spline. For a spline of order 4, the penalization of its integrated squared second order derivative forces it to have a small second order part and to have small third order derivative differences around each knot.

Consequently, smoothing splines are a particular case of O'Sullivan splines, with $k=n$ knots placed at the data points x_i. Whereas smoothing splines perform a trade-off between interpolation (for $\lambda\to 0$) and linear regression (for $\lambda\to\infty$), this method performs a trade-off between unpenalized B-splines (for $\lambda\to 0$) and linear fit ($\lambda\to\infty$). It enjoys the same computational efficiency as smoothing splines, since the penalty function is also 4-banded. It is in fact more efficient since we usually define way fewer knots then there are data points ($n\leq k$). This makes O'Sullivan splines a tool of choice for penalized spline regression. In the next section, a further simplification of this method is presented.

4.1.2.3 P-splines

Eilers and Marx (1996) introduced a penalized spline regression method called *P-splines* (for *penalized splines*). Its simplicity made it the most widely used penalized spline regression model, having been used in many applications and extended to more general settings (Eilers et al., 2015; Currie et al., 2004; Wand and Ormerod, 2010; Jiang and Carriere, 2014).

Consider the same framework as the previous section: the regression spline is a B-spline of any order k. P-splines are defined with knots equidistant over $[a,b]$.

Recall that $\Delta\alpha_j\triangleq\alpha_j-\alpha_{j-1}$ denotes the difference operator and $\Delta^p\triangleq\Delta^{p-1}\circ\Delta$, $p\geq 2$ denotes the p-th composition of the difference operator with itself. P-splines are defined as the solution of the penalized problem

$$\underset{\boldsymbol{\alpha}}{\arg\min}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\alpha}\|^2+\lambda\sum_{j=k}^{q+k}(\Delta^{k-2}\alpha_j)^2. \tag{4.10}$$

This penalty has an intuitive explanation, at least with the order 2. In this paragraph, consider $q=2$, i.e. piecewise linear functions. Recall that $\hat{f}=\sum_j\alpha_j B_{j,2}(x)$ and that over $[t_{j-1},t_j)$, $B_{j_0,2}(x)$ has

slope -1, $B_{j_0-1,2}(x)$ has slope 1 and all the other B-splines are equal to zero. Then at t_{j_0}, the slope shift of $\hat{f}(x)$ is $\hat{\alpha}_{j_0} - \hat{\alpha}_{j_0-1}$. When $(\Delta_2 \hat{\alpha}_{j_0})^2$ is small, $\hat{\alpha_{j_0}} \simeq \hat{\alpha}_{j_0-1}$, and the slope shift of $\hat{f}(x)$ at t_{j_0-1} vanishes. This enforces the spline to have small variations in slope at each knot, and throughout $[a,b]$. This result generalizes to other orders, and in particular cubic splines will have small differences in their third order derivatives.

The penalized problem (4.10) gives the explicit estimate

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

where $\boldsymbol{D}$ is the $q \times (q+k)$ matrix of the operator Δ^{k-2}. The entries of $\boldsymbol{D}$ are the signed binomial coefficients. Consequently, $\boldsymbol{D}^T\boldsymbol{D}$ is much simpler to compute than the penalty matrix $\boldsymbol{\Omega}$ used in smoothing splines or O'Sullivan splines, which depend on the knots. The difference in computational cost and complexity increases for higher orders of the spline.

We will see that the penalty used in P-splines is however quite close to that of O'Sullivan splines. For comparison we also assume $k = 4$ in this paragraph. From (De Boor, 1978, p. 115)'s formula of spline derivatives, we have the identity

$$(t_j - t_{j-1})B'_{j,k}(x) = k\big(B_{j,k-1}(x) - B_{j+1,k-1}(x)\big).$$

Using this formula, the O'Sullivan spline penalty rewrites

$$\int_{x_1}^{x_n} \sum_j \left\{\alpha_j B''_{j,4}(x)\right\}^2 dx \propto \int_{x_1}^{x_n} \left\{\sum_j \Delta^2\alpha_j B_{j,2}(x)\right\}^2 dx \tag{4.11}$$

$$= \int_{x_1}^{x_n} \sum_j (\Delta^2\alpha_j)^2 B_{j,2}(x)^2 dx + \int_{x_1}^{x_n} \sum_j \Delta^2\alpha_j \Delta^2\alpha_{j-1} B_{j,2}(x)B_{j-1,2}(x)dx \tag{4.12}$$

$$= c_1 \sum_j (\Delta^2\alpha_j)^2 + c_2 \sum_j \Delta^2\alpha_j \Delta^2\alpha_{j-1}, \tag{4.13}$$

where the indices are defined implicitly by the order of the spline in each sum. The constants c_1 and c_2 are integrals of second order splines over $[x_j, x_{j+1})$; they are proportional to $2/3$ and $1/6$ respectively. These results follow from simple algebra and are not detailed further (see Eilers and Marx, 1996, Section 3).

From this result we see that where the P-Spline penalty for the j-th knot is proportional to $(\Delta^2\alpha_j)^2$, that of O'Sullivan is proportional to $2/3(\Delta^2\alpha_j)^2 + 1/6\Delta^2\alpha_j\Delta^2\alpha_{j-1}$. These two penalties are quite similar for many values of $\boldsymbol{\alpha}$ (and even more so as $\Delta^3\boldsymbol{\alpha}$ gets small). This gives an understanding of the close relation between the two methods. We don't investigate this topic further (see Wand and Ormerod, 2010).

P-splines are praised for their simplicity. In comparison, O'Sullivan splines are somewhat numerically less simple, although Wand and Ormerod (2010) provides explicit formula for the penalty matrix in (4.9). O'Sullivan splines have been generalized to any order, so both methods stand equal in terms of generality. P-splines can be deemed too restrictive for their choice of setting equal knots. But in fact the penalization that it enforces will compensate for knots that are set too close (simulation work by Ruppert, 2002, show that above a certain threshold, the number of knots has little impact on the performance of P-splines).

4.2 Spline regression with automatic knot selection

In this section, I include the preprint "Spline regression with automatic knot selection".

Spline Regression with Automatic Knot Selection

Vivien Goepp *
MAP5 (CNRS UMR 8145), Université Paris Descartes
Olivier Bouaziz
MAP5 (CNRS UMR 8145), Université Paris Descartes,
and
Grégory Nuel
LPSM (CNRS UMR 8001), Sorbonne Université

September 4, 2019

Abstract

In this paper we introduce a new method for automatically selecting knots in spline regression. The approach consists in setting a large number of initial knots and fitting the spline regression through a penalized likelihood procedure called adaptive ridge. The proposed method is similar to penalized spline regression methods (e.g. P-splines), with the noticeable difference that the output is a sparse spline regression with a small number of knots. We show that our method – called A-spline, for *adaptive splines* – compares favorably with other knot selection methods: it runs way faster and has close to equal predictive performance. A-spline is applied both to simulated and real datasets. A fast and publicly available implementation in `R` is provided along with this paper. The `R` code and datasets used for simulations and real data analysis are available in Supplementary Materials.

Keywords: Spline Regression, B-splines, Penalized Likelihood, Adaptive Ridge, Bandlinear Systems, Changepoint Detection.

*`vivien.goepp@parisdescartes.fr`

1 Introduction

After witnessing great developments in the past decades (see Wahba, 1990; Hastie et al., 2001; Ruppert et al., 2009; Wood, 2017) spline regression has become a tool of choice for semiparametric regression: it is restrictive enough to benefit from the simplicity of parametric estimation and general enough to accurately approximate a large range of smooth functions. For a while, the knots over which the splines are built was arbitrary chosen by the data analyst. Yet the number of knots has an important influence in the resulting fit: with not enough knots the regression is underfitted and with too many knots it is overfitted. Choosing the knots' position is also important since uniformly distributed knots can lead to overfitting in an area where there are few points and underfitting in an area where there are many points.

During the 1990's, penalization methods have been developed to overcome this difficulty. The idea is to set (too) many knots and to control overfitting by penalizing over the spline parameters. In smoothing splines (see Hastie et al., 2001, Section 5), knots are set at each data point and the spline's wiggliness is controlled by penalizing over the integrated squared second order derivative $\int \{f''(t)\}^2 dt$. The smoothing spline estimate has a closed-form expression and is computationally efficient. O'Sullivan (1986) generalized smoothing splines to an arbitrary choice of knots, allowing to set fewer knots than the sample size. Later, Eilers and Marx (1996); Marx and Eilers (1998) introduced P-splines, based on a penalty over the finite order differences of the parameters. This penalization is closely related to that of O'Sullivan (see Eilers and Marx, 1996, Section 3) and is a generalization of Whittaker (1922)'s graduation method, which can be seen as a P-spline of order 0 with knots placed at data points. See Eilers et al. (2015) and citations therein for a review of P-splines and see Wand and Ormerod (2010) and Eilers et al. (2015, Appendix A) for a comparison between O'Sullivan's splines and P-splines. Smoothing splines are implemented in the `R` packages `gam` (Hastie et al., 2001; Hastie, 2018) and `mgcv` (Wood, 2017) and P-splines are implemented in the `R` package `pspline`.

These regularized spline regression methods are simple and computationally fast. However, a spline with fewer knots is easier to interpret, which in many cases is a desired goal. Indeed a knot corresponds to a discontinuity in the spline's kth derivative (where k is the

spline's order) and thus to a shift in the trend of the estimated function. As pointed out by Wand (2000), setting a very large number of knots and exploring the set of splines defined on any subset of the knots is not tractable in practice. Friedman (1991) has developed a multivariate variable selection technique called MARS (Multivariate Adaptive Regression Splines). It uses a recursive partitioning of the domain and sequentially selects the knots with a forward/backwards step size procedure (see Friedman and Silverman (1989) and (Hastie et al., 2001, Section 9.4) for details). Luo and Wahba (1997) have later developed a closely related approach called Hybrid Adaptive Splines which uses a forward stepwise procedure and fits penalized splines instead of using a backward procedure. Other paths have been taken to solve this computationally intensive problem. Namely, Jamrozik et al. (2010) have offered to estimate the best location of knots using a differential evolution algorithm. However, their approach was limited to a number of knots varying between 4 and 7 and to splines of order 1.

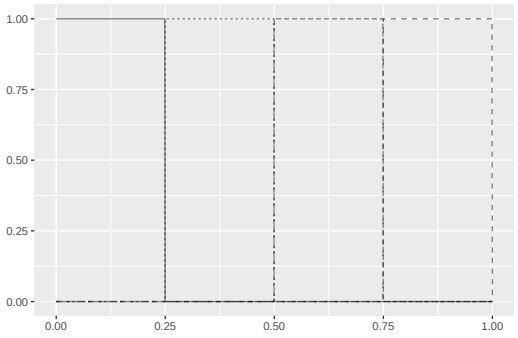
A different approach is to jointly estimate the spline coefficients and the position of the knots, as well as their number. The splines estimated by allowing the knots' position to vary are called free-knot splines. Jupp (1978) was among the first to introduce a method for free knot spline fitting. The squared residuals are regarded as a function of both the knot positions and the spline coefficients and the procedure uses a reparametrization of the knots' location and the Levenberg-Marquardt optimization algorithm (Marquardt, 1963) to avoid staying in local minima. Lindstrom (1999) improved on this method by adding a penalty over the dispersion of the knots. Secondly, Bayesian methods have also been widely used to select the knots' number and position, which are viewed as priors. In this context, Denison et al. (1998) developed a reversible jump Monte Carlo-based method (RJMCMC, see Green, 1995). For each knot set, unpenalized least square regression is used to fit the spline and the RJMCMC algorithm is used to roam the space of possible knot number and positions. The best knots are selected using a model averaging approach. Biller (2000) extended this approach to the case where the coefficients are also estimated in a Bayesian framework; DiMatteo et al. (2001) and Holmes and Mallick (2001) presented developments around the same lines, but limited to 3^{rd} order splines. Leitenstorfer and Tutz (2007) introduced a method of knot selection using families of radial functions (in practice Gaussian densities)

instead of splines. It uses a boosting algorithm to iteratively select the best fitting function among a predefined number of candidate functions, to estimate its parameter, and to add it to the current fit. Finally, penalized estimation has been used previously by Osborne et al. (1998) to perform knot selection. This work uses power series basis splines and selection is performed using the Lasso (Tibshirani, 1996). This penalty is only used for knot selection and an unpenalized fit is used to infer the coefficients selected by the Lasso. Stepwise selection has been used for a long time for knot selection (see Smith, 1982; Stone et al., 1997). A formulation of the procedure for one-dimensional splines is available in Wand (2000). Not many of these methods for knot selection have a currently available implementation in `R`. The method developed by DiMatteo et al. (2001) was implemented in `C` with an `R` wrapper by Wallstrom et al. (2008) and is publicly available[1]. The package `freeknotsplines` implements Spiriti et al. (2013)'s method for knot selection.

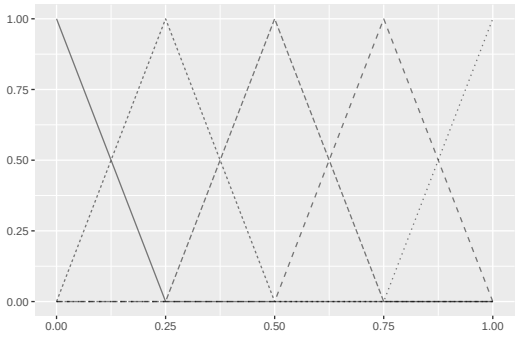
In this article we introduce a new computationally efficient method – called A-splines, for *adaptive splines* – to automatically select the number and position of the knots from the data. Contrarily to most free-knot spline methods, we set a high number of evenly distributed initial knots and use a penalized likelihood approach to gradually remove the least relevant knots. It is based on a regularization method with an approximate L_0 norm penalty. Therefore, it benefits from the simplicity and computational speed of penalized-based methods and is orders of magnitude faster to run than comparable knot-selection methods.

Section 2 gives a short summary of B-splines and B-spline regression. Section 3 introduces our spline regression method. In Section 4, our method is extended to the generalized linear model. Section 5 deals with the choice of the bias-variance tradeoff parameter. Section 6 compares the prediction performance of our model to comparable methods through a simulation study. Section 7 gives some details about the fast implementation of the fitting algorithm. Finally, A-spline is illustrated on several real datasets in Section 8.

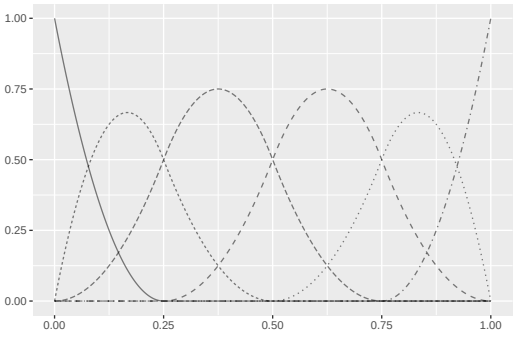
[1] `https://www.jstatsoft.org/article/view/v026i01`



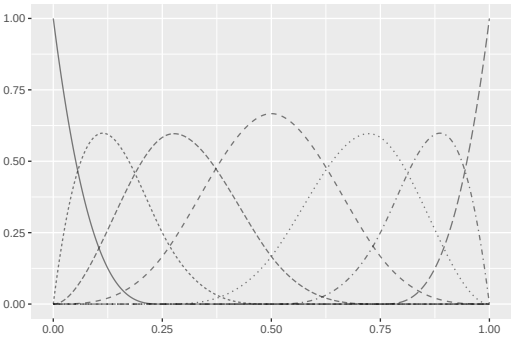
(a) Order 0 B-splines



(b) Order 1 B-splines



(c) Order 2 B-splines



(d) Order 3 B-splines

Figure 1: Bases of B-spline of degree 0 to 3 (Panels a to d) with 3 knots: $(0.25, 0.5, 0.75)$. Note that with 3 knots, there are 4 splines in the basis of degree 0 and 7 splines in the basis of degree 3.

2 B-spline Regression

2.1 B-spline Basis

In this section we recall the definition and some basic properties of splines and B-splines. Throughout this work, let $t_1, \ldots, t_k$ be the ordered knots included in a real interval $[a, b]$. A spline of order $p \geq 1$ is a piecewise polynomial function of degree $p-1$ such that its derivatives up to order $p-2$ are continuous at every knot $t_1, \ldots, t_k$. In the following, we will refer to the degree $q \geq 0$ of a spline and not to its order $q+1$. The set of splines of degree $q \geq 0$ over the knots $\boldsymbol{t} = (t_1, \ldots, t_k)$ is a vector space of dimension $q + k + 1$.

A possible choice of spline basis is the truncated power basis: $\{x^0, \ldots, x^q, (x - t_1)_+^q, \ldots,$

$(x - t_k)_+^q\}$, where $(u)_+ = \max(u, 0)$. The first $q+1$ functions of the basis are polynomials and the other k functions are truncated polynomials of degree q. Decomposing a spline into the truncated power basis brings out powers of large numbers, which lead to rounding errors and numerical inaccuracies (De Boor, 1978, p. 85).

Schoenberg introduced (Schoenberg, 1946; Curry and Schoenberg, 1966) a spline basis called B-splines – for *Basic*-splines. This spline basis provides more stable computations of spline regression (see de Boor, 1972). A B-spline is a spline which is non-zero over $[x_k, x_{k+q+1}]$ for some k. For $i = 1, \ldots, q+k+1$, the i-th B-spline of degree q is noted $B_{i,q}(x)$ and is defined by

$$B_{i,q}(x) = \frac{x - t_i}{t_{i+q} - t_i} B_{i,q-1}(x) + \frac{t_{i+q+1} - x}{t_{i+q+1} - t_{i+1}} B_{i+1,q-1}(x) \quad \text{if} \quad q > 0$$

and $B_{i,0}(x) = \mathbb{1}_{t_i \leq x < t_{i+1}}$. Important properties of a B-spline are: (i) the B-spline is non-zero over an interval spanning $q+2$ knots; (ii) at a point, only $q+1$ B-splines are non-zero; (iii) $B_{i,q}(x) \in [0, 1]$. An illustration of B-spline bases of degree 0 to 3 is given in Figure 1. In practice, B-splines can be computed using the function `bSpline` from the `R` package `splines2` (Wang and Yan, 2017).

2.2 B-Spline Regression

Let $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$ be the univariate data and consider the non-parametric regression setting

$$y_i = f(x_i) + \varepsilon_i, \quad 1 \leq i \leq n, \tag{1}$$

where f is a "smooth" function and where ε_i are i.i.d. Gaussian errors. The function f is estimated by a spline over an interval $[a, b]$ containing all x_is. Fitting the data consists in minimizing the sum of squares

$$\mathrm{SS}(\boldsymbol{a}, \boldsymbol{t}) = \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{q+k+1} a_j B_{j,q}(x_i) \right\}^2, \tag{2}$$

where $\boldsymbol{a} = (a_1, \ldots, a_{q+k+1})$ is the B-spline coefficients, and $\boldsymbol{t}$ are kept as parameters to highlight that the fitting procedure depends on the choice of the knots. This is the framework of ordinary least squares regression with design matrix $\boldsymbol{B} = [B_{j,q}(x_i)]_{i,j}$ and parameter $\boldsymbol{a}$:

$$\mathrm{SS}(\boldsymbol{a}, \boldsymbol{t}) = \|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}\|_2^2. \tag{3}$$

3 Automatic Selection of Knots

3.1 Model selection

When there are many knots, spline regression is prone to overfitting. In the extreme case, when there as as many parameters as data points, the fitted spline interpolates the data. In this paper, we propose to estimate the spline which makes the best tradeoff between model dimension (i.e. number of knots) and goodness of fit. To this effect, we choose a high number of equally spaced initial knots and penalize over the number of knots. When a B-spline is defined over the knots $t_1, \ldots, t_k$ and is such that $\Delta^{q+1} a_{j^*} = 0$ for some j^*, it can be reparametrized as a B-spline over the knots $t_1, \ldots, t_{j^*-1}, t_{j^*+1}, \ldots t_k$. Consequently, we penalize over the number of non-zero $(q+1)$-order differences

$$\frac{\lambda}{2} \sum_{j=q+2}^{k} \|\Delta^{q+1} a_j\|_0, \tag{4}$$

where $\|.\|_0$ is the L_0 norm, i.e. $\|x\|_0 = 0$ if $x = 0$ and $\|x\|_0 = 1$ otherwise, and where the parameter $\lambda > 0$ tunes the tradeoff between goodness of fit and regularity. This penalty allows to remove a knot t_{j^*} that is not relevant for the regression, to merge the adjacent intervals $[t_{j^*-1}, t_{j^*})$ and $[t_{j^*}, t_{j^*+1})$, and to continue the fitting procedure with a spline defined over the remaining knots. When $\lambda \to 0$, the fitted function is a B-spline with all knots $t_1, \ldots, t_k$ and when $\lambda \to \infty$, the fitted function is a polynomial of degree q.

