UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXI

Some developments in semiparametric and cross-classified multilevel models

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Francesco Pauli

Co-supervisore: Prof. Nicola Torelli

Dottoranda: Gioia Di Credico

30 November 2018

Abstract

Our work has developed from a real epidemiological problem. The carcinogenic effect of cigarette smoking on head and neck cancer has been extensively studied in the literature highlighting a non linear dose-response relationship. Recently, the use of linear regression splines, within semiparametric models framework, has allowed an improvement in the evaluation of the association between smoking habits and head and neck cancer. Our work focuses on the development of a methodology able to improve several aspects of the estimation of the aforementioned relationship. In particular, the approximation of the spline function, represented by truncated linear basis, has been refined by addressing the problem of estimating two key quantities for the definition of a spline function: the number and position of the knots. The proposed methodology uses a Bayesian approach. We then focused on developing a streamlined methodology applicable to generalised linear models for cross-classified data. In particular, the steps necessary to calculate the covariance matrix are optimised with respect to one of the two random effects, allowing a computational gain both in terms of time and memory usage. The proposed algorithms are applied in the context of the inferential variational methods in detail to the mean field variational Bayes.

Sommario

Il nostro lavoro si è sviluppato a partire da un problema epidemiologico reale. L'effetto carcinogenico del fumo di sigaretta sui tumori testa-collo è stato ampiamente studiato in letteratura evidenziando una relazione dose-risposta non lineare. Recentemente, l'utilizzo di spline lineari di regressione nell'ambito di modelli semiparametrici, ha permesso un miglioramento nella valutazione dell'associazione tra fumo e tumori testa-collo. Il nostro lavoro si concentra sullo sviluppo di una metodologia capace di migliorare la stima della suddetta relazione sotto diversi aspetti. In particolare, l'approssimazione della funzione spline, rappresentata attraverso basi lineari troncate, è stata affinata affrontando il problema di stima di due quantità chiave per sua la definizione: il numero e la posizione dei nodi. La metodologia proposta si serve di un approccio Bayesiano. Successivamente ci siamo concentrati sullo sviluppo di una metodologia streamline applicabile a modelli lineari generalizzati per dati con struttura cross-classified. In particolare, gli step necessari al calcolo della matrice di covarianza vengono ottimizzati rispetto ad uno dei due effetti random permettendo un guadagno computazionale sia in termini di tempo che di utilizzo della memoria. Gli algoritmi proposti vengono applicati nell'ambito dei metodi variazionali inferenziali nel dettaglio al mean field variational Bayes.

Ai sogni senza tempo,
alle impressioni di un momento

Acknowledgements

The three years of my PhD passed quickly and have been full of professional and personal experiences. Both offered me the opportunity to face with new and interesting points of view on my research. I would like to thank my supervisors professors Francesco Pauli and Nicola Torelli extraordinary men for their brilliance, optimism and support especially in my hardest moments. Luigino Dal Maso and Jerry Polesel, for their essential contribution and motivational presence. Valeria Edefonti, strong and tireless mind and wonderful mother. My brilliant mentor during my visiting period in Australia: professor Matt Wand. All the friendly colleagues I met in Sydney, I tasted the limitless cultural variety of the world, all thanks to you. The professors Monica Chiogna and Nicola Sartori, irreplaceable and always available presences. Patrizia Piacentini for her excellent work and her endless patience. My colleagues of the XXXI and XXX cycle, with whom I shared moments of great empathy. Claudia, Lucia, Leonardo, Daniela e Chiara, colleagues but mostly friends. Anna and Luca, precious friends, thanks for all the good time we have spent together. Above all, thanks to Federico, to my grandparents and to my beloved extended family. Even if I am far from home, I felt all your love and support. Of course, I can not forget my two purring and travelling cats: Fuliggine and Timao. Every person I met during this journey has contributed to this work but it would not be possible to list them one by one. I thank you all from the depth of my heart.

Contents

List of Figures

List of Tables

Introduction

Overview

The thesis is organised in three main sections. The first part of the work develops starting from an epidemiological issue, that is, to handle piecewise linear relationship between the response and one or more exposures and to detect possible change points.

Using data from the International Head and Neck Cancer Epidemiology consortium (INHANCE, 2004), we estimated a logistic regression model to study the association between the outcome, head and neck cancer, and two risk factors, duration and intensity of cigarette smoking (Dal Maso *et al.*, 2016). The dataset collects 35 case-control studies conducted in several Countries for a total of almost 26,000 cases and 38,000 controls.

Head and neck cancers include pharynx, oral cavity and larynx sites. We perform a stratified analysis separating current and former smokers and also dividing larynx from pharynx and oral cavity sites because of the different associative pattern of each stratum. Moreover the analysis is stratified by alcohol intensity consumption, measured in drinks per day, in current smokers and pharynx and oral cavity sites. Confounders, such as age, sex and socio-economic factors, need to be included in the model since controls are not matched to the cases in our studies. The association measure is the adjusted odds ratio in which the effect of risk factors is net of confounders (Hosmer Jr and Lemeshow, 2005).

Association between head and neck cancers and cigarette smoking has been deeply studied in the epidemiological literature, highlighting evidence in favour of departures from linearity (Lubin *et al.*, 2009). The most common approach is to include exposures as categorical variables. This choice adds flexibility but may lead to a loss of information and efficiency. It may also cause biased results because of the implied assumption of constant risk within each category.

The two risk factors, measured in year of cigarette consumption and in daily average number of cigarettes smoked, are modelled through a bivariate regression spline function, while confounders enter linearly in the logistic model (Dal Maso *et al.*, 2016). This choice

allows to explore the different effect of the two exposures on the outcome. Truncated linear basis represents the bivariate regression spline offering a direct interpretation of the knot locations as change points in slope of the risk surface. Splines are highly flexible, in fact, varying the number and position of knots may lead to extremely different shapes and a major risk is to overfit the data (Ruppert *et al.*, 2003).

The first approach used is to fix the number of knots between 1 and 3, according to biological reasons. Standard approaches to choose the location of knots that have been tested are: knots on the quantiles of the predictors distribution, uniformly distributed knots on the range of the independent variables and user defined knots following a priori information. Since fitted models are non-nested, they have been compared by AIC. However, these comparisons do not lead to clear cut conclusions, the competing models lead to similar values of the criteria probably due to the roughness of the objective function which leads to several local maxima. Hence, we decided to consider knots positions as unknown parameters.

This choice turned the problem into a non linear optimisation problem. For a fixed number of knots, the optimal knot locations and regression parameters were jointly estimated within the Bayesian approach, with prior distributions expressing plausible values of knot locations and regression parameters (Carpenter *et al.*, 2017). The best model is chosen as the one that minimises the Watanabe-Akaike Information Criterion (WAIC) (Gelman *et al.*, 2014).

The second section of our work addresses the issue of estimating the number of knots in a semiparametric regression model when univariate regression splines are used to describe one of the predictors (Ruppert *et al.*, 2003). The aim is to avoid to fit of a high number of models to choose from using information criteria. Indeed, estimate several models with many regression parameters and location of knots unknown may be computationally intensive. Two main classes of methodologies can be found in the statistical literature to face the issue. The first applies variable selection procedures to choose from a fixed set of knots (Smith and Kohn, 1996). The second class includes trans-dimensional methods employing samplers that allow for varying dimension of the parameter space (Denison *et al.*, 1998; DiMatteo *et al.*, 2001).

The proposed methodology is characterised by a two step procedure. In the first step we select the optimal number of knots considering a large, possibly, overparameterised model. In the second step we fit the final model condition on the number of knots by simultaneously estimating locations of knots and regression and spline coefficients. The concept underlying the proposed methodology is to perform variable selection on the basis functions, for this purpose we employ one of the most common approaches in

Bayesian literature: that based on the definition of spike-and-slab priors (O'Hara *et al.*, 2009). We compare the well-known stochastic search variable selection methodology (SSVS) with our proposal, based on a modification of the SSVS. Our method adapts the SSVS approach by assuming the mixing proportion parameters to be dependent on the knot locations. In this way, if there is no evidence in favour of the presence of a knot, the mixing proportion parameter linked to that knot will have a posterior distribution highly concentrated near zero. The methodology can be easily extended to any generalised linear model (McCullagh and Nelder, 1983).

The last section concentrates on the estimation of Gaussian longitudinal/multilevel models with the streamlined variational inference. In the literature this technique has been applied on nested data structure, while we explore the crossed-classified one (Lee and Wand, 2016a). The approach heads towards the improvement of the inversion of the matrix needed to compute the covariance matrix of the regression and the of the crossed random effects parameters. In the nested case the structure of the covariance matrix is characterised by sparsity which allows to simplify computations taking into account the inversion only of the sub-blocks of interest (Nolan and Wand, 2018).

Dealing with crossed random effects, the structure of the inverse of the covariance matrix is less sparse because of dependencies among sub-blocks of interest that cannot be simplified as in the nested case. The key assumption to the streamlined inference results in the crossed random effects model, is to keep the number of groups for one random effect relatively low, while the other number of groups can be extremely large. This is the case, e.g., when we are dealing with a questionnaire with 10 or 20 items submitted to thousands of people. Clearly, for a low number of items, the higher the number of subjects, the more we gain in terms of computational complexity.

Main contributions of the thesis

Our work supplies workable solutions to the three issues faced during the research period. In particular, for a fixed number of knots, we developed a Bayesian methodology to estimate the knot locations in a semiparametric logistic model using bivariate linear regression splines. Advantage of the methodology is mainly its ability to obtain a direct estimate of the location of the knots jointly with the spline parameters taking into account also the effects of all the other confounders. The methodology confirms results from the published literature on the association between cigarette smoking and head and neck cancer. It also helps to shed light on the intense epidemiological debate about the use of joint risk factors instead of cumulative one (Peto, 2012).

In order to reduce the computational effort of the method proposed in the first part of our work, we introduce a new methodology in two steps to estimate both the number and position of the knots in univariate regression splines with truncated linear basis. Even if more investigations are needed to better evaluate and extend our method, it gives better results in terms of estimation of the parameters and in terms of convergence of the algorithm when compared to the SSVS approach.

Lastly, the streamlined inference for Gaussian crossed random effects models turns out to be a faster but accurate alternative to the MCMC methods. Our results allow to compute sub-blocks of interest of covariance matrix of the parameters improving the efficiency of the algorithm. In detail, solution to the least squares problem gives us the possibility to apply the fast and stable QR decomposition on small sub-block of interest, restricting also the amount of storage memory needed for the matrix inversion.

Chapter 1

Semiparametric models for the effects of smoking-intensity and duration on H&N cancer: multicenter study

1.1 Head and neck cancer

It is well known that the main risk factor for the head and neck malignant pathologies is the tobacco use, especially in the form of cigarette smoking. The causal relationship between this exposure and the head and neck cancers (HNC) has been well explored in the literature (IARC, 1986). Cigarette smoking is a multidimensional phenomenon that can be described from several perspectives (Leffondré *et al.*, 2002). Besides the classification in never, former and current smoker, smoking behaviours in surveys can be described more accurately by quantitative aspects, e.g. number of cigarettes smoked per day, age at start cigarette smoking, duration of smoking or years since quit cigarette smoking.

A vivid epidemiological debate concentrates on the measure to use in the evaluation of the dose-response effect on the risk of HNC. When both intensity and duration of smoking are available, cumulative measure is the most common choice in the epidemiological cancer studies (Lubin *et al.*, 2007; Lubin and Caporaso, 2013). This measure is computed as average daily intensity times the duration of exposure and it is based on strong assumptions on the impact of the two risk factors. Moreover, assumptions of linearity underneath the use of cumulative measures in epidemiological studies are extensively described by Smith (1992).

The association between tobacco smoking and HNC risk has been evaluated by considering duration and intensity as either separate or interacting predictors, e.g. cross-product or pack-years (Lubin *et al.*, 2010; Hashibe *et al.*, 2007; Lubin and Caporaso, 2013). Pack-years is the most used continuous metric that synthesises information on quantity and duration of exposure. In this case, the risk factor is summarised by a lifetime cumulative measure which expresses the number of cigarette packs a subject smoked during his life. The use of pack-year measure implies that the interacting risk factors have the same impact (Leffondré *et al.*, 2002). Namely, the risk for a subject who smokes 10 cigarettes every day for 20 years is assumed to be the same as a subject who smokes 40 cigarettes per day for 5 years (Peto, 2012).

However, some authors criticise this choice maintaining that this measure does not aptly fit well-known biological results and that smoking behaviour are not well characterised. Indeed, for a fixed pack-year value, the risk remarkably varies with the exposure duration (Leffondré *et al.*, 2002; Peto, 2012).

Modelling the association between the outcome and two continuous and interacting risk factors increases the complexity of the model but also adds flexibility relaxing the restraining assumptions behind pack-years measure. Recently, bivariate spline models have been proposed in the epidemiological literature to estimate the association between HNC risk and alcohol consumption and cigarettes smoking (Dal Maso *et al.*, 2016).

Indeed,

1. the two exposures are allowed to determine disease risk in their separate or interacting role;

2. non-linearity in the dose-response relationship can be modelled.

In the following sections, we propose to re-evaluate the joint effects of intensity and duration of cigarette smoking on HNC risk using bivariate spline models. After the definition of the model, we describe frequentist and Bayesian approaches to evaluate the presence of departures from linearity in the dose-response relationship. We conclude the chapter with the analysis of HNC data from the International Head and Neck Cancer Epidemiology (INHANCE) consortium (INHANCE, 2004).

1.2 Model

In case-control studies, the association between head and neck neoplasms and tobacco consumption is generally estimated using the multiple logistic regression model, a specific type of generalised linear model (GLM). The association measure is usually

the adjusted odds ratio (OR) in which the effect of risk factor(s) is net of confounding. Confounders are defined as variables associated both with the response variable and the risk factor. Adjusted analysis prevents from apparent associations between the outcome and the exposures taking into account differences in the baseline characteristics of the sample (Rothman *et al.*, 2008; Pearl, 2009).

The model can be fitted with continuous or categorical exposure. The former specification assumes a linear relation between the predictor and conditional log odds which is not always biological plausible, especially for high exposure levels. The latter specification adds flexibility but may lead to a loss of information due to the categorisation of variables. Moreover, it may cause a loss of efficiency and bias in results because of the implied assumption of constant risk within each category (Dal Maso *et al.*, 2016). Different categorisation choices may lead to different results and the number of parameters that have to be estimated increases with the number of cross-product categories requiring also larger dataset.

Lately, a more flexible method based on the use of spline functions has been introduced in the epidemiological literature (Polesel *et al.*, 2005, 2008; Gasparrini *et al.*, 2010, 2017; Dal Maso *et al.*, 2016).

In Gasparrini *et al.* (2010) the objective is to model non-linear effects between exposure and response using univariate (Polesel *et al.*, 2005, 2008) or bivariate (Gasparrini *et al.*, 2010, 2017; Dal Maso *et al.*, 2016) splines with fixed degree. When no penalisation is used, the choice of the number of knots, is based on the minimization of an information criterion (Polesel *et al.*, 2005, 2008; Gasparrini *et al.*, 2010; Dal Maso *et al.*, 2016). The main difference with the proposed methodology concerns the choice of the knots location. Indeed, in our case, it is not fixed but a parameter to be estimated and, in particular, the setting of our methodology allows us to interpret each knot as a change of slope in the estimated risk surface.

Starting from the definition of a semiparametric generalised linear model (Eq. 2.1), we specify a semiparametric logistic model for the Bernoulli-distributed random variable $\mathbf{Y}$, that is

$$\text{logit}\left(\pi_i\right) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{z}_i\boldsymbol{\alpha} + f\left(x_i, w_i\right), \quad \text{for } i = 1, \ldots, n, \tag{1.1}$$

where the conditional mean $\pi = Pr\left(\mathbf{Y} = 1|\mathbf{Z}\right)$ expresses the probability of success and $\text{logit} : (0,1) \to \mathbb{R}$, is the canonical link function which maps probabilities onto the real line. Predictors $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_p)$ are parametrically estimated adjustment variables, $\mathbf{X}$ and $\mathbf{W}$ are continuous risk factors, evaluated through an arbitrary smooth function $f : \mathbb{R}^2 \to \mathbb{R}$.

Among all the numerous possible specifications of smooth function, we choose to define the function f as a regression spline using truncated linear basis representation.

Criteria for choosing take account of epidemiological reasons. Indeed, from a biological point of view, the number of changes in the risk pattern is generally low. Exposure may have a protective effect, usually at low level, e.g. moderate alcohol consumption has been proven to have a protective effect on several diseases. Saturation or threshold effect usually occurs at high doses of the exposure. Both these effects affect the dose-response curve, for a single exposure, or surface, for two exposures, producing changes on the slope.

Knots of the spline function, represented by truncated linear basis, can be interpreted as thresholds in the risk pattern. Therefore both number and position of knots have an important and meaningful interpretation in the model and truncated linear basis offer a direct interpretation of the parameters. Moreover, keeping a low number of knots restricts the well-known weaknesses of the truncated power bases (Ruppert *et al.*, 2003). Thus, the two exposures are modelled by a joined piecewise polynomial of a linear degree, with constraints for continuity at each join point.

The bivariate truncated linear basis is computed as the tensor product of two univariate truncated linear bases (Eq. 2.2), that is

$$\begin{aligned} f(\mathbf{x},\mathbf{w}) =& \beta_0 + \beta_1\mathbf{x} + \beta_2\mathbf{w} + \beta_3\mathbf{x}\mathbf{w} + \\ & \sum_{k_x=1}^{K_x} \gamma_{k_x}(\mathbf{x}-\xi_{k_x})_+ + \sum_{k_w=1}^{K_w} \gamma_{k_w}(\mathbf{w}-\xi_{k_w})_+ + \\ & \sum_{k_x=1}^{K_x} \gamma_{2,k_x}(\mathbf{x}-\xi_{k_x})_+\mathbf{w} + \sum_{k_w=1}^{K_w} \gamma_{2,k_w}(\mathbf{w}-\xi_{k_w})_+\mathbf{x} + \\ & \sum_{k_x=1}^{K_x}\sum_{k_w=1}^{K_w} \gamma_{3,k_x,k_w}(\mathbf{x}-\xi_{k_x})_+(\mathbf{w}-\xi_{k_w})_+, \end{aligned} \tag{1.2}$$

where ξ_{k_x} and ξ_{k_w} are the positions of the k_x-th and k_w-th knot and K_x and K_w are the total numbers of knots.

The likelihood function is defined as

$$L(\boldsymbol{\theta}|\boldsymbol{\xi},\mathbf{y},\mathbf{Z}) = \prod_{i=1}^{n} \pi(\mathbf{z}_i,x_i,w_i,\boldsymbol{\xi})^{y_i}(1-\pi(\mathbf{z}_i,x_i,w_i,\boldsymbol{\xi}))^{(1-y_i)},$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})$ is the vector of parameters that has to be estimated, and

$$\pi(\mathbf{z}_i,x_i,w_i,\boldsymbol{\xi}) = \frac{e^{\mathbf{z}_i\boldsymbol{\alpha}+f(x_i,w_i)}}{1+e^{\mathbf{z}_i\boldsymbol{\alpha}+f(x_i,w_i)}}$$

is the inverse-logit or logistic function.

Given number and position of knots parameters estimation reduces to the maximum likelihood estimate (MLE) that is numerically computed using the Newton-Raphson iterative method, since the score equations are not linear in the parameters and there is no close form solution (Hastie and Tibshirani, 1990).

The first tested approach is to fix the number of knots between 1 and 3. Then choose the knots position using standard criteria, such as quantiles of the predictors distribution, uniformly distributed knots on the range of the independent variables and user defined knots following a priori information.

Since fitted models are non-nested, they have been compared by the widely known Akaike information criterion (AIC) (Akaike, 1974), but a clear discrimination among all the fitted models was not evident. In particular, models with the same number of knots but different knot locations have very similar scores. Moreover, the number of available competitive models increases with wider sets of epidemiologically meaningful locations, especially when bivariate splines are considered. Fitting all available competitive models may easily become computationally unfeasible, unless one recurs to approximation techniques, such as variational Bayesian methods. However, this would not solve model selection difficulties in determining both the number and location of knots (Rosenberg *et al.*, 2003).

1.3 Free-knots splines

Another approach is to consider the knot location as a parameter that has to be estimated. This would lead us to improve the results in terms of spline adaptation but the knot location estimation problem is a non linear optimisation problem. Various methodologies proposed in the literature start from a defined set of knots trying to find the best subset of knots performing variable selection (Sect. 2.1).

Given the number of knots, we want to simultaneously estimate regression coefficients, spline coefficients, and knot locations.

1.3.1 Frequentist approach

In the first free-knots methodology applied we adapt the one proposed by Mao and Zhao (2003) to our semiparametric logistic model. The authors choose the B-spline basis representation with cubic degree and they estimate both regression parameters and knot location using a Newton-Raphson iterative algorithm. In our case, MLE of the parameters vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})$ can be found maximising the log likelihood

function

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \{y_i \log \pi(\mathbf{z}_i, x_i, w_i) + (1 - y_i) \log(1 - \pi(\mathbf{z}_i, x_i, w_i))\}.$$

Dealing with truncated linear basis representation, the choice of the optimisation algorithm falls on Nelder-Mead derivative-free optimisation methods since the continuous derivative requirement for the Newton-Raphson techniques is not satisfied.

The model selection step proposed by the authors employs for the selection of the optimal number of knots, a modified generalised cross-validation statistic (Mao and Zhao, 2003). In our case it can be adapted as

$$\mathrm{GCV}(r) = \frac{-2l(\hat{\boldsymbol{\theta}})n}{(n-r)^2},$$

where $r = (4 + 3(K_x + K_w) + (K_x K_w) + p)$ is the number of relevant parameters in the model and $-2l(\hat{\boldsymbol{\theta}})$ represents the deviance. The procedure has been tested on the univariate and bivariate spline case. Initialisation values have been chosen both at random and as equally spaced on the range of the risk factors. In both cases, optimised knots positions were extremely close to the initialisation values. Similarly to the fixed knots approach, AIC and GCV criteria turn out to have similar values. As the authors pointed out, we observed that the likelihood surface has several local maxima and this leads to apparent solutions strongly conditioned on starting values (Mao and Zhao, 2003). Moreover, performance deterioration in unconstrained Nelder-Mead simplex algorithms is well known as the dimension of the parameter space increases (Han and Neumann, 2006).

Another free-knots approach tested has been the bounded optimal knots. The paper from where we started is the one by Molinari *et al.* (2004). This method, as the previous one, starts from a fixed number of knots and estimates regression coefficients along with the knots location subject to constraints. In particular,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}, \\ \mathbf{l} \leq \boldsymbol{\xi} \leq \mathbf{u}}} \sum_{i=1}^{n} \{y_i \log \pi(\mathbf{z}_i, x_i, w_i) + (1 - y_i) \log(1 - \pi(\mathbf{z}_i, x_i, w_i))\},$$

where $\mathbf{l}$ and $\mathbf{u}$ are the lower and upper bound of the intervals, called windows, on which bounded maximisation for the knots position has to be computed. The authors describe the algorithm to construct the windows such that coalescent knots, namely replicated knots, and "lethargy" property are prevented. "Lethargy" property concerns poor convergence of the algorithm near the boundaries (Jupp, 1975). According to the

authors, this problem can not be solved using few knots.

Definition of windows is influenced by two user-defined parameters: ϵ and ρ. The former is the minimum distance between two knots and between the extremes and the smallest and biggest knots. The latter governs the minimum allowed difference between intervals, to limit the number of candidates (Molinari *et al.*, 2004).

For a fixed number of knots between 1 and 3, models have been fitted and compared through AIC and GCV. Even if some knot location estimates concentrate around biologically relevant values, estimated models showed to suffer from "lethargy" problem. Moreover, increasing the number of parameters, leads to unsatisfying results. As a matter of fact, knots position estimates converge at few decimal digits from the initial values when confounders are taken into account.

R functions tested are `optim` (R Core Team, 2018), `BBoptim` (Varadhan and Gilbert, 2009), and `nmkb` (Varadhan *et al.*, 2018).

1.3.2 Bayesian approach

Unlike frequentist inference, the Bayesian approach assumes the parameters of the model as random quantities, described through prior distributions which specify the *a priori* available information. Within Bayesian methodology, the model synthesises prior information and evidence from the data into the posterior distribution of the parameters. Bayesian inference is based on simulations from the posterior distribution that generate an empirical distribution of the parameters.

The Bayesian approach allows us to estimate knots locations starting from the overall set of confounding variables, jointly estimate knots locations and regression parameters and easily formalise constraints on the knots location through the definition of suitable prior distributions.

Bayesian inference computes posterior distribution according to the Bayes' theorem resulting in updating information about the parameters through evidence from the observed data.

$$\pi(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = \frac{L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X})\pi(\boldsymbol{\theta})}{\int_\Theta L(\mathbf{u}|\mathbf{y},\mathbf{X})\pi(\mathbf{u})\mathrm{d}\mathbf{u}},$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})$ is the parameter vector and π represents the product of the prior distributions.

In detail, we chose the prior distributions to express vague knowledge of plausible values of the parameters. For the knots locations, we assumed uniformly distributed

priors on the range of the linked risk factor

$$\xi_{k_x} \sim Unif(\min(\mathbf{x}), \max(\mathbf{x})), \quad \text{subject to} \quad \xi_{k_x} \leq \xi_{k_x+1}, \quad \text{for } k_x = 1, \ldots, K_x,$$

and

$$\xi_{k_w} \sim Unif(\min(\mathbf{w}), \max(\mathbf{w})), \quad \text{subject to} \quad \xi_{k_w} \leq \xi_{k_w+1}, \quad \text{for } k_w = 1, \ldots, K_w.$$

Following Gelman *et al.* (2008), we assumed Student-t distributions for the regression parameters, assuming that the continuous predictors are standardised and the dichotomous ones are mean centred

$$\boldsymbol{\alpha} \overset{\text{ind}}{\sim} t(3, 10), \quad \boldsymbol{\beta} \overset{\text{ind}}{\sim} t(3, 2.5), \quad \text{and } \boldsymbol{\gamma} \overset{\text{ind}}{\sim} t(3, 2.5).$$

In logistic regression with standardised predictors, coefficients are unlikely to be extremely large, thus this prior distribution choice represents a more vague and conservative information than the one we actually have.

Posterior inference was then obtained combining information from the prior distributions and the likelihood function within the described Bayesian model via Markov Chain Monte Carlo (MCMC) simulation.

We simulated our model using Stan for its flexibility and efficiency both on the computational and methodological perspective. Stan is based on the NUTS (No-U-Turn Sampler) (Hoffman and Gelman, 2014) algorithm, which is a MCMC algorithm based on the adaptive Hamiltonian Monte Carlo. Adaptive because it automatically sets all the parameters of the algorithm used to simulate a proposal, namely the step size and the number of leapfrog steps.

Models are fitted with an increasing number of knots until no evidence in favour of an extra knot is detected in diagnostic tools such as trace plots. In particular, given the computational burden of the problem, we consistently scaled up to the more complex models including an extra knot only when convergence diagnostics for the simplest models provided reassuring results.

Among the fitted models with different number of knots, we chose the best model as the one that minimises the Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010; Gelman *et al.*, 2014) among the convergent models. WAIC is fully Bayesian since it computes the predictive accuracy of a model evaluating the entire posterior distribution on the simulated values of the parameters, taking into account also the uncertainty of the parameters (Piironen and Vehtari, 2017). Asymptotically, WAIC coincides to the

leave-one-out cross validation (LOO-CV) (Watanabe, 2010).

Dealing with small dataset and a high number of competing models, the risk of overfitting in the selection procedure is present. In this situations, it is advisable to look at different model selection criteria, such as reference predictive method or projection method (Piironen and Vehtari, 2017). From a computational point of view, WAIC is definitely lighter and easier to obtain, since it does not need of a reference model to be computed. Moreover, the risk of overfitting is smaller when WAIC is used to compare models testing the inclusion of one variable at time to the selection among all possible combinations of predictors (Piironen and Vehtari, 2017).

1.3.2.1 Example of model selection procedure

Here we show an example of the model selection procedure for the larynx cancer data in which we examine the behaviour of the model in several diagnostic tools. Description of data and analysis are presented in Section 1.4.1.

It is good practice to check both diagnostic tools for the NUTS algorithm, such as energy plot and divergent transitions, and diagnostic tools related to the MCMC draws, such as trace plots, the posterior distribution of the parameters and $\hat{R}$ statistic (Gabry and Mahr, 2018).

As an example, Figure 1.1 shows trace plots for the case of larynx cancer model, including confounders and one knot on the intensity (cigarettes/day) and one on the duration (years of cigarette smoking). Plots show simulations of 4 chains each of $4,000$ iterations, respectively $2,000$ for the warm-up phase and $2,000$ for the sampling. Chains are initialised at different starting values. Trace plots reveal well mixing chains and clear convergence around 27 cigarettes/day for intensity knot parameter and around 31 years for the duration knot parameter.

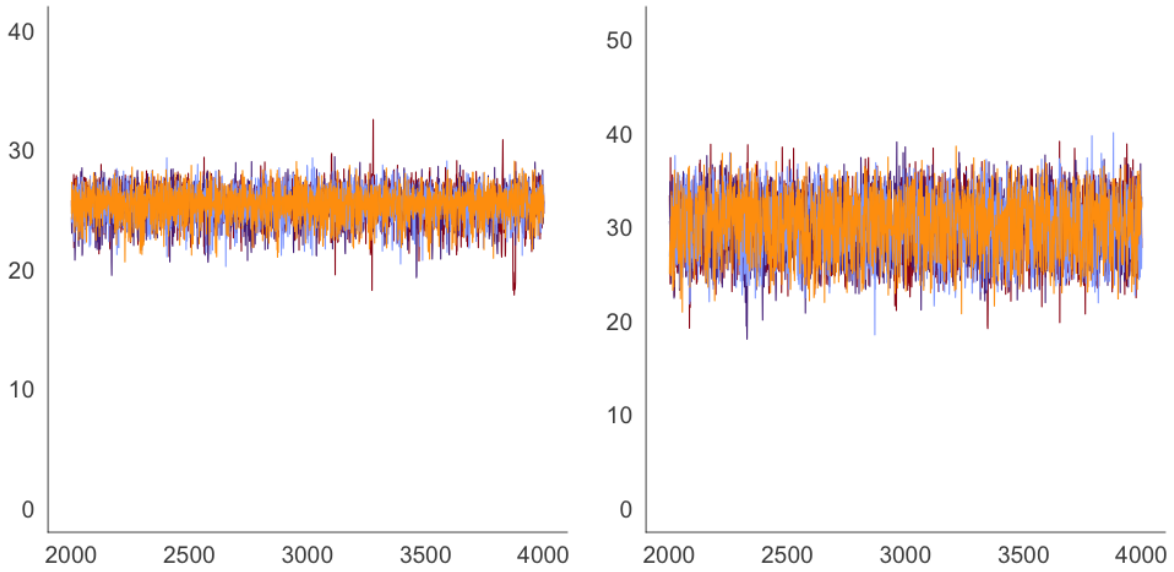

FIGURE 1.1: Trace plots for the intensity (left) and duration (right) knot location parameters in the larynx site semiparametric logistic model estimated with one knot on cigarettes/day and one knot on years of cigarette smoking duration. The chains mix well and converge quickly.

