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 XXXV

Approximate inference for misspecified additive and mixed regression models

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Mauro Bernardi

Co-supervisore: Prof. Laura M. Sangalli

Co-supervisore: Prof. Alessio Farcomeni

Dottorando: Cristian Castiglione

January 11th, 2023

Abstract

Nowadays, the increasing dimension and complexity of real data problems pose hard theoretical and practical challenges to researchers working in any field of science. The study of complex non-linear, eventually non-observable, phenomena deserves the development of new mathematical and statistical techniques able to explain the actual empirical evidence, taking into account of different sources of information. On the other hand, technological constraints may limit our capabilities to process massive multidimensional datasets in a reasonable amount of time and not exceeding the available memory space. For these reasons, the continuous development of new flexible models, reliable estimation methods and efficient algorithms is of prominent importance from an applied point of view. In this thesis, we try to address some of these methodological issues, proposing two original contributions, an algorithmic one and a modelling one.

In the first part of this thesis, we consider the estimation of robust regression models within a Bayesian inferential framework. In this case, we propose a new deterministic variational approximation for general posterior distributions, which build upon alternative methods in the literature, improving their performances in terms of accuracy. We then discuss several extensions and generalizations, so that to enlarge the range of application of the proposed method.

In the second part of this thesis, we study a new quantile regression model for heterogeneous data gathered on spatial, possibly complex, domains. To this end, we adopt a nonparametric penalized regression approach with differential regularization, which allow us to incorporate additional spatial information in the form of a partial differential equation. Upon that, we develop a new computational estimation method and we analyze the theoretical and empirical properties of the proposed estimator, showing its comparative advantages with respect to state-of-the-art methods in the literature.

Sommario

Al giorno d'oggi, la crescente dimensionalità e complessità dei dati generati da problemi applicativi reali pone nuove sfide teoriche e pratiche ai ricercatori che operano in qualsiasi campo della scienza. Lo studio di fenomeni non lineari, e tal volta non osservabili, richiede lo sviluppo di nuove tecniche matematiche e statistiche in grado di spiegare l'evidenza empirica attuale tenendo conto di varie fonti d'informazione. D'altra parte, vincoli tecnologici possono limitare le nostre capacità di elaborazione d'insiemi di dati complessi in un tempo ragionevole e senza eccedere lo spazio di memoria disponibile. Per queste ragioni, il continuo sviluppo di modelli flessibili, metodi di stima robusti e algoritmi efficienti è di primaria importanza dal punto di vista applicativo. In questa tesi, cerchiamo di affrontare alcuni di questi problemi metodologici, proponendo due contributi originali, il primo algoritmico e il secondo modellistico.

Nella prima parte di questa tesi, viene considerata la stima di modelli di regressione robusti in una cornice inferenziale Bayesiana. In questo caso, proponiamo una nuova approssimazione variazionale deterministica per distribuzioni a posteriori generalizzate, la quale, costruendo su metodi esistenti in letteratura, ne migliora le prestazioni in termini di accuratezza. Discutiamo poi numerose estensioni e generalizzazioni, al fine di ampliare lo spettro di applicazione di tale metodo.

Nella seconda parte di questa tesi, studiamo un nuovo modello di regressione quantilica per dati raccolti su domini spaziali, possibilmente complessi. A tal fine, proponiamo un approccio di regressione nonparametrica penalizzata con regolarizzazione differenziale, la quale permette d'incorporare informazione spaziale aggiuntiva in forma di equazioni alle derivate parziali. Sviluppiamo poi questo problema, sia in termini computazionali che di analisi teorica ed empirica delle proprietà dello stimatore, mostrandone limiti e vantaggi comparati rispetto a metodi alternativi presenti in letteratura.

To my family

Acknowledgements

Among all the people who supported me during my PhD, I first want to give my deepest thank you to Mauro. You always encouraged me since I was working on my master thesis, transmitting me your passion and enthusiasm for research. I am really grateful of have had the opportunity to work with you.

I extend my sincere gratitude to Laura and Eleonora, for all the time you dedicated to me. Your advises, suggestions and rigorous scientific approach have been very instructive to me.

I also wish to acknowledge Alessio for his always useful and stimulating feedbacks. I hope to meet you soon in person.

I am also very thankful to Nicola Sartori, coordinator of the PhD program. I thank you for the willingness and kindness you always demonstrated to me and to all may colleagues in these challenging years, especially during the most complex periods of the pandemic.

My warmest hug goes to the friends of the XXXV cycle, who shared with me this adventure and who made this time so special. Thank you all Beppe, Caizhu, Erika, Fariborz, Mara, Marco, Nicolas, Pietro, Touqeer.

I am also thankful to all the XXXIV cycle. Even if I started my PhD one year after you, I have always felt a little bit part of your group. In particular, I want to thank Emanuele, Mattia and Lorenzo for your friendship and for the uncountable stimulating discussions we made together.

I could never forget also all the other friends I met during my PhD, and especially I would like to remember Andrea, Dede, Meme, Tommy and Luca. Thank you for your inestimable worth advises and stimulating feedbacks, I wish to become a brilliant researcher as you are.

I am, and I will always be, thankful to my family, to Mum and Dad, to my grandparents, and to Roberto for raising me and giving me all the support and love I needed. I love you so much.

Finally, I just want to thank my hometown, bachelor and master friends. You accompanied me in all my experiences, even if I always bother you with statistics and politics, I love you so much.

February 28th, 2023

Cristian Castiglione

Contents

List of Figures **xii**

List of Tables **xvii**

List of Algorithms **xix**

Notational conventions **xix**

Introduction **1**

Overview 3

Main contributions of the thesis 4

1 Minimum risk estimation and quantile regression **7**

1.1 Minimum risk estimators 7

1.2 Quantile regression in a nutshell 9

1.2.1 Asymmetric-Laplace pseudo-likelihood 11

1.2.2 Bayesian quantile regression 12

1.3 Estimation methods 14

1.3.1 Expectation-maximization 14

1.3.2 Gibbs sampling 17

1.3.3 Mean field variational Bayes 19

2 Non-conjugate regression via variational belief updating **25**

2.1 Introduction 25

2.2 Motivating problems and results 26

2.2.1 Variational data-augmentation 26

2.2.2 Parametric variational Bayes 32

2.3 Approximate belief updating 34

2.3.1 Model specification 35

2.3.2 Variational approximation 36

2.3.3 Optimization algorithm 39

2.4 Relevant models 40

2.4.1 Quantile regression 40

2.4.2 Expectile regression 42

2.4.3 Support vector classification 44

2.4.4 Support vector regression 46

2.4.5 Exponential family 47

2.5 Possible extensions 49
2.5.1 Additive models 49
2.5.2 Inducing shrinkage priors 50
2.5.3 Dynamic linear models 51
2.5.4 Spatial random fields 51
2.5.5 Frequentist mixed models 52
2.6 Simulation studies 53
2.6.1 Performance measures 54
2.6.2 First simulation setup: semiparametric regression 55
2.6.3 Second simulation setup: random intercept models 61
2.7 Probabilistic Load Forecasting 64

3 Spatial quantile regression with differential regularization 69
3.1 Introduction 69
3.2 Spatial quantile regression model 70
3.2.1 Problem-specific information and PDE specification 71
3.2.2 Infinite-dimensional estimation problem 72
3.3 Functional EM algorithm 75
3.3.1 Algorithm derivation 75
3.3.2 Constrained formulation 78
3.4 Finite element discretization 81
3.4.1 Finite element basis expansion 81
3.4.2 Finite-dimensional estimator 82
3.4.3 Finite-dimensional EM algorithm 84
3.4.4 Smoothing parameter selection 85
3.5 Large sample properties 86
3.5.1 Infinite-dimensional estimator 86
3.5.2 Finite-dimensional estimator 87
3.6 Inclusion of covariates 88
3.7 Simulation studies 92
3.7.1 First simulation setup: field over irregular domain 92
3.7.2 Second simulation setup: anisotropic field over regular domain . . 93
3.8 Switzerland rainfall data 97

Conclusions 103

Appendix A 109
A.1 Integration results 109
A.2 Optimal distributions and evidence lower bound 117

Appendix B 121
B.1 EM algorithm and discretization 121
B.2 Large sample properties 124

Bibliography **137**

List of Figures

1.1 Graphical representation of the variational Bayes search of the optimal density $q^*(\boldsymbol{\theta})$ within the functional space $\mathcal{Q}$ under two scenarios: (left) $\pi(\boldsymbol{\theta}|\mathbf{y}) \in \mathcal{Q}$, (right) $\pi(\boldsymbol{\theta}|\mathbf{y}) \notin \mathcal{Q}$. 20

2.1 Quantile regression ($\tau = 0.75$): ψ_r and Ψ_r functions for different values of η and ν. From left to right: $r = 0, 1, 2$. 41
2.2 Expectile regression ($\tau = 0.75$): ψ_r and Ψ_r functions for different values of η and ν. From left to right: $r = 0, 1, 2$. 43
2.3 Support vector classification: ψ_r and Ψ_r functions for different values of η and ν. From left to right: $r = 0, 1, 2$. 45
2.4 Support vector regression ($\epsilon = 1$): ψ_r and Ψ_r functions for different values of η and ν. From left to right: $r = 0, 1, 2$. 46
2.5 Poisson regression: ψ_r and Ψ_r functions for different values of η and ν. From left to right: $r = 0, 1, 2$. 48
2.6 Logistic regression: ψ_r and Ψ_r functions for different values of η and ν. From left to right: $r = 0, 1, 2$. 48
2.7 Posterior pointwise predictions (solid lines) and credibility bands (dashed lines) for the 90% quantile regression model described in the text. The estimates are obtained using a dataset from simulation setting A. (left) MFVB against MCMC. (right) SVB against MCMC. 58
2.8 Top left: evidence lower bound evolution over the algorithm iterations. Others: optimal posterior density functions of the parameters in the 90% quantile regression model described in the text. The estimates are obtained using one dataset from simulation setting A. The solid blue line is for MCMC, the dashed red one for SVB and the dash-dotted green one for MFVB. The percentages denote the correspondent marginal accuracy scores defined in (2.26). 59
2.9 Boxplots of the summary statistics for the simulation study described in the text (Section 2.6.1 and 2.6.2). Each column corresponds to a model. Each row corresponds to a performance index. Within each panel, we find three groups of paired boxplot which correspond to the three simulation settings, i.e., A, B, C (see Equation 2.27 and Table 2.2). Each boxplot is calculated over 100 replications, i.e. 100 simulated datasets. 60
2.10 Sampling distribution of the marginal accuracy score defined in (2.26) for the simulation setup described in Section 2.6.3. Each row corresponds to a model, each column corresponds to a simulation setting. The left column is for setting A, the right column is for setting B. 62

2.11 Sampling distribution of the $\log_{10}$-transformed execution time measured in seconds for the simulation setup described in Section 2.6.3. Each row corresponds to a model, each column corresponds to a simulation setting. The left column is for setting A, the right column is for setting B. 63
2.12 Pairwise scatter-plots of the power load consumption (in MWh) against the available covariates in the dataset, that are: (first row) the time (in days), the day of the year, the day of the week, (second row) the lagged power load, the temperature (in Celsius scale) and the smoothed temperature (in Celsius scale). 64
2.13 Distribution of the accuracy scores for the 19 estimated quantiles. For each quantile level, the boxplot is calculated over the individual accuracy scores of the 72 regression parameters in the model. 65
2.14 Estimated non-linear marginal effects (and credibility bands) for the available covariates in the analysis of 95% conditional quantile of the power load in the GEFCom2014 dataset. Top rows: MCMC estimated effects (blue curves) against MFVB estimated effects (orange curves). Bottom rows: MCMC estimated effects (blue curves) against SVB estimated effects (red curves). 66

3.1 A linear finite element basis function on a triangular mesh. 82
3.2 Horseshoe domain for the first simulation study. First row: mean surface, data locations, triangular discretization of the domain. Second row: standard deviation surfaces corresponding to three simulation settings A, B and C. 93
3.3 Boxplots of the RMSE over 100 replicates of the estimated spatial quantile fields in the first simulation setup. Each column corresponds to a quantile level (10%, 25%, 50%, 75%, 90%). Each row corresponds to a simulation setting (A, B, C). 94
3.4 Estimated 10%-quantile surfaces for one dataset generated in the first simulation setup, setting C. 95
3.5 Square domain for the second simulation setup. First row: mean surfaces in settings A, B, C. Second row: standard deviation surfaces in settings A, B, C. Third row: data locations and triangular discretization of the domain. 96
3.6 Boxplots of the RMSE over 100 replicates of the estimated spatial quantile fields in the second simulation setup. Each column corresponds to a quantile level (10%, 25%, 50%, 75%, 90%). Each row corresponds to a simulation setting (A, B, C). The colors correspond to different families of estimating methods. 97
3.7 Estimated 90%-quantile surfaces for one dataset generated in the second simulation setup, setting C. 98
3.8 Left: Switzerland rainfall data and triangular discretization of the domain. Colored dots indicate the measurement locations. Color and dimension of the dots are proportional to the rainfall intensity. Right: residual versus fitted value plot and normal quantile-quantile plot of the residuals. The fitted values and the residuals are obtained by using the anisotropic spatial regression model by Bernardi *et al.* (2018). 99

3.9 Estimated mean and τ-quantile fields for the Switzerland rainfall data. . 100

List of Tables

2.1 References for the estimation algorithms used in the estimation of the models considered along all the simulation studies in Section 2.6. 54
2.2 Non-linear functions characterizing the three simulation settings described in the text. 56
2.3 Performance measure comparison between SVB, MFVB and Laplace approximation based on the 3 simulation setting outlined in the text and in Table 2.2. 57

List of Algorithms

1 EM algorithm for quantile regression 17
2 Gibbs sampling for quantile regression 19
3 MFVB algorithm for quantile regression 23
4 SVB algorithm for variational inference in model (2.12) with prior (2.13) 39
5 Functional EM algorithm for nonparametric spatial quantile regression . 85
6 Functional EM algorithm for semiparametric spatial quantile regression . 91

Notational conventions

Throughout this Ph.D. thesis, lower-case Roman and Greek letters denote scalars. Lower-case Roman and Greek letters in boldface denote vectors. Unless specified otherwise, they are always assumed to be column vectors. Upper-case Roman and Greek letters in boldface denote matrices, whose entries are indicated with double-index subscripts. Blackboard-bold Roman letters indicate numeric sets.

We now list and clarify some notational conventions adopted in this Ph.D. thesis.

Vectors and matrices	
$\mathbb{N}$ (and $\mathbb{N}_+$)	The set of natural (and positive) numbers.
$\mathbb{R}$ (and $\mathbb{R}_+$)	The set of real (and positive) numbers.
$\mathbb{R}^d$ (and $\mathbb{R}^d_+$)	The set of real vectors of dimension d, for $d \in \mathbb{N}$.
$\mathbb{R}^{p\times q}$	The set of real matrices of dimension $p \times q$, for $p, q \in \mathbb{N}$.
a_i and $[\boldsymbol{a}]_i$	The i-th element of vector $\boldsymbol{a} \in \mathbb{R}^q$ for $i = 1, \dots, d$.
$\mathbf{A}_{ij}$ and $[\mathbf{A}]_{ij}$	The (i, j)-th element of matric $\mathbf{A} \in \mathbb{R}^{p\times q}$, for $i = 1, \dots, p$ and $j = 1, \dots, q$.
$\boldsymbol{a}_{-i}$	The vector of dimension $d-1$ obtained removing the i-th element of $\boldsymbol{a} \in \mathbb{R}^d$, for $d \in \mathbb{N}$.
$\mathbf{0}$ and $\mathbf{1}$	Vectors full of zeros and ones, respectively.
$\mathbf{O}$	Matrix full of zeros with generic dimensions.
$\mathbf{I}$	Identity matrix with generic dimensions.
$\boldsymbol{a}^\top$ and $\mathbf{A}^\top$	Transpose of a vector $\boldsymbol{a} \in \mathbb{R}^d$, or of a matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$.
$\mathbf{A}^{-1}$	Inverse of a non-singular matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, which satisfies $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.
$\mathrm{stack}(\cdot)$	For column vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$, $\mathrm{stack}(\boldsymbol{a}_1, \dots, \boldsymbol{a}_n)$ returns a unique vector stacking by column all its arguments.
$\mathrm{diag}(\cdot)$	For a vector $\boldsymbol{a} \in \mathbb{R}^d$, $\mathrm{diag}(\boldsymbol{a})$ returns a diagonal matrix with diagonal equal to $\boldsymbol{a}$.
$\mathrm{blockdiag}(\cdot)$	For a sequence of square matrices $\mathbf{A}_1, \dots, \mathbf{A}_n$, $\mathrm{blockdiag}(\mathbf{A}_1, \dots, \mathbf{A}_n)$ returns a block-diagonal matrix with diagonal blocks equal to $\mathbf{A}_1, \dots, \mathbf{A}_n$.
$\mathrm{vec}(\cdot)$	For a matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$, $\mathrm{vec}(\mathbf{A})$ returns the column vectorization of $\mathbf{A}$, that is a vector stacking all the columns of $\mathbf{A}$ from left to right.

$\mathrm{trace}(\cdot)$	For a square matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, $\mathrm{trace}(\mathbf{A})$ returns its trace $\sum_{i=1}^{d} \mathbf{A}_{ii}$.
$\det(\cdot)$	For a square matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$. $\det(\mathbf{A})$ returns its determinant.
$\Vert\cdot\Vert$	Generic sign denoting the norm of a vector or matrix.
$\Vert\cdot\Vert_p$	For a vector $\boldsymbol{a} \in \mathbb{R}^d$, $\Vert\boldsymbol{a}\Vert_p = (\sum_{j=1}^{d} \vert a_j\vert^p)^{1/p}$ returns the ℓ^p norm of $\boldsymbol{a}$. For a matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$, $\Vert\mathbf{A}\Vert_p = (\sum_{i=1}^{p}\sum_{j=1}^{q} \vert A_{ij}\vert^p)^{1/p}$ returns the ℓ^p norm of $\mathbf{A}$.
$\Vert\cdot\Vert_\infty$	For a vector $\boldsymbol{a} \in \mathbb{R}^d$, $\Vert\boldsymbol{a}\Vert_\infty = \max_j \vert a_j\vert$ returns the ℓ^∞ norm of $\boldsymbol{a}$. For a vector $\mathbf{A} \in \mathbb{R}^{p\times q}$, $\Vert\mathbf{A}\Vert_\infty = \max_{i,j} \vert A_{ij}\vert$ returns the ℓ^∞ norm of $\mathbf{A}$.
$\otimes$	For two matrices $\mathbf{A} \in \mathbb{R}^{p\times q}$ and $\mathbf{B} \in \mathbb{R}^{n\times m}$, $\mathbf{A}\otimes\mathbf{B}$ returns the $pn \times qm$ matrix obtained by the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$.
$+, -, \odot, /$	Elementwise operations between vectors and matrices with compatible dimensions.

Special symbols

$\equiv$	For any A and B, $A \equiv B$ means "A is defined as B".
$\leftarrow$	For any A and B, $A \leftarrow B$ means "A is updated to B".
$\propto$	For any A and B, $A \propto B$ means "A is proportional to B, up to a multiplicative constant".
$\approx$	For any A and B, $A \propto B$ means "A is approximated by B".
$\sim$	For a random variable x and a distribution D, $x \sim D$ means "x is distributed as D ".

Hilbert spaces

$L^p(\Omega)$	Space of functions $f : \Omega \to \mathbb{R}$ such that $\int_\Omega \vert f\vert^p < \infty$, for $p \in \mathbb{N}$.
$H^p(\Omega)$	Sobolev space of functions $f \in L^2(\Omega)$ having p weak derivatives in $L^2(\Omega)$, for $p \in \mathbb{N}$.
$\Vert\cdot\Vert_V$	For any function $f \in V$, $\Vert f\Vert_V$ returns a properly defined norm, or seminorm, of f in V.
∇f	For a scalar field $f : \mathbb{R}^d \to \mathbb{R}$, $\nabla f = \partial f/\partial \boldsymbol{x}$ returns the gradient of f.
$\nabla^2 f$	For a scalar field $f : \mathbb{R}^d \to \mathbb{R}$, $\nabla^2 f = \partial^2 f/\partial\boldsymbol{x}\partial\boldsymbol{x}^\top$ returns the Hessian of f.
Δf	For a scalar field $f : \mathbb{R}^d \to \mathbb{R}$, $\Delta f = \sum_{j=1}^{d} \partial^2 f/\partial x_j^2$ returns the Laplacian of f.
$\mathrm{div}(f)$	For a vector field $f : \mathbb{R}^d \to \mathbb{R}^d$, $\mathrm{div}(f) = \sum_{j=1}^{d} \partial f_j/\partial x_j$ returns the divergence of f.

Probability and statistics

$\pi(\cdot)$	A generic probability density function.
$q(\cdot)$	A generic approximate density function.
$\mathbb{P}(\cdot)$	For a generic random event E, $\mathbb{P}(E)$ returns its probability.
$\mathbb{E}(\cdot)$	For a generic random variable x, $\mathbb{E}(x)$ returns its expected value.

$\mathrm{Var}(\cdot)$	For a generic random variable x, $\mathrm{Var}(x)$ returns its variance. For a generic d-dimensional random vector $\boldsymbol{x}$, $\mathrm{Var}(\boldsymbol{x})$ returns its $d \times d$ variance-covariance matrix.
$\mathrm{Cov}(\cdot,\cdot)$	For two generic random variables x and y, $\mathrm{Cov}(x, y)$ returns their covariance. For two generic p- and q-dimensional random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})$ returns their $p \times q$ cross-covariance matrix.

Special functions	
$\mathbb{I}(\cdot)$ and $\mathbb{I}_A(\cdot)$	For any set A, $\mathbb{I}_A(x) = \mathbb{I}(x \in A)$ is the indicator function of A, which is equal to 1 if $x \in A$, zero otherwise.
$\phi(\cdot)$ and $\phi_d(\cdot)$	Probability density functions of a univariate and d-variate standard Gaussian random variables.
$\Phi(\cdot)$ and $\Phi_d(\cdot)$	Cumulative density functions of a univariate and d-variate standard Gaussian random variables.
$\phi_d(\cdot\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Probability density functions of a multivariate Gaussian random variable with mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and variance-covariance matrix $\boldsymbol{\Sigma} \in \mathbb{S}^d_{++}$.
$\delta_a(\cdot)$	The Dirac delta function centered in the point $a \in \Omega$, such that $f(a) = \int_\Omega f(t)\delta_a(t)\mathrm{d}t$, for any $f : \Omega \to \mathbb{R}$.
$\Gamma(\cdot)$	The Euler's Gamma function. $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$, $x > 0$.
$B(\cdot,\cdot)$	The Beta function. $B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t$, $x, y, > 0$.
$K_\nu(\cdot)$	Modified Bessel function of the second kind.
$\mathrm{logit}(\cdot)$	The logistic function, equal to $\mathrm{logit}(x) = \log\{x/(1-x)\}$, $x \in (0, 1)$.
$\mathrm{expit}(\cdot)$	The inverse of the logistic function, equal to $\mathrm{expit}(x) = e^x/(1+e^x)$, $x \in \mathbb{R}$.
$\rho_\tau(\cdot)$	The quantile check function, equal to $\rho_\tau(x) = x\{\tau - \mathbb{I}(x < 0)\}$ or equivalently $\rho_\tau(x) = \frac{1}{2}\lvert x\rvert + (\tau - \frac{1}{2})x$, for $x \in \mathbb{R}$ and $\tau \in (0, 1)$.

Probability distributions	
$\mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate Gaussian with mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and positive semi-definite variance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. Density function: $\pi(\boldsymbol{x}) = \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{x} \in \mathbb{R}^p$.
$\mathrm{N}(\mu, \sigma^2)$	Univariate Gaussian with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Density function: $\pi(x) = \phi(x; \mu, \sigma^2)$, $x \in \mathbb{R}$.
$\mathrm{t}(\mu, \sigma, \nu)$	Student t distribution with degrees of freedom $\nu > 0$, location $\mu \in \mathbb{R}$ and scale $\sigma > 0$. Density function: $\pi(x) = \{1 + (x-\mu)^2/\nu\sigma^2\}^{-(\nu+1)/2}/\sqrt{\nu}B(\frac{1}{2}, \frac{\nu}{2})$, $x \in \mathbb{R}$.
$\mathrm{AL}(\mu, \sigma, \tau)$	Asymmetric-Laplace with asymmetry parameter $\tau \in (0, 1)$, location $\mu \in \mathbb{R}$ and scale $\sigma > 0$. Density function: $\pi(x) = \tau(1-\tau)\exp\{-\rho_\tau(x-\mu)/\sigma\}/\sigma$, $x \in \mathbb{R}$.
$\mathrm{Exp}(\lambda)$	Exponential with rate $\lambda > 0$. Density function: $\pi(x) = \lambda\exp(-\lambda x)$, $x > 0$.
$\mathrm{IG}(A, B)$	Inverse-Gamma with shape $A > 0$ and rate $B > 0$.

	Density function: $\pi(x) = B^A x^{-A-1} \exp(-B/x)/\Gamma(A)$, $x > 0$.
$\mathrm{IN}(\mu, \lambda)$	Inverse-Gaussian distribution with location $\mu > 0$, and scale $\lambda > 0$. Density function: $\pi(x) = \exp\{-\lambda(x-\mu)^2/2\mu^2 x\}(\lambda/2\pi x^3)^{1/2}$, $x > 0$.
$\mathrm{GIG}(\nu, A, B)$	Generalized-Inverse-Gaussian distribution with parameters $\nu > 0$, $A > 0$ and $B > 0$. Density function: $\pi(x) = (A/B)^{\nu/2} x^{\nu-1} \exp\{-\frac{1}{2}(Ax + B/x)\}/2K_\nu(\sqrt{AB})$, $x > 0$.
$\mathrm{U}(A, B)$	Uniform distribution over $[A, B]$. Density function: $\pi(x) = 1/\lvert B - A \rvert$, $x \in [A, B]$.
$\mathrm{Be}(\pi)$	Bernoulli distribution with $\pi \in (0, 1)$. Mass function: $\pi(x) = \pi^x (1-\pi)^{1-x}$, $x \in \{0, 1\}$.
$\mathrm{Pois}(\lambda)$	Poisson distribution with rate $\lambda > 0$. Mass function: $\pi(x) = \lambda^x e^{-\lambda}/x!$, $x \in \mathbb{N}$.

Chapter-specific additional notation is described along with the Ph.D. thesis. Notation overlapping between different chapters has to be intended chapter-specific.

Introduction

Overview

In empirical studies, researchers are often interested in analyzing the behavior of a response variable, given the information of a set of covariates, by specifying a regression function for the conditional mean. However, it is generally recognized that the mean provides little or no information about the conditional distribution of the response when samples significantly deviate from standard assumptions, such as homoscedasticity and Gaussianity. In particular, heteroscedasticity, skewness, lepto- or platy-kurtic tails are characteristics often present in real data that can shadow the relationship between the response variable and the covariates, as postulated by the conditional mean. Imposing further restrictive parametric assumptions on the data generating mechanism could be a solution; this is the answer, for instance, of generalized linear models, which replace the Gaussian distribution with a more appropriate law belonging to the exponential family. Robust procedures based on minimum risk criteria instead aim at providing an adequate description of the characteristics of the conditional distribution without relying on, possibly misspecified, parametric assumptions.

Among robust minimum risk procedures, quantile regression plays a fundamental role in statistical theory and modeling. Its ability to flexibly describe non-trivial features of the conditional distribution, along with robustness to outlier contamination and invariance to monotone transformations, motivates the widespread popularity gained by conditional quantile methods over the years, since their introduction by Koenker and Bassett (1978). Moreover, the strict relationship between quantiles, estimating equations and minimum risk procedures entails a rich theoretical characterization of the associated estimator and also highlights the connections with other popular risk based models, such as expectiles (Newey and Powell, 1987; Efron, 1991), M-quantiles (Breckling and Chambers, 1988) and support vector machines (Vapnik, 1998). On the other hand, the non-regular properties of the quantile loss function, such as non-differentiability, piecewise flat curvature, and lack of conjugacy, pose severe obstacles to the estimation and extension of quantile models in more general settings, including penalized, additive and mixed models. Because of these features, regression quantiles constitute a remarkable prototype of a non-regular, robust model defined through a minimum risk criterion, which is worth investigating in order to improve existing inferential procedures and propose new methodologies.

In this thesis, we present some new developments on additive and mixed effect regression models where the no probabilistic assumptions are made upon the data generating mechanism of the response variable. We instead rely on risk based models where the misfit between the estimates and the data is measured through a loss function, with a particular focus on quantile regression. We then aim at providing some new inferential and computational tools for handling non-linear, heterogeneous effects of the available covariates on a response variable.

The organization of the thesis is the following. In Chapter 1, we introduce quantile regression and related robust regression models, with a particular focus on computational methods based on data-augmentation, from both frequentist and Bayesian points of view; the common notation adopted in the rest of the thesis is also introduced here. In Chapter 2, we discuss the first original contribution of this work, which consists of a new estimation algorithm for risk-based additive and mixed regression models. In Chapter 3, we discuss the second contribution of the thesis, which is a new robust nonparametric model for dealing with data collected over complex spatial domains with possibly non-isotropic and non-stationary behaviors.

Main contributions of the thesis

The thesis is made by two main chapters. We now discuss the contributions of each of them.

Non-conjugate regression via variational belief updating

After a brief review of the literature, Chapter 2 is devoted to propose a new variational approximation method (Ormerod and Wand, 2010) for additive and mixed models defined through a minimum risk criterion. We devote a particular attention to the Bayesian formulation of the considered models, relying on the so-called Bayesian belief updating literature (Bissiri *et al.*, 2016). However, we show that such an approach is agnostic to the inferential perspective we prefer and, thus, it can also be employed for estimation and inference in frequentist mixed models.

The approximation we propose belongs to the (semi)parametric variational Bayes paradigm and, in particular, builds upon the works of Knowles and Minka (2011), Tan and Nott (2013) and Wand (2014) on variational message passing for Gaussian variational approximations. The corresponding coordinate ascent fixed-point algorithm we develop only involves closed form algebraic operations and univariate numerical quadratures, when no analytic solutions are available. This leads to a scalable optimization routine for the estimation of a broad class of models, including generalized linear mixed models. For instance, we here consider the estimation of quantile and expectile regression, support vector machines for both regression and classification problems, binomial regression with logistic link, and Poisson regression with logaritmic link.

One of the main benefits of our approach is to allow for non-differentiable loss functions and non-conjugate priors, without requiring stochastic approximations or model-specific data-augmentation strategies (Dempster *et al.*, 1977). Indeed, our method directly approximates the posterior distribution of the parameters (or the marginal likelihood in the frequentist case) on the original parameter space. Under mild regularity conditions, the proposed approach is theoretically guaranteed to improve the posterior approximation of existing data-augmented mean field variational Bayes (Wand *et al.*, 2011; McLean and Wand, 2019) methods in the Kullback-Leibler metric.

Generalizations accounting for additive models, shrinkage priors, dynamic and spatial models are also discussed, providing a unifying framework for statistical learning that covers a wide range of applications.

The performances of our algorithm and approximation are then assessed through an extensive simulation study and a read data application, in which we compare our proposal with Markov chain Monte Carlo and conjugate mean field variational Bayes in terms of posterior approximation accuracy, signal reconstruction, and execution time.

Chapter 2 is organized as follows. In Section 2.2, we introduce a first motivating result and the general setting of parametric and semiparametric variational inference. In Section 2.3, we describe the class of models we consider, we introduce our approximation and its properties, and we deliver a pseudo-code formulation of our coordinate ascent algorithm. In Section 2.4, we provide some remarkable examples of models that can be handled within our approach; thus, for all of them, we provide the quantities needed for implementing the proposed algorithm. In Section 2.5, we discuss some possible extensions and different model specifications. In Section 2.6, we assess the quality of the approximation via an extensive simulation study. Finally, in Section 2.7, we present a real data problem concerned with the probabilistic load forecasting of the electric power consumption in US.

Spatial quantile regression with differential regularization

In Chapter 3, we propose a nonparametric quantile regression model for spatially referenced data, extending spatial regression with differential regularization by Sangalli *et al.* (2013) and Azzimonti *et al.* (2014). The proposed method allows us to incorporate external physical knowledge in the estimation of the conditional quantile surface, whenever this information can be formulated as an elliptic partial differential equation (PDE; Evans, 2010). Such a construction permits dealing with stationary and non-stationary anisotropic diffusion effects, unidirectional flows, and mixed boundary conditions. We can also handle complex planar domains characterized by strong concavities, holes, and physical barriers.

The novelty of our methodology is threefold. First, we introduce a broad class of physically-informed quantile regression models, based on a penalized loss criterion. In doing this, we trade off a goodness-of-fit measure and a roughness penalization depending on the PDE specification. Secondly, we propose an innovative functional expectation-maximization algorithm (Dempster *et al.*, 1977) in order to estimate unknown functional surfaces. The infinite-dimensional solution of such an optimization is then discretized by means of finite element methods (see, e.g., Quarteroni, 2017), and a model selection criterion based on such a discretized estimator is proposed. Finally, we provide a theoretical characterization of both the infinite- and finite-dimensional PDE quantile estimators, proving existence, consistency and asymptotic normality.

We then study the empirical performances of the proposed method by means of extensive simulation experiments, in which we compare our model with alternative state-of-the-art approaches in the literature. In doing so, we consider different scenarios in terms of domain shape, quantile field characteristics and distributional features, so that to provide a complete picture of the comparative advantages and limitations of our methodology.

Chapter 3 is organized as follows. In Section 3.2, we introduce the spatial quantile regression model with PDE regularization, along with the associated infinite-dimensional estimation problem. In Section 3.3, we propose an appropriate functional estimation algorithm, and we characterize its solution at each iteration. In Section 3.4, we introduce the finite element method to discretize the infinite-dimensional estimator. In Section 3.5, we study the large-sample properties of our estimators, proving consistency and asymptotic normality under different assumptions. In Section 3.6, we extend the pure nonparametric model to a semiparametric additive formulation, including the effect of space-varying covariates. In Section 3.7, we present two simulation studies in which we compare our method to alternative state-of-the-art approaches under different data scenarios. Finally, in Section 3.8, we employ our method to analyze a benchmark data set concerning rainfall measurements in Switzerland.

Chapter 1

Minimum risk estimation and quantile regression

1.1 Minimum risk estimators

Let us suppose to be interested in the unknown parameter $\boldsymbol{\theta} \in \Theta$, which describes some latent feature of the random variable $y \in \mathcal{Y}$, distributed according to the unknown probability law Π, denoted by $y \sim \Pi$. We further suppose that there exists a risk function $R : \Theta \to \mathbb{R}_+$ such that the *true value* of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_0$, can be expressed as the minimizer

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\ R(\boldsymbol{\theta}), \qquad \text{where} \qquad R(\boldsymbol{\theta}) = \mathbb{E}\{L(y,\boldsymbol{\theta})\} = \int_{\mathcal{Y}} L(y,\boldsymbol{\theta})\,\Pi(\mathrm{d}y), \tag{1.1}$$

with $L : \mathcal{Y} \times \Theta \to \mathbb{R}_+$ denoting a loss function that measures the misfit between y and $\boldsymbol{\theta}$. Thanks to this formulation, $\boldsymbol{\theta}_0$ is uniquely determined by the loss function L and the probability distribution Π.

In empirical studies we are almost never able to precisely determine Π, and thus $\boldsymbol{\theta}_0$ has to be estimated from a finite dimensional sample $y_1, \dots, y_n$ randomly generated from $y \sim \Pi$. The sample counterpart of problem (1.1) can thus be defined as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\ R_n(\boldsymbol{\theta}), \qquad \text{where} \qquad R_n(\boldsymbol{\theta}) = \mathbb{E}_n\{L(y,\boldsymbol{\theta})\} = \frac{1}{n}\sum_{i=1}^{n} L(y_i,\boldsymbol{\theta}), \tag{1.2}$$

where $R_n(\cdot)$ is the empirical risk and $\mathbb{E}_n(\cdot)$ is the sample expectation calculated with respect to the probability measure Π_n, which assigns probability mass $1/n$ to each point y_i. The solution of the sample risk minimization, $\hat{\boldsymbol{\theta}}$, is called M-estimator or minimum risk estimator of $\boldsymbol{\theta}_0$ (Boos and Stefanski, 2013). Within the class of M-estimators we find, for example, the maximum likelihood method (Severini, 2000) and the generalized method of moments (Hall, 2005).

Depending on the knowledge we have on the data generating mechanism, different loss functions L can be considered for estimating the unknown parameter $\boldsymbol{\theta}$, or a suitable reparametrization of it. Let us suppose, for instance, that the probability law $\Pi(\cdot)$

likely belongs to a known family of parametric distributions $\Pi(\cdot|\boldsymbol{\theta})$ indexed by $\boldsymbol{\theta} \in \Theta$ and having density function $\pi(y|\boldsymbol{\theta})$, that is there exists $\boldsymbol{\theta}_0 \in \Theta$ such that

$$\Pi(\cdot) = \Pi(\cdot\,|\boldsymbol{\theta}_0) \in \big\{\Pi(\cdot\,|\boldsymbol{\theta}) :\ \Pi(\mathrm{d}y|\boldsymbol{\theta}) = \pi(\mathrm{d}y|\boldsymbol{\theta})\,\mathrm{d}y,\ \boldsymbol{\theta} \in \Theta\big\}.$$

Then, the most natural candidate for $L(y, \boldsymbol{\theta})$ is the negative log-likelihood function $L(y, \boldsymbol{\theta}) = -\log \pi(y|\boldsymbol{\theta})$, which leads to the so-called maximum likelihood estimator (Severini, 2000). Notice that, here and elsewhere, we use $\pi(y|x)$ for indicating either the density function of y indexed by the parameter x, or the conditional density function of y given the random parameter x. This permits us to lighten the formulas and avoid confusing changes of notation when moving between frequentist and Bayesian formulations.

A different situation arises when it is not possible to specify a reasonable generative model for y and, therefore, a proper likelihood for $\boldsymbol{\theta}$; in these cases, alternative loss functions must be considered in order to obtain a robust, coherent inference on $\boldsymbol{\theta}$. For example, the mean of y, say $\boldsymbol{\theta}_0 = \mathbb{E}(y)$, can always be obtained by minimizing the risk associated to the squared error loss $L(y, \boldsymbol{\theta}) = (y - \boldsymbol{\theta})^2$; then the induced estimator is the solution of a least-squares problem and corresponds to the sample mean $\hat{\boldsymbol{\theta}} = \mathbb{E}_n(y)$. Similarly, the median of y, say $\boldsymbol{\theta}_0 = \inf\{x \in \mathbb{R} : \mathbb{P}(y < x) = 1/2\}$, is associated to the absolute error loss $L(y, \boldsymbol{\theta}) = |y - \boldsymbol{\theta}|$. More generally, the τ-th quantile of y, i.e., $\boldsymbol{\theta}_0 = \inf\{x \in \mathbb{R} :\ \mathbb{P}(y < x) \geq \tau\}$, corresponds to the asymmetrically weighted absolute error loss $L(y, \boldsymbol{\theta}) = \frac{1}{2}|y - \boldsymbol{\theta}| - (\tau - \frac{1}{2})(y - \boldsymbol{\theta})$, for $\tau \in (0, 1)$.

Aside from extremely rare situations, in most of the cases M-estimators do not enjoy closed form solutions and, thus, we must rely on iterative optimization methods to solve the empirical risk problem in (1.2). Under classical differentiability conditions on L, Newton-type algorithms provide an elegant and efficient answer to this problem by iterating until convergence the updating formula

$$\boldsymbol{\theta}^{(k+1)} \leftarrow \boldsymbol{\theta}^{(k)} - \big[\nabla^2_{\boldsymbol{\theta}} R_n(\boldsymbol{\theta}^{(k)})\big]^{-1} \big[\nabla_{\boldsymbol{\theta}} R_n(\boldsymbol{\theta}^{(k)})\big],$$

where $\nabla_{\boldsymbol{\theta}} R_n(\boldsymbol{\theta}) = \mathbb{E}_n\big[\nabla_{\boldsymbol{\theta}} L(y, \boldsymbol{\theta})\big]$ and $\nabla^2_{\boldsymbol{\theta}} R_n(\boldsymbol{\theta}) = \mathbb{E}_n\big[\nabla^2_{\boldsymbol{\theta}} L(y, \boldsymbol{\theta})\big]$ denote the gradient vector and Hessian matrix of R_n. If the loss function L is convex and coercive, the convergence of the sequence $\{\boldsymbol{\theta}^{(k)}\}$ to the global minimizer $\hat{\boldsymbol{\theta}}$ is guaranteed (Nocedal and Wright, 2006; Lange, 2010).

Moreover, under mild additional regularity assumptions (Stefanski and Boos, 2002), the asymptotic M-estimator $\hat{\boldsymbol{\theta}}$ is consistent and normally distributed:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathrm{d}} \mathrm{N}_p(\mathbf{0}_p, \mathbf{G}^{-1}),$$

where $\mathbf{G} = \mathbf{H}\,\mathbf{J}^{-1}\mathbf{H}$ is called Godambe information matrix, $\mathbf{H} = \mathbb{E}\big[\nabla^2_{\boldsymbol{\theta}} L(y, \boldsymbol{\theta}_0)\big]$ is the sensitivity matrix and $\mathbf{J} = \mathrm{Var}\big[\nabla_{\boldsymbol{\theta}} L(y, \boldsymbol{\theta}_0)\big]$ is the variability matrix. Approximations of the sample distribution of $\hat{\boldsymbol{\theta}}$ can thus be obtained by replacing $\mathbf{H}$ and $\mathbf{J}$ with their empirical counterparts, say $\mathbf{H}_n = \mathbb{E}_n\big[\nabla^2_{\boldsymbol{\theta}} L(y, \hat{\boldsymbol{\theta}})\big]$ and $\mathbf{J}_n = \mathrm{Var}_n\big[\nabla_{\boldsymbol{\theta}} L(y, \hat{\boldsymbol{\theta}})\big]$, which typically are made available as a side-product of Newton and quasi-Newton algorithms.

A prevalent application of the M-estimation theory is concerned with regression models, providing a flexible, robust alternative to likelihood based regression under model misspecification (Stefanski and Boos, 2002). Here, in particular, we consider linear regression models predicting the response variable $y_i \in \mathcal{Y} \subseteq \mathbb{R}$ via a linear regression function, also called linear predictor, $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, where $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of covariates and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown coefficients to be estimated. The empirical risk function then takes the form

$$R_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n L(y_i, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n L(y_i, \eta_i),$$

where the loss function L depends on the regression parameters $\boldsymbol{\beta}$ only through the linear predictor η_i. Whenever L is twice differentiable with respect to η_i, the gradient and Hessian of $R_n(\boldsymbol{\beta})$ are given by

$$\nabla_{\boldsymbol{\beta}} R_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \frac{\partial L}{\partial \eta_i}, \qquad \nabla^2_{\boldsymbol{\beta}} R_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \frac{\partial^2 L}{\partial \eta_i^2}.$$

Leveraging such a convenient gradient-Hessian structure, the Newton algorithm may be efficiently employed by solving a sequence of weigthed least-squares problems, giving rise to the so-called iteratively reweighted least-squares (McCullagh and Nelder, 1989) algorithm for non-linear regression problems.

Notice that, if minimal differentiability conditions on L are not met, we can not perform the estimation via Newton methods and, moreover, we need to rephrase the standard asymptotic theory, since the first and second order derivatives $\partial L / \partial \eta_i$ and $\partial^2 L / \partial \eta_i^2$ no longer exist. This is the case, for instance, of the quantile regression estimator, which is the main subject of the following sections and constitutes a benchmark model of primary importance throughout this thesis.

1.2 Quantile regression in a nutshell

Quantile regression is a statistical model defined by the minimization of an asymmetrically weighted absolute error loss and usually employed to explore the relationship between the quantiles of a response variable and a set of available covariates. Since quantiles provides a much richer description of a sample distribution than mean, quantile-based methods offer an appealing alternative to classical least-squares regression and, more generally, to mean-based regression. We refer the reader to Koenker (2005) and Koenker *et al.* (2018) for an exhaustive review of the literature.

In a quantile regression framework, the linear predictor $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ represents the τ-th quantile of the conditional distribution of $y_i \in \mathcal{Y} \subseteq \mathbb{R}$ given $\mathbf{x}_i \in \mathbb{R}^p$ for $\tau \in (0, 1)$, namely we assume η_i is such that $\mathbb{P}(y_i \leq \eta_i | \mathbf{x}_i) = \tau$ for any i. The observed response realizations y_i are generated by an unknown absolute continuous distribution and are conditionally independent given $\mathbf{x}_i$. Thus, the minimum risk estimator of $\boldsymbol{\beta} \in \mathbb{R}^p$ for a linear quantile regression model is given by

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} R_n(\boldsymbol{\beta}), \qquad R_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \eta_i) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \eta_i), \tag{1.3}$$

where

$$\rho_\tau(x) = x\{\tau - \mathbb{I}(x < 0)\} = \tfrac{1}{2}|x| + (\tau - \tfrac{1}{2})x \tag{1.4}$$

is the quantile loss function, also called check function, or pinball loss, and $\mathbb{I}(\cdot)$ denotes the usual indicator function. For any choice of τ, the estimator (1.3) is not available in closed form and, because of the non-differentiability of (1.4), it must be obtained via non-smooth optimization methods.