However, the penalty in Equation (4) is non differentiable and the estimation is therefore computationally non-tractable. An approximation method for the L_0 norm is introduced in the next section to overcome this difficulty.

3.2 Adaptive ridge

Following the work from Rippe et al. (2012) and Frommlet and Nuel (2016), we approximate the L_0 norm by using an iterative procedure called Adaptive Ridge. The new objective function is the weighted penalized sum of squares

$$\mathrm{WPSS}\left(\boldsymbol{a}, \lambda\right) = \|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}\|_2^2 + \frac{\lambda}{2} \sum_{j=q+2}^{q+k+1} w_j \left(\Delta^{q+1} a_j\right)^2, \tag{5}$$

where $\Delta a_j = a_j - a_{j-1}$ is the first order difference operator, $\Delta^i a_j = \Delta^{i-1}\Delta a_j$, and w_j are positive weights. The penalty is close to the L_0 norm penalty when the weights are iteratively computed from the previous values of the parameter $\boldsymbol{a}$ following the formula:

$$w_j = \left(\left(\Delta^{q+1} a_j\right)^2 + \varepsilon^2\right)^{-1},$$

where $\varepsilon > 0$ is a small constant. Indeed the function $x \mapsto x^2/\left(x^2 + \varepsilon^2\right)$ approximates the function $x \mapsto \|x\|_0$ if ε is sufficiently small. In practice, one typically sets $\varepsilon = 10^{-5}$ (Frommlet and Nuel, 2016). At convergence, $\left(\Delta^{q+1} a_j\right)^2 w_j \simeq \|\Delta^{q+1} a_j\|_0$ gives a measure of how relevant the j-th knot is. One chooses a threshold of 10^{-2} and selects the knots with a weighted differences higher than 0.99, which we note t_j^{sel}. The number of selected knots will be noted k_λ, such that the number of parameters of the selected spline is $q + k_\lambda + 1$. Since the selected knots are present in breakpoints of the curve, one then fits unpenalized B-splines over the knots $\boldsymbol{t}^{\text{sel}}$, as explained in Section 2.2. Consequently, this method provides a regression model that is both regularizing and simple, in the sense that the model dimension is small.

We note that Frommlet and Nuel (2016) give a more general formula for the weights that allows to approximate any L_p norm, for $p > 0$. In particular, the L_1 norm could be chosen, which induces both shrinkage and selection of the coefficient. This penalty was already used in this context by Tibshirani (1996), which used in the Lasso for knot selection.

WPSS (a, λ) of Equation (5) easily rewrites

$$\|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}\|_2^2 + \lambda \boldsymbol{D}^T \boldsymbol{W} \boldsymbol{D} \boldsymbol{a}, \tag{6}$$

where $\boldsymbol{W} = \mathrm{diag}\,(\boldsymbol{w})$ and $\boldsymbol{D}$ is the matrix representation of the difference operator Δ^{q+1}. The minimization of WPSS is explicit:

$$\hat{\boldsymbol{a}} = \left(\boldsymbol{B}^T \boldsymbol{B} + \lambda \boldsymbol{D}^T \boldsymbol{W} \boldsymbol{D}\right)^{-1} \boldsymbol{B}^T \boldsymbol{y}. \tag{7}$$

The adaptive ridge procedure is detailed in Algorithm 1.

The penalty term is conveniently written with the circulating matrix $\boldsymbol{D}$. However, for computational efficiency, $\boldsymbol{D}$ is never computed and instead we implement a fast computation algorithm for the penalty term. More details about the implementation are given in Section 7.

Algorithm 1 Adaptive Ridge Procedure for Spline Regression

Input: $\boldsymbol{x}, \boldsymbol{y}, \lambda$

Output: $\hat{\boldsymbol{a}}$

function ADAPTIVE-SPLINE $(\boldsymbol{x}, \boldsymbol{y}, \lambda)$
  $\boldsymbol{a} \leftarrow \mathbf{0}$; $\boldsymbol{w} \leftarrow \mathbf{1}$
  while not converge **do**
    $\boldsymbol{a}^{\text{new}} \leftarrow \arg\min_{\boldsymbol{a}} \text{WPSS}(\boldsymbol{a}, \lambda)$
    $w_j \leftarrow \left(\left(\Delta^{q+1} a_j^{\text{new}}\right)^2 + \varepsilon^2\right)^{-1}$
    $\boldsymbol{a} \leftarrow \boldsymbol{a}^{\text{new}}$
  end while
  Compute $\boldsymbol{t}^{\text{sel}}$ using $\left(\Delta^{q+1}\boldsymbol{a}\right)^2 \boldsymbol{w}$
  $\hat{\boldsymbol{a}} \leftarrow \arg\min_{\boldsymbol{a}} \text{SS}\left(\boldsymbol{a}, \boldsymbol{t}^{\text{sel}}\right)$
  return $\hat{\boldsymbol{a}}$
end function

Relation to P-Splines fitting It is interesting to note that A-splines are closely related to P-splines (Eilers and Marx, 1996), whose objective function writes

$$\text{PSS}(\boldsymbol{a}, \lambda) = \text{SS}(\boldsymbol{a}) + \frac{\lambda}{2} \sum_{j=p+1}^{k+q+1} \left(\Delta^p a_j\right)^2, \tag{8}$$

where the difference order p is a parameter to be chosen. Thus, the implementation of A-splines can be seen as a weighted P-splines fitting. The philosophies of A-splines and P-splines are however very different. P-splines avoid choosing the best knots by penalizing over the differences of the coefficients. Instead, we directly choose the best knots for spline regression.

4 Generalized Linear Model

Spline regression has also been used to fit values in the general linear model setting, like in Eilers and Marx (1996); Hastie et al. (2001). In this section, we extend A-spline regression to the generalized linear model. The goal is to estimate $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{y}|\boldsymbol{x}] = g^{-1}(\boldsymbol{B}\boldsymbol{a})$, where g is

the canonical link function and the variance of $\boldsymbol{y}$ is a function V of $\boldsymbol{\mu}$: $\mathrm{Var}[y] = V(\boldsymbol{\mu})$. Like the linear model, $\boldsymbol{\mu}$ can be estimated using spline regression. The generalized linear model is fitted using the Iteratively Reweighted Least Squares (IRLS) algorithm (McCullagh and Nelder, 1989, Section 2.5). With weighted penalization, the IRLS iteration writes:

$$\hat{\boldsymbol{a}}^{(k+1)} = \left(\boldsymbol{B}^T\boldsymbol{\Omega}^{(k)}\boldsymbol{B} + \lambda\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D}\right)^{-1}\boldsymbol{B}^T\left(\boldsymbol{\Omega}^{(k)}\boldsymbol{B}\hat{\boldsymbol{a}}^{(k)} + \boldsymbol{y} - \boldsymbol{\mu}^{(k)}\right) \tag{9}$$

where k is the step index and $\boldsymbol{\Omega}^{(k)}$ is the diagonal matrix with entries

$$\omega_{i,i}^{(k)} = \frac{1}{V\left(\mu_i^{(k)}\right)g'\left(\mu_i^{(k)}\right)^2},$$

with $\mu_i^{(k)} = g^{-1}\left(\boldsymbol{B}_i\hat{\boldsymbol{a}}^{(k)}\right)$. In practice, the estimation procedure in Algorithm 1 remains the same, except that WPSS is minimized by the Newton-Raphson procedure given in Equation (9).

5 Choice of the Penalty Constant

In this section, one selects the penalty that performs the best trade-off between goodness of fit and regularity. A first criterion is the AIC, which was used by Eilers and Marx (1996) in a similar context:

$$\mathrm{AIC}(\lambda) = \mathrm{SS}\left(\hat{\boldsymbol{a}}_\lambda\right) + 2\left(q + k_\lambda + 1\right). \tag{10}$$

A different criterion is the Bayesian Information Criterion (BIC) (see Schwarz, 1978):

$$\mathrm{BIC}\left(\lambda\right) = \mathrm{SS}\left(\hat{\boldsymbol{a}}_\lambda\right) + \left(q + k_\lambda + 1\right)\log n. \tag{11}$$

Bayesian criteria maximize the posterior probability $\mathrm{P}(\mathcal{M}_\lambda|\mathrm{data}) \propto \mathrm{P}(\mathrm{data}|\mathcal{M}_\lambda)\pi(\mathcal{M}_\lambda)$, where $\mathrm{P}(\mathrm{data}|\mathcal{M}_\lambda)$ is the integrated likelihood and $\pi\left(\mathcal{M}_\lambda\right)$ is the prior distribution on the model $\mathcal{M}_\lambda$. This problem is equivalent to minimizing $-2\log\mathrm{P}(\mathcal{M}_\lambda|\mathrm{data})$. By integration

$$\mathrm{P}(\mathcal{M}_\lambda|\mathrm{data}) = \int_{\boldsymbol{a}} \mathrm{P}(\mathrm{data}|\mathcal{M}_\lambda, \boldsymbol{a})\pi(\boldsymbol{a})d\boldsymbol{a},$$

where $\mathrm{P}\left(\mathrm{data}|\mathcal{M}_\lambda, \boldsymbol{a}\right)$ is the likelihood and $\pi(\boldsymbol{a})$ is the prior distribution of the parameter, which is taken constant in the following. Thus Bayesian criteria are defined as

$$-2\log\mathrm{P}\left(\mathcal{M}_\lambda|\mathrm{data}\right) = \mathrm{SS}(\hat{\boldsymbol{a}}_\lambda) + \left(q + k_\lambda + 1\right)\log n - 2\log\pi(\mathcal{M}_\lambda) + \mathcal{O}_\mathrm{P}(1).$$

The BIC is the Bayesian criterion obtained when one chooses a uniform prior on the model: $\pi(\mathcal{M}_\lambda) = 1$. As explained by Żak-Szatkowska and Bogdan (2011), a uniform prior on the model is equivalent to a binomial prior on the model dimension. Therefore, the BIC tends to give too much importance to models of dimensions around $\frac{q+k+1}{2}$. Since the adaptive knot selection is performed with a large number of initial knots, this will result in underpenalized estimators. To this effect, Chen and Chen (2008) have developed an extended Bayesian information criterion called EBIC_0. The EBIC_0 criterion is defined by choosing:

$$\pi(M_\lambda) = \binom{q+k+1}{q+k_\lambda+1}^{-1}$$

and

$$\text{EBIC}_0(\lambda) = \text{SS}(\hat{\boldsymbol{a}}_\lambda) + (q+k_\lambda+1)\log n + 2\log\binom{q+k+1}{q+k_\lambda+1}. \tag{12}$$

The EBIC_0 assigns the same *a priori* probability to all models of same dimension. Therefore the EBIC_0 will tend to choose sparse models even with a high number of initial knots. These criteria's selection performances are compared in the next section through a simulation study.

6 Simulation Study

6.1 Comparing the Selection Criteria for A-spline

A simulation study has been conducted to compare the performances of the three criteria. Data are simulated as follows. The x_i are taken uniformly over $[0,1]$ and y_i are simulated using Equation (1), where f is a known function and $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. We use four different functions: the *Bump* function

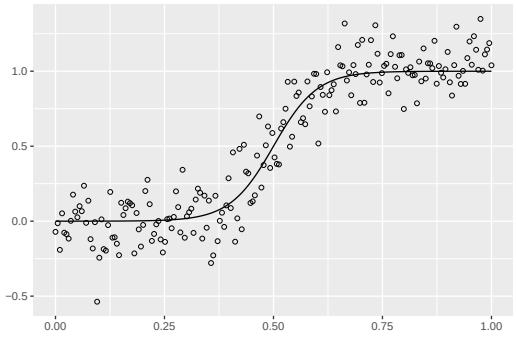
$$f_1(x) = 0.4\left(x + 2\exp\left[-\left\{16(x-0.5)\right\}^2\right]\right),$$

the *Logit* function

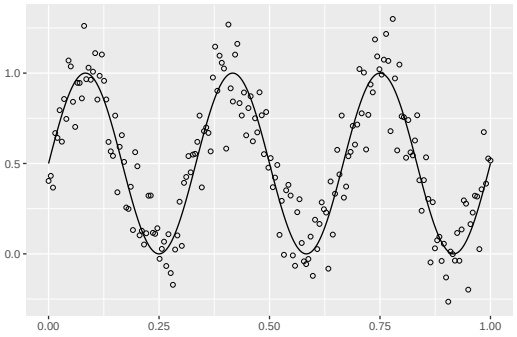
$$f_2(x) = \frac{1}{1+\exp\left\{-20(x-0.5)\right\}},$$

the *Sine* function

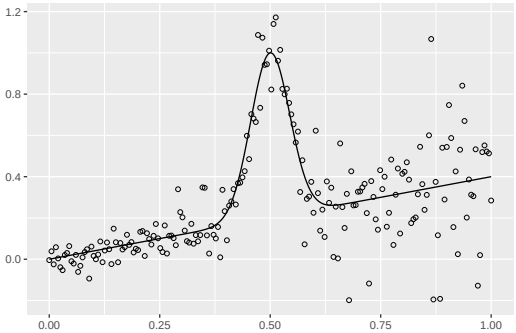
$$f_3(x) = 0.5\sin(6\pi x) + 0.5,$$



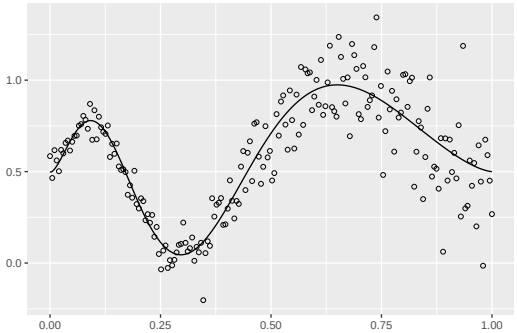
(a) Logit Function



(b) Sine Function



(c) Bump Function



(d) SpaHet Function

Figure 2: Simulated data using functions Logit (a), Sine (b), Bump (c) and SpaHet (d), in solid line. Each dataset has size 200. The errors are chosen homoscedastic ($\sigma = 0.15$) for (a) and (b) and heteroscedastic ($\sigma_i = \left(0.3x_i + 0.2\sqrt{x_i}\right)^2$) for (c) and (d).

Sample size	AIC	BIC	EBIC
50	0.02220	**0.02**	0.02418
100	0.00754	0.00324	**0.00248**
200	0.00285	0.00136	**0.00127**
400	0.00131	**0.00071**	0.00072

(a) Logit Function

Sample size	AIC	BIC	EBIC
50	0.02239	**0.02001**	0.02459
100	0.00755	0.00486	**0.00458**
200	0.00316	**0.00231**	0.00247
400	0.00156	**0.00132**	0.00141

(b) Sine Function

Sample size	AIC	BIC	EBIC
50	0.02000	**0.01801**	0.02211
100	0.00735	0.00627	**0.00479**
200	0.00354	0.00234	**0.00217**
400	0.00177	0.00106	**0.001**

(c) Bump Function

Sample size	AIC	BIC	EBIC
50	0.02082	**0.01784**	0.02138
100	0.00727	0.00509	**0.00371**
200	0.00333	0.00194	**0.00161**
400	0.00170	0.00081	$\mathbf{8e-04}$

(d) SpaHet Function

Table 1: Mean squared errors of adaptive spline regression for different selection criteria and for different sample sizes. Different datasets are simulated using four different functions: the Bump function (a), the Logit Function (b), the Sine function (c) and the SpaHet function (d). The smallest value of each row is highlighted in bold.

and the *SpaHet* – for *spatially heterogeneous* – function

$$f_4(x) = \sqrt{x(1-x)}\sin\left(\frac{2\pi\left(1+2^{-3/5}\right)}{x+2^{-3/5}}\right)+0.5.$$

These functions were used by Wand (2000) and Ruppert (2002) in similar contexts for benchmarking the efficiency of spline regression. The functions f_1 to f_4 have been rescaled in order to vary in $[0,1]$, so that all simulation cases have similar signal-to-noise ratios. We choose homoscedastic errors $\sigma_i = 0.15$ for the functions *Logit* and *Sine* and heteroscedastic errors for the *Bump* and *SpaHet* functions: $\sigma_i = \left(0.3x_i + 0.2\sqrt{x_i}\right)^2$, so that the variance increases from 0 when $x = 0$ to 0.25 when $x = 1$. Data are simulated with sample sizes 50, 100, 200, and 400. Illustration of the functions and of the simulated data are given in Figure 2. For each example 500 datasets were simulated. A-splines are fitted and we compare the Mean Squared Error (MSE) of the estimated function for the three criteria:

$$\|f-\widehat{f}\|_2^2 = \int_0^1 \left(f(x)-\widehat{f}(x)\right)^2 dx.$$

The median MSEs are displayed in Table 1 for each value of the sample size. For all functions and for all criteria, the MSE decreases with the sample size, as is expected. The comparison between the criteria brings the same conclusions for all four functions: the BIC and EBIC_0 always perform better than the AIC. Moreover, note that the EBIC_0 always outperforms the BIC for the sample size 100, and performs almost as well for the sample size 200. In conclusion, the BIC and EBIC_0 are to be preferred over the AIC and overall, the EBIC_0 seems a better choice than the BIC. In the remaining of this work, A-splines will be used with the EBIC_0 criterion.

6.2 Comparing A-spline with knot selection methods

In this section, the performance of A-spline is compared to that of other knot selection methods. As mentioned in the introduction, only two knot selection methods are currently available in `R`. DiMatteo et al. (2001)'s method, implemented by Wallstrom et al. (2008), is essentially an RJMCMC over the space of knot number and position. Spiriti et al. (2013)'s method, implemented in the package `freeknotsplines` initiates the knots uniformly and uses an optimization algorithm to search for the best location of each knot between its two neighboring knots. The method performs four passes through the list of ordered knots: two passes in increasing order and two in decreasing order. Two choices are provided for the optimization algorithm: a blind search or a genetic algorithm (see Haupt and Haupt, 2004). In the genetic algorithm, the crossover step consists in keeping the leftmost knots of one parent and the rightmost knots of the other parent. The mutation step consists in choosing one knot at random and sampling it uniformly between its neighboring knots. The selection function is set to either the residual sum of squares or the general cross validation (GCV). With both algorithms, the number of knots is chosen using an adjusted GCV criterion. Of the two optimization algorithms, we will use the genetic algorithm in our simulation setting.