Other diagnostic tools do not highlight convergence issues in the algorithm and in the MCMC draws. There is no presence of divergent transitions, histograms of the marginal energy almost completely overlap and the $\hat{R}$ statistics are in the converge range of values for all the parameters (Tab. 1.1). Thus, it is advisable to estimate models increasing the number of knots.

Parameter	$\hat{R}$	n_{eff}	mean	sd	2.5%	50%	97.5%
ξ_{x1}	1	3,897	25.4	1.4	22.3	25.5	27.8
ξ_{w1}	1	3,693	30.2	3.3	23.9	30.5	35.8

TABLE 1.1: Posterior distributions of the knot location parameters for intensity and duration predictors. The semiparametric logistic model is estimated with one knot on duration and one on intensity variables. $\hat{R}$ and n_{eff} statistics confirm convergence of the chains and the quality of simulations for practical purposes.

Results for the model on larynx data estimated with an extra knot on the duration variable are shown in the following plots. Analysing trace plot in Figure 1.2, we notice that the chains of the knot location for intensity distinctly converge around 25 cigarettes/day. Conversely, the duration trace plots show that when one parameter takes values around 32 years of cigarette smoking, the other parameter moves towards upper (or lower) range values and it does not converge on a specific value (Fig. 1.3).

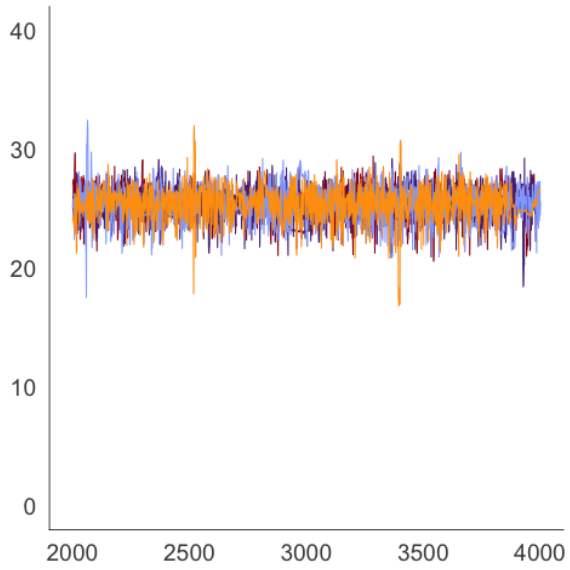

FIGURE 1.2: Trace plot for the intensity knot location parameter in the larynx site semiparametric logistic model estimated with one knot on cigarettes/day and two knots on years of cigarette smoking duration. Simulations from the sampling step are shown. The chain mixes well and converges quickly.

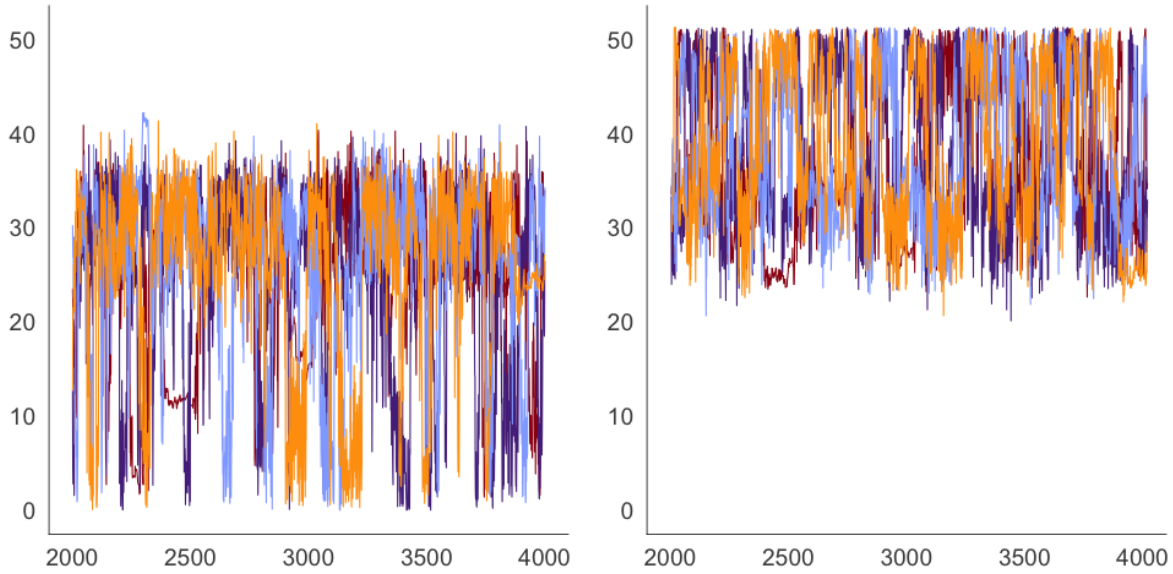

FIGURE 1.3: Trace plot for the duration knot location parameters in the larynx site semiparametric logistic model estimated with one knot on cigarettes/day and two knots on years of cigarette smoking duration. Simulations from the sampling step are shown. The chains show convergence issues due to overparameterisation.

Moreover, the n_{eff} statistic highlights highly correlated draws (Tab. 1.2). Inference based on these samples may lead to misleading and unreliable results.

Examining the posterior distributions of the two knots location parameters for the duration variable (Fig. 1.4), emerges a heavy left tail for the posterior distribution of the

first knot location parameter ξ_{w1}, represented by a solid line, and a bimodal posterior distribution for the second knot location parameter ξ_{w2}, depicted with a dashed line. The mode of the posterior distribution of ξ_{w1} coincides with the first mode of the posterior distribution of ξ_{w2}.

Parameter	$\hat{R}$	n_{eff}	mean	sd	2.5%	50%	97.5%
ξ_{w1}	1.02	213	25.2	9.5	2.6	27.7	36.8
ξ_{w2}	1.03	181	37.9	8.2	24.6.3	36.9	50.7

TABLE 1.2: Posterior distributions of the knot location parameters for the duration predictor. The semiparametric logistic model is estimated with two knots on duration and one on intensity variables. n_{eff} statistic highlights highly correlated draws leading to poor quality for practical purposes.

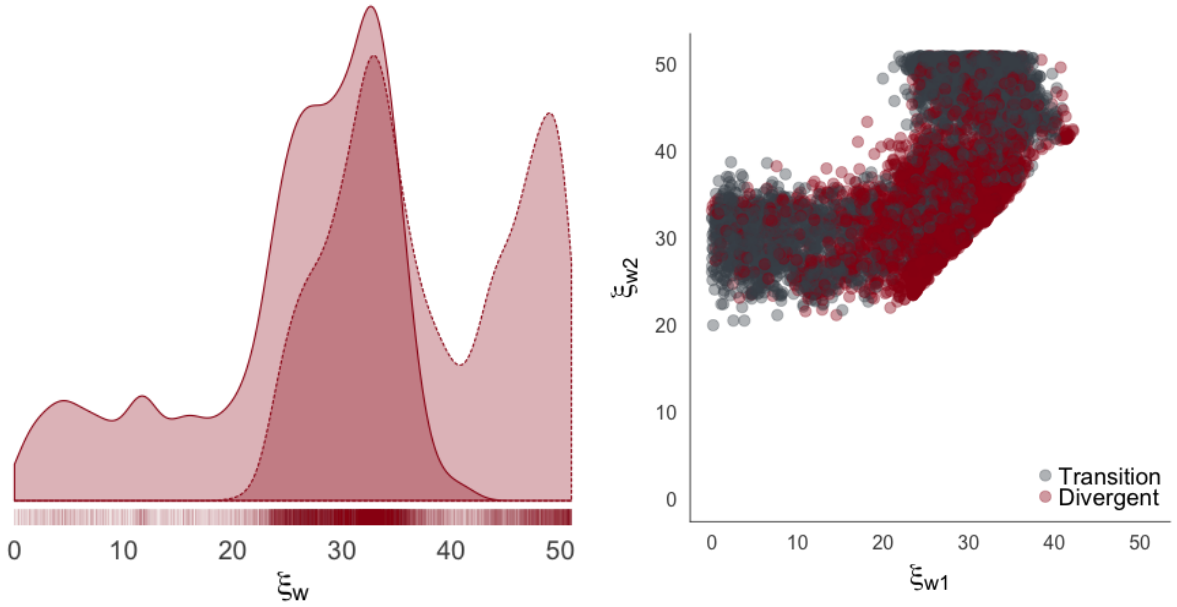

FIGURE 1.4: Posterior distribution of the knot location parameters related to the duration predictor (left). The solid line represents the posterior distribution of the first knot location parameter, while the dashed line represents the distribution of the second knot. The scatterplot on the right of the two knots location parameters highlights a region where divergent transitions are highly concentrated.

The interval of values around this mode is also characterised by a high number of divergent transitions (Fig. 1.4). The design matrix changes each time the knot location parameter is updated, thus, potentially, at each iteration. When the knot location parameters assume values close to each other, the produced columns of the design matrix are, by construction, (almost) collinear. When a knots location is estimated close to the lower bound of the predictor range, the design matrix is characterised by (almost) collinear columns with the predictor variable. On the other side, when the knots location

is estimated close to the upper bound of the predictor range, the generated variables are equal to zero. In our context, mainly the first behaviour forced by the ordered constraint imposed on the knot location parameters, leads to design matrix with near deficient rank with consequent estimation issues.

1.4 Application

Our application concentrates on the neoplasms of the pharynx and oral cavity sites and, separately, of larynx site, since they show a different associative pattern with the tobacco smoking habits.

1.4.1 Multicenter case-control studies

The International Head and Neck Cancer Epidemiology (INHANCE) consortium was established in 2004 to elucidate the aetiology of HNC through pooled analyses of individual-level data from several studies on a large scale (Hashibe *et al.*, 2007).

Several aspects of tobacco smoking and HNC risk have been previously investigated within the consortium (Hashibe *et al.*, 2007, 2009; Lubin *et al.*, 2009). From the INHANCE consortium pooled dataset (version 1.5), we extracted all the available case-control studies (35 studies) that collected information on cigarette smoking status, intensity, and duration at individual level (INHANCE, 2004). Information was harmonised at the study coordinating centre (Hashibe *et al.*, 2007). Although definitions varied by study, never smokers were those who never smoked regularly, or smoked for a very short period of time (generally less than one year) (Hashibe *et al.*, 2007). Former smokers were defined as those who had abstained from any type of smoking since at least 12 months before cancer diagnosis (cases) or interview (controls).

1.4.1.1 Selection of subjects

The INHANCE protocol allowed inclusion of invasive cancer cases of the oral cavity, oropharynx, hypopharynx, oral cavity or pharynx not otherwise specified, larynx, or unspecified HNC in the original studies. Cases with cancers of the salivary glands or of the nasal cavity/ear/paranasal sinuses were excluded (Hashibe *et al.*, 2007). The original study sample included $25,865$ head and neck cancer cases and $37,248$ controls, giving a total of $63,113$ subjects.

We conducted subjects' selection according to the following main steps, excluding: cases with unspecified (95 subjects) or overlapping head and neck cancers (331 subjects), subjects reporting other smoking habits (i.e., cigar, pipe, and cigarillo) than cigarettes, to avoid risk distortion due to the use of other tobacco products (Lubin *et al.*, 2009) (6, 255 subjects); subjects with missing information on duration and/or intensity of cigarette smoking (1, 897 subjects); all subjects from studies that included, after the previous selection steps, only never (i.e., Japan $1988 - 2000$) (822 subjects) or current (i.e., France $1987 - 1992$) smokers (457 subjects).

In order to prevent potential estimation distortion at the highest levels of the exposure distributions (due to small numbers of subjects or information bias in heavy tobacco consumers), we further excluded subjects reporting the highest 5% of cigarette smoking intensity (> 40 cigarettes per day) or duration (> 51 years), which ends up in the exclusion of 14% head and neck cases and 6% controls. After all the described selection steps, our analysis included 33 studies with 48, 104 subjects (18, 260 head and neck cancer cases and 29, 844 controls) (Tab. 1.3). If a study reported to carry out a case-control matching, separate controls were matched for oral cavity and pharynx (OCP) cancer cases combined and laryngeal cancer studies. In detail, there were: 5, 423 cancers of the oral cavity; 6, 261 pharyngeal cancer cases (4, 648 oropharyngeal and 1, 613 hypopharyngeal cancers cases); 1, 633 unspecified oral cavity/pharynx cancers (giving a total of 13, 317 OCP cancer cases combined), and 4, 943 laryngeal cancers. Among all the current and former smokers, only-cigarette subjects are selected, excluding who ever smoked pipe, cigar or cigarillo (Lubin *et al.*, 2009). The choice has been suggested by the lack of complete information about the simultaneous consumption of different tobacco products.

1.4.1.2 Data

The binary response variable specifies presence or absence of malignant pathologies. Risk factors describe duration and intensity of exposure: years of cigarette smoking and average daily number of smoked cigarettes (Tab. 1.4). Table 1.3 and Table 1.4 show selected characteristics of cases and controls, according to the variables included in the models as potential confounders and risk factors. Selected confounders are study design variables (sex, age, race), education (as a proxy of social status and income, important determinants of risk of head and neck cancers), study, smoking status and alcohol consumption variables (status, intensity and duration), one of the most important risk factor for cancers of the upper aerodigestive sites after tobacco smoking (IARC, 1986). The latter adjustment is supported by the evidence of an effect of both the duration

and the intensity of alcohol consumption on the analysed cancers (IARC, 1986). All the adjustment variables are categorical. Approximately 70% of the subjects were white. Studies from Europe contributed with approximately 44% of subjects; 31% of subjects were from the United States, whereas the remaining subjects were from Latin America (14%) and Asia (11%). Six studies provided cases of OCP cancer only (Tab. 1.3).

	Controls	(%)	OCP	(%)	Larynx	(%)
	29,844		13,317		4,943	
Sex						
Female	9,599	(32)	3,786	(28)	730	(15)
Male	20,245	(68)	9,531	(72)	4,213	(85)
Age						
<40	2,063	(7)	708	(5)	94	(2)
40 to 44	2,045	(7)	823	(6)	207	(4)
45 to 49	3,018	(10)	1,692	(13)	471	(10)
50 to 54	4,231	(14)	2,321	(17)	787	(16)
55 to 59	5,044	(17)	2,627	(20)	1,025	(21)
60 to 64	4,726	(16)	2,185	(16)	1,011	(20)
65 to 69	4,181	(14)	1,489	(11)	754	(15)
70 to 74	3,044	(10)	922	(7)	413	(8)
≥75	1,492	(5)	550	(4)	181	(4)
Race						
White	21,462	(72)	9,076	(68)	3,627	(73)
Black	976	(3)	683	(5)	169	(3)
Hispanic	421	(1)	149	(1)	41	(1)
Asian and Pacific Islanders	3,849	(13)	1,293	(10)	77	(2)
Others and Brazilians	3,136	(11)	2,116	(16)	1,029	(21)
Study						
Aviano	802	(3)	288	(2)	128	(3)
Baltimore	163	(1)	123	(1)	29	(1)
Beijing	377	(1)	322	(2)	0	(0)
Boston	473	(2)	339	(3)	75	(2)
Buffalo	863	(3)	282	(2)	113	(2)
Central Europe	730	(2)	238	(2)	295	(6)
France Multicen. (1989-1991)	255	(1)	163	(1)	247	(5)

France Multicen. (2001-2007)	2,964	(10)	1,343	(10)	366	(7)
Germany-Heidelberg	644	(2)	0	(0)	172	(3)
Germany-Saarland	83	(0)	53	(0)	22	(0)
HOTSPOT	59	(0)	63	(0)	0	(0)
Houston	738	(2)	561	(4)	119	(2)
International Multicenter	1,450	(5)	1,053	(8)	0	(0)
Iowa	600	(2)	344	(3)	58	(1)
Italy Multicenter	2,506	(8)	685	(5)	421	(9)
Japan (2001-2005)	2,817	(9)	392	(3)	71	(1)
Latin America	1,446	(5)	981	(7)	612	(12)
Los Angeles	905	(3)	273	(2)	70	(1)
Milan (1984-1989)	1,413	(5)	161	(1)	215	(4)
Milan (2006-2009)	669	(2)	118	(1)	162	(3)
MSKCC	115	(0)	61	(0)	20	(0)
New York Multicenter	1,246	(4)	818	(6)	202	(4)
North Carolina (1994-1997)	154	(1)	91	(1)	31	(1)
North Carolina (2002-2006)	982	(3)	594	(4)	283	(6)
Puerto Rico	410	(1)	182	(1)	0	(0)
Rome	350	(1)	98	(1)	173	(3)
Sao Paulo	1,519	(5)	1,042	(8)	406	(8)
Seattle (1985-1995)	465	(2)	317	(2)	0	(0)
Seattle-Leo	371	(1)	264	(2)	123	(2)
Switzerland	824	(3)	311	(2)	104	(2)
Tampa	789	(3)	115	(1)	48	(1)
US Multicenter	936	(3)	710	(5)	0	(0)
Western Europe	1,726	(6)	932	(7)	378	(8)
Education						
No education	1,078	(4)	726	(5)	118	(2)
≤Junior high school	10,456	(35)	4,674	(35)	2,371	(48)
Some high school	5,330	(18)	2,751	(21)	922	(19)
High school graduate	3,883	(13)	1,883	(14)	717	(14)
Technical school, some college	4,825	(16)	1,989	(15)	496	(10)
≥College graduate	4,272	(14)	1,294	(10)	319	(6)
Drinking status						
Never user	8,068	(27)	2,279	(17)	578	(12)
Former user	3,072	(10)	2,521	(19)	833	(17)

Current user	14,210	(48)	6,943	(52)	2,442	(49)
Missing	4,494	(15)	1,574	(12)	1,091	(22)
Alcohol drinking intensity						
0-<1	17,207	58)	4,899	37)	1,433	(29)
1-<5	8,913	30)	4,099	31)	1,757	(35)
≥5	2,925	10)	3,814	29)	1,626	(33)
Missing	799	3)	505	4)	127	(3)
Alcohol duration (years)						
Never drinkers	8,068	(27)	2,279	(17)	578	(12)
0-<20	2,984	(10)	1,182	(9)	297	(6)
20-<40	9,427	(32)	5,422	(41)	1,848	(37)
40-<60	6,013	(20)	2,920	(22)	1,435	(29)
≥60	412	(1)	176	(1)	123	(2)
Missing	2,940	(10)	1,338	(10)	662	(13)

TABLE 1.3: Selected adjustment variables included in the models. We excluded subjects with missing values on age, sex and race. Missing values for education were imputed according to Hashibe *et al.* (2007).

Table 1.4 shows the distribution of cigarette smoking habits for OCP cancer, laryngeal cancer, and controls. Never smokers were 21% of OCP and 7% of laryngeal cancers, versus 45% of controls. Current smokers were 66% of laryngeal cancer cases, 57% of OCP cancer cases, and 26% of controls. The prevalence of former smokers was similar in cases and controls, but the frequency of people who quit cigarette smoking 10 years ago or more was about 52% among HNC cases and 72% among controls.

	Controls	(%)	OCP	(%)	Larynx	(%)
Cigarette smoking status						
Never user	13,347	(45)	2,791	(21)	330	(7)
Former user	8,792	(29)	2,909	(22)	1,353	(27)
Current user	7,705	(26)	7,617	(57)	3,260	(66)
Cigarette smoking intensity						
Never user	13,347	(45)	2,791	(21)	330	(7)
≥1-15	7,199	(24)	2,814	(21)	1,054	(21)
>15-25	6,166	(21)	4,483	(34)	2,083	(42)
>26-40	3,132	(10)	3,229	(24)	1,476	(30)

Cigarette smoking duration						
Never user	13,347	(45)	2,791	(21)	330	(7)
≥1-25	7,214	(24)	2,019	(15)	604	(12)
>25-35	4,360	(15)	3,370	(25)	1,330	(27)
>35-51	4,923	(16)	5,137	(39)	2,679	(54)
Time since quit smoking						
≥1-<10	2,300	(26)	1,310	(45)	626	(46)
≥10	3,655	(72)	1,522	(52)	711	(53)
Missing	137	(2)	77	(3)	16	(1)

TABLE 1.4: Smoking habits related variables. Time since quitting cigarette smoking variable is included only in the former smokers strata analysis.

1.4.2 Results

We perform a stratified analysis separating current and former smokers and also dividing larynx from OCP sites because of the different associative pattern of each stratum. Moreover, the analysis is stratified by alcohol intensity consumption, measured in drinks per day, in current smokers and OCP sites.

A joint posterior distribution on knot locations and regression parameters was separately simulated for each cancer site, smoking status stratum, and combination of number of knots and it was based on the $8,000$ iterations ($2,000$ iterations times 4 chains) of the corresponding sampling step. Diagnostics criteria, including trace plots of the marginal chains, $\hat{R}$ of single parameters ($1 < \hat{R} < 1.05$), divergent transitions (not present), and energy plots (histograms completely overlapped), were satisfied for most models and reassured that the chains converged and the parameter space was fully explored for any parameter (Gabry *et al.*, 2017; Carpenter *et al.*, 2017).

Results are presented on the log odds ratio scale and surfaces are showed through perspective plots (three-dimensional graphs where we reported the ORs associated with any combinations of the two exposures) and contour plots (two-dimensional graphs showing iso-risk curves that identify combinations of cigarette smoking intensity and duration with the same cancer risk).

Once the optimal combination of knots locations was identified for each cancer site and stratum, we calculated the corresponding ORs of cancers of the OCP and larynx, together with the 95% credible intervals (CIs), from the marginal posterior distribution of the parameters for the two exposures.

Figure 1.5 and Figure 1.6 show the mesh and contour plots for cancers of the OCP and larynx among current cigarette smokers. Knot locations are indicated with thicker black lines, which represent a changing slope in risk surfaces. For both cancer sites, the best model was characterised by one knot for duration (at 33 years for OCP cancer and 30 years for laryngeal cancer) and one knot for intensity (at 16 and 25 cigarettes/day, respectively).

For any level of either exposure, risks of both cancer sites increased more with duration than with intensity. For OCP cancer, given a fixed value of ~ 40 pack-years, an OR~ 6 was reached after a duration of ≥ 40 years by smokers of ≥ 20 cigarettes/day, whereas the OR was ~ 4 after a duration of ≥ 20 years by smokers of ≥ 40 cigarettes/-day (Fig. 1.5). Notably, for any duration of up to 10 years, the ORs were always < 2, regardless of cigarette smoking intensity (Fig. 1.5). In contrast, smoking duration modified OCP cancer risk at any levels of smoking intensity; however, for very low intensities, ORs> 2 were reached only with long durations (i.e. > 40 years). In addition, for a fixed value of 20 pack-years, an intensity of 40 cigarettes/day and a duration of 10 years led to an OR of ~ 2, whereas the OR was equal to ~ 4 with a duration of 40 years by smokers of 10 cigarettes/day.

For laryngeal cancer, ORs> 20 were found for intensities of > 20 cigarettes/day and durations of > 28 years (Fig. 1.6). Moreover, ORs> 10 of laryngeal cancer were reached by current smokers of > 20 cigarettes/day only when duration was > 20 years, but they were not reached for any duration < 15 years in any level of intensity. Finally, the OR was 6.2 for smokers of 40 cigarettes/day for 10 years, but it was higher (between 9 and 10) for 20 cigarettes/day smoked for 20 years, or for 10 cigarettes/day smoked for 40 years.

Figures 1.7–1.9 show the joint effect of current smoking intensity and duration in strata of alcohol consumption. Among never drinkers (< 1 drink/day), the shape of the risk surface is similar to the one presented for all alcohol intensities together (Fig. 1.5), but the ORs were generally lower: the ORs were all less than 2 for any intensity of cigarette smoking and durations ≤ 15 years, whereas an OR> 5 was observed only after about 25 years of duration or more (Fig. 1.7). However, the shape of the surface and/or the ORs of the joint effect of duration and intensity are different when light (Fig. 1.8) and heavy (Fig. 1.9) drinkers are considered.

As a comparison, ORs and their corresponding 95% CI for each cancer site in current smokers were estimated within the Bayesian logistic regression model assuming step functions (Tab. 1.5, main ORs). These estimates were compared with the range of OR estimates derived from spline models for each joint category of duration and intensity

(Tab. 1.5, bracketed ORs). For both cancer sites, all Min-Max ranges included the OR estimates obtained from the Bayesian logistic regression, and this replication reassures that the spline model is valid. In addition, within the examined categories, the step-function intervals widely overlapped with the Min-Max ranges of the ORs from spline models, but failed to capture the variability in the ORs. For instance, for OCP cancer in current smokers of $26-40$ cigarettes/day for $36-51$ years, the categorical OR was 8.4 (95% CI: $8.0-8.9$), whereas the OR varied from 7.1 to 10.6 under the spline model approach for the same exposure categories (Tab. 1.5).

The same pattern emerged at any combination of duration and intensity for both cancer sites for former smokers who quit 10 years ago or more (Figs. 1.10 and 1.11). Among former smokers who quit more than 10 years ago, no ORs>4 were observed for OCP cancer (Fig. 1.10), and the ORs were approximately halved, as compared to the same levels of duration and intensity in current smokers (Fig. 1.5). A 1/3 reduction of risk was also found for laryngeal cancer in long-term former smokers, with reductions in the ORs that varied depending on the different combinations of intensity and duration (Fig. 1.11). These estimates were derived from models that included one knot for intensity (27 cigarettes/day) for laryngeal cancer and no knots for OCP cancer.

Some estimated surfaces seem to present a decreasing OR after the knots for increasing levels of risk factors, e.g. OCP cancer site for current smokers (Fig. 1.5) and OCP cancer site stratified by alcohol (Figs. 1.7 and 1.9). This not has to be interpreted in the wrong way. Indeed, examining the two-dimensional 95% credible intervals, it is easy to note that the estimates near the boundary regions, for high exposure levels, are characterised by an increasing variability due to the lower number of subjects (Appendix A). Moreover, the non decreasing OR is always included in the estimated 95% credible intervals (Appendix A).

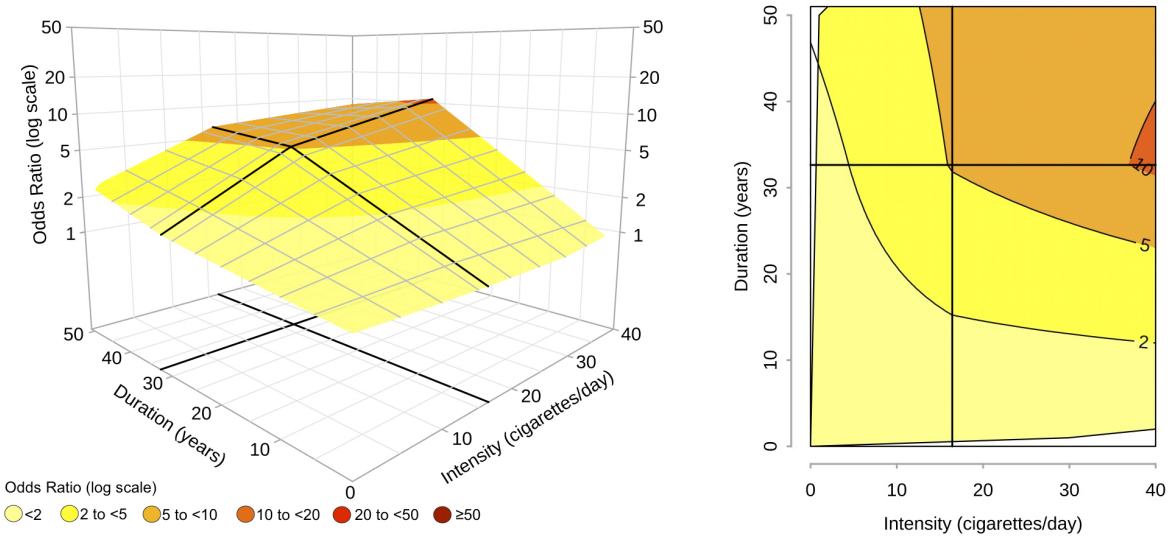

FIGURE 1.5: Current smokers - stratified analysis by oral cavity and pharynx sites. On the grid, black thicker lines represent knot locations: 16 cigarettes/day and 33 years of duration for oral and pharyngeal cancer. Dark grey lines in contour plots indicate iso-risk curves at defined levels of risk.