For these reasons, since its introduction by Koenker and Bassett (1978), quantile regression has always been strictly related to linear programming theory and non-smooth convex analysis (Koenker, 2005). Actually, the quantile estimator $\hat{\boldsymbol{\beta}}$ may be alternatively defined as the solution of the following linear programming problem

$$\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}, \hat{\mathbf{v}} = \operatorname*{argmin}_{\boldsymbol{\beta},\mathbf{u},\mathbf{v}} \left\{\tau \mathbf{1}_n^\top \mathbf{u} + (1-\tau)\mathbf{1}_n^\top \mathbf{v}\right\} \quad \text{subject to} \quad \begin{array}{l} \mathbf{u} + \mathbf{v} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}, \\ \mathbf{u} \geq \mathbf{0},\ \mathbf{v} \geq \mathbf{0}, \end{array} \tag{1.5}$$

where $\mathbf{y} = (y_1, \dots, y_n)^\top$ is the response vector, $\mathbf{X}$ is the design matrix obtained stacking by row all the vectors $\mathbf{x}_i^\top$, while $\mathbf{u}$ and $\mathbf{v}$ represent the positive and negative parts of the residual vector $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. We may say that, like the Euclidean geometry of the least-squares estimator, the polyhedral nature of quantile regression plays a crucial role in the theoretical characterization of such method; see, e.g., Portnoy and Koenker (1997). Indeed, thanks to the minimum risk representation (1.3) and the linear programming representation (1.5), the quantile estimator $\hat{\boldsymbol{\beta}}$ inherits the following properties (Koenker and Bassett, 1978):

1. if $\mathbf{X}$ is full rank, there exists at least one solution $\hat{\boldsymbol{\beta}}(h)$ such that $\hat{\boldsymbol{\beta}}(h) = \mathbf{X}_h^{-1}\mathbf{y}_h$, where $h \subset \{1, \dots, n\}$ a subset of indices of dimension p;
2. any solution $\hat{\boldsymbol{\beta}}$ belongs to the closed convex hull generated by all the solutions having the form $\hat{\boldsymbol{\beta}}(h) = \mathbf{X}_h^{-1}\mathbf{y}_h$;
3. any solution $\hat{\boldsymbol{\beta}}$ is a global minimizer of (1.3);
4. $\hat{\boldsymbol{\beta}}$ is equivariant with respect to location and scale transformations of the response;
5. $\hat{\boldsymbol{\beta}}$ is equivariant with respect to non-singular linear transformations of the design.

Notice that Properties 1 and 2 implicitly state the non-uniqueness of the sample quantile regression estimator $\hat{\boldsymbol{\beta}}$ which, though, exists finite thanks to the convexity and coercitivity of the check function (1.4).

Moreover, under standard regularity conditions (see, e.g., Koenker, 2005, Chapter 4), $\hat{\boldsymbol{\beta}}$ is consistent and enjoys an asymptotic Gaussian distribution:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathrm{d}} \mathrm{N}_p(\mathbf{0}_p, \mathbf{G}^{-1}),$$

where, similarly to standard M-estimators, the asymptotic variance-covariance matrix $\mathbf{G}^{-1}$ takes the sandwich form $\mathbf{G}^{-1} = \tau(1-\tau)\mathbf{D}_1^{-1}\mathbf{D}_0\mathbf{D}_1^{-1}$, with

$$\mathbf{D}_0 = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top, \qquad \mathbf{D}_1 = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\pi_i\mathbf{x}_i\mathbf{x}_i^\top.$$

Here, we denote by π_i the true unknown density function of y_i given $\mathbf{x}_i$ evaluated at the true conditional quantile $\mathbf{x}_i^\top\boldsymbol{\beta}_0$.

1.2.1 Asymmetric-Laplace pseudo-likelihood

An alternative formulation of the quantile problem in (1.3) has been proposed by Yu and Moyeed (2001), which showed that the check function $\rho_\tau(\cdot)$ in (1.4) is the negative log-kernel of an Asymmetric-Laplace distribution (Kotz *et al.*, 2001). Then, the regression quantiles in (1.3) can be rephrased as the maximum likelihood estimator relative to the misspecified model

$$y_i|\boldsymbol{\theta} \sim \mathrm{AL}(\eta_i, \sigma_\varepsilon^2, \tau), \qquad \eta_i = \mathbf{x}_i^\top\boldsymbol{\beta}, \qquad i = 1,\dots,n, \tag{1.6}$$

where $\eta_i \in \mathbb{R}$ is a location parameter, σ_ε^2 is a scale parameter, and $\tau \in (0,1)$ is a skewness parameter. The working probability density function of y_i given $\mathbf{x}_i$ implied by model (1.6) is then given by

$$\pi(y_i|\boldsymbol{\theta}) = \tau(1-\tau)\exp\big\{-\rho_\tau(y_i-\eta_i)/\sigma_\varepsilon^2\big\}/\sigma_\varepsilon^2,$$

where we denote by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_\varepsilon^2)$ the vector of unknown parameters in the model. Assuming conditional independence, multiplying all the individual terms and taking the logarithm, we obtain the log-likelihood function $\ell(\boldsymbol{\theta};\mathbf{y}) = \sum_{i=1}^n \log\pi(y_i|\boldsymbol{\theta})$, which is maximized at $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}_\varepsilon^2)$, where $\hat{\boldsymbol{\beta}}$ is defined as at (1.3) and $\hat{\sigma}_\varepsilon^2$ is given by

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}). \tag{1.7}$$

The introduction of the dispersion parameter σ_ε^2 does not affect the properties of the quantile estimator $\hat{\boldsymbol{\beta}}$, but may provide useful insights relative to the residual variability of $\varepsilon_i = y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}$ measured in the quantile loss scale.

As recognized by Kozumi and Kobayashi (2011) and proved by Kotz *et al.* (2001), any Asymmetric-Laplace density can be written as a location-scale convolution of a Guassian density with an Exponential kernel, that is

$$\frac{\tau(1-\tau)}{\sigma_\varepsilon^2}\exp\{-\rho_\tau(\varepsilon)/\sigma_\varepsilon^2\} = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2\omega}}\exp\left\{-\frac{(\varepsilon - a_1\omega)^2}{2a_2^2\sigma_\varepsilon^2\omega}\right\}\frac{e^{-\omega/\sigma_\varepsilon^2}}{\sigma_\varepsilon^2}\,\mathrm{d}\omega \tag{1.8}$$

where $a_1 = \frac{(1-2\tau)}{\tau(1-\tau)}$ and $a_2^2 = \frac{2}{\tau(1-\tau)}$ are non-stochastic constants completely determined by the value of τ. As a consequence, the working model in (1.6) is stochastically equivalent to the conditional Gaussian specification

$$y_i|\omega_i;\boldsymbol{\theta} \sim \mathrm{N}(\eta_i + a_1\omega_i, a_2^2\sigma_\varepsilon^2\omega_i), \qquad \omega_i|\boldsymbol{\theta} \sim \mathrm{Exp}(1/\sigma_\varepsilon^2), \qquad i = 1, \dots, n, \tag{1.9}$$

where $\mathrm{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, whereas $\mathrm{Exp}(\lambda)$ denotes the Exponential distribution with mean $1/\lambda > 0$. The completed log-likelihood formulation relative to the augmented model (1.9) for the i-th pair (y_i, ω_i) is then given by

$$\log \pi(y_i, \omega_i|\boldsymbol{\theta}) = -\frac{3}{2}\log\sigma_\varepsilon^2 - \frac{1}{2}\log\omega_i - \frac{\omega_i}{\sigma_\varepsilon^2} - \frac{(y_i - \eta_i - a_1\omega_i)^2}{2a_2^2\sigma_\varepsilon^2\omega_i}, \tag{1.10}$$

which, conditionally on ω_i, exhibits a familiar quadratic expression.

The pseudo-likelihood formulation of the quantile estimation problem in (1.6) together with the conditional Gaussian representation (1.9) provide the basic ingredient for the Bayesian formulation of quantile regression, which has been considered, among others, by Yu and Moyeed (2001) and Kozumi and Kobayashi (2011).

1.2.2 Bayesian quantile regression

Bayesian inference is concerned with the updating of a subjective prior belief about the parameter $\boldsymbol{\theta} \in \Theta$ to the posterior using the data information brought by the likelihood function. Such an updating is made possible by the Bayes theorem, which provides a practical rule to combine the prior $\pi(\boldsymbol{\theta})$ with the likelihood $\pi(\mathbf{y}|\boldsymbol{\theta})$, that is

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}, \boldsymbol{\theta})}{\pi(\mathbf{y})} = \frac{\pi(\boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta})}{\pi(\mathbf{y})}. \tag{1.11}$$

The numerator $\pi(\mathbf{y}, \boldsymbol{\theta}) = \pi(\boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta})$ is the model joint density function and the denominator

$$\pi(\mathbf{y}) = \int_\Theta \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}, \tag{1.12}$$

is the marginal likelihood, also called the evidence of the model in the machine learning literature. For most statistical models, the exact marginalization over Θ required by $\pi(\mathbf{y})$ is not available, as well as, an analytical expression for $\pi(\boldsymbol{\theta}|\mathbf{y})$. For this reason, the development of reliable and efficient integration techniques is one of the most important, and maybe most challenging, objective of computational Bayesian statisticians.

A Bayesian specification of the quantile regression model in (1.3) can be obtained by combining the Asymmetric-Laplace pseudo-likelihood in (1.6) with a prior distribution for $\boldsymbol{\beta}$ and σ_ε^2 reflecting our subjective beliefs about the parameters. A standard choice considered in literature is

$$\boldsymbol{\beta}|\sigma_\varepsilon^2 \sim \mathrm{N}_p(\mathbf{0}_p, \sigma_\varepsilon^2\sigma_\beta^2\mathbf{R}^{-1}), \qquad \sigma_\varepsilon^2 \sim \mathrm{IG}(A_\varepsilon, B_\varepsilon) \tag{1.13}$$

where $\mathrm{IG}(A, B)$ denotes the Inverse-Gamma distribution with shape $A > 0$ and rate $B > 0$, while $\sigma_\beta^2 > 0$ is a fixed parameter controlling the prior variance of $\boldsymbol{\beta}$, and $\mathbf{R}$ is a non-stochastic positive semi-definite matrix controlling the prior correlation structure of the $\boldsymbol{\beta}$ coefficients. Alternative prior specification can be considered for inducing robustness, shrinkage or sparsity effects on the posterior estimates.

Assuming likelihood (1.6) and prior (1.13), the unnormalized posterior distribution can be computed using the Bayes formula (1.11):

$$\begin{aligned}\log \pi(\mathbf{y}, \boldsymbol{\theta}) = & -\log \Gamma(A_\varepsilon) + A_\varepsilon \log B_\varepsilon - (A_\varepsilon - 1) \log \sigma_\varepsilon^2 - B_\varepsilon / \sigma_\varepsilon^2 \\ & - \tfrac{p}{2} \log(2\pi\sigma_\varepsilon^2\sigma_\beta^2) - \tfrac{1}{2} \operatorname{logdet}(\mathbf{R}) - \tfrac{1}{2} \boldsymbol{\beta}^\top \mathbf{R} \boldsymbol{\beta} / \sigma_\varepsilon^2 \sigma_\beta^2 \\ & + n \log \tau(1-\tau) - n \log \sigma_\varepsilon^2 - \mathbf{1}_n^\top \rho_\tau(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) / \sigma_\varepsilon^2 .\end{aligned} \tag{1.14}$$

From a frequentist point of view, the above log-posterior can also be interpreted as a penalized log-likelihood function subject to a Ridge-type regularization for $\boldsymbol{\beta}$ (Hastie *et al.*, 2009).

Differently from Bayesian linear model, in the quantile regression framework, prior and likelihood are not conjugate and, therefore, the normalizing constant $\pi(\mathbf{y})$ is not available in closed form. This fact, along with the non-differentiability of the log-likelihood, may complicate posterior analysis and computations. A possible solution has been suggested by Kozumi and Kobayashi (2011), which proposed to base the posterior inference on the conditional Gaussian representation of the Asymmetric-Laplace distribution (1.9). Doing so, the augmented posterior distribution factorizes as

$$\pi(\boldsymbol{\omega}, \boldsymbol{\theta} | \mathbf{y}) = \frac{\pi(\mathbf{y}, \boldsymbol{\omega}, \boldsymbol{\theta})}{\pi(\mathbf{y})} = \frac{\pi(\boldsymbol{\theta}) \pi(\boldsymbol{\omega} | \boldsymbol{\theta}) \pi(\mathbf{y} | \boldsymbol{\omega}, \boldsymbol{\theta})}{\pi(\mathbf{y})} \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^n \pi(\omega_i | \boldsymbol{\theta}) \pi(y_i | \omega_i, \boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ represents the vector of global parameters we are interested in, whereas $\boldsymbol{\omega}$ is a vector of auxiliary local parameters. The resulting unnormalized log-posterior can be written as

$$\begin{aligned}\log \pi(\mathbf{y}, \boldsymbol{\omega}, \boldsymbol{\theta}) = & -\log \Gamma(A_\varepsilon) + A_\varepsilon \log B_\varepsilon - (A_\varepsilon - 1) \log \sigma_\varepsilon^2 - B_\varepsilon / \sigma_\varepsilon^2 \\ & - \tfrac{p}{2} \log(2\pi\sigma_\varepsilon^2\sigma_\beta^2) - \tfrac{1}{2} \operatorname{logdet}(\mathbf{R}) - \tfrac{1}{2} \boldsymbol{\beta}^\top \mathbf{R} \boldsymbol{\beta} / \sigma_\varepsilon^2 \sigma_\beta^2 \\ & - \tfrac{n}{2} \log(2\pi a_2^2) - \tfrac{3}{2} n \log \sigma_\varepsilon^2 - \tfrac{1}{2} \mathbf{1}_n^\top \log \boldsymbol{\omega} - \mathbf{1}_n^\top \boldsymbol{\omega} / \sigma_\varepsilon^2 \\ & - \tfrac{1}{2} \, \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - a_1 \boldsymbol{\omega} \|_{\mathbf{W}}^2 / \sigma_\varepsilon^2 ,\end{aligned} \tag{1.15}$$

where $\mathbf{W} = \operatorname{diag}[1/\boldsymbol{\omega}]/a_2^2$ is a diagonal weighting matrix and $\|\boldsymbol{x}\|_{\mathbf{A}}^2 = \boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}$ is the squared norm of $\boldsymbol{x}$ induced by the symmetric positive semi-definite matrix $\mathbf{A}$. This way, local conjugacy between likelihood and priors is restored and all the full-conditional posterior densities of the parameters belong to standard families of probability distributions (see, e.g., Kozumi and Kobayashi, 2011). A straightforward benefit of such a representation is thus to facilitate the derivation of model-specific algorithms for point estimation, posterior sampling and posterior approximation. Some remarkable examples are presented and discussed in the next section.

Asymptotic properties of Bayesian quantile regression based on the misspecified Asymmetric-Laplace likelihood have been studied by Sriram *et al.* (2013), which prove its frequentist consistency under mild regularity conditions.

1.3 Estimation methods

Historically, the first and most popular approach for quantile regression estimation was to employ standard linear programming tools, such as simplex algorithms (Koenker and Bassett, 1978) and primal-dual interior-point methods (Koenker and Ng, 2005; Portnoy and Koenker, 1997), to solve the polyhedral problem (1.5). Such an approach has been successfully employed for handling several generalizations of the basic quantile regression model. Some examples are sparse and constrained quantile regression (Koenker and Ng, 2005), high-dimensional quantile regression (Wang *et al.*, 2012), quantile smoothing spline with ℓ^1 and ℓ^2 penalties (Koenker *et al.*, 1994; Ng, 1996; Bosch *et al.*, 1995), and spatial quantile smoothing with total variation regularization (Koenker and Mizera, 2004).

Despite its efficiency, stability and broad applicability, linear programming does not provide any insights about variability and uncertainty of the estimates; moreover, it involves a possibly high-dimensional constrained optimization which is not always easy to generalize for other model specifications. For instance, these drawbacks pose severe limits when it comes to estimate quantile mixed models for heterogeneous dependent data (Geraci and Bottai, 2014; Geraci, 2014), or Bayesian quantile regression models (Yu and Moyeed, 2001; Kozumi and Kobayashi, 2011).

As an alternative to linear programming, many authors proposed to solve the quantile regression problem by using a convergent iteratively reweighted least squares algorithms based on local quadratic approximations of the quantile loss function. Such an approach has been explored by, e.g., Hunter and Lange (2000), Yue and Rue (2011) and Fasiolo *et al.* (2021a). A different, but rather similar, possibility is to consider estimation methods based on the Asymmetric-Laplace pseudo-likelihood formulation in (1.6) and its data-augmented representation (1.8). In the rest of this section, we review frequentist and Bayesian estimating methods based this approach, which constitute a basic ingredient for understanding the original contributions proposed in Chapter 2 and 3.

1.3.1 Expectation-maximization

The expectation-maximization algorithm (EM; Dempster *et al.*, 1977; McCullagh and Nelder, 1989) is a popular approach to the iterative maximization of complex log-likelihood functions, for which other standard approaches, such as Newton algorithm, can not be applied or may encounter nontrivial obstacles. For instance, in classical quantile regression, the non-differentiability of the objective function (1.3) does not allow for a straightforward application of Newton and quasi-Newton algorithms, and we need model-specific method to perform the optimization (see, e.g., the interior point method by Koenker and Ng, 2005).

The intuition behind the EM paradigm is closely related to the concepts of missing information and completed data. We assume that the likelihood maximization is made difficult by the fact that we observe only an incomplete realization from a joint model, that, if completely observed, would exhibit a joint likelihood with a simple functional form. The EM approach thus tries to exploit the relationship between the marginal and completed likelihoods in order to iteratively search the maximum likelihood estimator.

More formally, let us denote by $\pi(\mathbf{y}, \boldsymbol{\omega}|\boldsymbol{\theta})$ the joint density function for the completed data $\{\mathbf{y}, \boldsymbol{\omega}\}$, where $\mathbf{y}$ is the observed data vector and $\boldsymbol{\omega}$ is the missing data vector; let $\pi(\mathbf{y}|\boldsymbol{\theta})$ be the marginal distribution of $\mathbf{y}$ obtained by integrating out $\boldsymbol{\omega}$ from the above joint density, that is

$$\pi(\mathbf{y}|\boldsymbol{\theta}) = \int_\Omega \pi(\mathbf{y}, \boldsymbol{\omega}|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\omega} = \int_\Omega \pi(\mathbf{y}|\boldsymbol{\omega}; \boldsymbol{\theta})\, \pi(\boldsymbol{\omega}|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\omega}.$$

Then, thanks to the Jensen inequality, if $\pi(\mathbf{y}, \boldsymbol{\omega}; \boldsymbol{\theta})$ is log-concave in $\boldsymbol{\omega}$, we obtain

$$\underline{\ell}(\boldsymbol{\theta}; \mathbf{y}) = \mathbb{E}\{\log \pi(\mathbf{y}, \boldsymbol{\omega}|\boldsymbol{\theta})\} \leq \log \mathbb{E}\{\pi(\mathbf{y}, \boldsymbol{\omega}|\boldsymbol{\theta})\} = \ell(\boldsymbol{\theta}; \mathbf{y}), \tag{1.16}$$

where the expectation is calculated with respect to the conditional distribution of $\boldsymbol{\omega}$ given $\mathbf{y}$ and $\underline{\ell}(\boldsymbol{\theta}; \mathbf{y})$ is a function of $\boldsymbol{\theta}$ and $\mathbf{y}$ bounding the log-likelihood $\ell(\boldsymbol{\theta}; \mathbf{y})$ from below. EM thus prescribes to iteratively maximize $\underline{\ell}(\boldsymbol{\theta}; \mathbf{y})$ in order to implicitly optimize $\ell(\boldsymbol{\theta}; \mathbf{y})$. Given a current estimate of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^{(k)}$, each iteration of an EM algorithm updates the actual guess $\boldsymbol{\theta}^{(k)}$ to the new value $\boldsymbol{\theta}^{(k+1)}$ by executing an expectation (E) step and a maximization (M) step:

$$\begin{aligned} &\text{E-step} \qquad \underline{\ell}^{(k)}(\boldsymbol{\theta}; \mathbf{y}) \leftarrow \mathbb{E}^{(k)}\{\log \pi(\mathbf{y}, \boldsymbol{\omega}|\boldsymbol{\theta})\}, && (1.17)\\ &\text{M-step} \qquad \boldsymbol{\theta}^{(k+1)} \leftarrow \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\ \underline{\ell}^{(k)}(\boldsymbol{\theta}; \mathbf{y}), && (1.18)\end{aligned}$$

where the expectation $\mathbb{E}^{(k)}(\cdot)$ is taken with respect to the conditional distribution $\pi(\boldsymbol{\omega}|\mathbf{y}; \boldsymbol{\theta}^{(k)})$ obtained at the previous iteration of the algorithm.

As proved by Dempster *et al.* (1977) and further discussed by McLachlan and Krishnan (2008) and Lange (2010), each iteration of (1.17) and (1.18) produces a non-decreasing increment of the likelihood, that is

$$\ell(\boldsymbol{\theta}^{(k)}; \mathbf{y}) \leq \ell(\boldsymbol{\theta}^{(k+1)}; \mathbf{y}). \tag{1.19}$$

Moreover, each sequence of EM estimators $\{\boldsymbol{\theta}^{(k)}\}$ almost surely converges in the limit to a local maximizer of the likelihood. If $\ell(\boldsymbol{\theta}; \mathbf{y})$ is strongly convex in $\boldsymbol{\theta}$, and therefore its optimum is unique, the global fixed point of the algorithm corresponds to the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$.

In order to apply the EM principle to the Asymmetric-Laplace representation (1.6) of quantile regression, we first need to obtain the conditional distribution of $\boldsymbol{\omega}$ given $\mathbf{y}$ in the augmented model (1.9). As proved by, e.g., Kozumi and Kobayashi (2011) and Tian *et al.* (2014), conditionally on y_i, each latent variable ω_i is independent on ω_j,

$j \neq i$, moreover, its conditional distribution is

$$\omega_i | y_i; \boldsymbol{\theta} \sim \mathrm{GIG}\left(\frac{1}{2}, \frac{a_1^2 + 2a_2^2}{a_2^2 \sigma_\varepsilon^2}, \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{a_2^2 \sigma_\varepsilon^2} \right), \tag{1.20}$$

or, equivalently,

$$\omega_i^{-1} | y_i; \boldsymbol{\theta} \sim \mathrm{IN}\left(\frac{(a_1^2 + 2a_2^2)^{1/2}}{|y_i - \mathbf{x}_i^\top \boldsymbol{\beta}|}, \frac{a_1^2 + 2a_2^2}{a_2^2 \sigma_\varepsilon^2} \right),$$

where $\mathrm{GIG}(\nu, A, B)$ denotes the Generalize-Inverse-Gaussian distribution, with parameters $\nu, A, B > 0$, whereas $\mathrm{IN}(\mu, \lambda)$ denotes the Inverse-Gaussian distribution with mean parameter $\mu > 0$ and scale parameter $\lambda > 0$ (see, e.g., Jørgensen, 1982). Adopting the notation $\mu_{\omega_i}^{(k)} = \mathbb{E}^{(k)}(\omega_i)$ and $\mu_{1/\omega_i}^{(k)} = \mathbb{E}^{(k)}(1/\omega_i)$, defining $\varepsilon_i = y_i - \mathbf{x}_i^\top \boldsymbol{\beta}$, and calculating the expectation of $\log \pi(y_i, \omega_i | \boldsymbol{\theta})$ with respect to $\pi(\omega_i | y_i; \boldsymbol{\theta}^{(k)})$, we obtain the lower bound

$$\underline{\ell}^{(k)}(\boldsymbol{\theta}; \mathbf{y}) = -\frac{3n}{2} \log \sigma_\varepsilon^2 - \frac{1}{2a_2^2 \sigma_\varepsilon^2} \sum_{i=1}^n \left\{ \varepsilon_i^2 \, \mu_{1/\omega_i}^{(k)} - 2a_1 \, \varepsilon_i + (a_1^2 + 2a_2^2) \mu_{\omega_i}^{(k)} \right\} + \text{const},$$

which can be alternatively expressed as

$$\underline{\ell}^{(k)}(\boldsymbol{\theta}; \mathbf{y}) = -\frac{3n}{2} \log \sigma_\varepsilon^2 - \frac{a_1^2 + 2a_2^2}{2a_2^2 \sigma_\varepsilon^2} \mathbf{1}_n^\top \boldsymbol{\mu}_\omega^{(k)} - \frac{1}{2\sigma_\varepsilon^2} (\mathbf{z}^{(k)} - \boldsymbol{\eta})^\top \mathbf{W}^{(k)} (\mathbf{z}^{(k)} - \boldsymbol{\eta}) + \text{const},$$

where $\mathbf{W}^{(k)} = \mathrm{diag}\big[\boldsymbol{\mu}_{1/\omega}^{(k)}\big] / a_2^2$ is a diagonal weighting matrix, $\mathbf{z}^{(k)} = \mathbf{y} - a_1 / \boldsymbol{\mu}_{1/\omega}^{(k)}$ is a pseudo-data vector, and "const" is a constant term not depending on $\boldsymbol{\theta}$ and σ_ε^2. As shown by Tian *et al.* (2014), the expectations $\mu_{\omega_i}^{(k)}$ and $\mu_{1/\omega_i}^{(k)}$ can be calculated in closed form as

$$\mu_{\omega_i}^{(k)} = \{\mu_{1/\omega_i}^{(k)}\}^{-1} + \frac{a_2^2 \sigma_\varepsilon^{2(k)}}{a_1^2 + 2a_2^2}, \qquad \mu_{1/\omega_i}^{(k)} = \frac{(a_1^2 + 2a_2^2)^{1/2}}{|y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(k)}|}.$$

This complete the specification of $\underline{\ell}^{(k)}(\boldsymbol{\theta}; \mathbf{y})$ and, thus, the E-step.

Assuming that all the diagonal elements of $\mathbf{W}^{(k)}$ are strictly positive and bounded away from ∞, the M-step of the algorithm is analytically available and may be obtained via the weighted least squares update

$$\boldsymbol{\beta}^{(k+1)} = \big(\mathbf{X}^\top \mathbf{W}^{(k)} \mathbf{X}\big)^{-1} \mathbf{X}^\top \mathbf{W}^{(k)} \mathbf{z}^{(k)}. \tag{1.21}$$

Similarly, the updated value of σ_ε^2 can be obtained by maximizing $\underline{\ell}^{(k)}(\boldsymbol{\theta}; \mathbf{y})$, or, alternatively, by using the maximum likelihood estimator (1.7). This second solution is typically to be preferred, since it leads to a faster convergence, preserving at the same time the non-decreasing property of the EM sequence.

Algorithm 1 provides a pseudo-code description of the fundamental steps of the quantile EM procedure outlined so far. Doing this, we consider a Bayesian, or penalized, formulation of the quantile regression problem, with a Gaussian prior distribution $\boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{0}_p, \sigma_\varepsilon^2 \sigma_\beta^2 \mathbf{R}^{-1})$. At convergence, the algorithm provides a maximum *a posteriori*, or

Algorithm 1 EM algorithm for quantile regression

```
Require: $\tau, \mathbf{y}, \mathbf{X}$
  Initialize $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_\varepsilon^2$;
  while convergence is not reached do
    $\hat{\boldsymbol{\mu}}_{1/\omega} \leftarrow (a_1^2 + 2a_2^2)^{1/2} / |\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}|$;
    $\hat{\mathbf{W}} \leftarrow \mathrm{diag}\big[\hat{\boldsymbol{\mu}}_{1/\omega}\big]/a_2^2$;   $\hat{\mathbf{z}} \leftarrow \mathbf{y} - a_1 \hat{\boldsymbol{\mu}}_{1/\omega}^{-1}$;
    $\hat{\boldsymbol{\beta}} \leftarrow (\mathbf{X}^\top \hat{\mathbf{W}}\, \mathbf{X} + \mathbf{R}/\sigma_\beta^2)^{-1} \mathbf{X}^\top \hat{\mathbf{W}}\, \hat{\mathbf{z}}$;
    $\hat{\sigma}_\varepsilon^2 \leftarrow \big\{\frac{1}{2}\hat{\boldsymbol{\beta}}^\top \mathbf{R}\, \hat{\boldsymbol{\beta}}/\sigma_\beta^2 + \mathbf{1}_n^\top \rho_\tau(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\big\}/(\frac{p}{2} + n)$;
  end while
```

penalized maximum likelihood estimate of $\boldsymbol{\beta}$. The frequentist unpenalized version of the algorithm can be recovered by setting $\sigma_\beta^2 \to \infty$. The convergence may be assessed by monitoring the relative change of the parameters and the Asymmetric-Laplace pseudo-likelihood.

1.3.2 Gibbs sampling

Despite its simple, elegant formulation and its theoretical coherence, Bayesian inference has been almost unexplored for practical applications since the recent development and diffusion of high-performance computers. The intrinsic complexity of calculating multivariate integrals over high-dimensional spaces made Bayesian calculations unpractical even for small data problems, with a moderate number of unknown parameters. A revolutionary turning point for Bayesian computation was the introduction of Markov chain Monte Carlo (MCMC) algorithms for posterior simulation. In its essence, MCMC is a family of procedures to sample a sequence of dependent realizations $\{\boldsymbol{\theta}^{(k)}\}$ from a stationary Markov chain, whose ergodic distribution converges in the limit to the true posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$. Then, after a certain number of iterations needed for the chain to reach the convergence, also called *burnin* period, the samples $\{\boldsymbol{\theta}^{(k)}\}$ can be used to evaluate posterior integrals via Monte Carlo approximation:

$$\mathbb{E}\{m(\boldsymbol{\theta})|\mathbf{y}\} = \int_\Theta m(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\,\mathrm{d}\boldsymbol{\theta} \approx \frac{1}{K}\sum_{k=1}^{K} m(\boldsymbol{\theta}^{(k)}).$$

Thanks to the ergodic theorem (Norris, 1998), the Monte Carlo estimator converges in probability to the true expectation with an error that may be controlled by increasing the number of simulated values.

Among other methods, the Gibbs sampling algorithm (Casella and George, 1992) plays a crucial role in the MCMC literature and is still one of the most effective method for posterior simulation when hierarchical Bayesian models with conjugate prior distributions are of interest. Its application is particularly convenient for situations in which the parameter vector $\boldsymbol{\theta}$ may be partitioned in sub-blocks, say $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_H\}$, so that all the associated full-conditional densities belong to notable probability families for which efficient sampling routines exist. Hereafter, we denote with $\pi(\boldsymbol{\theta}_h|\text{rest}) = \pi(\boldsymbol{\theta}_h|\boldsymbol{\theta}_{-h}, \mathbf{y})$ the full-conditional density of $\boldsymbol{\theta}_h$, where "rest" denotes the conditioning set, containing the data $\mathbf{y}$ and all the parameters except $\boldsymbol{\theta}_h$, say $\boldsymbol{\theta}_{-h} = \boldsymbol{\theta} \setminus \boldsymbol{\theta}_h$. After the (k)-th iteration of the Gibbs sampling algorithm, a new value of $\boldsymbol{\theta}$ is drawn by cycling over the following conditional iteration

$$\boldsymbol{\theta}_h^{(k+1)} \sim \pi^{(k)}(\boldsymbol{\theta}_h|\text{rest}), \qquad h = 1, \dots, H.$$

In the case where the model is expanded by an additional set of working variables $\boldsymbol{\omega}$, the Gibbs sampling scheme can be still applied by employing an additional conditional step at each iteration, that is

$$\begin{aligned} \boldsymbol{\theta}_h^{(k+1)} &\sim \pi^{(k)}(\boldsymbol{\theta}_h|\text{rest}), \qquad && h = 1, \dots, H. \\ \omega_i^{(k+1)} &\sim \pi^{(k)}(\omega_i|\text{rest}), \qquad && i = 1, \dots, n. \end{aligned}$$

As for any MCMC algorithm, the first values of the chain must be discarded as burnin period.

Similarly to the EM algorithm, Gibbs sampling can be conveniently employed for Bayesian quantile regression problems leveraging the Asymmetric-Lalace likelihood (1.6) and its augmented representation (1.9). For the sake of simplicity, we here consider only Bayesian regression models with prior distributions of the form (1.13). Following Kozumi and Kobayashi (2011), the full conditional distribution for the parameter $\boldsymbol{\beta}$ is given by

$$\begin{aligned} \boldsymbol{\beta} \mid \text{rest}^{(k)} &\sim \mathrm{N}_p(\boldsymbol{\mu}_\beta^{(k)}, \boldsymbol{\Sigma}_\beta^{(k)}), \\ \boldsymbol{\mu}_\beta^{(k)} &= \boldsymbol{\Sigma}_\beta^{(k)} \mathbf{X}^\top \mathbf{W}^{(k)} \mathbf{z}^{(k)} / \sigma_\varepsilon^{2(k)}, \\ \boldsymbol{\Sigma}_\beta^{(k)} &= \sigma_\varepsilon^{2(k)} \big(\mathbf{X}^\top \mathbf{W}^{(k)} \mathbf{X} + \mathbf{R}/\sigma_\beta^2 \big)^{-1}, \end{aligned} \tag{1.22}$$

where $\mathbf{W}^{(k)} = \mathrm{diag}\big[1/\boldsymbol{\omega}^{(k)}\big]/a_2^2$ and $\mathbf{z}^{(k)} = \mathbf{y} - a_1 \boldsymbol{\omega}^{(k)}$. The full-conditional distribution for the parameter σ_ε^2 is given by

$$\begin{aligned} \sigma_\varepsilon^2 \mid \text{rest}^{(k)} &\sim \mathrm{IG}(A_\varepsilon^{(k)}, B_\varepsilon^{(k)}), \\ A_\varepsilon^{(k)} &= A_\varepsilon + \tfrac{1}{2} p + \tfrac{3}{2} n, \\ B_\varepsilon^{(k)} &= B_\varepsilon + \tfrac{1}{2} \, \big\| \boldsymbol{\beta}^{(k)} \big\|_{\mathbf{R}}^2 / \sigma_\beta^2 + \mathbf{1}_n^\top \boldsymbol{\omega}^{(k)} + \tfrac{1}{2} \big\| \mathbf{z}^{(k)} - \mathbf{X}\boldsymbol{\beta}^{(k)} \big\|_{\mathbf{W}^{(k)}}^2, \end{aligned} \tag{1.23}$$

where $\|\boldsymbol{x}\|_{\mathbf{A}}^2 = \boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}$ is the squared norm of $\boldsymbol{x}$ induced by the symmetric positive semi-definite matrix $\mathbf{A}$. Whereas, the full-conditional distribution for the auxiliary variables $\boldsymbol{\omega}$ takes the form of a factorized Generalized-Inverse-Gaussian law, as provided in (1.20).

Hence, in order to sample from the augmented posterior $\pi(\boldsymbol{\omega}, \boldsymbol{\beta}, \sigma_\varepsilon^2|\mathbf{y})$, we can employ the Gibbs sampling scheme summarized in Algorithm 2, which cycles over the conditional distribution (1.20), (1.22) and (1.23).

Algorithm 2 Gibbs sampling for quantile regression

```
Require: $\tau, \mathbf{y}, \mathbf{X}$
  Initialize $\boldsymbol{\beta}$ and $\sigma_\varepsilon^2$;
  while convergence is not reached do
    $\hat{A}_\omega \leftarrow (a_1^2 + 2a_2^2)/a_2^2\sigma_\varepsilon^2$;   $\hat{B}_\omega \leftarrow \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/a_2^2\sigma_\varepsilon^2$;   $\boldsymbol{\omega} \sim \mathrm{GIG}(\frac{1}{2}, \hat{A}_\omega, \hat{B}_\omega)$;
    $\hat{\mathbf{W}} \leftarrow \mathrm{diag}\big[1/\boldsymbol{\omega}\big]/a_2^2$;   $\hat{\mathbf{z}} \leftarrow \mathbf{y} - a_1\boldsymbol{\omega}$;
    $\hat{\boldsymbol{\Sigma}}_\beta \leftarrow \sigma_\varepsilon^2(\mathbf{X}^\top\hat{\mathbf{W}}\mathbf{X} + \mathbf{R}/\sigma_\beta^2)^{-1}$;   $\hat{\boldsymbol{\mu}}_\beta \leftarrow \hat{\boldsymbol{\Sigma}}_\beta\mathbf{X}^\top\hat{\mathbf{W}}\,\hat{\mathbf{z}}/\sigma_\varepsilon^2$;   $\boldsymbol{\beta} \sim \mathrm{N}_p(\hat{\boldsymbol{\mu}}_\beta, \hat{\boldsymbol{\Sigma}}_\beta)$;
    $\hat{B}_\varepsilon \leftarrow B_\varepsilon + \frac{1}{2}\boldsymbol{\beta}^\top\mathbf{R}\,\boldsymbol{\beta}/\sigma_\beta^2 + \mathbf{1}_n^\top\boldsymbol{\omega} + \frac{1}{2}\mathbf{1}_n^\top(\mathbf{y} - a_1\boldsymbol{\omega} - \mathbf{X}\boldsymbol{\beta})^2/a_2^2\boldsymbol{\omega}$
    $\hat{A}_\varepsilon \leftarrow A_\varepsilon + \frac{p}{2} + \frac{3}{2}n$;   $\sigma_\varepsilon^2 \sim \mathrm{IG}(\hat{A}_\varepsilon, \hat{B}_\varepsilon)$
  end while
```

1.3.3 Mean field variational Bayes

Variational methods are a family of deterministic techniques for making approximate inference in complex statistical models depending on a set of stochastic latent parameters. Even though, in principle, the use of variational methods is not exclusive of Bayesian statistics, they are mostly employed for posterior approximation of Bayesian hierarchical models, as an efficient alternative to expensive simulation-based methods, such as MCMC. Thereby, for the sake of exposition, we here focus our treatment on variational approximations in a Bayesian context.

Following the so-called density transformation approach (Ormerod and Wand, 2010), variational approximate inference on the true posterior is made by replacing $\pi(\boldsymbol{\theta}|\mathbf{y})$ with a convenient density function $q(\boldsymbol{\theta})$ belonging to a tractable functional space $\mathcal{Q}$. The optimal approximating distribution $q^*(\boldsymbol{\theta})$ is then selected by minimizing some measure of divergence between $q(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|\mathbf{y})$, i.e.

$$q^*(\boldsymbol{\theta}) = \underset{q \in \mathcal{Q}}{\mathrm{argmin}}\; D\{q(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\}. \tag{1.24}$$

Different choices of divergence measure D and functional space $\mathcal{Q}$ give rise to alternative variational approximations. Among others, its worth mentioning mean field variational Bayes (Ormerod and Wand, 2010; Blei *et al.*, 2017), Gaussian variational approximation (Ormerod and Wand, 2012) and expectation-propagation (Minka, 2005).

Specifically, we here consider the family of variational Bayes techniques, which is by far the most common form of variational approximation in literature (see Bishop, 2006; Ormerod and Wand, 2010; Blei *et al.*, 2017). Within the variational Bayes paradigm, the optimal variational density $q^*(\boldsymbol{\theta})$ is defined as the minimizer of the Kullback-Leibler (KL) divergence

$$\mathrm{KL}\{q(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\} = -\int_\Theta q(\boldsymbol{\theta}) \log\left\{\frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta})}\right\} \mathrm{d}\boldsymbol{\theta}, \tag{1.25}$$

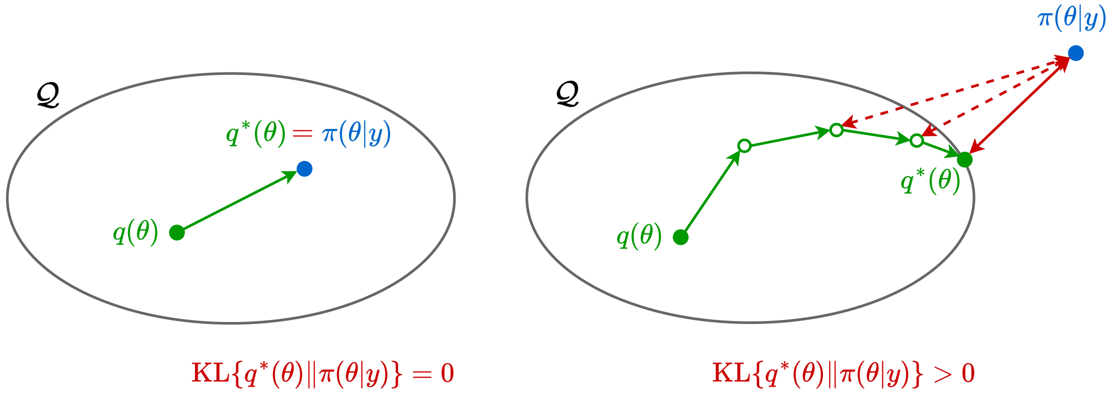
$\mathcal{Q}$
$q^*(\theta) = \pi(\theta|y)$
$q(\theta)$
$\mathrm{KL}\{q^*(\theta)\|\pi(\theta|y)\} = 0$
$\pi(\theta|y)$
$\mathcal{Q}$
$q^*(\theta)$
$q(\theta)$
$\mathrm{KL}\{q^*(\theta)\|\pi(\theta|y)\} > 0$

FIGURE 1.1: Graphical representation of the variational Bayes search of the optimal density $q^*(\boldsymbol{\theta})$ within the functional space $\mathcal{Q}$ under two scenarios: (left) $\pi(\boldsymbol{\theta}|\mathbf{y}) \in \mathcal{Q}$, (right) $\pi(\boldsymbol{\theta}|\mathbf{y}) \notin \mathcal{Q}$.

or, equivalently, as the maximizer of the lower bound on the marginal log-likelihood, also called evidence lower bound (ELBO),

$$\underline{\ell}\{\mathbf{y}; q(\boldsymbol{\theta})\} = \int_\Theta q(\boldsymbol{\theta}) \log \left\{ \frac{\pi(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\theta}. \tag{1.26}$$

For any choice of $q(\boldsymbol{\theta})$, the gap between the Kullback-Leibler divergence and the evidence lower bound is constant and corresponds to the marginal log-likelihood, or evidence of the model, say

$$\log \pi(\mathbf{y}) = \ell(\mathbf{y}) = \underline{\ell}\{\mathbf{y}; q(\boldsymbol{\theta})\} + \mathrm{KL}\{q(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\}. \tag{1.27}$$

Hence, recalling that $\mathrm{KL}\{q(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\} \geq 0$ for any density $q(\boldsymbol{\theta})$, equating 0 almost surely if and only if $q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})$ (Kullback and Leibler, 1951), we have

$$\underline{\ell}\{\mathbf{y}; q(\boldsymbol{\theta})\} \leq \ell(\mathbf{y}),$$

which motivates the name *lower bound* on the marginal log-likelihood for the quantity (1.26).

The second ingredient of variational inference is the choice of the functional space $\mathcal{Q}$. Notice that, if $\pi(\boldsymbol{\theta}|\mathbf{y}) \in \mathcal{Q}$, the unique minimizer of the Kullback-Leibler divergence will correspond to the target posterior $q^*(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})$, which guarantees the coherence of the variational Bayes approach under the exact identification of the posterior space. If, otherwise, $\pi(\boldsymbol{\theta}|\mathbf{y}) \notin \mathcal{Q}$, as it is almost always the case in practical situations, the optimal variational Bayes posterior $q^*(\boldsymbol{\theta})$ corresponds to the projection of $\pi(\boldsymbol{\theta}|\mathbf{y})$ onto the functional space $\mathcal{Q}$ measured in the Kullback-Leibler metric. Figure 1.1 provide a graphical representation of these two scenarios.

Hence, the choice of $\mathcal{Q}$ must be guided by a trade-off between accuracy and complexity of the solution, that is: the more general and precise the approximation, the less tractable the computations. In this, time constraints, computational resources and memory space must be taken into account when specifying $\mathcal{Q}$ and designing a corresponding variational optimization algorithm. Tractability of variational methods may be achieved by imposing a suitable factorization of the approximated posterior over a pre-specified partition of the parameter vector, so that to disentangle the complex dependence structures induced by the posterior distribution. Such an approach gives rise the mean field variational Bayes (MFVB) method, which is characterized by the product restriction

$$\mathcal{Q} = \left\{ q(\boldsymbol{\theta}) : q(\boldsymbol{\theta}) = \prod_{h=1}^{H} q_h(\boldsymbol{\theta}_h) \right\}, \tag{1.28}$$

for $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_H\}$ being a partition of $\boldsymbol{\theta}$. Very fine partitions impose restrictive posterior independence; on the other hand, too conservative partitions would not afford the tractability of the optimization.

Starting from a posterior guess $q^{(k)}(\boldsymbol{\theta})$, exploiting factorization (1.28) and maximizing the evidence lower bound with respect to $q_h(\boldsymbol{\theta}_h)$, the optimal coordinate-wise approximation for the h-th sub-vector $\boldsymbol{\theta}_h$ takes the nonparametric form

$$q_h^{(k+1)}(\boldsymbol{\theta}_h) \propto \exp\left[\mathbb{E}_{-h}^{(k)}\{\log \pi(\boldsymbol{\theta}_h|\text{rest})\}\right], \qquad h = 1, \dots, H, \tag{1.29}$$

where $\pi(\boldsymbol{\theta}_h|\text{rest})$ is the full-conditional density of $\boldsymbol{\theta}_h$ and $\mathbb{E}_{-h}^{(k)}(\cdot)$ denotes the variational expectation calculated with respect to the density $q_{-h}^{(k)}(\boldsymbol{\theta}_{-h}) = \prod_{\ell \neq h} q_\ell^{(k)}(\boldsymbol{\theta}_\ell)$. The iterative refinement of the variational posterior $q^{(k)}(\boldsymbol{\theta})$ through (1.29) gives rise to the coordinate ascent variational inference (CAVI) algorithm for MFVB approximate inference. For a detailed proof of this and related results on MFVB inference, we refer the reader to Bishop (2006), Ormerod and Wand (2010) and Blei *et al.* (2017).