We use the same simulation setting as in the previous section. Figure 3 represents the MSE distribution over 500 replications for every sample size and for every function. We see that the three methods display close to equal performances. Across all sample sizes and all functions, the BARS seems to perform better than the genetic optimization

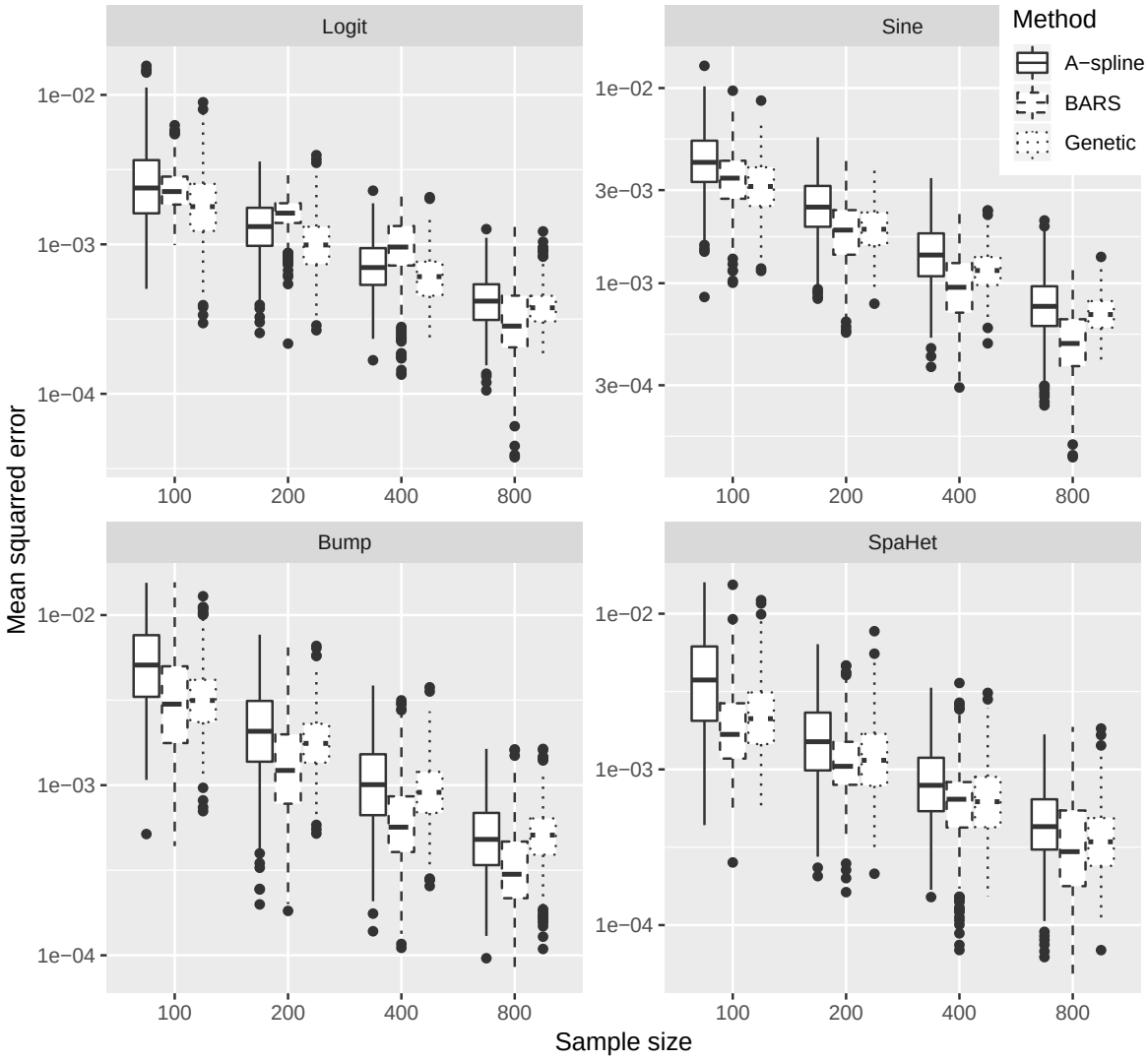


Figure 3: Mean squared errors of A-spline (solid line), BARS (dashed line), and Spiriti's genetic optimization (dotted line) estimates for different sample sizes: 50, 100, 200, and 400. The simulations are performed with the *Bump*, *Logit*, *Sine*, and *SpaHet* functions and repeated 500 times.

method, which in turn performs slightly better than the A-spline. However these differences are very small compared to the variation of the MSE inside each method, which remains important even for high sample sizes. Moreover, with sample size 800, the A-spline performs almost to as well as the genetic optimization method. Furthermore, we compare A-spline's predictive performance with P-splines through a similar simulation study, as shown in the Supplementary Materials. It results that P-splines have lower MSE on average, but the difference between the two become negligible for large data sets ($n \sim 400$). It is worth noticing that the comparison also illustrates how A-splines provide models that are easier to interpret.

The three methods compared here have very different computational costs. Simulations were carried on an Intel Core i3 CPU running at 2.3 GHz. For a sample size of 400, the median running time over 500 repetitions is 0.28 seconds for A-spline, while it is 2.66 seconds and 123.02 seconds for the BARS and genetic optimization methods respectively. This illustrates the computational efficiency of A-spline compared to state-of-the-art knot-selection methods.

In conclusion, the BARS and Spiriti's genetic optimization knot selection methods perform slightly better than A-spline, but this comes at the cost of inconveniently intensive computation.

7 Practical Implementation

In this section, the implementation of A-splines is explained in details. Particular attention has been brought to the computation of matrix products. Consequently, fitting A-splines is almost instantaneous: on a standard laptop, it takes 1.3 seconds with 200 initial knots and 5000 data points. In the next three sections, several bottlenecks in the computation of A-splines are addressed. Matrix products computations are accelerated using an `Rcpp` (Eddelbuettel, 2013) implementation. An `R` implementation of the A-spline estimation procedure is publicly available in the package `aspline`[2].

Let us note that the design matrix only appears in the regression model through $\boldsymbol{B}^T\boldsymbol{B}$ and $\boldsymbol{B}^T\boldsymbol{y}$, so apart from the computation of $\boldsymbol{B}$, $\boldsymbol{B}^T\boldsymbol{B}$, and $\boldsymbol{B}^T\boldsymbol{y}$, which is done only once,

[2] `github.com/goepp/aspline`

the algorithm does not depend on the sample size.

7.1 Adaptive Spline Regression with Several Penalties

The penalty constant λ tunes the tradeoff between goodness of fit and regularity. To choose the optimal λ, regression is performed for a sequence of penalties $\boldsymbol{\lambda} = (\lambda_\ell), 1 \leq \ell \leq L$ and a criterion is used to determine which regression model to select. Computing the procedure for a series of values of λ significantly increases the computing time. Note that a small variation of λ yields a small variation of $\hat{\boldsymbol{a}}_\lambda = \arg\min_a \mathrm{WPSS}(\boldsymbol{a}, \lambda)$. Consequently, $\hat{\boldsymbol{a}}_{\lambda_\ell}$ is a good initial point for the minimization of $\mathrm{WPSS}(\boldsymbol{a}, \lambda_{\ell+1})$. Making use of this *hot start* significantly speeds up the minimization of $\mathrm{WPSS}(\boldsymbol{a}, \lambda_{\ell+1})$ and thus decreases the computation time of the adaptive ridge procedure. This implementation of the adaptive ridge is introduced in Rippe et al. (2012) and Frommlet and Nuel (2016) and a similar idea is used in the implementation of the LASSO in the package `glmnet` (Friedman et al., 2010).

7.2 Fast Computation of the Weighted Penalty

The matrix inversion in Equations (7) and (9) is the computational bottleneck of the adaptive ridge procedure. The matrix $\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D}$ is symmetric and q-banded, and as noticed by Wand and Ormerod (2010), so is $\boldsymbol{B}^T\boldsymbol{B}$. Consequently, the inversion is done using Cholesky decomposition and back-substitution, as implemented in the package `bandsolve`[3]. This reduces the temporal complexity from $\mathcal{O}\left((k+q+1)^3\right)$ to $\mathcal{O}\left((k+q+1)(q+2)\right)$. For example, if $k = 50$ and $q = 3$, the computation time will be reduced by a factor 500. It is important to note that the matrices $\boldsymbol{W}$ and $\boldsymbol{D}$ are not stored in memory: only the vector $\boldsymbol{w}$ and the first row of $\boldsymbol{D}$ are used. This leads to improvements in spatial complexity, the details of which are not given here.

7.3 Fast Computation of the Weighted Design Matrix

In the setting of generalized linear regression, the matrix product $\boldsymbol{B}^T\boldsymbol{\Omega}\boldsymbol{B}$ in Equation (9) is computed at each iteration of the Newton-Raphson procedure. Since the design matrix has

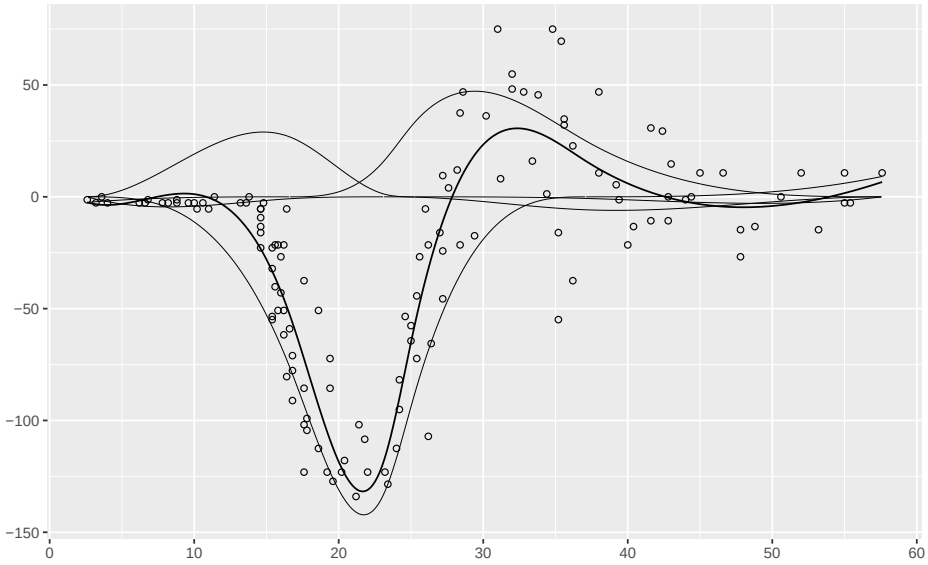
[3] `github.com/monneret/bandsolve`

n rows, this operation makes the generalized linear regression computationally expensive for large datasets. Fortunately $\boldsymbol{B}$ is sparse: it has $q+1$ non-zero elements in each row. Due to this structure, the product $\boldsymbol{B}^T\boldsymbol{\Omega}\boldsymbol{B}$ only has $(q+k+1)(q+1)$ non-zero entries. Each entry takes $\mathcal{O}\left(\frac{n}{k}\right)$ operations to compute on average. Thus the matrix product can be computed with a $\mathcal{O}\left((q+k+1)(q+1)n/k\right)$ temporal complexity, compared to the $\mathcal{O}\left((q+k+1)^2 n\right)$ complexity of the naive implementation. For instance, even with $q=3$ and $k=50$, this implementation is faster by a factor ~ 700.

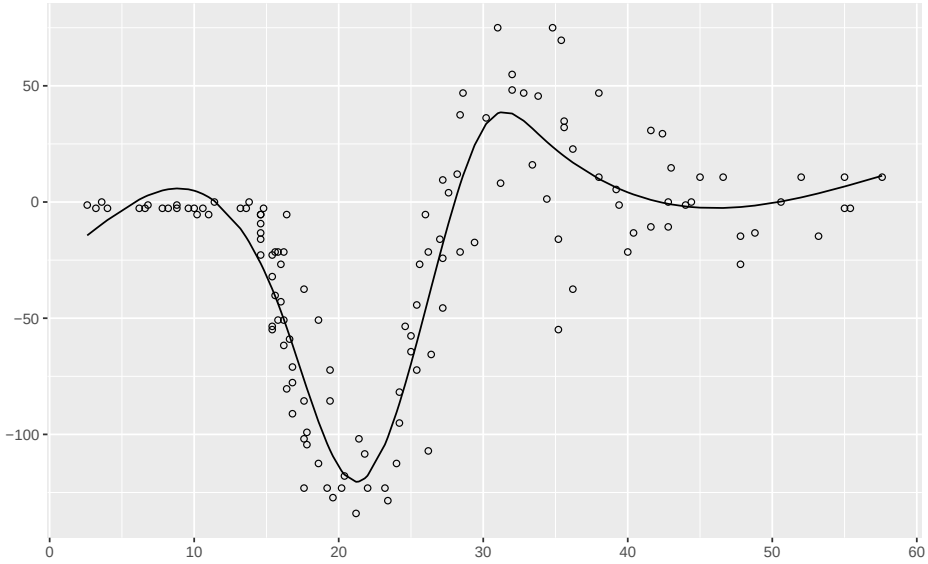
8 Real Data Applications

Our method is illustrated with several real data applications. All the data sets used in this section and the code used to produce the figures are available in Supplementary Materials.

We first present a dataset of simulated motorcycle accidents used to crash-test helmets. The data consists of 132 observations of helmet acceleration (in units of g) measured along time after impact (in *milliseconds*). These data have being used as illustration of spline regression by Silverman (1985) and Eilers and Marx (1996) and are available in Härdle (1990). A history of the data set and its use in spline regression is given in de Souza and Heckman (2013). This dataset represents a good test for non parametric regression since the variance of the errors varies a great deal and there are several breakdown moments in the data. A-spline regression and BARS regression with order $q=3$ are performed (Figures 4a and Figure 4b respectively). In Figure 4a, the solid lines represent the estimated fit and the dashed lines represent the decomposition of the fit onto the B-spline family. The two estimations are almost equal; the A-spline fit seems visually "smoother" than BARS, especially around the points of high variation. Many non-parametric regression methods have been criticised for fitting a function which increases slightly between 15 and 20 ms, as is the case for BARS, and (to a lesser extent) for A-spline. The rational is that since the helmet is subject to no exterior force before the impact ($\sim$ 15 ms), the *real* signal is equal to zero before that time, and the fitted function should not display variations in that region. However our work aims to find a function that can simply represent a set of points, regardless of whether they respect some constraint about the real function. We could imagine extensions of spline regression that would incorporate such *a priori* knowledge on



(a) A-splines



(b) BARS

Figure 4: Motorcycle helmet crash data: helmet acceleration (unit of g) as a function of time (in ms). A-spline (a) regression and BARS (b) regression are fitted. In Figure (a), the bold lines represent the estimates and grey lines represent the decomposition of the estimates onto the B-spline bases.

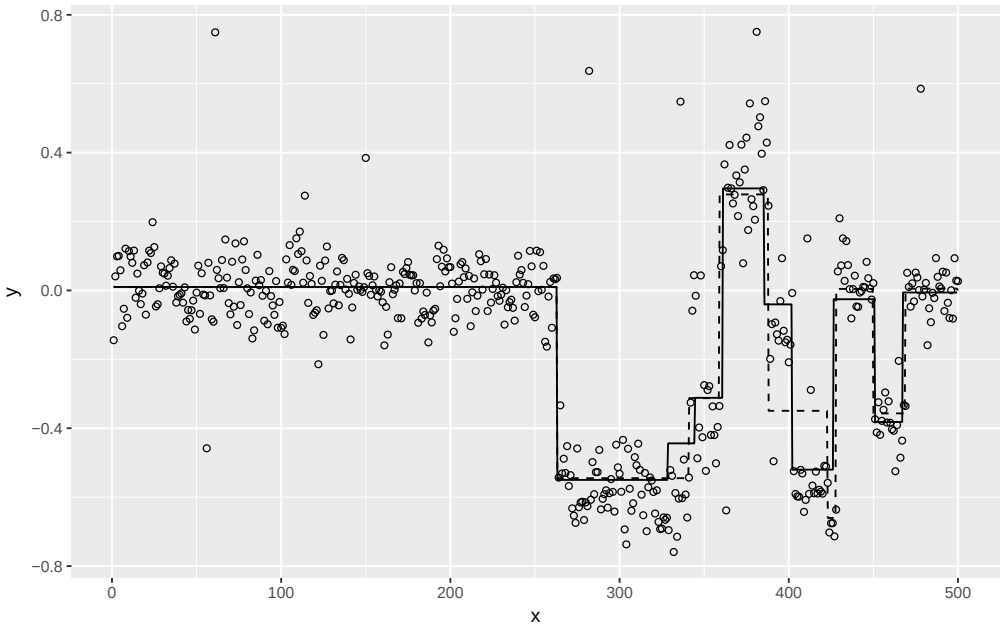


Figure 5: aCGH data of bladder cancer: probes 1 through 500. A-splines of order 0 are fitted (solid line) as well as the mean values fitted using the PELT changepoint detection method (dashed line).

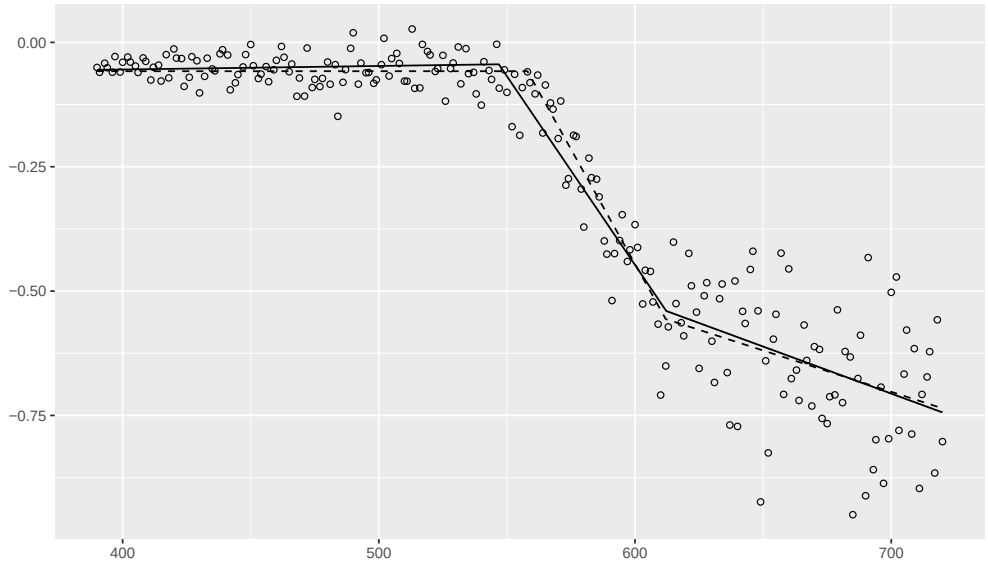


Figure 6: LIDAR data: log-ratio of light intensity as a function of the travelled distance. A-splines of order 1 (solid line) and Multivariate Adaptive Regression Splines (dashed lines) are fitted.

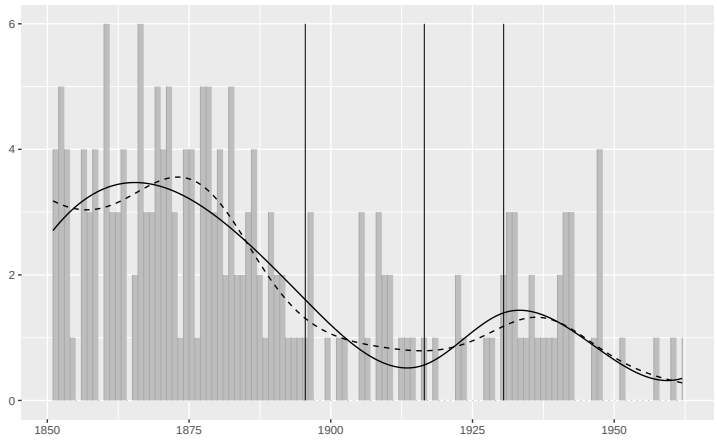


Figure 7: Yearly number of coal accidents in Britain (grey bars) with P-splines regression (dashed curve) A-spline regression (solid curve). The three knots selected by A-splines are represented by vertical lines.

the signal; this is beyond the scope of this article. A-spline regression has selected only 5 knots as relevant, and thus the fitted function is a linear combination of $5 + 3 + 1 = 9$ splines.