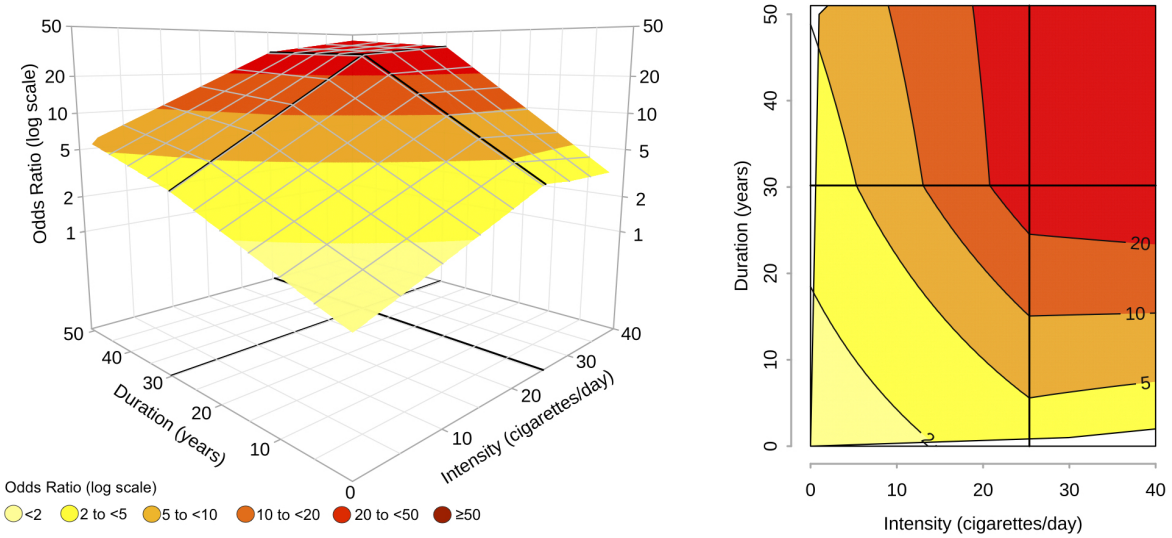

FIGURE 1.6: Current smokers - stratified analysis by larynx site. On the grid, black thicker lines represent knot locations: 25 cigarettes/day and 30 years of duration for larynx cancer. Dark grey lines in contour plots indicate iso-risk curves at defined levels of risk.

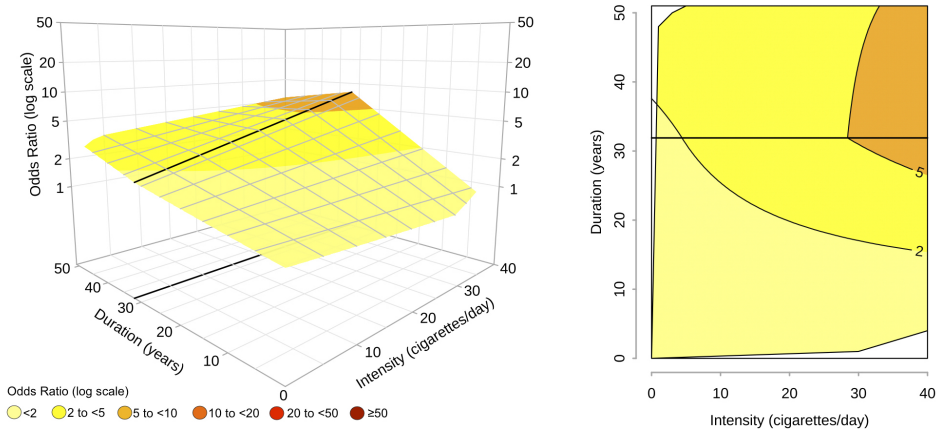

FIGURE 1.7: Current smokers - stratified analysis by oral cavity and pharynx sites and never drinkers. On the grid, black thicker line represents the knot location at 32 years of duration. Dark grey lines in contour plots indicate iso-risk curves at defined levels of risk.

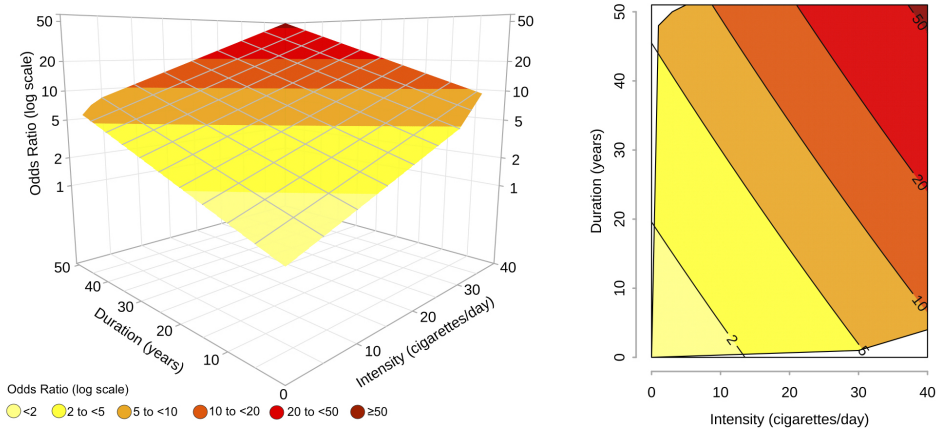

FIGURE 1.8: Current smokers - stratified analysis by oral cavity and pharynx sites and light drinkers. Dark grey lines in contour plots indicate iso-risk curves at defined levels of risk.

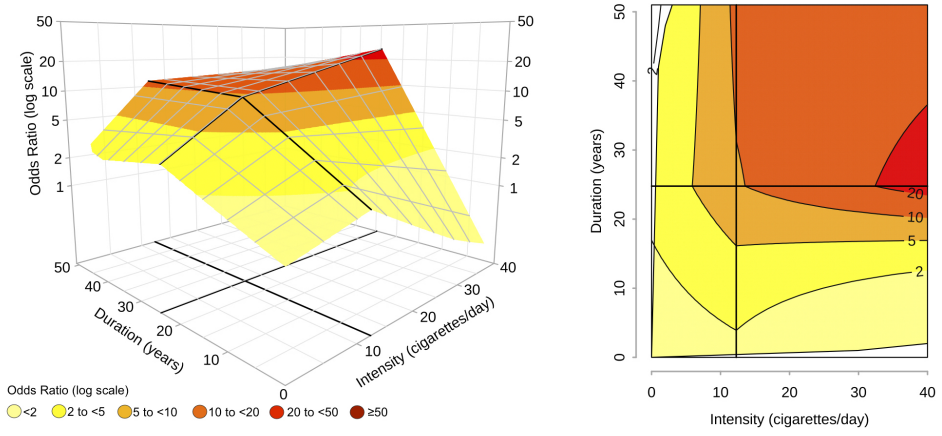

FIGURE 1.9: Current smokers - stratified analysis by oral cavity and pharynx sites and heavy drinkers. On the grid, black thicker lines represent knot locations: 12 cigarettes/day and 25 years of duration for oral and pharyngeal cancer. Dark grey lines in contour plots indicate iso-risk curves at defined levels of risk.

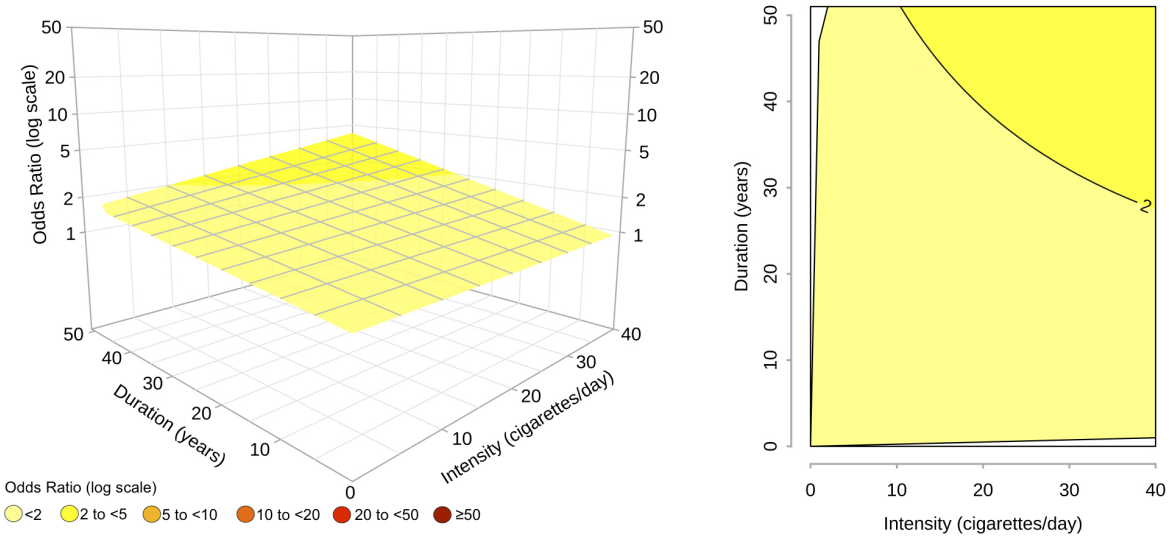

FIGURE 1.10: Former smokers who quit smoking more than 10 years ago - stratified analysis by oral cavity and pharynx sites. Dark grey lines in contour plots indicate iso-risk curves at defined levels of risk.

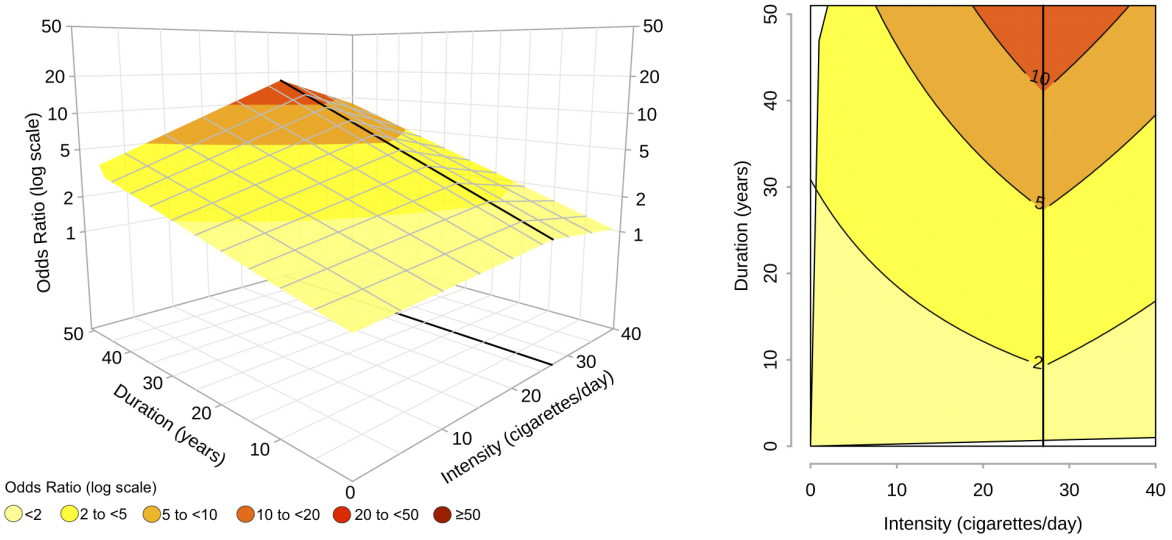

FIGURE 1.11: Former smokers who quit smoking more than 10 years ago - stratified analysis by larynx sites. On the grid, the black thicker line represents the knot location: 27 cigarettes/day for laryngeal cancer. Dark grey lines in contour plots indicate iso-risk curves at defined levels of risk.

Cancer type		Intensity (cigarettes/day) 1-15 Ca:Co	1-15 OR (95%CI) [Min-Max]	16-25 Ca:Co	16-25 OR (95%CI) [Min-Max]	26-40 Ca:Co	26-40 OR (95%CI) [Min-Max]
Oral cavity and pharynx							
	Duration (years)						
	1-25	322:1,046	1.5 (1.4-1.7) [0.9-3.1]	371:663	2.7 (2.4-2.9) [0.9-4.2]	191:270	3.3 (2.8-3.9) [0.8-5.9]
	26-35	583:911	2.9 (2.7-3.1) [1.4-4.8]	1,138:954	5.1 (4.9-5.4) [3.5-6.9]	849:466	7.2 (6.7-7.8) [4.6-11.0]
	36-51	912:1,175	3.7 (3.5-3.9) [1.6-5.8]	1,838:1,409	6.0 (5.8-6.2) [5.2-7.1]	1,413:694	8.4 (8.0-8.9) [7.1-10.6]
Larynx							
	Duration (years)						
	1-25	75:908	3.8 (3.3-4.4) [1.1-8.8]	112:576	9.4 (8.3-10.5) [2.3-20.1]	70:247	14.8 (12.2-17.8) [2.8-23.1]
	26-35	183:812	7.3 (6.9-7.9) [2.9-12.6]	423:845	17.2 (16.8-17.6) [10.1-29.6]	319:415	28.5 (27-30.2) [22.4-37.7]
	36-51	394:1,064	8.9 (8.8-9.1) [3.9-15.3]	966:1,274	19.5 (19.3-19.8) [13.9-31.0]	718:616	33.9 (33.5-34.3) [30.9-43.2]

TABLE 1.5: Odds ratios (ORs)a and 95% credible intervals (CIs) of OCP and laryngeal cancer in current smokers, for the joint effect of intensity (cigarettes/day) and duration (years) of cigarette smoking estimated through step-function as compared with results from bivariate spline models. Min and Max represent the lowest and the highest OR values estimated for any combinations of intensity and duration by bivariate spline models. Fitted models included adjustment for age, sex, race, study, education, drinking status, drinking intensity, and drinking duration. The reference category was defined as "Never smokers" and it includes $2,791$ cases and $13,139$ controls for the analysis on OCP cancer and 330 cases and $11,433$ controls for that on laryngeal cancer.

1.5 Discussion

Our large international pooled analysis shows that cigarette smoking duration and intensity do not increase HNC risk to the same extent, but the effect is greater for those having a longer duration of cigarette smoking. Bivariate regression spline models have proven to be a successful approach in exploring the separate and joint effects of intensity and duration of cigarette smoking. So, when possible, we should consider models that allow this differential impact of intensity and duration on risk to be taken into account (Peto, 2012; Lubin and Caporaso, 2013).

For both cancer sites examined, the dose-response relationship is still far from being linear, with a steeper increase with duration and a possible plateaux indicating a "saturation effect" in smokers with ≥ 20 years of duration and with ≥ 30 cigarettes/day.

In previous studies, an interaction term between alcohol consumption and cigarette smoking was added to the models. Here, we explored this effect by referring to a stratified analysis according to different levels of alcohol consumption and we provided further evidence that alcohol acts as a substantial modifier of the association between OCP cancer risk and the joint effect of cigarette intensity and duration.

We apply a novel Bayesian approach to jointly estimate the optimal knot locations and the ORs of HNC for the joint effect of intensity and duration in a bivariate context. After examining various frequentist solutions to the optimal choice of knot locations (Mao and Zhao, 2003; Molinari *et al.*, 2004), we opted for the Bayesian approach that allowed us to put the knots where the data suggested, once we had taken the entire set of confounding variables into consideration.

The proposed approach is also applicable to other epidemiological situations where continuous exposures in their potential interaction affect disease risk. Criteria for choosing the best fitting models are still questionable (Piironen and Vehtari, 2017), but all of them must take into account previous epidemiological evidence suggesting that, when the exposure of interest is cigarette smoking, the number of changes in the risk pattern (i.e., knots) is likely to be at most two. Indeed, the exposure is supposed to have a protective or null effect at lower levels, e.g., for the association between alcohol and HNC (Polesel *et al.*, 2005) and/or a saturation effect on the risk at the highest intensity levels (Schöllnberger *et al.*, 2006), as well as the expected increase in risk at the intermediate levels of consumption. This ends up in spline models with either one or two knots and corresponding two or three changes in the slope of the OR surface. Within the Bayesian framework, we were able to choose the optimal number of knots (up to 2) and knot

locations by model comparison based on information criteria, instead of just comparing models with fixed number and location of the knots.

Among study limitations, the retrospective study design and the self-reported smoking history were the most relevant ones. Indeed, misclassification may occur, especially among heavy smokers. Furthermore, smoking intensity may vary over time and by age of exposure. However, its estimates are often based on the self-reported average number of cigarettes smoked each day; these two aspects can easily lead to appreciable error in measuring the true mean intensity of exposure over one's lifetime. To reduce possible information bias, we excluded subjects reporting higher cigarette intensity and/or duration from the present analysis. In addition, to avoid bias due to the use of other tobacco products, we excluded subjects reporting use of tobacco products other than cigarettes.

Finally, our Bayesian approach was computationally time consuming, asking for several hours of server computing for each model fitted.

Chapter 2

Bayesian estimation of number and position of knots in regression splines

2.1 Introduction

When modelling the relationship between a response and some (continuous) covariates the linearity assumption turns out to be too restrictive in many contexts. Naive solutions to overcome this limitation such as categorisation of the predictor or its polynomial representation have well-known drawbacks. The assumptions and complications underlying the choice of using predictors as categorical variables are described in Section 1.2. Polynomial regression can be a flexible solution to model a non linear relation- ship among the outcome and the predictor especially when a visual inspection of the variables is available. Unfortunately, in complex models this approach can be unfeasible and selecting the polynomial degree may be challenging. Indeed, the higher the degree, the higher the risk of overfitting. Using polynomial bases, each observation affects the entire curve, thus they are characterised by a high sensitivity to outliers which are usually not easy to detect.

A viable alternative is represented by spline functions. They are defined as piecewise polynomials with a fixed degree whose joint points are called knots. Splines are highly flexible and are described as an excellent approximation tools (de Boor, 2001). In fact, varying the number and position of knots may lead to extremely different shapes and a major risk is to overfit the data. The two factors that affect the most the flexibility of the spline function are the number of distinct knots and the polynomial degree. For a given polynomial degree, the well-known bias-variance trade-off is controlled by the

knots: the higher the number of knots, the lower the bias of the estimated function.

To manage the flexibility of the regression spline, a classical approach consists in using an optimising criterion with a suitable penalisation to control the roughness of the function. In this case, the number and position of the knots is not as crucial, e.g. potential knots are placed evenly-spaced on the predictor range or on the quantiles of the predictor. Then, the optimal number of knots is selected as the minimiser of some criteria through cross-validation or, alternatively, a penalisation term on the spline coefficients is added. In the literature, the latter option is usually preferred since the former setting may require high computational effort depending on the model complexity. The penalisation term usually acts on the curvature of the function leading to shrunk spline coefficients. The term that calibrates the amount of penalisation can be selected using cross-validation criteria, AIC, GCV or their derivations (Ruppert *et al.*, 2003; Wood, 2006).

Assuming that the number and position of knots may have an important and substantial interpretation, other techniques proposed in the literature include the use of variable selection to choose basis function (Smith and Kohn, 1996), or employing samplers that allow for varying dimension of the parameter (Denison *et al.*, 1998; DiMatteo *et al.*, 2001). Variable selection is usually applied to a set of knots to choose from, thus number and positions of knots are determined and fixed. The latter class of methods allows to estimate both the number and the position of the knots but at the cost of greater methodological complexity and possible convergence issues. Moreover, transdimensional methods are known to be time consuming especially when the dimension of the parameter space is high.

Here we consider estimation of number and position of knots following one of the most recent approaches to variable selection in a Bayesian context. Estimating the positions of the knots is not an easy task and, for a fixed degree, regression coefficients and locations of knots have to be estimated simultaneously, turning the standard estimation procedure into a non linear optimisation problem.

In the sequel, we propose a Bayesian method to estimate the number and position of knots with a two-step procedure in the generalised linear model framework. A description of the method is presented in the following section. Hence, simulation study on synthetic data is introduced in the subsequent section. We conclude the chapter with an application to real data using bivariate splines.

2.2 Methods

2.2.1 Model

Generalised linear models are characterised by three main elements: an exponential family distribution of the dependent variable, a link function and a linear predictor which incorporates information from the independent variables (McCullagh and Nelder, 1983). The model describes the linear relationship between the expected value of the response variable, transformed by the link function, and the linear predictor.

Consider the semiparametric generalised linear model

$$\begin{aligned} \mathbb{E}[y_i] &= g^{-1}(\eta_i) \\ \eta_i &= \mathbf{z}_i\alpha + f(x_i), \quad \text{for } i = 1, \ldots, n, \end{aligned} \tag{2.1}$$

where $\mathbf{Y}$ is the dependent variable, g is the link function and η is the linear predictor. Furthermore, $\mathbf{Z}$ is the covariates vector that enters linearly in the model, α is the vector of regression coefficients and $\mathbf{X}$ is a continuous variable affecting the response through a smooth function $f : \mathbb{R} \to \mathbb{R}$, described with a spline with few knots.

We restrict our analysis to those situations in which a low number of knots can be adequate and their positions are directly interpretable and of specific interest for the analysis. This is the case, for example, when truncated power basis (TPB) of order one is used since in this case positions of knots represent changing points for the slope. One of the main drawbacks of truncated power basis representation is that the basis is not orthogonal, which can lead to numerical instability and slow convergence of the optimisation algorithm. Keeping a low number of knots alleviates the issue (Ruppert *et al.*, 2003).

Let then

$$f(\mathbf{x}) = \beta_0 + \beta_1\mathbf{x} + \sum_{k=1}^{K} \gamma_k(\mathbf{x} - \xi_k)_+, \tag{2.2}$$

where ξ_k is the position of the k-th knot and K is the total number of knots, and

$$(x - \xi_k)_+ = \begin{cases} x - \xi_k, & \text{if } x \geq \xi_k \\ 0, & \text{otherwise} \end{cases}$$

is the truncated linear function. Given the number of knots, parameters estimation reduces to maximum likelihood estimate through optimisation algorithms such as derivations of Newton-Raphson method (Hastie and Tibshirani, 1990).

A usual approach is to choose the knot locations using standard criteria (such as quantiles of the predictor distribution, uniformly distributed knots on the range of the independent variables and user-defined knots following a priori information (Ruppert

et al., 2003), estimate models with a different number and location of knots and compare them through standard criteria, such as AIC, BIC or GCV. This procedure often results in a not clear discrimination among all competing models.

2.2.2 Free-knot regression splines

In order to enhance the fit of the model, a possible extension is to consider locations of knots as parameters to be estimated along with other regression coefficients. In such a case, within a maximum likelihood approach, exploration of the objective function surface could locate local maxima leading to apparent solutions strongly dependent on starting values. A Bayesian specification of the model and exploration of the posterior distribution, possibly by using MCMC simulations, could prove in this case much more effective.

2.2.2.1 No Variable selection approach

Our aim is to estimate both number and location of knots, thus a preliminary idea is to estimate several models with free knot locations and with increasing but fixed number of knots. Prior information available on the knots is that they are constrained to be ordered. For this reason, we define their prior distributions as Uniform on the range of the variable, that is

$$\xi_k \sim Unif(\min(\mathbf{x}), \max(\mathbf{x})), \quad \text{subject to} \quad \xi_k \leq \xi_{k+1}, \quad \text{for } k = 1, \ldots, K.$$

Diffuse priors on the regression and spline coefficients are chosen. In particular,

$$\boldsymbol{\alpha} \overset{\text{ind}}{\sim} N(0, \sigma_\alpha),$$

and

$$\boldsymbol{\beta} \overset{\text{ind}}{\sim} N(0, \sigma_\beta),$$

where both σ_α and σ_β are selected, according to the range of the variables, such that the prior distribution is weakly informative. We will refer to this model as the no variable selection (NVS) model.

Models with an increasing number of knots were compared on the basis of diagnostic tools such as trace plots and $\hat{R}$ to check convergence of parameters, and information criteria were used to choose the best model among the estimated ones. The main drawback of this procedure is a large number of models that one needs to consider and the implied computational effort in high dimensional problems.

As an example of the described procedure, we refer to the application in Section 1.4.

2.2.2.2 Stochastic search variable selection approach

Results obtained on simulated data on the basis of diagnostic tools show that the NVS approach can lead to reasonable estimates and that convergence of chains for knot related parameters is univocal only if the number of specified knots is lower or equal to the true one.

This prompted us to consider a two-step procedure:

- select the optimal number of knots considering a large, possibly, overparameterised model,
- fit the final model on a restricted set of knots by simultaneously estimating locations of knots and regression and spline coefficients.

In the first step, we estimate a model having more knots than reasonably warranted. This leads to an overparameterised model where the posterior of some knot locations are expected to concentrate at the limits of the predictor range.

To assess convergence of the spline parameters, our advice is to run several chains and look at the results of each chain separately. Indeed, overparameterising the model may lead to chains which converge at different points. As an example, suppose that the true number of knots is two and we simulate two chains to fit the model with 5 ordered knots. It can happen that in the first chain the first and the second knot parameters, say ξ_1 and ξ_2, converge on the values of the true knots, while in the second chain the second and third knots parameters, ξ_2 and ξ_3, converge on the right values. Posterior inference on the combined chains would lead to not reliable results for the parameters, while looking at the posterior results distinctly for each chain would let us to properly recognise the presence of two knots.

Since each knot location is uniquely linked to a spline coefficient, we evaluate the presence of a knot based on the analysis of the associated coefficient posterior distribution.

The concept underlying the proposed methodology is to perform variable selection on the basis functions, for this purpose we employ one of the most common approaches in Bayesian literature: that based on the definition of spike-and-slab priors.

Several versions have been proposed in the literature (O'Hara *et al.*, 2009) but, generally speaking, prior distributions for the regression coefficients are defined with a spike component, usually highly concentrated around zero, and a diffused slab part. This is the case of the stochastic search variable selection approach (SSVS), that defines

a mixture distribution for each parameter that has to be selected (O'Hara *et al.*, 2009). This type of methodology gives us the opportunity to evaluate the presence of a variable through the marginal posterior distribution of the mixing proportion.

Starting from the NVS model specification, we set a prior distribution on each spline parameter γ_k such that

$$\pi(\gamma_k|\lambda_k) = \lambda_k N(0, \sigma_{sl}) + (1 - \lambda_k)N(0, \sigma_{sp}),$$

where the mixing proportion $\lambda_k \sim Beta(a, b)$, with $a = b$. Standard deviations of the two mixture components, σ_{sl} and σ_{sp}, are chosen to be respectively large and small. Appropriate values have to be evaluated taking into account the unit of measurement of dependent and independent variables.

Our method adapts the modified SSVS approach by assuming λ_k to be dependent on the knot location ξ_k. The prior distributions of the ordered knots remain defined as Uniform on the support of the variable X and independent from both the mixing proportion $\boldsymbol{\lambda}$ and the coefficient $\boldsymbol{\gamma}$. Each coefficient γ_k, conditioned on the mixing parameter λ_k follows the same mixture distribution of two components specified in the SSVS approach described above, while each element of the mixing proportion vector $\boldsymbol{\lambda}$ is now defined as:

$$\lambda_k|\xi_k \sim Beta(a, b_k),$$

where a is a positive but very small value and $b_k : [\min(\mathbf{x}); \max(\mathbf{x})] \rightarrow [a; 1 + a]$ is a U-shaped even function of the knot location which returns values close to $1 + a$ when the knot is near the boundaries of the variable, while it is almost uniform and close to a elsewhere. In practice, the prior for the mixing parameter swings between a beta U-shaped distribution when the knot location is on plausible values and a beta distribution highly concentrated on zero when the knot is close to the boundaries (Fig. 2.1).

All the other prior distributions remain defined as in the previous model specification.

In the next section, we compare results from (i) the proposed method, named later on SSVSξ, with (ii) the ones obtained from the SSVS approach and (iii) the same model without a variable selection procedure, NVS. Moreover, we tested our approach on data simulated with different signal to noise ratios.

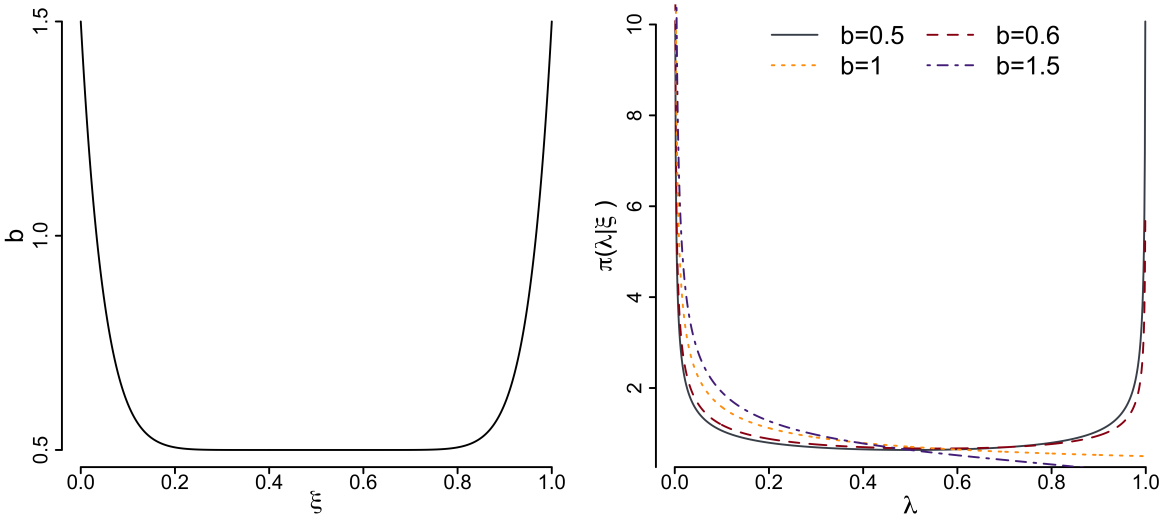

FIGURE 2.1: Plot of the hyper-parameter b varying ξ (left). When the knot location is estimated close to the boundary region of the predictor, the hyper-parameter b takes increasing value on the range $[0.5; 1.5]$. Consequently, the prior distribution on λ moves from a horseshoe shaped distribution to concentrate on values close to zero (right).

2.3 Preliminary results and simulation study

We simulate data from the linear regression model

$$y_i = 6 + 2x_i - 5(x_i - 2.7)_+ + 8(x_i - 4.3)_+ + \varepsilon_i, \quad \text{for } i = 1, \ldots, 500,$$

where $\varepsilon_i \overset{\text{ind}}{\sim} N(0, 3)$ and the predictor $\mathbf{X}$ is uniformly defined in the interval $[0; 10]$. Two knots are placed respectively in 2.67 and 4.33. We set the parameter a of the mixing proportions $\boldsymbol{\lambda}$ for the SSVS and the SSVSξ models equal to 0.5. Moreover, we chose σ_{sl} equal to 100 and σ_{sp} equal to 0.1. Standard deviations of the prior distributions on spline coefficients and intercept were chosen equal to 100.