MFVB shares many properties and similarities with both EM algorithm and Gibbs sampling. It is based on a conditioning principle which leverages the structure of the posterior full-conditional densities, allowing for closed form updates under conjugate priors. As a consequence, it can be easily extended to accommodate for augmented models $\pi(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y})$ with mean field approximation $q(\boldsymbol{\omega}, \boldsymbol{\theta}) = \prod_{i=1}^{n} q_i(\omega_i) \prod_{h=1}^{H} q_h(\boldsymbol{\theta}_h)$. In such cases, the coordinate updates of $q_i^{(k+1)}(\omega_i)$ and $q_h^{(k+1)}(\boldsymbol{\theta}_h)$ are given by

$$\begin{aligned}
q_h^{(k+1)}(\boldsymbol{\theta}_h) &\propto \exp\left[\mathbb{E}_{-\boldsymbol{\theta}_h}^{(k)}\{\log \pi(\boldsymbol{\theta}_h|\text{rest})\}\right], \qquad && h = 1, \dots, H, \\
q_i^{(k+1)}(\omega_i) &\propto \exp\left[\mathbb{E}_{-\omega_i}^{(k)}\{\log \pi(\omega_i|\text{rest})\}\right], \qquad && i = 1, \dots, n.
\end{aligned}$$

CAVI then provides a stable optimization routine, which almost surely converges to a local optimum of the objective functional; moreover, each iteration of the algorithm produces a non-increasing sequence of lower bound values, that is

$$\underline{\ell}\{\mathbf{y}; q^{(k)}(\boldsymbol{\theta})\} \leq \underline{\ell}\{\mathbf{y}; q^{(k+1)}(\boldsymbol{\theta})\}. \tag{1.30}$$

A straightforward implementation of the mean field principle for Bayesian quantile regression models with Asymmetric-Laplace likelihoods have been provided by Wand *et al.* (2011) and McLean and Wand (2019). In the following, we give a sketch of their results under the minimal product restriction

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \approx q(\boldsymbol{\omega}, \boldsymbol{\beta}, \sigma_\varepsilon^2) = q(\boldsymbol{\omega})\, q(\boldsymbol{\beta})\, q(\sigma_\varepsilon^2),$$

which leads to the induced factorization $q(\boldsymbol{\omega}) = \prod_{i=1}^n q(\omega_i)$. Then, thanks to the analytic expression of the full-conditional distribution (1.22), the optimal update for $q^*(\boldsymbol{\beta})$ at the $(k+1)$-th iteration of the CAVI algorithm is $q^{(k+1)}(\boldsymbol{\beta}) \sim \mathrm{N}_p(\boldsymbol{\mu}_\beta^{(k)}, \boldsymbol{\Sigma}_\beta^{(k)})$, with mean and variance

$$\boldsymbol{\mu}_\beta^{(k)} = \mu_{1/\sigma_\varepsilon^2}^{(k)} \boldsymbol{\Sigma}_\beta^{(k)} \mathbf{X}^\top \mathbf{W}^{(k)} \mathbf{z}^{(k)}, \qquad \boldsymbol{\Sigma}_\beta^{(k)} = \big\{\mu_{1/\sigma_\varepsilon^2}^{(k)}\big\}^{-1} \big(\mathbf{X}^\top \mathbf{W}^{(k)} \mathbf{X} + \mathbf{R}/\sigma_\beta^2\big)^{-1}, \tag{1.31}$$

where $\mathbf{W}^{(k)} = \mathrm{diag}\big[\boldsymbol{\mu}_{1/\omega}^{(k)}\big]/a_2^2$ and $\mathbf{z}^{(k)} = \mathbf{y} - a_1/\boldsymbol{\mu}_{1/\omega}^{(k)}$. From full-conditional distribution (1.23), we can derive the optimal density $q^{(k+1)}(\sigma_\varepsilon^2) \sim \mathrm{IG}(A_\varepsilon^{(k)}, B_\varepsilon^{(k)})$, which has variational parameters

$$A_\varepsilon^{(k)} = A_\varepsilon + \tfrac{1}{2}p + \tfrac{3}{2}n, \qquad B_\varepsilon^{(k)} = B_\varepsilon + C_1^{(k)} + C_2^{(k)}, \tag{1.32}$$

and

$$\begin{aligned} C_1^{(k)} &= \tfrac{1}{2}\Big\{\big\|\boldsymbol{\mu}_\beta^{(k)}\big\|_{\mathbf{R}}^2 + \mathrm{trace}\big[\mathbf{R}\boldsymbol{\Sigma}_\beta^{(k)}\big]\Big\}/\sigma_\beta^2, \\ C_2^{(k)} &= \tfrac{1}{2}\Big\{\boldsymbol{\mu}_{1/\omega}^{(k)\top} \boldsymbol{\mu}_{\varepsilon^2}^{(k)} - 2\lambda \mathbf{1}_n^\top \boldsymbol{\mu}_\varepsilon^{(k)} + (a_1^2 + 2a_2^2)\mathbf{1}_n^\top \boldsymbol{\mu}_{1/\omega}^{(k)}\Big\}/a_2^2, \end{aligned}$$

with $\boldsymbol{\mu}_\varepsilon^{(k)} = \mathbf{y} - \mathbf{X}\boldsymbol{\mu}_\beta^{(k)}$ and $\boldsymbol{\mu}_{\varepsilon^2}^{(k)} = \big\{\boldsymbol{\mu}_\varepsilon^{(k)}\big\}^2 + \mathrm{trace}\big\{\mathbf{X}\,\boldsymbol{\Sigma}_\beta^{(k)}\mathbf{X}^\top\big\}$. Finally, taking the variational expectation of the log-density associated to full-conditional distribution (1.20), we get the approximation $q^{(k+1)}(\omega_i) \sim \mathrm{GIG}(\frac{1}{2}, A_{\omega_i}^{(k)}, B_{\omega_i}^{(k)})$, $i = 1, \dots, n$, where

$$A_{\omega_i}^{(k)} = \mu_{1/\sigma_\varepsilon^2}^{(k)}(a_1^2 + 2a_2^2)/a_2^2, \qquad B_{\omega_i}^{(k)} = \mu_{1/\sigma_\varepsilon^2}^{(k)} \mu_{\varepsilon_i^2}^{(k)}/a_2^2. \tag{1.33}$$

Form such an approximation, we can also calculate the expectations $\mu_{\omega_i}^{(k)} = \mathbb{E}_q^{(k)}(\omega_i)$ and $\mu_{1/\omega_i}^{(k)} = \mathbb{E}_q^{(k)}(1/\omega_i)$ using closed form results for GIG distributions (Jørgensen, 1982).

Finally, exploiting the closed form of the variational distributions in (1.31), (1.32) and (1.33), we can analytically evaluate the evidence lower bound, that is

$$\begin{aligned} \underline{\ell}(\mathbf{y}; q^{(k)}) = &- \tfrac{1}{2}\, \mu_{1/\sigma_\varepsilon^2}^{(k)} \big\{\boldsymbol{\mu}_{1/\omega}^{(k)\top} \boldsymbol{\mu}_{\varepsilon^2}^{(k)} - 2a_1 \mathbf{1}_n^\top \boldsymbol{\mu}_\varepsilon^{(k)} + (a_1^2 + 2a_2^2)\mathbf{1}_n^\top \boldsymbol{\mu}_\omega^{(k)}\big\}/a_2^2 \\ &+ \tfrac{n}{2} \log A_\omega^{(k)} + \mathbf{1}_n^\top \big\{(A_\omega^{(k)} \boldsymbol{B}_\omega^{(k)})^{1/2} - \tfrac{1}{2}\, A_\omega^{(k)} \boldsymbol{\mu}_\omega^{(k)} - \tfrac{1}{2}\, \boldsymbol{B}_\omega^{(k)} \odot \boldsymbol{\mu}_{1/\omega}^{(k)}\big\} \\ &+ \tfrac{1}{2} \,\mathrm{logdet}(\boldsymbol{\Sigma}_\beta^{(k)}) - \tfrac{1}{2}\mu_{1/\sigma_\varepsilon^2}^{(k)} \big\{\|\boldsymbol{\mu}_\beta^{(k)}\|_{\mathbf{R}}^2 + \mathrm{tr}(\mathbf{R}\boldsymbol{\Sigma}_\beta^{(k)})\big\}/\sigma_\beta^2 \\ &- \log\big\{\Gamma(A_\varepsilon)/\Gamma(A_\varepsilon + \tfrac{p}{2} + \tfrac{3}{2}n)\big\} + A_\varepsilon \log(B_\varepsilon/B_\varepsilon^{(k)}) \\ &- (\tfrac{p}{2} + \tfrac{3}{2}n) \log B_\varepsilon^{(k)} - (B_\varepsilon - B_\varepsilon^{(k)})\mu_{1/\sigma_\varepsilon^2}^{(k)} + \mathrm{const}, \end{aligned}$$

where $\mathrm{const} = -\frac{n}{2}\log(2\pi a_2^2) - \frac{p}{2}\log\sigma_\beta^2 - \frac{1}{2}\,\mathrm{logdet}(\mathbf{R}) + \frac{p}{2}$ is a constant term not depending on the parameters and on the variational distributions.

```
**Algorithm 3** MFVB algorithm for quantile regression
Require: $\tau, \mathbf{y}, \mathbf{X}, \sigma^2_\beta, A_\varepsilon, B_\varepsilon$
    Initialize $\hat{\boldsymbol{\mu}}_\beta$, $\hat{\boldsymbol{\Sigma}}_\beta$, $\hat{\mu}_{1/\sigma^2_\varepsilon}$, $\hat{\boldsymbol{\mu}}_\omega$, $\hat{\boldsymbol{\mu}}_{1/\omega}$, $\hat{\boldsymbol{\mu}}_\varepsilon$, $\hat{\boldsymbol{\mu}}_{\varepsilon^2}$;
    while convergence is not reached do
        $\hat{\boldsymbol{\mu}}_{1/\omega} \leftarrow (a_1^2 + 2a_2^2)^{1/2} / \hat{\boldsymbol{\mu}}_{\varepsilon^2}^{1/2}$;    $\hat{\boldsymbol{\mu}}_\omega \leftarrow \hat{\boldsymbol{\mu}}_{1/\omega}^{-1} + \hat{\mu}_{1/\sigma^2_\varepsilon}^{-1} a_2^2/(a_1^2 + 2a_2^2)$;
        $\hat{\mathbf{W}} \leftarrow \mathrm{diag}\big[\hat{\boldsymbol{\mu}}_{1/\omega}\big]/a_2^2$;    $\hat{\mathbf{z}} \leftarrow \mathbf{y} - a_1 \hat{\boldsymbol{\mu}}_{1/\omega}^{-1}$;
        $\hat{\boldsymbol{\Sigma}}_\beta \leftarrow \hat{\mu}_{1/\sigma^2_\varepsilon}^{-1} (\mathbf{X}^\top \hat{\mathbf{W}} \mathbf{X} + \mathbf{R}/\sigma^2_\beta)^{-1}$;    $\hat{\boldsymbol{\mu}}_\beta \leftarrow \hat{\mu}_{1/\sigma^2_\varepsilon} \hat{\boldsymbol{\Sigma}}_\beta \mathbf{X}^\top \hat{\mathbf{W}} \hat{\mathbf{z}}$;
        $\hat{\boldsymbol{\mu}}_\varepsilon \leftarrow \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\mu}}_\beta$;    $\hat{\boldsymbol{\mu}}_{\varepsilon^2} \leftarrow \hat{\boldsymbol{\mu}}_\varepsilon^2 + \mathrm{diag}\big[\mathbf{X}\hat{\boldsymbol{\Sigma}}_\beta \mathbf{X}^\top\big]$;
        $C_1 \leftarrow \frac{1}{2}\big\{\hat{\boldsymbol{\mu}}_\beta^\top \mathbf{R} \hat{\boldsymbol{\mu}}_\beta + \mathrm{trace}\big[\mathbf{R}\hat{\boldsymbol{\Sigma}}_\beta\big]\big\}/\sigma^2_\beta$;
        $C_2 \leftarrow \frac{1}{2}\big\{\hat{\boldsymbol{\mu}}_{1/\omega}^\top \hat{\boldsymbol{\mu}}_{\varepsilon^2} - 2\lambda \mathbf{1}_n^\top \hat{\boldsymbol{\mu}}_\varepsilon + (\lambda^2 + 2\delta^2)\mathbf{1}_n^\top \hat{\boldsymbol{\mu}}_\omega\big\}/\delta^2$;
        $\hat{A}_\varepsilon \leftarrow A_\varepsilon + \frac{p}{2} + \frac{3}{2}n$;    $\hat{B}_\varepsilon \leftarrow B_\varepsilon + C_1 + C_2$;    $\hat{\mu}_{1/\sigma^2_\varepsilon} \leftarrow \hat{B}_\varepsilon/\hat{A}_\varepsilon$;
    end while
```

The recursive execution of the closed form updates in (1.31), (1.32) and (1.33) give rise to the CAVI routine summarized in Algorithm 3.

Chapter 2

Non-conjugate regression via variational belief updating

2.1 Introduction

The increasing prevalence of big volume and velocity data, eventually coming from different data sources, entails a great opportunity but also a major challenge of modern data analysis and, in particular, of Bayesian statistics. The computational burden required by Markov chain Monte Carlo simulation, that is the state-of-the-art approach to Bayesian inference (Gelman *et al.*, 2013), is often not compatible with time constraints and memory limits. Therefore, in the last two decades many efforts have been spent to develop alternative estimation methods not involving posterior simulation. In this context, optimization-based algorithms play an important role because of their ability to provide a reasonably good approximation of the posterior, while keeping a high level of efficiency. Within this family, two remarkable examples are the Laplace approximation (Wolfinger, 1994), along with its integrated nested generalization (Rue *et al.*, 2009), and the variational inference approach, which includes, among others, local variational approximation (Jaakkola and Jordan, 2000), mean field variational Bayes (Ormerod and Wand, 2010; Blei *et al.*, 2017), expectation propagation (Minka, 2013; Bishop, 2006, Chapter 10), black-box stochastic variational inference (Ranganath *et al.*, 2014; Kucukelbir *et al.*, 2017; Ong *et al.*, 2018), and natural gradient stochastic variational inference (Hoffman *et al.*, 2013; Khan and Lin, 2017; Khan and Rue, 2021). All of these are mainly concerned with the estimation of hierarchical Bayesian models with a regular likelihood function, often belonging to the exponential family. For instance, both Laplace approximation and stochastic variational inference require for the log-likelihood function to be differentiable with continuity from one to three times, depending on the implementation. Local variational approximations leverage the concept of convex duality and require for the log-likelihood to have Lipschitz continuous gradient. On the other hand, mean field variational Bayes and expectation propagation are employed when the hierarchical model can be described through a Bayesian factor graph with locally conjugate nodes, like for the Gibbs sampling algorithm.

The joint lack of smoothness and conjugacy constitutes one of the main issues when it comes to approximating a posterior density, and, in particular, when the parameters of interest do not index a known family of probability distributions but, instead, are defined according to a minimum risk criterion. In these cases, the minimal regularity conditions for an estimation problem to be well-defined may not be satisfied, as it happens for support vector machines (Vapnik, 1998), quantile and expectile regression (Koenker, 2005).

The present work aims to introduce a unified variational methodology, alternative to data-augmented methods, for approximating the general posterior distribution of a Bayesian regression model, which combines our subjective prior beliefs with the information coming from an empirical risk function. Doing this, we take the subjective perspective introduced by Bissiri *et al.* (2016), which proved the theoretical coherence of updating a prior to the posterior using the negative loss as if it was the kernel of a proper log-likelihood. In this context, we also mention the work of Alquier *et al.* (2016) and Alquier and Ridgway (2020), which investigated both the finite- and large-sample properties of variational approximations to Gibbs posteriors, while Wang and Blei (2019a) established the frequentist consistency and asymptotic normality of variational Bayes under model misspecification.

The range of applications of the proposed method virtually permits dealing with all the standard regression and classification models predicting the response variable through a linear predictor, including generalized linear models. Here, particular attention is devoted to loss functions characterized by a non-regular behavior and for which standard Bayesian approximation techniques can not be directly employed, or may suffer severe drawbacks. Our approach combines the efficiency and modularity of mean field variational Bayes with Gaussian variational approximations to deal with parameters not having a conjugate distribution. Previously, a similar strategy, known as semiparametric variational Bayes (Knowles and Minka, 2011; Wand, 2014; Rohde and Wand, 2016), has been used for the estimation of Bayesian generalized linear models, heteroscedastic regression models and Gaussian process regression by, e.g., Ormerod and Wand (2012), Tan and Nott (2013), Wand (2014), Luts and Wand (2015), Menictas and Wand (2015), and Khan and Lin (2017).

2.2 Motivating problems and results

2.2.1 Variational data-augmentation

Data-augmentation is a powerful tool that permits representing complex pseudo-likelihood functions as the marginal density of a joint model, that is only partially observed (see Chapter 1). This technique, first formalized and popularized by Dempster *et al.* (1977), has been successfully used in a number of different contexts for the estimation of complex statistical models. Some examples are finite-mixture models (Frühwirth-Schnatter, 2006), probit regression (Albert and Chib, 1993), logit regression (Polson *et al.*, 2013), quantile regression (Kozumi and Kobayashi, 2011), support vector machines (Polson and Scott, 2011), sparse regression (Carvalho *et al.*, 2010; Griffin and Brown, 2011; Bhattacharya *et al.*, 2015).

Such an approach provides several benefits: it often restores the conjugacy between likelihood and priors, it may regularize non-smooth optimization problems, it opens the door to several iterative estimation algorithms with closed-form updating formulas, and, finally, it provides a fascinating probabilistic interpretation in terms of missing information and completed data. On the other hand, data augmentation and related methods often suffer significant drawbacks related to the enlarging of the parameter space with a number of unknown variables that is proportional to the sample size. Indeed, it is well recognized how EM, MCMC and MFVB algorithms based on augmented representations of the original likelihood are typically prone to slow convergence, poor mixing, strong autocorrelation, stacking on local optima, and an increasing computational inefficiency at the growing of the sample size; see, e.g., Lewandowski *et al.* (2010) for EM, Duan *et al.* (2018) and Johndrow *et al.* (2019) for MCMC, Neville *et al.* (2014) and Loaiza-Maya *et al.* (2022) for MFVB.

In the variational literature, pitfalls related to data-augmentation methods have been empirically studied by Neville *et al.* (2014), which also provided a geometrical interpretation and a theoretical analysis of marginal and augmented variational approximations under simple statistical models. Being aware of the risks entailed by data-augmentation, many authors proposed alternative methods to circumvent these problems, improving standard approximations while maintaining computational tractability. Some examples are given by Fasano *et al.* (2022) and Loaiza-Maya *et al.* (2022).

In this context, it is worth understanding if and how there is a convenience working with augmented models instead of relying on a marginal formulation. To this end, we assume that (i) both the marginal and the completed models are available, (ii) a variational approximation can be produced both for the marginal and the joint models, and (iii) there is no interest in estimating the unknown distribution of the missing data. Under these assumptions, we compare the posterior approximation accuracy obtained under the marginal and the joint variational approximations using the Kullback-Leibler divergence. We then provide a general theoretical result which establishes the sub-optimality of variational approximations based on data-augmented models.

Let us assume that the (pseudo-)likelihood function $\pi(\mathbf{y}|\boldsymbol{\theta})$ can be expressed as the convolution of an augmented likelihood $\pi(\mathbf{y}|\boldsymbol{\omega},\boldsymbol{\theta})$ with a mixing density $\pi(\boldsymbol{\omega}|\boldsymbol{\theta})$, which may be improper, namely

$$\pi(\mathbf{y}|\boldsymbol{\theta}) = \int_{\Omega} \pi(\mathbf{y}|\boldsymbol{\omega},\boldsymbol{\theta})\, \pi(\boldsymbol{\omega}|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\omega}, \tag{2.1}$$

where $\boldsymbol{\omega} \in \Omega$ is a non-observed latent vector.

Let us suppose that the interest of the analysis is to estimate the marginal posterior distribution for the parameters $\boldsymbol{\theta} \in \Theta$, while $\boldsymbol{\omega} \in \Omega$ can be considered as a vector of nuisance latent variables. The augmented representation (2.1) is here considered as a computational tool employed to simplify the calculations within approprietely designed model-specific estimation algorithms.

We define the marginal and the augmented approximations, respectively, $q_{\text{M}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\text{M}}$ and $q_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta}) \in \mathcal{Q}_{\text{A}}$. Further, we define $q_{\text{A}}(\boldsymbol{\theta}) = \int_{\Omega} q_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\omega}$ as the marginal density obtained by integrating out $\boldsymbol{\omega}$ from the augmented approximation $q_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta})$. Then, following a variational Bayes methodology, the optimal densities $q^*_{\text{M}}(\boldsymbol{\theta})$ and $q^*_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta})$ are obtained by minimizing the following Kullback-Leibler divergences

$$q^*_{\text{M}}(\boldsymbol{\theta}) = \underset{q \in \mathcal{Q}_{\text{M}}}{\operatorname{argmin}}\, \mathrm{KL}\{q(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\}, \quad q^*_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta}) = \underset{q \in \mathcal{Q}_{\text{A}}}{\operatorname{argmin}}\, \mathrm{KL}\{q(\boldsymbol{\omega}, \boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y})\}. \tag{2.2}$$

Here, we assume for $q^*_{\text{M}}(\boldsymbol{\theta})$ and $q^*_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta})$ to be the global minimizers of the respective optimization problems. Henceforth, we do not consider possible local minima.

In order to establish some connection between these two approximations, along with the respective divergences, some *compatibility* assumption has to be imposed over $\mathcal{Q}_{\text{M}}$ and $\mathcal{Q}_{\text{A}}$.

Assumption 1 (Compatibility)**.** *We say that* $\mathcal{Q}_{\text{A}}$ *is a density space* compatible *with* $\mathcal{Q}_{\text{M}}$ *if any marginal distribution* $q_{\text{A}}(\boldsymbol{\theta})$ *induced by the marginalization of* $\boldsymbol{\omega} \in \Omega$ *from* $q_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta})$ *belongs to* $\mathcal{Q}_{\text{M}}$. *That is*

$$\mathcal{Q}_{\text{A}} = \left\{ q_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta}) :\ q_{\text{A}}(\boldsymbol{\theta}) = \int_{\Omega} q_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\omega} \in \mathcal{Q}_{\text{M}} \right\}.$$

For instance, if we assume for $q_{\text{M}}(\boldsymbol{\theta})$ to be a Gaussian density, also $q_{\text{A}}(\boldsymbol{\theta})$ should be a Gaussian density in order to keep the compatibility condition satisfied. Similarly, if we do not specify any parametric assumption, but instead we assume a mean field factorization $q_{\text{M}}(\boldsymbol{\theta}) = \prod_k q_{\text{M}}(\boldsymbol{\theta}_k)$ over a given partition of $\boldsymbol{\theta}$, because of the compatibility, also $q_{\text{A}}(\boldsymbol{\theta}) = \prod_k q_{\text{A}}(\boldsymbol{\theta}_k)$ will factorize according to the same partition.

Under this hypothesis, we can characterize the behavior of the Kullback-Leibler divergence under the marginal and augmented approximation schemes using the following results.

Theorem 2.1. *Let* $\mathcal{Q}_{\text{M}}$ *and* $\mathcal{Q}_{\text{A}}$ *be compatible spaces as defined in Assumption 1. Let* $q^*_{\text{M}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\text{M}}$ *and* $q^*_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta}) \in \mathcal{Q}_{\text{A}}$ *be the optimal approximating distributions defined as in (2.2). Then, we have*

$$\mathrm{KL}\{q^*_{\text{M}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\} \leq \mathrm{KL}\{q^*_{\text{A}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\} \leq \mathrm{KL}\{q^*_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y})\}. \tag{2.3}$$

Proof. Let us start by considering the first inequality in Theorem 2.1. Let $q^*_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta}) \in \mathcal{Q}_{\text{A}}$, $q^*_{\text{A}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\text{M}}$ and $q^*_{\text{M}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\text{M}}$ be the optimal variational approximations defined in (2.2). Since $q^*_{\text{M}}(\boldsymbol{\theta})$ is the global minimum of the Kullback-Leibler divergence calculated with respect to the marginal space $\mathcal{Q}_{\text{M}}$, we have

$$\mathrm{KL}\{q^*_{\text{M}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\} \leq \mathrm{KL}\{\tilde{q}_{\text{M}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\},$$

for any density function $\tilde{q}_{\text{M}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\text{M}}$ and, in particular, for $\tilde{q}_{\text{M}}(\boldsymbol{\theta}) = q^*_{\text{A}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\text{M}}$. This proves the first inequality in Theorem 2.1.

Let us move to the second inequality. Let $q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) = q_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta})/q_{\text{A}}(\boldsymbol{\theta})$ be the conditional density function of $\boldsymbol{\omega}$ given $\boldsymbol{\theta}$ with respect to the q_{A}-measure. Then, from the Jensen inequality follows that

$$\log \pi(\boldsymbol{\theta}|\mathbf{y}) = \log \int_{\Omega} q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \left\{ \frac{\pi(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y})}{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\omega} \geq \int_{\Omega} q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \log \left\{ \frac{\pi(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y})}{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\omega}.$$

Thus, by substitution, we end up with

$$\begin{aligned} \int_{\Theta} q_{\text{A}}(\boldsymbol{\theta}) \log \left\{ \frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{q_{\text{A}}(\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\theta} &= \int_{\Theta} q_{\text{A}}(\boldsymbol{\theta}) \left[\log \pi(\boldsymbol{\theta}|\mathbf{y}) - \log q_{\text{A}}(\boldsymbol{\theta}) \right] \mathrm{d}\boldsymbol{\theta} \\ &\geq \int_{\Theta} q_{\text{A}}(\boldsymbol{\theta}) \left[\int_{\Omega} q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \log \left\{ \frac{\pi(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y})}{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\omega} - \log q_{\text{A}}(\boldsymbol{\theta}) \right] \mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{2.4}$$

Now, because of the identity

$$\log q_{\text{A}}(\boldsymbol{\theta}) = \underbrace{\left[\int_{\Omega} q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\omega} \right]}_{=1} \log q_{\text{A}}(\boldsymbol{\theta}) = \int_{\Omega} q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \, \log q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\omega},$$

the last integral in (2.4) becomes

$$\begin{aligned} &\int_{\Theta} q_{\text{A}}(\boldsymbol{\theta}) \left[\int_{\Omega} q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \log \left\{ \frac{\pi(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y})}{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\omega} - \int_{\Omega} q(\boldsymbol{\omega}|\boldsymbol{\theta}) \, \log q_{\text{A}}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\theta} \\ &\quad = \iint_{\Omega \times \Theta} q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \, q_{\text{A}}(\boldsymbol{\theta}) \left[\log \left\{ \frac{\pi(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y})}{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})} \right\} - \log q_{\text{A}}(\boldsymbol{\theta}) \right] \mathrm{d}\boldsymbol{\omega} \, \mathrm{d}\boldsymbol{\theta} \\ &\quad = \iint_{\Theta \times \Omega} q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \, q_{\text{A}}(\boldsymbol{\theta}) \log \left\{ \frac{\pi(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y})}{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \, q_{\text{A}}(\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\omega} \, \mathrm{d}\boldsymbol{\theta} \\ &\quad = \iint_{\Theta \times \Omega} q_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta}) \log \left\{ \frac{\pi(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y})}{q_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\omega} \, \mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{2.5}$$

Observing that the left integral in (2.4) corresponds to the negative Kullback-Leibler divergence $-\mathrm{KL}\{q_{\text{A}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\}$, while the last integral in (2.5) corresponds to $-\mathrm{KL}\{q_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y})\}$, we have

$$\mathrm{KL}\{q_{\text{A}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\} \leq \mathrm{KL}\{q_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y})\}.$$

This concludes the proof of Theorem 2.1. □

Remark 2.2. Inequalities (2.4) and (2.5) hold only under the *compatibility* Assumption 1 made upon $\mathcal{Q}_{\text{M}}$ and $\mathcal{Q}_{\text{A}}$, which permits to write $q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})$ as a proper conditional distribution generated by $q_{\text{A}}(\boldsymbol{\omega}, \boldsymbol{\theta})$ and $q_{\text{A}}(\boldsymbol{\theta})$. Whenever the compatibility is not satisfied there are no guarantees that Theorem 2.1 still holds.

In the following, we show this fact by using a counterexample.

Example 2.1. *Let us consider a generic augmented space* $\mathcal{Q}_{\text{A}}$ *such that* $\pi(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y}) \in \mathcal{Q}_{\text{A}}$. *On the opposite, we consider* $\mathcal{Q}_{\text{M}}$ *such that* $\pi(\boldsymbol{\theta}|\mathbf{y}) \notin \mathcal{Q}_{\text{M}}$. *This way,* $\mathcal{Q}_{\text{A}}$ *and* $\mathcal{Q}_{\text{M}}$ *are* *not compatible by construction, since there exists at least one element of* $\mathcal{Q}_{\text{A}}$ *whose marginal density does not belong to* $\mathcal{Q}_{\text{M}}$*, namely* $\pi(\boldsymbol{\theta}|\mathbf{y}) = \int_\Omega \pi(\boldsymbol{\omega},\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\omega}$*. Recall that the Kullback-Leibler divergence is always non-negative, i.e.* $\mathrm{KL}(q \parallel \pi) \geq 0$*, and reaches 0 if and only if* $q = \pi$ *almost everywhere, namely* $\mathrm{KL}(\pi \parallel \pi) = 0$*. Then, there exists at least one element of* $\mathcal{Q}_{\text{A}}$ *such that* $\mathrm{KL}\{q_{\text{A}}(\boldsymbol{\omega},\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega},\boldsymbol{\theta}|\mathbf{y})\} = 0$*, which corresponds to* $q_{\text{A}}(\boldsymbol{\omega},\boldsymbol{\theta}) = \pi(\boldsymbol{\omega},\boldsymbol{\theta}|\mathbf{y})$*, whereas* $\mathrm{KL}\{q_{\text{M}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\} > 0$ *for any* $q_{\text{M}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\text{M}}$*, since* $\pi(\boldsymbol{\theta}|\mathbf{y}) \notin \mathcal{Q}_{\text{M}}$*. This means that*

$$\underbrace{\mathrm{KL}\{q_{\text{M}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\}}_{>0} \not\leq \underbrace{\mathrm{KL}\{\pi(\boldsymbol{\omega},\boldsymbol{\theta}|\mathbf{y}) \parallel \pi(\boldsymbol{\omega},\boldsymbol{\theta}|\mathbf{y})\}}_{=0}, \qquad \forall\, q_{\text{M}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\text{M}},$$

which contradicts the inequality in Theorem 2.1.

In order to better contextualize Theorem 2.1, we provide the following corollary, which characterizes the inequalities in (2.3) separating the strict inequality cases from the equality ones.

Corollary 2.3. *Under the same assumptions of Theorem 2.1, the equality case*

$$\mathrm{KL}\{q^*_{\text{M}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\} = \mathrm{KL}\{q^*_{\text{A}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\} = \mathrm{KL}\{q^*_{\text{A}}(\boldsymbol{\omega},\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega},\boldsymbol{\theta}|\mathbf{y})\}$$

holds if, and only if, $\pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\, q^*_{\text{A}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\text{A}}$ *for* $q^*_{\text{A}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\text{M}}$ *and, moreover,* $q^*_{\text{A}}(\boldsymbol{\omega},\boldsymbol{\theta}) = \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\, q^*_{\text{A}}(\boldsymbol{\theta})$*. On the opposite, if* $\pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\, q^*_{\text{A}}(\boldsymbol{\theta}) \notin \mathcal{Q}_{\text{A}}$*, we have*

$$\mathrm{KL}\{q^*_{\text{M}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\} < \mathrm{KL}\{q^*_{\text{A}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\} < \mathrm{KL}\{q^*_{\text{A}}(\boldsymbol{\omega},\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega},\boldsymbol{\theta}|\mathbf{y})\}.$$

Proof. ($\Rightarrow$) Let consider the augmented variational approximation of the form $q_{\text{A}}(\boldsymbol{\omega},\boldsymbol{\theta}) = \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\, q_{\text{A}}(\boldsymbol{\theta})$. Then, the Kullback-Leibler divergence in the augmented space is

$$\begin{aligned}
&\mathrm{KL}\{\pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\, q_{\text{A}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega},\boldsymbol{\theta}|\mathbf{y})\} = \\
&\quad = -\iint_{\Theta\times\Omega} \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\, q_{\text{A}}(\boldsymbol{\theta}) \log\left\{\frac{\pi(\boldsymbol{\omega},\boldsymbol{\theta}|\mathbf{y})}{\pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\, q_{\text{A}}(\boldsymbol{\theta})}\right\} \mathrm{d}\boldsymbol{\omega}\,\mathrm{d}\boldsymbol{\theta} \\
&\quad = -\iint_{\Theta\times\Omega} \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\, q_{\text{A}}(\boldsymbol{\theta}) \log\left\{\frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{q_{\text{A}}(\boldsymbol{\theta})}\right\} \mathrm{d}\boldsymbol{\omega}\,\mathrm{d}\boldsymbol{\theta} \\
&\quad = -\int_\Theta \left[\int_\Omega \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\,\mathrm{d}\boldsymbol{\omega}\right] q_{\text{A}}(\boldsymbol{\theta}) \log\left\{\frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{q_{\text{A}}(\boldsymbol{\theta})}\right\} \mathrm{d}\boldsymbol{\theta} \\
&\quad = -\int_\Theta q_{\text{A}}(\boldsymbol{\theta}) \log\left\{\frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{q_{\text{A}}(\boldsymbol{\theta})}\right\} \mathrm{d}\boldsymbol{\theta} = \mathrm{KL}\{q_{\text{A}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\}
\end{aligned}$$

The optimal approximation $q^*_{\text{A}}(\boldsymbol{\omega},\boldsymbol{\theta}) = \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\, q^*_{\text{A}}(\boldsymbol{\theta})$ is thus the solution to the following variational problem $q^*_{\text{A}}(\boldsymbol{\theta}) = \operatorname{argmin}_{q_{\text{A}} \in \mathcal{Q}_{\text{M}}} \mathrm{KL}\{q_{\text{A}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\}$ which, by definition, corresponds to $q^*_{\text{M}}(\boldsymbol{\theta})$.

($\Leftarrow$) Let us consider the variational approximation $q_{\text{A}}(\boldsymbol{\omega},\boldsymbol{\theta}) = q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})\, q_{\text{A}}(\boldsymbol{\theta})$ such that $q_{\text{A}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\text{M}}$ and $q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \in \mathcal{Q}_c$, where $\mathcal{Q}_c$ denotes the space of variational conditional densities $q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})$ such that $q_{\text{A}}(\boldsymbol{\omega},\boldsymbol{\theta}) \in \mathcal{Q}_{\text{M}}$. Thus, the optimal variational distribution

$q^*_{\text{A}}(\boldsymbol{\omega},\boldsymbol{\theta})$ is the solution of the following optimization problem

$$\begin{aligned}
&\min_{q(\boldsymbol{\omega},\boldsymbol{\theta})\in\mathcal{Q}_{\text{A}}} \text{KL}\{q_{\text{A}}(\boldsymbol{\omega},\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega},\boldsymbol{\theta}|\mathbf{y})\} \\
&\quad = \min_{q_{\text{A}}(\boldsymbol{\theta})\in\mathcal{Q}_{\text{M}}} \left[\min_{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})\in\mathcal{Q}_c} \text{KL}\{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})\, q_{\text{A}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})\} \right].
\end{aligned}$$

The later Kullback-Leibler divergence may also be written as

$$\begin{aligned}
&\text{KL}\{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})\, q_{\text{A}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})\} \\
&\quad = -\iint_{\Theta\times\Omega} q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})\, q_{\text{A}}(\boldsymbol{\theta}) \log\left\{ \frac{\pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\,\pi(\boldsymbol{\theta}|\mathbf{y})}{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})\, q_{\text{A}}(\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{\omega} \\
&\quad = -\int_{\Theta} q_{\text{A}}(\boldsymbol{\theta}) \left[\int_{\Omega} q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \log\left\{ \frac{\pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})}{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\theta} - \int_{\Theta} q_{\text{A}}(\boldsymbol{\theta}) \log\left\{ \frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{q_{\text{A}}(\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\theta} \\
&\quad = \int_{\Theta} q_{\text{A}}(\boldsymbol{\theta})\, \text{KL}\{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\}\, \mathrm{d}\boldsymbol{\theta} + \text{KL}\{q_{\text{A}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\} \\
&\quad = \mathbb{E}_{q_{\text{A}}}\left[\text{KL}\{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\}\right] + \text{KL}\{q_{\text{A}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\}.
\end{aligned}$$

This reformulation is useful to highlight the fact that minimizing $\text{KL}\{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})\, q_{\text{A}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega},\boldsymbol{\theta}|\mathbf{y})\}$ with respect to $q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})$ is equivalent to minimizing $\text{KL}\{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\}$ with respect to $q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})$ keeping $q_{\text{A}}(\boldsymbol{\theta})$ fixed. Therefore, if $\pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y}) \in \mathcal{Q}_c$, $q^*_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) = \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})$ is the unique minimizer of $\text{KL}\{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\}$ and, furthermore, we have $\text{KL}\{\pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y}) \parallel \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\} = 0$ (Kullback and Leibler, 1951). On the other hand, if $\pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y}) \notin \mathcal{Q}_c$, $\text{KL}\{q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})\} > 0$ for any $q_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) \in \mathcal{Q}_c$ and also $\text{KL}\{q_{\text{A}}(\boldsymbol{\omega},\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\omega},\boldsymbol{\theta}|\mathbf{y})\} > \text{KL}\{q_{\text{A}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\}$. This concludes the proof. □

Starting from the right inequality in (2.3), Theorem 2.1 states that the divergence in the augmented space is always higher (or equal) than the divergence in the marginal space. The difference between the two can then be interpreted as the loss of information required for approximating a larger model keeping fixed the available set of data.

On the other hand, the left-hand side inequality (2.3) in Theorem 2.1 establishes that a loss of information in the augmented parameter space, i.e. $\Omega \times \Theta$, also reflects on a worsening of the approximation in the marginal space of interest, i.e. Θ. So that, under Assumption 1, marginal approximations dominate data-augmentation based approximations in the Kullback-Leibler metric.

Corollary 2.3 then states that the equivalence in the Kullback-Leibler metric may be reached if, and only if, the augmented approximation $q^*_{\text{A}}(\boldsymbol{\omega},\boldsymbol{\theta})$ takes the form $q^*_{\text{A}}(\boldsymbol{\omega},\boldsymbol{\theta}) = q^*_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta})\, q^*_{\text{A}}(\boldsymbol{\theta})$, where $q^*_{\text{A}}(\boldsymbol{\omega}|\boldsymbol{\theta}) = \pi(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{y})$ is actually the true full-conditional density function of $\boldsymbol{\omega}$ given $\boldsymbol{\theta}$ and $\mathbf{y}$. Similar observations have already been studied in literature, even if in a less general form (see, e.g., Theorem 1 and Corollary 1 by Loaiza-Maya *et al.*, 2022), and they are at the base of some recent variational approximations for models with many latent variables (Loaiza-Maya *et al.*, 2022; Fasano *et al.*, 2022).

These facts motivate the development of methods alternative to data-augmented mean field variational Bayes algorithms for posterior approximation, especially once different approaches lead to approximations having the same functional form.

2.2.2 Parametric variational Bayes

When the priors are not conjugate with the likelihood under study and, thus, mean field approximations do not enjoy closed form solutions, variational Bayes approximation can still be employed by imposing additional restrictions on the functional space $\mathcal{Q}$. Specifically, parametric variational inference circumvents the lack of conjugacy by imposing the parametric restriction

$$\mathcal{Q} = \Big\{ q(\boldsymbol{\theta}) : \ q(\boldsymbol{\theta}) = q(\boldsymbol{\theta}; \boldsymbol{\xi}), \ \boldsymbol{\xi} \in \Xi \Big\}, \tag{2.6}$$

for some user-specified family $q(\boldsymbol{\theta}; \boldsymbol{\xi})$ indexed by the variational parameter vector $\boldsymbol{\xi} \in \Xi$. The optimal variational approximation $q^*(\boldsymbol{\theta})$ can then be obtained by solving

$$q^*(\boldsymbol{\theta}) = \underset{q \in \mathcal{Q}}{\operatorname{argmax}} \ \underline{\ell}\{\mathbf{y}; q(\boldsymbol{\theta}; \boldsymbol{\xi})\}.$$

If $q(\boldsymbol{\theta}; \boldsymbol{\xi})$ is uniquely identified by its parameter vector $\boldsymbol{\xi}$, the optimization above can be equivalently expressed into finding the optimal value $\hat{\boldsymbol{\xi}}$, such that

$$\hat{\boldsymbol{\xi}} = \underset{\boldsymbol{\xi} \in \Xi}{\operatorname{argmax}} \ \underline{\ell}(\mathbf{y}; \boldsymbol{\xi}), \tag{2.7}$$

where $\underline{\ell}(\mathbf{y}; \boldsymbol{\xi}) = \underline{\ell}\{\mathbf{y}; q(\boldsymbol{\theta}; \boldsymbol{\xi})\}$. Under mild regularity conditions, the optimization (2.7) can be performed via general purpose optimizers, like quasi-Newton algorithms (Nocedal and Wright, 2006), conjugate gradient (Honkela *et al.*, 2010; Rohde and Wand, 2016) and stochastic search (Hoffman *et al.*, 2013; Kucukelbir *et al.*, 2017).

In the special case of variational densities belonging to the Exponential Family (EF) with canonical parameter $\boldsymbol{\xi}$, denoted by $q(\boldsymbol{\theta}; \boldsymbol{\xi}) \sim \mathrm{EF}(\boldsymbol{\xi})$, Knowles and Minka (2011) and Wand (2014) proposed an alternative, and more efficient, maximization scheme based on a fixed-point variational message passing procedure. Specifically, we consider the exponential family variational approximation

$$q(\boldsymbol{\theta}; \boldsymbol{\xi}) = H(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\xi}^\top T(\boldsymbol{\theta}) - A(\boldsymbol{\xi})\right\},$$

where $\boldsymbol{\xi}$ is the vector of canonical parameters, $H(\boldsymbol{\theta})$ is the *base measure*, $T(\boldsymbol{\theta})$ is the *natural* sufficient statistics and $A(\boldsymbol{\xi})$ is a twice differentiable, convex, *log-partition* function. Then, according to Knowles and Minka (2011) and Tan and Nott (2013), the optimal variational message passing iteration for climbing the evidence lower bound (1.26) is given by

$$\boldsymbol{\xi}^{(k+1)} \leftarrow \left[\mathrm{Var}_q^{(k)}\{T(\boldsymbol{\theta})\}\right]^{-1} \left[\nabla_{\boldsymbol{\xi}} \, \mathbb{E}_q^{(k)}\{\log \pi(\mathbf{y}, \boldsymbol{\theta})\}\right]. \tag{2.8}$$

As pointed out by Tan and Nott (2013) and Rohde and Wand (2016), the updating formula (2.8) constitutes a fixed-point iteration, which iteratively solve the first order equation $\nabla_{\boldsymbol{\xi}} \, \underline{\ell}(\mathbf{y}; \boldsymbol{\xi}) = 0$. Under local convexity assumptions on $\log \pi(\mathbf{y}, \boldsymbol{\theta})$, (2.8) is guaranteed to converge to a local stationary point of $\underline{\ell}(\mathbf{y};\boldsymbol{\xi})$ with a local convergence rate which depends on the curvature of the evidence lower bound in a neighborhood of the maximum.