The second illustrative example uses a dataset of array Comparative Genomic Hybridization (aCGH) profiles for 57 bladder tumor samples (see Stransky et al., 2006, for references and access to the data). This dataset was used by Bleakley and Vert (2011) in the similar context of changepoint detection. The data represent the log-ratio of DNA quantities along 2215 probes. For the illustration, the 500 first observations of individual 1's aCGH profile are used. We fit a spline of order 0, i.e. a piecewise constant function. Indeed, A-splines of order 0 perform a regression with changepoint detection of the data, which is a desired goal for these data. The fitted spline is represented in solid line in Figure 5. The estimated function performs a satisfying estimation of the changepoints and of the mean values over each interval. Our regression method estimated 9 changepoints, each corresponding to a shift in the mean value of the signal. Our method is compared to a popular changepoint detection algorithm (dashed line of Figure 5) called PELT (Killick et al., 2012). We used the `R` package `changepoint.np` (Haynes et al., 2016) . This method detects 8 changepoints, all of which correspond to a changepoint detected by the A-spline regression.

The third example is based on the LIDAR data (Sigrist et al., 1994; Holst et al., 1996), which is used by Ruppert et al. (2003) to illustrate regression methods. The data come from a light detection and ranging (LIDAR) experiment. It consists of 221 observations of log-ratio of measured light intensities between two sources, as a function of the distance travelled by the light before being reflected (in *meters*). The data are available in the `R` package `SemiPar` and are represented in Figure 6.

The y-variable looks to be linear on three intervals: it is almost constant before $x = 550$ m, decrease steeply between $x = 550$ m and $x = 600$ m, after which the slope increases. To highlight these shifts in slope, splines of order 1 (i.e. piecewise linear functions) are chosen to fit the data. The A-spline fit displays two slope changes, at $x = 567$ m and $x = 607$ m. These moments visually correspond to the two biggest shifts in slope. We also fit Friedman (1991)'s MARS procedure (in dashed line, Figure 6) and compare it to A-splines. We use an implementation of the procedure in the `R` package `earth`. This method also selects two breakpoints of the slope, at $x = 558$ and $x = 612$, which are very close to the breakpoints detected by A-splines.

The last example uses the data of the registered number of disasters in British coal mines per year between the years 1850 and 1962 (Diggle and Marron, 1988). The number of coal disasters in each year is assumed to be Poisson distributed and the mean of the distribution is fitted using a Poisson regression. The data are fitted using A-spline regression of order 3. The fitted curve $\hat{\boldsymbol{\mu}} = g^{-1}(\boldsymbol{B}\hat{\boldsymbol{a}})$ is given in Figure 7. The 3 selected knots are represented by vertical dashed lines. The regression is now compared with P-splines (in dashed lines), which yields a similar estimation – although less regularized.

9 Conclusion

In this paper we introduce a method called A-spline (for *adaptive spline*) performing spline regression which automatically selects the number and position of the knots. For that purpose, we set a large number of initial knots and use an iterative penalized likelihood approach (the adaptive ridge) to sequentially remove the unnecessary knots. The model achieving the best bias-variance tradeoff is selected using the EBIC_0.

Our method yields sparse models which are more interpretable than classical penalized

spline regressions (e.g. P-splines). Other methods in the literature offer to jointly fit the spline and select the knot's position. Through simulations, we highlight that our method performs almost as well as these methods, while being way less computationally intensive (by at least a factor 100).

Since knots correspond to a shift in the trend of the spline, their position is an important information about the data. To illustrate this, we apply A-splines to several real data sets. When using A-spline with low order splines (e.g. 0 or 1), the approach allows performing changepoint detection. Indeed, A-splines of order 0 fit a piecewise constant function to the data and hence detect changepoint in terms of mean. A-spline of order 1 fits a piecewise linear continuous function (i.e. a continuous broken line) that detects changepoints in terms of slope. For splines of degree 3, the knots indicate shifts in the third order derivative of the underlying function.

A fast implementation of A-spline is provided in `R`. Thanks to this, the computation of A-spline is very fast (~ 1 sec for $n \sim 10000$ and $k \sim 1000$ on a standard laptop), even when fitting generalized linear models with large sample sizes.

Our work can be naturally generalized to multivariate data using multidimensional B-splines. Moreover, we limited our work to using B-splines for the sake of simplicity. But a variety of other splines can be used instead. For example M-splines, which are a basis of non-negative splines, could be used for fitting non-negative functions (e.g. densities) and I-splines, which are a basis of monotonous splines, would yield a sparse isotonic regression model. Finally, our method can be used for non-parametric transformation of variables. In particular, splines of order 0 could provide an automatic categorization of continuous covariates in regression models.

Supplementary Materials

The Supplementary Materials are available online as a single archive. They include all the code necessary to replicate the simulations and the real data applications present in this paper. The package `aspline` implementing A-splines is also given in Supplementary Materials. We refer to `https://github.com/goepp/aspline` for the latest version of the

package. Below is a description the files included in the Supplementary Materials.

Data sets

Motorcycle helmet crash data set (`helmet.csv`).

aCGH data set of bladder cancer (`bladder.csv`).

LIDAR measurement data set (`lidar.csv`).

Coal mine accident data set (`coal.csv`).

Code

simu_hetero Simulations with heteroscedastic errors (`simu_hetero.R`)

simu_homo Simulations with homoscedastic errors (`simu_homo.R`)

real_data_ Illustration of A-splines on real data sets and comparison with other spline regression methods. (`real_data_helmet.R`, `real_data_bladder.R`, `real_data_lidar.R`, `real_data_coal.R`)

Files

`R` **package** A-spline package (`aspline-master.zip`)

Appendix Simulation study comparing the A-spline and B-spline methods (`appendix.pdf`)

Readme Description of the Supplementary Materials and of their use (`readme.txt`).

References

Biller, C. (2000), 'Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models', *Journal of Computational and Graphical Statistics* **9**(1), 122–140.

Bleakley, K. and Vert, J.-P. (2011), 'The Group Fused Lasso for Multiple Change-Point Detection', *arXiv preprint arXiv:1106.4199* .

Chen, J. and Chen, Z. (2008), 'Extended Bayesian information criteria for model selection with large model spaces', *Biometrika* **95**(3), 759–771.

Curry, H. and Schoenberg, I. (1966), 'On Pòlya Frequency Functions IV: The Fundamental Spline Functions and Their Limits', *Journal d'Analyse Mathematique* **17**(1), 71–107.

de Boor, C. (1972), 'On calculating with B-splines', *Journal of Approximation Theory* **6**(1), 50–62.

De Boor, C. (1978), *A Practical Guide to Splines*, Vol. 27, Springer-Verlag New York.

de Souza, C. P. E. and Heckman, N. E. (2013), 'Switching Nonparametric Regression Models and the Motorcycle Data revisited', *arXiv preprint arXiv:1305.2227* .

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998), 'Automatic Bayesian curve fitting', *Journal of the Royal Statistical Society: Series B* **60**(2), 333–350.

Diggle, P. and Marron, J. S. (1988), 'Equivalence of Smoothing Parameter Selectors in Density and Intensity Estimation', *Journal of the American Statistical Association* **83**(403), 793–800.

DiMatteo, I., Genovese, C. R. and Kass, R. E. (2001), 'Bayesian Curve-Fitting with Free-Knot Splines', *Biometrika* **88**(4), 1055–1071.

Eddelbuettel, D. (2013), *Seamless R and C++ Integration with Rcpp*, Springer New York, New York, NY.

Eilers, P. H. C. and Marx, B. D. (1996), 'Flexible Smoothing with B-splines and Penalties', *Statistical Science* **11**(2), 89–102.

Eilers, P. H. C., Marx, B. D. and Durbán, M. (2015), 'Twenty Years of P-splines', *Statistics and Operations Research Transactions* **39**(2), 149–186.

Friedman, J. H. (1991), 'Multivariate Adaptive Regression Splines', *The Annals of Statistics* **19**(1), 1–67.

Friedman, J. H. and Silverman, B. W. (1989), 'Flexible Parsimonious Smoothing and Additive Modeling', *Technometrics* **31**(1), 3–21.

Friedman, J., Hastie, T. and Tibshirani, R. (2010), 'Regularization Paths for Generalized Linear Models via Coordinate Descent', *Journal of Statistical Software* **33**(1), 1–22.

Frommlet, F. and Nuel, G. (2016), 'An Adaptive Ridge Procedure for L0 Regularization', *PLoS ONE* **11**(2), e0148620.

Green, P. J. (1995), 'Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination', *Biometrika* **82**, 711–732.

Härdle, W. (1990), *Applied Nonparametric Regression*, number 19, Cambridge university press.

Hastie, T. (2018), 'Gam: Generalized Additive Models'.

Hastie, T., Friedman, J. and Tibshirani, R. (2001), *The Elements of Statistical Learning*, Springer Series in Statistics, 2nd edn, Springer New York.

Haupt, R. L. and Haupt, S. E. (2004), *Practical Genetic Algorithms*, 2nd ed edn, John Wiley, Hoboken, N.J.

Haynes, K., Killick, R., Fearnhead, P. and Eckley, I. (2016), 'Changepoint.np: Methods for Nonparametric Changepoint Detection'.

Holmes, C. C. and Mallick, B. K. (2001), 'Bayesian Regression with Multivariate Linear Splines', *Journal of the Royal Statistical Society: Series B* **63**(1), 3–17.

Holst, U., Hössjer, O., Björklund, C., Ragnarson, P. and Edner, H. (1996), 'Locally Weighted Least Squares Kernel Regression and Statistical Evaluation of LIDAR Measurements', *Environmetrics* **7**(4), 401–416.

Jamrozik, J., Bohmanova, J. and Schaeffer, L. (2010), 'Selection of Locations of Knots for Linear Splines in Random Regression Test-Day Models', *Journal of Animal Breeding and Genetics* **127**(2), 87–92.

Jupp, D. L. B. (1978), 'Approximation to Data by Splines with Free Knots', *SIAM Journal on Numerical Analysis* **15**(2), 328–343.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012), 'Optimal detection of changepoints with a linear computational cost', *Journal of the American Statistical Association* **107**(500), 1590–1598.

Leitenstorfer, F. and Tutz, G. (2007), 'Knot selection by boosting techniques', *Computational Statistics & Data Analysis* **51**(9), 4605–4621.

Lindstrom, M. J. (1999), 'Penalized Estimation of Free-Knot Splines', *Journal of Computational and Graphical Statistics* p. 21.

Luo, Z. and Wahba, G. (1997), 'Hybrid Adaptive Splines', *Journal of the American Statistical Association* **92**(437), 107–116.

Marquardt, D. W. (1963), 'An Algorithm for Least-Squares Estimation of Nonlinear Parameters', *Journal of the Society for Industrial and Applied Mathematics* **11**(2), 431–441.

Marx, B. D. and Eilers, P. H. (1998), 'Direct Generalized Additive Modeling with Penalized Likelihood', *Computational Statistics & Data Analysis* **28**(2), 193–209.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 2 edn, Chapman and Hall.

Osborne, M. R., Presnell, B. and Turlach, B. A. (1998), 'Knot Selection for Regression Splines via the LASSO', *Computing Science and Statistics* pp. 44–49.

O'Sullivan, F. (1986), 'A Statistical Perspective on Ill-Posed Inverse Problems', *Statistical Science* **1**(4), 502–518.

Rippe, R. C. A., Meulman, J. J. and Eilers, P. H. C. (2012), 'Visualization of Genomic Changes by Segmented Smoothing Using an L0 Penalty', *PLoS ONE* **7**(6), e38230.

Ruppert, D. (2002), 'Selecting the Number of Knots for Penalized Splines', *Journal of Computational and Graphical Statistics* **11**(4), 735–757.

Ruppert, D., Wand, M. and Carroll, R. J. (2009), 'Semiparametric Regression During 2003–2007', *Electronic Journal of Statistics* **3**, 1193–1256.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge University Press.

Schoenberg, I. J. (1946), 'Contributions to the problem of approximation of equidistant data by analytic functions. Part B. On the problem of osculatory interpolation. A second class of analytic approximation formulae', *Quarterly of Applied Mathematics* **4**(2), 112–141.

Schwarz, G. (1978), 'Estimating the Dimension of a Model', *The Annals of Statistics* **6**(2), 461–464.

Sigrist, M. W., Winefordner, J. D. and Kolthoff, I. (1994), *Air Monitoring by Spectroscopic Techniques*, Vol. 127, John Wiley & Sons.

Silverman, B. W. (1985), 'Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting', *Journal of the Royal Statistical Society, Series B* **47**, 1–52.

Smith, P. L. (1982), Curve fitting and modeling with splines using statistical variable selection techniques, Technical Report NASA Report 166034, NASA, Langley Research Center.

Spiriti, S., Eubank, R., Smith, P. W. and Young, D. (2013), 'Knot selection for least-squares and penalized splines', *Journal of Statistical Computation and Simulation* **83**(6), 1020–1036.

Stone, C. J., Hansen, M. H., Kooperberg, C. and Truong, Y. K. (1997), 'Polynomial Splines and their Tensor Products in Extended Linear Modeling', *The Annals of Statistics* **25**(4), 1371–1425.

Stransky, N., Vallot, C., Reyal, F., Bernard-Pierrot, I., de Medina, S. G. D., Segraves, R., de Rycke, Y., Elvin, P., Cassidy, A., Spraggon, C., Graham, A., Southgate, J., Asselain, B., Allory, Y., Abbou, C. C., Albertson, D. G., Thiery, J. P., Chopin, D. K., Pinkel, D. and Radvanyi, F. (2006), 'Regional Copy Number–Independent Deregulation of Transcription in Cancer', *Nature Genetics* **38**(12), 1386–1396.

Tibshirani, R. (1996), 'Regression Shrinkage and Selection via the Lasso', *Journal of the Royal Statistical Society. Series B* **58**(1), 267–288.

Wahba, G. (1990), *Spline Models for Observational Data*, Vol. 59, Society for Industrial and Applied Mathematics.

Wallstrom, G., Liebner, J. and Kass, R. E. (2008), 'An Implementation of Bayesian Adaptive Regression Splines (BARS) in C with S and R Wrappers', *Journal of statistical software* **26**(1), 1.

Wand, M. P. (2000), 'A Comparison of Regression Spline Smoothing Procedures', *Computational Statistics* **15**(4), 443–462.

Wand, M. P. and Ormerod, J. T. (2010), 'On Semiparametric Regression with O'Sullivan Penalised Splines', *Australian & New Zealand Journal of Statistics* **52**(2), 239–239.

Wang, W. and Yan, J. (2017), 'Splines2: Regression Spline Functions and Classes'.

Whittaker, E. T. (1922), 'On a New Method of Graduation', *Proceedings of the Edinburgh Mathematical Society* **41**, 63–75.

Wood, S. N. (2017), *Generalized Additive Models: An Introduction with R*, 2 edn, Chapman and Hall/CRC.

Żak-Szatkowska, M. and Bogdan, M. (2011), 'Modified Versions of the Bayesian Information Criterion for Sparse Generalized Linear Models', *Computational Statistics & Data Analysis* **55**(11), 2908–2924.

4.3 Comparison of A-spline with P-spline

The following pages consist in the supplementary materials to the previous section. In this section, I provide additional simulations aimed at comparing the A-splines with the P-splines, both qualitatively and quantitatively.

Appendix to the article "Spline Regression with Automatic Knot Selection"

Vivien Goepp *
MAP5 (CNRS UMR 8145), Université Paris Descartes
Olivier Bouaziz
MAP5 (CNRS UMR 8145), Université Paris Descartes,
and
Grégory Nuel
LPSM (CNRS UMR 8001), Sorbonne Université

In this Appendix, the performance of A-splines is compared to a penalized spline regression method. For the sake of simplicity, we limit our study to the comparison with P-splines (Eilers and Marx, 1996), which is one of the most successful penalized spline regression methods. The goal of A-splines is fit a model that is sparser and consequently more interpretable then that of P-splines. Therefore, A-splines are not required to have a better predictive performance than P-splines. This simulation study aims to illustrate that the loss in performance is acceptable when using A-splines instead of P-splines. We use the same simulation setting as in Section 6.2 of the article. We use the EBIC_0 criterion to select the penalty.

Figure 1 represents the fitted functions with A-splines and P-splines for the four functions with datasets of size 200. The thick lines represent the estimated functions; the thin lines represent the splines' basis decomposition. With every function, A-spline and P-spline yield similar estimates. The basis decomposition highlights that A-spline selects very sparse models, which are also simpler. Over the 500 replications, A-spline selects a median number of 9 splines for the *Bump* function, 6 for the *Logit* function, 11 for the *Sine* function, and 7 for the *SpaHet* function. An important remark is that when there are few knots, their location carry an important information about the underlying signal. Indeed,