We run 10 chains with $2,000$ iterations each. Posterior inference is based on the last $1,000$ draws of each chain. To support the complete exploration of the posterior distribution, initial values for the locations of the knots are chosen widely spread on the range of the predictor variable $\mathbf{X}$. Spline coefficients and intercept are initialised at zero. The three models are fitted with a different number of knots, respectively with 2, 5 and 10 knots. The interest lies in the parameter estimates, both spline coefficients and knot locations, and in the analysis of the chains behaviour.

The number of knots can be chosen in the SSVS and SSVSξ models looking at the plots in Figure 2.2. Plots in the first row refer to the SSVS models, while plots in the

second row refer to the SSVSξ approach. The x-axis represents the specified number of knots in the overparameterised models, while the y-axis represents the posterior mean of the mixing proportion. Vectors of posterior means are sorted in descending order and each line corresponds to one chain. In both models performing variable selection, the selected number of relevant knots is always equal to 2, even if the SSVSξ approach makes a slightly clearer distinction with respect to the classic SSVS method.

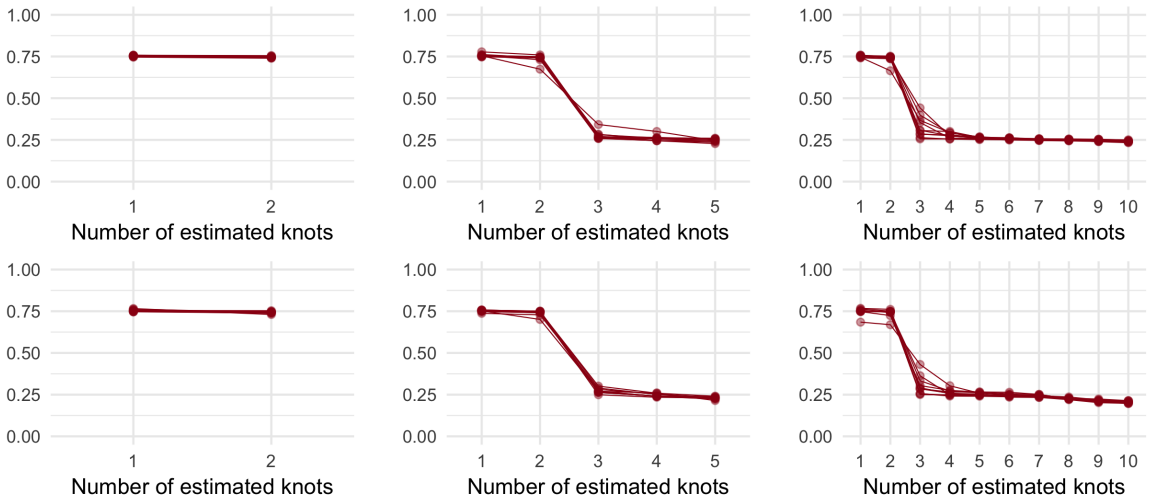

FIGURE 2.2: Posterior means of the mixing parameters $\boldsymbol{\lambda}$ in the overparameterised SSVS (first row) and SSVSξ (second row) models with 2 (left column), 5 (middle column) and 10 (right column) estimated knots. We run 10 chains with $1,000$ iterations in the sampling step. Each line represents a chain.

The second step of the procedure is to estimate the models with the selected number of knots.

The three models are compared by means of diagnostic tools, such as trace plots, $\hat{R}$, effective sample size (n_{eff}) and analysis of marginal posterior distributions. In Table 2.1 we report estimates for the SSVSξ model, parameter estimates are close to the true parameter values. The greatest discrepancies among the model results are on the order of one decimal point. For the three models, $\hat{R}$ statistics equal to 1 suggest that the chains show good mixing, but differences in the neff estimates highlight a lower estimate stability of SSVS model compared to the other two fitted models. According to this limited evidence, SSVSξ approach should be chosen to perform the proposed procedure to estimate the number and location of the knots. Among the three tested models, SSVSξ gives us the best results in terms of estimation of the parameters and in terms of convergence of the algorithm.

Parameter	true	mean	sd	2.5%	50%	97.5%	$\hat{R}$	$n_{eff}^{SSVS\xi}$	n_{eff}^{SSVS}	n_{eff}^{NVS}
β_0	6	6.6	0.6	5.4	6.6	7.6	1.0	4,644	762	3,313
β_1	2	2.1	0.4	1.3	2.1	3.1	1.0	3,039	667	2,171
γ_1	-5	-4.5	0.7	-5.8	-4.5	-3.3	1.0	2,290	287	3,468
γ_2	8	7.4	0.6	6.3	7.3	8.5	1.0	2,704	410	2,690
ξ_1	2.7	2.4	0.2	1.9	2.4	2.8	1.0	3,183	3,105	2,066
ξ_2	4.3	4.4	0.1	4.2	4.4	4.5	1.0	4,634	597	5,196
λ_1		0.7	0.3	0.1	0.8	1.0	1.0	9,387	3,717	
λ_2		0.7	0.3	0.1	0.8	1.0	1.0	8,299	7,027	

TABLE 2.1: Posterior distributions of the SSVSξ model parameters of the model simulated choosing the true number of knots (2). The model is estimated with the true number of knots. $\hat{R}$ and n_{eff} statistics for the three models. Discrepancies among model results are on the order of one decimal point. Substantial differences are detected between the n_{eff} statistics of the three models.

2.3.1 Simulation study

We decided to test performance of our approach fitting linear models on simulated data with different signal to noise ratio. Signal to noise ratio is defined as the variance of the signal over the variance of the random error (Friedman *et al.*, 2001). The lower the signal to noise ratio, the higher the noise level in the data and the more difficult is to detect the knots.

The method is tested on nine synthetic datasets generated with different combinations of number of knots, respectively 0, 2 and 5, and varying levels of signal to noise ratio from low to high (Fig. 2.3). Synthetic data are simulated from the linear regression models

$$
\begin{aligned}
y_i &= 6 + 2x_i + \varepsilon_i, \\
y_i &= 6 + 2x_i - 5(x_i - 2.7)_+ + 8(x_i - 4.3)_+ + \varepsilon_i, \\
y_i &= 6 + 2x_i - 6(x_i - 2.2)_+ + 7(x_i - 2.9)_+ - 3(x_i - 3.8)_+ \\
&\quad + 4(x_i - 4.4)_+ - 3(x_i - 5.3)_+ + \varepsilon_i,
\end{aligned}
$$

for $i = 1, \ldots, 500$, where $\varepsilon_i \overset{\text{ind}}{\sim} N(0, \sigma_\varepsilon)$ and the standard deviation σ_ε is determined according to the chosen level of signal to noise ratio. The predictor $\mathbf{X}$ is always defined on the interval $[0; 10]$. Parameters of the prior distributions are defined as in Section 2.3.

Simulated datasets are shown in Figure 2.3 with increasing combinations of number of knots, by row, and increasing signal to noise ratio, by column. The number of knots used to simulate the data is zero for the three plots in the first row, two for the plots in the second row and five for the plots in the third row. Plots are ordered by column

with an increasing level of signal to noise ratio. For each linear model, we simulated 10 chains of 2,000 iterations estimating 2, 5 and 10 knots.

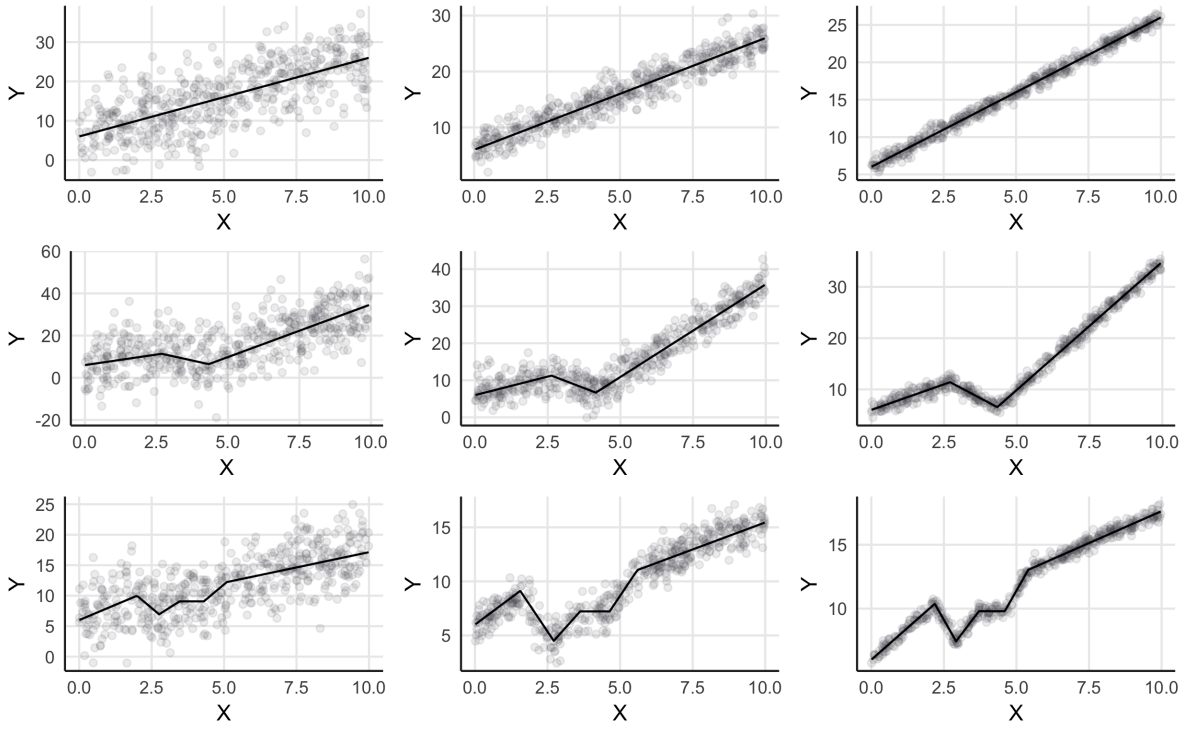

FIGURE 2.3: Simulated datasets and true conditional mean (black line) with increasing combinations of number of knots, by row, and signal to noise ratio, by column. The number of knots used to simulate the data is zero (first row), two (second row) and five (third row). The lower the signal to noise ratio (first column), the higher the noise level in the data (third column).

Figures 2.4–2.6 show results of the fitted SSVSξ models using data with low signal to noise ratio level (first column of Fig. 2.3). The number of estimated knots is correct for the model with no true knots (first row of Fig. 2.4), even when the model is heavily overparameterised. For an increasing number of true knots in the model (second and third row of Fig. 2.4), the posterior means of the mixing parameters $\boldsymbol{\lambda}$ suggest to chose one or two knots in the model fitted on the data with two true knots, and one or zero for the model fitted on the data with five true knots. An inspection of the posterior distributions of the knots location parameters confirms number of knots suggested by the posterior means of $\boldsymbol{\lambda}$ in the case of no true knots (first row of Fig. 2.5). In the other two cases the evidence supports the option with few knots, respectively one and zero in the two and five true knots cases. The peak close to the boundary of the predictor range is associated, by construction, to a mixing parameter $\boldsymbol{\lambda}$ concentrated on zero.

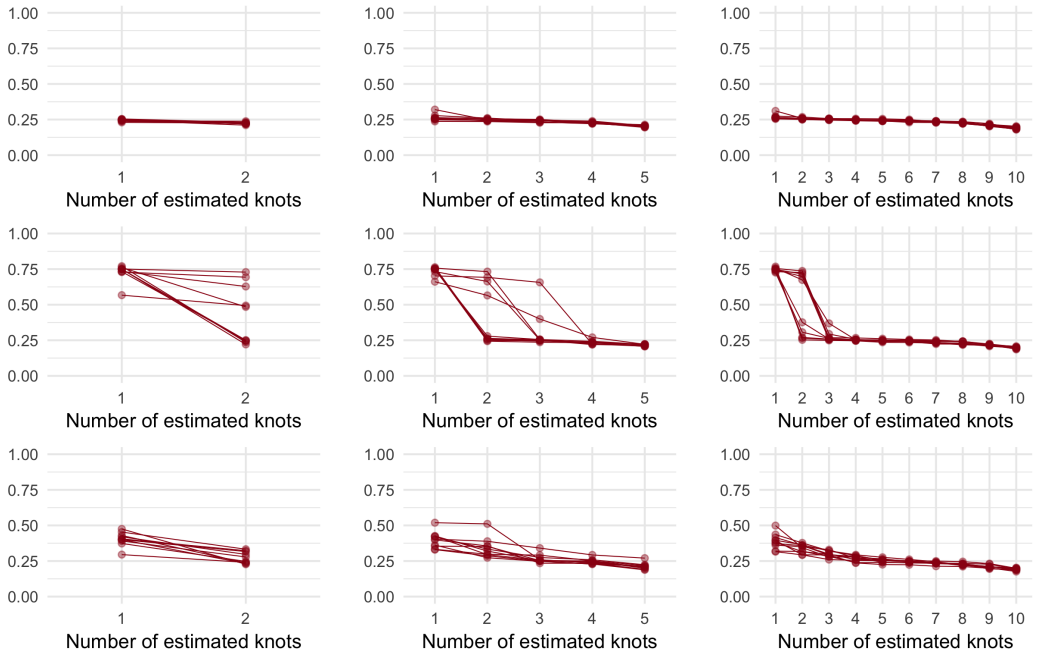

FIGURE 2.4: Posterior means of the mixing parameters $\boldsymbol{\lambda}$ in the overparameterised SSVSξ models with a low signal to noise ratio. The number of estimated knots increases by column, while the number of knots in the simulated data increases by row, respectively zero knots, 2 knots, and 5 knots. Each line represents a chain.

Figure 2.6 shows that approximations to the true signals (black lines) are good also for the overparameterised models. The five true knots case, represented in the last row, highlights more difficulty in the estimation of the underlying signal. However, it is important to stress that this is the most arduous setting for our method characterised by noisy data and a high number of knots closely placed to each other. Indeed, looking at the data, it is hard to recognise both number and location of knots even knowing where they are.

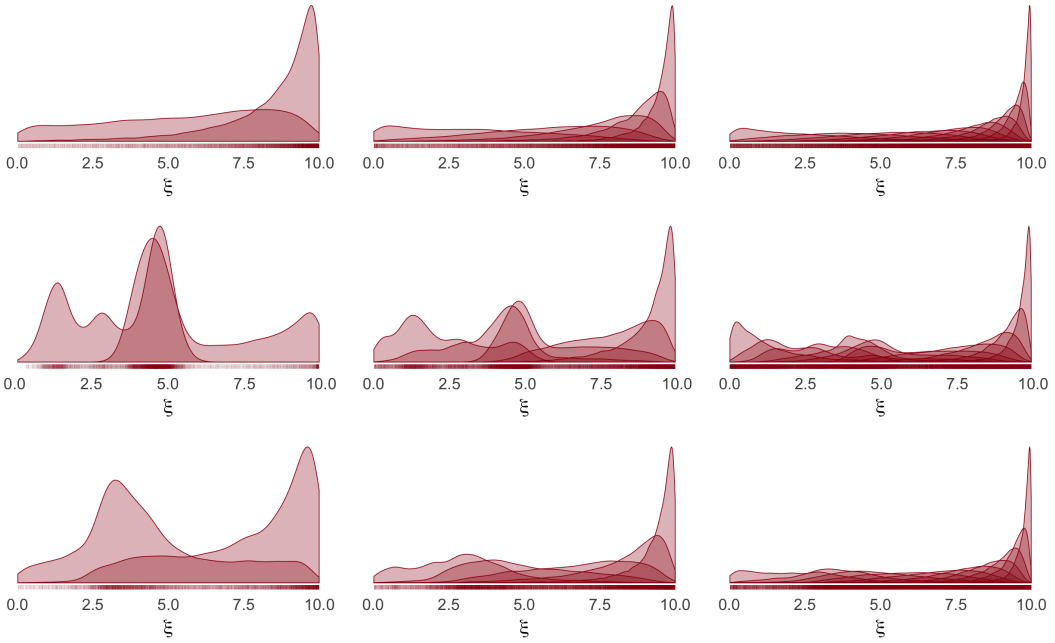

FIGURE 2.5: Posterior distributions of the knot location parameters $\boldsymbol{\xi}$ in the overparameterised SSVSξ models with a low signal to noise ratio. The number of estimated knots increases by column, while the number of knots in the simulated data increases by row, respectively zero knots, 2 knots, and 5 knots. The rug drawn along the axis highlights the highest density regions the chains visited.

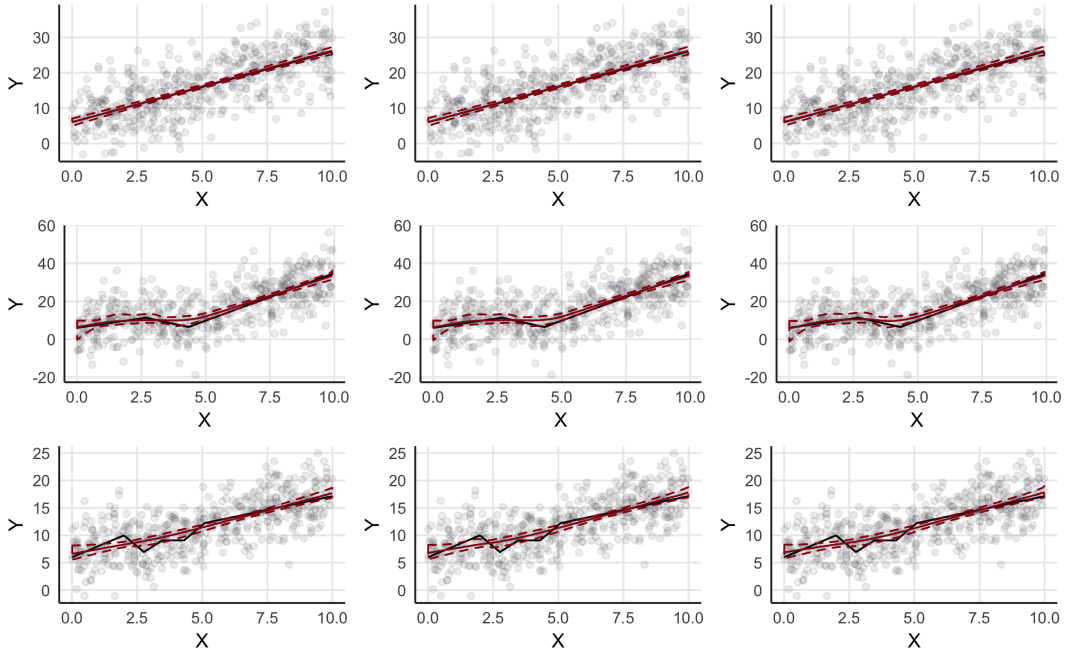

FIGURE 2.6: Fitted conditional mean (red solid line) with 95% credible interval (red dashed 95% lines). The true conditional mean is represented by a black line and data are characterised by a low signal to noise ratio. The overparameterised SSVSξ models are estimated with 2 (first column), 5 (second column) and 10 knots (third column). The number of true knots increases by row, respectively zero knots, 2 knots, and 5 knots.

Figures 2.7–2.9 show results of the fitted SSVSξ models using data with moderate signal to noise ratio level (second column of Fig. 2.3). The number of estimated knots is correct for the model with no true knots (first row of Fig. 2.7) for all the 10 simulated chains. For a number of true knots in the model equal to two (second row of Fig. 2.7), the posterior means of the mixing parameters $\boldsymbol{\lambda}$ clearly suggests to chose two knots in the model fitted on the data with two true knots. When the number of true knots increases to five, the posterior means of the mixing parameters $\boldsymbol{\lambda}$ suggest to select two or three knots (third row of Fig. 2.7). An inspection of the posterior distributions of the knots location parameters confirms the number of knots suggested by the posterior means of $\boldsymbol{\lambda}$ in the case of zero and two true knots (Fig. 2.8). In the last case, the evidence supporting the third knots is more tenuous if compared to one of the first two knots. However, since the third knot is not close to the boundaries of the predictor range, we suggest to select three knots and compare the results with those obtained from the model estimated with two knots.

Figure 2.9 shows that approximations to the true signals (black solid lines) are good. In the models estimated using synthetic data with five knots (third row of Fig. 2.9) the fitted line is closer to the true signal for the overparameterised models.

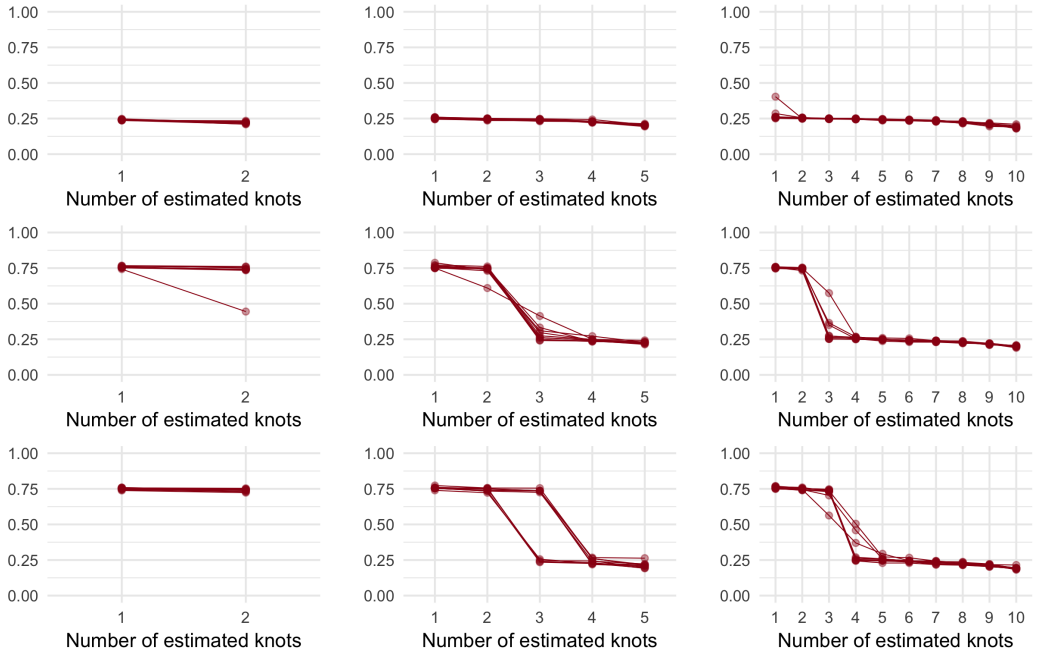

FIGURE 2.7: Posterior means of the mixing parameters $\boldsymbol{\lambda}$ in the overparameterised SSVSξ models with a moderate signal to noise ratio. The number of estimated knots increases by column, while the number of knots in the simulated data increases by row, respectively zero knots, 2 knots, and 5 knots. Each line represents a chain.

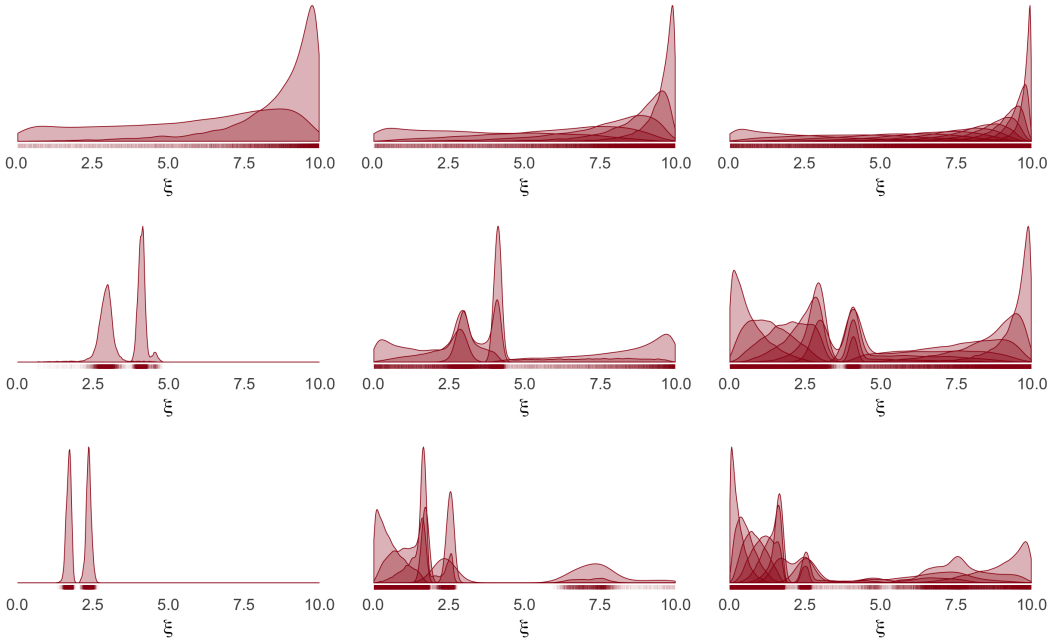

FIGURE 2.8: Posterior distributions of the knot location parameters $\boldsymbol{\xi}$ in the over-parameterised SSVSξ models with a moderate signal to noise ratio. The number of estimated knots increases by column, while the number of knots in the simulated data increases by row, respectively zero knots, 2 knots, and 5 knots. The rug drawn along the axis highlights the highest density regions the chains visited.

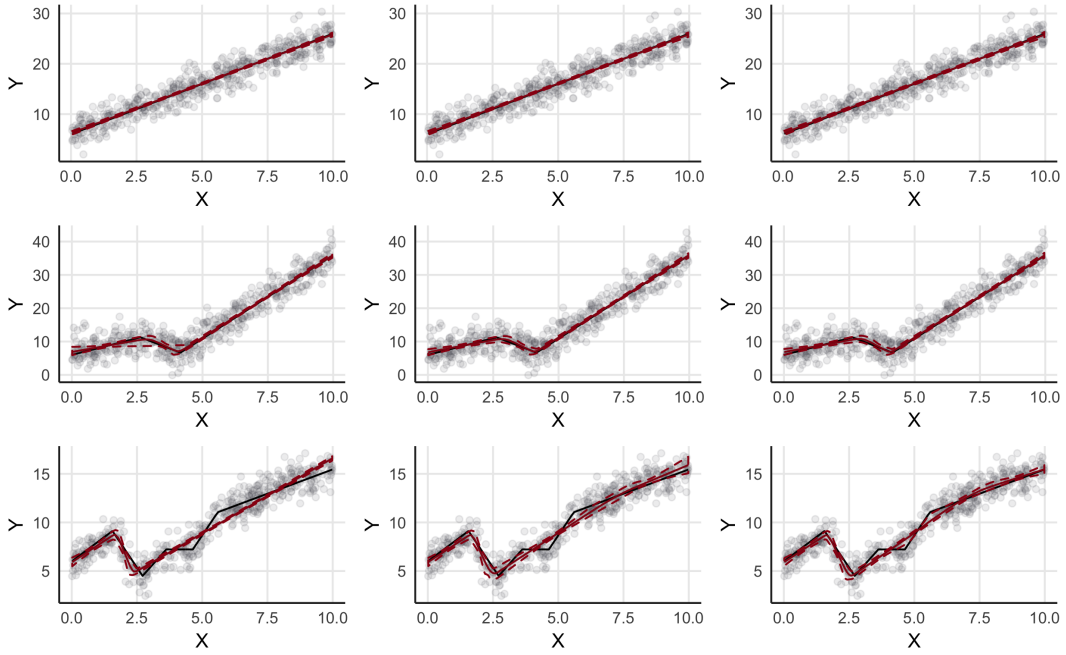

FIGURE 2.9: Fitted conditional mean (red solid line) with 95% credible interval (red dashed 95% lines). The true conditional mean is represented by a black line and data are characterised by a moderate signal to noise ratio. The overparameterised SSVSξ models are estimated with 2 (first column), 5 (second column) and 10 knots (third column). The number of true knots increases by row, respectively zero knots, 2 knots, and 5 knots.

Lastly, Figures 2.10–2.12 show results of the fitted SSVSξ models using data with high signal to noise ratio level (third column of Fig. 2.3). The number of estimated knots coincides with the true number of knots in the data for the models with no and two true knots (first and second row of Fig. 2.10). In the model fitted on the data with five true knots, the evidence in favour of three knots is weaker than in the previous scenario with lower signal to noise ratio level. This is confirmed also by the inspection of the posterior distributions of the knots location parameters.

Figure 2.12 shows that approximations to the true signals (black solid lines) are good also in this case. Even when the signal to noise ratio is high, the model is not able to estimate the true five knots locations and number. Partially, it may be due to the non orthogonality of the truncated linear basis. Nevertheless, the method is able to find the two more evident knots, so an in-depth analysis of the strength of the change points evidence may be interesting.

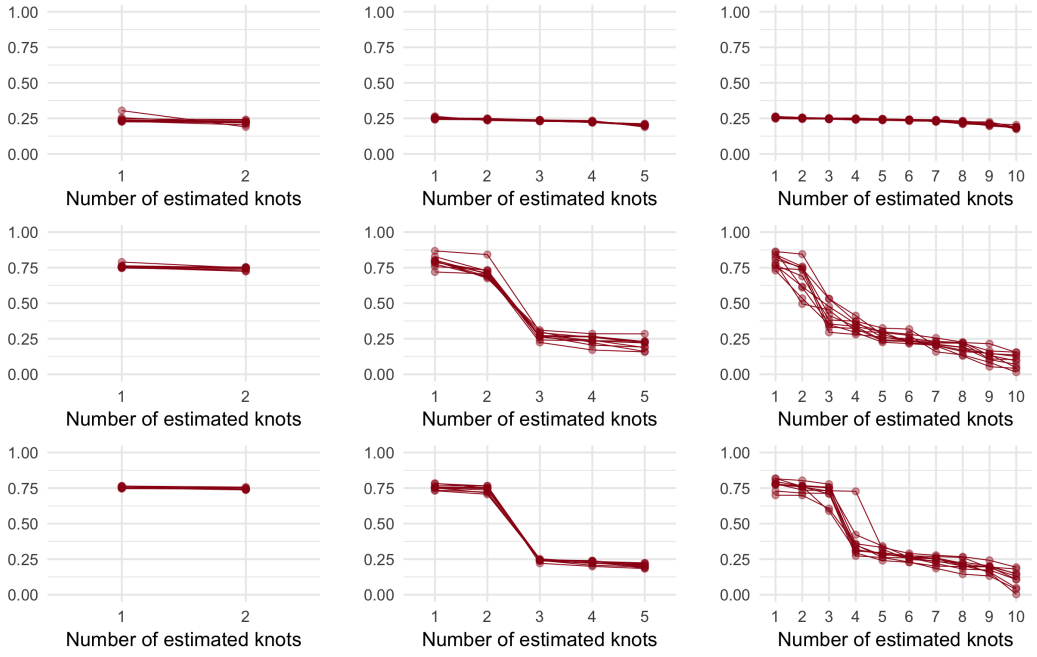

FIGURE 2.10: Posterior means of the mixing parameters $\boldsymbol{\lambda}$ in the overparameterised SSVSξ models with a high signal to noise ratio. The number of estimated knots increases by column, while the number of knots in the simulated data increases by row, respectively zero knots, 2 knots, and 5 knots. Each line represents a chain.