Let us now consider the Gaussian variational approximation $q(\boldsymbol{\theta};\boldsymbol{\xi}) \equiv \phi(\boldsymbol{\theta};\boldsymbol{\mu},\boldsymbol{\Sigma})$, where the canonical parameter is defined by $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^\top, \boldsymbol{\xi}_2^\top)^\top$, with $\boldsymbol{\xi}_1 = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $\boldsymbol{\xi}_2 = -\frac{1}{2}\text{vech}(-2\boldsymbol{\Sigma}^{-1})$. Then, following Wand (2014) and Rohde and Wand (2016), employing the parametrization $(\boldsymbol{\mu},\boldsymbol{\Sigma}) \mapsto (\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)$, calculating the variational message passing update (2.8), and reparametrizing back to $(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) \mapsto (\boldsymbol{\mu},\boldsymbol{\Sigma})$, the optimal update for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\mu}^{(k+1)} \leftarrow \boldsymbol{\mu}^{(k)} - \big[\mathbf{H}^{(k)}\big]^{-1}\mathbf{g}^{(k)}, \qquad \boldsymbol{\Sigma}^{(k+1)} \leftarrow -\big[\mathbf{H}^{(k)}\big]^{-1}. \tag{2.9}$$

Here, $\mathbf{g}^{(k)} = \nabla_{\boldsymbol{\mu}}\,\underline{\ell}(\mathbf{y};\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)})$ and $\mathbf{H}^{(k)} = \nabla^2_{\boldsymbol{\mu}}\,\underline{\ell}(\mathbf{y};\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)})$ denote the gradient and Hessian of $\underline{\ell}(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$ calculated with respect to $\boldsymbol{\mu}$ and evaluated at the current estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. It is worth noting that, under suitable differentiability conditions on the log-likelihood function $\ell(\boldsymbol{\theta};\mathbf{y}) = \log\pi(\mathbf{y}|\boldsymbol{\theta})$, the above gradient and Hessian may be calculated as

$$\begin{aligned}
\mathbf{g}^{(k)} &= \nabla_{\boldsymbol{\mu}}\,\mathbb{E}_q^{(k)}\{\log\pi(\mathbf{y},\boldsymbol{\theta})\} = \mathbb{E}_q^{(k)}\big\{\nabla_{\boldsymbol{\theta}}\log\pi(\mathbf{y},\boldsymbol{\theta})\big\},\\
\mathbf{H}^{(k)} &= \nabla^2_{\boldsymbol{\mu}}\,\mathbb{E}_q^{(k)}\{\log\pi(\mathbf{y},\boldsymbol{\theta})\} = \mathbb{E}_q^{(k)}\big\{\nabla^2_{\boldsymbol{\theta}}\log\pi(\mathbf{y},\boldsymbol{\theta})\big\};
\end{aligned}$$

see, e.g., Opper and Archambeau (2009), Ormerod and Wand (2012) and Tan and Nott (2013). This fact highlights a different interpretation on (2.9), which can be viewed as a variational implementation of the Newton-Rapson algorithm used for finding the maximum likelihood estimator of $\boldsymbol{\theta}$.

In the variational literature, formula (2.9) is also known as Knowles-Minka-Wand update, from the authors that firstly connected such an updating rule with the variational message passing theory (Knowles and Minka, 2011; Wand, 2014).

Differently from conjugate mean field variational Bayes, the Newton-like step (2.9) does not guarantee for the semiparametric variational Bayes optimization scheme to converge monotonically to the maximum. For this reason, in this work we take a slightly different approach, employing a modified update for $\boldsymbol{\mu}$, that scales the searching direction $\mathbf{d}^{(k)} = -\big[\mathbf{H}^{(k)}\big]^{-1}\mathbf{g}^{(k)}$ with a step-size parameter $\rho \in (0,1]$, that is $\boldsymbol{\mu}^{(k+1)} \leftarrow \boldsymbol{\mu}^{(k)} + \rho\,\mathbf{d}^{(k)}$. The step-length ρ can then be determined using an efficient line-search algorithm that enforces the Armijo-Wolfe conditions to be satisfied (Nocedal and Wright, 2006). This improvement has two main advantages: first, it stabilizes the iterative optimization, preventing the risk of jumping too far from the optimum; second, it helps in scaling up the convergence speed, adaptively calibrating the step-size to the shape of the lower bound surface.

For Bayesian models where only a subset of the parameters have a non-conjugate prior, mean field variational Bayes and parametric variational approximations can be combined within a unified semiparametric variational algorithm. More formally, let $\boldsymbol{\theta} = (\boldsymbol{\phi},\boldsymbol{\psi})$ be a partition of the parameter vector, with sub-partitions $\boldsymbol{\phi} = (\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_{N_\phi})$

and $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_{N_\psi})$. Then, assuming the factorization

$$\mathcal{Q} = \left\{ q(\boldsymbol{\theta}) : \; q(\boldsymbol{\theta}) = \prod_{h=1}^{N_\phi} q_h(\boldsymbol{\phi}_h) \prod_{\ell=1}^{N_\psi} q_\ell(\boldsymbol{\psi}_\ell; \boldsymbol{\xi}_\ell), \; \boldsymbol{\xi}_\ell \in \Xi_\ell \right\},$$

and the parametric restriction $q(\boldsymbol{\psi}_\ell; \boldsymbol{\xi}_\ell) \sim \mathrm{EF}(\boldsymbol{\xi}_\ell)$, the optimal variational approximation $q^*(\boldsymbol{\theta})$ may be obtained iterating until convergence the coordinatewise optimization

$$\begin{aligned} q_h^{(k+1)}(\boldsymbol{\phi}_h) &\leftarrow \exp\left[\mathbb{E}_{-\boldsymbol{\phi}_h}^{(k)}\{\log \pi(\boldsymbol{\phi}_h|\mathrm{rest})\}\right], && h = 1, \dots, N_\phi, \\ \boldsymbol{\xi}_\ell^{(k+1)} &\leftarrow \left[\mathrm{Var}_q^{(k)}\{T_\ell(\boldsymbol{\psi}_\ell)\}\right]^{-1} \left[\nabla_{\boldsymbol{\xi}_\ell} \mathbb{E}_q^{(k)}\{\log \pi(\mathbf{y}, \boldsymbol{\theta})\}\right], && \ell = 1, \dots, N_\psi. \end{aligned}$$

This approach has been employed by Tan and Nott (2013) for the estimation of generalized linear mixed models with a partially non-centered parametrization, by Wand (2014) and Menictas and Wand (2015) for the estimation of heteroscedastic regression models, and by Luts and Ormerod (2014) for the estimation of regression models for counting data.

2.3 Approximate belief updating

We now suppose to be interested in the unknown parameter $\boldsymbol{\theta} \in \Theta$, which describes some latent features of the random variable $y \in \mathcal{Y} \subseteq \mathbb{R}$. Let $y \sim \Pi$ be distributed according to the probability law Π defined over $\mathcal{Y}$. Then, we define $\boldsymbol{\theta}$ as the minimizer of the risk

$$R(\boldsymbol{\theta}) = \mathbb{E}\{L(y, \boldsymbol{\theta})\} = \int_{\mathcal{Y}} L(y, \boldsymbol{\theta}) \, \Pi(\mathrm{d}y), \tag{2.10}$$

with $L : \mathcal{Y} \times \Theta \to \mathbb{R}_+$ denoting a loss function measuring the misfit between y and $\boldsymbol{\theta}$.

Differently from classical M-estimation framework discussed in Chapter 1, here we are interested in combining a risk-based description of the data behavior with a subjective prior distribution on $\boldsymbol{\theta}$. In a Bayesian vein, we represent our subjective beliefs about $\boldsymbol{\theta}$ through the prior distribution $\pi(\boldsymbol{\theta})$. Then, when we observe a random sample $\mathbf{y} = (y_1, \dots, y_n)^\top \in \mathcal{Y}^n$ generated by $y \sim \Pi$, the prior $\pi(\boldsymbol{\theta})$ can be coherently updated to the posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$ by means of the generalized Bayes formula (Bissiri *et al.*, 2016):

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta}) \, \pi(\mathbf{y}|\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) \exp\{-n R_n(\boldsymbol{\theta})/\alpha\}, \tag{2.11}$$

where $\pi(\mathbf{y}|\boldsymbol{\theta}) = \exp\{-n R_n(\boldsymbol{\theta})/\alpha\}$ denotes the pseudo-likelihood function induced by the empirical risk $R_n(\boldsymbol{\theta}) = \mathbb{E}_n\{L(y, \boldsymbol{\theta})\}$, which approximates the integral in (2.10). The parameter $1/\alpha > 0$ controls the weight of the risk function relative to the prior and is often called *temperature* in the Gibbs posterior and PAC Bayes literature (Alquier *et al.*, 2016).

Whenever a proper log-likelihood is considered for the specification of the loss $L(y, \boldsymbol{\theta})$, the most natural candidate for the temperature parameter is $\alpha = 1$. In all the other cases, the elicitation of α is more delicate and, to some extent, arbitrary. The choice of α

can be driven by inferential considerations, for example, calibrating α to guarantee some frequentist coverage level of the posterior confidence intervals. This selection problem has been faced, among others, by Germain *et al.* (2016) and Bissiri *et al.* (2016). Despite the practical and theoretical importance of choosing an appropriate value for α, this is not the main concern of this work. For this reason, hereafter, we will consider α as it was selected in advance either by an arbitrary decision of the researcher, or by using some objective selection criteria.

The generalized Bayes update (2.11) is particularly important when it comes to estimating robust regression models for an arbitrary loss function $L(y, \boldsymbol{\theta})$, not being necessarily associated with the kernel of a probability density function. In what follows we leverage this representation for dealing with a wide range of mixed prediction models within the same theoretical framework.

2.3.1 Model specification

Supervised mixed regression and classification models aim at predicting the response variable $y_i \in \mathcal{Y}$ through a deterministic transformation of the covariate vectors $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{z}_i \in \mathbb{R}^d$, for $i = 1, \dots, n$. To this end, we define the linear predictor $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{u}$, with $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{u} \in \mathbb{R}^d$ being the fixed- and random-effect regression parameters, respectively. We assume for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ to be the maximizers of the negative empirical risk function, i.e. the pseudo log-likelihood,

$$\log \pi(\mathbf{y}|\boldsymbol{\theta}) = -\frac{n}{\alpha} R_n(\boldsymbol{\theta}) = -\frac{1}{\alpha} \sum_{i=1}^n L(y_i, \boldsymbol{\theta}) = -\frac{n}{\alpha} \log \sigma_\varepsilon^2 - \frac{1}{\alpha \sigma_\varepsilon^2} \sum_{i=1}^n \psi(y_i, \eta_i), \tag{2.12}$$

where $\psi : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ is a continuous function measuring the misfit between the i-th data point y_i and the corresponding linear predictor η_i; σ_ε^2 is a dispersion parameter measuring the variability of the marginal prediction error calculated in the loss scale; α is a fixed temperature parameter calibrating the relative weight of the risk function.

In the same vein as generalized linear models, which are based upon the specification of an exponential family, a linear predictor and a link function, the risk formulation in (2.12) encapsulates a large number of regression models defined by a loss function and a linear predictor. Moreover, for an appropriate specification of ψ, generalized linear models can be included in our model specification and σ_ε^2 can be interpreted as the dispersion parameter of an exponential dispersion family. More details are outlined in Section 2.4.

We complete the model specification by introducing a prior distribution that reflects our subjective beliefs about the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{u}, \sigma_u^2, \sigma_\varepsilon^2)$:

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathrm{N}_p\big(\mathbf{0}_p, \sigma_\beta^2 \mathbf{I}_p\big), & \sigma_\varepsilon^2 &\sim \mathrm{IG}(A_\varepsilon, B_\varepsilon), \\ \boldsymbol{u}|\sigma_u^2 &\sim \mathrm{N}_d\big(\mathbf{0}_d, \sigma_u^2 \mathbf{R}^{-1}\big), & \sigma_u^2 &\sim \mathrm{IG}(A_u, B_u), \end{aligned} \tag{2.13}$$

where $\sigma_\beta^2, A_\varepsilon, B_\varepsilon, A_u, B_u > 0$ are fixed user-specified prior parameters; $\sigma_u^2 > 0$ is an unknown scale parameter controlling the marginal variability of the random effect vector

$\boldsymbol{u}$; while $\mathbf{R} \in \mathbb{R}^{d \times d}$ is a non-stochastic positive semi-definite matrix determining the prior conditional correlation structure among the elements of $\boldsymbol{u}$.

For models where σ_ε^2 is not required, its prior distribution can be set as a Dirac delta function centered in 1, i.e. $\pi(\sigma_\varepsilon^2) = \delta_1(\sigma_\varepsilon^2)$. In this way, we obtain a formulation coherent with the limiting case $\mathbb{E}(\sigma_\varepsilon^2) \to 1$, $\text{Var}(\sigma_\varepsilon^2) \to 0$, which corresponds to an Inverse-Gamma distribution with parameters $A_\varepsilon \to \infty$, $B_\varepsilon \to \infty$.

Alternative prior distributions for the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{u}$ in (2.13) may be considered, including more involved hierarchical structures, sparse and robust distributions, as well as spatial and temporal dependence. In the same way, the specification of the prior laws for the scale parameters σ_ε^2 and σ_u^2 in (2.13) may be generalized to account for alternative distributions, weakly informative and non-informative priors. This choice is particularly delicate in mixed regression modelling, and it has been extensively discussed by Gelman (2006), which we refer to for further details. Just for a matter of exposition, we here consider only prior laws of the form (2.13), and we defer to Section 2.5 for a deeper discussion on model extensions and generalizations.

As proved by Bissiri *et al.* (2016), the most rational way to update our prior knowledge about $\boldsymbol{\theta}$ using the information coming from the empirical risk (2.10) is to combine prior and pseudo-likelihood through the Bayes theorem, as in Equation (2.11). The resulting generalized posterior density is then given by

$$\pi(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_u^2, \sigma_\varepsilon^2 | \mathbf{y}) \propto \pi(\mathbf{y} | \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2)\, \pi(\boldsymbol{\beta})\, \pi(\boldsymbol{u} | \sigma_u^2)\, \pi(\sigma_u^2)\, \pi(\sigma_\varepsilon^2). \tag{2.14}$$

Typically, such a posterior can not be normalized analytically, even when the pseudo-likelihood is conditionally conjugate with the prior distributions. For this reason, an efficient and general approximation algorithm is needed to perform posterior inference on (2.14) without changing the approximation scheme every time that a new ψ-function is considered.

2.3.2 Variational approximation

For approximating the conditional density function (2.14) we rely on a semiparametric variational approach; we thus impose the minimal product restriction

$$\pi(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_u^2, \sigma_\varepsilon^2 | \mathbf{y}) \approx q(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_u^2, \sigma_\varepsilon^2) = q(\boldsymbol{\beta}, \boldsymbol{u})\, q(\sigma_u^2, \sigma_\varepsilon^2),$$

which leads to the induced factorization $q(\sigma_u^2, \sigma_\varepsilon^2) = q(\sigma_u^2)\, q(\sigma_\varepsilon^2)$, because of the conditional independence structure of the posterior distribution. Furthermore, we restrict the density $q(\boldsymbol{\beta}, \boldsymbol{u}) \sim \text{N}_m(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ to be multivariate Gaussian with mean vector and covariance matrix

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \hat{\boldsymbol{\mu}}_\beta \\ \hat{\boldsymbol{\mu}}_u \end{bmatrix}, \qquad \hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{\beta\beta} & \hat{\boldsymbol{\Sigma}}_{\beta u} \\ \hat{\boldsymbol{\Sigma}}_{u\beta} & \hat{\boldsymbol{\Sigma}}_{uu} \end{bmatrix},$$

where the subscripts refer to the p- and d-dimensional sub–blocks corresponding to $\boldsymbol{\beta}$ and $\boldsymbol{u}$, so that $\hat{\boldsymbol{\mu}}_\beta \in \mathbb{R}^p$, $\hat{\boldsymbol{\Sigma}}_{\beta\beta} \in \mathbb{R}^{p \times p}$ and $\hat{\boldsymbol{\Sigma}}_{\beta u} \in \mathbb{R}^{p \times d}$. We denote with $m = p + d$ the total number of regression parameters in the model. Such a parametric assumption leads to a Gaussian variational distribution for the linear predictor, that is $q(\boldsymbol{\eta}) \sim \mathrm{N}_n(\hat{\boldsymbol{\eta}}, \hat{\mathbf{V}})$ with mean and variance

$$\mathbb{E}_q(\boldsymbol{\eta}) = \hat{\boldsymbol{\eta}} = \mathbf{C}\hat{\boldsymbol{\mu}}, \qquad \mathrm{Var}_q(\boldsymbol{\eta}) = \hat{\mathbf{V}} = \mathbf{C}\hat{\boldsymbol{\Sigma}}\mathbf{C}^\top.$$

Here, $\mathbf{C} = \big[\mathbf{X}, \mathbf{Z}\big]$ is the $n \times m$ completed design matrix having i-th row equal to $\mathbf{c}_i^\top = (\mathbf{x}_i^\top, \mathbf{z}_i^\top)$. Moreover, we denote by $\hat{\boldsymbol{\nu}}^2$ the vector containing the diagonal elements of $\hat{\mathbf{V}}$, that is $\hat{\nu}_i^2 = \hat{V}_{ii}$.

As shown in Appendix A.2, a direct application of the variational formulas (1.29) and (2.9) gives rise to the optimal variational updates summarized in the following optimal distributions:

- $q^{(k+1)}(\sigma_\varepsilon^2) \sim \mathrm{IG}(A_\varepsilon^{(k)}, B_\varepsilon^{(k)})$ where

$$A_\varepsilon^{(k)} \leftarrow A_\varepsilon + n/\alpha, \qquad B_\varepsilon^{(k)} \leftarrow B_\varepsilon + \mathbf{1}_n^\top \boldsymbol{\Psi}_0^{(k)}/\alpha;$$

- $q^{(k+1)}(\sigma_u^2) \sim \mathrm{IG}(A_u^{(k)}, B_u^{(k)})$ where

$$A_u^{(k)} \leftarrow A_u + d/2, \qquad B_u^{(k)} \leftarrow B_u + \tfrac{1}{2}\, \boldsymbol{\mu}_u^{(k)\top} \mathbf{R} \boldsymbol{\mu}_u^{(k)} + \tfrac{1}{2}\,\mathrm{trace}\big[\mathbf{R}\,\boldsymbol{\Sigma}_{uu}^{(k)}\big];$$

- $q^{(k+1)}(\boldsymbol{\beta}, \boldsymbol{u}) \sim \mathrm{N}_m(\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$ with $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Sigma}^{(k)}$ being the fixed-point of the Knowles-Minka-Wand recursion (2.9). Thanks to Proposition A.2 in Appendix A.1, the gradient vector $\mathbf{g}^{(k)}$ and the Hessian matrix $\mathbf{H}^{(k)}$ for our model specification take the form

$$\begin{aligned} \mathbf{g}^{(k)} &= -\bar{\mathbf{R}}^{(k)} \boldsymbol{\mu}^{(k)} - \mu_{1/\sigma_\varepsilon^2}^{(k)}\, \mathbf{C}^\top \boldsymbol{\Psi}_1^{(k)}/\alpha, \\ \mathbf{H}^{(k)} &= -\bar{\mathbf{R}}^{(k)} - \mu_{1/\sigma_\varepsilon^2}^{(k)}\, \mathbf{C}^\top \mathrm{diag}\big[\boldsymbol{\Psi}_2^{(k)}\big]\, \mathbf{C}/\alpha, \end{aligned} \tag{2.15}$$

  where $\bar{\mathbf{R}}^{(k)} = \mathrm{blockdiag}\big[\sigma_\beta^{-2}\mathbf{I}_p, \mu_{1/\sigma_u^2}^{(k)}\mathbf{R}\big]$, $\mu_{1/\sigma_u^2}^{(k)} = \mathbb{E}_q^{(k)}(1/\sigma_u^2)$ and $\mu_{1/\sigma_\varepsilon^2}^{(k)} = \mathbb{E}_q^{(k)}(1/\sigma_\varepsilon^2)$.

Here, the i-th element of the n-dimensional vector $\boldsymbol{\Psi}_r^{(k)} = (\Psi_{r,1}^{(k)}, \dots, \Psi_{r,n}^{(k)})^\top$ is defined as

$$\Psi_{r,i}^{(k)} = \Psi_r(y_i, \eta_i^{(k)}, \nu_i^{(k)}) = \mathbb{E}_q^{(k)}\big\{\psi_r(y_i, \eta_i)\big\}, \quad r = 0, 1, 2, \quad i = 1, \dots, n, \tag{2.16}$$

where $\psi_r(y, \eta)$ is the r-th order weak derivative of $\psi(y, \eta)$ with respect to η. The existence and regularity properties of Ψ_r are provided in the following theorem.

Theorem 2.4. *Let $\psi(y, \eta)$ be a function such that, for any $y \in \mathcal{Y}$ and $r = 1, \dots, R$, the r-th order weak derivative $\psi_r(y, \eta)$ is well-defined. Then, the following statements hold:*

(1) Ψ_r has infinitely many continuous derivatives with respect to η and ν;

(2) if ψ_r is continuous in η, then $\Psi_r(y, \eta, \nu) \to \psi_r(y, \eta)$ as $\nu \to 0$;

(3) if ψ_0 is convex in η, then Ψ_0 is jointly convex with respect to η and ν;

(4) if ψ_0 is convex in η, then $\psi_0(y, \eta) \leq \Psi_0(y, \eta, \nu)$ for any $\nu \geq 0$.

Proof. Theorem 2.4 follows as a consequence of Proposition A.1 in Appendix A.1. ☐

Remark 2.5. Theorem 2.4 does not require the loss function ψ to satisfy classical regularity assumptions, like differentiability, since the weak derivative ψ_r may exist even though ψ is not r-times differentiable all over its domain. This fact permits us to employ our variational approximation even for models with non-regular loss functions.

It is worth noting that, for any dimension of p and d, only univariate numerical integrations are required in the calculation of (2.15), since

$$\Psi_r(y, \eta, \nu) = \int_{-\infty}^{+\infty} \psi_r(y, x)\, \phi(\eta; \eta, \nu^2)\, \mathrm{d}x.$$

This fact leads to a scalable optimization routine, not depending on cumbersome high-dimensional integration problems. Clearly, efficiency and stability in the calculation of (2.15) highly depend on the algorithm used for evaluating the n expectations in (2.16). In our experience, whenever no analytic solutions are available, adaptive Gauss-Hermite quadrature (Liu and Pierce, 1994) leads to fast calculations and robust results.

Further, it is worth emphasizing the generality of the formulas in (2.15), that, together with (2.9), goes far beyond the few examples presented in the following section, for which we found simple analytic solutions. It indeed includes, as special cases, the algorithms proposed by e.g. Tan and Nott (2013), Wand (2014) and Luts and Ormerod (2014) for logistic, Poisson and Negative-Binomial regression models. Moreover, it is strongly connected with the Gaussian variational approximation proposed by Ormerod and Wand (2012) for estimating frequentist generalized mixed models.

The last ingredient of our variational method is the objective function, say the evidence lower bound on the marginal log-likelihood, which, at the $(k+1)$-th iteration of the algorithm, takes the closed form expression

$$\begin{aligned} \underline{\ell}(\mathbf{y}; q^{(k+1)}) = \\ & - \mu^{(k)}_{1/\sigma^2_\varepsilon}\, \mathbf{1}_n^\top \mathbf{\Psi}_0^{(k)}/\alpha + \tfrac{1}{2}\, \mathrm{logdet}(\mathbf{\Sigma}^{(k)}) - \tfrac{1}{2}\, \boldsymbol{\mu}^{(k)\top} \bar{\mathbf{R}}^{(k)}\, \boldsymbol{\mu}^{(k)} - \tfrac{1}{2}\, \mathrm{trace}\big[\bar{\mathbf{R}}^{(k)} \mathbf{\Sigma}^{(k)}\big] \\ & - \log\big\{\Gamma(A_u)/\Gamma(A_u^{(k)})\big\} + A_u \log(B_u/B_u^{(k)}) - \tfrac{d}{2} \log B_u^{(k)} - (B_u - B_u^{(k)})\, \mu^{(k)}_{1/\sigma^2_u} \\ & - \log\big\{\Gamma(A_\varepsilon)/\Gamma(A_\varepsilon^{(k)})\big\} + A_\varepsilon \log(B_\varepsilon/B_\varepsilon^{(k)}) - \tfrac{n}{\alpha} \log B_\varepsilon^{(k)} - (B_\varepsilon - B_\varepsilon^{(k)})\, \mu^{(k)}_{1/\sigma^2_\varepsilon} \\ & + \mathrm{const}, \end{aligned} \tag{2.17}$$

for $\bar{\mathbf{R}}^{(k)} = \mathrm{blockdiag}\big[\sigma_\beta^{-2}\mathbf{I}_p, \mu^{(k)}_{1/\sigma^2_u}\mathbf{R}\big]$. Here, “const” denotes a constant additive term not depending on the variational distributions. See Appendix A.2 for the outline of the derivation of the lower bound formula (2.17).

Remark 2.6. Let us assume that all the variational parameters in (2.17) but $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are fixed. Then, Proposition 2.4 guarantees the joint differentiability and concavity of $\underline{\ell}(\mathbf{y}; q)$ as a function of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$. Therefore, under the assumptions of Proposition A.1 the updating iteration (2.9) locally converges to a global maximum of the evidence lower bound (2.17). See, e.g., Rohde and Wand (2016) and Lange (2010), Chapter 12.

2.3.3 Optimization algorithm

The recursive refinement of the optimal distributions gives rise to the coordinate ascent variational inference routine summarized in Algorithm 4. We assess the algorithm convergence by looking at the relative change of the variational parameters and the lower bound (2.17). A well-behaved variational Bayes algorithm is expected to produce a non-decreasing sequence of lower bound values, thereby, providing a practical role for monitoring the convergence and detecting pathological behaviors of the algorithm. At the end of the estimation process, the evidence lower bound can also be used for model selection purposes, being a variational approximation of the true marginal log-likelihood.

Algorithm 4 SVB algorithm for variational inference in model (2.12) with prior (2.13)

```
Require: $\alpha, \mathbf{y}, \mathbf{X}$
Require: $\sigma_\beta^2, A_u, B_u, A_\varepsilon, B_\varepsilon$
  Initialize $\hat{A}_\varepsilon, \hat{B}_\varepsilon, \hat{A}_u, \hat{B}_u, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$;
  while convergence is not reached do
      Evaluate $\hat{\boldsymbol{\Psi}}_0$, $\hat{\boldsymbol{\Psi}}_1$, $\hat{\boldsymbol{\Psi}}_2$;    $O(nm^2)$
      $\hat{A}_\varepsilon \leftarrow A_\varepsilon + n/\alpha$;   $\hat{B}_\varepsilon \leftarrow B_\varepsilon + \mathbf{1}_n^\top \hat{\boldsymbol{\Psi}}_0/\alpha$;    $O(n)$
      $\hat{A}_u \leftarrow A_u + d/2$;   $\hat{B}_u \leftarrow B_u + \frac{1}{2}\{\hat{\boldsymbol{\mu}}_u^\top \mathbf{R}\, \hat{\boldsymbol{\mu}}_u + \mathrm{trace}[\mathbf{R}\, \hat{\boldsymbol{\Sigma}}_{uu}]\}$;    $O(d^2)$
      $\hat{\mu}_{1/\sigma_u^2} \leftarrow \hat{A}_u/\hat{B}_u$;   $\hat{\mu}_{1/\sigma_\varepsilon^2} \leftarrow \hat{A}_\varepsilon/\hat{B}_\varepsilon$;    $O(1)$
      $\bar{\mathbf{R}} \leftarrow \mathrm{blockdiag}[\sigma_\beta^{-2}\mathbf{I}_p, \hat{\mu}_{1/\sigma_u^2}\mathbf{R}]$;    $O(p + d^2)$
      $\hat{\mathbf{g}} \leftarrow -\bar{\mathbf{R}}\hat{\boldsymbol{\mu}} - \hat{\mu}_{1/\sigma_\varepsilon^2}\mathbf{C}^\top \hat{\boldsymbol{\Psi}}_1/\alpha$;    $O(nm)$
      $\hat{\mathbf{H}} \leftarrow -\bar{\mathbf{R}} - \hat{\mu}_{1/\sigma_\varepsilon^2}\mathbf{C}^\top \mathrm{diag}[\hat{\boldsymbol{\Psi}}_2]\mathbf{C}/\alpha$;    $O(nm^2)$
      $\rho \leftarrow \mathrm{LineSearch}(\hat{\mathbf{g}}, \hat{\mathbf{H}})$;   $\hat{\boldsymbol{\Sigma}} \leftarrow -\hat{\mathbf{H}}^{-1}$;   $\hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{\mu}} - \rho\, \hat{\mathbf{H}}^{-1}\hat{\mathbf{g}}$;    $O(m^3)$
  end while
```

In Algorithm 4, the notation LineSearch($\hat{\mathbf{g}}, \hat{\mathbf{H}}$) is used to denote a function taking as arguments the actual objective function $\underline{\ell}$, the gradient vector $\hat{\mathbf{g}}$ and Hessian matrix $\hat{\mathbf{H}}$ and returning the selected step-size parameter $\rho \in (0, 1]$ as an output. Such a routine can be easily implemented using an iterative backtracking method (Nocedal and Wright, 2006) until the step–length satisfies some minimal requirements, such as the Armijo-Wolfe conditions.

Assuming for $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{R}$ to be dense matrices, the number of flops required by one iteration of the algorithm is of order $O(nm^2 + m^3)$, which is equivalent to expectation-maximization, Gibbs sampling and mean field variational Bayes when applied to models of the form (2.12). However, if some sparsity patterns are observed, efficient sparse linear algebra routines may help in calculating $\hat{\mathbf{H}}^{-1}$ and $\hat{\mathbf{H}}^{-1}\hat{\mathbf{g}}$, turning down the complexity of the algorithm to a lower order. In such cases, the computational gain depends on the specific sparsity patterns and implementation details. Notice that we are here assuming that the calculation of $\boldsymbol{\Psi}_r$ is dominated by the evaluation of $\hat{\eta}_i = \mathbf{c}_i^\top \hat{\boldsymbol{\mu}}$ and $\hat{\nu}_i^2 = \mathbf{c}_i^\top \hat{\boldsymbol{\Sigma}}\, \mathbf{c}_i$, $i = 1, \dots, n$, that is of order $\mathcal{O}(nm^2)$. In our experience, this assumption is satisfied for all the statistical models we tried, even when no analytic solutions are available and numerical quadrature is required.

An important feature of the computational framework described above is that it is shared across all the linear models defined through minimum-risk criteria. What distinguishes different models is the specification of the Ψ-vectors, which bring the individual information about the local behavior of the expected loss function, similarly as the gradient and Hessian of the log-likelihood do in the classical penalized re-weighted iterated least squares algorithm. However, differently from standard gradient-based methods, we do not require the loss function to be differentiable, since the weak derivatives of ψ may exist even though the proper derivatives of ψ are not well-defined. Moreover, from Proposition 2.4 it follows that Ψ_0, Ψ_1, Ψ_2 are smooth functions of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$, since the convolution of any function with a Gaussian kernel produces a smooth transformation. More details and proofs can be found in Appendix A.1.

For a stable numerical implementation, we suggest truncating the values of $\boldsymbol{\Psi}_2$ from below to a small positive constant, say 10^{-6}, to prevent some elements of the vector to approach zero. In our numerical experiments, this correction never had a significant impact on the final solution, but often helped the convergence in the very early epochs of the iterative optimization routine.

2.4 Relevant models

As pointed out in Section 2.3, Algorithm 4 provides a general recipe for performing variational inference on a wide range of Bayesian models. In this section, we present some remarkable examples coming from both machine learning and statistical literature. In particular, we show how to specify the Ψ-vectors for different settings ranging from quantile and expectile regression (Sections 2.4.1 and 2.4.2) to support vector machines (Section 2.4.3 and 2.4.4). Finally, we move to the exponential family case (Section 2.4.5) for discussing the estimation of generalized linear models. In doing this, we make an extensive use of non-standard integration results summarized in Appendix A.1, which constitute the key results for calculating the necessary Ψ-functions in closed form.

2.4.1 Quantile regression

As introduced in Chapter 1, quantile regression (Koenker and Bassett, 1978; Koenker, 2005) is a statistical model which aims at estimating the τ-th conditional quantile of y_i given the available covariates $\mathbf{x}_i$ and $\mathbf{z}_i$. Here we consider the Bayesian mixed model generalization of classical quantile regression by specifying the linear predictor, i.e., the quantile regression function, as $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{u}$ where $\boldsymbol{\beta}$ is a vector of fixed effects and $\boldsymbol{u}$ is a vector of random effects.

Previously, Bayesian quantile regression has been considered by Yu and Moyeed (2001), which proposed a Metropolis-Hastings algorithm for Markov chain Monte Carlo inference; Kozumi and Kobayashi (2011) instead proposed a Gibbs sampling method (Algorithm 2) for efficient posterior sampling; while Wand *et al.* (2011) and McLean

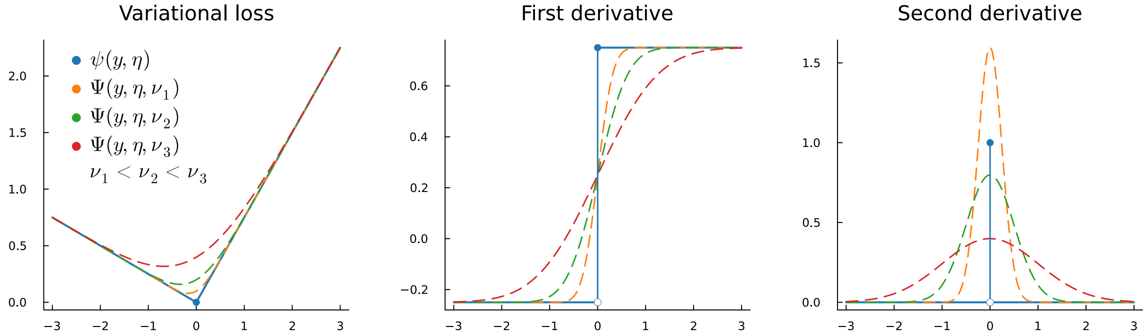

FIGURE 2.1: Quantile regression ($\tau = 0.75$): ψ_r and Ψ_r functions for different values of η and ν. From left to right: $r = 0, 1, 2$.

and Wand (2019) implemented a mean field variational Bayes procedure (Algorithm 3) for approximate inference on Asymmetric-Laplace likelihood models. Frequentist linear mixed quantile regression models have been considered by Geraci and Bottai (2014) and Geraci (2014), which proposed to evaluate the marginal pseudo-likelihood via adaptive Gaussian and Laplace quadrature.

Here, we take a different approach employing a semiparametric variational Bayes approximation and using Algorithm 4 to estimate quantile regression as a particular instance of the model defined in (2.12). Recalling the definition of the minimum risk criterion in (1.3), the ψ-function for the quantile regression model can be set as

$$\psi(y,\eta) = (y-\eta)\{\tau - \mathbb{I}(y \leq \eta)\}. \tag{2.18}$$

The specification of Algorithm 4 is then completed by providing the explicit solution for the Ψ-functions defined in (2.16).

Proposition 2.7. *The Ψ-functions for the quantile regression model are*

$$\begin{aligned}
\Psi_0(y,\eta,\nu) &= (y-\eta)\big\{\Phi(y;\eta,\nu^2) + \tau - 1\big\} + \nu^2\,\phi(y;\eta,\nu^2),\\
\Psi_1(y,\eta,\nu) &= 1 - \tau - \Phi(y;\eta,\nu^2),\\
\Psi_2(y,\eta,\nu) &= \phi(y;\eta,\nu^2).
\end{aligned}$$

Proof. Consider the quantile check loss $\psi(y,\eta) = \tilde{\psi}(y-\eta)$ along with its first and second weak derivatives:

$$\begin{aligned}
\tilde{\psi}_0(x) &= \tfrac{1}{2}|x| + (\tau - \tfrac{1}{2})x = x\{\tau - \mathbb{I}(x<0)\},\\
\tilde{\psi}_1(x) &= \tfrac{1}{2}\mathrm{sign}(x) + (\tau - \tfrac{1}{2}) = \tau - \mathbb{I}(x<0),\\
\tilde{\psi}_2(x) &= \delta_0(x).
\end{aligned}$$

Then, taking the expectation with respect to $x \sim \mathrm{N}(\mu, \nu^2)$, we get

$$\begin{aligned}
\mathbb{E}\{\tilde{\psi}_0(x)\} &= \tau\mathbb{E}(x) - \mathbb{E}\{x\,\mathbb{I}(x < 0)\} = \mu\{\tau - \Phi(0; \mu, \nu^2)\} + \nu^2\,\phi(0; \mu, \nu^2), \\
\mathbb{E}\{\tilde{\psi}_1(x)\} &= \tau - \mathbb{E}\{\mathbb{I}(x < 0)\} = \tau - \Phi(0; \mu, \nu^2), \\
\mathbb{E}\{\tilde{\psi}_2(x)\} &= \mathbb{E}\{\delta_0(x)\} = \phi(0; \mu, \nu^2).
\end{aligned}$$

Substituting $\mu = y - \eta$ and simplifying the results, we obtain the closed form expressions for Ψ_r, $r = 0, 1, 2$, provided in Proposition 2.7. □

As shown in Figure 2.1, the check function (2.18) is a piecewise linear loss weighting positive and negative errors differently. This characteristic is maintained by the variational loss introduced in Proposition 2.7, which has an asymmetric linear tail behavior, but an almost quadratic smooth trajectory in a neighborhood of the minimum. The variational convolution in (2.16) thus introduces a smoothing effect that regularizes the behavior of Ψ and its derivatives, even though the proper derivatives of ψ are not well-defined. Figure 2.1 thus provides an intuitive, graphical representation of the properties of Ψ stated in Theorem 2.4, such as smoothness, convexity, majorization of ψ and convergence to ψ as $\nu \to 0$.

2.4.2 Expectile regression

Expectile regression (Newey and Powell, 1987; Efron, 1991) is a statistical model alternative to Quantile Regression, which replaces the minimization of an asymmetrically weighted sum of absolute errors with an asymmetric squared error criterion. The Expectile Regression loss is then given by

$$\psi(y, \eta) = (y - \eta)^2 |\tau - \mathbb{I}(y \leq \eta)|. \tag{2.19}$$

The linear predictor $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{u}$ minimizing the expectile risk is called conditional τ-expectile of y_i given $\mathbf{x}_i$ and $\mathbf{z}_i$. Expectiles provide a generalization of the mean in the same way as quantiles provide an extension of the median, indeed, for $\tau = 0.5$, we obtain $\psi(y, \eta) = (y - \eta)^2$.

Expectiles arise to be particularly important in financial applications when it comes to measuring risks. Indeed, differently from the quantile check function (2.18), the expetile loss (2.19) leads to a coherent subadditive elicitable risk measure, as proved by Bellini and Bignozzi (2015) and Ziegel (2016).

Classical Bayesian inference on expectile regression models can be performed via Metropolis-Hastings simulation from the posterior distribution using as working likelihood an Asymmetric-Gaussian model. See, for example, the works of Sobotka and Kneib (2012), Xing and Qian (2017) and Waldmann *et al.* (2017). To the best of our knowledge, no conditionally conjugate stochastic representation of the Asymmetric-Gaussian model have been proposed in the literature. Therefore, expectation-maximization, Gibbs sampling, or mean field variational Bayes algorithms are not available as an alternative to Metropolis-Hastings .

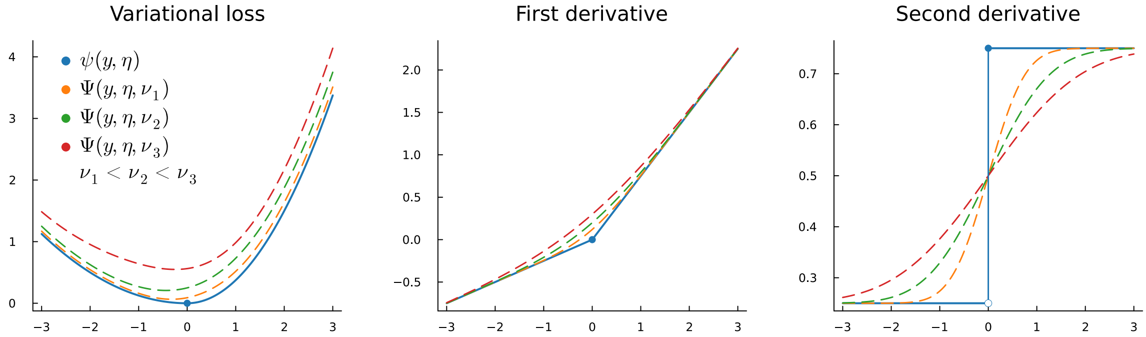

FIGURE 2.2: Expectile regression ($\tau = 0.75$): ψ_r and Ψ_r functions for different values of η and ν. From left to right: $r = 0, 1, 2$.

Like for the models presented so far, we can adopt Algorithm 4 in order to estimate the approximate posterior distribution of the conditional τ-expectile of y_i. To do so we consider the specification of the Ψ-functions (Figure 2.2) outlined in Proposition 2.8.

Proposition 2.8. *The Ψ-functions for the expectile regression model are*

$$\begin{aligned}
\Psi_0(y,\eta,\nu) &= \tfrac{1}{2}\big\{(y-\eta)^2 + \nu^2\big\}\big\{(1-\tau) - (1-2\tau)\,\Phi(y;\eta,\nu^2)\big\} \\
&\quad - \tfrac{1}{2}(1-2\tau)(y-\eta)\nu^2\,\phi(y;\eta,\nu^2), \\
\Psi_1(y,\eta,\nu) &= -(y-\eta)\big\{(1-\tau) - (1-2\tau)\,\Phi(y;\eta,\nu^2)\big\} \\
&\quad + (1-2\tau)\,\nu^2\,\phi(y;\eta,\nu^2), \\
\Psi_2(y,\eta,\nu) &= (1-\tau) - (1-2\tau)\,\Phi(y;\eta,\nu^2).
\end{aligned}$$

Proof. Consider the expectile loss $\psi(y,\eta) = \tilde{\psi}(y-\eta)$, along with its first and second weak derivatives:

$$\begin{aligned}
\tilde{\psi}_0(x) &= \tfrac{1}{2}x^2|\tau - \mathbb{I}(x<0)| = \tfrac{1}{2}x^2\big\{\tau + (1-2\tau)\mathbb{I}(x<0)\big\}, \\
\tilde{\psi}_1(x) &= x\,|\tau - \mathbb{I}(x<0)| = x\big\{\tau + (1-2\tau)\mathbb{I}(x<0)\big\}, \\
\tilde{\psi}_2(x) &= |\tau - \mathbb{I}(x<0)| = \tau + (1-2\tau)\,\mathbb{I}(x<0).
\end{aligned}$$

Here, we use the identity $|\tau - \mathbb{I}(x<0)| = \tau + (1-2\tau)\mathbb{I}(x<0)$, which permits to simplify the following calculations. Next, we marginalize out x by assuming the distribution $x \sim \mathrm{N}(\mu,\nu^2)$, leading to

$$\begin{aligned}
\mathbb{E}\big\{\tilde{\psi}_0(x)\big\} &= \tfrac{1}{2}\tau\mathbb{E}(x^2) + \tfrac{1}{2}(1-2\tau)\,\mathbb{E}\big\{x^2\,\mathbb{I}(x<0)\big\} \\
&= \tfrac{1}{2}\tau(\mu^2+\nu^2) + \tfrac{1}{2}(1-2\tau)\big\{(\mu^2+\nu^2)\Phi(0;\mu,\nu^2) - \mu\nu^2\phi(0;\mu,\nu^2)\big\} \\
&= \tfrac{1}{2}(\mu^2+\nu^2)\big\{\tau + (1-2\tau)\Phi(0;\mu,\nu^2)\big\} - \tfrac{1}{2}(1-2\tau)\mu\nu^2\phi(0;\mu,\nu^2).
\end{aligned}$$

In the same way, we get

$$\begin{aligned}\mathbb{E}\big\{\tilde{\psi}_1(x)\big\} &= \tau\mathbb{E}(x) + (1-2\tau)\,\mathbb{E}\big\{x\,\mathbb{I}(x<0)\big\} \\ &= \tau\mu + (1-2\tau)\big\{\mu\Phi(0;\mu,\nu^2) - \nu^2\phi(0;\mu,\nu^2)\big\} \\ &= \mu\big\{\tau + (1-2\tau)\Phi(0;\mu,\nu^2)\big\} - (1-2\tau)\nu^2\phi(0;\mu,\nu^2),\end{aligned}$$

and

$$\mathbb{E}\big\{\tilde{\psi}_2(x)\big\} = \tau + (1-2\tau)\,\mathbb{E}\big\{\mathbb{I}(x<0)\big\} = \tau + (1-2\tau)\,\Phi(0;\mu,\nu^2).$$

Substituting $\mu = y-\eta$ and simplifying the results, we obtain the closed form expressions for Ψ_r, $r = 0,1,2$, provided in Proposition 2.8. □

As shown in Figure 2.2, the variational loss averaging mapping ψ to Ψ helps to regularize the derivatives of the original loss (2.19), producing a new objective function having continuous derivatives up to the second order.

2.4.3 Support vector classification

Support vector classifiers Vapnik, 1998 are a family of statistical models which predict the binary response variable $y_i \in \mathcal{Y} = \{-1,+1\}$ using the decision function $\mathrm{sign}(\eta_i)$ with $\eta_i = \mathbf{x}_i^\top\boldsymbol{\beta} + \mathbf{z}_i^\top\boldsymbol{u}$. The support vector estimation problem attempt to find the optimal decision function in the covariate space by pushing further apart observations with different labeling. In doing so, it maximizes the *margin*, namely the distance in the covariate space, between the sets $\{i : y_i = -1\}$ and $\{i : y_i = +1\}$.

In terms of the minimum risk formalization, the loss function characterizing the support vector estimation problem is the so-called Hinge loss, that is defined as

$$\psi(y,\eta) = 2\max(0, 1-y\eta). \tag{2.20}$$

Regularized non-linear support vector machines may be included in our model specification by representing the random effect covariate vector $\mathbf{z}_i$ as a sequence of basis functions. In this way, we can jointly learn the posterior distribution of the basis coefficients and the smoothing parameter, $1/\sigma_u^2$, controlling the amount of regularization needed to prevent overfitting phenomena.

A complete Bayesian treatment of the support vector estimation problem has been discussed by Polson and Scott (2011), which introduced a popular data-augmentation technique to represent the Hinge pseudo-likelihood as a location-scale mixture of Gaussian distributions, similarly to the quantile regression representation (1.8). The authors then proposed a Gibbs sampler and an expectation-maximization algorithm to estimate the model parameters. Luts and Ormerod (2014) considered a similar approach and leveraged the conditional conjugacy of the augmented model to approximate the posterior distribution with a mean field variational Bayes approach.