*vivien.goepp@parisdescartes.fr

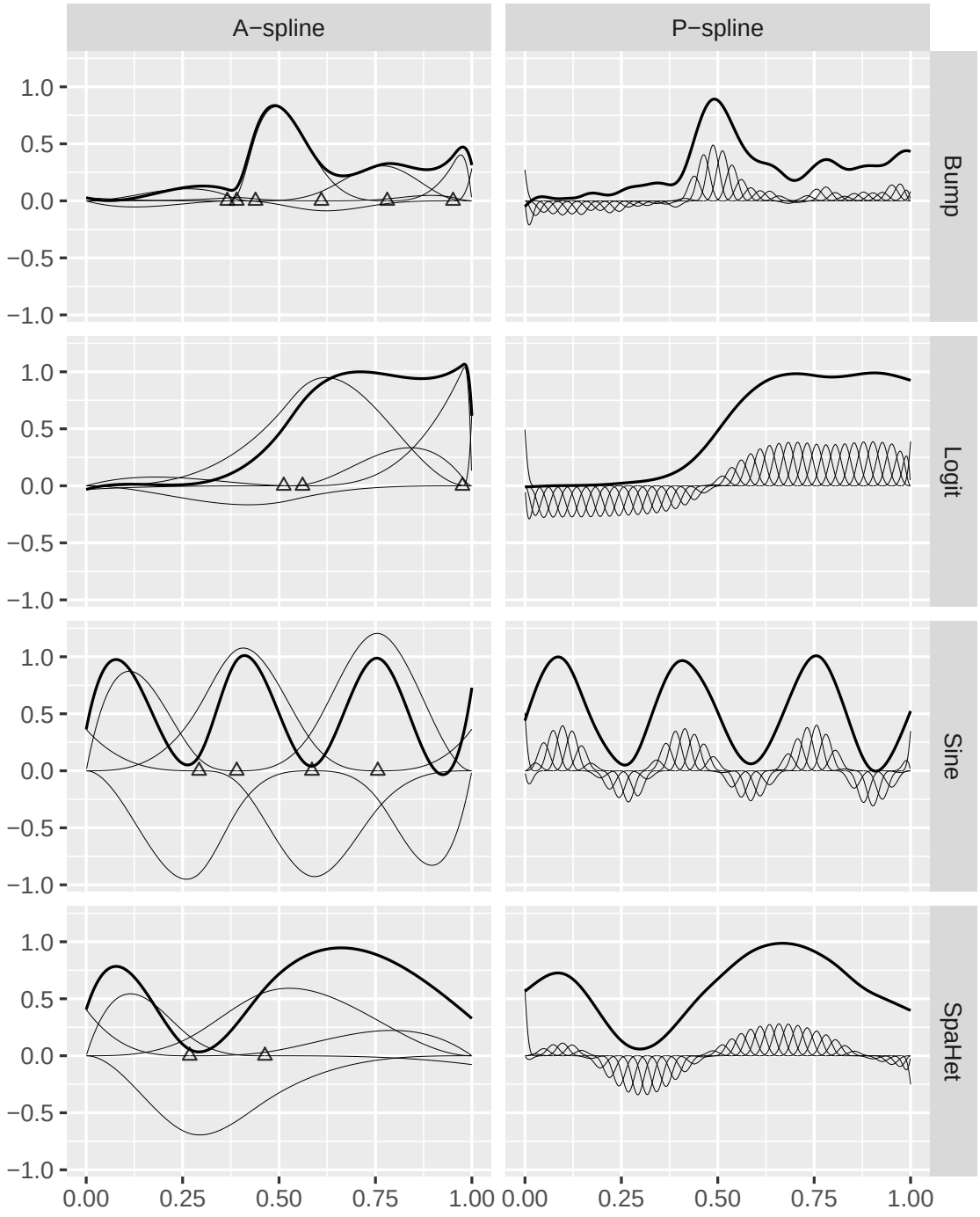
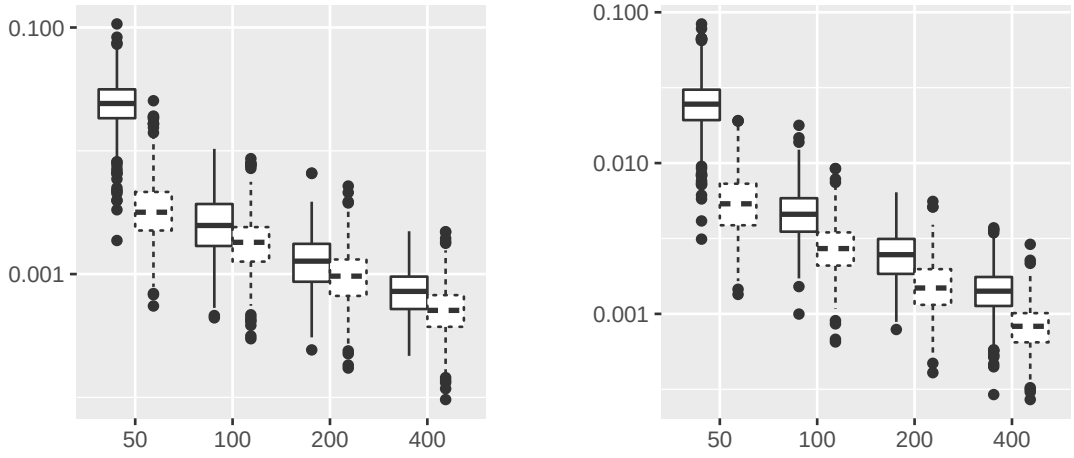
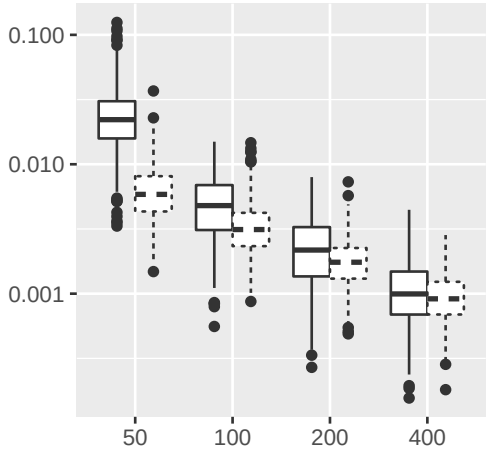


Figure 1: A-spline and P-spline regressions over different functions (tick lines). Basis decomposition of the fitted splines are represented in thin lines. For the A-spline regression, triangles represent the selected knots. The sample size is 200.

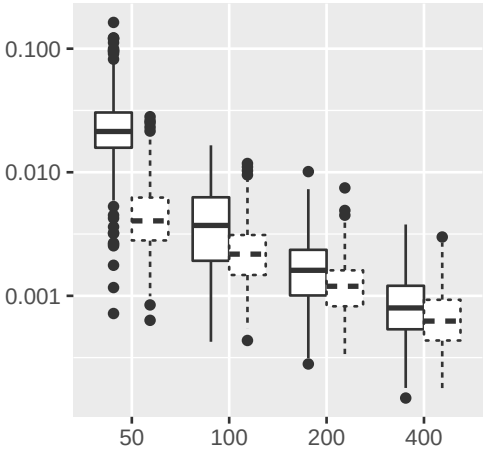


(a) Logit

(b) Sine



(c) Bump



(d) SpaHet

Figure 2: Mean squared errors of A-spline (solid line) and P-spline (dashed line) estimates for different sample sizes: 50, 100, 200, and 400. The simulations are performed with the *Bump*, *Logit*, *Sine*, and *SpaHet* functions and repeated 500 times.

the knots selected by A-spline are mostly located at *shifts* – when they exist – in the signal: for the *Bump* function, the knots corresponds to places in the data with a lot variation, and for the *Logit* function, two knots are placed at the inflection point.

In order to verify that A-spline's gain in interpretability is not detrimental to its predictive performance, we compare the Mean Square Errors between the two methods. Figure 2 shows the MSE for A-splines (solid lines) and P-splines (dotted lines) for every sample size and every function. It shows that for sample size 50, P-splines performs on average better than A-splines. For greater sample size however, A-splines performs almost as well as P-splines. These results are the same for every function. In conclusion, P-splines has better predictive performance for small sample sizes, but for data sets of size 200 and above, A-splines and P-splines turn out to have comparable predictive performance.

References

Eilers, P. H. C. and Marx, B. D. (1996), 'Flexible Smoothing with B-splines and Penalties', *Statistical Science* **11**(2), 89–102.

Chapter 5

Segmentation of spatial data

Contents

5.1 Introduction 136
5.2 A model for spatial segmentation 137
5.2.1 Using graphical data for spatial segmentation 137
5.2.2 Segmentation on a graph 138
5.2.3 The adaptive ridge algorithm on a graph 139
5.3 Simulation 142
5.4 Real data application: Overweight prevalence in the Netherlands 146
5.5 Conclusion 148



Figure 5.1: Map of cholera outbreaks in the neighborhood of Soho (London) by John Snow (1854). Crosses are water wells, dots are cholera outbursts.

5.1 Introduction

Spatial statistics plays a prominent role in epidemiology. The founding of epidemiology is said to occur with studies of spatial data of contamination. In 1854, the physician John Snow studies the geographical distribution of cholera outbreaks during the cholera epidemic of 1854 in the London neighborhood of Soho. He represented each outbreaks on a city map and over imposed the locations of water pumps in this neighborhood (see Figure 5.1). It enabled to identified a public water pump as the origin of the epidemic. This is the first record of a geographical analysis being used to epidemiology, and John Snow is considered to be the founding father of epidemiology.

Nowadays, the study of disease with respect to the geographical location of the patients has become widespread, and it plays a major role in epidemiology. Spatial (or geographical) epidemiology is based on the comprehension that there are geographical factors that are connected to a disease or to health indicators. It is interested in the detection and study of *(i)* "disease clusters" which are limited regions with unusually high values of a health indicator and *(ii)* "geographic correlation studies", which are the study of the effect of geographical location on the risk of disease onset. The data for these studies can be of various types, leading to different statistical tools to answer the epidemiological questions at hand. We refer to Lawson et al. (2016); Lawson (2006) for a review of the field and to Elliott and Wartenberg (2004); Rezaeian et al. (2007) for a summary of the tools and stakes thereof.

A regular problem in spatial statistics is the regularization of the spatial data. Most regularization methods perform a spatial smoothing of the data. These methods offer some advantages: (a) reduction of spatial covariance, (b) higher interpretability of the resulting map, and (c) interpolation. One of the most popular approaches to spatial smoothing is *kriging*, or Gaussian process regression (Cressie, 1993, Section 3). It assumes that the spatial data are the realization of a Gaussian process, with a specific covariance structure. The Gaussian process can be computed over the whole map using the covariance process at the observed points and the values of the process at the observed points. Since the Gaussian process is a Gaussian random variable at any point, the process's mean and variance can be estimated, and the whole distribution of the process is estimated. Computationally, this estimate requires the inversion of the covariance matrix, whose dimension is the sample size. This method is thus easy to compute and to interpret, and the assumption of an underlying Gaussian process is reasonable in many cases in absence of other information. The choice of covariance kernel tunes the estimated process: the smoother the covariance kernel, the smoother the estimated process. Other techniques of spatial smoothing have been developed; a popular approach is kernel smoothing (Wand and Jones, 1995).

Spatial smoothing has many advantages. In some cases, one may want to obtain a segmented es-

timation of the spatial distribution instead of a smoothed estimation, in cases where the underlying spatial effect is assumed to be discontinuous by nature. In many public health indicators, the spatial information is a proxy for a very located information, that is discontinuous by nature. In cities, the discretization into small neighborhoods is often very significant to tackle the spatial inequalities. When adjusted for professional category, income, wealth, and other quality-of-life variables, the neighborhood in which we live codes for the sole quality of life of each neighborhood provides, which is often a discontinuous variable. The same principle applies to demographic and epidemiologic studies where the quality of life is an important explanatory variable. This variable is of discrete nature and is indirectly present in the spatial distribution of the individuals. As an example, demographic studies on longevity focus namely on finding specific geographic areas where the longevity is unexpectedly high (Poulain et al., 2004) – called *blue zones*. These areas are discrete by nature, and such studies use a spatial division into (administrative) areas to identify blue zones.

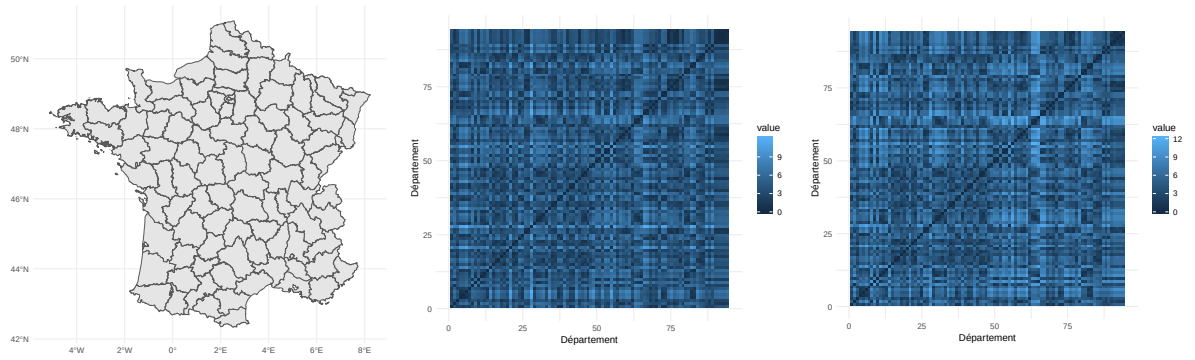
Our work is considering the case where the spatial distribution is indexed by a finite number of geographical units, called "areas" here (sometimes also called geographical sites). These areas represent a partition of the spatial domain. They often correspond to an administrative division of a territory, for instance census tracts, municipalities, or counties. Many studies present this type of data. This is due to the way data are collected by the administration (as is the case for demography) or to the imperative of privacy (as is the case in epidemiology). The areas are defined as polygons on $\mathbb{R}^2$. This work is built on solely using the planar graph of the adjacency structure of the areas as the information representing the spatial distribution. Doing so can seem to throw away information about the geographical proximity between areas, namely because two areas can be geographically close on average (i.e. for the Euclidean distance) and not be adjacent. However when the division into areas is regular enough, as is the case for most administrative unit types, the adjacency between areas is a good measure of proximity (see Section 5.2.1). Examples of studies based on the adjacency of areas are popular in epidemiology (van de Kassteele et al., 2017) and demographics (Poulain et al., 2004). This field of spatial statistics has been widely studied, including modelizations using Gaussian distributions and Markov random fields (Cressie, 1993, see Sections 6.3 and 6.4). We refer to Cressie and Read (1989) and (Cressie and Chan, 1989) for applications of such models to real-world data.

In this context, we introduce a new model for estimation of the spatial effect with segmentation based only on the adjacency of the areas. To the best of our knowledge, this problem has not been tackled in this context. We use a regularization method by penalized likelihood to perform segmentation over the graph. This approach can be seen as a generalization of the fused lasso-type regularization used in image denoising (Tibshirani et al., 2005), where the adjacency structure of pixels is a lattice. As in previous chapters, we use the Adaptive Ridge to penalize over the differences between adjacent areas. Our method yields a segmented estimate of the spatial effect in a computationally efficient way, because the complexity depends on the number of areas and not on the number of individuals. Particular attention has been brought to speeding up the computation, which is tractable up to a large number of areas. Our work is organized as follows. Section 5.2 introduces the model. In Section 5.3, we illustrate on simulated data that out method performs well to recover areas of constant spatial effect. Moreover, our method is shown to perform well also in the case where the true (unknown) effect has a smooth variation over space. Illustration with a real-data example of obesity prevalence in the Netherlands is provided in Section 5.4. The implementation of our method in `R` is publicly available (`github.com/goepp/graphseg`).

5.2 A model for spatial segmentation

5.2.1 Using graphical data for spatial segmentation

Our work is built on the following assumption: the areas provide a spatial discretization adapted to the sample distribution. This implies that the areas are not too small with respect to the overall spatial distribution, and not too large either. When the areas are too large (or equivalently when



(a) *Départements* (b) Euclidian distances (c) Shortest path distances

Figure 5.2: (a) Division of continental France into 94 *départements* of roughly uniform size. (b) Euclidean distances between the centroids of each *départements*, using the latitude and longitude as coordinates. (c) Shortest path distances between *départements*, using the adjacency graph.

there are too few areas), most of the sample is located in few areas, and a lot of information is lost. When the areas are too small (or equivalently when there are too many areas), many areas have no individuals and thus carry no information on the outcome variable. This case is not problematic in our work, since our objective is to find groups of adjacent areas with similar effects. However the total number of areas is a limiting factor for computation efficiency. Moreover, we assume that the areas are roughly equal in size, so that the discretization into geographical units is roughly uniform over space. In applications where there are several connected components (like islands), one can consider each component as a separate problem an perform segmentation independently from each other. If one wishes to introduce a notion of proximity between the connected components, one can connect every pair of components by artificially adding an edge between their closest areas (using the Euclidean distance). This case is not considered in the remainder of this work. In the forthcoming, we assume that the areas form a planar graph with only one connected component.

Under these conditions, we can see that the distance between two areas is well approximated by its adjacency distance, that is the minimal number of other areas one needs to cross to join the two areas. This is illustrated with the following example. We use the *départements* of France as the geographical areas (data obtained at `github.com/gregoiredavid/france-geojson`, collected from the French national statistical institute (INSEE) and the French national geographical institute (IGN)). We exclude the overseas *départements* as well as the two *départements* forming the island of Corsica. The 94 *départements* are represented in Figure 5.2a, numbered using their codes. Since *départements* number 7 and 8 are removed, the numbering of the *départements* number 9 and more are offset by 2. We make the assumption that the distance given by the shortest path on the adjacency gives a good approximation of the geographic (i.e. Euclidean) distance between the areas – in practice, we will take the distances between the centroids of the areas. We compare the similarity matrices of these two distances, whose (i, j)-th entry is the distance between the areas i and j. There is a strong visual similarity between the two similarity matrices: see Figures 5.2b and 5.2c. We could test whether the two distances are similar in a sense to be defined. This is beyond the scope of this study.

5.2.2 Segmentation on a graph

Let n be the sample size and $y_i, i = 0, \dots, n$ be the response variable for the n individuals. Let $\mathscr{D} \in \mathbb{R}^2$ be the connected subset of $\mathbb{R}^2$ representing the spatial domain of interest. Let $\boldsymbol{A} = (A_j)_{1 \leq j \leq p}$ denote the areas and p denote the number of areas. The areas form a partition of $\mathscr{D}$: $\forall\, j, A_j \subset \mathscr{D}$, $\forall\, j, j', A_j \cap A_{j'} = \emptyset$, and $\cup_j A_j = \mathscr{D}$. Each individual is assumed to be in only one area: $\sum_{i=1}^{n} \mathbb{1}_{i \in A_j} = 1$. Let $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ be

the adjacency graph of $(A_j)_j$: each vertex $v \in \mathcal{V}$ corresponds to an area A_j and there exists and edge between two edges (v, v') if their corresponding areas have a border in common (areas connected by only one point are not said to be adjacent). The areas (A_j) are identified with their nodes $\mathcal{V}$. The effect of each area over the response variable is noted $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$.

Our method applies to segmenting any area-based signal $\boldsymbol{\gamma}$. When this signal is itself an estimate over the individuals in each area, our model includes its estimated variance $\hat{\boldsymbol{\sigma}}^2 = (\hat{\sigma}^2)$ as a measure of the relative importance to give to each area.

We assume that $\hat{\gamma}_j \sim \mathcal{N}\left(\theta_j, \hat{\sigma}_j^2\right)$ for each area j. In this context of meta-analysis, $\hat{\gamma}_j$ are considered as observations and θ_j as parameters. The negative log-likelihood simply writes:

$$\ell(\boldsymbol{\theta}) = -\log\left[\sum_{j=1}^{p} \frac{1}{\sqrt{2\pi}\hat{\sigma}_j} \exp\left(-\frac{\left(\hat{\gamma}_j - \theta_j\right)^2}{2\hat{\sigma}_j^2}\right)\right], \tag{5.1}$$

In order to enforce the values of $\boldsymbol{\gamma}$ which are both adjacent and close in value to have close estimated $\boldsymbol{\theta}$, we use a penalized likelihood estimated defined as the minimizer of:

$$\ell^{\text{pen}}(\boldsymbol{\theta}, \boldsymbol{\nu}) = \ell(\boldsymbol{\theta}) + \frac{\kappa}{2} \sum_{j,k} p(\theta_j - \theta_k) \tag{5.2}$$

with respect to $\boldsymbol{\theta}$. where the sum is taken over all adjacent nodes (j, k), $\kappa > 0$ is the penalty constant, and p is a zero inducing penalty function (for instance a non-concave penalty, see Section 1.1.7). Since we penalize over the differences of values between adjacent edges, many adjacent areas will be estimated to have the same value. Consequently, the penalized estimate is piecewise constant over few connected components of the graph. In this regard, this penalization procedure over a graph is related to the graphical lasso (Friedman et al., 2008), which penalizes over the non-zero entries of the sensitivity matrix.

5.2.3 The adaptive ridge algorithm on a graph

To ensure segmentation of $\boldsymbol{\theta}$, we must take a zero-inducing penalty. We choose the L_0 Adaptive Ridge penalty, an iterative method developed by Rippe et al. (2012) and Frommlet and Nuel (2016). We define a matrix of weights $\boldsymbol{\nu} = (\nu_{j,k})_{1 \leq j,k \leq p}$ between adjacent areas (the areas that are not adjacent have a weight $\nu_{j,k}$ set to zero by convention). The L_0 adaptive ridge procedure is defined iteratively. The current estimate is defined as the minimizer of

$$\ell^{\text{pen}}(\boldsymbol{\theta}) = \sum_{j=1}^{p} \frac{(\theta_j - \hat{\gamma}_j)^2}{2\sigma_j^2} + \kappa \sum_{j,k} \nu_{j,k} \left(\theta_j - \theta_k\right)^2. \tag{5.3}$$

The weights are initialized at the beginning of the iteration and are adapted at each step following the formula:

$$\nu_{j,k} = \frac{1}{\left(\theta_j - \theta_k\right)^2 + \varepsilon^2}. \tag{5.4}$$

where $\varepsilon > 0$ is a small numerical constant. We iterate between estimating $\boldsymbol{\theta}$ using (5.3) with fixed weights and adapting the weights using the previously estimated values of $\boldsymbol{\theta}$. At convergence, the weighted differences $\delta_{j,k} \triangleq \nu_{j,k}(\theta_j - \theta_q)^2$ is almost equal to either zero (if the two values are set equal) or one (if the two values are set different). Using a cutoff threshold, the weighed differences are rounded to zero or one, which defines a set of connected components of the graph, that we call "regions". These connected component are the segmentation of the areas into groups of constant spatial effect. The adaptive ridge procedure performs a model selection amongst all the possible division of the graph into connected components. This effect $\boldsymbol{\theta}$ is then estimated over each connected component by unpenalized estimation The weighted differences $\boldsymbol{\delta}$ is used to diagnose the convergence of the algorithm. The whole procedure is given in Algorithm 6.