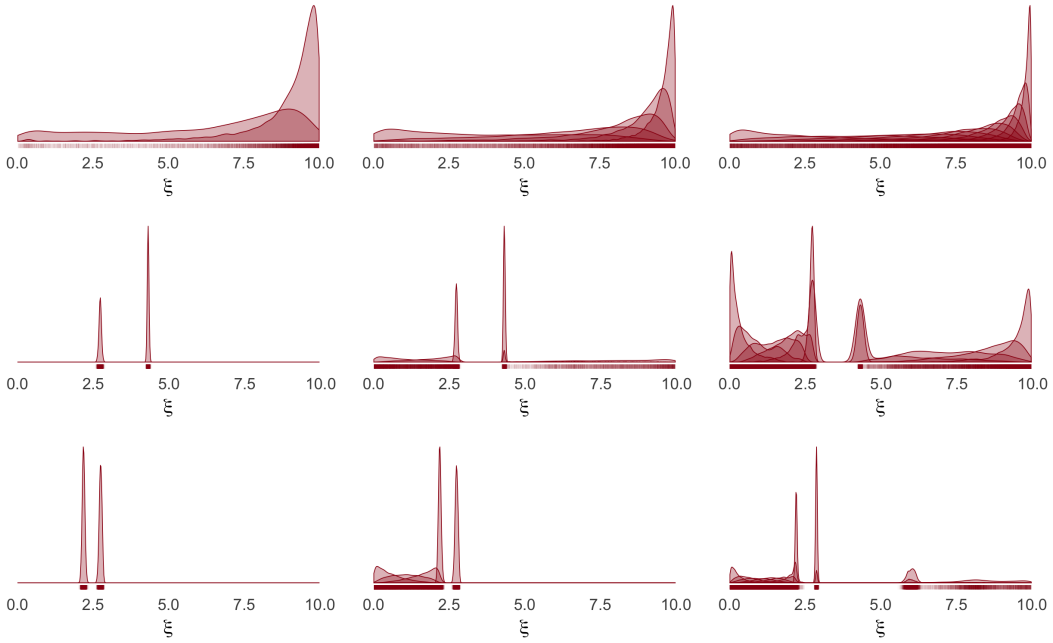

FIGURE 2.11: Posterior distributions of the knot location parameters $\boldsymbol{\xi}$ in the overparameterised SSVSξ models with a high signal to noise ratio. The number of estimated knots increases by column, while the number of knots in the simulated data increases by row, respectively zero knots, 2 knots, and 5 knots. The rug drawn along the axis highlights the highest density regions the chains visited.

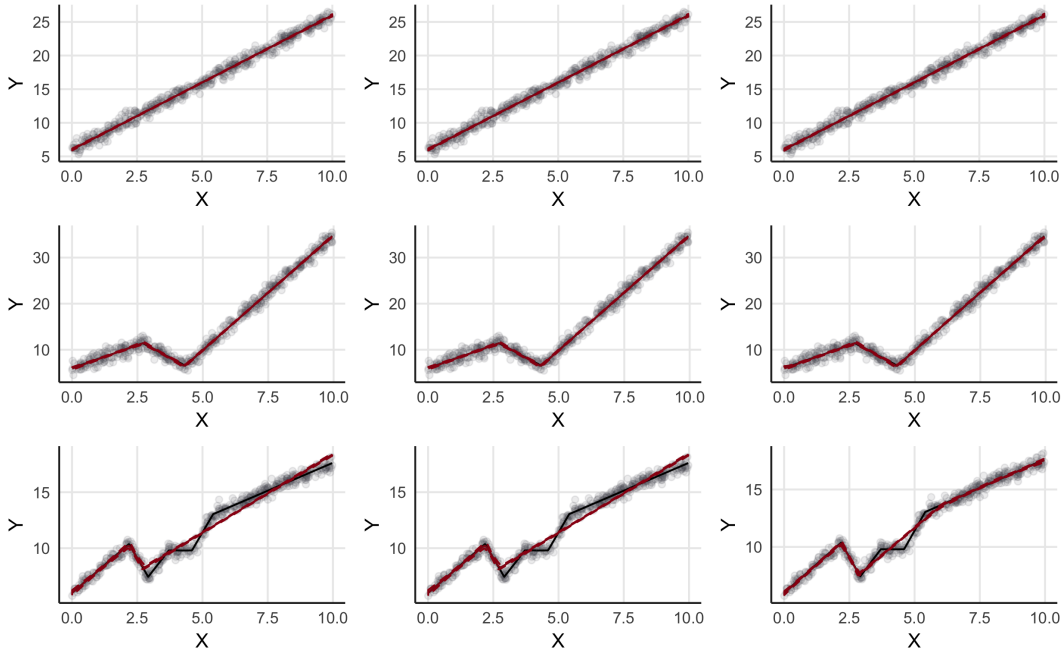

FIGURE 2.12: Fitted conditional mean (red solid line) with 95% credible interval (red dashed 95% lines). The true conditional mean is represented by a black line and data are characterised by a high signal to noise ratio. The overparameterised SSVSξ models are estimated with 2 (first column), 5 (second column) and 10 knots (third column). The number of true knots increases by row, respectively zero knots, 2 knots, and 5 knots.

Once the number of knots has been selected the final model can be fitted with the most appropriate specification and methodology. Knots locations emerged in the posterior distribution of the knots locations, can be used as initial values for optimisation algorithms, e.g. the ones used in the frequentist approaches (Sect. 1.3.1).

2.4 Application

In this section we apply the proposed SSVSξ methodology to estimate the semiparametric logistic model presented in Section 1.4 using data on cigarettes smoking habits and larynx cancer from the INHNACE Consortium described in Section 1.4.1.2.

2.4.1 Bivariate extension of the SSVSξ model

We specify a semiparametric logistic model for the Bernoulli-distributed random variable $\mathbf{Y}$, as in Equation 1.1

$$\text{logit}\left(\pi_i\right) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{z}_i\boldsymbol{\alpha} + f\left(x_i, w_i\right), \quad \text{for } i = 1, \ldots, n.$$

Adjustment variables $\mathbf{Z}$ enter linearly in the model. The variables $\mathbf{X}$ intensity and $\mathbf{W}$ duration of cigarettes smoking enter in the model as a bivariate linear spline function represented through the truncated linear basis as in Equation 1.2

$$\begin{aligned} f(\mathbf{x}, \mathbf{w}) =& \beta_0 + \beta_1\mathbf{x} + \beta_2\mathbf{w} + \beta_3\mathbf{x}\mathbf{w} + \\ & \sum_{k_x=1}^{K_x} \gamma_{k_x}(\mathbf{x} - \xi_{k_x})_+ + \sum_{k_w=1}^{K_w} \gamma_{k_w}(\mathbf{w} - \xi_{k_w})_+ + \\ & \sum_{k_x=1}^{K_x} \gamma_{2,k_x}(\mathbf{x} - \xi_{k_x})_+\mathbf{w} + \sum_{k_w=1}^{K_w} \gamma_{2,k_w}(\mathbf{w} - \xi_{k_w})_+\mathbf{x} + \\ & \sum_{k_x=1}^{K_x}\sum_{k_w=1}^{K_w} \gamma_{3,k_x,k_w}(\mathbf{x} - \xi_{k_x})_+(\mathbf{w} - \xi_{k_w})_+. \end{aligned}$$

Prior distributions on the knots positions are defined as before as Uniform on the range of the related predictor, subject to ordered constraint

$$\xi_{k_x} \sim Unif(\min(\mathbf{x}), \max(\mathbf{x})), \quad \text{subject to} \quad \xi_{k_x} \leq \xi_{k_x+1}, \quad \text{for } k_x = 1, \ldots, K_x,$$

and

$$\xi_{k_w} \sim Unif(\min(\mathbf{w}), \max(\mathbf{w})), \quad \text{subject to} \quad \xi_{k_w} \leq \xi_{k_w+1}, \quad \text{for } k_w = 1, \ldots, K_w.$$

Prior distributions on the regression coefficients $\boldsymbol{\alpha}$ and on the spline coefficients $\boldsymbol{\beta}$ are defined as weakly informative

$$\boldsymbol{\alpha} \overset{\text{ind}}{\sim} t(3, 10), \qquad \boldsymbol{\beta} \overset{\text{ind}}{\sim} t(3, 2.5),$$

while prior distributions on the spline coefficients $\boldsymbol{\gamma}$ and the prior distributions on the mixing parameters $\boldsymbol{\lambda}$ change in

$$\begin{aligned}
\pi(\gamma_{k_x}|\lambda_{k_x}) &= \lambda_{k_x} N(0, 100) + (1 - \lambda_{k_x}) N(0, 0.1), & \lambda_{k_x}|\xi_{k_x} &\sim Beta(0.5, b_{k_x}),\\
\pi(\gamma_{k_w}|\lambda_{k_w}) &= \lambda_{k_w} N(0, 100) + (1 - \lambda_{k_w}) N(0, 0.1), & \lambda_{k_w}|\xi_{k_w} &\sim Beta(0.5, b_{k_w}),\\
\pi(\gamma_{2,k_x}|\lambda_{2,k_x}) &= \lambda_{2,k_x} N(0, 100) + (1 - \lambda_{2,k_x}) N(0, 0.1), & \lambda_{2,k_x}|\xi_{k_x} &\sim Beta(0.5, b_{k_x}),\\
\pi(\gamma_{2,k_w}|\lambda_{2,k_w}) &= \lambda_{2,k_w} N(0, 100) + (1 - \lambda_{2,k_w}) N(0, 0.1), & \lambda_{2,k_w}|\xi_{k_w} &\sim Beta(0.5, b_{k_w}),
\end{aligned}$$

and

$$\begin{aligned}
\pi(\gamma_{3,k_x,k_w}|\lambda_{3,k_x,k_w}) &= \lambda_{3,k_x,k_w} N(0, 100) + (1 - \lambda_{3,k_x,k_w}) N(0, 0.1),\\
\lambda_{3,k_x,k_w}|\xi_{k_x,k_w} &\sim Beta(0.5, \min(b_{k_x}, b_{k_w})),
\end{aligned}$$

where $b_{k_x} : [\min(\mathbf{x}); \max(\mathbf{x})] \to [0.5; 1.5]$ and $b_{k_w} : [\min(\mathbf{w}); \max(\mathbf{w})] \to [0.5; 1.5]$.

Due to the high number of parameters that have to be estimated and the dimension of the dataset, we chose to test the bivariate extension of the SSVSξ methodology fixing the number of estimated knots to two on both risk factors. We run 10 chains with $2,000$ iterations each. Initial values for the knot location parameters are chosen uniformly spread on the linked predictor range, while regression and spline parameters are initialised at zero. The mixing parameters are initialised at 0.9 to support complete exploration of the posterior distribution.

Results shown in Figure 2.13 and Figure 2.14 support the choices made using information criteria presented in Section 1.4.2. Indeed, there is evidence in favour of one knot both for intensity and duration of cigarettes smoking variables (Fig. 2.13). Higher uncertainty in the mixing parameter for the duration variable (left, Fig. 2.13) is related to the higher variability observed in the trace plot of the knots location (Fig. 1.1) if compared with the intensity variable results. As Rosenberg *et al.* (2003) pointed out, regression splines are locally sensitive to the data and we can take advantage of this characteristic to better understand analysed data. Posterior distributions of the knot locations also confirm previous results (Fig. 2.14). In particular, the knot location for the intensity variable was estimated at 25 cigarettes/day, while the knot location for duration variable at 30 years of cigarettes smoking, as highlighted by the rug along the

x-axis.

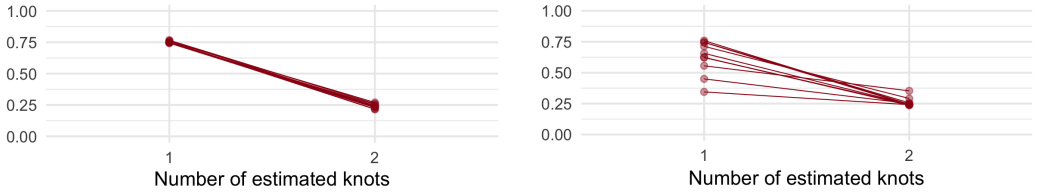

FIGURE 2.13: Posterior means of the mixing parameters λ_x (left) and λ_w (right) in the overparameterised SSVSξ bivariate model using larynx data from INHANCE Consortium. Each line represents a chain.

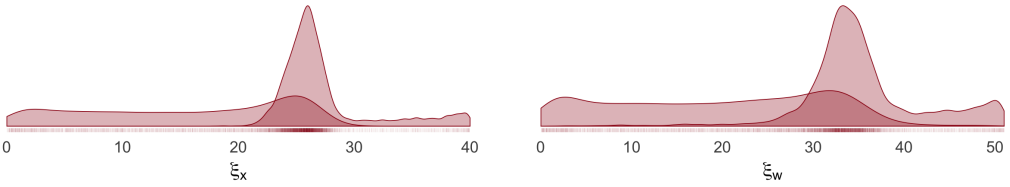

FIGURE 2.14: Posterior distributions of the knot location parameters ξ_x for intensity, cigarettes/day, (left) and ξ_w duration, years of cigarettes smoking, (right) in the overparameterised SSVSξ bivariate model using larynx data from INHANCE Consortium. The rug drawn along the axis highlights the highest density regions the chains visited.

Figure 2.15 show the fitted surface using $10,000$ draws from the overparameterised model with two knots on intensity and two knots on duration variables. Posterior means of the knots location parameters are 16 and 28 cigarettes/day and 19 years and 35 of duration. Comparing the perspective plot and the contour plot with the ones in Figure 1.6, we notice that estimated surfaces are very similar apart from small differences on boundary regions. However, posterior inference based on simulations from the overparameterised model leads to unreliable results for the parameters of interest.

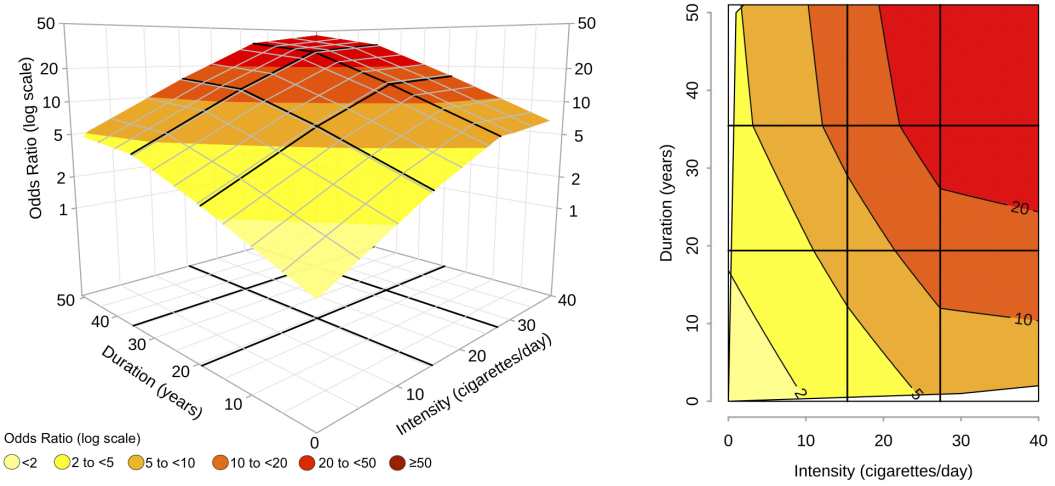

FIGURE 2.15: Current smokers - larynx site. The surface is estimated through the overparameterised SSVSξ model with two knots on intensity and two knots on duration variables. On the grid, black thicker lines represent knot locations: 16 and 28 cigarettes/day and 19 years and 35 of duration for larynx cancer. Dark grey lines in contour plots indicate iso-risk curves at defined levels of risk.

2.5 Discussion

The proposed methodology aims to solve the problem of estimating the number and positions of knots in semiparametric regression models with linear splines. A well-known variable selection technique has been adapted in order to estimate the presence or absence of knots in possible overparameterised models. Once that the number of knots is selected, the appropriate model can be fitted with the preferred technique. Moreover, the method gives us a first guess on the knot locations through the inspection of the marginal posterior distributions of the knots location parameters. This can be useful in the initialisation step of algorithms with difficulties in exploring entirely the parameter space, especially in high dimensional problems.

In terms of computational complexity, this methodology requires a higher number of parameters to be estimated if compared with one model as specified in the NVS approach or, equivalently in the bivariate case, in Section 1.3.2. Consequently, it is more time consuming. On the other hand, this approach requires the estimate of only one model to select the number on knots, while the NVS approach needs estimating a possibly large number of models, especially in the bivariate case, and, moreover, is based on information criteria which have been criticised in the statistical literature (Piironen and Vehtari, 2017).

Lastly, in order to compute the WAIC or LOO criteria we need to include additional steps in the simulations. This has two main consequences: a negligible increment in the

sampling in terms of time, but a remarkable increase of the memory needed to store the simulations. As an example we refer to the larynx data, the fitted model in Section 1.3.2 with one knot on each risk factor was about 2Gb versus 8Mb of the model estimated with the SSVSξ approach.

A deeper examination of the effects of the spike and slab priors on the spline coefficients of the model is of primary interest for future works. Moreover, it will be also considered the possibility to make inference based on the results of the first step of the SSVSξ methodology for descriptive purposes of the phenomenon in analysis.

The proposed simulation study is based on a simple linear regression model that allows us to stress the methodology by focusing attention on some of its critical aspects. In particular, we decided to test if the method is able to estimate the correct number of knots even if they are many and close together or when the signal-to-noise ratio is very low. The methodology is designed for situations in which the number of expected slope changes is limited, e.g. reasonable in many epidemiological studies. When we expect an high number of knots, this methodology may be not appropriate, but despite this, the knots corresponding to the most evident slope changes are correctly identified.

The methodology was also tested on the semiparametric logistic model with bivariate spline proposed in the first chapter using real data. In this case, the model structure is much more complex if compared with the one of the simulation study. The method confirms the number and location of the knots identified by the selection through information criteria. Changes in the slope of the surface are estimated together with the other parameters and this represents the main innovative result of our methodology, taking also into account that it can handle the high dimensionality of the problem.

Chapter 3

Streamlined inference for generalised linear models with crossed random effects

3.1 Introduction

Multilevel or hierarchical models are a generalisation of the linear and generalised linear models in which group structure of the data can be taken into account in the estimation process. Fitting a model with pooled data, ignoring the group structure, does not allows considering variation between groups. Assuming a probabilistic distribution on some regression coefficients, multilevel models estimate an average regression line and group-level variances. The varying coefficients, intercept and/or the slopes, are usually called random effects, while the others coefficients are called fixed effects.

Hierarchical models can handle several group structure of the data, e.g. nested or crossed structure. A classical example of two-level nested structure is represented by students within classes, but also higher hierarchy levels can be considered (Gelman and Hill, 2006). Crossed structure occurs, for example, when a questionnaire is submitted to several subjects. A detailed introduction on crossed random effects models is provided by Baayen *et al.* (2008).

When the number of groups increases or when nonparametric extensions on the group-specific curve are required, the estimation process can easily become slow or unfeasible.

The streamlined variational inference has recently been applied to longitudinal/multilevel model to overcome these estimation difficulties. In the literature, this technique has been applied on nested data structure, e.g. in Lee and Wand (2016a,b); Jeon *et al.*

(2017), we will explore the crossed one.

In the following sections, we present the solution to the inversion of a sparse multilevel matrix problem for the crossed random effects case. Later, we consider a mean field variational Bayes approach to fit the model and make inference on the fixed and random effects coefficients. After the definition of the model under the Bayesian model assumptions, we compute the mean field variational Bayes approximation to the posterior distribution. Lastly, we compare, in terms of computational time, the streamlined mean field algorithm with the Hamiltonian Monte Carlo (HMC) procedure.

3.2 Frequentist approach

3.2.1 Gaussian crossed random effects model

Under the frequentist assumption, the general form of the Gaussian response linear mixed model with two crossed random effects is

$$\mathbf{y}_{i'i}|\mathbf{u}'_{i'}, \mathbf{u}_i \overset{\text{ind}}{\sim} N(\mathbf{X}_{i'i}\boldsymbol{\beta} + \mathbf{Z}'_{i'i}\mathbf{u}'_{i'} + \mathbf{Z}_{i'i}\mathbf{u}_i, \mathbf{R}),$$

$$\mathbf{u}'_{i'} \sim N(\mathbf{0}, \boldsymbol{\Sigma}'), \quad \mathbf{u}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \quad \text{for } i' = 1, \ldots, m', \quad i = 1, \ldots, m,$$

where $\mathbf{y}_{i'i}$ is the vector $n_{i'i} \times 1$ of the continuous response variable, $\mathbf{X}_{i'i}$ is the $n_{i'i} \times p$ design matrix, $\mathbf{Z}'_{i'i}$ and $\mathbf{Z}_{i'i}$, respectively of dimension $n_{i'i} \times q'$ and $n_{i'i} \times q$, are the random effects matrices, $\boldsymbol{\beta}$ is the vector $p \times 1$ of the fixed effect coefficients, $\mathbf{u}'_{i'}$ and $\mathbf{u}_i$, respectively of dimension $q' \times 1$ and $q \times 1$, are the vectors of the random effects coefficients, $\mathbf{R} = \sigma^2_\varepsilon \mathbf{I}$ is the variance of the noise and $\boldsymbol{\Sigma}'$ and $\boldsymbol{\Sigma}$ are the $q' \times q'$ and $q \times q$ covariance matrices of the random effects.

Best linear unbiased predictor (BLUP) is used to estimate the vector of the random effects parameters and its covariance matrix

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}' \\ \hat{\mathbf{u}} \end{bmatrix} = (\mathbf{C}^\intercal \mathbf{R}^{-1} \mathbf{C} + \mathbf{D})^{-1} \mathbf{C}^\intercal \mathbf{R}^{-1} \mathbf{y},$$

$$\mathbb{C}ov \left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}' - \mathbf{u}' \\ \hat{\mathbf{u}} - \mathbf{u} \end{bmatrix} \right) = \mathbf{A}^{-1} = (\mathbf{C}^\intercal \mathbf{R}^{-1} \mathbf{C} + \mathbf{D})^{-1},$$

where the matrix $\mathbf{C} = [\mathbf{X}, \mathbf{Z}', \mathbf{Z}]$ is composed of the design matrix $\mathbf{X}$ and the random effects matrices $\mathbf{Z}'$ and $\mathbf{Z}$. For a generic number of groups for the two random effects, the matrix $\mathbf{C}$ is an $n \times (p + m'q' + mq)$ matrix. Let the number of groups for each random effect be $m' = 2$ and $m = 3$, then the $\mathbf{C}$ matrix can be rearranged to have the following structure

$$\mathbf{C} = \left[\begin{array}{c|cc|ccc} \mathbf{X}_{11} & \mathbf{Z}'_{11} & \mathbf{0} & \mathbf{Z}_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{X}_{21} & \mathbf{0} & \mathbf{Z}'_{21} & \mathbf{Z}_{21} & \mathbf{0} & \mathbf{0} \\ \mathbf{X}_{12} & \mathbf{Z}'_{12} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_{12} & \mathbf{0} \\ \mathbf{X}_{22} & \mathbf{0} & \mathbf{Z}'_{22} & \mathbf{0} & \mathbf{Z}_{22} & \mathbf{0} \\ \mathbf{X}_{13} & \mathbf{Z}'_{13} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_{13} \\ \mathbf{X}_{23} & \mathbf{0} & \mathbf{Z}'_{23} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_{23} \end{array}\right].$$

Lastly, the matrix $\mathbf{D} = \text{diag}(\mathbf{0}, \boldsymbol{\Sigma}', \boldsymbol{\Sigma})$ is the block diagonal matrix composed of the covariance matrices of the random effects.

3.2.1.1 General linear system solution

The block symmetric matrix $\mathbf{A}$, of dimension $(p + q'm' + qm) \times (p + q'm' + qm)$, is characterised by a sparse structure

$$\mathbf{A} = \left[\begin{array}{c|ccc|ccc} \mathbf{A}_{11} & \mathbf{A}'_{12,1} & \dots & \mathbf{A}'_{12,m'} & \mathbf{A}_{12,1} & \dots & \mathbf{A}_{12,m} \\ \hline \mathbf{A}'^{\intercal}_{12,1} & \mathbf{A}'_{22,1} & 0 & 0 & \mathbf{M}_{11} & \dots & \mathbf{M}_{1m} \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} & \vdots & \ddots & \vdots \\ \mathbf{A}'^{\intercal}_{12,m'} & 0 & 0 & \mathbf{A}'_{22,m'} & \mathbf{M}_{m'1} & \dots & \mathbf{M}_{m'm} \\ \hline \mathbf{A}^{\intercal}_{12,1} & \mathbf{M}^{\intercal}_{11} & \dots & \mathbf{M}^{\intercal}_{m'1} & \mathbf{A}_{22,1} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{A}^{\intercal}_{12,m} & \mathbf{M}^{\intercal}_{1m} & \dots & \mathbf{M}^{\intercal}_{m'm} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{22,m} \end{array}\right].$$

More in detail,

$$\mathbf{A}_{11} = \frac{1}{\sigma^2_\varepsilon} \sum_{i=1}^{m} \sum_{i'=1}^{m'} \mathbf{X}^{\intercal}_{i'i} \mathbf{X}_{i'i},$$

$$\mathbf{A}'_{12,i'} = \frac{1}{\sigma^2_\varepsilon} \sum_{i=1}^{m} \mathbf{X}^{\intercal}_{i'i} \mathbf{Z}'_{i'i} \quad \text{and} \quad \mathbf{A}'_{22,i'} = \frac{1}{\sigma^2_\varepsilon} \sum_{i=1}^{m} \mathbf{Z}'^{\intercal}_{i'i} \mathbf{Z}'_{i'i} + \boldsymbol{\Sigma}'_{i'}, \quad \text{for } i' = 1, \dots, m',$$

$$\mathbf{A}_{12,i} = \frac{1}{\sigma^2_\varepsilon} \sum_{i'=1}^{m'} \mathbf{X}^{\intercal}_{i'i} \mathbf{Z}_{i'i} \quad \text{and} \quad \mathbf{A}_{22,i} = \frac{1}{\sigma^2_\varepsilon} \sum_{i'=1}^{m'} \mathbf{Z}^{\intercal}_{i'i} \mathbf{Z}_{i'i} + \boldsymbol{\Sigma}_i, \quad \text{for } i = 1, \dots, m,$$

$$\mathbf{M}_{i'i} = \frac{1}{\sigma^2_\varepsilon} \mathbf{Z}'^{\intercal}_{i'i} \mathbf{Z}^{\intercal}_{i'i} \quad \text{for } i' = 1, \dots, m', \quad i = 1, \dots, m,$$

where $\mathbf{A}_{11}$ is a symmetric matrix of dimension $p \times p$, each of the $\mathbf{A}'_{12,i'}$, for $i' = 1, \ldots, m'$, block is a matrix of dimension $p \times q'$ and each of the $\mathbf{A}_{12,i}$ for $i = 1, \ldots, m$, block is a matrix of dimension $p \times q$. Each of the $\mathbf{A}'_{22,i'}$, for $i' = 1, \ldots, m'$, symmetric sub-matrix has dimension $q' \times q'$ and each of the $\mathbf{A}_{22,i}$, for $i = 1, \ldots, m$, symmetric sub-matrix has dimension $q \times q$. Each of the $\mathbf{M}_{i',i}$, for $i' = 1, \ldots, m'$ and $i = 1, \ldots, m$, sub-blocks has dimension $q' \times q$.