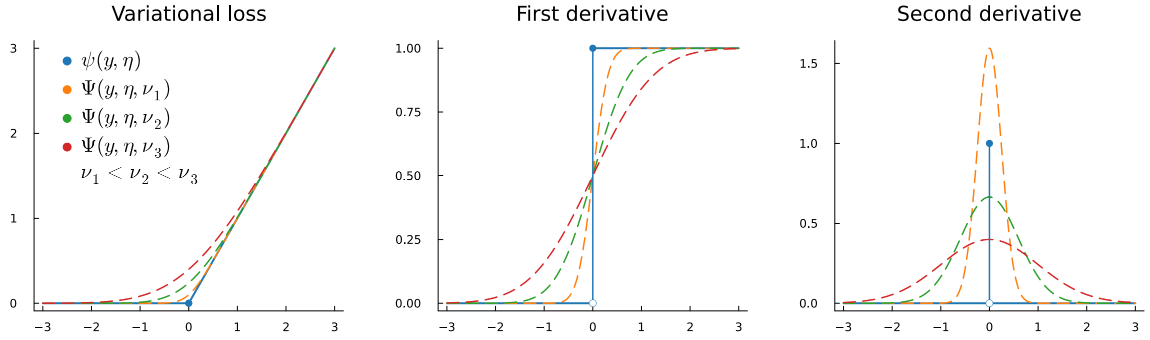

FIGURE 2.3: Support vector classification: ψ_r and Ψ_r functions for different values of η and ν. From left to right: $r = 0, 1, 2$.

Unlike existing methods, the major advantage of the semiparametric variational Bayes approach proposed in Section 2.3 consists of avoiding to transform the pseudo-likelihood and to introduce additional auxiliary variables in the model specification. Instead, we use an approximation scheme which does not transform the loss function (2.20) and the associated posterior distribution. Then, all the parameters in the model can be estimated using Algorithm 4 and specifying the Ψ-vectors according to the following proposition.

Proposition 2.9. *The Ψ-functions for the support vector classification model are*

$$\begin{aligned}
\Psi_0(y,\eta,\nu) &= 2\big\{(1-y\eta)\,\Phi(1;y\eta,\nu^2) + \nu^2\,\phi(1;y\eta,\nu^2)\big\},\\
\Psi_1(y,\eta,\nu) &= -2y\,\Phi(1;y\eta,\nu^2),\\
\Psi_2(y,\eta,\nu) &= 2\,\phi(1;y\eta,\nu^2).
\end{aligned}$$

Proof. By the definition of hinge loss function, we have that $\psi(y,\eta) = \tilde\psi(1-y\eta)$, where the first three weak derivatives of $\tilde\psi(\cdot)$ are

$$\begin{aligned}
\tilde\psi_0(x) &= |x| + x = 2\{x - x\,\mathbb{I}(x<0)\},\\
\tilde\psi_1(x) &= \operatorname{sign}(x) + 1 = 2\{1 - \mathbb{I}(x<0)\},\\
\tilde\psi_2(x) &= 2\,\delta_0(x).
\end{aligned}$$

Here, we use the identities $2\max(0,x) = |x| + x$, $|x| = x - 2x\mathbb{I}(x<0)$, $\operatorname{sign}(x) = 1 - 2\,\mathbb{I}(x<0)$. Let $x \sim \mathrm{N}(\mu,\nu^2)$, then, applying Propositions A.3 and A.4, we have

$$\begin{aligned}
\mathbb{E}\{\tilde\psi_0(x)\} &= 2\,\mathbb{E}(x) - 2\,\mathbb{E}\{x\,\mathbb{I}(x<0)\} = 2\mu\{1-\Phi(0;\mu,\nu^2)\} + 2\nu^2\phi(0;\mu,\nu^2),\\
\mathbb{E}\{\tilde\psi_1(x)\} &= 2 - 2\,\mathbb{E}\{\mathbb{I}(x<0)\} = 2\{1-\Phi(0;\mu,\nu^2)\},\\
\mathbb{E}\{\tilde\psi_2(x)\} &= \mathbb{E}\{\delta_0(x)\} = 2\phi(0;\mu,\nu^2).
\end{aligned}$$

Now, defining $\mu = 1 - y\eta$, we have

$$\Phi(0;\mu,\nu) = 1 - \Phi(1;y\eta,\nu^2), \quad \phi(0;\mu,\nu) = \phi(1;y\eta,\nu^2),$$

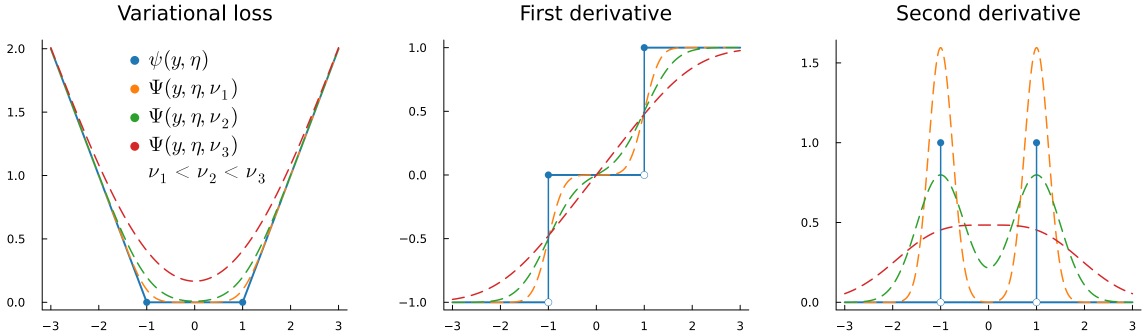

FIGURE 2.4: Support vector regression ($\epsilon = 1$): ψ_r and Ψ_r functions for different values of η and ν. From left to right: $r = 0, 1, 2$.

Simplifying the results, we obtain the closed form expressions for Ψ_r, $r = 0, 1, 2$, provided in Proposition 2.9. This concludes the proof. □

Figure 2.3 shows the behavior of both ψ and Ψ along with their first and second weak derivatives for different values of the variational parameters.

2.4.4 Support vector regression

Support vector regression (Vapnik, 1998) is a robust prediction model which extends the maximum margin approach to regression problem. This finds the best linear predictor $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{u}$ to fit the data $y_i \in \mathcal{Y} = \mathbb{R}$ by estimating the parameters $\boldsymbol{\beta}$ and $\boldsymbol{u}$ in such a way to minimize the ϵ-insensitive loss (2.21), which, in our model specification, corresponds to the ψ-function

$$\psi^{\mathrm{SVR}}(y, \eta) = 2 \max(0, |y - \eta| - \epsilon). \tag{2.21}$$

Either frequentist and Bayesian estimation procedures have been developed for Support Vector Regression models. Among others, Zhu *et al.* (2012) and Zhu *et al.* (2014) adopted a double data-augmentation strategy to develop an expectation-maximization algorithm for penalized pseudo-likelihood maximization and a Gibbs sampling algorithm for posterior simulation. Mean field variational Bayes may be used as well for posterior approximation in the augmented parameter space, even though we are not aware of any work proposing such a method in the literature. In this context, our proposal is to directly approximate the marginal posterior distribution (2.14) using Algorithm 4 and defining the Ψ-functions (Figure 2.4) according to the following proposition.

Proposition 2.10. *The Ψ-functions for the support vector regression model are*

$$\begin{aligned}
\Psi_0(y, \eta, \nu) &= 2\big\{(y_\epsilon^- - \eta)\, \Phi(y_\epsilon^+; \eta, \nu^2) + \nu^2\, \phi(y_\epsilon^+; \eta, \nu^2) \\
&\qquad + (y_\epsilon^+ - \eta)\, \Phi(y_\epsilon^-; \eta, \nu^2) + \nu^2\, \phi(y_\epsilon^-; \eta, \nu^2)\big\} \\
\Psi_1(y, \eta, \nu) &= 2\{1 - \Phi(y_\epsilon^+; \eta, \nu^2) - \Phi(y_\epsilon^-; \eta, \nu^2)\}, \\
\Psi_2(y, \eta, \nu) &= 2\{\phi(y_\epsilon^+; \eta, \nu^2) + \phi(y_\epsilon^-; \eta, \nu^2)\},
\end{aligned}$$

where $y_\epsilon^+ = y + \epsilon$ *and* $y_\epsilon^- = y - \epsilon$, *for* $\epsilon \geq 0$.

Proof. By the definition of ϵ-insensitive loss function, we have that $\psi(y, \eta) = \tilde{\psi}(y - \eta)$, where the first three weak derivatives of $\tilde{\psi}(\cdot)$ are

$$\begin{aligned}\tilde{\psi}_0(x) &= |x - \epsilon| + (x - \epsilon) + |x + \epsilon| - (x + \epsilon),\\ \tilde{\psi}_1(x) &= \mathrm{sign}(x - \epsilon) + \mathrm{sign}(x + \epsilon),\\ \tilde{\psi}_2(x) &= 2\delta_0(x - \epsilon) + 2\delta_0(x + \epsilon).\end{aligned}$$

Let $x \sim \mathrm{N}(\mu, \nu^2)$ and use, again, Proposition A.4 for calculating

$$\begin{aligned}\mathbb{E}\,|x - \epsilon| + \mathbb{E}(x - \epsilon) &= 2(\mu - \epsilon)\{1 - \Phi(-\epsilon; \mu, \nu^2)\} + 2\sigma^2\phi(-\epsilon; \mu, \nu^2),\\ \mathbb{E}\,|x + \epsilon| - \mathbb{E}(x + \epsilon) &= 2(\mu + \epsilon)\{1 - \Phi(+\epsilon; \mu, \nu^2)\} + 2\sigma^2\phi(+\epsilon; \mu, \nu^2).\end{aligned}$$

These lead to the expected value

$$\begin{aligned}\mathbb{E}\big\{\tilde{\psi}_0^{\mathrm{SVR}}(x)\big\} = 2\big[&(\mu - \epsilon)\{1 - \Phi(-\epsilon; \mu, \nu^2)\} + \nu^2\phi(-\epsilon; \mu, \nu^2)\\ &+ (\mu + \epsilon)\{1 - \Phi(+\epsilon; \mu, \nu^2)\} + \nu^2\phi(+\epsilon; \mu, \nu^2)\big].\end{aligned}$$

Similarly, using $\mathrm{sign}(x) = 1 - 2\,\mathbb{I}(x < 0)$, we have

$$\begin{aligned}\mathbb{E}\big\{\tilde{\psi}_1(x)\big\} &= 2 - 2\,\mathbb{E}\{\mathbb{I}(x < \epsilon)\} - 2\,\mathbb{E}\{\mathbb{I}(x < -\epsilon)\}\\ &= 2 - 2\,\Phi(+\epsilon; \mu, \nu^2) - 2\,\Phi(-\epsilon; \mu, \nu^2),\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\big\{\tilde{\psi}_2(x)\big\} &= 2\,\mathbb{E}\{\delta_{+\epsilon}(x)\} + 2\,\mathbb{E}\{\delta_{-\epsilon}(x)\}\\ &= 2\,\phi(+\epsilon; \mu, \nu^2) + 2\,\phi(-\epsilon; \mu, \nu^2).\end{aligned}$$

Substituting $\mu = y - \eta$ and simplifying the results, we obtain the closed form expressions for Ψ_r, $r = 0, 1, 2$, provided in Proposition 2.10. □

Figure 2.4 shows the behavior of both ψ and Ψ along with their first and second weak derivatives for different values of the variational parameters.

2.4.5 Exponential family

The exponential family is a wide class of distributions that includes, among others, the Gaussian, Gamma, Binomial and Poisson probability laws. It constitutes the theoretical foundation of generalized linear models (McCullagh and Nelder, 1989) and is characterized by a probability density function of the form

$$\pi(y_i|\xi_i) = H(y_i)\exp\big\{y_i\,\xi_i - A(\xi_i)\big\}, \qquad i = 1, \dots, n,$$

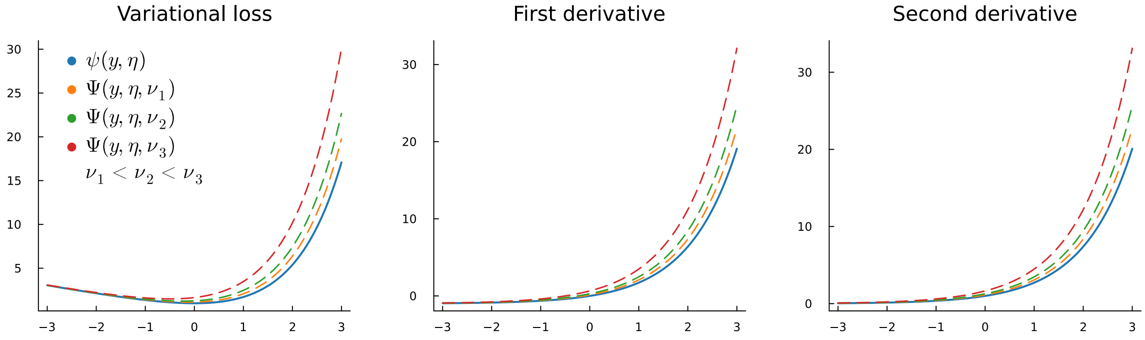

FIGURE 2.5: Poisson regression: ψ_r and Ψ_r functions for different values of η and ν. From left to right: $r = 0, 1, 2$.

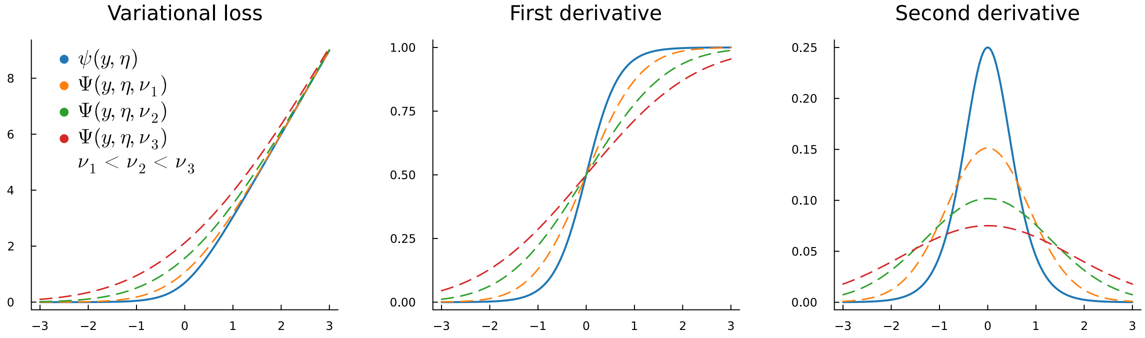

FIGURE 2.6: Logistic regression: ψ_r and Ψ_r functions for different values of η and ν. From left to right: $r = 0, 1, 2$.

where ξ_i is the so-called *canonical* parameter, $A(\cdot)$ and $H(\cdot)$ are, respectively, the *log-partition* function and the *base measure* specific to the members of the family. The canonical parameter ξ_i is then linked with the linear predictor through the equation $g^{-1}(\eta_i) = A'(\xi_i)$, where $g(\cdot)$ is a bijective link function and $A'(\cdot)$ is the first order derivative of the convex, two times differentiable map $A(\cdot)$.

Assuming for simplicity that a canonical link function is considered, i.e. $g^{-1}(\cdot) = A'(\cdot)$ and $\xi_i = \eta_i$, the ψ-loss associated to the exponential family log-likelihood takes the form

$$\psi(y, \eta) = -y\eta + A(\eta).$$

The variational expectations Ψ_r^{EF}, $r = 0, 1, 2$, are thus given by

$$\begin{aligned}
\Psi_0(y, \hat{\eta}, \hat{\nu}) &= -y\,\hat{\eta} + \mathbb{E}_q\big\{A(\eta)\big\}, \\
\Psi_1(y, \hat{\eta}, \hat{\nu}) &= -y + \mathbb{E}_q\big\{A'(\eta)\big\}, \\
\Psi_2(y, \hat{\eta}, \hat{\nu}) &= \mathbb{E}_q\big\{A''(\eta)\big\}.
\end{aligned}$$

Depending on the shape of $A(\cdot)$, the integrals $\mathbb{E}_q\{A(\eta)\}$, $\mathbb{E}_q\{A'(\eta)\}$ and $\mathbb{E}_q\{A''(\eta)\}$ may be computed analytically or approximated via univariate quadrature. For instance, in the Poisson regression model, where $A(\eta) = \exp(\eta)$, the explicit solutions

$$\mathbb{E}_q\{A(\eta)\} = \mathbb{E}_q\{A'(\eta)\} = \mathbb{E}_q\{A''(\eta)\} = \exp(\hat{\eta} + \hat{\nu}^2/2),$$

are available. Differently, in Binomial regression with logistic link function, where $A(\eta) = \log(1 + e^\eta)$, we may employ an adaptive Gauss-Hermite quadrature for an efficient and stable calculation, as discussed in the supplementary material of Ormerod and Wand (2012).

There exists a huge literature on both frequentist and Bayesian generalized linear mixed models. From a variational perspective, the most relevant contributions related to our work are: Ormerod and Wand (2012) (frequentist variational approximations for Poisson and Bernoulli mixed models), Tan and Nott (2013) (variational message passing for generalized linear mixed models), Wand (2014) (simplified multivariate normal updated for non-conjugate variational message passing), Luts and Wand (2015) (semiparametric variational Bayes for count response data). Further, we suggest the books of Ruppert *et al.* (2003), Gelman and Hill (2006), Wood (2017) for a comprehensive treatment of the theory of generalized linear mixed models and their semiparametric additive extensions.

2.5 Possible extensions

The semiparametric variational Bayes methodology discussed so far can be easily extended to accommodate more structured model specifications, including, e.g., additive models, inducing shrinkage priors, dynamic linear models and spatial random fields. In this Section we present a brief discussion about each one of these generalizations, showing how to modify and use Algorithm 4 for the estimation of a wide class of flexible models.

The present exposition is inspired by Hodges (2014), which we refer to for an exhaustive treatment of the models introduced in the following.

2.5.1 Additive models

Additive models (Wood, 2017) are a straightforward generalization of model (2.12), which permits dealing with multiple heterogeneous random effects $\boldsymbol{u}_1, \dots, \boldsymbol{u}_H$ within the same linear predictor. This way, for each observation we have

$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \sum_{h=1}^{H} \mathbf{z}_{i,h}^\top \boldsymbol{u}_h, \quad \boldsymbol{u}_h | \sigma_h^2 \sim \mathrm{N}_{d_h}(\mathbf{0}_{d_h}, \sigma_h^2 \mathbf{R}_h^{-1}), \quad \sigma_h^2 \sim \mathrm{IG}(A_h, B_h), \tag{2.22}$$

with $\mathbf{z}_{i,h} \in \mathbb{R}^{d_h}$, $\boldsymbol{u}_h \in \mathbb{R}^{d_h}$, $h = 1, \dots, H$, $i = 1, \dots, n$. The total number of random effects is $d = d_1 + \dots + d_H$.

The model specified in (2.22) can be represented in a more compact way as $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{u}$ where $\mathbf{z}_i = (\mathbf{z}_{i,1}^\top, \dots, \mathbf{z}_{i,H}^\top)^\top \in \mathbb{R}^d$ and $\boldsymbol{u} = (\boldsymbol{u}_1^\top, \dots, \boldsymbol{u}_H^\top)^\top \in \mathbb{R}^d$. Then, we may use Algorithm 4 for performing parameter estimation. The only difference we have to account for is concerned with the presence of multiple variance parameters $\sigma_1^2, \dots, \sigma_H^2$. However, under the mean field factorization

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_1^2, \dots, \sigma_H^2, \sigma_\varepsilon^2) = q(\boldsymbol{\beta}, \boldsymbol{u})\, q(\sigma_1^2) \times \cdots \times\, q(\sigma_H^2)\, q(\sigma_\varepsilon^2),$$

we are able to explicitly calculate the optimal variational approximation at the $(k+1)$-th iteration of the algorithm, say $q^{(k+1)}(\sigma_h^2)$, $h = 1, \dots, H$, that is an Inverse-Gamma density with variational parameters

$$A_h^{(k)} \leftarrow A_h + d_h/2, \qquad B_h^{(k)} \leftarrow B_h + \tfrac{1}{2}\boldsymbol{\mu}_h^{(k)\top}\mathbf{R}_h\, \boldsymbol{\mu}_h^{(k)} + \tfrac{1}{2}\mathrm{trace}\big[\mathbf{R}_h \boldsymbol{\Sigma}_{hh}^{(k)}\big]$$

Such a modified algorithm still maintains the same computational complexity of Algorithm 4 and all the updates are available in closed form.

2.5.2 Inducing shrinkage priors

The automatic identification of non-relevant covariates is an omnipresent issue in statistics. From a Bayesian point of view, this challenge has been faced with a number of different approaches, which are all concerned with the specification of a convenient prior distribution. In principle, we wish to define a prior law able to operate an aggressive shrinkage toward zero for the non-relevant effects, while not introducing much *bias* on the estimates of the relevant coefficients.

According to the recent literature on continuous shrinkage prior, we propose to extend model (2.13) according to the following Gaussian scale-mixture specification:

$$\boldsymbol{u}|\boldsymbol{\lambda}^2, \delta^2 \sim \mathrm{N}_d(\mathbf{0}_d, \delta^2\boldsymbol{\Lambda}^2), \quad \boldsymbol{\Lambda}^2 = \mathrm{diag}(\boldsymbol{\lambda}^2), \quad \boldsymbol{\lambda} \sim \pi(\boldsymbol{\lambda}), \quad \delta \sim \pi(\delta),$$

where $\boldsymbol{\lambda}$ is a vector of local scale parameters and δ is a global scale parameter. Depending on the particular specification of $\pi(\boldsymbol{\lambda})$ and $\pi(\delta)$ different models are obtained and, as a consequence, also different variational approximations arise. See, among others, the Bayesian Lasso (Park and Casella, 2008), the Horseshoe (Carvalho *et al.*, 2010), the Normal-Exponential-Gamma (Griffin and Brown, 2011), the generalized double-Pareto (Armagan *et al.*, 2013), the adaptive Bayesian Lasso (Leng *et al.*, 2014), and the Dirichlet-Laplace (Bhattacharya *et al.*, 2015) models.

Assuming for the approximate posterior distribution the mean field factorization

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\lambda}^2, \delta^2, \sigma_\varepsilon^2) = q(\boldsymbol{\beta}, \boldsymbol{u})\, q(\boldsymbol{\lambda}^2)\, q(\delta^2)\, q(\sigma_\varepsilon^2),$$

the optimal densities $q^{(k+1)}(\boldsymbol{\beta}, \boldsymbol{u})$ and $q^{(k+1)}(\sigma_\varepsilon^2)$ maintain the same structure described in Section 2.3, where $\mu_{1/\sigma_u^2}^{(k)}\mathbf{R}$ in Equation (2.15) is replaced by $\mathrm{diag}\big[\mu_{1/\delta^2}^{(k)}\, \boldsymbol{\mu}_{1/\lambda^2}^{(k)}\big]$. However, using such a factorization, the approximating densities $q^{(k+1)}(\boldsymbol{\lambda}^2)$ and $q^{(k+1)}(\delta^2)$ might not enjoy analytic solutions, even in the case of conjugate prior distributions for $\pi(\boldsymbol{u}|\boldsymbol{\lambda}, \delta)$, $\pi(\boldsymbol{\lambda})$ and $\pi(\delta)$. Though, reliable and stable approximations may be obtained by employing clever integration strategies, as shown by, e.g., Neville *et al.* (2014) in the Horseshoe, Normal-Exponential-Gamma and generalized double Pareto models.

2.5.3 Dynamic linear models

Dynamic linear models (Triantafyllopoulos, 2021) generalize the static specification (2.12) by introducing a latent transition equation that determines the evolution of the random effects over time, that is

$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{u}_i, \quad \boldsymbol{u}_{i+1} = \mathbf{T}\boldsymbol{u}_i + \boldsymbol{v}_i, \quad \boldsymbol{v}_i | \sigma_u^2 \sim \mathrm{N}_d(\mathbf{0}_d, \sigma_u^2 \mathbf{R}^{-1}),$$

with $i = 1, \dots, n$ denoting the time index. We complete the model by specifying the initial distribution $\boldsymbol{u}_0 \sim \mathrm{N}_d(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Here, $\mathbf{T}$ is a $d \times d$ transition matrix, $\mathbf{R}$ is a $d \times d$ precision matrix, while $\boldsymbol{v}_i$ is a $d \times 1$ vector of latent innovations.

Stacking all the random effects by column, i.e. $\boldsymbol{u} = (\boldsymbol{u}_1^\top, \dots, \boldsymbol{u}_i^\top, \dots, \boldsymbol{u}_n^\top)^\top \in \mathbb{R}^{nd}$, we obtain a prior law for $\boldsymbol{u}$ that is a rank-deficient Gaussian distribution with block-tridiagonal precision matrix. Moreover, we can express the linear predictor as $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + (\boldsymbol{e}_i \otimes \mathbf{z}_i)^\top \boldsymbol{u}$, where $\boldsymbol{e}_i$ is the i-th column of an n-dimensional identity matrix and $\otimes$ denotes the Kronecker product. This joint representation (Chan and Jeliazkov, 2009) allows us to approximate the posterior distribution by employing Algorithm 4. Kalman filter and smoother routines (Durbin and Koopman, 2012) can then be used for an efficient numerical evaluation of $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Sigma}^{(k)}$.

Since in the dynamic specification the dimension of the latent states grows together with the number of observed data, also the asymptotic computational complexity of the algorithm increases. Assuming for simplicity that no fixed covariates are included, i.e. $m = d$, the complexity of the algorithm is dominated by Kalman filter and smoother, which require $O(nm^3)$ flops each.

2.5.4 Spatial random fields

Latent Gaussian random fields (Rue *et al.*, 2009) are the standard tool in Bayesian hierarchical modelling for dealing with spatially correlated data. Let us assume that the observations have been gathered on a set of n spatial locations $\mathbf{p}_1, \dots, \mathbf{p}_n$, which lie in the spatial domain Γ. Then, we can account for spatial dependence in the data by specifying the following mixed model for the linear predictor:

$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + u(\mathbf{p}_i), \quad u(\cdot) \sim \text{Gaussian random field over } \Gamma.$$

Depending on the shape of the domain Γ and on the assumptions made upon $u(\cdot)$, different models arise.

For instance, if Γ is a finite discrete spatial domain, like an areal map or a finite network, common specifications for $\boldsymbol{u} = (u(\mathbf{p}_1), \dots, u(\mathbf{p}_n))$ are simultaneous and conditional autoregressive models (Cressie, 2015), or Gaussian Markov random fields (Rue and Held, 2005). All of these lead to a multivariate Gaussian prior distribution $\boldsymbol{u} | \sigma_u^2 \sim$

$\mathrm{N}_n(\mathbf{0}_n, \sigma_u^2 \mathbf{R}^{-1})$, where the form and the sparsity pattern of the precision matrix $\mathbf{R}$ are determined by the selected model.

Gaussian random fields over continuous domains can be dealt as well by approximating $u(\cdot)$ with a penalized basis expansion: $u(\cdot) \approx \boldsymbol{u}^\top \mathbf{z}(\cdot)$. Here, $\mathbf{z}(\cdot) = (z_1(\cdot), \dots, z_d(\cdot))^\top$ is a vector of locally supported basis functions defined over Γ, while $\boldsymbol{u}|\sigma_u^2 \sim \mathrm{N}_d(\mathbf{0}_d, \sigma_u^2 \mathbf{R}^{-1})$ is a Gaussian Markov random field with sparse precision $\mathbf{R}$. In the literature this representation is used, e.g., for discretizing a wide class of Wittle-Matérn fields implicitly defined as the solution to a fractional stochastic partial differential equation (Lindgren *et al.*, 2011, 2022).

Both the discrete and the continuous cases fit in formulation (2.13) and, hence, they can be efficiently estimated using Algorithm 4. Doing this, a careful management of sparse linear algebra routines is necessary for avoiding the explicit inversion of high-dimensional sparse matrices.

2.5.5 Frequentist mixed models

In the Bayesian literature, variational approximations have been successfully employed in a wide range of applications and for a broad class of models. Their accuracy, efficiency and broad applicability made them very popular in the Bayesian community; however, their usage in frequentist statistics is still limited, and their finite- and large-sample properties are mostly unexplored.

An interesting extension of the variational approach proposed in Section 2.3 is to consider the estimation of mixed effect models within a frequentist inferential framework. Consider, for instance, the mixed model

$$\mathbf{y}|\boldsymbol{u};\boldsymbol{\theta} \sim \pi(\mathbf{y}|\boldsymbol{u};\boldsymbol{\theta}), \qquad \boldsymbol{u}|\boldsymbol{\theta} \sim \mathrm{N}_d(\mathbf{0}_d, \sigma_u^2 \mathbf{R}^{-1}),$$

where $\pi(\mathbf{y}|\boldsymbol{u};\boldsymbol{\theta}) = \exp\{-nR_n(\boldsymbol{u},\boldsymbol{\theta})/\alpha\}$ is the pseudo-likelihood function defined as in (2.12), $\boldsymbol{u}$ is a vector of random effects, and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_\varepsilon^2, \sigma_u^2)$ is a vector of fixed parameters. Then, the maximum pseudo-likelihood estimator of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$, is defined as the maximizer of the marginal log-likelihood

$$\ell(\boldsymbol{\theta};\mathbf{y}) = \log \int_{\mathbb{R}^d} \pi(\mathbf{y},\boldsymbol{u}|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{u} = \log \int_{\mathbb{R}^d} \pi(\mathbf{y}|\boldsymbol{u};\boldsymbol{\theta})\,\pi(\boldsymbol{u}|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{u}. \tag{2.23}$$

Except for trivial cases, (2.23) is not available in closed form, as well as its maximizier, which must be computed via iterative optimization and numerical integration. Indeed, the evaluation of $\ell(\boldsymbol{\theta};\mathbf{y})$ inherits all the difficulties entailed by the integration of a posterior distribution in a Bayesian setting. Furthermore, classical approaches, such as restricted maximum likelihood (Patterson and Thompson, 1971), penalized quasi-likelihood (Breslow and Clayton, 1993), or Laplace approximation (Wolfinger, 1994), can not be applied whenever the risk function $R_n(\cdot)$ does not satisfy second order differentiability conditions. A remarkable example is given by the work of Geraci and Bottai (2014) and Geraci (2014), which proposed to estimate frequentist linear quantile mixed models by using a careful combination of multivariate quadrature and non-smooth optimization.

However, by the Jensen inequality, for any probability density function $q(\boldsymbol{u}) \in \mathcal{Q}$, we have

$$\ell(\boldsymbol{\theta}; \mathbf{y}) \geq \underline{\ell}(\boldsymbol{\theta}; \mathbf{y}, q) = \int_{\mathbb{R}^d} q(\boldsymbol{u}) \log \frac{\pi(\mathbf{y}, \boldsymbol{u} | \boldsymbol{\theta})}{q(\boldsymbol{u})} \, \mathrm{d}\boldsymbol{u};$$

thus, an alternative estimator of $\boldsymbol{\theta}$ may be defined following a variational approach and, in particular, adopting a Gaussian variational approximation (Ormerod and Wand, 2012). Doing so, we consider the parametric density transformation $q(\boldsymbol{u}) \equiv q(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and we define the corresponding variational estimator as

$$\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}} = \underset{\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}}{\operatorname{argmax}} \, \underline{\ell}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{y}),$$

where $\underline{\ell}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{y}) \equiv \underline{\ell}\{\boldsymbol{\theta}; \mathbf{y}, q(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\}$.

Under the same hypothesis of Theorem 2.4, the lower bound $\underline{\ell}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{y})$ is a smooth concave function with respect to $\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$. Thus, in this new formulation, the estimator of $\boldsymbol{\theta}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ may be obtained by standard numerical optimization algorithms or, alternatively, by using a fixed-point algorithm, like the one we proposed in Section 2.3.3. In this context, an opportune modification of Algorithm 4 may be interpreted as a generalized variational implementation of the EM algorithm, which cycles over the following steps

$$\begin{aligned}
&\text{V-step:} \qquad && \text{update } \boldsymbol{\mu}^{(k)} \text{ and } \boldsymbol{\Sigma}^{(k)}; \\
&\text{E-step:} \qquad && \underline{\ell}^{(k)}(\boldsymbol{\theta}; \mathbf{y}) \leftarrow \mathbb{E}_q^{(k)}\{\log \pi(\mathbf{y}, \boldsymbol{u} | \boldsymbol{\theta})\}; \\
&\text{M-step:} \qquad && \text{update } \boldsymbol{\theta}^{(k+1)} \text{ so that } \underline{\ell}^{(k)}(\boldsymbol{\theta}^{(k)}; \mathbf{y}) \leq \underline{\ell}^{(k)}(\boldsymbol{\theta}^{(k+1)}; \mathbf{y});
\end{aligned}$$

where $\mathbb{E}_q^{(k)}(\cdot)$ denotes the expectation calculated with respect to the density $q^{(k)}(\boldsymbol{u}) = q(\boldsymbol{u}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$.

Such an approach has been proposed by Ormerod and Wand (2012) for the estimation of binary and Poisson linear mixed models, while Hall *et al.* (2011a,b) studied the asymptotic properties of the resulting estimator for a simple random intercept Poisson regression model. Further theoretical developments in the asymptotic analysis of variational tempered posterior distributions and variational misspecified models have been studied by Alquier *et al.* (2016), Alquier and Ridgway (2020), Wang and Blei (2019b), Wang and Blei (2019c).

2.6 Simulation studies

In the following numerical examples we only consider synthetic datasets in order to reliably assess the empirical qualities of the semiparametric variational Bayes (SVB) approach discussed in Section 2.3. The models we consider for this analysis are: quantile regression (QR), expectile regression (ER), support vector regression (SVR), support vector classification (SVC) and logistic regression (Logit).

The performances of our method are then compared with alternative approaches developed in the literature; for all the models, we approximate the posterior via Markov chain Monte Carlo (MCMC), conjugate mean field variational Bayes (MFVB), or Laplace approximation when MFVB is not available. Here, MCMC is used as a proxy of the *true* posterior distribution. Except for expectile regression, all the other models considered here enjoy a conditional Gaussian representation, thereby MCMC and MFVB algorithms are based upon a data-augmentation strategy similar to that shown in Section 1.3 for quantile regression. For the Laplace approximation, the optimal densities are obtained in a transformed space such that all the parameters in the model belong in an unconstrained support. See Table 2.1 for the detailed references for all the models and estimation methods considered.

All the numerical routines used for the estimation have been implemented in `Julia` (Version 1.7.1). The simulations have been performed on a `Dell XPS 15` laptop with 4.7 gigahertz processor and 32 gigabytes of random access memory.

2.6.1 Performance measures

The simulations we propose rely on regression tasks, where an underling regression function $f(x)$ is estimated using a semiparametric model. To evaluate the ability of our method in reconstructing the original signal, we calculate the integrated absolute error (IAE) between the estimated and the true curve, that is defined as

$$\mathrm{IAE}(\hat{f}) = \int_0^1 |\hat{f}(x) - f(x)|\,\mathrm{d}x. \tag{2.24}$$

Table 2.1: References for the estimation algorithms used in the estimation of the models considered along all the simulation studies in Section 2.6.

Model	Method	Reference
Expectile regression	MCMC	Waldmann *et al.* (2017)
	Laplace	Standard BFGS optimization
Quantile regression	MCMC	Kozumi and Kobayashi (2011)
	MFVB	Wand *et al.* (2011)
Support vector regression	MCMC	Extension of Polson and Scott (2011)
	MFVB	Extension of Luts and Ormerod (2014)
Support vector classification	MCMC	Polson and Scott (2011)
	MFVB	Luts and Ormerod (2014)
Logistic regression	MCMC	Polson *et al.* (2013)
	MFVB	Durante and Rigon (2019)

For the quantile/expectile model, we compare the predictions with the true underling quantile/expectile. For support vector regression, instead, we compare the prediction with the true mean function. For the support vector classification, we compare the true decision function with the estimated one. While for logistic regression, we compare the true and estimated probability of success.

The posterior approximation accuracy is quantified by means of four accuracy measures: the relative absolute error (RAE) on the posterior mean vector and on the posterior variance-covariance, the average marginal accuracy score (Acc) and the evidence lower bound (ELBO) obtained at the end of the optimization. The absolute relative errors are calculated as

$$\mathrm{RAE}(\hat{\boldsymbol{\mu}}) = \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty / \|\boldsymbol{\mu}\|_\infty, \quad \mathrm{RAE}(\hat{\boldsymbol{\Sigma}}) = \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty / \|\boldsymbol{\Sigma}\|_\infty, \tag{2.25}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the *true* mean vector and variance matrix, evaluated via Monte Carlo approximation. Then, the average accuracy score is given by

$$\mathrm{Acc}(q^*) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{Acc}_k(q_k^*), \qquad \mathrm{Acc}_k(q_k^*) = 1 - \frac{1}{2}\int_{-\infty}^{+\infty} |q_k^*(\vartheta_k) - p(\vartheta_k|\mathbf{y})| \, \mathrm{d}\vartheta_k, \tag{2.26}$$

where K is the total number of regression parameters in the model and ϑ_k is the k-th element of the parameter vector $\boldsymbol{\vartheta}$. All these metrics are normalized and can be interpreted as proportions lying in $[0, 1]$.

In addition to the accuracy measures described so far, we also gathered the execution time in seconds and the number of iterations needed for all the algorithms to reach the convergence.

2.6.2 First simulation setup: semiparametric regression

In this first simulation study, the synthetic data are generated according to the non-linear model

$$y_i | x_i \sim \begin{cases} \mathrm{N}(\mu_i, \sigma_i^2) & \text{for quantile and expectile regression,} \\ \mathrm{t}(\mu_i, \sigma, \nu) & \text{for support vector regression,} \\ \mathrm{Be}(\pi_i) & \text{for logistic and support vector classification,} \end{cases} \tag{2.27}$$

where

$$\mu_i = f(x_i), \qquad \log(\sigma_i) = g(x_i), \qquad \mathrm{logit}(\pi_i) = h(x_i), \qquad x_i \sim \mathrm{U}(0,1),$$

are deterministic non-linear functions, $\mathrm{t}(\mu, \sigma, \nu)$ is the t distribution with location $\mu \in \mathbb{R}$, scale $\sigma > 0$ and degrees of freedom $\nu > 0$, $\mathrm{Be}(\pi)$ is the Bernoulli distribution with probability parameter $\pi \in (0, 1)$, $\mathrm{U}(0, 1)$ is the Uniform distribution on the interval $[0, 1]$ and $\mathrm{logit}(x) = \log\{x/(1 - x)\}$ is the inverse of the logistic transformation. Three specifications are considered for the non-linear functions in (2.27), named A, B, C. These

Table 2.2: Non-linear functions characterizing the three simulation settings described in the text.

Setting	Non-linear functions
A	$f(x) = 1.6\sin(3\pi x^2)$ $g(x) = -0.6 + 0.5\cos(4\pi x)$ $h(x) = 1.74\sin(3\pi x^2) - 1.076$
B	$f(x) = -1.02x + 0.018x^2 + 0.4\phi(x; 0.38, 0.08) + 0.08\phi(x; 0.75, 0.03)$ $g(x) = -0.25 + 0.15x^2 - 0.5\phi(x; 0.2, 0.1)$ $h(x) = -1.357x + 0.024x^2 + 0.532\phi(x; 0.38, 0.08) + 0.106\phi(x; 0.75, 0.03) - 0.003$
C	$f(x) = \sin(3\pi x^3) + 1.02x + 0.01x^2 + 0.4\phi(x; 0.38, 0.08)$ $g(x) = -0.4 + 0.3x^2 + \cos(3\pi x) - 0.5\phi(x; 0.2, 0.1)$ $h(x) = 0.91\sin(3\pi x^3) + 0.929x + 0.009x^2 + 0.364\phi(x; 0.38, 0.08) - 1.076$

are shown in Table 2.2. For each setting, we generated 100 independent datasets having the same number of observations, $n = 500$.

For all the considered scenarios, we model the linear predictor $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{u}$ using a mixed model based penalized spline. The covariate vectors $\mathbf{x}_i = (1,\, x_i)^\top$ and $\mathbf{z}_i = (z_1(x_i),\, \dots,\, z_d(x_i))^\top$ represent an orthogonalized O'Sullivan spline basis expansion (Wand and Ormerod, 2008), with corresponding fixed and random effect coefficients $\boldsymbol{\beta} = (\beta_0,\, \beta_1)^\top$ and $\boldsymbol{u} = (u_1, \dots, u_d)^\top$. The dimension of the basis expansion is $d = 40$. The prior distributions of $\boldsymbol{\beta}$, $\boldsymbol{u}$, σ_ε^2, σ_u^2 are specified as in equation (2.13), where $\sigma_\beta^2 = 10^6$, $A_\varepsilon = A_u = 2.0001$ and $B_\varepsilon = B_u = 1.0001$. These correspond to Inverse-Gamma distributions having mean $\mathbb{E}(\sigma_\varepsilon^2) = \mathbb{E}(\sigma_u^2) = 1$ and variance $\mathrm{Var}(\sigma_\varepsilon^2) = \mathrm{Var}(\sigma_u^2) = 10^3$. In each setting, for the heteroscedastic data we estimate quantile and expectile regression models with $\tau = 0.9$; for the homoscedastic t-distributed data we estimate a support vector regression model with $\epsilon = 0.01$; for the Bernoulli data we estimate logistic and support vector classification models. Algorithm 4 is stopped when the relative change of both the lower bound and the variational parameters fall bellow 10^{-4}.

Table 2.3 reports the accuracy measures described in Section 2.6.1 for each model, algorithm and setting. Only one dataset is considered for such an analysis. In terms of prediction accuracy and signal reconstruction (fifth column), semiparametric variational Bayes, mean field variational Bayes and Laplace approximation perform quite similarly. However, in terms of posterior approximation accuracy the situation is different. The evidence lower bound (fourth column) of semiparametric variational Bayes is always higher than mean field variational Bayes, indicating a better approximation of the joint posterior density. This also reflects on the average accuracy score (last column), indeed semiparametric variational Bayes outperforms mean field variational Bayes in almost all the simulations and for all the considered models, while being slightly less precise than Laplace approximation for expectile regression.

The improvement of semiparametric variational Bayes over mean field approximation is mainly due to a more precise quantification of the posterior variance of the regression coefficients, as can be seen by looking at the relative absolute errors in Table 2.3 (sixth

Table 2.3: Performance measure comparison between SVB, MFVB and Laplace approximation based on the 3 simulation setting outlined in the text and in Table 2.2.

Setting	Model	Method	ELBO	IAE(f)	RAE($\boldsymbol{\mu}$)	RAE($\boldsymbol{\Sigma}$)	Acc(q)
A	ER	Laplace		0.4757	0.0651	0.2114	0.9743
		SVB	249.3842	0.4450	0.1282	0.2466	0.9687
	QR	MFVB	-626.4552	0.7423	0.2833	0.5066	0.7809
		SVB	-621.0120	0.7402	0.0498	0.2412	0.8760
	SVR	MFVB	-547.6501	0.1265	0.1543	0.3785	0.8903
		SVB	-544.8723	0.1280	0.0267	0.1346	0.9324
	SVC	MFVB	-666.3824	0.0340	0.2958	0.7426	0.7537
		SVB	-664.1825	0.0300	0.1372	0.5491	0.8565
	Logit	MFVB	-334.9531	0.0619	0.0430	0.2558	0.9568
		SVB	-334.8914	0.0608	0.0315	0.1993	0.9570
B	ER	Laplace		0.5350	0.0347	0.1762	0.9678
		SVB	217.1924	0.5126	0.0479	0.1839	0.9669
	QR	MFVB	-656.9304	0.8259	0.1201	0.4614	0.8488
		SVB	-652.4221	0.8261	0.0360	0.1990	0.9136
	SVR	MFVB	-552.2767	0.1399	0.1625	0.4148	0.8724
		SVB	-549.2169	0.1398	0.0258	0.1490	0.9139
	SVC	MFVB	-573.2292	0.1080	0.1767	0.4468	0.8731
		SVB	-571.7809	0.1080	0.1051	0.3035	0.8915
	Logit	MFVB	-307.3833	0.0438	0.1293	0.3642	0.9318
		SVB	-307.1958	0.0424	0.1021	0.2995	0.9444
C	ER	Laplace		0.7467	0.1233	0.4780	0.9456
		SVB	89.9416	0.7107	0.1983	0.5413	0.9277
	QR	MFVB	-776.0003	1.0132	0.1453	0.5421	0.8220
		SVB	-771.6341	1.0200	0.0619	0.2812	0.8668
	SVR	MFVB	-638.7977	0.1105	0.0558	0.3575	0.8964
		SVB	-636.0827	0.1139	0.0218	0.1610	0.9372
	SVC	MFVB	-578.2343	0.0260	0.1806	0.4983	0.8328
		SVB	-575.9832	0.0260	0.0918	0.3328	0.8785
	Logit	MFVB	-291.6064	0.0347	0.0407	0.2926	0.9520
		SVB	-291.3853	0.0342	0.3261	0.1977	0.9567

and seventh columns). This fact is also highlighted by Figures 2.7 and 2.8, which show the predictive distribution and the marginal posterior density functions for some parameters in the model. Both figures refer to a quantile regression model estimated over one dataset from simulation setting A. Mean field variational Bayes systematically overshrinks the posterior variability producing narrow credibility bands for the estimated curve, see Figure 2.7. On the other hand, semiparametric variational Bayes tends to mimic in a more accurate way the Markov chain Monte Carlo posterior (Figure 2.8), producing reliable credibility intervals and prediction bands (Figure 2.7).