Algorithm 6 Adaptive ridge procedure over a graph

1: **function** ADAPTIVE-RIDGE($\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\sigma}}^{\boldsymbol{2}}$)
2: $\boldsymbol{\theta}^{\text{old}} \leftarrow \mathbf{0}$
3: $v_{i,j}^{\text{old}} \leftarrow \mathbb{1}_{i \text{ and } j \text{ are adjacent}}$
4: $\delta_{j,k}^{\text{old}} \leftarrow v_{j,k}(\theta_j - \theta_k)^2$
5: **while** not converge **do**
6: $\boldsymbol{\theta} \leftarrow \arg\min_{\boldsymbol{\theta}} \ell^{\text{pen}}(\boldsymbol{\theta}, \boldsymbol{v}^{\text{old}})$
7: $v_{j,k} \leftarrow \left((\theta_j - \theta_k)^2 + \varepsilon^2\right)^{-1}$
8: $\delta_{j,k} \leftarrow v_{j,k}(\theta_j - \theta_k)^2$
9: **if** $\|\boldsymbol{\delta} - \boldsymbol{\delta}^{\text{old}}\| < 10^{-8}$ **then**
10: break
11: **end if**
12: $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}$
13: $\boldsymbol{v}^{\text{old}} \leftarrow \boldsymbol{v}$
14: $\boldsymbol{\delta}^{\text{old}} \leftarrow \boldsymbol{\delta}$
15: **end while**
16: **return** $\boldsymbol{\theta}$
17: **end function**

The last point is the numerical minimization of Equation 5.3 (Line 6 in Algorithm 6). We first reformulate (5.2) using the following property. Let $\boldsymbol{L}$ be the Laplacian matrix of the weighted adjacency graph, that is, $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{V}$, where $\boldsymbol{D}$ is the degree matrix of the graph and $\boldsymbol{V} = (v_{j,k})_{j,k}$ is the matrix of weights between adjacent areas. Then the following identity (see Mohar, 1997, Proposition 2.2) holds:

$$\sum_{j,k} v_{j,k}(\theta_j - \theta_k)^2 = \boldsymbol{\theta}^T \boldsymbol{L}\boldsymbol{\theta}$$

and thus (5.2) rewrites

$$\ell^{\text{pen}}(\boldsymbol{\theta}) = \|\boldsymbol{W}^T(\boldsymbol{\theta} - \boldsymbol{\Gamma})\|^2 + \kappa\boldsymbol{\theta}^T\boldsymbol{L}\boldsymbol{\theta}. \tag{5.5}$$

Where

$$\boldsymbol{W} = \left(\frac{1}{\sqrt{2\sigma_j^2}}\right)_j$$

is a vector of weights of the areas (not to be mixed up with the weights between areas, noted $v_{j,k}$). The minimizer of this quantity is explicit:

$$\arg\min_{\boldsymbol{\theta}} \ell^{\text{pen}}(\boldsymbol{\theta}) = (\boldsymbol{W}\boldsymbol{W}^T + \kappa\boldsymbol{L})^{-1}\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{\Gamma}. \tag{5.6}$$

Consequently, the minimization of $\ell^{\text{pen}}(\boldsymbol{\theta})$ in Line 6 of Algorithm 6 is explicit. This significantly improves the computational speed of our method. Note that the computation of (5.6) is the bottleneck of our method, the complexity thereof depends on the dimension of $\boldsymbol{L}$, that is, on p. Hence the limiting factor of our algorithm is the number of areas.

Remark. Our method is closely related to the problem of dividing an image into connected clusters of equal brightness. (This is not to be mistaken for the clustering of images into unconnected clusters of equal brightness, called image thresholding.) Consider a graph that is a lattice (or a square grid graph): each area has 4 neighboring areas (except those on the "border" and in the "corners", which have 3 and 2 neighboring areas, respectively). Since an image is a grid of pixels, the lattice is the adjacency structure of the image, where each area is a pixel. Then penalizing over the adjacent areas (see Equation 5.3) is like penalizing over the differences of adjacent pixels. This is the approach taken by

penalty-based denoising methods, such as total variation (Rudin et al., 1992b). The image processing community usually uses a penalization of the L_1 norm of the differences of adjacent pixel values. But this choice is mainly motivated by the development of methods for solving this problem numerically, like the coordinate descent (Friedman et al., 2007) or the primal-dual algorithm (Chambolle and Pock, 2011). We could instead apply another penalty, like that of the L_0 adaptive ridge. Consequently, the present work generalizes directly to segmentation of images (into a piecewise constant image). This is however beyond the scope of this chapter.

Choice of the trade-off constant The penalized likelihood approach makes use of a penalty constant: κ. In this section, we tackle the choice of this tradeoff parameter. In practice, the choice of κ is done *a posteriori*: we run the penalized estimation for a series of values of κ, and then select the best value. We denote $\boldsymbol{\kappa} = (\kappa^{(1)}, \cdots, \kappa^{(L)})$, the sequence of values of the penalty constant, in increasing order. As previously explained, the adaptive ridge performs model selection: to each $\kappa^{(l)}$ corresponds one model. Several values of the penalty constant can induce the same model and according to the procedure of the adaptive ridge, if two models are the same, their estimate are the same. Consequently the criteria used for model selection depend only on the model dimension.

We use the Bayesian Information Criterion (BIC, see Schwarz, 1978):

$$BIC(m) = 2\ell(\hat{\boldsymbol{\theta}}_m) + m \log n, \tag{5.7}$$

where m is the model dimension (i.e., the number of estimated regions) and $\hat{\boldsymbol{\theta}}_m$ denotes here its corresponding estimate.

Remark. This criterion is shown to perform well in simulations and on real data. However, the BIC does not ensure that the selected models are of small enough dimensions, especially when p gets too large. This is explained here. The BIC is the criterion that maximizes the posterior probability of the model conditionally on the data, in the Bayesian framework. The BIC implicitly sets a uniform prior over all the models. Now consider what this prior implies on the dimension of the model: there are very few models of small dimension (close to 1) and of maximal dimension (close to p) and there are many models of medium dimension (close to $p/2$). This implies that the *a priori* probability of models of small dimensions is very small and consequently, the BIC will very probably not select very sparse models. When p gets large, the BIC fails to select sparse models (in practice, with ~ 20 parameters), which is a desired goal. We note that other well-known model selection criteria are less penalizing than the BIC, so are of no use here.

One can add a corrective prior term to the BIC which would assign equal probability to all model dimensions. This requires to enumerate the number of models for each dimension. That is: for $k \in \{1, \cdots, p\}$, we want to compute the number of ways of dividing the planar graph $\mathcal{G}$ in k connected components. There seems to be no easy approach of this problem, which, interestingly, is not mentioned in the literature on planar graphs. Finding an exact formula seems out of reach. I believe there are good hopes to develop an algorithm that computes this enumeration. However *(i)* finding a polynomial-time algorithm seems much harder a problem *(ii)* even in that case, such an algorithm could take a long time to run on a large graph ($p \gtrsim 1000$). The study of this problem has been relegated as potential future work.

Visual inspection. As an alternative to the BIC, we propose to use visual inspection to choose the best-fitting penalty. This does not provide an automated and non-arbitrary way of selecting the best model and as such, is not a satisfying method. Instead, we advise the practitioner to use visual inspection as a verification that the BIC selects a "coherent" model. Concretely, our method is applied on a sequence of penalty and one can consider that the resulting estimate is a sequence of segmentations, ranging from the least regularized to the most regularized. In this case, the practitioner can inspect if some segmented regions correspond to a sensible division or if they are isolated areas which



Figure 5.3: Iris of the city of Paris used for simulation

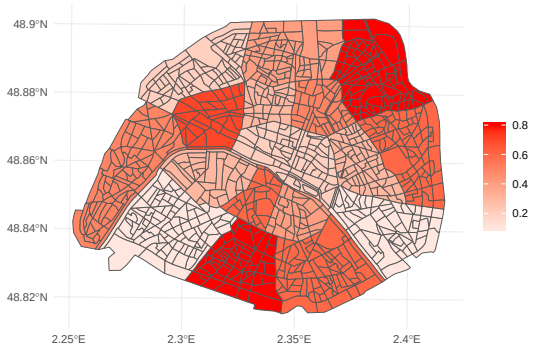
may have been detected due to the method's instability in selection (we refer to Fan and Li, 2001, Section 2 for an introductory discussion on the selection instability of penalized methods).

5.3 Simulation

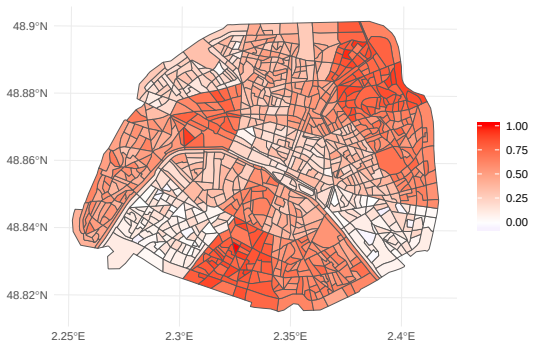
We illustrate the model on several simulation settings. We use the spatial structure of *iris* (for *Ilots Regroupés pour l'Information Statistique*, i.e. "grouped islets for statistical information"), a geographical division of the French territory used by the national French statistical institute. There are 16100 iris areas dividing the territory into roughly uniform areas geometrically and demographically. The iris data can be openly accessed from the French government's website[1]; more information about iris can be found from the INSEE's website[2]. We choose iris because they are a good example of uniform division into administrative areas, as is the case in many problems in spatial statistics. Moreover, they are numerous enough to illustrate the numerical efficiency of our approach and the iris are connected (in the sense of graph connectivity). We limit the domain to the city of Paris, which comprises 987 iris of roughly equal size (see Figure 5.3). For the sake of the visual representation, we removed the 5 disproportionally large iris that constitute two forests east and west of Paris. We simulate a spatial distribution $\boldsymbol{\gamma}$ of different types, leading to different simulation settings. In both simulation settings, the variance term $\hat{\sigma}^2$ is set to 1 for each area.

First simulation setting. We defined the observed spatial effect to be piecewise constant over a small number of regions. These regions are the arrondissements of Paris. We merge the arrondissements 1 and 2 together as well as the 3rd and 4th arrondissements, so that there are a total of 18

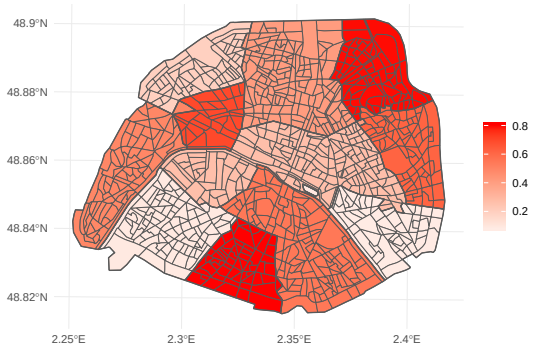
[1] `www.data.gouv.fr/fr/datasets/contours-iris-insee-ign/`
[2] `www.insee.fr/fr/information/2017499`



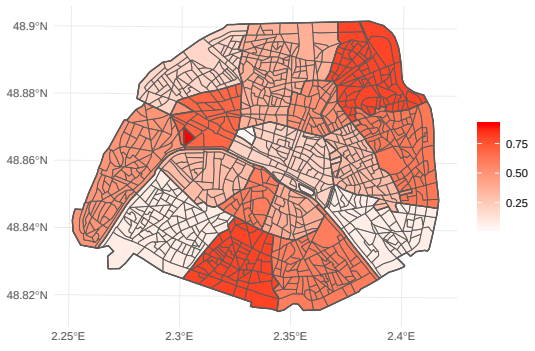
(a) Piecewise constant signal $\boldsymbol{\alpha}$



(b) Spatial signal $\boldsymbol{\gamma}$



(c) Segmented signal $\boldsymbol{\theta}$ using the BIC



(d) Segmented signal $\boldsymbol{\theta}$ with 18 regions

Figure 5.4: Result of simulation setting 1: piecewise constant effect $\boldsymbol{\alpha}$ (a), noisy piecewise constant effect with additive Gaussian noise $\boldsymbol{\gamma}$ (b), segmented estimate using the L_0 adaptive ridge penalization $\boldsymbol{\theta}$ with the BIC (c), and with 18 regions (d).

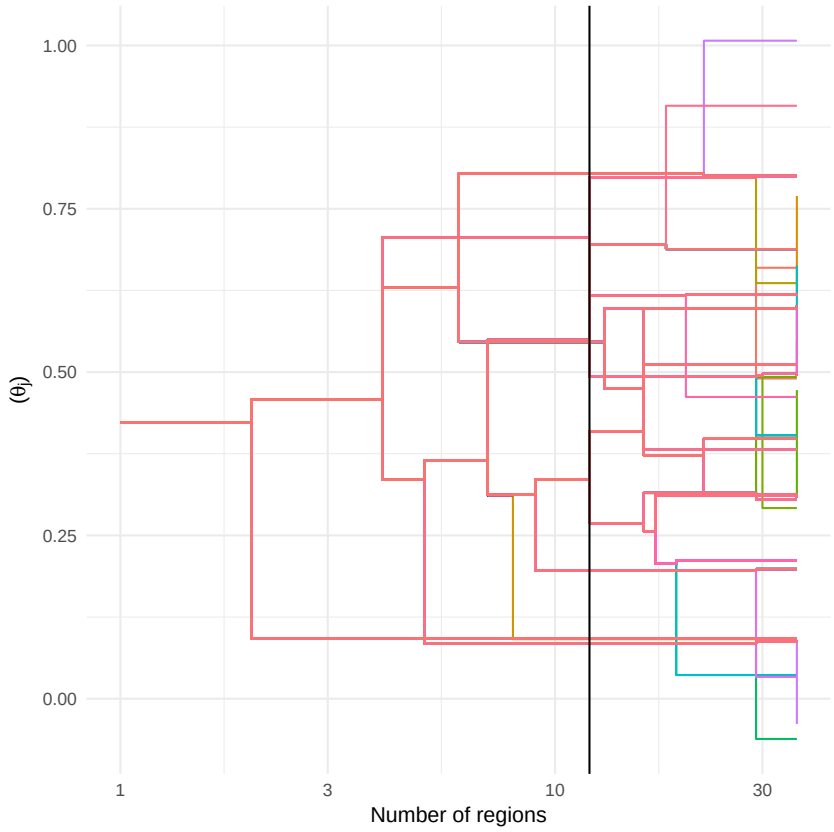


Figure 5.5: Regularization path in simulation setting 1: trajectories of the estimated $(\theta_j)_j$ for different values of κ. For a better visualization, the x axis gives the number of estimated parameters instead of κ. The vertical line represents the estimate selected by the BIC. Each color represents a different area.

regions of comparable size, noted $(R_k)_{1\leq k\leq 18}$ in the increasing order of the arrondissement numbering. The aim of this simulation setting is to quantify how well our method can recover these regions. The spatial term is defined as

$$\boldsymbol{\gamma}_j = \alpha_j + \varepsilon_j, \tag{5.8}$$

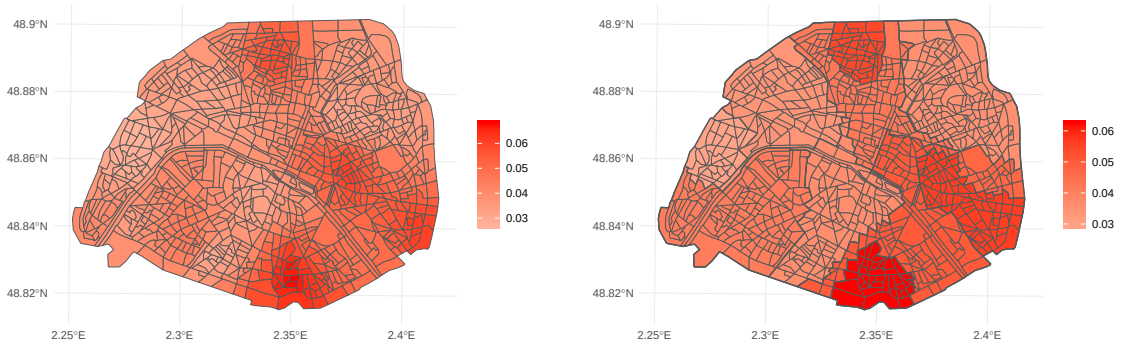
where $\boldsymbol{\alpha}_j = \sum_{k=1}^{18} \mathbb{1}_{A_j \in R_k} \eta_k$ is the effect over A_j, $\boldsymbol{\eta} = (\eta_k)_{1\leq k\leq K}$ is the vector of values taken by this function, and $\boldsymbol{\varepsilon} = (\varepsilon_j)_{1\leq j\leq p}$ is a vector of iid Gaussian errors:

$$\varepsilon_j \sim \mathcal{N}(0, 0.25). \tag{5.9}$$

We set $\boldsymbol{\eta} = (4,2,4,6,3,7,3,5,3,1,6,8,1,5,2,4,8,6)$. This piecewise constant "function" is illustrated in Figure 5.4a and the observed effect is given in Figure 5.4b.

The resulting estimate obtained using our method is given in Figure 5.4c. Our method accurately estimates most regions R_j. The limits between the regions are accurately estimated, except for a total number of 7 areas. However, the BIC tends to underestimate the total number of regions: it selects 12 regions but there are 18 regions.

To illustrate that the segmentation method performs well if the correct number of regions is correctly selected. We represent in Figure 5.4d the segmented estimate with a penalty constant κ chosen such that there are 18 selected regions. This estimate is close to the original piecewise constant signal; the only difference in terms of segmentation being that the arrondissements 9 and 18 are fused together, as well as 1/2 and 3/4. Since this information is not known *a priori*, this estimate cannot be used to assess the method's performance in selection. It is however used here to highlight that the correct segmentation can appear somewhere on the regularization path. This indicates that with



(a) Spatial signal $\boldsymbol{\gamma}$ (b) Segmented signal $\boldsymbol{\theta}$

Figure 5.6: (a) Smooth spatial signal $\boldsymbol{\gamma}$ used in simulation setting 1 (b) and segmented estimate $\boldsymbol{\theta}$ using the BIC.

additional work done on finding a model selection method more fitted to this problem, our method could show an even good performance in selecting regions of constant effect.

We represent the regularization path of the estimation procedure in Figure 5.5. Each line represent the estimate value for on area across different values of κ, or rather for different values of number of estimated parameters, which is equivalent (the greater κ the fewer regions are estimated). The vertical line represents the model chosen by the BIC. The path illustrates the functioning of the adaptive ridge: for high values of κ, only one region is estimated, and as κ decreases, there are sudden shifts in the number of estimated regions. The breaking point where an additional region is estimated corresponds to a splitting of a colored line in two. The trajectories illustrate that the segmentations are encapsulated models: once two regions have been fused together (as κ increases), the region is never split again for higher values of κ. Consequently, the dendrogram provided that the regularization path forms is a hierarchical clustering of all the areas.