For most models we are interested only in sub-blocks of the covariance matrix$\mathbf{A}^{-1}$, thus, calculations can be streamlined computing the inverse of the sub-blocks of interest

$$\mathbf{A}^{-1} = \left[\begin{array}{c|ccc|ccc} \mathbf{A}^{11} & \mathbf{A}'^{12,1} & \ldots & \mathbf{A}'^{12,m'} & \mathbf{A}^{12,1} & \ldots & \mathbf{A}^{12,m} \\ \hline \mathbf{A}'^{12,1\intercal} & \mathbf{A}'^{22,1} & \times & \times & \mathbf{M}^{11} & \ldots & \mathbf{M}^{1m} \\ \vdots & \times & \ddots & \times & \vdots & \ddots & \vdots \\ \mathbf{A}'^{12,m'\intercal} & \times & \times & \mathbf{A}'^{22,m'} & \mathbf{M}^{m'1} & \ldots & \mathbf{M}^{m'm} \\ \hline \mathbf{A}^{12,1\intercal} & \mathbf{M}^{11\intercal} & \ldots & \mathbf{M}^{m'1\intercal} & \mathbf{A}^{22,1} & \times & \times \\ \vdots & \vdots & \ddots & \vdots & \times & \ddots & \times \\ \mathbf{A}^{12,m\intercal} & \mathbf{M}^{1m\intercal} & \ldots & \mathbf{M}^{m'm\intercal} & \times & \times & \mathbf{A}^{22,m} \end{array}\right].$$

The non-zero blocks of $\mathbf{A}$ correspond to the sub-blocks of $\mathbf{A}^{-1}$ that are of interest. In particular, from the structure of the $\mathbf{A}^{-1}$ matrix we recognise:

$$\mathbf{A}^{11} = \mathbb{C}ov(\hat{\boldsymbol{\beta}}),$$

$$\mathbf{A}'^{12,i'} = \mathbb{E}[\hat{\boldsymbol{\beta}}(\hat{\mathbf{u}}'_{i'} - \mathbf{u}'_{i'})^\intercal] \quad \text{and} \quad \mathbf{A}'^{22,i'} = \mathbb{C}ov(\hat{\mathbf{u}}'_{i'} - \mathbf{u}'_{i'}), \quad \text{for } i' = 1, \ldots, m',$$

$$\mathbf{A}^{12,i} = \mathbb{E}[\hat{\boldsymbol{\beta}}(\hat{\mathbf{u}}_i - \mathbf{u}_i)^\intercal] \quad \text{and} \quad \mathbf{A}^{22,i} = \mathbb{C}ov(\hat{\mathbf{u}}_i - \mathbf{u}_i), \quad \text{for } i = 1, \ldots, m,$$

$$\mathbf{M}^{i'i} = \mathbb{E}[(\hat{\mathbf{u}}'_{i'} - \mathbf{u}'_{i'})(\hat{\mathbf{u}}_i - \mathbf{u}_i)^\intercal] \quad \text{for } i' = 1, \ldots, m', \quad i = 1, \ldots, m.$$

In the nested model with two levels, the $\mathbf{A}$ matrix can be rearranged in an arrowhead block matrix. This simplifies the computations of the BLUPs allowing the streamlined inversion of the matrix $\mathbf{A}$ (Nolan and Wand, 2018). Indeed, the solution to the sparse linear system $\mathbf{Ax} = \mathbf{a}$, where the $\mathbf{A}$ matrix is defined as an arrowhead block matrix, permits to isolate the results of the sub-blocks of interest reducing the computational complexity of the algorithm. Theorem 1, exposed in Nolan and Wand (2018), can be applied to any arrowhead block matrix and, in particular, it suits the estimation of the covariance matrix in the two-level nested model.

Dealing with crossed effects, the covariance matrix $\mathbf{A}^{-1}$ shows a less sparse structure compared to the one in the nested case. In particular, the presence of the $\mathbf{M}_{i'i}$ sub-blocks introduces dependencies that cannot be easily simplified solving the linked linear

system. This leads us to implement a partial streamlined solution adapting the two-level nested solution to the crossed effects structure.

The key assumption to the streamlined inference results in the crossed random effects model is to keep the number of groups for one random effect, say m', relatively low, while the other number of groups, m, can be extremely large. This is the case, e.g., when we are dealing with a questionnaire with 10 or 20 items submitted to thousands of people. Clearly, for a low number of items, the higher the number of subjects, the more we gain in terms of computational complexity.

In the following simplified representation of the $\mathbf{A}$ matrix, the thick lines reveal an arrowhead structure of the matrix.

$$\mathbf{A} = \left[\begin{array}{c|c|c} \mathbf{A}_{11} & \mathbf{A}'_{12} & \mathbf{A}_{12} \\ \hline \mathbf{A}'^{\intercal}_{12} & \mathbf{A}'_{22} & \mathbf{M} \\ \hline \mathbf{A}^{\intercal}_{12} & \mathbf{M}^{\intercal} & \mathbf{A}_{22} \end{array}\right].$$

Hence, the sub-block of the $\mathbf{A}$ matrix whose inversion can be streamlined is the block-diagonal sub-matrix $\mathbf{A}_{22}$. Indeed, time complexity linearly grows with the increase in the number of groups m.

Thus, we present the streamlined inversion of the sparse matrix $\mathbf{A}$ as the solution to the linear system

$$\left[\begin{array}{c|c|c} \mathbf{A}_{11} & \mathbf{A}'_{12} & \mathbf{A}_{12} \\ \hline \mathbf{A}'^{\intercal}_{12} & \mathbf{A}'_{22} & \mathbf{M} \\ \hline \mathbf{A}^{\intercal}_{12} & \mathbf{M}^{\intercal} & \mathbf{A}_{22} \end{array}\right] \left[\begin{array}{c} \mathbf{x}_1 \\ \hline \mathbf{x}'_2 \\ \hline \mathbf{x}_2 \end{array}\right] = \left[\begin{array}{c} \mathbf{a}_1 \\ \hline \mathbf{a}'_2 \\ \hline \mathbf{a}_2 \end{array}\right],$$

where $\mathbf{x}_1$ and $\mathbf{a}_1$ are $p \times 1$, $\mathbf{x}'_2$ and $\mathbf{a}'_2$ are $q'm' \times 1$ and $\mathbf{x}_2$ and $\mathbf{a}_2$ are $qm \times 1$.

Applying Theorem 1 of Nolan and Wand (2018), the solution set to the streamlined computation of $\mathbf{A}^{-1}$ in the crossed random effects model is given by

$$\left[\begin{array}{c|c} \mathbf{A}^{11} & \mathbf{A}'^{12} \\ \hline \mathbf{A}'^{12\intercal} & \mathbf{A}'^{22} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}'_{12} \\ \hline \mathbf{A}'^{\intercal}_{12} & \mathbf{A}'_{22} \end{array}\right] - \sum_{i=1}^{m} \left[\begin{array}{c} \mathbf{A}_{12,i} \\ \hline \mathbf{M}_{\cdot i} \end{array}\right] \mathbf{A}^{-1}_{22,i} \left[\mathbf{A}^{\intercal}_{12,i} \;\middle|\; \mathbf{M}^{\intercal}_{\cdot i}\right],$$

$$\left[\begin{array}{c} \mathbf{A}^{12,i} \\ \hline \mathbf{M}^{\cdot i} \end{array}\right] = -\left(\mathbf{A}^{-1}_{22,i} \left[\mathbf{A}^{\intercal}_{12,i} \;\middle|\; \mathbf{M}^{\intercal}_{\cdot i}\right] \left[\begin{array}{c|c} \mathbf{A}^{11} & \mathbf{A}'^{12} \\ \hline \mathbf{A}'^{12\intercal} & \mathbf{A}'^{22} \end{array}\right]\right)^{\intercal}, \quad \text{for } i = 1, \ldots, m,$$

and

$$\mathbf{A}^{22,i} = \mathbf{A}^{-1}_{22,i}\left(\mathbf{I} - \left[\mathbf{A}^{\intercal}_{12,i} \;\middle|\; \mathbf{M}^{\intercal}_{\cdot i}\right] \left[\begin{array}{c} \mathbf{A}^{12,i} \\ \hline \mathbf{M}^{\cdot i} \end{array}\right]\right), \quad \text{for } i = 1, \ldots, m.$$

While, the solution set to the linear system is

$$\begin{bmatrix} \mathbf{x}_1 \\ \hline \mathbf{x}'_2 \end{bmatrix} = \left[\begin{array}{c|c} \mathbf{A}^{11} & \mathbf{A}'^{12} \\ \hline \mathbf{A}'^{12\intercal} & \mathbf{A}'^{22} \end{array}\right] \left(\begin{bmatrix} \mathbf{a}_1 \\ \hline \mathbf{a}'_2 \end{bmatrix} - \sum_{i=1}^{m} \begin{bmatrix} \mathbf{A}_{12,i} \\ \hline \mathbf{M}_{\cdot i} \end{bmatrix} \mathbf{A}_{22,i}^{-1} \mathbf{a}_{2,i} \right),$$

$$\begin{bmatrix} \mathbf{x}_{2,i} \end{bmatrix} = \mathbf{A}_{22,i}^{-1} \left(\mathbf{a}_{2,i} - \left[\mathbf{A}_{12,i}^{\intercal} \;\middle|\; \mathbf{M}_{\cdot i}^{\intercal} \right] \begin{bmatrix} \mathbf{x}_1 \\ \hline \mathbf{x}'_2 \end{bmatrix} \right), \quad \text{for } i = 1, \ldots, m.$$

The algorithm takes in input the sub-blocks of the $\mathbf{A}$ matrix and the $\mathbf{a}$ vector and returns the inverse of the sub-block of interest of the $\mathbf{A}$ matrix and the solution set $\mathbf{x}$. Nolan *et al.* (2018) illustrate the algorithm for the two-levels nested model. It can be easily adapted to the crossed case, redefining the sub-block of the $\mathbf{A}$ matrix as previously shown. In this way, we are able to compute the streamlined inverse of a matrix with sparse structure as the inverse of the covariance matrix in the crossed random effects case. Clearly, when applied to a generic matrix, interest has to be on the inverse of non-zero sub-blocks of the matrix.

3.2.1.2 Least squares form solution

In Nolan and Wand (2018), the second result concentrates on the BLUPs as minimiser of the least squares problem. We can write the vector of coefficients as the minimiser of the form

$$\left\| \mathbf{b} - \mathbf{B} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u}' \\ \mathbf{u} \end{bmatrix} \right\|^2 = \left(\mathbf{b} - \mathbf{B} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u}' \\ \mathbf{u} \end{bmatrix} \right)^{\intercal} \left(\mathbf{b} - \mathbf{B} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u}' \\ \mathbf{u} \end{bmatrix} \right),$$

whose solution can be written as

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}' \\ \hat{\mathbf{u}} \end{bmatrix} = \left(\mathbf{B}^{\intercal} \mathbf{B} \right)^{-1} \mathbf{B}^{\intercal} \mathbf{b}.$$

The study of the appropriate structure of the $\mathbf{B}$ matrix, as declared in the least squares problem, allows taking advantage of the useful QR-decomposition in the matrix inversion problem. Profit of this result involves a more efficient way to compute both the fixed and random effects coefficients and their covariance matrix.

As for the two-levels nested case presented in Nolan and Wand (2018), also in the crossed effects model, the $\mathbf{B}$ matrix has the following structure

$$\mathbf{B} = \left[\begin{array}{c|cccc} \mathbf{B}_{.1} & \dot{\mathbf{B}}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{B}_{.2} & \mathbf{0} & \dot{\mathbf{B}}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_{.m} & \mathbf{0} & \mathbf{0} & \dots & \dot{\mathbf{B}}_m \end{array}\right]. \tag{3.1}$$

Except for slight differences, the solution of the sub-matrices $\mathbf{B}_{.i}$ and $\dot{\mathbf{B}}_i$ keeps the same structure of the two-levels nested case in Nolan and Wand (2018). In particular, the sub-matrices of $\mathbf{B}$ through which we can express the BLUPs solution in the crossed random effects case are

$$\mathbf{b}_i = \begin{bmatrix} \frac{\mathbf{y}_{1i}}{\sigma_\varepsilon} \\ \vdots \\ \frac{\mathbf{y}_{m'i}}{\sigma_\varepsilon} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \mathbf{B}_{.i} = \begin{bmatrix} \frac{\mathbf{X}_{1i}}{\sigma_\varepsilon} & \frac{\mathbf{Z}'_{1i}}{\sigma_\varepsilon} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathbf{X}_{m'i}}{\sigma_\varepsilon} & \mathbf{0} & \dots & \frac{\mathbf{Z}'_{m'i}}{\sigma_\varepsilon} \\ \mathbf{0} & \boldsymbol{\Sigma}'^{\frac{1}{2}} & \dots & \boldsymbol{\Sigma}'^{\frac{1}{2}} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix}, \dot{\mathbf{B}}_i = \begin{bmatrix} \frac{\mathbf{Z}_{1i}}{\sigma_\varepsilon} \\ \vdots \\ \frac{\mathbf{Z}_{m'i}}{\sigma_\varepsilon} \\ \mathbf{0} \\ \boldsymbol{\Sigma}^{\frac{1}{2}} \end{bmatrix}, \quad \text{for } i = 1, \dots, m,$$

where $\mathbf{B}_{.i}$ is $\left(\sum_{i'=1}^{m'} n_{i'i} + q' + q\right) \times (p + q'm')$, $\dot{\mathbf{B}}_i$ is $\left(\sum_{i'=1}^{m'} n_{i'i} + q' + q\right) \times q$ and $\mathbf{b}_i$ is $\left(\sum_{i'=1}^{m'} n_{i'i} + q' + q\right) \times 1$.

The $\mathbf{A}^{-1}$ matrix can be computed through the inverse of the $\mathbf{B}$ matrix using the QR-decomposition. Indeed, the following result holds

$$\mathbf{B}^\intercal\mathbf{B} = \left[\begin{array}{c|ccc} \sum_{i=1}^{m} \mathbf{B}_{.i}^\intercal\mathbf{B}_{.i} & \mathbf{B}_{.1}^\intercal\dot{\mathbf{B}}_1 & \dots & \mathbf{B}_{.m}^\intercal\dot{\mathbf{B}}_m \\ \hline \dot{\mathbf{B}}_1^\intercal\mathbf{B}_{.1} & \dot{\mathbf{B}}_1^\intercal\dot{\mathbf{B}}_1 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \dot{\mathbf{B}}_m^\intercal\mathbf{B}_{.m} & \mathbf{0} & \dots & \dot{\mathbf{B}}_m^\intercal\dot{\mathbf{B}}_m \end{array}\right] = \mathbf{C}^\intercal\mathbf{R}^{-1}\mathbf{C} + \mathbf{D} = \mathbf{A}.$$

More in detail, the non-zero elements of four sub-blocks of the $\mathbf{B}^\intercal\mathbf{B}$ matrix, highlighted by the thick lines, are defined as

$$\sum_{i=1}^{m}\mathbf{B}_{.i}^\intercal\mathbf{B}_{.i} = \frac{1}{\sigma_\varepsilon^2}\left[\begin{array}{c|ccc} \sum_{i=1}^{m}\sum_{i'=1}^{m'}\mathbf{X}_{i'i}^\intercal\mathbf{X}_{i'i} & \sum_{i=1}^{m}\mathbf{X}_{1i}^\intercal\mathbf{Z}'_{1i} & \dots & \sum_{i=1}^{m}\mathbf{X}_{m'i}^\intercal\mathbf{Z}'_{m'i} \\ \hline \sum_{i=1}^{m}\mathbf{Z}_{1i}'^\intercal\mathbf{X}_{1i} & \sum_{i=1}^{m}\mathbf{Z}_{1i}'^\intercal\mathbf{Z}'_{1i}+\boldsymbol{\Sigma}' & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{m}\mathbf{Z}_{m'i}'^\intercal\mathbf{X}_{m'i} & \mathbf{0} & \dots & \sum_{i=1}^{m}\mathbf{Z}_{m'i}'^\intercal\mathbf{Z}'_{m'i}+\boldsymbol{\Sigma}' \end{array}\right]$$
$$= \left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}'_{12} \\ \hline \mathbf{A}_{12}'^\intercal & \mathbf{A}'_{22} \end{array}\right],$$

$$\mathbf{B}_{.i}^\intercal\dot{\mathbf{B}}_i = \frac{1}{\sigma_\varepsilon^2}\left[\begin{array}{c} \sum_{i'=1}^{m'}\mathbf{X}_{i'i}^\intercal\mathbf{Z}_{i'i} \\ \hline \mathbf{Z}_{1i}'^\intercal\mathbf{Z}_{1i} \\ \vdots \\ \mathbf{Z}_{m'i}'^\intercal\mathbf{Z}_{m'i} \end{array}\right] = \left[\begin{array}{c} \mathbf{A}_{12,i} \\ \hline \mathbf{M}_{.i} \end{array}\right], \quad \text{for } i=1,\dots,m,$$

and

$$\dot{\mathbf{B}}_i^\intercal\dot{\mathbf{B}}_i = \frac{1}{\sigma_\varepsilon^2}\sum_{i'=1}^{m'}\mathbf{Z}_{i'i}^\intercal\mathbf{Z}_{i'i}+\boldsymbol{\Sigma} = \left[\mathbf{A}_{22,i}\right], \quad \text{for } i=1,\dots,m.$$

In the algorithm described in Nolan and Wand (2018), the QR-decomposition is applied to the $\dot{\mathbf{B}}_i$ sub-blocks in order to compute the inverse of the non zero sub-blocks of the $\mathbf{A}_{22}$ matrix.

3.3 Mean field variational Bayes approach

In recent years, there has been a growing interest in variational approximations methods applied to statistical inference problems. Variational methods are approximation techniques based on the variational principle mostly applied in the physics branch of quantum mechanics. The idea is to choose some functions among a specified class of functions, through which maximising a certain quantity of interest that depends on that functions. Restrictions on the class of functions cause approximation but enhance tractability (Ormerod and Wand, 2010). Variational methods employed on Bayesian inference problems are called variational Bayes and are mostly used to take advantage of their ability in approximating intractable or high-dimensional integrals.

Variational methods and Monte Carlo Markov Chain

Variational Bayes methods concentrate on the approximation of the joint posterior distribution, especially in complex statistical models. Indeed, some situations are difficult to handle using exact inference. It is the case of models with intractable posterior density function or with specific dependence structure among parameters. The former scenario requires the application of approximation techniques since exact inference is infeasible. In the latter example, high dependence among some groups of parameters leads to an increase of algorithm complexity resulting in an increase of the computational cost and time needed to achieve results with exact algorithms (Jordan *et al.*, 1999).

In the statistical literature, Monte Carlo methods are the most applied approximation methods to overcome unmanageable situations with exact inference. These techniques are characterised by flexibility, ease of implementation and proven theoretical convergence (Robert and Casella, 2004). Monte Carlo methods are based on the construction of an ergodic Markov chain whose stationary distribution is the posterior distribution. Approximate inference in MCMC is made on the samples drawn from the Markov chain. However, facing situations with high and complex dependencies among parameters, Monte Carlo algorithms may suffer from poor mixing and slow convergence. Variational Bayes techniques turn out to be an efficient alternative approach. As distinct from MCMC based on sampling, variational methods evaluate the inferential problem using optimisation algorithms. In recent studies, this class of methods appears to be faster than MCMC in several situations, especially dealing with large-scale data. Compared to MCMC, the set-up of the algorithm in variational methods is usually more complex. Concurrently, statistical properties of variational methods are still under examination. As an example, some studies highlighted that variational techniques may underestimate the variance of the posterior distribution, that however is not always of primary interest for the analysis or, under other circumstances, a trade-off between speed and accuracy is needed. For a comprehensive introduction of variational Bayes methods and a detailed comparison with MCMC techniques, we refer to Blei *et al.* (2017).

3.3.1 Mean Field Variational Bayes

Early developments of the class of variational Bayes methods emerged since the 1980s in the statistical physics literature. Lately, in 1990s, these methods have been used in machine learning problems (Jordan *et al.*, 1999), but their success, applied to statistics, arrived in 2010s. According to the Bayes theorem, for a generic Bayesian model the posterior distribution for the continuous parameter vector $\boldsymbol{\theta} \in \Theta$, is defined as

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})},$$

where $\mathbf{y}$ is the observed vector, $p(\mathbf{y}|\boldsymbol{\theta})$ represents the likelihood function, $p(\boldsymbol{\theta})$ the prior distributions and $p(\mathbf{y})$ is the normalising constant or marginal likelihood, that is

$$p(\mathbf{y}) = \int_{\Theta} p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$

The marginal likelihood is usually not available in closed form or is computationally intensive. In variational inference, the posterior density function $p(\boldsymbol{\theta}|\mathbf{y})$ is approximated through a variational distribution called approximating density function $q(\boldsymbol{\theta})$, defined on the parameter space Θ.

Working on the logarithmic scale and dividing and multiplying the marginal likelihood by the variational distribution $q(\boldsymbol{\theta})$, we get

$$\begin{aligned}
\log p(\mathbf{y}) &= \log \int_{\Theta} q(\boldsymbol{\theta}) \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} \\
&= \log \mathbb{E}_q \left[\frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right] \\
&\geq \mathbb{E}_q \left[\log \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right],
\end{aligned}$$

where the inequality holds for the log-concavity of the likelihood function and the Jensen's inequality (Jeon *et al.*, 2017). The last term represents a lower bound of the marginal likelihood generated by the auxiliary distribution that we can rewrite as follows

$$\begin{aligned}
\mathbb{E}_q \left[\log \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right] &= \mathbb{E}_q[\log p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})] - \mathbb{E}_q[\log q(\boldsymbol{\theta})] \\
&= \mathbb{E}_q[\log p(\boldsymbol{\theta}|\mathbf{y})p(\mathbf{y})] - \mathbb{E}_q[\log q(\boldsymbol{\theta})] \\
&= \mathbb{E}_q[\log p(\boldsymbol{\theta}|\mathbf{y})] + \mathbb{E}_q[\log p(\mathbf{y})] - \mathbb{E}_q[\log q(\boldsymbol{\theta})] \\
&= \mathbb{E}_q[\log p(\mathbf{y})] - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})),
\end{aligned}$$

where $\mathbb{E}_q[\log p(\mathbf{y})]$ is actually the logarithm of the marginal likelihood and KL indicates the Kullback-Leibler divergence which is defined as

$$\mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) = \int_{\Theta} q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}\mathrm{d}\boldsymbol{\theta},$$

which is a quantity greater or equal to zero. Thus, to summarise, the following result holds

$$\log p(\mathbf{y}) = \mathbb{E}_q \left[\log \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right] - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})).$$

The gap between the lower bound and the marginal likelihood is minimised when the

variational distribution $q(\boldsymbol{\theta})$ minimises the Kullback-Leibler divergence, that is when it is equal to the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$.

As previously said, the main difficulty in exact inference is the intractability of the marginal likelihood, and variational inference bypasses the problem applying restrictions on the class of distributions $\mathcal{Q}$. Mean field approximation applies the product restriction on the class of distribution $\mathcal{Q}$ in order to achieve tractability, that is

$$q^*(\boldsymbol{\theta}) = \underset{q\in\mathcal{Q}}{\operatorname{argmin}}\, \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})),$$

and

$$\mathcal{Q} = \left\{ q(\boldsymbol{\theta}) : q(\boldsymbol{\theta}) = \prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i) \text{ for some partition of } \boldsymbol{\theta} \right\},$$

where M is the total number of parameters (Lee and Wand, 2016a). From the previous equation it is clear that the main idea behind variational inference is to approximate the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ through an approximating density function $q(\boldsymbol{\theta})$ that generates a tractable lower bound on the marginal likelihood $p(\mathbf{y})$, subject to some product density restrictions (Ormerod and Wand, 2010).

The chosen product structure can play an important role in the accuracy of the resulting inference. Indeed, if the partition does not reflect the posterior dependence structure among parameters, the accuracy may be low (Ormerod and Wand, 2010).

3.3.2 Bayesian Gaussian crossed random effects model

The crossed random effects model under the Bayesian assumptions is defined as

$$\begin{gathered}
\mathbf{y}_{i'i}|\boldsymbol{\beta}, \mathbf{u}'_{i'}, \mathbf{u}_i, \sigma^2_\varepsilon \overset{\text{ind}}{\sim} N(\mathbf{X}_{i'i}\boldsymbol{\beta} + \mathbf{Z}'_{i'i}\mathbf{u}'_{i'} + \mathbf{Z}_{i'i}\mathbf{u}_i, \mathbf{R}), \quad \boldsymbol{\beta} \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta), \\
\mathbf{u}'_{i'} \sim N(\mathbf{0}, \boldsymbol{\Sigma}'), \quad \mathbf{u}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \quad \text{for } i' = 1, \ldots, m', \quad i = 1, \ldots, m, \\
\sigma^2_\varepsilon|a_\varepsilon \sim \text{Inverse-}\chi^2(1, 1/a_\varepsilon), \quad a_\varepsilon \sim \text{Inverse-}\chi^2(1, 1/A^2_\varepsilon), \\
\boldsymbol{\Sigma}'|a'_{\boldsymbol{\Sigma}',1}, \ldots, a'_{\boldsymbol{\Sigma}',q'} \sim \text{Inverse-Wishart}(\nu' + q' - 1, 2\nu'\text{diag}(1/a'_{\boldsymbol{\Sigma}',1}, \ldots, 1/a'_{\boldsymbol{\Sigma}',q'})), \\
a'_{\boldsymbol{\Sigma}',1}, \ldots, a'_{\boldsymbol{\Sigma}',q'} \overset{\text{ind}}{\sim} \text{Inverse-}\chi^2(1, 1/A'^2_{\boldsymbol{\Sigma}'}), \\
\boldsymbol{\Sigma}|a_{\boldsymbol{\Sigma},1}, \ldots, a_{\boldsymbol{\Sigma},q} \sim \text{Inverse-Wishart}(\nu + q - 1, 2\nu\text{diag}(1/a_{\boldsymbol{\Sigma},1}, \ldots, 1/a_{\boldsymbol{\Sigma},q})), \\
a_{\boldsymbol{\Sigma},1}, \ldots, a_{\boldsymbol{\Sigma},q} \overset{\text{ind}}{\sim} \text{Inverse-}\chi^2(1, 1/A^2_{\boldsymbol{\Sigma}}),
\end{gathered} \tag{3.2}$$

where, for $i' = 1, \ldots, m'$ and $i = 1, \ldots, m$, $\mathbf{y}_{i'i}$ is $n_{i'i} \times 1$, $\mathbf{X}_{i'i}$ is $n_{i'i} \times p$, $\boldsymbol{\beta}$ is $p \times 1$, $\mathbf{Z}'_{i'i}$ is $n_{i'i} \times q'$ and $\mathbf{Z}_{i'i}$ is $n_{i'i} \times q$, $\mathbf{u}'_{i'}$ is $q' \times 1$, $\mathbf{u}_i$ is $q \times 1$, $\boldsymbol{\Sigma}'$ is $q' \times q'$, $\boldsymbol{\Sigma}$ is $q \times q$ and $\mathbf{R} = \sigma^2_\varepsilon \mathbf{I}$.

Moreover, the hyper-parameter $\mathbf{\Sigma}_\beta$ is defined as symmetric and positive definite and ν, A_ε, $A'_{\mathbf{\Sigma}'}$, and $A_{\mathbf{\Sigma}}$ are greater than zero. The main difference with the model under frequentist assumptions is the definition of the parameters as random variables with their density functions and hyper-parameter specifications.

Prior choice for the variance and covariance matrix parameters are defined as in Wand (2017), similarly to the marginally non informative prior distributions for covariance matrices described in Huang *et al.* (2013). This specification, characterised by conditional conjugacy, allows to obtain an Half-t prior on standard deviation parameter, suggested by Gelman *et al.* (2006) for weakly informative prior on the variance parameters. Moreover, it leads to Uniform priors on correlation parameters (Huang *et al.*, 2013).

3.3.3 Directed acyclic graph

The directed acyclic graph (DAG) shows the hierarchical structure of the Bayesian crossed random effects model and highlights the conditional distributional relationship among the variables. The directed feature is due to the one-headed arrows and it is called acyclic since it is not possible to visit the same variable twice. Lack of edges between nodes represents conditional independence in the joint distribution. Stochastic quantities, called nodes, differ between parameters or unknown quantities, represented by white circles, and data or observed quantities, depicted as shaded circles. Figure 3.1 Figure 3.1 shows the DAG related to the crossed random effects in Equation 3.2.

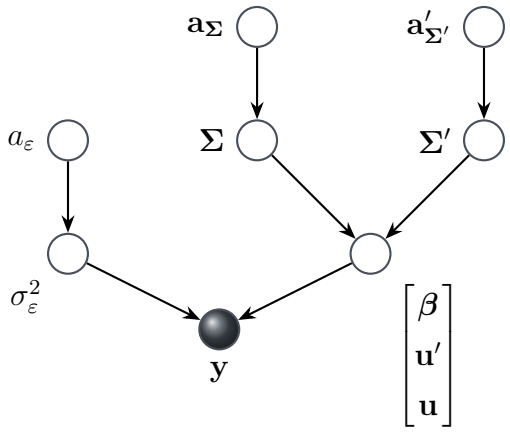

FIGURE 3.1: Directed acyclic graph for the crossed random effects model under the Bayesian approach. Nodes, represented by a white circle, correspond to random and auxiliary variables. The shaded node refers to the data vector, while the edges specify conditional dependencies.

The nodes $\mathbf{a}'_{\mathbf{\Sigma}'}$ and $\mathbf{a}_{\mathbf{\Sigma}}$ correspond to the random vectors $\left[a'_{\mathbf{\Sigma}',1}, \ldots, a'_{\mathbf{\Sigma}',q'}\right]$ and

$[a_{\boldsymbol{\Sigma},1}, \ldots, a_{\boldsymbol{\Sigma},q}]$. Similarly, the nodes $\mathbf{u}'$ and $\mathbf{u}$ are respectively defined as the random vectors $[\mathbf{u}'_1, \ldots, \mathbf{u}'_{m'}]$ and $[\mathbf{u}_1, \ldots, \mathbf{u}_m]$. Given a node, looking at the directed edges departing from and arriving at it, we can identify its parents, co-parents and children. This set of kinship is called Markov blanket and it is useful to compute the full conditional distribution of a given node. Indeed, the distribution of a node given all the other variables in the DAG is equal to the distribution of the node given its Markov blanket.

3.3.4 Mean field variational Bayes approximations

The mean field product restriction applied is

$$q(\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}, a_\varepsilon, a'_{\boldsymbol{\Sigma}',1}, \ldots, a'_{\boldsymbol{\Sigma}',q'}, a_{\boldsymbol{\Sigma},1}, \ldots, a_{\boldsymbol{\Sigma},q}) q(\sigma^2_\varepsilon, \boldsymbol{\Sigma}', \boldsymbol{\Sigma}),$$

that is, apart from induced factorisation (Bishop, 2006), the minimal factorisation in our approximation (Lee and Wand, 2016a). Variational distributions for the parameters of the model in Equation 3.2 are presented below. Detailed computations of the variational distribution are showed for the a_ε hyper-parameter, for all the other parameters we refer to Appendix B.