As we discussed in Section 2.2, under suitable conditions, the evidence lower bound

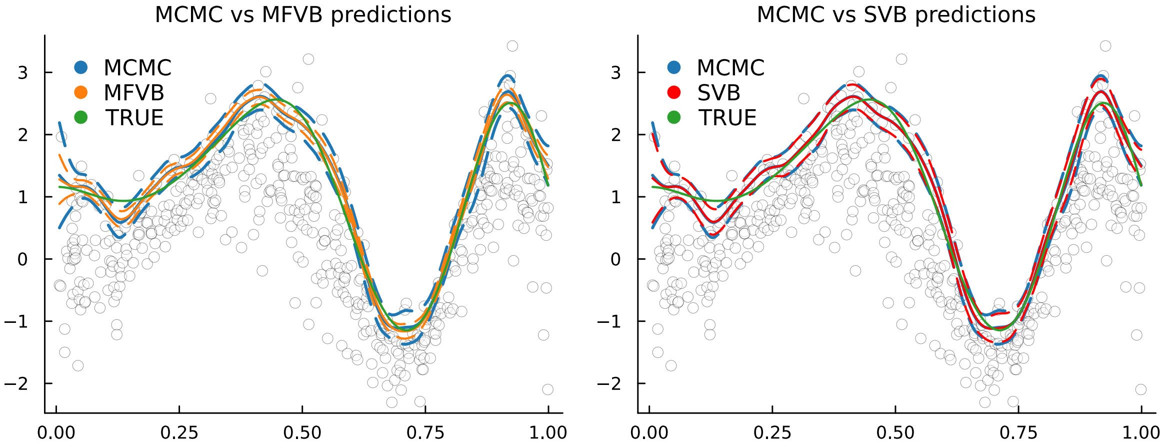

FIGURE 2.7: Posterior pointwise predictions (solid lines) and credibility bands (dashed lines) for the 90% quantile regression model described in the text. The estimates are obtained using a dataset from simulation setting A. (left) MFVB against MCMC. (right) SVB against MCMC.

for a marginal variational approach is always higher than the evidence lower bound obtained with data-augmented mean field approximation (Table 2.3 and Figure 2.8). This fact constitutes an empirical evidence in favor of Theorem 2.1, since Assumption 1 is satisfied and the inequality $\underline{\ell}\{\mathbf{y}; q^*_{\text{SVB}}(\boldsymbol{\theta})\} \geq \underline{\ell}\{\mathbf{y}; q^*_{\text{MFVB}}(\boldsymbol{\theta})\}$ implies

$$\text{KL}\{q^*_{\text{SVB}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\} \leq \text{KL}\{q^*_{\text{MFVB}}(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{y})\}.$$

To verify that Assumption 1 is accommodated, we just observe that both the approximations $q^*_{\text{SVB}}(\boldsymbol{\theta})$ and $q^*_{\text{MFVB}}(\boldsymbol{\theta})$ factorize according to the same partition, i.e. $q(\boldsymbol{\theta}) = q(\boldsymbol{\beta}, \boldsymbol{u})q(\sigma^2_u)q(\sigma^2_\varepsilon)$, and they take the same functional form, i.e. $q(\boldsymbol{\beta}, \boldsymbol{u})$ is Gaussian, while $q(\sigma^2_u)$ and $q(\sigma^2_\varepsilon)$ are Inverse-Gamma.

Figure 2.9 summarizes the results obtained by replicating the analysis described so far over a set of 100 datasets for each simulation setting. Coherently with the previous findings, semiparametric variational Bayes uniformly outperforms conjugate mean field variational Bayes in terms of marginal posterior approximation (first row). On the other hand, semiparametric variational inference has a slightly worse accuracy score than Laplace approximation for the expectile regression model. Still, this loss in the marginal accuracy for the expectile model does not have a significant impact on the relative errors for the posterior mean and variance, which are statistically equivalent for the two approximations (third and fourth rows). The performance in terms of signal reconstruction is almost equivalent among different approximation methods (second row).

For what concerns the computational complexity, both the number of iterations and the execution times obtained with the proposed approach are competitive with the alternatives methods in literature (fifth and sixth rows, Figure 2.9). This evidence is consistent with the theoretical computational complexity derived for Algorithm 4, i.e. $O(nm^2 + m^3)$, that is the same as Laplace approximation and data-augmented mean

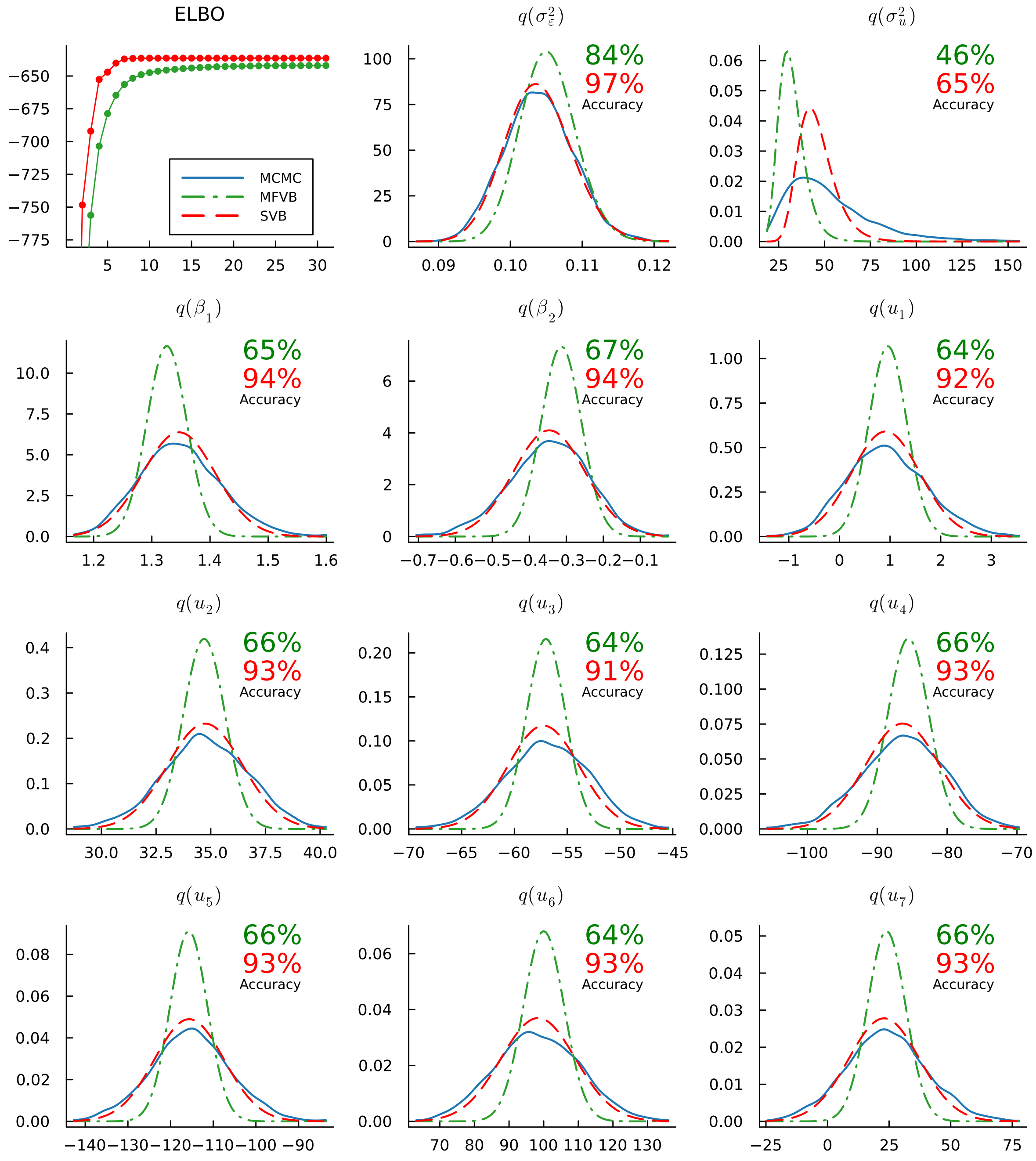

FIGURE 2.8: Top left: evidence lower bound evolution over the algorithm iterations. Others: optimal posterior density functions of the parameters in the 90% quantile regression model described in the text. The estimates are obtained using one dataset from simulation setting A. The solid blue line is for MCMC, the dashed red one for SVB and the dash-dotted green one for MFVB. The percentages denote the correspondent marginal accuracy scores defined in (2.26).

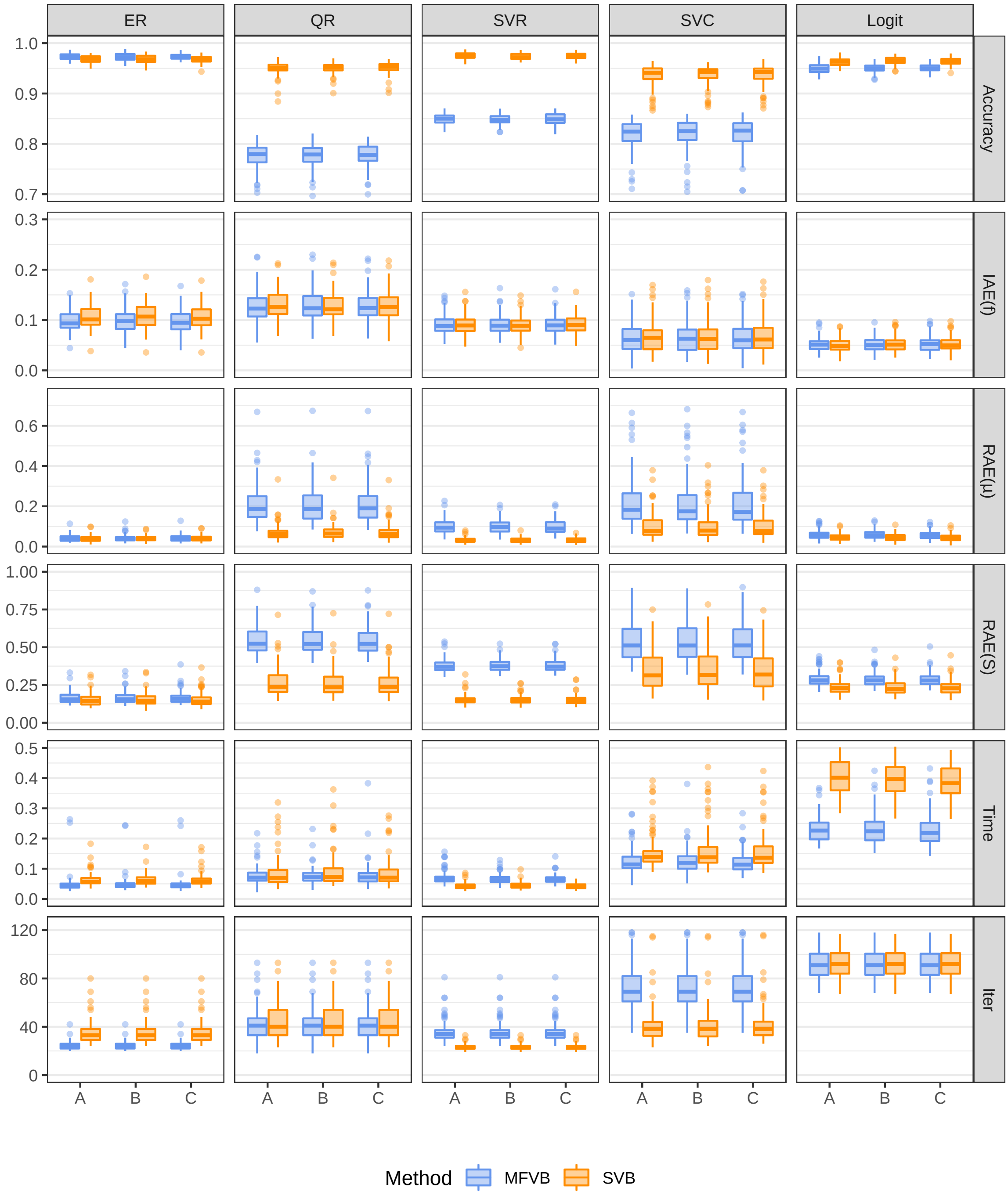

FIGURE 2.9: Boxplots of the summary statistics for the simulation study described in the text (Section 2.6.1 and 2.6.2). Each column corresponds to a model. Each row corresponds to a performance index. Within each panel, we find three groups of paired boxplot which correspond to the three simulation settings, i.e., A, B, C (see Equation 2.27 and Table 2.2). Each boxplot is calculated over 100 replications, i.e. 100 simulated datasets.

field variational Bayes for the models considered here.

2.6.3 Second simulation setup: random intercept models

In this second simulation study, we assess the relative quality of the proposed variational approximation for a random intercept model when the sample size and the number of parameters change. We consider two simulation setups: in the first one, setting A, the number of parameters is fixed and the sample dimension grows; in the second one, setting B, the sample size is kept fixed and the number of parameters increases.

In both the scenarios, the considered data generating mechanism is analogous to the one presented in (2.27): we have a heteroscedastic Gaussian model for quantile and expectile regression, a homoschedastic t distributed model for support vector regression, and a Bernoulli model for support vector classification and logistic regression. In formulas,

$$y_{ij}|x_{ij} \sim \begin{cases} \mathrm{N}(\mu_{ij}, \sigma_{ij}^2) & \text{for quantile and expectile regression,} \\ \mathrm{t}(\mu_{ij}, \sigma, \nu) & \text{for support vector regression,} \\ \mathrm{Be}(\pi_{ij}) & \text{for logistic and support vector classification.} \end{cases}$$

A random intercept specification is considered for μ_{ij}, σ_{ij} and π_{ij}, that is

$$\mu_{ij} = \beta_0 + \beta_1 x_{ij} + u_j, \qquad \log(\sigma_{ij}) = \gamma_0 + \gamma_1 x_{ij} + v_j, \qquad \mathrm{logit}(\pi_{ij}) = \mu_{ij},$$

where the indices $i = 1, \dots, n$ and $j = 1, \dots, d$ identify, respectively, the i-th subject and the j-th group under study. The fixed effect parameters β_0, β_1 and γ_0, γ_1 are generated according to a $\mathrm{N}(0, 1/2)$ distribution, the random intercepts $u_1, \dots, u_d$ and $v_1, \dots, v_d$ are generated according to a $\mathrm{N}(0, 1/4)$ distribution, while $\sigma = 1/10$ and $\nu = 4$.

In simulation setting A, the considered sample sizes are $n = 250, 500, 1000, 2000, 4000$, and the fixed number of groups is $d = 20$. In simulation setting B, the fixed sample size is $n = 500$ and the random intercept groups are $d = 5, 10, 25, 50, 100$. For each simulation setting, sample dimension and number of groups, we generate 100 datasets using the sampling design described so far.

For the estimation, we specify the linear predictor as $\eta_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{u}$, where $\mathbf{x}_{ij}^\top = (1, x_{ij})$ is a covariate vector and $\mathbf{z}_{ij}^\top$ is a $1 \times d$ selection vector associated to the j-th group, whose j-th entry is equal to 1 and all the others are 0.

The prior distributions of $\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2, \sigma_u^2$ are specified as in equation (2.13), where $\sigma_\beta^2 = 10^4$, $A_\varepsilon = A_u = 2.0001$ and $B_\varepsilon = B_u = 1.0001$. For the heteroscedastic data, we estimate quantile and expectile regression models with $\tau = 0.9$: for the homoscedastic t-distributed data, we estimate a support vector regression model with $\epsilon = 0.01$; for the Bernoulli data, we estimate a support vector classification model and a logistic regression.

In Figure 2.10, we show the boxplots of the average accuracy scores (2.26) calculated over a bunch of 100 datasets per each scenario, both under simulation setting A (left)

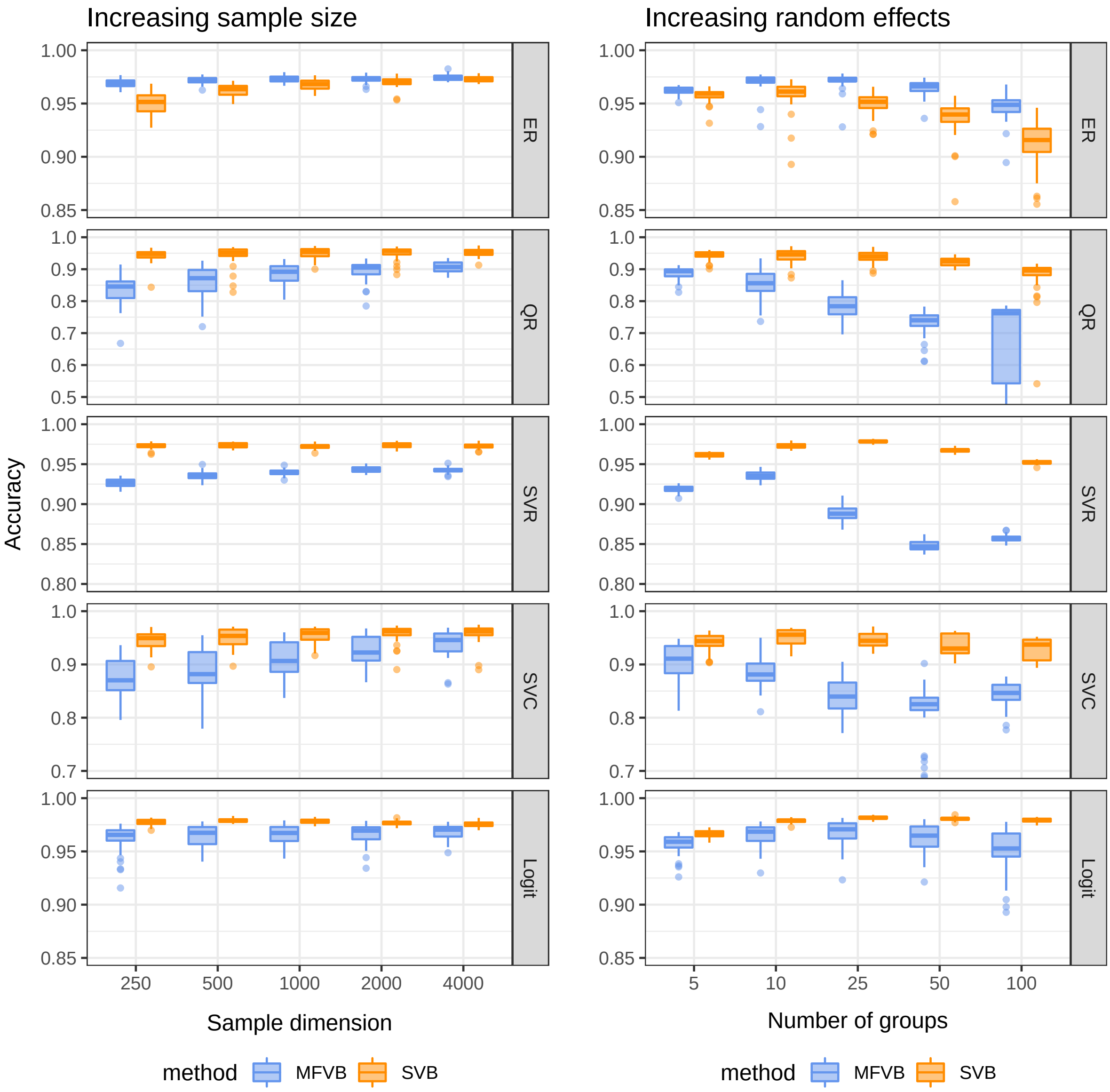

FIGURE 2.10: Sampling distribution of the marginal accuracy score defined in (2.26) for the simulation setup described in Section 2.6.3. Each row corresponds to a model, each column corresponds to a simulation setting. The left column is for setting A, the right column is for setting B.

and B (right). As we might expect, at the increase of the sample dimension (setting A), the accuracy scores improve both for semiparametric and mean field variational approximations. The opposite happens when the sample size is kept fixed and the number of groups, i.e., the number of parameters, grows (setting B). Furthermore, we observe a clear dominance of semiparametric variational Bayes over conjugate mean field variational Bayes in all the considered simulation setups, except for expectile regression, which is compared with Laplace approximation. Such a behavior confirms our findings in Section 2.6.2.

For what concerns the computation efficiency, Figure 2.11 provides the boxplots of the $\log_{10}$-transformed execution times measured for each model considered in simulation

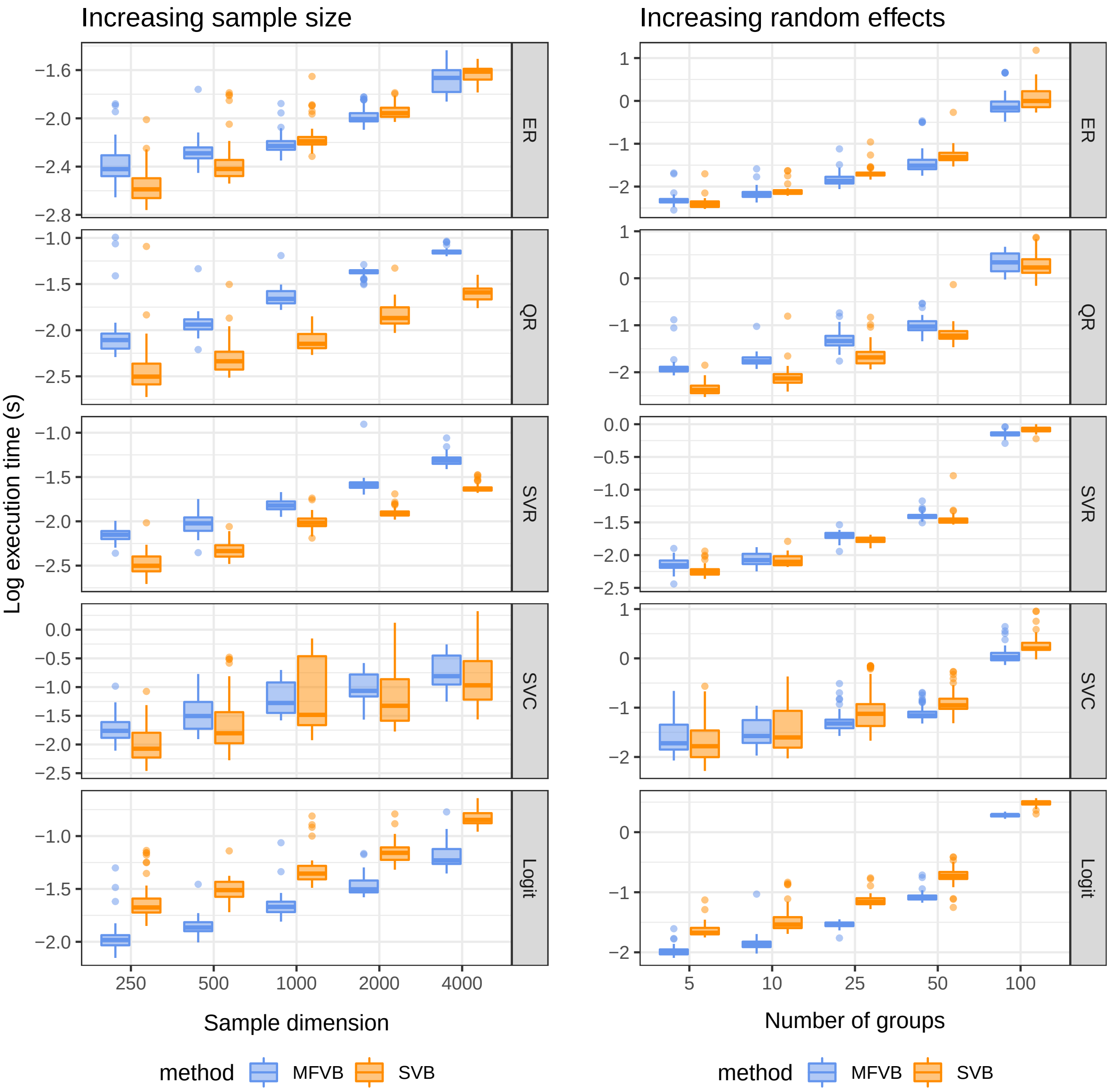

Figure 2.11: Sampling distribution of the $\log_{10}$-transformed execution time measured in seconds for the simulation setup described in Section 2.6.3. Each row corresponds to a model, each column corresponds to a simulation setting. The left column is for setting A, the right column is for setting B.

settings A and B. In both the scenarios, the execution time increase with the dimension of the problem, or in terms of sample size, or in terms of number of random effect parameters. For all the models and scenarios, our semiparametric variational routine reaches a computational time relative to or lower than the mean field competitor. The only exception is logistic regression, for which our algorithm systematically takes more time than mean field coordinate ascent, while maintaining the same computational complexity. This is due to the fact that we do not have closed form integration results for the Ψ functions in the logistic regression case (see Section 2.4.5), and we need to rely on numerical quadrature techniques. Therefore, each iteration of the algorithm requires a higher, but fixed, number of operations than conjugate mean field variational Bayes.

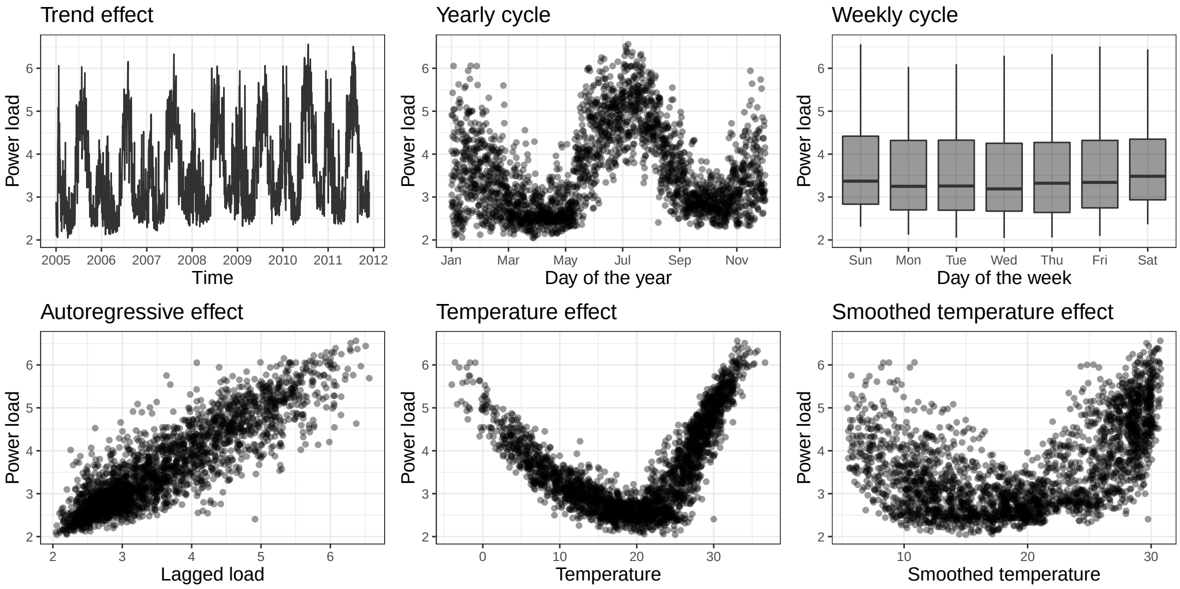

FIGURE 2.12: Pairwise scatter-plots of the power load consumption (in MWh) against the available covariates in the dataset, that are: (first row) the time (in days), the day of the year, the day of the week, (second row) the lagged power load, the temperature (in Celsius scale) and the smoothed temperature (in Celsius scale).

2.7 Probabilistic Load Forecasting

In this section, we present a real data problem concerning the forecast of the global electricity load consumption. In such a context, the data are usually highly dominated by non-stationary trends, multiple seasonal cycles of different lengths, like daily, weekly and monthly patterns, heteroscedasticity and by the presence of extreme values. Therefore, for the management of the power supply, it is of critical importance to understand and predict the behavior of the distribution of the load consumption, especially during exceptional events. To this end, a nonparametric density forecasting approach can be taken by pooling the information coming from several quantile estimates. Each conditional quantile can then be expressed as a non-linear function of the available meteorological, economic and social information, for example, using an additive model specification.

Here we consider the data used in the load forecasting track of the Global Energy Competition 2014 (GEFCom2014, Hong *et al.*, 2016), which have already been analyzed by, e.g., Gaillard *et al.* (2016) and Fasiolo *et al.* (2021a). The dataset collects the half-hourly load consumption and temperatures over the period going from January 2005 to December 2011. Our aim is then to estimate in a semiparametric way 19 equally spaced conditional quantiles between $\tau = 0.05$ and $\tau = 0.95$. This way, we provide an approximation of the conditional distribution of the load consumption at each time without imposing any parametric assumption. As Fasiolo *et al.* (2021a), we only consider the time interval between 11:30 and 12:00 a.m., but a similar analysis can be performed for all the remaining periods of the day, as it is common in literature. Figure 2.12 portrays how the observed power load is associated with some relevant covariates in the dataset.

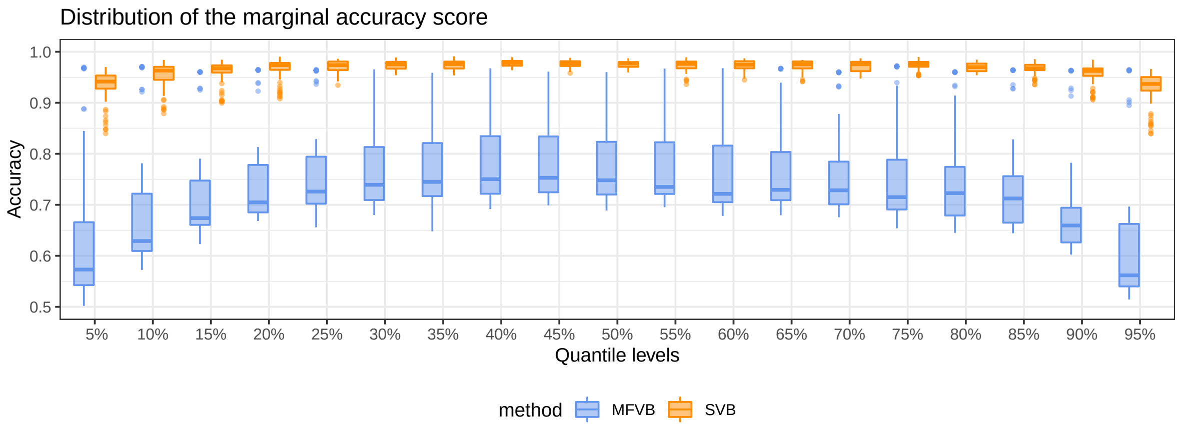

FIGURE 2.13: Distribution of the accuracy scores for the 19 estimated quantiles. For each quantile level, the boxplot is calculated over the individual accuracy scores of the 72 regression parameters in the model.

The τ-quantile of the power load, namely $\texttt{Load}_t$, is then modelled according to an additive specification, as described in Section 2.5.1. We only consider here the variables selected by Gaillard *et al.* (2016) and used also by Fasiolo *et al.* (2021a), obtaining so the linear predictor:

$$\begin{aligned}\eta_i = \beta_0 + f_1^4(\texttt{time}_i) + f_2^7(\texttt{week\_day}_i) + f_3^{20}(\texttt{year\_cycle}_i) \\ + f_4^{10}(\texttt{lagged\_load}_i) + f_5^{15}(\texttt{temperature}_i) + f_6^{15}(\texttt{smooth\_temp}_i),\end{aligned}$$

where $f_h^r(x) = \boldsymbol{u}_h^\top \mathbf{z}_h(x)$ is an orthogonal O'Sullivan spline basis expansion of rank r (Wand and Ormerod, 2008), having coefficients $\boldsymbol{u}_h = (u_{h,1}, \dots, u_{h,r})^\top$ and basis functions $\mathbf{z}_h(x) = (z_{h,1}(x), \dots, z_{h,r}(x))^\top$; $\texttt{week\_day}_i$ is a variable indicating the day of the week; $\texttt{temperature}_i$ is the hourly temperature; $\texttt{smooth\_temp}_i$ is the smoothed temperature, obtained as a moving average of the current and previous values of the temperature calculated using a weighting proportion of 0.05 and 0.95, respectively; $\texttt{year\_cycle}_i$ is a cyclic variable indicating the position within the year; $\texttt{lagged\_load}_i$ is the power load observed at the same time of the previous day; and $\texttt{time}_i$ is a trend variable indicating the time point.

We set diffuse prior distributions for all the parameters in the model, that is: $\sigma_\beta^2 = 10^6$, $A_\varepsilon = 2.0001$, $B_\varepsilon = 1.0001$, $A_h = 2.0001$, $B_h = 1.0001$, $h = 1, \dots, 5$. The parameters are then estimated using Markov chain Monte Carlo (Kozumi and Kobayashi, 2011), mean field variational Bayes (Wand *et al.*, 2011) and semiparametric variational Bayes (Algorithm 4).

The outcome of our analysis confirms an excellent performance of the proposed semiparametric variational approach in approximating the Monte Carlo posterior, as shown in Figures 2.13 and 2.14. The accuracy scores for all the considered quantile levels are all very close to 1 and almost always exceed 0.9 for semiparametric variational Bayes. On the other hand, mean field variational Bayes has a median accuracy centered around 0.7,

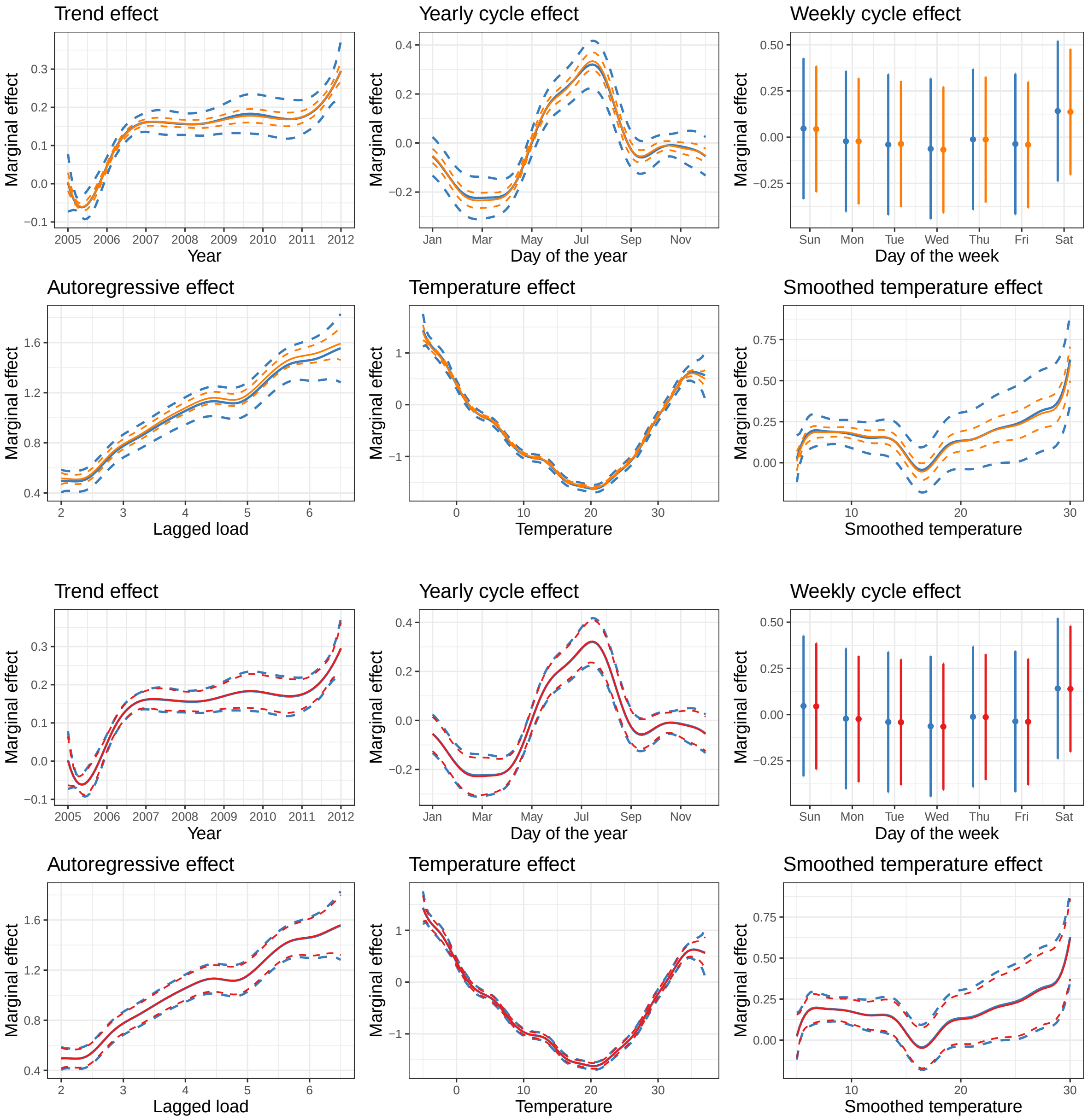

FIGURE 2.14: Estimated non-linear marginal effects (and credibility bands) for the available covariates in the analysis of 95% conditional quantile of the power load in the GEFCom2014 dataset. Top rows: MCMC estimated effects (blue curves) against MFVB estimated effects (orange curves). Bottom rows: MCMC estimated effects (blue curves) against SVB estimated effects (red curves).

with many values falling below 0.6, especially for the most extreme quantiles. Therefore, also in a real data example, semiparametric variational Bayes strongly outperforms conjugate mean field approximation.

The goodness of the approximation of our semiparametric variational approach is also confirmed by a graphical analysis of the estimated non-linear marginal effects of the covariates shown in Figure 2.14. The estimated curves, as well as the credibility bands, obtained using an *exact* Markov chain Monte Carlo approach or our semiparametric variational inference method are almost indistinguishable, even for extreme quantile levels.

Chapter 3

Spatial quantile regression with differential regularization

3.1 Introduction

In this work we are interested in modelling the heterogeneous effects of complex spatial phenomena on the quantiles of a response variable. Spatial anisotropy, flows, external perturbations and non-stationarity effects are well known sources of dependence in spatial data, whose impact can largely differ across quantiles. In addition, boundary constraints and non-trivial geometries of the sampling domain may introduce additional levels of complexity. An interesting example is given by meteorological and climate data. Temperature, pressure and precipitations often manifest local anisotropy and are influenced by wind streams and local characteristics of the geographical morphology. The presence of mountains, woods, or lakes, for instance, plays an important role in determining the regional behavior of weather. Moreover, such kind of data typically manifests a heteroscedastic skewed distribution, with local changes in space and time. Modelling these complex features is of prominent interest in a number of scientific fields, like physics, engineering, ecology, biology and, of course, spatial statistics.

In the spatial and quantile regression literature many authors studied the problem of estimating nonparametric and semiparametric regression models in one- and two-dimensional domains. For example, Koenker *et al.* (1994) and Ng (1996) considered quantile smoothing spline models with total variation regularization, and they proposed a linear programming algorithm for parameter estimation. Bosch *et al.* (1995) studied quantile regression with cubic smoothing spline, and estimated the parameters with an interior point algorithm. He *et al.* (1998) generalized the quantile smoothing approach to a bivariate setting, introducing a tensor product basis expansion. Koenker and Mizera (2004) extended the total variation regularization method to spatial regression quantiles by combining the so-called penalized triogram basis expansion with a linear programming algorithm. Univariate and bivariate quantile smoothing splines based on linear programming and interior point methods are implemented in the `R` package `quantreg` (Koenker, 2021). Alternative approaches to non-linear quantile regression are the local linear estimator by Yu and Jones (1998) and Hallin *et al.* (2009), and the reproducing kernel Hilbert space estimator by Li *et al.* (2007). More recently, Fasiolo *et al.* (2021a) proposed a fast calibrated framework for estimating additive and mixed quantile regression models, which is implemented in the `R` package `qgam` (Fasiolo *et al.*, 2021b). This leverages and extends the capabilities of the popular `R` package `mgcv` (Wood, 2017), enabling a wide range of univariate and bivariate smoothers in a quantile regression context. In particular, we mention thin plate spline smoothing (see, e.g., Wood, 2003) and soap film smoothing (Wood *et al.*, 2008), which are state-of-the-art smoothers for non-linear regression over unbounded and complex planar domains, respectively. Among spatial smoothers, it is also worth mentioning the bivariate penalized spline approach by, e.g., Ramsay (2002), Lai and Wang (2013) and Wang *et al.* (2020), which demonstrated a high degree of flexibility in handling spatial fields over domains with a complex boundary structure.

We here propose a nonparametric quantile regression model for spatially referenced data. In particular, we extend spatial regression with differential regularization by Sangalli *et al.* (2013), and Azzimonti *et al.* (2014), in order to estimate the conditional quantile of a spatially distributed response variable. The proposed method allows us to incorporate external physical knowledge in the estimation of the conditional quantile surface, whenever this information can be formulated as an elliptic partial differential equation (PDE; Evans, 2010). Such construction permits dealing with stationary and non-stationary anisotropic diffusion effects, unidirectional flows, and mixed boundary conditions. We can also handle complex planar domains characterized by strong concavities, holes, and physical barriers. An additional benefit of our approach is to allow for several extensions along different directions. For instance, we could consider data with space-time dependence, data gathered on smooth manifolds, data observed over areal regions; see, e.g., Sangalli (2021) for a comprehensive review of theory, extensions and applications of spatial regression methods with PDE regularization.

The novelty of our methodology is threefold. First, we introduce a broad class of physically-informed quantile regression models, based on a penalized loss criterion. In doing this, we trade off a goodness-of-fit measure and a roughness penalization depending on the PDE specification. Secondly, we propose an innovative parameter estimation algorithm following the expectation-maximization (EM) approach (Dempster *et al.*, 1977). The infinite-dimensional solution of such an optimization is then discretized by means of finite element methods (see, e.g., Quarteroni, 2017). Finally, we provide a theoretical characterization of both the infinite- and finite-dimensional PDE quantile estimators, proving existence, consistency and asymptotic normality.

3.2 Spatial quantile regression model

Consider n spatial locations $\mathbf{p}_1, \ldots, \mathbf{p}_n$ collected over the bounded region $\Omega \subset \mathbb{R}^2$, with regular boundary $\partial\Omega \in C^2(\mathbb{R})$. At each site $\mathbf{p}_i$ we observe a realization y_i, $i = 1, \ldots, n$, of a real random variable. We assume $y_1, \ldots, y_n$ are independent given the spatial locations, and y_i has an absolutely continuous distribution with probability density function $\pi_{y_i|\mathbf{p}_i}(y)$, cumulative density function $\Pi_{y_i|\mathbf{p}_i}(y)$, and quantile function $Q_{y_i|\mathbf{p}_i}(\tau) = \Pi^{-1}_{y_i|\mathbf{p}_i}(\tau) = \inf\{y \in \mathbb{R} : \Pi_{y_i|\mathbf{p}_i}(y) \geq \tau\}$, for any probability level $\tau \in (0,1)$. We further assume the following nonparametric spatial model for the τ-quantile of y_i

$$Q_{y_i|\mathbf{p}_i}(\tau) = f(\mathbf{p}_i), \quad \mathbf{p}_i \in \Omega, \quad i = 1, \dots, n, \tag{3.1}$$

where $f : \Omega \to \mathbb{R}$ is an unknown smooth deterministic field. Our aim is to estimate f, taking advantage of the available problem-specific information.

3.2.1 Problem-specific information and PDE specification

We here assume to have problem-specific prior knowledge about the phenomenon of interest, that can be described in terms of a PDE. Specifically, we consider linear second-order elliptic PDEs of the form

$$Lf = u \quad \text{in} \quad \Omega, \tag{3.2}$$

with squared integrable forcing term $u : \Omega \to \mathbb{R}$ and diffusion-transport-reaction operator L, defined as

$$Lf = -\text{div}(\mathbf{K}\nabla f) + \mathbf{b} \cdot \nabla f + cf. \tag{3.3}$$

Specifically, $\mathbf{K} \in \mathbb{R}^{2\times 2}$ denotes a symmetric positive definite diffusion tensor, $\mathbf{b} \in \mathbb{R}^2$ is a transport vector and $c \geq 0$ is a reaction term. Further, we denote by $\text{div}(\cdot)$ the divergence of a vector field, by ∇ the gradient of a scalar field, and by "$\cdot$" the standard scalar product, so that

$$\text{div}(\mathbf{K}\nabla f) = \sum_{i,j=1}^{2} K_{ij} \frac{\partial}{\partial p_i}\frac{\partial f}{\partial p_j} \qquad \text{and} \qquad \mathbf{b} \cdot \nabla f = \sum_{j=1}^{2} b_j \frac{\partial f}{\partial p_j}.$$

The PDE coefficients (i.e., $\mathbf{K}, \mathbf{b}, c$ and u) may vary over the domain, as smooth functions of $\mathbf{p} \in \Omega$. This permits to characterize a wide class of possibly non-stationary fields, manifesting different local behaviors. For instance, with the diffusion tensor $\mathbf{K}$ we can model non-stationary anisotropic diffusion effects; with the transport term $\mathbf{b}$ we can model non-stationary unidirectional effects, like flows over the domain; with the reaction term c we can model local shrinking effects. Finally, with the forcing term u we can model local perturbation of the field f from its homogeneous state; when $u = 0$, equation (3.2) is said homogeneous.