Second simulation setting. In the first simulation setting, $\boldsymbol{\gamma}$ is constant over a set of regions and our method satisfyingly estimates these regions. In order to test the performance of our model it is necessary to evaluate it in a setting where the underlying signal is not piecewise constant. What it means for a signal over a graph to be "smoothed" is defined here. Signal processing on graphs have been developed using the shortest-path distance. We refer to Shuman et al. (2013) for a review on this topic. The Laplacian of the graph plays a major role in the study of signals on graphs. This work naturally extends the results in signal processing of signals with discrete support. For example when the graph is a two-dimensional lattice, the Laplacian of the graph is the same as the Laplacian operator for discrete two-dimensional signals. We can apply filters to graph-based signals as for the signals indexed by integers (see Kalofolias, 2016, and references therein).

In this simulation, we define the spatial signal as the smoothing of $\boldsymbol{\gamma}^{(0)}$ obtained by heat filtering (Zhang and Hancock, 2008). This writes simply

$$\boldsymbol{\gamma} = \exp(-s\boldsymbol{L})\boldsymbol{\gamma}^{(0)}, \tag{5.10}$$

where the exponential is the matrix exponential, $\boldsymbol{L}$ is the Laplacian of the graph, and $s \in (0,\infty)$ is the smoothing parameter. We take $s = 2$ here. The resulting spatial effect is represented in Figure 5.6a.

The estimated segmentation is represented in Figure 5.6b. The BIC has selected 13 regions and their estimated effect $\boldsymbol{\theta}$ has similar values. In fact, the "segmented" estimate (Figure 5.6b) displays a gradient of values and is close to the underlying smooth effect $\boldsymbol{\gamma}$. This highlights that our method is robust to models where the true spatial effect is not in fact piecewise constant.

5.4 Real data application: Overweight prevalence in the Netherlands

We apply our method to the analysis of overweight prevalence indicators in the Netherlands. The data comprises 387,195 persons in the Netherlands. The individuals are spread over a domain spanning 415 Dutch municipalities and divided into 11,432 areas, which are neighborhoods. We have limited our studies to the region of Utrecht, comprising $p = 2955$ areas. The response variable is binary: it equals 1 is an individual is overweight and 0 if he is not. There are 11 explanatory variables (outside of the area): 6 qualitative variables (sex, ethnicity, marital status, household type, household income source, and home ownership) and 5 quantitative variables (age, household size, household capital, household income, and neighborhood urbanization). An additive binomial regression model is used to link $\boldsymbol{\eta} = \text{logit}(\mathbb{E}\left[\boldsymbol{y}|\boldsymbol{x}\right])$ (the prevalence of being overweight) to the covariates and to the area. The model is a generalized additive model (Wood, 2017):

$$\boldsymbol{\eta} = \boldsymbol{X}^{(1)\,T}\boldsymbol{\beta}^{(1)} + \sum_{j=1}^{p} f_j(\boldsymbol{x}_j^{(2)})\beta_j^{(2)} + \sum_{j=1}^{p} \mathbb{1}_{i\in j}\gamma_j,$$

where (f_j) are (known) spline functions, $\boldsymbol{X}^{(1)}$ is the (binary-coded) design matrix for the 6 qualitative variables, $\boldsymbol{X}^{(2)}$ is the design matrix for the spline regression for the 5 quantitative variables, and $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ are their corresponding variables. Finally, $\boldsymbol{\gamma}$ is the log-odds-ratio of the area location. Conditionally on the choice of the spline knots, the model writes

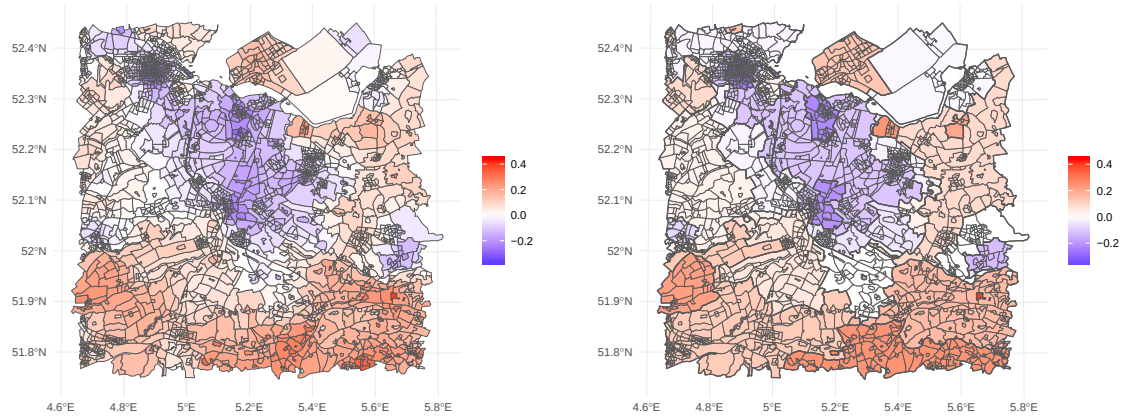
$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{p} \mathbb{1}_{i\in j}\gamma_j, \tag{5.11}$$

where $\boldsymbol{X} = (\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$ is the design matrix for all covariates except the area location. This model was fitted and I have used $\hat{\boldsymbol{\gamma}}$ and its estimated variance $\hat{\boldsymbol{\sigma}^2}$. From the theory on logistic regression, we know that

$$\hat{\sigma}_j^2 \triangleq \text{Var}\left[\hat{\gamma}_j\right] = \left(\sum_{i=1}^{n} \mathbb{1}_{i\in j}\frac{\hat{p}_i}{1-\hat{p}_i}\right)^{-1}, \tag{5.12}$$

where $\hat{p}_i$ is the estimate probability that individual i is overweight. The computation of $\hat{\boldsymbol{\sigma}}$ was made by people with access to the design matrix. From (5.12) we see that it is not possible to recover personal information from the estimated variance, as long as there are enough individuals in each area. Thus, $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\sigma}})$ can be communicated publicly without concerns of data privacy.

Remark. Define $\tilde{\boldsymbol{X}}$, the extended design matrix obtained by column-concatenation of $\boldsymbol{X}^{(1)}$ and the matrix with (i,j)-th entry $\mathbb{1}_{i\in j}$. Equation 5.11 is a linear model with design matrix $\tilde{\boldsymbol{X}}$ and parameter $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$. We could define a regularized estimate of $\boldsymbol{\gamma}$ by introducing a penalty term over $\boldsymbol{\gamma}$ in this linear model. We take another two-fold approach. We first estimate $\boldsymbol{\beta}$ from (5.11), considering $\boldsymbol{\gamma}$ as a nuisance parameter. We obtain an estimated $\hat{\boldsymbol{\gamma}}$ corresponding to the spatial effects from the latter step. Then, we perform a regularization of $\boldsymbol{\gamma}$ using the previously estimated spatial effects $\hat{\boldsymbol{\gamma}}$, its estimated variance $\hat{\boldsymbol{\sigma}}^2$, and the adjacency graph. This has some advantages. The first one is regarding the privacy of the data. The second step of spatial segmentation can be performed without having to know the potentially sensitive data present in $\boldsymbol{X}$. This methodology allows to apply the segmentation method to many studies where access to privacy-sensitive data is impossible, as is the case here.



(a) Unregularized values: $\hat{\boldsymbol{\gamma}}$

(b) Segmented values: $\hat{\boldsymbol{\theta}}$

Figure 5.7: Unregularized (a) and segmented (b) values of overweight prevalence score in the Netherlands, adapted for 11 ethnic, financial, and social covariates.

Results. The vector $\hat{\boldsymbol{\gamma}}$ is represented in Figure 5.7a. The segmented estimate $\hat{\boldsymbol{\theta}}$ is represented in Figure 5.7b. We identify 40 regions of constant overweight prevalence using the BIC. The values with a higher risk (with respect to the overall risk) of being overweight are coded in red, the ones with the lower risk of being overweight are coded in blue. Our method highlights three disconnected regions with a specifically low overweight risk: two in light blue, in the central left and central right part of the map and in the center of the map one very large light blue zone, which includes three smaller regions with a very low overweight risk. The two light blue regions correspond to urban areas around the cities of Gouda (west-most region) and Arnhem (east-most region). The big blue region consists of a densely populated area spanning from Amsterdam to Utrecht. The three dark blue regions correspond almost exactly to the cities of Amsterdam (north-west), Hilversum (center-north) and Utrecht (center-south). These three urban regions represent sites where the quality of life is better, since individuals are less likely to be overweight, even when adjusted for a series of important financial, social, and demographic variables.

The red regions span across the west, east and (mostly) south of the Amsterdam-Utrecht axis. We identify two major dark red regions: one south of the city of Gouda (west on the map) and one in the south-east, the two being rural areas. There are also a few notable informations that can be seen from Figure 5.7b. There is a (light) red area in the eastern suburbs of Amsterdam, which seems like an important message for public health policy. Moreover there are few areas that have a different overweight prevalence than their neighboring areas. Further investigation by public health specialist need be carried about these specific neighborhoods to determine if there is indeed an abnormally increase risk of overweight or if this phenomenon is an artifact due to statistical uncertainty.

To conclude, the segmented estimate obtained by our method (Figure 5.7b) is not so different from the unregularized estimate (Figure 5.7b) and has some advantages:

- It defines few zones with different risks.
- It removes the arbitrariness of having to define what regions are part of a blue zone or a red zone.
- It provides a more strikingly visual message which is easier to communicate with.

5.5 Conclusion

Many public health studies now require precise geographic information to enforce policies and take action. This faces two issues: that of the privacy issues that go with sensitive (medical) personal information and that of finding models that make sense of the sometimes very geographically precise data. These two problems are solved by using a post-treatment of the spatial effect that estimates a segmentation into constant regions. We show that using a Bayesian framework, this approach only requires data aggregated at a unit of spatial resolution, called "area" (for instance, neighborhoods). We develop an approach to perform this segmentation using the adjacency structure of the areas. We show through simulations that *(i)* it performs well to recover the initial segmentation in the case where the underlying spatial signal is piecewise constant *(ii)* it is robust to other models for instance smoothly varying spatial signals. We found no record of another statistical method performing segmentation over an adjacency graph. Consequently, we had no reference method to compare to.

We illustrated our method on the motivating example for this work: an application to the study of prevalence of overweight individuals in the Amsterdam-Utrecht area. We have used data from a pre-existing epidemiological study (van de Kassteele et al., 2017) and perform segmentation of the parameter coding for the spatial effect. Note that the personal data of individuals was not accessible in this case, and our method does not require to rerun the logistic regression of the original study. Instead, we perform segmentation of the estimated parameters.

The computation time is very sensible to the number of areas, which is the limiting factor. The method is computation-intensive due to important number of possible edges in a planar graph, which makes the adjacency matrix a computational burden. We highlight that our method scales relatively well with the number of areas in practical applications, up to $p \simeq 1000$. For that number of areas, the typical computation time is 7 minutes on a laptop equipped with an Intel Core™i3 CPU. An `R` implementation of the method is available in the package `graphseg`[3]. It makes use of sparse matrices and warm start to boost up computations.

Some work is still needed for this method to be fully validated. Future works include a quantitative analysis of the performance in selection and estimation. An extension of the BIC to take into account the specifics of the model seems to be of interest. The EBIC (for *Extended BIC* Chen and Chen, 2008) is an extension of the BIC that includes a Bayesian prior on the model to give equal importance between the models of different dimension. This does not apply directly to our graphical model. We would require to compute the number of ways to divide a (planar) graph into k connected components, for varying values of k. Work in that direction has been undertaken, and would significantly add to the performance of our method.

[3] github.com/goepp/graphseg

Chapter 6

Conclusion

The adaptive ridge and its applications

This thesis provides several applications of the adaptive ridge in age-period-cohort analysis, spline regression, and spatial epidemiology. When I started this thesis, the tentative title of my work was "Latent classes in survival analysis and heterogeneity of response to cancer treatment". The initial project was to study different approaches to tackle heterogeneity in survival analysis. My work was not initially meant to be focused on the adaptive ridge, but it evolved in this direction.

My work began with the study of the possibilities to extend the work of Bouaziz and Nuel (2017) to more general frameworks in survival analysis. Age-period-cohort analysis was our first choice to generalize the PCH model to bidimensional hazard. Age-period-cohort analysis is concerned with inference of follow-up data, where in most cases, regularization is needed. The factor models implicitly perform regularization of the log-hazard as they infer the effect of the age, period, and cohort variables. But they are too restrictive to infer the interaction between the effects of these variables. I have applied the fused adaptive ridge method to two models: the direct estimation of the log hazard (Chapter 2) and age-cohort-interaction model (Chapter 3). The former does not infer the effects of age, period, and cohort but instead performs segmentation of the log-hazard into regions of equal value. The latter provides an extension of the age-cohort model (or whatever two-variable model) and performs regularization of the interaction term.

The work on spline regression, in Chapter 3, was motivated by a discussion with G. Nuel. We thought that using a sparsity-inducing penalty on the coefficient of a spline would enforce non-relevant knots to be selected out. At that time, I knew little about P-splines, which performs a ridge penalty over the difference of the coefficients. Only when implementing the adaptive ridge on this problem did I understand that it makes an iterated use of the same penalized likelihood estimation as the P-splines. The resulting estimate is, to my knowledge, the fastest procedure to perform selection of the knots.

Finally, I have had the occasion to apply the adaptive ridge on a widely different topic: spatial epidemiology (Chapter 5). This work was motivated by the application to the study of overweight prevalence (see Section 5.4). When discussing with the author of van de Kassteele et al. (2017), I realized that segmentation of the spatial effect would enable to detect a simpler and more interpretable spatial effect of the prevalence of being overweight. The data used in the generalized additive model in van de Kassteele et al. (2017) is sensitive and I could not have access to the design matrix. Thus I developed a method that used only the estimate for each area (i.e. spatial unit) and an estimate of its variance. This method performs a fused adaptive ridge over a graph: the graph of adjacency of the areas. This work can be seen as a generalization of the fused adaptive ridge over a grid, which is used in Chapters 2 and 3. I believe the method I developed, albeit somewhat simple, can be of great interest for the epidemiologist: a segmented spatial signal gives a clear message, which is important to take decisions of public health policy.

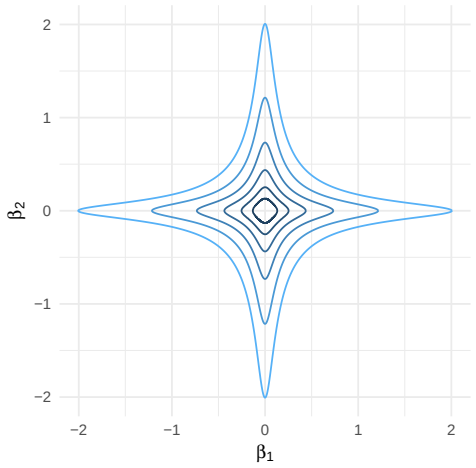


Figure 6.1: Level set of the square log penalty $\sum_{j=1}^{2} \log(\beta_j^2 + \delta^2)$ with $\delta^2 = 10^{-2}$.

Why the adaptive ridge? The adaptive ridge is just one in a wide range possible model selection methods. The adaptive ridge was introduced under different names in different contexts, but has only been introduced as an iterative penalized maximum likelihood methods in Frommlet and Nuel (2016). Contrarily to the adaptive lasso, it requires many iteration to provide a sparse estimate. It is however easier to implement, since its uses the L_2 norm penalty at each iteration, which in many cases has an explicit solution. I chose to use this method because it seemed like a good trade-off between computational cost and ease of implementation.

However, no proof has yet been given of the asymptotic properties of the adaptive ridge. Thus, an important perspective of the present thesis is the theoretical study of the adaptive ridge. The next section presents the framework for this analysis.

Perspective: the adaptive ridge and its properties

In this Thesis I have used the adaptive ridge extensively. This method was first introduced as a numerical trick by Rippe et al. (2012) and an investigation into its numerical performance was provided by Frommlet and Nuel (2016). In this section, we lay the ground for a statistical study of this method. In a first part, we demonstrate that the L_0 adaptive ridge can be equivalently defined as the LQA of a penalized estimate, as defined by Fan and Li (2001). The corresponding penalty is a function that has interesting properties. It can be seen as a relaxation of the log penalty, which was already introduced in the LLA one-step estimator by Zou and Li (2008). This framework gives good insight into the strength of the adaptive ridge: it gives up the non-concave property in exchange for a gain in computational cost. Indeed, the oracle properties of non-concave penalties (Fan and Li, 2001) do not apply for the adaptive ridge because its function has an inflexion point.

Adaptive ridge as the LQA of a penalized estimate

Define the "square-log" penalty

$$p(|\theta|) = \log(\theta^2 + \delta^2), \tag{6.1}$$

where $\delta^2 \ll 1$ is a small numerical constant. This function is not a non-concave penalty (see Section 1.1.7): it is convex on $(-\delta, \delta)$. However, this penalty converges pointwise towards the penalty $\log(\theta^2) = 2\log(|\theta|)$. Since δ is set to a very small value in practice, Equation 6.1 can be seen as an heuristic approximation of the log penalty.

The level set of the square log penalty in two dimensions is represented in figure 6.1.

Around zero, the square log penalty is a convex function and does not induce sparsity. For greater values of (β_1, β_2), the square-log penalty is concentrated around the axes. This function has no singularity, so it enforces the penalized estimate to have coordinates set to almost zero, but not equal to zero.

We will show in this section that the Local Quadratic Approximation (see Section 1.3.2.1) of the square-log penalty is the L_0 adaptive ridge. Since the square-log penalty is not a non-concave function, the LQA of the square-log penalty is not proven to have the oracle properties that non-concave penalties enjoy. But writing the L_0 adaptive lasso as the MM Optimization procedure with the square-log penalty allows to prove that the procedure is stable and to give a criterion to ensure that the procedure converges.

From the definition (1.42), the local quadratic approximating function of the square-log penalty writes

$$q(\theta|\theta^{(k)}) = \frac{\theta^2}{{\theta^{(k)}}^2 + \delta^2} + \log({\theta^{(k)}}^2 + \delta^2) - \frac{{\theta^{(k)}}^2}{{\theta^{(k)}}^2 + \delta^2}. \tag{6.2}$$

Hence the dominating function of the penalized nll writes

$$g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}) = \ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \frac{\beta_j^2}{{\beta_j^{(k)}}^2 + \delta^2} + \lambda \sum_{j=1}^{p} D(\beta_j^{(k)}), \tag{6.3}$$

where $D(\beta_j^{(k)}) = \log({\beta_j^{(k)}}^2 + \delta^2) - ({\beta_j^{(k)}}^2)/({\beta_j^{(k)}}^2 + \delta^2)$ is constant with respect to β_j. This $D(\beta_j^{(k)})$ is constant in (6.3), we can neglect this term in the minimization of $g(\boldsymbol{\beta}_j^{(k)})$.

In the minimization step of the MM optimization, we minimize $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$. The criteria for the MM procedure is $g(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}) \leq g(\boldsymbol{\beta}^{(k)}|\boldsymbol{\beta}^{(k)}) = f(\boldsymbol{\beta}^{(k)})$. In practice, we minimize g with the Newton-Raphson algorithm (or another second-order derivative-based method). These methods are unstable when the second-order derivative of $\ell(\boldsymbol{\beta})$ is too close to zero. It is therefore important to verify that the minimization step has decrease the function g. Equation 6.3 gives the following criteria for a candidate $\boldsymbol{\beta}^{(k+1)}$:

$$\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \frac{\beta_j^2}{{\beta_j^{(k)}}^2 + \delta^2} \leq \ell(\boldsymbol{\beta}^{(k)}) + \lambda \sum_{j=1}^{p} \frac{{\beta_j^{(k)}}^2}{{\beta_j^{(k)}}^2 + \delta^2}. \tag{6.4}$$

From the theory on MM optimization, if this criteria is met at each step, the procedure is stable and converges to a local minima.