The first step for the derivation of the variational distribution concerns the full conditional distribution that is defined as

$$\begin{aligned}
p(a_\varepsilon|\text{rest}) &= p(a_\varepsilon|\text{Markov blanket of } a_\varepsilon) \\
&= p(a_\varepsilon|\sigma^2_\varepsilon) \\
&\propto p(\sigma^2_\varepsilon, a_\varepsilon) \\
&= p(\sigma^2_\varepsilon|a_\varepsilon) p(a_\varepsilon) \\
&\propto a_\varepsilon^{-1/2} a_\varepsilon^{-3/2} exp\left\{ -\frac{\sigma_\varepsilon^{-2} + A_\varepsilon^{-2}}{2a_\varepsilon} \right\}.
\end{aligned}$$

Hence, the full conditional for the parameter a_ε is

$$a_\varepsilon|\sigma^2_\varepsilon \sim \text{Inverse-}\chi^2 \left(2, \frac{1}{\sigma^2_\varepsilon} + \frac{1}{A^2_\varepsilon} \right).$$

The q^* density on log scale is

$$\begin{aligned}
\log q^*(a_\varepsilon) &= \mathbb{E}_{q(rest)} \log p(a_\varepsilon | rest) + \text{const} \\
&= \mathbb{E}_{q(\sigma_\varepsilon^2)} \log p(a_\varepsilon | \sigma_\varepsilon^2) + \text{const} \\
&= \mathbb{E}_{q(\sigma_\varepsilon^2)} \left[-2 \log a_\varepsilon - \left(\frac{1}{\sigma_\varepsilon^2} + \frac{1}{A_\varepsilon^2} \right) \frac{1}{2a_\varepsilon} \right] + \text{const} \\
&= -2 \log a_\varepsilon - \left(\mathbb{E}_{q(\sigma_\varepsilon^2)} \left[\frac{1}{\sigma_\varepsilon^2} \right] + \frac{1}{A_\varepsilon^2} \right) \frac{1}{2a_\varepsilon} + \text{const}.
\end{aligned}$$

Thus, the approximate density function for the a_ε hyper-parameter is given by

$$q^*(a_\varepsilon) \sim \text{Inverse-}\chi^2 \left(2, \mathbb{E}_{q(\sigma_\varepsilon^2)} \left[\frac{1}{\sigma_\varepsilon^2} \right] + \frac{1}{A_\varepsilon^2} \right).$$

With similar computations we get

$$q^*(a'_{\boldsymbol{\Sigma}',i'}) \sim \text{Inverse-}\chi^2 \left(1, \frac{1}{A^2_{\boldsymbol{\Sigma}'}} + 2\nu' \mathbb{E}_{q(\boldsymbol{\Sigma}')} \left[\boldsymbol{\Sigma}'^{-1}_{i'i'} \right] \right),$$

and, equivalently,

$$q^*(a_{\boldsymbol{\Sigma},i}) \sim \text{Inverse-}\chi^2 \left(1, \frac{1}{A^2_{\boldsymbol{\Sigma}}} + 2\nu \mathbb{E}_{q(\boldsymbol{\Sigma})} \left[\boldsymbol{\Sigma}^{-1}_{ii} \right] \right).$$

The approximate density functions for $\boldsymbol{\Sigma}'$ and $\boldsymbol{\Sigma}$ are given by

$$q^*(\boldsymbol{\Sigma}') \sim \text{Inverse-Wishart} \left(m' + \nu' + q' - 1, \mathbb{E}_{q(\mathbf{u}', a'_{\boldsymbol{\Sigma}',1}, \ldots, a'_{\boldsymbol{\Sigma}',q'})} [\mathbf{B}'] \right),$$

where $\mathbf{B}' = \sum_{i'=1}^{m'} \mathbf{u}'_{i'} \mathbf{u}'^{\intercal}_{i'} + 2\nu' \text{diag} \left(1/a'_{\boldsymbol{\Sigma}',1}, \ldots, 1/a'_{\boldsymbol{\Sigma}',q'} \right)$, and

$$q^*(\boldsymbol{\Sigma}) \sim \text{Inverse-Wishart} \left(m + \nu + q - 1, \mathbb{E}_{q(\mathbf{u}, a_{\boldsymbol{\Sigma},1}, \ldots, a_{\boldsymbol{\Sigma},q})} [\mathbf{B}] \right),$$

where $\mathbf{B} = \sum_{i=1}^{m} \mathbf{u}_i \mathbf{u}_i^{\intercal} + 2\nu \text{diag} \left(1/a_{\boldsymbol{\Sigma},1}, \ldots, 1/a_{\boldsymbol{\Sigma},q} \right)$.

Lastly, we have

$$\begin{aligned}
q^*(\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}) \sim N \Bigg(& \left(\mathbf{C}^{\intercal} \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \left[\left(\sigma_\varepsilon^2 \mathbf{I} \right)^{-1} \right] \mathbf{C} + \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \left[\boldsymbol{\Lambda}^{-1} \right] \right)^{-1} \mathbf{C}^{\intercal} \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \left[\left(\sigma_\varepsilon^2 \mathbf{I} \right)^{-1} \right] \mathbf{y}, \\
& \left(\mathbf{C}^{\intercal} \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \left[\left(\sigma_\varepsilon^2 \mathbf{I} \right)^{-1} \right] \mathbf{C} + \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \left[\boldsymbol{\Lambda}^{-1} \right] \right)^{-1} \Bigg),
\end{aligned}$$

where where $\mathbf{C} = [\mathbf{X}, \mathbf{Z}', \mathbf{Z}]$, $\mathbf{a} = \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u}' \\ \mathbf{u} \end{bmatrix}$ and $\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Sigma}_\beta & 0 & 0 \\ 0 & \boldsymbol{\Sigma}' & 0 \\ 0 & 0 & \boldsymbol{\Sigma} \end{pmatrix}$, and

$$q^*(\sigma_\varepsilon^2) \sim \text{Inverse-}\chi^2\left(n+1, \mathbb{E}_{q(a_\varepsilon)}\left[\frac{1}{a_\varepsilon} + \|\mathbf{y} - \mathbf{Ca}\|^2\right]\right).$$

A coordinate ascending algorithm is used to update the parameters values of the optimal q-density functions, indeed they are related to each other. Derived solutions in Section 3.2.1.1 and Section 3.2.1.2 are applied in the MFVB routine in order to update the parameters of the following approximating densities, denoted later on with a notation easier to follow

$$q^*(\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}) \sim N\left(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\mathbf{u}',\mathbf{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\mathbf{u}',\mathbf{u})}\right),$$

$$q^*(\sigma_\varepsilon^2) \sim \text{Inverse-}\chi^2\left(\kappa_{q(\sigma_\varepsilon^2)}, \lambda_{q(\sigma_\varepsilon^2)}\right),$$

and

$$q^*(\boldsymbol{\Sigma}') \sim \text{Inverse-Wishart}\left(m' + \nu' + q' - 1, \boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma}')}\right),$$
$$q^*(\boldsymbol{\Sigma}) \sim \text{Inverse-Wishart}\left(m + \nu + q - 1, \boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})}\right).$$

In particular, the mean vector and covariance matrix of $q^*(\boldsymbol{\beta}, \mathbf{u}', \mathbf{u})$ have the form

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\mathbf{u}',\mathbf{u})} = \left(\mu_{q(1/\sigma_\varepsilon^2)}\mathbf{C}^\intercal\mathbf{C} + \begin{bmatrix} \boldsymbol{\Sigma}_\beta^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m'} \otimes \mathbf{M}_{q(\boldsymbol{\Sigma}'^{-1})} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_m \otimes \mathbf{M}_{q(\boldsymbol{\Sigma}^{-1})} \end{bmatrix}\right)^{-1},$$
$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\mathbf{u}',\mathbf{u})} = \mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\mathbf{u}',\mathbf{u})}\mathbf{C}^\intercal\mathbf{y},$$

and they can be computed as the solutions of a least squares problem form applying the algorithms described in Section 3.2.1.1 and Section 3.2.1.2.

Indeed, writing

$$\left\|\mathbf{b} - \mathbf{B}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\mathbf{u}',\mathbf{u})}\right\|^2,$$

where the sub-matrices of $\mathbf{B}$, with structure as in Equation 3.1, are defined as

$$\mathbf{b}_i = \begin{bmatrix} \mu_{q(1/\sigma_\varepsilon^2)}^{1/2}\mathbf{y}_{1i} \\ \vdots \\ \mu_{q(1/\sigma_\varepsilon^2)}^{1/2}\mathbf{y}_{m'i} \\ m^{-1/2}\boldsymbol{\Sigma}_\beta^{-1/2}\boldsymbol{\mu}_\beta \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \mathbf{B}_{.i} = \begin{bmatrix} \mu_{q(1/\sigma_\varepsilon^2)}^{1/2}\mathbf{X}_{1i} & \mu_{q(1/\sigma_\varepsilon^2)}^{1/2}\mathbf{Z}'_{1i} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{q(1/\sigma_\varepsilon^2)}^{1/2}\mathbf{X}_{m'i} & \mathbf{0} & \dots & \mu_{q(1/\sigma_\varepsilon^2)}^{1/2}\mathbf{Z}'_{m'i} \\ m^{-1/2}\boldsymbol{\Sigma}_\beta^{-1/2}\boldsymbol{\mu}_\beta & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{q(\boldsymbol{\Sigma}'^{-1})}^{1/2} & \dots & \mathbf{M}_{q(\boldsymbol{\Sigma}'^{-1})}^{1/2} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix},$$

and

$$\dot{\mathbf{B}}_i = \begin{bmatrix} \mu^{1/2}_{q(1/\sigma^2_\varepsilon)}\mathbf{Z}_{1i} \\ \vdots \\ \mu^{1/2}_{q(1/\sigma^2_\varepsilon)}\mathbf{Z}_{m'i} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{M}^{1/2}_{q(\boldsymbol{\Sigma}^{-1})} \end{bmatrix}, \quad \text{for } i = 1, \ldots, m,$$

leads to the following equivalences

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\mathbf{u}')} = \left[\begin{array}{c} \mathbf{x}_1 \\ \hline \mathbf{x}'_2 \end{array}\right], \quad \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\mathbf{u}')} = \left[\begin{array}{c|c} \mathbf{A}^{11} & \mathbf{A}'^{12} \\ \hline \mathbf{A}'^{12\intercal} & \mathbf{A}'^{22} \end{array}\right],$$

$$\boldsymbol{\mu}_{q(\mathbf{u}_i)} = \mathbf{x}_{2,i}, \quad \boldsymbol{\Sigma}_{q(\mathbf{u}_i)} = \mathbf{A}^{22,i}, \quad \left[\begin{array}{c} \mathbb{E}_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\mathbf{u}_i - \boldsymbol{\mu}_{q(\mathbf{u}_i)})^\intercal\} \\ \hline \mathbb{E}_q\{(\mathbf{u}' - \boldsymbol{\mu}_{q(\mathbf{u}')})(\mathbf{u}_i - \boldsymbol{\mu}_{q(\mathbf{u}_i)})^\intercal\} \end{array}\right] = \left[\begin{array}{c} \mathbf{A}^{12,i} \\ \hline \mathbf{M}^{\cdot i} \end{array}\right].$$

The streamlined MFVB algorithm implemented in Nolan *et al.* (2018) for the two-level linear mixed model can be easily extended to the crossed random effects case using results presented above.

3.4 Results and discussion

The streamlined MFVB algorithm has been tested on synthetic data varying the number of groups of the random effects, i.e. varying intercept and slope in both crossed effects. Fixing the number of iterations for the MFVB algorithm to 50, we fitted several models considering combinations of number of groups with m' always lower than m and increasing m up to $1,000$ groups. Models are estimated using a $2,8$ GHz Intel Core i7 processor with 16 Gb of RAM. Computational times are reported in Table 3.1.

Groups		**m**			
		10	100	300	1,000
	5	0.9 sec	8.6 sec	55.3 sec	513.4 sec
m'	10	1.9 sec	38.6 sec	183.3 sec	1809.5 sec
	20	-	124.1 sec	687.2 sec	6356.6 sec

TABLE 3.1: Computational times for the crossed random effects models fitted with streamlined MFVB algorithm varying number of groups.

Computational times, of the streamlined algorithm, are compared with the Stan implementation of the same Gaussian crossed random effects model. For $m' = 5$ and $m = 10$, the average time to simulate 1 chain of $2,000$ draws, with a sampling period

of $1,000$, is in the range of $2-3$ minutes. Rising the number of groups m up to 100, increases the computational time of the HMC simulations up to about 1 hour. Moreover, the accuracy of the estimated posterior distributions, comparing the MVBF fit with the HMC one is about 98%.

Concluding, streamlined MFVB approach allows us to obtain efficient, in terms of computational time and memory storage, and accurate approximate inference results for the Gaussian crossed random effects model. The simple structure of the model presented here is suitable to several extensions and applications taking into account that the greatest advantage is obtained when the number of groups in one random effect is kept low while the other grows.

Conclusions

The work presented in the thesis addresses two inferential issues regarding generalised linear models. The first aspect, described in the first two chapters, is about the estimation of number and position of knots in semiparametric generalised linear models using regression splines represented through truncated linear basis.

The first chapter presents a methodology that takes a step forward in estimating the knot locations, for a fixed number of knots. The proposed Bayesian methodology allows to jointly estimate regression and spline coefficients and knot locations using bivariate spline functions. The choice of the spline basis is motivated by practical reasons. Indeed, in linear regression splines, the knots can be interpreted as changes of slope in the estimated relation. Although truncated linear basis representation does not have optimal mathematical properties, it has proved to be an appropriate tool for estimating slope changes in contexts where the relationship is assumed to be piecewise linear and with few (2 or 3) change points. For example, in our epidemiological analysis, the detected change points were supported by a biological interpretation.

In the second chapter, we extend this methodology in order to estimate also the number of knots. The stochastic search variable selection approach has been adapted including dependence between the knot location and the related spline coefficient. The hierarchical mixture prior allows to identify the number of knots in the model and to have an initial guess of the change point locations. These approach has been tested through a simulation study and has been applied to the semiparametric logistic regression model with bivariate splines on larynx data with good results. The bivariate application confirmed results obtained from the model in the first chapter with a lower effort in terms of computational cost. Forthcoming developments involve comparing this procedure with alternative Bayesian approaches proposed in the literature and a deeper study of the effects of prior hierarchy on the model coefficients.

Possible and interesting future directions of research are manifold. Flexibility of our methodology allows being easily applied to situations in which it is useful to verify the presence and, the number, of change points in the relations. The development of an R package that makes the procedure easy to use and available for other practical

applications, e.g. suitable epidemiological, socio-economic or environmental data, is of great interest.

From a methodological point of view, other options for the basis representation, e.g. radial basis or B-spline, can be tested. Another aspect to investigate is to relax the piecewise linearity hypothesis of the splines taking into account higher order polynomial degree. First of all, it is necessary to study the meaning of the knots in quadratic or cubic spline regression, then, considering the degree of the spline as a parameter, it can be estimated extending the proposed methodology.

Moreover, referring to our practical epidemiological study, move to multivariate spline functions to jointly model the effect of more than two risk factors can be of great attractiveness leading to an even more flexible approach. Looking at some of our epidemiological applications, the inclusion of monotonicity constraints represents an interesting direction of work.

The second part of our work focused on the extension of the streamlined variational inference in Gaussian random effects model for nested structured data to the crossed structure case. Due to the less sparse structure of the matrix that needs to be inverted during the estimation problem, we were able to find a partial streamlined solution. Streamlined MFVB algorithm highlighted an almost linear increase in terms of computational time with the increase of the number of groups in one random effect. This approach can be extended also to other generalised linear models with crossed random effects structure. Possible extensions may involve, for example, different structures on the random effects covariance matrix or the specification of group-specific curves able to capture non-linearity. Moreover, the presented algorithms can be also applied to other variational Bayes techniques, e.g. variational message passing.

Appendix A

Two - dimensional credible intervals for estimated surfaces of OCP cancer site for current smokers (Fig. A.1) and OCP cancer site stratified by alcohol (Fig. A.2 and Fig. A.3), which seem to have a decreasing OR after the knots for increasing levels of risk factors. For a fixed level of one risk factor, the following plots show the two-dimensional 95% credible intervals of the fitted surface varying the other exposure. Fixed values of one exposure are chosen equal to the estimated knot location, the maximum value of the exposure and the mid point between them.

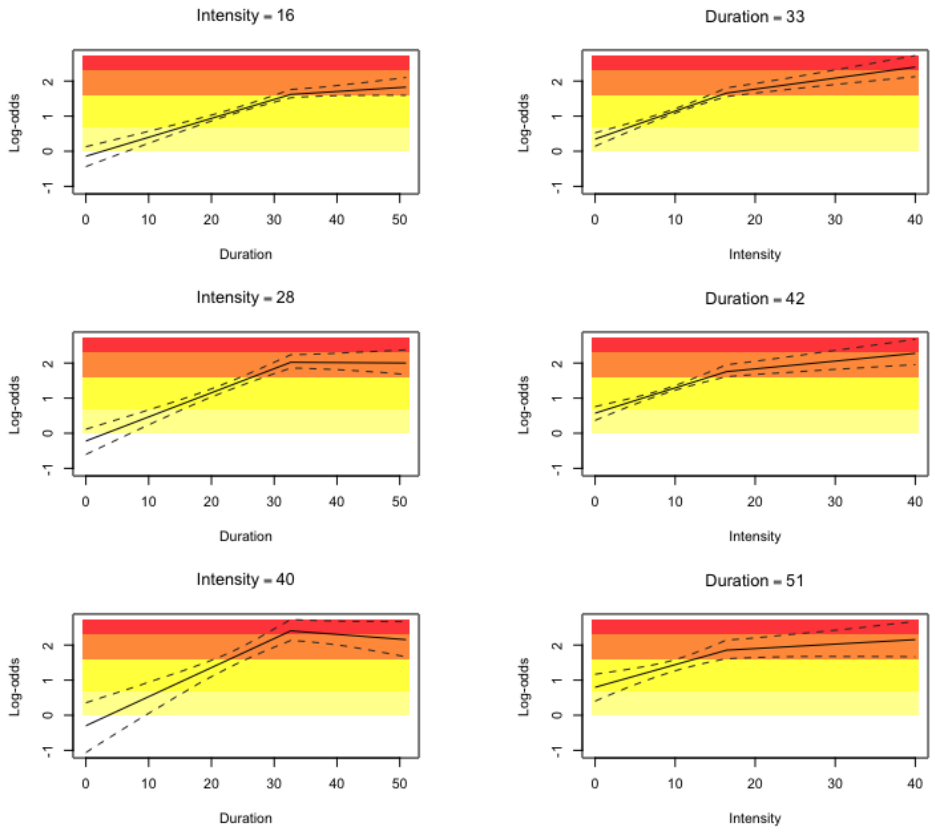

FIGURE A.1: Current smokers - stratified analysis by oral cavity and pharynx sites. For a fixed level of one risk factor, the plot shows the two-dimensional 95% credible intervals of the fitted surface varying the other exposure. Fixed values of one exposure are chosen equal to the estimated knot location, the maximum value of the exposure and a mid point between them. Results are shown in log-odds scale.

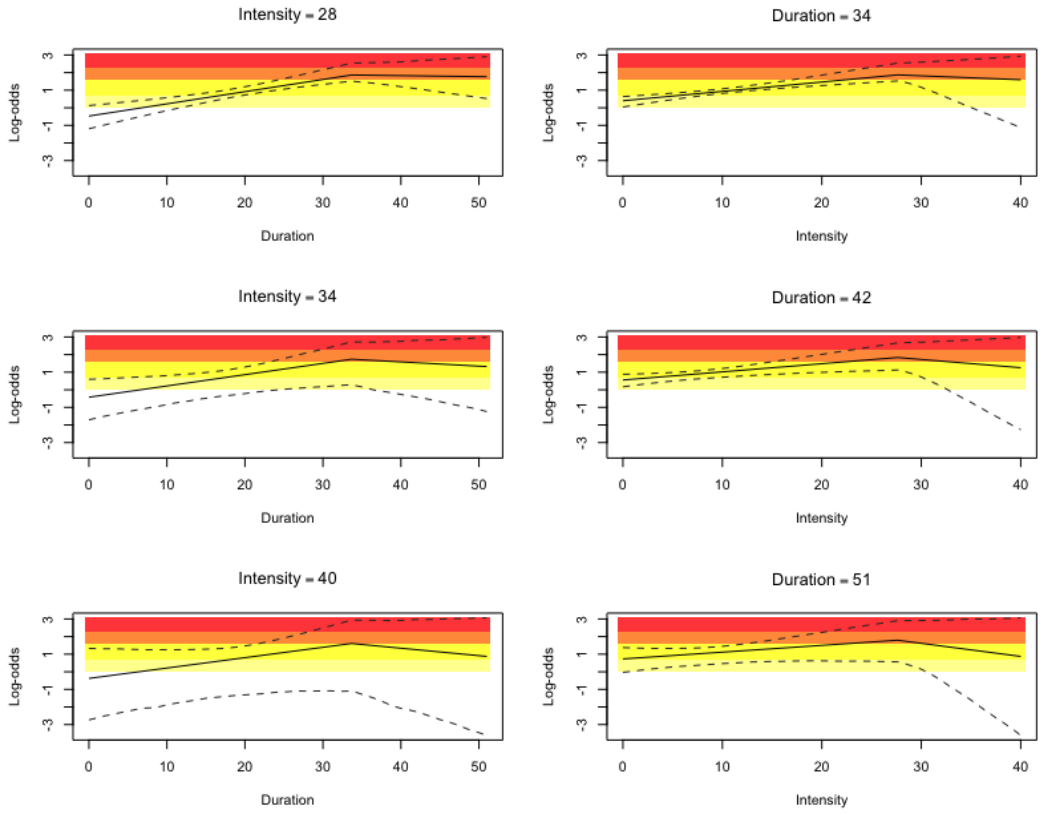

FIGURE A.2: Current smokers - stratified analysis by oral cavity and pharynx sites and and never drinkers. For a fixed level of one risk factor, the plot shows the two-dimensional 95% credible intervals of the fitted surface varying the other exposure. Fixed values of one exposure are chosen equal to the estimated knot location, the maximum value of the exposure and a mid point between them. Results are shown in log-odds scale.

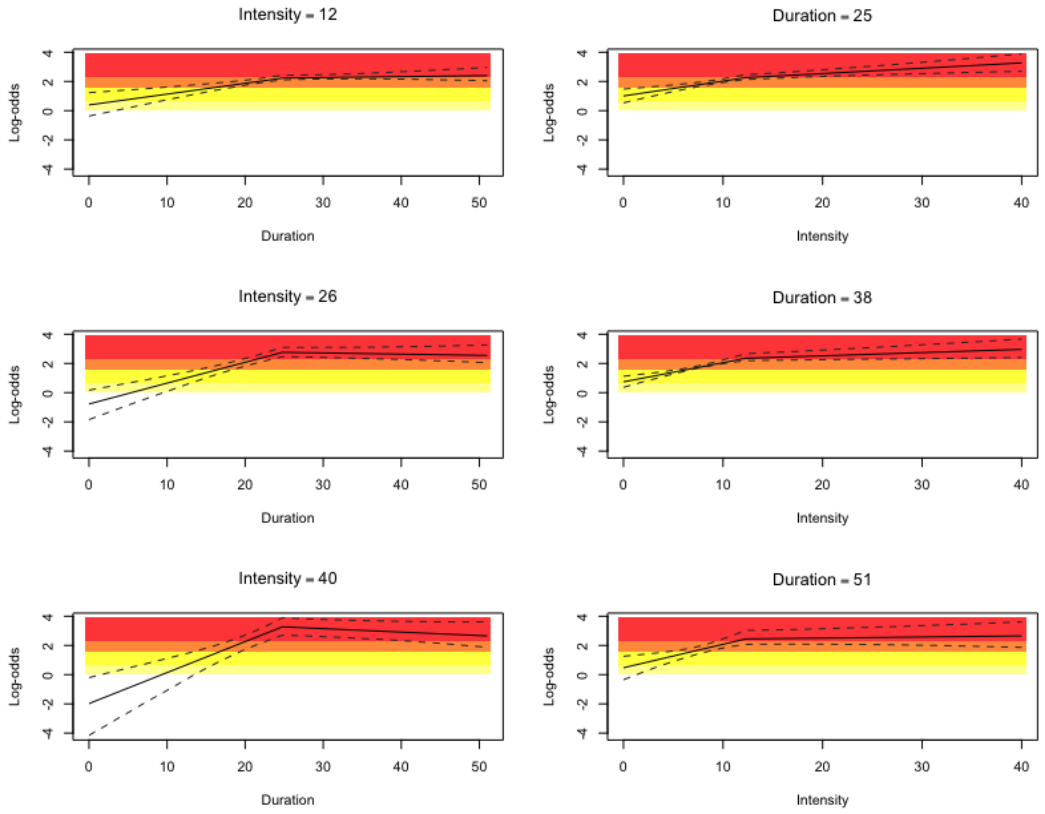

FIGURE A.3: Current smokers - stratified analysis by oral cavity and pharynx sites and and heavy drinkers. For a fixed level of one risk factor, the plot shows the two-dimensional 95% credible intervals of the fitted surface varying the other exposure. Fixed values of one exposure are chosen equal to the estimated knot location, the maximum value of the exposure and a mid point between them. Results are shown in log-odds scale.

Appendix B

Mean Field Variational Bayes approximate densities derivations for the crossed random effects model in Equation 3.2. Computations for the parameter a_ε are in Section 3.3.4.

$a'_{\mathbf{\Sigma}',1}, \ldots, a'_{\mathbf{\Sigma}',q'}$ and $a_{\mathbf{\Sigma},1}, \ldots, a_{\mathbf{\Sigma},q}$

The full conditional distribution is defined as

$$
\begin{aligned}
p(a'_{\mathbf{\Sigma}',1}, \ldots, a'_{\mathbf{\Sigma}',q'}) &= p(a'_{\mathbf{\Sigma}',1}, \ldots, a'_{\mathbf{\Sigma}',q'} | \text{Markov blanket of } a'_{\mathbf{\Sigma}',1}, \ldots, a'_{\mathbf{\Sigma}',q'}) \\
&= p(a'_{\mathbf{\Sigma}',1}, \ldots, a'_{\mathbf{\Sigma}',q'} | \mathbf{\Sigma}') \\
&\propto p(\mathbf{\Sigma}', a'_{\mathbf{\Sigma}',1}, \ldots, a'_{\mathbf{\Sigma}',q'}) \\
&= p(\mathbf{\Sigma}' | a'_{\mathbf{\Sigma}',1}, \ldots, a'_{\mathbf{\Sigma}',q'}) p(a'_{\mathbf{\Sigma}',1}, \ldots, a'_{\mathbf{\Sigma}',q'}) \\
&\propto |\mathbf{\Sigma}'|^{-\frac{\nu'+2q'}{2}} \exp\left\{ -\frac{1}{2} \text{tr}\left(2\nu' \text{diag}\left(1/a'_{\mathbf{\Sigma}',1}, \ldots, 1/a'_{\mathbf{\Sigma}',q'} \right) \mathbf{\Sigma}'^{-1} \right) \right\} \\
&\quad \times \prod_{i'=1}^{q'} a_{\mathbf{\Sigma}',i'}^{\prime -3/2} \exp\left\{ -\frac{1}{2a'_{\mathbf{\Sigma}',i'} A^2_{\mathbf{\Sigma}'}} \right\} \\
&\propto \prod_{i'=1}^{q'} a_{\mathbf{\Sigma}',i'}^{\prime -3/2} \exp\Bigg\{ -\frac{1}{2} \sum_{i'=1}^{q'} \left(\frac{1}{a'_{\mathbf{\Sigma}',i'} A^2_{\mathbf{\Sigma}'}} \right) \\
&\quad - \frac{1}{2} \text{tr}\left(2\nu' \text{diag}\left(1/a'_{\mathbf{\Sigma}',1}, \ldots, 1/a'_{\mathbf{\Sigma}',q'} \right) \mathbf{\Sigma}'^{-1} \right) \Bigg\} \\
&= \prod_{i'=1}^{q'} a_{\mathbf{\Sigma}',i'}^{\prime -3/2} \exp\left\{ -\frac{1}{2} \sum_{i'=1}^{q'} \left(\frac{1}{a'_{\mathbf{\Sigma}',i'} A^2_{\mathbf{\Sigma}'}} + 2\nu' \frac{\mathbf{\Sigma}'^{-1}_{i'i'}}{a'_{\mathbf{\Sigma}',i'}} \right) \right\} \\
&\propto \prod_{i'=1}^{q'} a_{\mathbf{\Sigma}',i'}^{\prime -3/2} \exp\left\{ -\frac{1}{2} \sum_{i'=1}^{q'} \frac{1}{a'_{\mathbf{\Sigma}',i'}} \left(\frac{1}{A^2_{\mathbf{\Sigma}'}} + 2\nu' \mathbf{\Sigma}'^{-1}_{i'i'} \right) \right\}.
\end{aligned}
$$

Hence,

$$
a'_{\mathbf{\Sigma}',i'} | \mathbf{\Sigma}'_{i'i'} \sim \text{Inverse-}\chi^2\left(1, \frac{1}{A^2_{\mathbf{\Sigma}'}} + 2\nu' \mathbf{\Sigma}'^{-1}_{i'i'} \right), \quad \text{for } i' = 1, \ldots, q'.
$$

The q^* density on log scale is

$$\begin{aligned}
\log q^*(a'_{\boldsymbol{\Sigma}',i'}) &= \mathbb{E}_{q(rest)} \log p(a'_{\boldsymbol{\Sigma}',i'}|rest) + \text{const} \\
&= \mathbb{E}_{q(\boldsymbol{\Sigma}')} \log p(a'_{\boldsymbol{\Sigma}',i'}|\boldsymbol{\Sigma}'_{i'i'}) + \text{const} \\
&= \mathbb{E}_{q(\boldsymbol{\Sigma}')} \left[-\frac{3}{2} \log a'_{\boldsymbol{\Sigma}',i'} - \frac{1}{a'_{\boldsymbol{\Sigma}',i'}} \left(\frac{1}{A^2_{\boldsymbol{\Sigma}'}} + 2\nu' \boldsymbol{\Sigma}'^{-1}_{i'i'} \right) \right] + \text{const} \\
&= -\frac{3}{2} \log a'_{\boldsymbol{\Sigma}',i'} - \frac{1}{a'_{\boldsymbol{\Sigma}',i'}} \left(\frac{1}{A^2_{\boldsymbol{\Sigma}'}} + 2\nu' \mathbb{E}_{q(\boldsymbol{\Sigma}')} \left[\boldsymbol{\Sigma}'^{-1}_{i'i'} \right] \right) + \text{const}.
\end{aligned}$$