In a PDE-based description of a spatial phenomenon, a second fundamental ingredient is the specification of the boundary conditions, that is

$$Bf = \gamma \quad \text{on} \quad \partial\Omega, \tag{3.4}$$

which describes the behavior of f on the boundary. Here, B is an appropriate differential operator, and $\gamma : \partial\Omega \to \mathbb{R}$ is a smooth non-homogeneous term on the boundary.

Different types of boundary conditions can be imposed; for the sake of simplicity, we here consider only Neumann boundaries, which are the most natural choice for linear second order PDEs. In particular, we specify the first order differential operator on the border of the domain

$$Bf = \mathbf{K}\nabla f \cdot \boldsymbol{\nu} \quad \text{on} \quad \partial\Omega, \tag{3.5}$$

where $\boldsymbol{\nu}$ denotes the outward normal vector of $\partial\Omega$. Depending on the problem at hand, other types of mixed boundary conditions can be used to model the boundary behavior of the phenomenon under study, including Dirichlet and Robin conditions; see, e.g., the work of Azzimonti *et al.* (2014) for the use of such conditions on simpler linear regression models with PDE regularization.

3.2.2 Infinite-dimensional estimation problem

Building on the work of Sangalli *et al.* (2013) and Azzimonti *et al.* (2014), we propose to estimate the unknown spatial field f corresponding to the τ-quantile by minimizing the penalized loss functional

$$J_{\tau,\lambda}(f) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\{y_i - f(\mathbf{p}_i)\} + \frac{\lambda}{2}\int_\Omega (Lf - u)^2, \tag{3.6}$$

where $\lambda > 0$ is a smoothing parameter, and $\rho_\tau(x) = \frac{1}{2}|x| + (\tau - \frac{1}{2})x$ is the so-called quantile check function, or pinball loss (Koenker and Bassett, 1978). The penalized objective functional (3.6) trades off an appropriate goodness-of-fit criterion, the sum of quantile loss functions, and a prior information criterion, the L^2 norm of the PDE misfit $Lf - u$. The first term permits to center the estimate on the correct quantile location, the regularization term instead penalizes the departures from the domain-specific characterization of the physical problem, so as to shrink the estimates toward the solution space of the PDE considered.

When no physical information is available, except that f has to be smooth, we can use a simple stationary and isotropic PDE specification in order to penalize the local curvature of the field f. Namely, we can consider the diffusion equation $\Delta f = 0$, where $\Delta f = \partial^2 f/\partial p_1^2 + \partial^2 f/\partial p_2^2$ is the Laplacian of f. The resulting regularization term, $\int_\Omega (\Delta f)^2$, is a possible generalization of the univariate smoothing spline penalties to bivariate smoothing problems. See, e.g., Ramsay (2002), Wood *et al.* (2008) and Sangalli *et al.* (2013). This Laplacian regularization corresponds to a particular case of differential regularization considered in (3.6), where the PDE parameters are set to $\mathbf{K} = \mathbf{I}$, $\mathbf{b} = \mathbf{0}$, $c = 0$ and $u = 0$, where $\mathbf{I}$ denotes the identity matrix with appropriate dimensions.

Some regularity conditions on f should now be introduced to guarantee the existence the quantities defined so far. We denote by $H^d(\Omega)$ the Sobolev space of order d, that is the space of functions in $L^2(\Omega)$ having d weak derivatives in $L^2(\Omega)$. Formally,

$$H^d(\Omega) = \{f \in L^2(\Omega) : \ D^\alpha f \in L^2(\Omega), \ \forall|\alpha| \leq d\},$$

where $D^\alpha f$ is the α-th weak derivative of f. The appropriate functional embedding for the estimation problem in (3.6) is the space $\mathcal{F}_\gamma(\Omega)$, defined as

$$\mathcal{F}_\gamma(\Omega) = \{f \in H^2(\Omega) : Bf = \gamma \text{ on } \partial\Omega\}.$$

Because of the Sobolev embedding theorem, any function $f \in \mathcal{F}_\gamma$ is continuous, since $\mathcal{F}_\gamma(\Omega) \subset H^2(\Omega) \subset C(\bar{\Omega})$ for $\Omega \subset \mathbb{R}^2$, where $C(\bar{\Omega})$ is the set of continuous functions over the closure $\bar{\Omega} = \Omega \cup \partial\Omega$. Thereby, assuming $f \in \mathcal{F}_\gamma$, the point-wise evaluation of f at the observed spatial locations in (3.6) is well-defined, as well as the differential operators in (3.3) and (3.5).

Under the previous assumptions, the estimation problem takes the following formulation.

Problem 1. *Find* $\hat{f} \in \mathcal{F}_\gamma$ *such that* $J(\hat{f}) = \inf_{f\in\mathcal{F}_\gamma} J(f)$.

Let us denote by $\mathcal{V}_\gamma(\Omega) = \{\hat{f} \in \mathcal{F}_\gamma(\Omega) : J(\hat{f}) = \inf_{f\in\mathcal{F}_\gamma} J(f)\}$ the space collecting all the fields $\hat{f}$ minimizing the objective functional (3.6), i.e., solving Problem 1. The existence of $\mathcal{V}_\gamma$ is guaranteed by the following proposition.

Proposition 3.1. *The solution space* $\mathcal{V}_\gamma$ *is a non-empty, closed, convex set. Moreover, any field* $\hat{f} \in \mathcal{V}_\gamma$ *is a global minimum of the functional* (3.6).

Proof. We first recall that any continuous convex function defined over a convex domain attains its minimum values within its domain (see, e.g., Lange, 2013, Proposition 6.5.1). Then, proving the first statement in Proposition 3.1 is equivalent to showing that $\mathcal{F}_\gamma$ is a closed, convex space, and $J(f)$ is a continuous, convex functional.

The closure and convexity of $\mathcal{F}_\gamma(\Omega) = \{f \in H^2(\Omega) : Bf = \gamma \text{ on } \partial\Omega\}$ follows from the vector space structure of $H^2(\Omega)$ and from the linearity of the differential operator B, that is

$$B\{\phi f + (1-\phi)g\} = \phi Bf + (1-\phi)Bg = \phi\gamma + (1-\phi)\gamma = \gamma, \qquad \forall\, \phi \in [0,1].$$

The continuity and convexity of $J(f)$ follow from the continuity and convexity of the quantile loss $\rho_\tau\{y_i - f(\mathbf{p}_i)\}$ (see, e.g., Koenker and Bassett, 1978) and the regularization term $\int_\Omega (Lf - u)^2$ (see, e.g., Azzimonti *et al.*, 2014).

Finally, we prove that $\mathcal{V}_\gamma(\Omega)$ is a closed, convex set. To do so, we define the sublevel set of $J(f)$ by

$$\mathcal{V}_\gamma(\Omega, t) = \big\{f \in \mathcal{F}_\gamma(\Omega) : J(f) \le t\big\},$$

for any t such that $\mathcal{V}_\gamma(\Omega, t)$ is non-empty. Let $f, g \in \mathcal{V}_\gamma(\Omega, t)$ and consider the convex combination

$$J\{\phi f + (1-\phi)g\} \le \phi J(f) + (1-\phi)J(g) \le \phi t + (1-\phi)t = t, \quad \forall\, \phi \in [0,1].$$

This implies the closure and convexity of $\mathcal{V}_\gamma(\Omega, t)$ for any t since $J(f)$ is a continuous functional. The closure and convexity of $\mathcal{V}_\gamma(\Omega)$ immediately follows by noting that

$$\mathcal{V}_\gamma(\Omega) = \big\{f \in \mathcal{F}_\gamma(\Omega) : J(\hat{f}) = \inf J(f)\big\} = \mathcal{V}_\gamma\big(\Omega, \inf J(f)\big).$$

This concludes the proof. □

Differently from the strictly convex optimization criteria discussed, e.g., by Sangalli *et al.* (2013) and Azzimonti *et al.* (2014) for simple linear models, or by Wilhelm and Sangalli (2016) for generalized linear models, in the quantile regression framework the solution to Problem 1 is not guaranteed to be unique for finite samples. However, it is worth nothing that all the estimators minimizing (3.6) are global minimizers and, therefore, reach the same value of the penalized loss functional (3.6). In the parametric quantile regression literature, this is a known property that naturally arises by exploiting the alternative representation of Problem 1 in terms of linear programming. For more details in the parametric context, we refer the reader to Koenker and Bassett (1978) and Koenker (2005).

The elements belonging to $\mathcal{V}_\gamma$ can also be characterized by means of a first-order necessary condition. This leverages on the convexity and continuity of the functional $J(f)$ without requiring any further regularity properties, such as uniform differentiability of the loss function. The main idea is to exploit the behavior of directional derivatives of (3.6) in a neighborhood of its minimum, and to use standard convex analysis results to ensure optimality (Rockafellar, 1997). Intuitively, any departure from the optimum must yield an increment of the functional and thereby leads to a non-negative effect on its directional derivatives. This is formally stated in the following proposition

Proposition 3.2. *Let $\hat{f} \in \mathcal{V}_\gamma$ be a minimum of (3.6). Then, $\hat{f}$ satisfies*

$$-\frac{1}{n}\sum_{i=1}^{n} \psi(\mathbf{p}_i)\, d\rho_\tau\{y_i - \hat{f}(\mathbf{p}_i), -\psi(\mathbf{p}_i)\} \geq -\lambda \int_\Omega (L\psi)(L\hat{f} - u), \quad \forall\, \psi \in \mathcal{F}_0,$$

where $d\rho_\tau(v, w)$ is the directional derivative of $\rho_\tau(\cdot)$ calculated in $v \in \mathbb{R}$ along the direction $w \in \mathbb{R}$, defined as

$$d\rho_\tau(v, w) = \begin{cases} \tau - \mathbb{I}(v < 0), & \text{if } v \neq 0, \\ \tau - \mathbb{I}(w < 0), & \text{if } v = 0. \end{cases}$$

Proof. Let $\hat{f} \in \mathcal{V}_\gamma$ be a minimizer of functional (3.6). For any $t \geq 0$ and $\psi \in \mathcal{F}_0$, we have $J(\hat{f}) \leq J(\hat{f} + t\psi)$. Hence, taking the limit for $t \downarrow 0$, we get that the Gateaux directional derivative of $J(\hat{f})$ along the direction ψ must be non-negative:

$$\partial_\psi J(\hat{f}) = \frac{\partial}{\partial t} J(\hat{f} + t\psi)\bigg|_{t=0} = \lim_{t \downarrow 0} \frac{J(\hat{f} + t\psi) - J(\hat{f})}{t} \geq 0. \tag{3.7}$$

Whenever $J(\cdot)$ is a differentiable functional, the above condition collapses into the first order equation $\partial_\psi J(\hat{f}) = 0$ for any $\psi \in \mathcal{F}_0$. This is the case, for instance, for generalized linear models with PDE regularization (Wilhelm and Sangalli, 2016).

Result 3.2 is a particular case of inequality (3.7), obtained when the functional form of $J(\cdot)$ is explicitly considered. That is

$$\partial_\psi J(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} \partial_\psi \rho_\tau(y_i - \hat{f}(\mathbf{p}_i)) + \frac{\lambda}{2}\partial_\psi \int_\Omega (L\hat{f} - u)^2 \geq 0.$$

As shown by, e.g., Azzimonti *et al.* (2014), the directional derivative for the penalty term is given by

$$\partial_\psi \int_\Omega (L\hat{f} - u)^2 = 2\int_\Omega (L\psi)(L\hat{f} - u).$$

On the other hand, the directional derivative for the quantile loss term (Koenker, 2005) is

$$\partial_\psi \rho_\tau\{y_i - f(\mathbf{p}_i)\} = -\psi(\mathbf{p}_i)\, d\rho_\tau\{y_i - \hat{f}(\mathbf{p}_i), -\psi(\mathbf{p}_i)\}.$$

Finally, the Gateaux derivative of $J(\cdot)$ is

$$\partial_\psi J(\hat{f}) = -\frac{1}{n}\sum_{i=1}^{n} \psi(\mathbf{p}_i)\, d\rho_\tau\{y_i - \hat{f}(\mathbf{p}_i), -\psi(\mathbf{p}_i)\} + \lambda \int_\Omega (L\psi)(L\hat{f} - u) \geq 0.$$

This concludes the proof. □

Proposition 3.2 characterizes the elements of the solution space $\mathcal{V}_\gamma$. However, it does not provide any direct way to find an explicit solution to the minimization of (3.6). The quantile estimator in Problem 1 must then be computed via iterative algorithms, like the one proposed in Section 3.3.

3.3 Functional EM algorithm

As commented at the end of the previous section, the theoretical characterization of Problem 1, given in Proposition 3.2, does not provide a direct way to find an estimator $\hat{f} \in \mathcal{V}_\gamma$, which must instead be computed via numerical methods. The approach here followed is to employ an EM algorithm (Dempster *et al.*, 1977; McLachlan and Krishnan, 2008) in order to approximate the optimization Problem 1 by a sequence of simpler problems, having closed form solutions. The optimizers of such sequence of problems converge in the limit to an element of the space $\mathcal{V}_\gamma$, because of the monotonic convergence property of the EM algorithm (see Section 1.3.1).

3.3.1 Algorithm derivation

We here give a sketch of the derivation of our functional EM algorithm, while referring to Appendix B.1 for further technical details. First, we recall the result by Yu and Moyeed (2001), which states the equivalence between the negative log-likelihood of an Asymmetric-Laplace model (Kotz *et al.*, 2001) with the quantile check function in (3.6). This suggests that solving Problem 1 corresponds to maximize the penalized log-likelihood functional assuming for the regression model the following non-Gaussian specification:

$$y_i = f(\mathbf{p}_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{AL}(0, \sigma_\varepsilon^2, \tau), \qquad i = 1, \dots, n, \tag{3.8}$$

where $\mathrm{AL}(\mu, \sigma, \tau)$ denotes the Asymmetric-Laplace law with, location $\mu \in \mathbb{R}$, scale $\sigma > 0$ and shape $\tau \in (0,1)$. The working probability density function of ε_i is then given by

$$\pi(\varepsilon_i|\boldsymbol{\theta}) = \tau(1-\tau)\exp\big\{-\rho_\tau(\varepsilon_i)/\sigma_\varepsilon^2\big\}/\sigma_\varepsilon^2,$$

and is such that $\int_{-\infty}^{0} \pi(\varepsilon|\boldsymbol{\theta})\,\mathrm{d}\varepsilon = \tau$. As a consequence, the penalized log-likelihood for the unknown parameter $\boldsymbol{\theta} = (f, \sigma_\varepsilon^2) \in \mathcal{F}_\gamma \times \mathbb{R}_+$ induced by the misspecified model specification (3.8) corresponds to

$$\ell_\lambda(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; y_i) - \frac{\lambda n}{2\sigma_\varepsilon^2} \int_\Omega (Lf - u)^2, \tag{3.9}$$

where $\ell(\boldsymbol{\theta}; y_i)$ denotes the i-th unpenalized contribution of the Laplace log-likelihood, being $\ell(\boldsymbol{\theta}; y_i) = -\log\sigma_\varepsilon^2 - \rho_\tau\{y_i - f(\mathbf{p}_i)\}/\sigma_\varepsilon^2$.

We then take advantage of the location-scale mixture representation of the Laplace distribution introduced in Section 1.2.1 (see also Kotz *et al.*, 2001), which permits to write the i-th error component in model (3.8) as a Gaussian random variable with hierarchical conditional distribution

$$\varepsilon_i|\omega_i; \boldsymbol{\theta} \sim \mathrm{N}\left(\frac{(1-2\tau)\omega_i}{\tau(1-\tau)}, \frac{2\sigma_\varepsilon^2\omega_i}{\tau(1-\tau)}\right), \quad \omega_i|\boldsymbol{\theta} \sim \mathrm{Exp}(1/\sigma_\varepsilon^2), \quad i = 1, \dots, n, \tag{3.10}$$

where $\mathrm{N}(\mu, \sigma^2)$ is the univariate Gaussian law with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, while $\mathrm{Exp}(\mu)$ is the Exponential law with mean $1/\mu > 0$.

Combining the completed log-likelihood functional relative to the model (3.8) and (3.10) with the PDE regularization term in (3.6), we obtain the penalized completed log-likelihood

$$\ell_\lambda(\boldsymbol{\theta}; \boldsymbol{\omega}, \mathbf{y}) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \omega_i, y_i) - \frac{\lambda n}{2\sigma_\varepsilon^2} \int_\Omega (Lf - u)^2,$$

where the i-th augmented data contribution is equal to

$$\ell(\boldsymbol{\theta}; \omega_i, y_i) = -\frac{3}{2}\log\sigma_\varepsilon^2 - \frac{1}{2}\log\omega_i - \frac{\omega_i}{\sigma_\varepsilon^2} - \frac{\tau(1-\tau)}{4\sigma_\varepsilon^2\omega_i}\left\{y_i - f(\mathbf{p}_i) - \frac{(1-2\tau)}{\tau(1-\tau)}\omega_i\right\}^2.$$

Provided that $\exp\{\ell(\boldsymbol{\theta}; y_i)\} = \int_0^\infty \exp\{\ell(\boldsymbol{\theta}; \omega_i, y_i)\}\,\mathrm{d}\omega_i$ (see, e.g., Kotz *et al.*, 2001), the penalized log-likelihood functional (3.9) can be maximized via EM algorithm (see, e.g., Section 1.3.1), iterating until convergence the expectation (E) and maximization (M) steps. At the $(k+1)$-th iteration of the procedure, the parameter estimates are updated as follows:

$$\begin{aligned} &\text{E-step:} \qquad \underline{\ell}_\lambda^{(k)}(\boldsymbol{\theta};\mathbf{y}) = \mathbb{E}^{(k)}\{\ell_\lambda(\boldsymbol{\theta};\boldsymbol{\omega},\mathbf{y})\}, \\ &\text{M-step:} \qquad \boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta}\in\Theta}{\text{argmin}}\ \underline{\ell}_\lambda^{(k)}(\boldsymbol{\theta};\mathbf{y}), \end{aligned}$$

with $\Theta = \mathcal{F}_\gamma \times \mathbb{R}_+$. The expectation in the E-step is taken with respect to the conditional distribution of $\omega_1, \dots, \omega_n$ given the observations $y_1, \dots, y_n$, keeping fixed the values of the spatial field $f^{(k)}$ and of the scale parameter $\sigma_\varepsilon^{2(k)}$ obtained at the k-th iteration of the algorithm.

As shown by, e.g., Kozumi and Kobayashi (2011) and Tian *et al.* (2014) and discussed in Section 1.3.1, *a posteriori* each latent factor ω_i is independently distributed according to a Generalized-Inverse-Gaussian probability law, as in (1.20). We can thus use the linearity of the expected value, exploit the analytic integration of ω_i, and discard all the additive terms not depending on the unknown f and σ_ε^2, in order to obtain a closed form expression for the expected penalized log-likelihood functional $\underline{\ell}_\lambda^{(k)}(\boldsymbol{\theta};\mathbf{y})$. Hence, at the (k)-th iteration of the algorithm, we have

$$\underline{\ell}_\lambda^{(k)}(\boldsymbol{\theta};\mathbf{y}) = -\frac{3}{2} n \log \sigma_\varepsilon^2 - \frac{(a_1^2 + 2a_2^2)}{2a_2^2 \sigma_\varepsilon^2} \mathbf{1}_n^\top \boldsymbol{\mu}_\omega^{(k)} - \frac{n}{2\sigma_\varepsilon^2} \tilde{J}_\lambda^{(k)}(f), \tag{3.11}$$

where $J_\lambda^{(k)}(f)$ is a quadratic functional of f not depending on σ_ε^2:

$$J_\lambda^{(k)}(f) = \frac{1}{n}(\mathbf{z}^{(k)} - \mathbf{f}_n)^\top \mathbf{W}^{(k)} (\mathbf{z}^{(k)} - \mathbf{f}_n) + \lambda \int_\Omega (Lf - u)^2. \tag{3.12}$$

Here, $\mathbf{z}^{(k)} = \mathbf{y} - (1 - 2\tau)|\mathbf{y} - \mathbf{f}_n^{(k)}|$ is a vector of working observations and $\mathbf{W}^{(k)} = \text{diag}(\mathbf{w}^{(k)})$ is a working weight matrix such that $1/\mathbf{w}^{(k)} = 2|\mathbf{y} - \mathbf{f}_n^{(k)}|$. Hereafter, we use the notation $\mathbf{f}_n = (f(\mathbf{p}_1), \dots, f(\mathbf{p}_n))^\top$ to indicate the vector containing the evaluation of any spatial field f at the n spatial locations $\mathbf{p}_1, \dots, \mathbf{p}_n$.

As suggested by the expression in (3.11), maximizing $\underline{\ell}_\lambda^{(k)}(\boldsymbol{\theta};\mathbf{y})$ with respect to f is equivalent to minimizing the quadratic functional $J_\lambda^{(k)}(f)$. Moreover, such an optimization does not depend on the value of σ_ε^2, therefore the exact optimization in the M-step of the algorithm is achieved by first profiling out f and, then, using the new value of $f^{(k)}$ to obtain the scale parameter $\sigma_\varepsilon^{2(k)}$. The update for σ_ε^2 can hence be evaluated by maximizing either

$$\underline{\ell}_\lambda^{(k)}(\sigma_\varepsilon^2;\mathbf{y}) = \underline{\ell}_\lambda^{(k)}(\boldsymbol{\theta};\mathbf{y})\big|_{f=f^{(k)}}, \qquad \text{or} \qquad \ell_\lambda^{(k)}(\sigma_\varepsilon^2;\mathbf{y}) = \ell_\lambda(\boldsymbol{\theta};\mathbf{y})\big|_{f=f^{(k)}};$$

both the alternatives lead to a genuine EM update with closed form solution, that are

$$\begin{aligned} \underline{\sigma}_\varepsilon^{2(k)} &= \underset{\sigma_\varepsilon^2\in\mathbb{R}_+}{\text{argmax}}\ \underline{\ell}_\lambda^{(k)}(\sigma_\varepsilon^2;\mathbf{y}) = \frac{2}{3n}\left\{\frac{(a_1^2+2a_2^2)}{2a_2^2}\mathbf{1}_n^\top\boldsymbol{\mu}_\omega^{(k)} + \frac{n}{2}J_\lambda^{(k)}(f^{(k)})\right\}, \\ \sigma_\varepsilon^{2(k)} &= \underset{\sigma_\varepsilon^2\in\mathbb{R}_+}{\text{argmax}}\ \ell_\lambda^{(k)}(\sigma_\varepsilon^2;\mathbf{y}) = \frac{1}{n}\sum_{i=1}^n \rho_\tau\{y_i - f^{(k)}(\mathbf{p}_i)\} + \frac{\lambda}{2}\int_\Omega (Lf^{(k)} - u). \end{aligned}$$

However, the second estimator is preferable in terms of efficiency and algorithm convergence rate, being the exact solution to the original maximum likelihood problem.

It is worth noting that functional (3.12) in the M-step of the algorithm is well-defined only in the case where all elements of the vector $1/\mathbf{w}^{(k)} = 2|\mathbf{y} - \mathbf{f}_n^{(k)}|$ are non-zero. Otherwise, the calculation of $\mathbf{W}^{(k)}$ might produce diagonal entries approaching ∞. Moreover, the M-step still involves an optimization over an infinite-dimensional space. However, differently from original Problem 1, the minimization of (3.12) is convex and quadratic, and thus can be appropriately solved by extending the procedure discussed in Sangalli *et al.* (2013) and Azzimonti *et al.* (2014) for simple linear regression models. The infinite-weight issue and the characterization of the infinite-dimensional formulation of the M-step of the algorithm are the main topics of the next section.

As a final remark, we observe that the value of the penalized log-likelihood functional (3.9) increases at each iteration of the EM algorithm, or at least it does not increase, yielding a stable monotonic convergence to a stationary point. Any solution thus belongs to $\mathcal{V}_\gamma$ thanks to the convexity and coercitivity of functional (3.6). See, e.g., Lange (2013), Chapters 9 and 12, for more details.

3.3.2 Constrained formulation

In the previous section we introduced an infinite-weight problem: there exist the possibility that some elements of $1/\mathbf{w}^{(k)}$ numerically approach zero. As a consequence, the corresponding entries of $\mathbf{w}^{(k)}$ might diverge, causing numerical instabilities in the calculation and optimization of functional (3.12). Such a behavior has already been observed in the analysis of Bayesian support vector machines by Polson and Scott (2011). As argued also in Polson and Scott (2011), values of $1/\mathbf{w}^{(k)}$ numerically close to zero do not indicate a pathological behavior of the algorithm. Actually, they arise in correspondence of support vector points, that are observations for which the complementary slackness conditions are active constraints in the Karush-Kuhn-Tucker formulation of support vector machines (Bishop, 2006, Chapter 7). See Section 3 of Polson and Scott (2011) for more details. In the quantile regression context we observe a similar phenomenon, that is related to the activation of some implicit constraints that arise when Problem 1 is rephrased in terms of linear programming (see, e.g., Koenker, 2005). Intuitively, if $1/w_i^{(k)} \to 0$, then the i-th squared pseudo-residual $w_i^{(k)}\{z_i^{(k)} - f(\mathbf{p}_i)\}^2$ will receive an infinite weight since $w_i^{(k)} \to \infty$. This behavior implicitly enforces the constraint $z_i^{(k)} - f(\mathbf{p}_i) = 0$ to be satisfied. A way to solve this issue, and overcome potential numerical instabilities, is to explicitly take into account the constraint induced by $1/w_i^{(k)} = 0$. To this end, we decouple the contribution of the unconstrained and constrained parts of the optimization, defining the partition $\mathbf{z}^{(k)} = \{\mathbf{z}_s^{(k)}, \mathbf{z}_{-s}^{(k)}\}$, where $\mathbf{z}_s^{(k)} = \{z_i^{(k)} : 1/w_i^{(k)} = 0\}$ and $\mathbf{z}_{-s}^{(k)} = \{z_i^{(k)} : 1/w_i^{(k)} > 0\}$. Under this setting, the objective functional (3.12) becomes

$$J_{-s}^{(k)}(f) = \frac{1}{n}(\mathbf{z}^{(k)} - \mathbf{f}_n)_{-s}^\top \mathbf{W}_{-s}^{(k)}(\mathbf{z}^{(k)} - \mathbf{f}_n)_{-s} + \lambda \int_\Omega (Lf - u)^2,$$

which is always well-defined, since we are no longer dividing by zero. Moreover, we define the constraint $(\mathbf{z}^{(k)} - \mathbf{f}_n)_s = \mathbf{0}$. Then, the M-step of the EM algorithm can be alternatively stated as follows.

Problem 2. *Find* $\tilde{f} \in \mathcal{F}_\gamma$ *such that* $J_{-s}^{(k)}(\tilde{f}) = \inf_{f\in\mathcal{F}_\gamma} J_{-s}^{(k)}(f)$ *subject to* $(\mathbf{z}^{(k)} - \tilde{\mathbf{f}}_n)_s = \mathbf{0}$.

Because of the Lagrange multiplier theorem (see, e.g., Nocedal and Wright, 2006, Chapter 13), searching a solution to Problem 2 is equivalent to minimizing the Lagrangian functional

$$\mathcal{L}^{(k)}(f, \boldsymbol{\eta}) = J_{-s}^{(k)}(f) + \boldsymbol{\eta}^\top(\mathbf{z}^{(k)} - \mathbf{f}_n)_s, \tag{3.13}$$

where $\boldsymbol{\eta} \in \mathbb{R}^{|s|}$ is a vector of Lagrange multipliers and $|s|$ is the number of active constraints. Hence, we can directly optimize (3.13) with respect to $\boldsymbol{\eta}$ and f, without imposing any explicit constraint.

In order to characterize the minimum of (3.13) in a variational formulation, we introduce the field $g = Lf - u \in L^2(\Omega)$, which represents the misfit of the PDE. We define the bilinear forms $R_1(\cdot,\cdot)$ and $R_0(\cdot,\cdot)$ and the linear operator $F(\cdot)$ as follows

$$\begin{aligned} R_1(\phi,\psi) &= \int_\Omega \big[(\mathbf{K}\nabla\phi)\cdot\nabla\psi + (\mathbf{b}\cdot\nabla\phi)\psi + c\phi\psi\big], \\ R_0(\phi,\psi) &= \int_\Omega \phi\psi, \quad F(\phi) = \int_\Omega u\phi + \int_{\partial\Omega}\gamma\phi, \end{aligned} \tag{3.14}$$

for any pair of functions $\phi, \psi \in H^1(\Omega)$. As we mention in Section 3.2, we here assume for simplicity Neumann boundary conditions, but similar formulas arise when Dirichlet or Robin conditions are imposed; see, e.g., the works of Azzimonti *et al.* (2014) and Arnone *et al.* (2019) for the linear regression case. Assuming that all these quantities are well-defined, the minimizer of the Lagrangian functional (3.13) satisfies the following proposition.

Proposition 3.3. *Let* $(\tilde{f}, \tilde{\boldsymbol{\eta}}) \in \mathcal{F}_\gamma \times \mathbb{R}^{|s|}$ *be a minimum of* $\mathcal{L}^{(k)}(f,\boldsymbol{\eta})$ *in* (3.13) *and let* $\tilde{g} = L\tilde{f} - u \in H^1(\Omega)$. *Then,* $(\tilde{f}, \tilde{g}, \tilde{\boldsymbol{\eta}})$ *is the solution of the following system of first order equations*

$$\begin{aligned} \tfrac{1}{n}(\boldsymbol{\psi}_n)_{-s}^\top \mathbf{W}_{-s}^{(k)}(\mathbf{f}_n)_{-s} - (\boldsymbol{\psi}_n)_s^\top\tilde{\boldsymbol{\eta}} + \lambda R_1(\psi,\tilde{g}) &= \tfrac{1}{n}(\boldsymbol{\psi}_n)_{-s}^\top\mathbf{W}_{-s}^{(k)}\mathbf{z}_{-s}^{(k)}, \\ R_1(\tilde{f},\phi) - R_0(\tilde{g},\phi) &= F(\phi), \\ (\mathbf{z}^{(k)} - \tilde{\mathbf{f}}_n)_s &= \mathbf{0}, \end{aligned} \tag{3.15}$$

for any pair of test functions $\phi, \psi \in \mathcal{F}_0$.

Proof. Let $\mathcal{L}(f,\boldsymbol{\eta}) = J_{-s}(f) + \boldsymbol{\eta}^\top(\mathbf{z} - \mathbf{f}_n)_s$ be the Lagrangian functional associated with the constrained optimization in Problem 2 and let $(\tilde{f}, \tilde{\boldsymbol{\eta}}) \in \mathcal{F}_\gamma \times \mathbb{R}^{|s|}$ be a minimizer of $\mathcal{L}(f,\boldsymbol{\eta})$. Then, because of the Lagrange multiplier theorem, $(\tilde{f}, \tilde{\boldsymbol{\eta}})$ must satisfy the first order conditions

$$\frac{\partial}{\partial\boldsymbol{\eta}}\mathcal{L}(\tilde{f},\tilde{\boldsymbol{\eta}}) = 0, \qquad \text{and} \qquad \frac{\partial}{\partial t}\mathcal{L}(\tilde{f} + t\psi, \tilde{\boldsymbol{\eta}})\bigg|_{t=0} = 0, \qquad \forall\, \psi \in \mathcal{F}_0.$$

The first equation just enforces the constraints, being

$$\frac{\partial}{\partial \boldsymbol{\eta}} \mathcal{L}(\tilde{f}, \tilde{\boldsymbol{\eta}}) = \frac{\partial}{\partial \boldsymbol{\eta}} \{ \tilde{\boldsymbol{\eta}}^\top (\mathbf{z} - \tilde{\mathbf{f}}_n)_s \} = (\mathbf{z} - \mathbf{f}_n)_s = 0.$$

The second equation requires that the Gateaux directional derivatives of $\mathcal{L}(f, \boldsymbol{\eta})$ with respect to f get nullified in all the possible directions $\psi \in \mathcal{F}_0$.

Observing that $\mathcal{L}(f, \boldsymbol{\eta})$ is quadratic in f, taking the directional derivative and equating to zero, we obtain the first order variational equation

$$\begin{aligned} &\frac{1}{n} (\boldsymbol{\psi}_n)_{-s}^\top \mathbf{W}_{-s} \, (\tilde{\mathbf{f}}_n)_{-s} + \lambda \int_\Omega (L\psi)(L\tilde{f}) = \\ &\qquad = \frac{1}{n} (\boldsymbol{\psi}_n)_{-s}^\top \mathbf{W}_{-s} \, \mathbf{z}_{-s} + (\boldsymbol{\psi}_n)_{-s}^\top \tilde{\boldsymbol{\eta}} + \lambda \int_\Omega (L\psi) u, \end{aligned} \tag{3.16}$$

which must hold for any $\psi \in \mathcal{F}_0$. We denote by $G(\psi, \tilde{f})$ and $T(\tilde{f})$ the left and right sides of the above equation (3.16), respectively; in this way, we can write the first order condition in the equivalent form

$$G(\psi, \tilde{f}) = T(\psi), \qquad \forall \psi \in \mathcal{F}_0. \tag{3.17}$$

If the parameters of the PDE are such that, for any $u \in L^2(\Omega)$, there exists a unique solution f of the PDE $Lf = u$, which, moreover, satisfies $f \in H^2(\Omega)$ (Assumption 2 in Azzimonti *et al.*, 2014), then $G(\cdot, \cdot)$ is a symmetric, continuous, coercive bilinear map, and $T(\cdot)$ is a continuous linear operator (Theorem 2 in Azzimonti *et al.*, 2014). As a consequence, thanks to the Lax-Milgram lemma (see, e.g., Quarteroni, 2017, Section 3.4.1), there exists a unique solution $\tilde{f} \in \mathcal{F}_\gamma$ to equation (3.17).

In order to obtain the first order optimality conditions as stated in Proposition 3.3, we now exploit the following equivalent weak variational formulation of the Euler-Lagrange equation (3.16):

$$\begin{aligned} \frac{1}{n} (\boldsymbol{\psi}_n)_{-s}^\top \mathbf{W}_{-s} \, (\tilde{\mathbf{f}}_n)_{-s} + \lambda \int_\Omega (L\psi) \tilde{g} &= \frac{1}{n} (\boldsymbol{\psi}_n)_{-s}^\top \mathbf{W}_{-s} \, \mathbf{z}_{-s} + (\boldsymbol{\psi}_n)_{-s}^\top \tilde{\boldsymbol{\eta}}, \\ \int_\Omega (L\tilde{f})\phi - \int_\Omega \tilde{g}\phi &= \int_\Omega u\phi, \end{aligned} \tag{3.18}$$

which must hold for any pair of test functions $\psi, \phi \in \mathcal{F}_0$, where $\tilde{g} = L\tilde{f} - u$. Hence, integrating by parts $\int_\Omega (L\psi)\tilde{g}$ and $\int_\Omega (L\tilde{f})\phi$, and using the definition of $R_1(\cdot, \cdot)$, $R_0(\cdot, \cdot)$ and $F(\cdot)$ in (3.14) we can recognize the weak variational formulation in Proposition (3.3).

Let us first consider the explicit formulas

$$\begin{aligned} \int_\Omega (L\psi)\tilde{g} &= \int_\Omega \big[-\operatorname{div}(\mathbf{K}\nabla\psi)\tilde{g} + (\mathbf{b} \cdot \nabla\psi)\tilde{g} + c\psi\tilde{g} \big], && \forall \psi \in \mathcal{F}_0, \\ \int_\Omega (L\tilde{f})\phi &= \int_\Omega \big[-\operatorname{div}(\mathbf{K}\nabla\tilde{f})\phi + (\mathbf{b} \cdot \nabla\tilde{f})\phi + c\tilde{f}\phi \big], && \forall \phi \in \mathcal{F}_0. \end{aligned}$$

Assuming $\tilde{g} = L\tilde{f} - u \in H^1(\Omega)$, using the first Green identity, the non-homogeneous Neumann boundary conditions for $\tilde{f}$ and the homogeneous boundary conditions for ψ, we get

$$\begin{aligned} -\int_\Omega \operatorname{div}(\mathbf{K}\nabla\psi)\tilde{g} &= \int_\Omega (\mathbf{K}\nabla\psi)\cdot\nabla\tilde{g}, \\ -\int_\Omega \operatorname{div}(\mathbf{K}\nabla\tilde{f})\phi &= \int_\Omega (\mathbf{K}\nabla\tilde{f})\cdot\nabla\phi - \int_{\partial\Omega}\gamma\phi, \end{aligned}$$

which lead to the identities

$$\int_\Omega (L\psi)\tilde{g} = R_1(\psi,\tilde{g}) \qquad \text{and} \qquad \int_\Omega (L\tilde{f})\phi = R_1(\tilde{f},\phi) - \int_{\partial\Omega}\gamma\phi$$

The proof is concluded by the noting that the above equations combined with the variational system in (3.18) give rise the final result in (3.15). □

Such a variational formulation of Problem 2 is the cornerstone to derivation of the finite element discretization we propose in Section 3.4.

3.4 Finite element discretization

In order to tackle the infinite-dimensional problem as stated in the weak formulation (3.15), we study a numerical solution within a proper finite-dimensional subspace. We consider a regular triangularization $\mathcal{T}_h$ of the original spatial domain with characteristic size h, where h is the maximum length of the triangle edges. In this way, Ω can be represented by the union of all triangles in $\mathcal{T}_h$, leading to the approximated domain Ω_h with polygonal boundary $\partial\Omega_h$. The discretization $\mathcal{T}_h$, also called mesh, is a fundamental tool in numerical analysis and engineering that permits to describe the geometry of possibly very complex domains, characterized by strong concavities, holes, or a curved nature.

3.4.1 Finite element basis expansion

Denote by $\mathcal{P}_r(\boldsymbol{\tau})$ the space of polynomial functions of order r over the triangle $T \in \mathcal{T}_h$ and define by $\mathcal{F}^r_\gamma(\Omega_h) \subset H^1(\Omega_h) \cap C(\bar{\Omega}_h)$ the finite-dimensional subspace

$$\mathcal{F}^r_\gamma(\Omega_h) = \{f_h \in C(\bar{\Omega}_h) : \ f_h|_T \in \mathcal{P}_r(T) \ \ \forall T \in \mathcal{T}_h, \ \ Bf_h = \gamma_h \text{ on } \partial\Omega_h\}.$$

where γ_h is the local r-th order polynomial interpolation of γ. Starting from the triangular discretization $\mathcal{T}_h$, we can thus define locally supported polynomial functions that provide a basis $\psi_1,\ldots,\psi_{N_h}$ for the $\mathcal{F}^r_\gamma(\Omega_h)$. If we consider piecewise linear functions, the elements of the basis expansion $\psi_1,\ldots,\psi_{N_h}$ have a one-to-one correspondence with the nodes of the mesh $\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{N_h}$, that are the vertices of the triangles. A graphical representation is provided in Figure 3.1. The evaluation of the i-th basis on the j-th

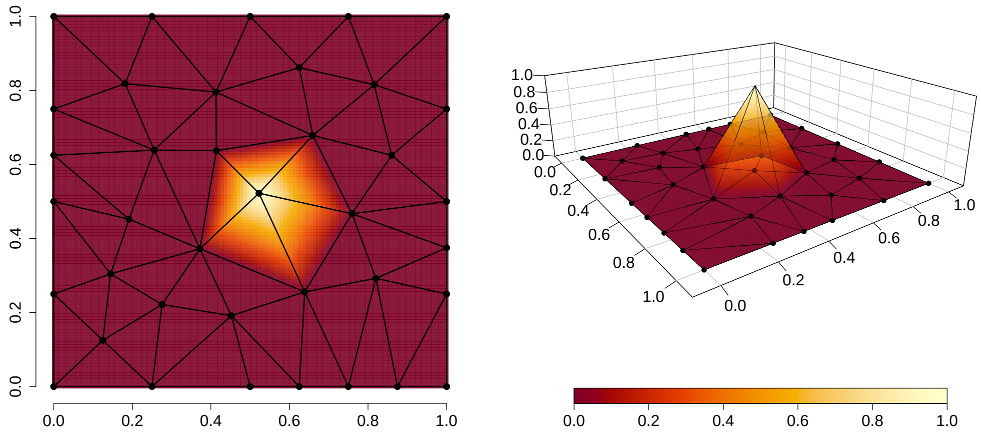

FIGURE 3.1: A linear finite element basis function on a triangular mesh.

node is given by $\psi_i(\boldsymbol{\xi}_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta with $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$, for any $i, j \in \{1, \dots, N_h\}$. Hence, any surface $f_h \in \mathcal{F}_{\gamma,h} \equiv \mathcal{F}_\gamma^1(\Omega_h)$ is uniquely determined by its values at the nodes:

$$f_h(\mathbf{p}) = \sum_{j=1}^{N_h} f_h(\boldsymbol{\xi}_j)\, \psi_j(\mathbf{p}) = \mathbf{f}^\top \boldsymbol{\psi}(\mathbf{p}),$$

where $\boldsymbol{\psi}(\mathbf{p}) = (\psi_1(\mathbf{p}), \dots, \psi_{N_h}(\mathbf{p}))^\top$ is the basis vector at the point $\mathbf{p} \in \Omega$, while $\mathbf{f} = (f_h(\boldsymbol{\xi}_1), \dots, f_h(\boldsymbol{\xi}_{N_h}))^\top$ is the coefficient vector of the basis expansion.

It is worth highlighting that the mesh can be constructed independently of the n data locations $\mathbf{p}_1, \dots, \mathbf{p}_n$. In fact, in some applications, this may be very useful, especially when the data locations have a spatial distribution which is far from been uniform over the domain, since a coarse mesh composed by triangles with very different dimensions and sharp angles may lead to numerical instabilities and poor approximations (see, e.g., Quarteroni, 2017). This may be the case, for instance, when the data are characterized by clustering effects or regional sparsity. In these situations, a regular mesh can be constructed using only a subset of the data locations and filling the remaining spatial regions with an almost uniform discretization, so that to improve the numerical stability of the solution and the accuracy of the approximation.

3.4.2 Finite-dimensional estimator

Let $\boldsymbol{\Psi}$ be the matrix evaluation of the N_h basis functions at the n data locations:

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\psi}(\mathbf{p}_1)^\top \\ \vdots \\ \boldsymbol{\psi}(\mathbf{p}_n)^\top \end{bmatrix} = \begin{bmatrix} \psi_1(\mathbf{p}_1) & \cdots & \psi_{N_h}(\mathbf{p}_1) \\ \vdots & & \vdots \\ \psi_1(\mathbf{p}_n) & \cdots & \psi_{N_h}(\mathbf{p}_n) \end{bmatrix},$$

so that, for any $f_h \in \mathcal{F}_{\gamma,h}$, we have $\mathbf{f}_n = \boldsymbol{\Psi}\mathbf{f}$. Let $\mathbf{R}_1$ be the $N_h \times N_h$ matrix evaluation of the bilinear form $R_1(\cdot,\cdot)$ over the finite element basis:

$$\mathbf{R}_1 = \int_{\Omega_h} \left[(\nabla\boldsymbol{\psi})\mathbf{K}(\nabla\boldsymbol{\psi})^\top + (\nabla\boldsymbol{\psi})(\mathbf{b}\boldsymbol{\psi}^\top) + c\boldsymbol{\psi}\boldsymbol{\psi}^\top\right],$$

where $\nabla\boldsymbol{\psi}$ is the Jacobian matrix of the vector field $\boldsymbol{\psi}$. Similarly, we define the mass matrix $\mathbf{R}_0 = \int_{\Omega_h} \boldsymbol{\psi}\boldsymbol{\psi}^\top$, being the discretization of $R_0(\cdot,\cdot)$. The notation $\mathbf{u} = \int_{\Omega_h} u_h\boldsymbol{\psi}$ and $\boldsymbol{\gamma} = \int_{\partial\Omega_h} \gamma_h\boldsymbol{\psi}$ represent the discretization of $F(\cdot)$. Then the finite element approximation of the system (3.15) in Proposition 3.3 is presented in the following proposition.