This framework motivates the study of the theoretical properties of the adaptive ridge.

List of Figures

1.1 Visualization of the ridge estimate as the projection of the OLS onto an L_2 norm ball. The ellipses are level curves of the quadratic form $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2$. The projection onto the circle of radius $t=1$ has the effect of shrinking the estimate's coordinates together. . . . 6
1.2 Illustration of L_q unit balls for different values of q. . . . 8
1.3 Illustration of the lasso estimation under orthogonal design. Projection onto two L_1 norm balls of radius t. As t decreases, the OLS will most probably end up in a greyed area and the lasso estimate will become sparse. . . . 9
1.4 Illustration of why the elastic net induces sparsity and not the L_q norm ($1<q<2$). Unit radius balls of the elastic-net norm ($\alpha=0.2$) and the $L_{1.2}$ norm. . . . 13
1.5 Bayesian priors on the parameter with the Ridge, Lasso, Elastic-net ($\alpha=0.5$), and Bridge ($q=0.5$) penalties. . . . 14
1.6 Several non-concave penalties with different values of λ. . . . 15
1.7 Thresholding functions in orthogonal design. . . . 16
1.8 All possible submodels with three variables, forming a Hasse diagram of the power-set of $\{1,2,3\}$. . . . 20
1.9 Local linear approximation (a, in dashed line) and local quadratic approximation (b, in dashed line) of the bridge penalty (solid line) $p(|\theta|)=|\theta|^{0.5}$ around the current point $\theta^{(k)}=0.5$. . . . 23
1.10 L_q penalties (solid lines) and their approximations by the Adaptive Ridge (dotted lines) with $\gamma=2$ and $\varepsilon=10^{-4}$. . . . 26
1.11 Approximations of L_q penalties by the Adaptive Ridge for different values of γ, and with $q=0$ and $\varepsilon=10^{-4}$. . . . 27
1.12 Unit ball of the group lasso norm with three variables and with $\mathscr{G}=\{\{1,2\},\{3\}\}$. We represent the 3d ball (a) and its restriction to the variables $\{\boldsymbol{x}_1,\boldsymbol{x}_2\}$ (b) and $\{\boldsymbol{x}_1,\boldsymbol{x}_3\}$ (c). . . 29
1.13 Unit ball for the L_1/L_2 mixed norm with overlapping groups $\{\{1,3\},\{2,3\}\}$. . . . 30
1.14 (a) Hierarchical structure of the overlapping group $\mathscr{G}=\{\{1,2,3\},\{2,3\},\{3\}\}$, (b) Unit ball of the corresponding mixed norm, (c) Projection of the unit ball on the plane $\boldsymbol{x}_2=0$. . . 32
1.15 Admissible solution set for the fused lasso penalty (dark grey) and for the lasso and total variation penalties (light grey). . . . 33

2.1 Lexis Diagrams in the Age-Period plane (a) and in the Age-Cohort plane (b). . . . 42

3.1 True hazard ($\lambda^*_{j,k}$) for simulation design. . . . 78
3.2 Estimates of the ACI model in Simulation Design 1. (a) Maximum likelihood estimate (b) Estimated hazard ($\log\lambda_{j,k}$) in the ACI model (c) Age and cohort effects ($\mu+\alpha_j+\beta_k$) in the ACI model (d) Interaction effect ($\delta_{j,k}$) from the ACI model (e) Age effects (α_j) (f) Cohort effects (β_k). . . . 82
3.3 Estimates in Simulation Design 2 with the ACI model: (a) True hazard, (b) AC model ($\log\lambda_{j,k}=\mu+\alpha_j+\beta_k$), (c) Estimated hazard in the ACI model, and (d) Interaction effect ($\delta_{j,k}$) in the ACI model. . . . 83

3.4 Estimates in Simulation Design 2 with the ACI model: (a) True hazard, (b) AC model ($\log \lambda_{j,k} = \mu + \alpha_j + \beta_k$), (c) Estimated hazard in the ACI model, and (d) Interaction effect ($\delta_{j,k}$) in the ACI model. 85

4.1 Representation of the truncated power basis $(T_{j,k,\boldsymbol{t}}(x))_{1\leq j\leq k+q}$ for $k = 3$ and $\boldsymbol{t} = (0.5, 1, 1.5)$. 92
4.2 Illustration of the numerical instability of the decomposition in truncated power basis, with a spline of order 2 supported by 3 knots (tick line) and its decomposition with 2-digit machine precision coefficients (dotted line). 93
4.3 B-spline bases of order 1 to 4 with three equally spaced knots $\boldsymbol{t} = c(0.25, 0.5, 0.75)$. 94
4.4 Spline regression with 9 knots placed at equi-distance (c, d) and at the quantiles (a, b), for two different data sets. 96

5.1 Map of cholera outbreaks in the neighborhood of Soho (London) by John Snow (1854). Crosses are water wells, dots are cholera outbursts. 136
5.2 (a) Division of continental France into 94 *départements* of roughly uniform size. (b) Euclidean distances between the centroids of each *départements*, using the latitude and longitude as coordinates. (c) Shortest path distances between *départements*, using the adjacency graph. 138
5.3 Iris of the city of Paris used for simulation 142
5.4 Result of simulation setting 1: piecewise constant effect $\boldsymbol{\alpha}$ (a), noisy piecewise constant effect with additive Gaussian noise $\boldsymbol{\gamma}$ (b), segmented estimate using the L_0 adaptive ridge penalization $\boldsymbol{\theta}$ with the BIC (c), and with 18 regions (d). 143
5.5 Regularization path in simulation setting 1: trajectories of the estimated $(\theta_j)_j$ for different values of κ. For a better visualization, the x axis gives the number of estimated parameters instead of κ. The vertical line represents the estimate selected by the BIC. Each color represents a different area. 144
5.6 (a) Smooth spatial signal $\boldsymbol{\gamma}$ used in simulation setting 1 (b) and segmented estimate $\boldsymbol{\theta}$ using the BIC. 145
5.7 Unregularized (a) and segmented (b) values of overweight prevalence score in the Netherlands, adapted for 11 ethnic, financial, and social covariates. 147

6.1 Level set of the square log penalty $\sum_{j=1}^{2} \log(\beta_j^2 + \delta^2)$ with $\delta^2 = 10^{-2}$. 150

List of Tables

1.1 Available implementations in R of the penalization-based variable selection methods for the linear model. 19

3.1 Mean squared errors of the AC model, the ACI model, and the maximum likelihood estimate. For each sample size, the mean squared error was computed over 100 repetitions. The smallest mean squared error in each row is highlighted in bold. 80

Bibliography

Hirotugu Akaike. A New Look at the Statistical Model Identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer, 1974.

A. Antoniadis. Wavelets in statistics: A review. *Journal of the Italian Statistical Society*, 6(2):97–130, 1997.

Francis Bach. Optimization with Sparsity-Inducing Penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2011.

Sergey Bakin. *Adaptive Regression and Model Selection in Data Mining Problems*. PhD thesis, 1999.

Olivier Bouaziz and Grégory Nuel. L0 Regularization for the Estimation of Piecewise Constant Hazard Rates in Survival Analysis. *Applied Mathematics*, 08(03):377–394, 2017.

Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK ; New York, 2004.

Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232–253, 2011.

Leo Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383, 1996.

Emmanuel J. Candès, Michael B. Wakin, and Stephen P. Boyd. Enhancing Sparsity by Reweighted 1 Minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008.

B. Carstensen. Age–Period–Cohort Models for the Lexis Diagram. *Statistics in Medicine*, 26(15):3018–3045, 2007.

Antonin Chambolle and Thomas Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

R. Chartrand. Iteratively reweighted algorithms for compressive sensing. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3869–3872, 2008.

J. Chen and Z. Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Jiahua Chen and Zehua Chen. EXTENDED BIC FOR SMALL-n-LARGE-P SPARSE GLM. *Statistica Sinica*, 22(2):555–574, 2012.

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic Decomposition by Basis Pursuit. *Society for Industrial and Applied Mathmatics*, 43(1):129–159, 2001.

D. Clayton and E. Schifflers. Models for Temporal Variation in Cancer Rates. II: Age–period–cohort models. *Statistics in Medicine*, 6(4):469–481, 1987.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Noel Cressie. *Statistics for Spatial Data*. Wiley Series in Probability and Statistics, 1993.

Noel Cressie and Ngai H Chan. Spatial Modeling of Regional Variables. *Journal of the American Statistical Association*, 84(406):393–401, 1989.

Noel Cressie and Timothy R. C. Read. Spatial Data Analysis of Regional Counts. *Biometrical Journal*, 31(6):699–719, 1989.

Iain D Currie, Maria Durban, and Paul HC Eilers. Smoothing and Forecasting Mortality Rates. *Statistical Modelling: An International Journal*, 4(4):279–298, 2004.

H.B. Curry and I.J Schoenberg. On Pòlya Frequency Functions IV: The Fundamental Spline Functions and Their Limits. *Journal d'Analyse Mathematique*, 17(1):71–107, 1966.

Linlin Dai, Kani Chen, Zhihua Sun, Zhenqiu Liu, and Gang Li. Broken adaptive ridge regression and its asymptotic properties. *Journal of Multivariate Analysis*, 168:334–351, 2018.

Carl de Boor. Package for Calculating with B-Splines. *SIAM Journal on Numerical Analysis*, 14(3): 441–472, 1977.

Carl De Boor. *A Practical Guide to Splines*, volume 27. Springer-Verlag New York, 1978.

Carl de Boor. Least Squares Cubic Spline Approximation, II - Variable Knots. Technical Report CSD TR 21, Purdue University, Lafayette, IN, 1986.

Camila P. E. de Souza and Nancy E. Heckman. Switching Nonparametric Regression Models and the Motorcycle Data revisited. *arXiv preprint arXiv:1305.2227*, 2013.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Paul Dierckx. *Curve and Surface Fitting with Splines*. Oxford University Press, 1995.

D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

Bradley Efron, Trevor Hastie, Iain Johnston, and Robert Tibshirani. Least Angle Regression. page 46, 2004.

Paul H. C. Eilers and Brian D. Marx. Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11(2):89–102, 1996.

Paul H C Eilers, Brian D Marx, and Maria Durbán. Twenty Years of P-splines. *Statistics and Operations Research Transactions*, 39(2):149–186, 2015.

Michael Elad and Alfred M Bruckstein. A Generalized Uncertainty Principle and Sparse Representation in Pairs of R^N Bases. page 20, 2001.

Paul Elliott and Daniel Wartenberg. Spatial Epidemiology: Current Approaches and Future Challenges. *Environmental Health Perspectives*, 112(9):998–1006, 2004.

A Fan. Comments on "Wavelets in Statistics: A Review". *Journal of the Italian Statistical Society*, 6(2): 131–138, 1997.

Jianqing Fan and Runze Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Thomas R. Fleming and David P. Harrington. *Counting Processes and Survival Analysis*, volume 169. John Wiley \& Sons, 2011.

Ildiko E Frank and Jerome H Friedman. A Statistical View of Some Chemometrics Regression Tools. 35(2):28, 1993.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, 9(3):432–441, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv:1001.0736 [math, stat]*, 2010a.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010b.

Florian Frommlet and Grégory Nuel. An Adaptive Ridge Procedure for L0 Regularization. *PLoS ONE*, 11(2):e0148620, 2016.

Wenjiang J Fu. Penalized Regressions: The Bridge Versus the Lasso. page 20, 1998.

George M. Furnival and Robert W. Wilson. Regressions by Leaps and Bounds. *Technometrics*, 16(4): 499, 1974.

Vivien Goepp, Jean-Christophe Thalabard, Grégory Nuel, and Olivier Bouaziz. Regularized Bidimensional Estimation of the Hazard Rate. *arXiv:1803.04853 [math, stat]*, 2018.

P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models*. Number 58 in Monographs on Statistics and Applied Probability. Chapman & Hall /CRC, Boca Raton, Fla., 1., crc press repr edition, 2000. OCLC: 248205275.

DJ Hand. Branch and bound in statistical data analysis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 30(1):1–13, 1981.

Wolfgang Härdle. *Applied Nonparametric Regression*. Number 19. Cambridge university press, 1990.

Trevor Hastie, Jerome Friedman, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, 2nd edition, 2001.

Arthur E Hoerl and Robert W Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970.

Jian Huang, Joel L. Horowitz, and Shuangge Ma. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*, 36(2):587–613, 2008.

Junzhou Huang, Tong Zhang, and Dimitris Metaxas. Learning with structured sparsity. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, Montreal, Quebec, Canada, 2009. ACM Press.

Peter J Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

David R. Hunter and Runze Li. Variable selection using MM algorithms. *The Annals of Statistics*, 33 (4):1617–1642, 2005.

Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group Lasso with Overlap and Graph Lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, Montreal, Quebec, Canada, 2009. ACM Press.

Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured Variable Selection with Sparsity-Inducing Norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.

Bei Jiang and Keumhee C. Carriere. Age-period-cohort models using smoothing splines: A generalized additive model approach: B. JIANG AND K. C. CARRIERE. *Statistics in Medicine*, 33(4):595–606, 2014.

David L. B. Jupp. Approximation to Data by Splines with Free Knots. *SIAM Journal on Numerical Analysis*, 15(2):328–343, 1978.

Vassilis Kalofolias. How to learn a graph from smooth signals. page 13, 2016.

Dongshin Kim, Sangin Lee, and Sunghoon Kwon. A unified algorithm for the non-convex penalized estimation: The ncpen package. *arXiv:1811.05061 [stat]*, 2018.

Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5): 1356–1378, 2000.

D. Kuang, B. Nielsen, and J. P. Nielsen. Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika*, 95(4):979–986, 2008.

Kenneth Lange. *Optimization*. Springer Texts in Statistics. Springer, New York, 2004.

Andrew Lawson. *Statistical Methods in Spatial Epidemiology*. Wiley Series in Probability and Statistics. Wiley, Chichester, England ; Hoboken, NJ, 2nd ed edition, 2006. OCLC: ocm62804712.

Andrew B Lawson, Sudipto Banerjee, Robert P Haining, and Maria Dolores Ugarte. Handbook of Spatial Epidemiology. *CRC Press*, page 704, 2016.

Sangin Lee, Sunghoon Kwon, and Yongdai Kim. A modified local quadratic approximation algorithm for penalized optimization problems. *Computational Statistics & Data Analysis*, 94:275–286, 2016.

Yi Lin and Hao Helen Zhang. Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models. *The Annals of Statistics*, 34(5):2272–2297, 2006.

E. Lund, V. Dumeaux, T. Braaten, A. Hjartaker, D. Engeset, G. Skeie, and M. Kumle. Cohort Profile: The Norwegian Women and Cancer Study–NOWAC–Kvinner og kreft. *International Journal of Epidemiology*, 37(1):36–41, 2008.

Julien Mairal, Rodolphe Jenatton, Guillaume Obozinski, and Francis Bach. Convex and Network Flow Optimization for Structured Sparsity. *Journal of Machine Learning Research*, 12:2681–2720, 2011.

S. G. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Elsevier/Academic Press, Amsterdam ; Boston, 3rd ed edition, 2009.

Ricardo A. Maronna, Douglas Martin, and Víctor J. Yohai. *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. Wiley, Chichester, reprinted with corr edition, 2006. OCLC: 845656286.

Peter McCullagh. Generalized Linear Models. *European Journal of Operational Research*, 16(3):285–292, 1984.

Geoffrey McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*, volume 382. John Wiley & Sons, 2007.

Allan Miller. *Subset Selection in Regression*. Monographs on Statistics and Applied Probability 95. Chapman & Hall/CRC, 2002.

Thomas Lumley based on Fortran code by Alan Miller. *Leaps: Regression Subset Selection*. 2017.

Bojan Mohar. Some applications of Laplace eigenvalues of graphs. In Geňa Hahn and Gert Sabidussi, editors, *Graph Symmetry*, pages 225–275. Springer Netherlands, Dordrecht, 1997.

B. K. Natarajan. Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, 24 (2):227–234, 1995.

Finbarr O'Sullivan. A Statistical Perspective on Ill-Posed Inverse Problems. *Statistical Science*, 1(4): 502–518, 1986.

Art B. Owen. A robust hybrid of lasso and ridge regression. In Joseph Stephen Verducci, Xiaotong Shen, and John Lafferty, editors, *Contemporary Mathematics*, volume 443, pages 59–71. American Mathematical Society, Providence, Rhode Island, 2006.

Michel Poulain, Giovanni Mario Pes, Claude Grasland, Ciriaco Carru, Luigi Ferrucci, Giovannella Baggio, Claudio Franceschi, and Luca Deiana. Identification of a geographic area characterized by extreme longevity in the Sardinia island: The AKEA study. *Experimental Gerontology*, 39(9): 1423–1429, 2004.

M. Rezaeian, G. Dunn, S. St Leger, and L. Appleby. Geographical epidemiology, spatial analysis and geographical information systems: A multidisciplinary glossary. *Journal of Epidemiology & Community Health*, 61(2):98–102, 2007.

Ralph C. A. Rippe, Jacqueline J. Meulman, and Paul H. C. Eilers. Visualization of Genomic Changes by Segmented Smoothing Using an L0 Penalty. *PLoS ONE*, 7(6):e38230, 2012.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a Regularized Path to a Maximum Margin Classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

Volker Roth and Bernd Fischer. The Group-Lasso for Generalized Linear Models: Uniqueness of Solutions and Efficient Algorithms. *Proceedings of the 25th international conference on Machine learning*, pages 848–855, 2008.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992a.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992b.

David Ruppert. Selecting the Number of Knots for Penalized Splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757, 2002.

Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

B. W. Silverman. Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting. *Journal of the Royal Statistical Society, Series B*, 47:1–52, 1985.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Paul Tseng. Coordinate Ascent for Maximizing Nondifferentiable Concave Functions. LIDS-P 1840, Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 1988.

Berwin A Turlach, William N Venables, and Stephen J Wright. Simultaneous Variable Selection. *Technometrics*, 47(3):349–363, 2005.

Jan van de Kassteele, Laurens Zwakhals, Oscar Breugelmans, Caroline Ameling, and Carolien van den Brink. Estimating the prevalence of 26 health-related indicators at neighbourhood level in the Netherlands using structured additive regression. *International Journal of Health Geographics*, 16 (1), 2017.

C. R. Vogel and M. E. Oman. Iterative Methods for Total Variation Denoising. *SIAM Journal on Scientific Computing*, 17(1):227–238, 1996.

Garrett Wahba. *Spline Models for Observational Data*, volume 59. Society for Industrial and Applied Mathematics, 1990.

M. P. Wand and M. C. Jones. *Kernel Smoothing*. Monographs on Statistics and Applied Probability. Chapman & hall edition, 1995.

M. P. Wand and J. T. Ormerod. On Semiparametric Regression with O'Sullivan Penalised Splines. *Australian & New Zealand Journal of Statistics*, 52(2):239–239, 2010.

Simon N Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition, 2017.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Fan Zhang and Edwin R. Hancock. Graph spectral image smoothing using the heat kernel. *Pattern Recognition*, 41(11):3328–3342, 2008.

Fuzhen Zhang, editor. *The Schur Complement and Its Applications*. Number 4 in Numerical Methods and Algorithms. Springer, New York, 2005.

Peng Zhao, Guilherme Rocha, and Bin Yu. Grouped and Hierarchical Model Selection through Composite Absolute Penalties. Technical Report 703, Department of Statistics, UC Berkeley, 2006.

Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.

Hui Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533, 2008.