Thus,

$$q^*(a'_{\boldsymbol{\Sigma}',i'}) \sim \text{Inverse-}\chi^2 \left(1, \frac{1}{A^2_{\boldsymbol{\Sigma}'}} + 2\nu' \mathbb{E}_{q(\boldsymbol{\Sigma}')} \left[\boldsymbol{\Sigma}'^{-1}_{i'i'} \right] \right).$$

Equivalently, for $a_{\boldsymbol{\Sigma},1}, \ldots, a_{\boldsymbol{\Sigma},q}$ we have

$$a_{\boldsymbol{\Sigma},i}|\boldsymbol{\Sigma}_{ii} \sim \text{Inverse-}\chi^2 \left(1, \frac{1}{A^2_{\boldsymbol{\Sigma}}} + 2\nu \boldsymbol{\Sigma}^{-1}_{ii} \right), \quad \text{for } i = 1, \ldots, q,$$

and

$$q^*(a_{\boldsymbol{\Sigma},i}) \sim \text{Inverse-}\chi^2 \left(1, \frac{1}{A^2_{\boldsymbol{\Sigma}}} + 2\nu \mathbb{E}_{q(\boldsymbol{\Sigma})} \left[\boldsymbol{\Sigma}^{-1}_{ii} \right] \right).$$

$\boldsymbol{\Sigma}'$

The full conditional distribution is defined as

$$
\begin{aligned}
p(\boldsymbol{\Sigma}'|\text{rest}) &= p(\boldsymbol{\Sigma}'|\text{Markov blanket of } \boldsymbol{\Sigma}') \\
&= p(\boldsymbol{\Sigma}'|\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}, a'_{\boldsymbol{\Sigma}',1}, \ldots, a'_{\boldsymbol{\Sigma}',q'}, \boldsymbol{\Sigma}) \\
&\propto p(\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}, a'_{\boldsymbol{\Sigma}',1}, \ldots, a'_{\boldsymbol{\Sigma}',q'}, \boldsymbol{\Sigma}', \boldsymbol{\Sigma}) \\
&\propto p(\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}, |\boldsymbol{\Sigma}', \boldsymbol{\Sigma}) p(\boldsymbol{\Sigma}'|a'_{\boldsymbol{\Sigma}',1}, \ldots, a'_{\boldsymbol{\Sigma}',q'}) p(\boldsymbol{\Sigma}|a_{\boldsymbol{\Sigma},1}, \ldots, a_{\boldsymbol{\Sigma},q}) \\
&\propto p(\boldsymbol{\beta}) p(\mathbf{u}'|\boldsymbol{\Sigma}') p(\mathbf{u}|\boldsymbol{\Sigma}) p(\boldsymbol{\Sigma}'|a'_{\boldsymbol{\Sigma}',1}, \ldots, a'_{\boldsymbol{\Sigma}',q'}) p(\boldsymbol{\Sigma}|a_{\boldsymbol{\Sigma},1}, \ldots, a_{\boldsymbol{\Sigma},q}) \\
&\propto p(\mathbf{u}'|\boldsymbol{\Sigma}') p(\boldsymbol{\Sigma}'|a'_{\boldsymbol{\Sigma}',1}, \ldots, a'_{\boldsymbol{\Sigma}',q'}) \\
&\propto |\boldsymbol{\Sigma}'|^{-\frac{m'}{2}} \exp\left\{-\frac{1}{2}\mathbf{u}'^{\intercal}\boldsymbol{\Sigma}'^{-1}\mathbf{u}'\right\} \\
&\quad |\boldsymbol{\Sigma}'|^{-\frac{\nu'+2q'}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(2\nu'\text{diag}\left(1/a'_{\boldsymbol{\Sigma}',1}, \ldots, 1/a'_{\boldsymbol{\Sigma}',q'}\right)\boldsymbol{\Sigma}'^{-1}\right)\right\} \\
&= |\boldsymbol{\Sigma}'|^{-\frac{m'+\nu'+2q'}{2}} \exp\Bigg\{-\frac{1}{2}\Bigg[\sum_{i'=1}^{m'} \text{tr}\left(\mathbf{u}'^{\intercal}_{i'}\boldsymbol{\Sigma}'^{-1}\mathbf{u}'_{i'}\right) \\
&\quad + \text{tr}\left(2\nu'\text{diag}\left(1/a'_{\boldsymbol{\Sigma}',1}, \ldots, 1/a'_{\boldsymbol{\Sigma}',q'}\right)\boldsymbol{\Sigma}'^{-1}\right)\Bigg]\Bigg\} \\
&= |\boldsymbol{\Sigma}'|^{-\frac{m'+\nu'+2q'}{2}} \exp\Bigg\{-\frac{1}{2}\Bigg[\text{tr}\sum_{i'=1}^{m'} \mathbf{u}'_{i'}\mathbf{u}'^{\intercal}_{i'}\boldsymbol{\Sigma}'^{-1} \\
&\quad + \text{tr}\left(2\nu'\text{diag}\left(1/a'_{\boldsymbol{\Sigma}',1}, \ldots, 1/a'_{\boldsymbol{\Sigma}',q'}\right)\boldsymbol{\Sigma}'^{-1}\right)\Bigg]\Bigg\} \\
&= |\boldsymbol{\Sigma}'|^{-\frac{m'+\nu'+2q'}{2}} \exp\Bigg\{-\frac{1}{2}\Bigg[\text{tr}\left(\sum_{i'=1}^{m'} \mathbf{u}'_{i'}\mathbf{u}'^{\intercal}_{i'} \right. \\
&\quad \left. + 2\nu'\text{diag}\left(1/a'_{\boldsymbol{\Sigma}',1}, \ldots, 1/a'_{\boldsymbol{\Sigma}',q'}\right)\right)\boldsymbol{\Sigma}'^{-1}\Bigg]\Bigg\}.
\end{aligned}
$$

Hence,

$$
\boldsymbol{\Sigma}'|\text{rest} \sim \text{Inverse-Whishart}\left(m' + \nu' + q' - 1, \sum_{i'=1}^{m'} \mathbf{u}'_{i'}\mathbf{u}'^{\intercal}_{i'} + 2\nu'\text{diag}\left(1/a'_{\boldsymbol{\Sigma}',1}, \ldots, 1/a'_{\boldsymbol{\Sigma}',q'}\right)\right).
$$

The q^* density on log scale is

$$\begin{aligned}
\log q^*(\mathbf{\Sigma}') &= \mathbb{E}_{q(rest)} \log p(\mathbf{\Sigma}'|rest) + \text{const} \\
&= \mathbb{E}_{q(\mathbf{u}', a'_{\mathbf{\Sigma}',1}, \ldots, a'_{\mathbf{\Sigma}',q'})} \log p(\mathbf{\Sigma}'|\mathbf{u}', a'_{\mathbf{\Sigma}',1}, \ldots, a'_{\mathbf{\Sigma}',q'}) + \text{const} \\
&= \mathbb{E}_{q(\mathbf{u}', a'_{\mathbf{\Sigma}',1}, \ldots, a'_{\mathbf{\Sigma}',q'})} \left[-\frac{m' + \nu' + 2q'}{2} \log |\mathbf{\Sigma}'| - \frac{1}{2} \text{tr} \left(\mathbf{B}' \mathbf{\Sigma}'^{-1} \right) \right] + \text{const} \\
&= -\frac{m' + \nu' + 2q'}{2} \log |\mathbf{\Sigma}'| - \frac{1}{2} \text{tr} \left(\mathbb{E}_{q(\mathbf{u}', a'_{\mathbf{\Sigma}',1}, \ldots, a'_{\mathbf{\Sigma}',q'})} [\mathbf{B}'] \, \mathbf{\Sigma}'^{-1} \right) + \text{const},
\end{aligned}$$

where $\mathbf{B}' = \sum_{i'=1}^{m'} \mathbf{u}'_{i'} \mathbf{u}'^{\intercal}_{i'} + 2\nu' \text{diag} \left(1/a'_{\mathbf{\Sigma}',1}, \ldots, 1/a'_{\mathbf{\Sigma}',q'} \right)$.

Thus,

$$q^*(\mathbf{\Sigma}') \sim \text{Inverse-Whishart} \left(m' + \nu' + q' - 1, \mathbb{E}_{q(\mathbf{u}', a'_{\mathbf{\Sigma}',1}, \ldots, a'_{\mathbf{\Sigma}',q'})} [\mathbf{B}'] \right).$$

Equivalently, for $\mathbf{\Sigma}$ we have

$$\mathbf{\Sigma}|\text{rest} \sim \text{Inverse-Whishart} \left(m + \nu + q - 1, \sum_{i=1}^{m} \mathbf{u}_i \mathbf{u}_i^{\intercal} + 2\nu \text{diag} \left(1/a_{\mathbf{\Sigma},1}, \ldots, 1/a_{\mathbf{\Sigma},q} \right) \right),$$

and

$$q^*(\mathbf{\Sigma}) \sim \text{Inverse-Whishart} \left(m + \nu + q - 1, \mathbb{E}_{q(\mathbf{u}, a_{\mathbf{\Sigma},1}, \ldots, a_{\mathbf{\Sigma},q})} [\mathbf{B}] \right),$$

where $\mathbf{B} = \sum_{i=1}^{m} \mathbf{u}_i \mathbf{u}_i^{\intercal} + 2\nu \text{diag} \left(1/a_{\mathbf{\Sigma},1}, \ldots, 1/a_{\mathbf{\Sigma},q} \right)$.

$\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}$

The full conditional distribution is defined as

$$
\begin{aligned}
p(\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}|\text{rest}) &= p(\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}|\text{Markov blanket of } \boldsymbol{\beta}, \mathbf{u}', \mathbf{u}) \\
&= p(\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}|\mathbf{y}, \sigma_\varepsilon^2, \boldsymbol{\Sigma}', \boldsymbol{\Sigma}) \\
&\propto p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}, \sigma_\varepsilon^2) p(\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}|\boldsymbol{\Sigma}', \boldsymbol{\Sigma}) \\
&\propto \exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{C}\mathbf{a})^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}(\mathbf{y}-\mathbf{C}\mathbf{a}) -\right\} \exp\left\{-\frac{1}{2}\mathbf{a}^\intercal\boldsymbol{\Lambda}^{-1}\mathbf{a}\right\} \\
&= \exp\left\{-\frac{1}{2}\left[(\mathbf{y}-\mathbf{C}\mathbf{a})^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}(\mathbf{y}-\mathbf{C}\mathbf{a}) + \mathbf{a}^\intercal\boldsymbol{\Lambda}^{-1}\mathbf{a}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\mathbf{y}^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{y} - 2\mathbf{y}^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{C}\mathbf{a}\right.\right. \\
&\quad \left.\left. + \mathbf{a}^\intercal\mathbf{C}^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{C}\mathbf{a} + \mathbf{a}^\intercal\boldsymbol{\Lambda}^{-1}\mathbf{a}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\mathbf{a}^\intercal\left(\mathbf{C}^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{C} + \boldsymbol{\Lambda}^{-1}\right)\mathbf{a}\right.\right. \\
&\quad \left.\left. - 2\mathbf{a}^\intercal\mathbf{C}^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{y} + \mathbf{y}^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{y}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\mathbf{a} - \left(\mathbf{C}^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{C} + \boldsymbol{\Lambda}^{-1}\right)^{-1}\mathbf{C}^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{y}\right)\right. \\
&\quad \left(\mathbf{C}^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{C} + \boldsymbol{\Lambda}^{-1}\right)\left(\mathbf{a} - \left(\mathbf{C}^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{C} + \boldsymbol{\Lambda}^{-1}\right)^{-1}\right. \\
&\quad \left.\left.\mathbf{C}^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{y}\right)\right\},
\end{aligned}
$$

where $\mathbf{C} = [\mathbf{X}, \mathbf{Z}', \mathbf{Z}]$, and $\mathbf{a} = \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u}' \\ \mathbf{u} \end{bmatrix}$ and $\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Sigma}_\beta & 0 & 0 \\ 0 & \boldsymbol{\Sigma}' & 0 \\ 0 & 0 & \boldsymbol{\Sigma} \end{pmatrix}$. Thus,

$$
\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}|\boldsymbol{\Sigma}', \boldsymbol{\Sigma} \sim N\left(\left(\mathbf{C}^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{C} + \boldsymbol{\Lambda}^{-1}\right)^{-1}\mathbf{C}^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{y}, \left(\mathbf{C}^\intercal \left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{C} + \boldsymbol{\Lambda}^{-1}\right)^{-1}\right).
$$

The q^* density on log scale is

$$\begin{aligned}
\log q^*(\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}) &= \mathbb{E}_{q(rest)} \log p(\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}|rest) + \text{const} \\
&= \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \log p(\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}|\boldsymbol{\Sigma}', \boldsymbol{\Sigma}) + \text{const} \\
&= \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \left[-\frac{1}{2} \left(\mathbf{a}^\intercal \left(\mathbf{C}^\intercal \left(\sigma_\varepsilon^2 \mathbf{I} \right)^{-1} \mathbf{C} + \boldsymbol{\Lambda}^{-1} \right) \mathbf{a} \right. \right. \\
&\quad \left. \left. -2\mathbf{a}^\intercal \mathbf{C}^\intercal \left(\sigma_\varepsilon^2 \mathbf{I} \right)^{-1} \mathbf{y} \right) \right] + \text{const} \\
&= -\frac{1}{2} \left(\mathbf{a}^\intercal \left(\mathbf{C}^\intercal \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \left[\left(\sigma_\varepsilon^2 \mathbf{I} \right)^{-1} \right] \mathbf{C} + \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \left[\boldsymbol{\Lambda}^{-1} \right] \right) \mathbf{a} \right. \\
&\quad \left. -2\mathbf{a}^\intercal \mathbf{C}^\intercal \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \left[\left(\sigma_\varepsilon^2 \mathbf{I} \right)^{-1} \right] \mathbf{y} \right) + \text{const.}
\end{aligned}$$

Thus,

$$\begin{aligned}
q^*(\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}) \sim N \Bigg(& \left(\mathbf{C}^\intercal \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \left[\left(\sigma_\varepsilon^2 \mathbf{I} \right)^{-1} \right] \mathbf{C} + \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \left[\boldsymbol{\Lambda}^{-1} \right] \right)^{-1} \mathbf{C}^\intercal \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \left[\left(\sigma_\varepsilon^2 \mathbf{I} \right)^{-1} \right] \mathbf{y}, \\
& \left(\mathbf{C}^\intercal \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \left[\left(\sigma_\varepsilon^2 \mathbf{I} \right)^{-1} \right] \mathbf{C} + \mathbb{E}_{q(\boldsymbol{\Sigma}', \boldsymbol{\Sigma})} \left[\boldsymbol{\Lambda}^{-1} \right] \right)^{-1} \Bigg).
\end{aligned}$$

σ_ε^2

The full conditional distribution is defined as

$$\begin{aligned}
p(\sigma_\varepsilon^2|\text{rest}) &= p(\sigma_\varepsilon^2|\text{Markov blanket of } \sigma_\varepsilon^2) \\
&= p(\sigma_\varepsilon^2|\mathbf{y}, a_\varepsilon, \boldsymbol{\beta}, \mathbf{u}', \mathbf{u}) \\
&\propto p(\mathbf{y}, \boldsymbol{\beta}, \mathbf{u}', \mathbf{u}, \sigma_\varepsilon^2, a_\varepsilon) \\
&\propto p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}', \mathbf{u}, \sigma_\varepsilon^2) p(\sigma_\varepsilon^2|a_\varepsilon) \\
&\propto |\sigma_\varepsilon^2 \mathbf{I}|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{C}\mathbf{a})^\intercal (\sigma_\varepsilon^2 \mathbf{I})^{-1} (\mathbf{y} - \mathbf{C}\mathbf{a}) \right\} (\sigma_\varepsilon^2)^{-3/2} \exp \left\{ -\frac{1}{2a_\varepsilon \sigma_\varepsilon^2} \right\} \\
&= (\sigma_\varepsilon^2)^{-\left(\frac{n+1}{2}\right)-1} \exp \left\{ -\frac{1}{2\sigma_\varepsilon^2} \left[\frac{1}{a_\varepsilon} + \|\mathbf{y} - \mathbf{C}\mathbf{a}\|^2 \right] \right\},
\end{aligned}$$

where $\mathbf{C} = [\mathbf{X}, \mathbf{Z}', \mathbf{Z}]$ and $\mathbf{a} = \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u}' \\ \mathbf{u} \end{bmatrix}$.

Hence,

$$\sigma_\varepsilon^2|a_\varepsilon \sim \text{Inverse-}\chi^2 \left(n + 1, \frac{1}{a_\varepsilon} + \|\mathbf{y} - \mathbf{C}\mathbf{a}\|^2 \right).$$

The q^* density on log scale is

$$\begin{aligned}\log q^*(\sigma_\varepsilon^2) &= \mathbb{E}_{q(rest)} \log p(\sigma_\varepsilon^2|rest) + \text{const} \\ &= \mathbb{E}_{q(a_\varepsilon)} \log p(\sigma_\varepsilon^2|a_\varepsilon) + \text{const} \\ &= -\frac{1}{2\sigma_\varepsilon^2}\mathbb{E}_{q(a_\varepsilon)}\left[\frac{1}{a_\varepsilon} + \|\mathbf{y} - \mathbf{Ca}\|^2\right] + \text{const.}\end{aligned}$$

Thus,

$$q^*(\sigma_\varepsilon^2) \sim \text{Inverse-}\chi^2\left(n+1, \mathbb{E}_{q(a_\varepsilon)}\left[\frac{1}{a_\varepsilon} + \|\mathbf{y} - \mathbf{Ca}\|^2\right]\right).$$

Bibliography

Akaike, H. (1974) A new look at the statistical model identification. *IEEE transactions on automatic control* **19**(6), 716–723.

Baayen, R. H., Davidson, D. J. and Bates, D. M. (2008) Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* **59**(4), 390–412.

Bishop, C. M. (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer New York.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017) Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**(518), 859–877.

de Boor, C. (2001) *A Practical Guide to Splines*. Springer New York.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017) Stan: A probabilistic programming language. *Journal of statistical software* **76**(1).

Dal Maso, L., Torelli, N., Biancotto, E., Di Maso, M., Gini, A., Franchin, G., Levi, F., La Vecchia, C., Serraino, D. and Polesel, J. (2016) Combined effect of tobacco smoking and alcohol drinking in the risk of head and neck cancers: a re-analysis of case–control studies using bi-dimensional spline models. *European Journal of Epidemiology* **31**(4), 385–393.

Denison, D., Mallick, B. and Smith, A. (1998) Automatic bayesian curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(2), 333–350.

DiMatteo, I., Genovese, C. R. and Kass, R. E. (2001) Bayesian curve-fitting with free-knot splines. *Biometrika* **88**(4), 1055–1071.

Friedman, J., Hastie, T. and Tibshirani, R. (2001) *The Elements of Statistical Learning*. Springer New York.

Gabry, J. and Mahr, T. (2018) *bayesplot: Plotting for Bayesian Models*. R package version 1.6.0.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M. and Gelman, A. (2017) Visualization in Bayesian workflow. *arXiv e-prints* .

Gasparrini, A., Armstrong, B. and Kenward, M. G. (2010) Distributed lag non-linear models. *Statistics in Medicine* **29**(21), 2224–2234.

Gasparrini, A., Scheipl, F., Armstrong, B. and Kenward, M. G. (2017) A penalized framework for distributed lag non-linear models. *Biometrics* **73**(3), 938–948.

Gelman, A. *et al.* (2006) Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1**(3), 515–534.

Gelman, A. and Hill, J. (2006) *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Hwang, J. and Vehtari, A. (2014) Understanding predictive information criteria for bayesian models. *Statistics and Computing* **24**(6), 997–1016.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y. S. (2008) A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics* **2**(4), 1360–1383.

Han, L. and Neumann, M. (2006) Effect of dimensionality on the nelder–mead simplex method. *Optimization Methods and Software* **21**(1), 1–16.

Hashibe, M., Brennan, P., Benhamou, S., Castellsague, X., Chen, C., Curado, M. P., Maso, L. D., Daudt, A. W., Fabianova, E., Wünsch-Filho, V. *et al.* (2007) Alcohol drinking in never users of tobacco, cigarette smoking in never drinkers, and the risk of head and neck cancer: pooled analysis in the international head and neck cancer epidemiology consortium. *Journal of the National Cancer Institute* **99**(10), 777–789.

Hashibe, M., Brennan, P., Chuang, S.-c., Boccia, S., Castellsague, X., Chen, C., Curado, M. P., Dal Maso, L., Daudt, A. W., Fabianova, E. *et al.* (2009) Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the international head and neck cancer epidemiology consortium. *Cancer Epidemiology Biomarkers & Prevention* **18**(2), 541–550.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. CRC Press.

Hoffman, M. and Gelman, A. (2014) The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**, 1593–1623.

Hosmer Jr, D. W. and Lemeshow, S. (2005) *Applied logistic regression*. John Wiley & Sons.

Huang, A., Wand, M. P. *et al.* (2013) Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis* **8**(2), 439–452.

IARC (1986) Tobacco smoking (IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans, vol. 38). *Lyon, France: IARCPress* .

INHANCE, C. (2004) International Head And Neck Cancer Epidemiology consortium. `http://http://www.inhance.utah.edu`.

Jeon, M., Rijmen, F. and Rabe-Hesketh, S. (2017) A variational maximization-maximization algorithm for generalized linear mixed models with crossed random effects. *Psychometrika* **82**(3), 693–716.

Jordan, M. I., Jaakkola, T. S., Saul, L. K. and Park, F. (1999) An Introduction to Variational Methods for Graphical Models. *Machine Learning* **37**(2), 183–233.

Jupp, D. L. (1975) The "lethargy" theorem - a property of approximation by γ-polynomials. *Journal of Approximation Theory* **14**(3), 204–217.

Lee, C. Y. Y. and Wand, M. P. (2016a) Streamlined mean field variational bayes for longitudinal and multilevel data analysis. *Biometrical Journal* **58**(4), 868–895.

Lee, C. Y. Y. and Wand, M. P. (2016b) Variational methods for fitting complex bayesian mixed effects models to health data. *Statistics in Medicine* **35**(2), 165–188.

Leffondré, K., Abrahamowicz, M., Siemiatycki, J. and Rachet, B. (2002) Modeling smoking history: A comparison of different approaches. *American Journal of Epidemiology* **156**(9), 813–823.

Lubin, J. and Caporaso, N. (2013) Misunderstandings in the misconception on the use of pack-years in analysis of smoking. *British Journal of Cancer* **108**(5), 1218–1220.

Lubin, J. H., Caporaso, N., Wichmann, H. E., Schaffrath-Rosario, A. and Alavanja, M. C. (2007) Cigarette smoking and lung cancer: modeling effect modification of total exposure and intensity. *Epidemiology* **18**(5), 639–648.

Lubin, J. H., Gaudet, M. M., Olshan, A. F., Kelsey, K., Boffetta, P., Brennan, P., Castellsague, X., Chen, C., Curado, M. P., Dal Maso, L. *et al.* (2010) Body mass index, cigarette smoking, and alcohol consumption and cancers of the oral cavity, pharynx, and larynx: modeling odds ratios in pooled case-control data. *American Journal of Epidemiology* **171**(12), 1250–1261.

Lubin, J. H., Purdue, M., Kelsey, K., Zhang, Z.-F., Winn, D., Wei, Q., Talamini, R., Szeszenia-Dabrowska, N., Sturgis, E. M., Smith, E. *et al.* (2009) Total exposure and exposure rate effects for alcohol and smoking and risk of head and neck cancer: a pooled analysis of case-control studies. *American Journal of Epidemiology* **170**(8), 937–947.

Mao, W. and Zhao, L. H. (2003) Free-knot polynomial splines with confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(4), 901–919.

McCullagh, P. and Nelder, J. (1983) *Generalised Linear Models*. London Chapman and Hall.

Molinari, N., Durand, J.-F. and Sabatier, R. (2004) Bounded optimal knots for regression splines. *Computational Statistics & Data Analysis* **45**(2), 159–178.

Nolan, T. H., Menictas, M. and Wand, M. P. (2018) Streamlined computing for variational inference with higher level random effects. *In progress* .

Nolan, T. H. and Wand, M. P. (2018) Solutions to sparse multilevel matrix problems. *In progress* .

O'Hara, R. B., Sillanpää, M. J. *et al.* (2009) A review of bayesian variable selection methods: what, how and which. *Bayesian Analysis* **4**(1), 85–117.

Ormerod, J. T. and Wand, M. P. (2010) Explaining variational approximations. *The American Statistician* **64**(2), 140–153.

Pearl, J. (2009) *Causality*. Cambridge University Press.

Peto, J. (2012) That the effects of smoking should be measured in pack-years: misconceptions 4. *British Journal of Cancer* **107**(3), 406–407.

Piironen, J. and Vehtari, A. (2017) Comparison of bayesian predictive methods for model selection. *Statistics and Computing* **27**(3), 711–735.

Polesel, J., Dal Maso, L., Bagnardi, V., Zucchetto, A., Zambon, A., Levi, F., La Vecchia, C. and Franceschi, S. (2005) Estimating dose-response relationship between ethanol and risk of cancer using regression spline models. *International Journal of Cancer* **114**(5), 836–841.

Polesel, J., Talamini, R., La Vecchia, C., Levi, F., Barzan, L., Serraino, D., Franceschi, S. and Dal Maso, L. (2008) Tobacco smoking and the risk of upper aero-digestive tract cancers: A reanalysis of case–control studies using spline models. *International Journal of Cancer* **122**(10), 2398–2402.

R Core Team (2018) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robert, C. P. and Casella, G. (2004) *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer New York.

Rosenberg, P. S., Katki, H., Swanson, C. A., Brown, L. M., Wacholder, S. and Hoover, R. N. (2003) Quantifying epidemiologic risk factors using non-parametric regression: model selection remains the greatest challenge. *Statistics in Medicine* **22**(21), 3369–3381.

Rothman, K. J., Greenland, S. and Lash, T. L. (2008) *Modern Epidemiology*. Lippincott Williams & Wilkins.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric regression*. Cambridge University Press.

Schöllnberger, H., Manuguerra, M., Bijwaard, H., Boshuizen, H., Altenburg, H., Rispens, S., Brugmans, M. and Vineis, P. (2006) Analysis of epidemiological cohort data on smoking effects and lung cancer with a multi-stage cancer model. *Carcinogenesis* **27**(7), 1432–1444.

Smith, M. and Kohn, R. (1996) Nonparametric regression using bayesian variable selection. *Journal of Econometrics* **75**(2), 317–343.

Smith, T. J. (1992) Occupational exposure and dose over time: limitations of cumulative exposure. *American Journal of Industrial Medicine* **21**(1), 35–51.

Varadhan, R. and Gilbert, P. (2009) BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software* **32**(4), 1–26.

Varadhan, R., University, J. H., Borchers, H. W. and Research., A. C. (2018) *dfoptim: Derivative-Free Optimization*. R package version 2018.2-1.

Wand, M. P. (2017) Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *Journal of the American Statistical Association* **112**(517), 137–168.

Watanabe, S. (2010) Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* **11**(Dec), 3571–3594.

Wood, S. N. (2006) *Generalized additive models: an introduction with R*. Chapman and Hall.

Gioia Di Credico

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4174
Cell. +39 334 9251327
e-mail: gioia.dicredico@phd.unipd.it

Current Position

Since October 2015;
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Some developments in semiparametric and cross-classified multilevel models
Supervisor: Prof. Francesco Pauli
Co-supervisor: Prof. Nicola Torelli.

Research interests

- Semiparametric and multilevel models
- Bayesian statistics
- Streamlined variational inference

Education

January 2011 – March 2014
Master (*laurea specialistica/magistrale***) degree in Statistics and Decision Sciences**.
University of Rome La Sapienza, Department of Statistical Sciences
Title of dissertation: "Classification of Mediterranean and Black Sea Lagoons using multimetric indices for benthic macroinvertebrates "
Supervisor: Prof. Giovanna Jona Lasinio
Final mark: 110/110 cum laude

October 2007 – December 2010
Bachelor degree (*laurea triennale***) in Statistics and Informatics**.
University of Rome La Sapienza, Department of Statistical Sciences
Title of dissertation: "Diseguaglianza economica e preferenze politiche: un confronto tra paesi" ("Economic inequalities and political preferences: a comparison among Countries")
Supervisor: Prof. Maria Grazia Pittau
Final mark: 110/110 cum laude.

Visiting periods

April 2018 – June 2018
University of Technology Sydney,
Sydney, Australia.
Supervisor: Prof. Matt Wand

September 2012 – March 2013
Université Paris Dauphine IX,
Paris, France.
Supervisor: Prof. Christian Robert

October 2009 – March 2010
Erasmus University Rotterdam,
Rotterdam, Netherland.

Work experience

July 2014 – January 2015
Major Bit Consulting S.r.l.
Statistical consultant at Poste Italiane.

Computer skills

- Language & software: Stan, R, SAS, laTex, Office, Winbugs
- Operating Systems: Windows, macOS, and Linux
- Databases: MySQL

Language skills

Language1: Italian (native language);
Language2: English (fluent);

Publications

Articles in journals

Di Credico, G. Pauli, F., Torelli N. (2018). "Bayesian estimation of number and position of knots in regression spline". *Book of Short Papers SIS 2018 (Abbruzzo, A., Piacentino, D., Chiodi, M., and Brentari, E., editors)*. ISBN: 9788891910233.

Di Credico, G., Edefonti, V., Polesel, J. et al. (2018). "Joint effects of intensity and duration of cigarette smoking on the risk of head and neck cancer: a bivariate spline model approach". *(In submission)*

Conference presentations

Di Credico, G. Pauli, F., Torelli N. (2018). Bayesian estimation of number and position of knots in regression splines. (Contributed) *49th Scientific meeting of the Italian Statistical Society (SIS)*, Palermo, Italy, 20–22 June 2018.

References

Prof. Francesco Pauli
Institution Università degli studi di Trieste - Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche
Address Via Tigor 22, Trieste
Phone: +39 040 558 2518
e-mail: francesco.pauli@deams.units.it

Prof. Nicola Torelli
Institution Università degli studi di Trieste - Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche
Address Via Tigor 22, Trieste
Phone: +39 040 558 7032
e-mail: nicola.torelli@deams.units.it