Proposition 3.4. *The finite element estimator* $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}}, \tilde{\boldsymbol{\eta}}) \in \mathbb{R}^{N_h} \times \mathbb{R}^{N_h} \times \mathbb{R}^{|s|}$ *is the solution of the linear system*

$$\begin{bmatrix} \frac{1}{n}\boldsymbol{\Psi}_{-s}^\top\mathbf{W}_{-s}^{(k)}\boldsymbol{\Psi}_{-s} & \lambda\mathbf{R}_1^\top & \boldsymbol{\Psi}_s^\top \\ \lambda\mathbf{R}_1 & -\lambda\mathbf{R}_0 & \mathbf{O} \\ \boldsymbol{\Psi}_s & \mathbf{O} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{f}} \\ \tilde{\mathbf{g}} \\ \tilde{\boldsymbol{\eta}} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\boldsymbol{\Psi}_{-s}^\top\mathbf{W}_{-s}^{(k)}\mathbf{z}_{-s}^{(k)} \\ \lambda(\mathbf{u}+\boldsymbol{\gamma}) \\ \mathbf{z}_s^{(k)} \end{bmatrix}. \tag{3.19}$$

Proof. From Proposition 3.3, any solution $(\tilde{f}_h, \tilde{g}_h, \tilde{\boldsymbol{\eta}})$ in the finite element space must satisfy the system of equations

$$\begin{aligned} \tfrac{1}{n}(\boldsymbol{\psi}_n)_{-s}^\top\mathbf{W}_{-s}^{(k)}(\tilde{\mathbf{f}}_n)_{-s} - (\boldsymbol{\psi}_n)_s^\top\tilde{\boldsymbol{\eta}} + \lambda R_1(\psi_h, \tilde{g}_h) &= \tfrac{1}{n}(\boldsymbol{\psi}_n)_{-s}^\top\mathbf{W}_{-s}^{(k)}\mathbf{z}_{-s}^{(k)}, \\ R_1(\tilde{f}_h, \phi_h) - R_0(\tilde{g}_h, \phi_h) &= F(\phi_h), \\ \mathbf{z}_s^{(k)} - (\tilde{\mathbf{f}}_n)_{-s} &= \mathbf{0}, \end{aligned}$$

for any $\psi_h, \phi_h \in \mathcal{F}_{0,h}$. Thanks to the finite element discretization, we have

$$\begin{array}{lll} (\boldsymbol{\psi}_n)_{-s} = \boldsymbol{\Psi}_{-s}\boldsymbol{\psi}, & (\tilde{\mathbf{f}}_n)_{-s} = \boldsymbol{\Psi}_{-s}\tilde{\mathbf{f}}, & (\tilde{\mathbf{f}}_n)_{-s} = \boldsymbol{\Psi}_s\tilde{\mathbf{f}}, \\ R_0(\psi_h, \tilde{g}_h) = \boldsymbol{\psi}^\top\mathbf{R}_0^\top\tilde{\mathbf{g}}, & R_1(\tilde{f}_h, \phi_h) = \mathbf{f}^\top\mathbf{R}_1^\top\boldsymbol{\phi}, & F(\phi_h) = (\mathbf{u}+\boldsymbol{\gamma})^\top\boldsymbol{\phi}, \end{array}$$

therefore the above system can be written as

$$\begin{aligned} \tfrac{1}{n}\boldsymbol{\psi}^\top\boldsymbol{\Psi}_{-s}^\top\mathbf{W}_{-s}^{(k)}\boldsymbol{\Psi}_{-s}\tilde{\mathbf{f}} - \boldsymbol{\psi}^\top\boldsymbol{\Psi}_s^\top\tilde{\boldsymbol{\eta}} + \lambda\boldsymbol{\psi}^\top\mathbf{R}_1^\top\tilde{\mathbf{g}} &= \tfrac{1}{n}\boldsymbol{\psi}^\top\boldsymbol{\Psi}_{-s}^\top\mathbf{W}_{-s}^{(k)}\mathbf{z}_{-s}^{(k)}, \\ \tilde{\mathbf{f}}^\top\mathbf{R}_1^\top\boldsymbol{\phi} - \tilde{\mathbf{g}}^\top\mathbf{R}_0\boldsymbol{\phi} &= (\mathbf{u}+\boldsymbol{\gamma})^\top\boldsymbol{\phi}, \\ \mathbf{z}_s^{(k)} - \boldsymbol{\Psi}_s\tilde{\mathbf{f}} &= \mathbf{0}, \end{aligned}$$

for any pair of vectors $\boldsymbol{\psi}, \boldsymbol{\phi} \in \mathbb{R}^{N_h}$. Since any function $f_h \in \mathcal{F}_{\gamma,h}$ is uniquely determined by its values on the nodes, i.e. by its coefficient vector, solving the above system is equivalent to finding $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}}, \tilde{\boldsymbol{\eta}}) \in \mathbb{R}^{N_h} \times \mathbb{R}^{N_h} \times \mathbb{R}^{|s|}$ such that

$$\begin{aligned} \tfrac{1}{n}\boldsymbol{\Psi}_{-s}^\top\mathbf{W}_{-s}^{(k)}\boldsymbol{\Psi}_{-s}\tilde{\mathbf{f}} - \boldsymbol{\Psi}_s^\top\tilde{\boldsymbol{\eta}} + \lambda\mathbf{R}_1^\top\tilde{\mathbf{g}} &= \tfrac{1}{n}\boldsymbol{\Psi}_{-s}^\top\mathbf{W}_{-s}^{(k)}\mathbf{z}^{(k)}, \\ \mathbf{R}_1\tilde{\mathbf{f}} - \mathbf{R}_0\tilde{\mathbf{g}} &= \mathbf{u}+\boldsymbol{\gamma}, \\ \mathbf{z}_s^{(k)} - \boldsymbol{\Psi}_s\tilde{\mathbf{f}} &= \mathbf{0}, \end{aligned}$$

which corresponds to (3.19). This concludes the proof. □

The system (3.19) in Proposition 3.4 admits closed form solution that can be expressed as

$$\begin{aligned}
\tilde{\mathbf{f}} &= (\mathbf{A}^{(k)})^{-1}\mathbf{d}^{(k)} + (\mathbf{A}^{(k)})^{-1}\boldsymbol{\Psi}_s^\top(\mathbf{B}^{(k)})^{-1}\big\{\mathbf{z}_s^{(k)} - \boldsymbol{\Psi}_s(\mathbf{A}^{(k)})^{-1}\mathbf{d}^{(k)}\big\},\\
\tilde{\mathbf{g}} &= \mathbf{R}_0^{-1}(\mathbf{R}_1\tilde{\mathbf{f}} - \mathbf{u} - \boldsymbol{\gamma}),\\
\tilde{\boldsymbol{\eta}} &= (\mathbf{B}^{(k)})^{-1}\big\{\mathbf{z}_s^{(k)} - \boldsymbol{\Psi}_s(\mathbf{A}^{(k)})^{-1}\mathbf{d}^{(k)}\big\}.
\end{aligned} \tag{3.20}$$

Here, we use the matrix notation

$$\begin{aligned}
\mathbf{A}^{(k)} &= \tfrac{1}{n}\boldsymbol{\Psi}_{-s}^\top\mathbf{W}_{-s}^{(k)}\boldsymbol{\Psi}_{-s} + \lambda\mathbf{P},\\
\mathbf{B}^{(k)} &= \boldsymbol{\Psi}_s(\mathbf{A}^{(k)})^{-1}\boldsymbol{\Psi}_s^\top,\\
\mathbf{d}^{(k)} &= \tfrac{1}{n}\boldsymbol{\Psi}_{-s}^\top\mathbf{W}_{-s}^{(k)}\mathbf{z}_{-s}^{(k)} + \lambda\mathbf{h},
\end{aligned}$$

where $\mathbf{P} = \mathbf{R}_1^\top\mathbf{R}_0^{-1}\mathbf{R}_1$ is the $N_h \times N_h$ penalty matrix that discretizes the penalty term in (3.6) and $\mathbf{h} = \mathbf{R}_1^\top\mathbf{R}_0^{-1}(\mathbf{u}+\boldsymbol{\gamma})$ is the $N_h \times 1$ bias vector induced by the non-homogeneous terms of equations (3.2) and (3.4). Moreover, $(\mathbf{A}^{(k)})^{-1}\mathbf{d}^{(k)}$ is the solution of the unconstrained optimization problem

$$\min_{\mathbf{f}\in\mathbb{R}^{N_h}} J_{-s}^{(k)}\big(\mathbf{f}^\top\boldsymbol{\psi}\big) = \min_{f_h\in\mathcal{F}_{\gamma,h}} J_{-s}^{(k)}(f_h).$$

The constrained estimator $\tilde{\mathbf{f}}$ is then obtained by projecting the unconstrained solution $(\mathbf{A}^{(k)})^{-1}\mathbf{d}^{(k)}$ onto the null space $\big\{\mathbf{f}\in\mathbb{R}^{N_h} : \mathbf{z}_s^{(k)} - \boldsymbol{\Psi}_s\mathbf{f} = \mathbf{0}\big\}$.

Remark 3.5. If $\mathbf{A}^{(k)}$ is non-singular and $\boldsymbol{\Psi}_s$ has full row rank, i.e., if $\mathbf{B}^{(k)}$ is non-singular, the estimator $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}}, \tilde{\boldsymbol{\eta}})$ is the unique solution to the linear system (3.19) in Proposition 3.4 (Nocedal and Wright, 2006, Chapter 16).

Remark 3.6. Because of the monotonic convergence of the EM algorithm (Lange, 2013, Chapters 9 and 12), in the limit for $k \to \infty$, the finite element estimator $\tilde{f}_h = \tilde{\mathbf{f}}^\top\boldsymbol{\psi}$ converges to $\hat{f}_h = \hat{\mathbf{f}}^\top\boldsymbol{\psi}$, which is the minimizer of $J(f_h)$ for $f_h \in \mathcal{F}_{\gamma,h}$.

The above results, and in particular Propositions 3.3 and 3.4, may be generalized in order to accommodate for additive and semiparamtric models, when also a set of fixed covariates is available, as we show in Section 3.6.

3.4.3 Finite-dimensional EM algorithm

Thanks to Proposition 3.4, we have all the ingredients to define a numerical routine approximating the quantile estimator defined in Problem 1. Algorithm 5 provides a pseudo-code description of the resulting EM algorithm with finite-element approximation of the spatial field f.

Notice that in the numerical treatment of the constraint Problem 2, it is not possible to precisely determine the set of indices $s = \{i : |y_i - f_h(\mathbf{p}_i)| = 0,\ i = 1,\dots,n\}$ because of rounding errors introduced by numerical approximations. Instead, we here consider the alternative definition $s = \{i : |y_i - f_h(\mathbf{p}_i)| \le \epsilon,\ i = 1,\dots,n\}$, where ϵ is a small tolerance parameter. In our numerical experiments we set $\epsilon = 10^{-6}$.

Algorithm 5 Functional EM algorithm for nonparametric spatial quantile regression

```
Require: $\tau, \lambda, \mathbf{y}, \boldsymbol{\Psi}, \mathbf{R}_0, \mathbf{R}_1$
  Initialize $\hat{\mathbf{f}}$, $\hat{\mathbf{g}}$, $\hat{\boldsymbol{\eta}}$ and $\hat{\sigma}_\varepsilon^2$;
  while convergence is not reached do
    $s \leftarrow \{i : |y_i - \boldsymbol{\psi}_i^\top \hat{\mathbf{f}}| \le \epsilon,\ i = 1, \dots, n\}$;
    $\hat{\mathbf{w}}_{-s} \leftarrow \frac{1}{2}|\mathbf{y} - \boldsymbol{\Psi}\hat{\mathbf{f}}|_{-s}^{-1}$;    $\hat{\mathbf{W}}_{-s} \leftarrow \mathrm{diag}(\hat{\mathbf{w}}_{-s})$;
    $\hat{\mathbf{z}}_{-s} \leftarrow \mathbf{y}_{-s} - (1 - 2\tau)|\mathbf{y} - \boldsymbol{\Psi}\hat{\mathbf{f}}|_{-s}$;    $\hat{\mathbf{z}}_s \leftarrow \mathbf{y}_s$;
    Solve $\begin{bmatrix} \frac{1}{n}\boldsymbol{\Psi}_{-s}^\top \hat{\mathbf{W}}_{-s} \boldsymbol{\Psi}_{-s} & \lambda \mathbf{R}_1^\top & \boldsymbol{\Psi}_s^\top \\ \lambda \mathbf{R}_1 & -\lambda \mathbf{R}_0 & \mathbf{O} \\ \boldsymbol{\Psi}_s & \mathbf{O} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{f}} \\ \hat{\mathbf{g}} \\ \hat{\boldsymbol{\eta}} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\boldsymbol{\Psi}_{-s}^\top \hat{\mathbf{W}}_{-s} \mathbf{z}_{-s} \\ \lambda(\mathbf{u} + \boldsymbol{\gamma}) \\ \mathbf{z}_s \end{bmatrix}$;
    $\hat{\sigma}_\varepsilon^2 \leftarrow \frac{1}{n}\mathbf{1}_n^\top \rho_\tau(\mathbf{y} - \boldsymbol{\Psi}\hat{\mathbf{f}}) + \frac{\lambda}{2}\,\hat{\mathbf{g}}^\top \mathbf{R}_0\,\hat{\mathbf{g}}$;
  end while
```

3.4.4 Smoothing parameter selection

A classical issue in penalized nonparametric regression modelling is the selection of the smoothing parameter λ, which controls the amount of regularization enforced on the estimates. In this work we select the value of λ that minimizes the approximated Generalized Cross-Validation (GCV) score. We rely on the definition of GCV for quantile regression problems used by Nychka *et al.* (1995), Yuan (2006), and Li *et al.* (2007), which in our case takes the form

$$\mathrm{GCV}(\lambda) = \sum_{i=1}^{n} \frac{\rho_\tau\{y_i - \hat{f}_h(\mathbf{p}_i)\}}{n - \mathrm{df}}, \tag{3.21}$$

where df denotes a measure of the effective degrees of freedom induced by the smoother $\hat{f}_h$. Notice that the GCV score (3.21) depends on the smoothing parameter through $\hat{f}_h \equiv \hat{f}_h(\lambda)$ and $\mathrm{df} \equiv \mathrm{df}(\lambda)$, which are implicit functions of λ.

In order to derive a closed-form expression for the effective degrees of freedom, as defined by Nychka *et al.* (1995), we first need to provide a convenient representation for the estimated surface at the n observed locations $\hat{\mathbf{f}}_n$. To this purpose we recall that, at the end of the optimization, the discrete quantile estimator $\hat{\mathbf{f}}$ in (3.20) can be written as a linear transformation of the pseudo-data vector $\hat{\mathbf{z}}$. Therefore, $\hat{\mathbf{f}}_n$ is a linear smoother and can be written as

$$\hat{\mathbf{f}}_n = \mathbf{S}\,\hat{\mathbf{z}} + \mathbf{r}, \tag{3.22}$$

for an appropriate smoothing matrix $\mathbf{S}$ and a correspondent residual vector $\mathbf{r}$. In particular, the smoothing matrix is partitioned as $\mathbf{S} = \left[\ \mathbf{S}_{-s},\ \mathbf{S}_s\ \right]$, where

$$\mathbf{S}_{-s} = \underbrace{\boldsymbol{\Psi}(\mathbf{I} - \mathbf{A}^{-1}\boldsymbol{\Psi}_s^\top \mathbf{B}^{-1}\boldsymbol{\Psi}_s)\mathbf{A}^{-1}\boldsymbol{\Psi}_{-s}\mathbf{W}_{-s}}_{n\times(n-|s|)}, \qquad \mathbf{S}_s = \underbrace{\boldsymbol{\Psi}\mathbf{A}^{-1}\boldsymbol{\Psi}_s^\top \mathbf{B}^{-1}}_{n\times|s|}, \tag{3.23}$$

and the residual vector is given by

$$\mathbf{r} = \lambda \boldsymbol{\Psi} \mathbf{A}^{-1} \boldsymbol{\Psi}_s^\top \mathbf{B}^{-1} \boldsymbol{\Psi}_s \mathbf{A}^{-1} \mathbf{R}_1^\top \mathbf{R}_0^{-1} (\mathbf{u} + \boldsymbol{\gamma}). \tag{3.24}$$

In this way we take advantage of the decomposition between unconstrained and constrained terms, namely $\mathbf{S}\hat{\mathbf{z}} + \mathbf{r} = \mathbf{S}_{-s}\hat{\mathbf{z}}_{-s} + \mathbf{S}_s\hat{\mathbf{z}}_s + \mathbf{r}$, generalizing the classical results for restricted least squares estimators (Greene and Seaks, 1991) to penalized smoothing problems.

Because of the linearity of (3.22), we can use the definition of effective degrees of freedom discussed, e.g., in Silverman (1985) and used by, e.g., Nychka *et al.* (1995), which is $\mathrm{df} = \mathrm{tr}(\mathbf{S})$, where $\mathrm{tr}(\cdot)$ is the trace of a matrix. After some simplifications (see Section B.1 of the online supplement material), the equivalent degrees of freedom take the explicit expression

$$\mathrm{df} = |s| + \mathrm{tr}\big\{\mathbf{A}^{-1}(\mathbf{I} - \boldsymbol{\Psi}_s^\top \mathbf{B}^{-1} \boldsymbol{\Psi}_s \mathbf{A}^{-1})(\tfrac{1}{n} \boldsymbol{\Psi}_{-s}^\top \mathbf{W}_{-s}^{-1} \boldsymbol{\Psi}_{-s})\big\}. \tag{3.25}$$

The latter formulation generalizes the definition of degrees of freedom for penalized linear smoothers by Wahba (1990) to penalized linear smoothers subject to equality constraints. Indeed if no active constraints were present, the set $\{i \in \mathbb{N} : 1/w_i = 0, 1 \leq i \leq n\}$ would be empty and, therefore, we would obtain

$$\mathrm{df} = \mathrm{tr}\big\{\mathbf{A}^{-1}(\tfrac{1}{n} \boldsymbol{\Psi}^\top \mathbf{W}^{-1} \boldsymbol{\Psi})\big\},$$

which is actually the standard definition of effective degrees of freedom for penalized weighted regression problems.

3.5 Large sample properties

We here study the asymptotic properties of both the infinite- and finite-dimensional quantile estimators. Doing so, we denote by $f_0 \in \mathcal{F}_\gamma$ the true τ-quantile field of Y_i given $\mathbf{p}_i$ and we define $g_0 = Lf_0 - u \in L^2(\Omega)$ as the misfit of the PDE relative to f_0.

3.5.1 Infinite-dimensional estimator

In order to ensure the identifiability of the estimator in Problem 1 and to establish its large-sample properties, we need to make some assumptions on the probability density function of y_i given $\mathbf{p}_i$ (Assumption 2), on the spatial distribution of the points $\mathbf{p}_i$ over Ω (Assumption 3) and on the asymptotic behavior of $\lambda = \lambda_n$ (Assumption 4) for n going to infinity. We denote by $\Pi_{\mathbf{p}}^n$ the empirical bivariate cumulative density function of the probability measure that assigns mass $1/n$ to each point $\mathbf{p}_i$. Let $\Pi_{\mathbf{p}}$ be the limiting distribution of the sequence $\Pi_{\mathbf{p}}^n$ and let $d_n = \sup_{\mathbf{p} \in \Omega} |\Pi_{\mathbf{p}}(\mathbf{p}) - \Pi_{\mathbf{p}}^n(\mathbf{p})|$ be the maximum difference between $\Pi_{\mathbf{p}}$ and $\Pi_{\mathbf{p}}^n$.

Assumption 2. *There exist* h_1, h_2 *such that, for any* $\mathbf{p} \in \Omega$, $0 < h_1 < \pi_{y|\mathbf{p}}(f_0(\mathbf{p})) < h_2 < \infty$.

Assumption 3. *The sequence d_n converges to* 0 *in probability, i.e.* $\Pi_{\mathbf{p}}^n$ *uniformly converges to* $\Pi_{\mathbf{p}}$*. Moreover,* $\Pi_{\mathbf{p}}$ *has probability density function* $\pi_{\mathbf{p}} \in C^\infty(\bar{\Omega})$ *such that, for all* $\mathbf{p} \in \Omega$, $0 < k_1 < \pi_{\mathbf{p}}(\mathbf{p}) < k_2 < \infty$*, for some constants* k_1, k_2*.*

Assumption 4. *The smoothing parameter* $\lambda = \lambda_n$ *is such that* $d_n/\lambda_n \to 0$ *and* $\lambda_n \to 0$*.*

Assumption 5. *The coefficients of the PDE* (3.2)*, i.e.,* $\mathbf{K}$*,* $\mathbf{b}$ *and* c*, are such that for any forcing term* $u \in L^2(\Omega)$ *there exists a unique solution of the PDE, which, moreover, belongs to* $H^2(\Omega)$

Assumption 2 ensures the existence of a well-behaved asymptotic estimator, preventing the quantile field f_0 lying in a region with almost null probability mass. Assumption 3 guarantees that the locations $\mathbf{p}_1, \dots, \mathbf{p}_n$ cover all the domain Ω with probability 1 for n going to infinity. Assumption 4 establishes the convergence speed of λ_n to 0 as n goes to infinity, in such a way to control the asymptotic behavior of the bias of the estimator. Assumption 5 provide the sufficient regularity conditions to guarantee the equivalence between the norms $\|Lf\|_{L^2}$ and $\|f\|_{H^2}$, which is fundamental to control the bias and variance of the infinite-dimensional estimator in an asymptotic regime.

Under Assumptions 2–5, we can now study the convergence of the asymptotic mean squared error (MSE) of the infinite-dimensional estimator $\hat{f}$. The following theorem shows that the nonparametric quantile estimator $\hat{f}$ is consistent in the L^2 norm under different Sobolev regularity conditions of the underlying true field, $f_0 \in H^2(\Omega)$ and $f_0 \in H^4(\Omega)$. Moreover, the MSE of $\hat{f}$ nearly achieves the optimal convergence rate for nonparametric estimators but for an infinitesimal factor $\delta > 0$, as small as desired.

Theorem 3.7. *Under Assumptions 2–5, if* $f_0 \in H^2(\Omega)$ *and* $Bf_0 = \gamma$*, setting* $\lambda = \lambda_n = n^{-2/3}$*, we have* $\mathrm{MSE}_{L^2}(\hat{f}) = O\big(n^{-2/3+\epsilon}\big)$*, for* $\epsilon > 0$ *as small as desired. If, in addition,* $g_0 \in H^2(\Omega)$*, setting* $\lambda = \lambda_n = n^{-2/5}$*, we have* $\mathrm{MSE}_{L^2}(\hat{f}) = O\big(n^{-4/5+\epsilon/2}\big)$*.*

Proof. See Section B.2 of the online supplementary material of this article. □

The proof of Theorem 3.7 builds upon the work of Arnone *et al.* (2022a), which studied the asymptotic bias, variance and consistency of spatial linear regression with PDE regularization and Dirichlet boundary conditions. Instead, we here generalize such an approach to quantile regression model with PDE regularization and Neumann boundaries.

3.5.2 Finite-dimensional estimator

We denote by $\mathbf{f}_0$ and $\mathbf{g}_0$ the evaluation vectors of f_0 and g_0 at the mesh knots $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{N_h}$. Furthermore, we define the $N_h \times N_h$ matrices

$$\mathbf{D}_{0,n} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{p}_i)\boldsymbol{\psi}(\mathbf{p}_i)^\top, \qquad \mathbf{D}_{1,n} = \frac{1}{n}\sum_{i=1}^{n} \pi_{y_i|\mathbf{p}_i}(f_0(\mathbf{p}_i))\, \boldsymbol{\psi}(\mathbf{p}_i)\boldsymbol{\psi}(\mathbf{p}_i)^\top,$$

where the subscript n highlights the dependence on the sample size. We here assume that the number of bases N_h and the triangulation $\mathcal{T}_h$ are fixed, and that the discretization is sufficiently fine to accurately describe the global and local characteristics of the underlying spatial quantile field.

A sufficient set of regularity conditions that guarantee the convergence of the discretized estimator to its asymptotic distribution is the following.

Assumption 6. *Either the knots of the triangulation are a subset of the data locations, i.e.,* $\{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{N_h}\} \subset \{\mathbf{p}_1, \dots, \mathbf{p}_n\}$*, or, for* n *large enough, there is at least one observation in the support of each basis function* $\psi_1, \dots, \psi_{N_h}$*.*

Assumption 7. *There exist positive definite matrices* $\mathbf{D}_0$ *and* $\mathbf{D}_1$ *such that* $\mathbf{D}_{0,n} \to \mathbf{D}_0$ *and* $\mathbf{D}_{1,n} \to \mathbf{D}_1$ *as* $n \to \infty$*.*

Assumption 6, along with Assumption 2, gives a sufficient condition for the non-singularity of the matrices $\mathbf{D}_{0,n}$ and $\mathbf{D}_{1,n}$. Assumption 7 guarantees the existence of a non-singular limit for the matrices $\mathbf{D}_{0,n}$ and $\mathbf{D}_{1,n}$. Under Assumptions 2, 6, 7, we can now study the convergence of the finite-element estimator $\hat{\mathbf{f}}$ toward $\mathbf{f}_0$ conditionally on the mesh knots $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{N_h}$. The following theorem establishes the asymptotic normality of the finite-dimensional estimator $\hat{\mathbf{f}}$ and it shows that the MSE of $\hat{\mathbf{f}}$ reaches the optimal rate of convergence for parametric estimators.

Theorem 3.8. *Under Assumptions 2, 6 and 7, if* $\lambda = \lambda_n = O(n^{-1/2})$ *with* $\sqrt{n}\,\lambda_n \to \lambda_0$*, then the* τ*-quantile estimator* $\hat{\mathbf{f}}$ *has asymptotic distribution*

$$\sqrt{n}(\hat{\mathbf{f}} - \mathbf{f}_0) \xrightarrow{\mathrm{d}} \mathrm{N}_{N_h}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu} = -\lambda_0 \mathbf{D}_1^{-1} \mathbf{R}_1^\top \mathbf{g}_0$ *and* $\boldsymbol{\Sigma} = \tau(1-\tau)\mathbf{D}_1^{-1}\mathbf{D}_0\,\mathbf{D}_1^{-1}$*. Moreover, under the same assumptions,* $\mathrm{MSE}_n(\hat{\mathbf{f}}) = O(n^{-1})$*. Finally, if* $\lambda = \lambda_n = o(n^{-1/2})$*,* $\sqrt{n}(\hat{\mathbf{f}} - \mathbf{f}_0)$ *is asymptotically unbiased.*

Proof. See Appendix B.2. ☐

Remark 3.9. For n large but finite and $\lambda = \lambda_n = O(n^{-1/2})$, a better approximation of the distribution of $\sqrt{n}(\hat{\mathbf{f}} - \mathbf{f}_0)$ is given by the Gaussian law $N_{N_h}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ with mean and variance

$$\begin{aligned}
\boldsymbol{\mu}_n &= -\sqrt{n}\,\lambda(\mathbf{D}_{1,n} + \lambda\mathbf{P})^{-1}\mathbf{R}_1^\top \mathbf{g}_0, \\
\boldsymbol{\Sigma}_n &= \tau(1-\tau)(\mathbf{D}_{1,n} + \lambda\mathbf{P})^{-1}\mathbf{D}_{0,n}(\mathbf{D}_{1,n} + \lambda\mathbf{P})^{-1}.
\end{aligned}$$

See the supplementary material for more details.

3.6 Inclusion of covariates

The nonparametric quantile regression method described in Section 3.4 of the main paper can be extended to include the effects of space-varying covariates, in a semiparametric regression framework. Let $\mathbf{X}$ be a $n \times q$ design matrix with i-th row

$\mathbf{x}_i^\top = (x_{i1}, \ldots, x_{iq})$, the vector of covariates observed at the location $\mathbf{p}_i$. We then consider the additive model formulation

$$Q_{y_i|\mathbf{x}_i,\mathbf{p}_i}(\tau) = \mathbf{x}_i^\top \boldsymbol{\beta} + f(\mathbf{p}_i), \quad i = 1, \ldots, n,$$

where $\boldsymbol{\beta} \in \mathbb{R}^q$ is an unknown vector of regression coefficients. We can obtain the semiparametric quantile estimator $(\hat{\boldsymbol{\beta}}, \hat{f})$ by minimizing the functional

$$J_{\tau,\lambda}(\boldsymbol{\beta}, f) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\{y_i - \mathbf{x}_i^\top\boldsymbol{\beta} - f(\mathbf{p}_i)\} + \frac{\lambda}{2}\int_\Omega (Lf - u)^2,$$

with respect to $\boldsymbol{\beta} \in \mathbb{R}^q$ and $f \in \mathcal{F}_\gamma$.

The EM algorithm proposed in Section 3.3 can be adapted to this new model specification. Aside from the presence of an additional component in the linear predictor, the E-step does not change, as well as the constrained formulation of the M-step. Defining $\mathbf{z}^{(k)} = \mathbf{y} - (1-2\tau)|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)} - \mathbf{f}_n^{(k)}|$, where $1/\mathbf{w}^{(k)} = 2|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k)} - \mathbf{f}_n^{(k)}|$, the estimator of $\boldsymbol{\beta}$ and f at a new iteration of the algorithm is updated by optimizing with respect to $\boldsymbol{\beta}$ and $\mathbf{f}$ the quadratic functional

$$J_{-s}^{(k)}(\boldsymbol{\beta}, f) = \frac{1}{n}(\mathbf{z}^{(k)} - \mathbf{X}\boldsymbol{\beta} - \mathbf{f}_n)_{-s}^\top \mathbf{W}_{-s}^{(k)} (\mathbf{z}^{(k)} - \mathbf{X}\boldsymbol{\beta} - \mathbf{f}_n)_{-s} + \lambda \int_\Omega (Lf - u),$$

under the constraint $(\mathbf{z}^{(k)} - \mathbf{X}\boldsymbol{\beta} - \mathbf{f}_n)_s = 0$, where $\mathbf{W}_{-s}^{(k)} = \operatorname{diag}(\mathbf{w}_{-s}^{(k)})$. The solution of such a minimization problem is then characterized by the following proposition.

Proposition 3.10. *Let* $(\tilde{\boldsymbol{\beta}}, \tilde{f}, \tilde{\boldsymbol{\eta}}) \in \mathbb{R}^q \times \mathcal{F}_\gamma \times \mathbb{R}^{|s|}$ *be a minimum of the Lagrangian functional*

$$\mathcal{L}^{(k)}(\boldsymbol{\beta}, f, \boldsymbol{\eta}) = J_{-s}^{(k)}(\boldsymbol{\beta}, f) + \boldsymbol{\eta}^\top(\mathbf{z}^{(k)} - \mathbf{X}\boldsymbol{\beta} - \mathbf{f}_n)_s, \tag{3.26}$$

and let $\tilde{g} = L\tilde{f} - u \in H^1(\Omega)$. *Then* $(\tilde{\boldsymbol{\beta}}, \tilde{f}, \tilde{g}, \tilde{\boldsymbol{\eta}})$ *must satisfy the following system of first order equations*

$$\begin{aligned} -\tfrac{1}{n}\mathbf{X}_{-s}^\top \mathbf{W}_{-s}^{(k)}(\mathbf{z}^{(k)} - \mathbf{X}\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{f}}_n)_{-s} - \mathbf{X}_s^\top \tilde{\boldsymbol{\eta}} &= 0, \\ -\tfrac{1}{n}(\boldsymbol{\psi}_n)_{-s}^\top \mathbf{W}_{-s}^{(k)}(\mathbf{z}^{(k)} - \mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{f}_n)_{-s} - (\boldsymbol{\psi}_n)_s^\top \tilde{\boldsymbol{\eta}} + \lambda R_1(\psi, \tilde{g}) &= 0, \\ R_1(\tilde{f}, \phi) - R_0(\tilde{g}, \phi) &= F(\phi), \\ (\mathbf{z}^{(k)} - \mathbf{X}\boldsymbol{\beta} - \mathbf{f}_n)_s &= \mathbf{0}, \end{aligned} \tag{3.27}$$

for any pair of test functions $\phi, \psi \in \mathcal{F}_0$.

Proof. Let $\mathcal{L}(\boldsymbol{\beta}, f, \boldsymbol{\eta}) = J_{-s}(\boldsymbol{\beta}, f) + \boldsymbol{\eta}^\top(\mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{f}_n)_s$ be the Lagrangian functional defined in (3.26). Let $(\tilde{\boldsymbol{\beta}}, \tilde{f}, \tilde{\boldsymbol{\eta}}) \in \mathbb{R}^q \times \mathcal{F}_\gamma \times \mathbb{R}^{|s|}$ be a minimizer of $\mathcal{L}(\boldsymbol{\beta}, f, \boldsymbol{\eta})$. Then, $(\tilde{\boldsymbol{\beta}}, \tilde{f}, \tilde{\boldsymbol{\eta}})$ must satisfy the first order equations

$$\frac{\partial}{\partial \boldsymbol{\eta}}\mathcal{L}(\tilde{\boldsymbol{\beta}}, \tilde{f}, \tilde{\boldsymbol{\eta}}) = 0, \qquad \frac{\partial}{\partial \boldsymbol{\beta}}\mathcal{L}(\tilde{\boldsymbol{\beta}}, \tilde{f}, \tilde{\boldsymbol{\eta}}) = 0, \qquad \frac{\partial}{\partial t}\mathcal{L}(\tilde{\boldsymbol{\beta}}, \tilde{f} + t\psi, \tilde{\boldsymbol{\eta}})\bigg|_{t=0} = 0,$$

for any $\psi \in \mathcal{F}_0$. The first two equations take the form

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\eta}} \mathcal{L}(\tilde{\boldsymbol{\beta}}, \tilde{f}, \tilde{\boldsymbol{\eta}}) &= (\mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{f}_n)_s = \mathbf{0}, \\
\frac{\partial}{\partial \boldsymbol{\beta}} \mathcal{L}(\tilde{\boldsymbol{\beta}}, \tilde{f}, \tilde{\boldsymbol{\eta}}) &= -\tfrac{1}{n} \mathbf{X}_{-s}^\top \mathbf{W}_{-s} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{f}_n)_{-s} - \mathbf{X}_s^\top \tilde{\boldsymbol{\eta}} = \mathbf{0}.
\end{aligned}$$

Moreover, thanks to the Lax-Milgram lemma (see, e.g., Quarteroni, 2017), if Assumption 2 in Azzimonti *et al.* (2014) is satisfied, the Lagrangian functional $\mathcal{L}(\boldsymbol{\eta}, f, \boldsymbol{\beta})$ has a unique minimizer, which satisfies the Euler-Lagrange equation

$$-\tfrac{1}{n} (\boldsymbol{\psi}_n)_{-s}^\top \mathbf{W}_{-s} (\mathbf{z} - \mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{f}}_n)_{-s} - (\boldsymbol{\psi}_n)_{-s}^\top \tilde{\boldsymbol{\eta}} + \lambda \int_\Omega (L\psi)(L\tilde{f} - u) = 0,$$

for any $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\eta}}) \in \mathbb{R}^q \times \mathbb{R}^{|s|}$ and $\psi \in \mathcal{F}_0$. This is equivalent to the system

$$\begin{aligned}
-\tfrac{1}{n} (\boldsymbol{\psi}_n)_{-s}^\top \mathbf{W}_{-s} (\mathbf{z} - \mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{f}}_n)_{-s} + \lambda \int_\Omega (L\psi)\tilde{g} &= (\boldsymbol{\psi}_n)_{-s}^\top \tilde{\boldsymbol{\eta}} + \lambda \int_\Omega (L\psi) u, \\
\int_\Omega (L\tilde{f})\phi - \int_\Omega \tilde{g}\phi &= \int_\Omega u\phi,
\end{aligned}$$

for any pair of test functions $\phi, \psi \in \mathcal{F}_0$. Hence, as in the proof of Proposition 3.3, we can integrate by parts and use the definition of $R_1(\cdot,\cdot)$, $R_0(\cdot,\cdot)$ and $F(\cdot)$ in (3.14) to obtain the weak variational formulation in Proposition 3.10. This concludes the proof. ☐

Analogously to the pure non-parametric case, after introducing the finite element discretization, the variational formulation (3.27) in Proposition 3.10 can be approximated by the linear system

$$\begin{bmatrix}
\frac{1}{n}\mathbf{X}_{-s}^\top \mathbf{W}_{-s}^{(k)} \mathbf{X}_{-s} & \frac{1}{n}\mathbf{X}_{-s}^\top \mathbf{W}_{-s}^{(k)} \boldsymbol{\Psi}_{-s} & \mathbf{O} & \mathbf{X}_s^\top \\
\frac{1}{n}\boldsymbol{\Psi}_{-s}^\top \mathbf{W}_{-s}^{(k)} \mathbf{X}_{-s} & \frac{1}{n}\boldsymbol{\Psi}_{-s}^\top \mathbf{W}_{-s}^{(k)} \boldsymbol{\Psi}_{-s} & \lambda \mathbf{R}_1^\top & \boldsymbol{\Psi}_s^\top \\
\mathbf{O} & \lambda \mathbf{R}_1 & -\lambda \mathbf{R}_0 & \mathbf{O} \\
\mathbf{X}_s & \boldsymbol{\Psi}_s & \mathbf{O} & \mathbf{O}
\end{bmatrix}
\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{f}} \\ \tilde{\mathbf{g}} \\ \tilde{\boldsymbol{\eta}} \end{bmatrix}
=
\begin{bmatrix}
\frac{1}{n}\mathbf{X}_{-s}^\top \mathbf{W}_{-s}^{(k)} \mathbf{z}_{-s}^{(k)} \\
\frac{1}{n}\boldsymbol{\Psi}_{-s}^\top \mathbf{W}_{-s}^{(k)} \mathbf{z}_{-s}^{(k)} \\
\lambda(\mathbf{u} + \boldsymbol{\gamma}) \\
\mathbf{z}_s^{(k)}
\end{bmatrix}. \tag{3.28}$$

The resulting EM routine based on such a representation is summarized in the pseudocode description of Algorithm 6.

In order to present the explicit solution of system (3.28), we introduce the completed design matrix $\bar{\boldsymbol{\Psi}} = \left[\mathbf{X},\, \boldsymbol{\Psi}\right]$. Further, let $\bar{\mathbf{A}}^{(k)} = \bar{\boldsymbol{\Psi}}_{-s}^\top \mathbf{W}_{-s}^{(k)} \bar{\boldsymbol{\Psi}}_{-s} + \lambda \bar{\mathbf{P}}$, $\bar{\mathbf{B}}^{(k)} = \bar{\boldsymbol{\Psi}}_s (\bar{\mathbf{A}}^{(k)})^{-1} \bar{\boldsymbol{\Psi}}_s^\top$ and $\bar{\mathbf{d}}^{(k)} = \bar{\boldsymbol{\Psi}}_{-s} \mathbf{W}_{-s}^{(k)} \mathbf{z}_{-s}^{(k)} + \lambda \bar{\mathbf{h}}$, where

$$\bar{\mathbf{P}} = \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{R}_1^\top \mathbf{R}_0^{-1} \mathbf{R}_1 \end{bmatrix}, \qquad \bar{\mathbf{h}} = \begin{bmatrix} \mathbf{0} \\ \mathbf{R}_1^\top \mathbf{R}_0^{-1} (\mathbf{u} + \boldsymbol{\gamma}) \end{bmatrix}, \tag{3.29}$$

Algorithm 6 Functional EM algorithm for semiparametric spatial quantile regression

```
Require: $\tau, \lambda, \mathbf{y}, \boldsymbol{\Psi}, \mathbf{R}_0, \mathbf{R}_1$
  Initialize $\hat{\mathbf{f}}$, $\hat{\mathbf{g}}$, $\hat{\boldsymbol{\eta}}$ and $\hat{\sigma}_\varepsilon^2$;
  while convergence is not reached do
    $s \leftarrow \{i : |y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} - \boldsymbol{\psi}_i^\top \hat{\mathbf{f}}| \leq \epsilon,\ i = 1, \dots, n\}$;
    $\hat{\mathbf{w}}_{-s} \leftarrow \frac{1}{2}\, |\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \boldsymbol{\Psi}\hat{\mathbf{f}}|_{-s}^{-1}; \quad \hat{\mathbf{W}}_{-s} \leftarrow \operatorname{diag}(\hat{\mathbf{w}}_{-s})$;
    $\hat{\mathbf{z}}_{-s} \leftarrow \mathbf{y}_{-s} - (1 - 2\tau)|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \boldsymbol{\Psi}\hat{\mathbf{f}}|_{-s}; \quad \hat{\mathbf{z}}_s \leftarrow \mathbf{y}_s$;
    $\mathbf{A}_{11} \leftarrow \begin{bmatrix} \frac{1}{n}\mathbf{X}_{-s}^\top \hat{\mathbf{W}}_{-s}\mathbf{X}_{-s} & \frac{1}{n}\mathbf{X}_{-s}^\top \hat{\mathbf{W}}_{-s}\boldsymbol{\Psi}_{-s} \\ \frac{1}{n}\boldsymbol{\Psi}_{-s}^\top \hat{\mathbf{W}}_{-s}\mathbf{X}_{-s} & \frac{1}{n}\boldsymbol{\Psi}_{-s}^\top \hat{\mathbf{W}}_{-s}\boldsymbol{\Psi}_{-s} \end{bmatrix}; \quad \mathbf{b}_2 \leftarrow \begin{bmatrix} \frac{1}{n}\mathbf{X}_{-s}^\top \hat{\mathbf{W}}_{-s}\mathbf{z}_{-s} \\ \frac{1}{n}\boldsymbol{\Psi}_{-s}^\top \hat{\mathbf{W}}_{-s}\mathbf{z}_{-s} \end{bmatrix}$;
    $\mathbf{A}_{21} \leftarrow \begin{bmatrix} \mathbf{O} & \lambda\mathbf{R}_1 \\ \mathbf{X}_s & \boldsymbol{\Psi}_s \end{bmatrix}; \quad \mathbf{A}_{22} \leftarrow \begin{bmatrix} -\lambda\mathbf{R}_0 & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}; \quad \mathbf{b}_2 \leftarrow \begin{bmatrix} \lambda(\mathbf{u} + \boldsymbol{\gamma}) \\ \mathbf{z}_s \end{bmatrix}$;
    Solve $\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{21}^\top \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}; \quad \hat{\boldsymbol{\beta}} \leftarrow \boldsymbol{x}_1; \quad \hat{\mathbf{f}} \leftarrow \boldsymbol{x}_2; \quad \hat{\mathbf{g}} \leftarrow \boldsymbol{x}_3; \quad \hat{\boldsymbol{\eta}} \leftarrow \boldsymbol{x}_4$;
    $\hat{\sigma}_\varepsilon^2 \leftarrow \frac{1}{n}\mathbf{1}_n^\top \rho_\tau(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \boldsymbol{\Psi}\hat{\mathbf{f}}) + \frac{\lambda}{2}\, \hat{\mathbf{g}}^\top \mathbf{R}_0\, \hat{\mathbf{g}}$;
  end while
```

are the completed penalty matrix and vector, respectively. Then, similarly to the pure nonparametric finite element estimator (3.20), we have

$$\begin{aligned} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{f}} \end{bmatrix} &= (\bar{\mathbf{A}}^{(k)})^{-1}\bar{\mathbf{d}}^{(k)} + (\bar{\mathbf{A}}^{(k)})^{-1}\bar{\boldsymbol{\Psi}}_s^\top (\bar{\mathbf{B}}^{(k)})^{-1}\{\mathbf{z}_s^{(k)} - \bar{\boldsymbol{\Psi}}_s(\bar{\mathbf{A}}^{(k)})^{-1}\bar{\mathbf{d}}^{(k)}\}, \\ \tilde{\mathbf{g}} = \mathbf{R}_0^{-1}(\mathbf{R}_1\tilde{\mathbf{f}} - \mathbf{u} - \boldsymbol{\gamma}), &\qquad \tilde{\boldsymbol{\eta}} = (\bar{\mathbf{B}}^{(k)})^{-1}\{\mathbf{z}_s^{(k)} - \bar{\mathbf{C}}(\bar{\mathbf{A}}^{(k)})^{-1}\bar{\mathbf{d}}^{(k)}\}. \end{aligned} \tag{3.30}$$

If $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are non-singular matrices, the discrete estimator solving the linear system (3.28) exists and is unique (Nocedal and Wright, 2006, Chapter 16).

Similarly to what proposed in Section 3.4.4, the smoothing parameter λ can be selected by minimizing the GCV score, which in this case is given by

$$\mathrm{GCV}(\lambda) = \sum_{i=1}^{n} \frac{\rho_\tau\{y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} - \hat{f}_h(\mathbf{p}_i)\}}{n - \mathrm{df}},$$

with effective degrees of freedom df_c. As a consequence of the linearity of the estimator $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{f}}) = (\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{f}})$ in (3.30) with respect to $\mathbf{z}$, we have

$$\mathrm{df} = |s| + \mathrm{tr}\{\bar{\mathbf{A}}^{-1}(\mathbf{I} - \bar{\boldsymbol{\Psi}}_s^\top \bar{\mathbf{B}}^{-1}\bar{\boldsymbol{\Psi}}_s\bar{\mathbf{A}}^{-1})(\tfrac{1}{n}\bar{\boldsymbol{\Psi}}_{-s}^\top \mathbf{W}_{-s}^{-1}\boldsymbol{\Psi}_{-s})\}.$$

The derivation of df_c is analogous to what presented in Section 3.4.4, and further detailed in Appendix B.1 for the pure non-parametric model.

3.7 Simulation studies

In this section we present the results of two simulation studies, in order to assess the performance of the proposed method and to compare it with alternative existing approaches. We consider heteroscedastic data generating models, defined as

$$y_i \sim \mathrm{N}(\mu_i, \sigma_i^2), \quad \mu_i = \mu(\mathbf{p}_i), \quad \sigma_i = \sigma(\mathbf{p}_i), \quad \mathbf{p}_i \in \Omega, \quad i = 1, \ldots, n, \tag{3.31}$$

with a spatial mean surface $\mu : \Omega \to \mathbb{R}$ and a standard deviation surface $\sigma : \Omega \to \mathbb{R}_+$. In the first simulation, we consider three combinations of mean and standard deviation surfaces defined over a non trivial horseshoe domain. In the second simulation setup we consider an anisotropic specification of the mean and standard deviation fields, defined upon a simple square domain.

In each scenario, we independently simulate 100 datasets with 500 observations each, according to the generative model (3.31). We then estimate 5 quantile surfaces, corresponding to levels $\alpha = 0.1, 0.25, 0.5, 0.75, 0.95$. We compare

- SQR-PDE: the proposed spatial quantile regression with the second order differential regularization $\int_\Omega \{\mathrm{div}(\mathbf{K}\nabla f)\}^2$ and homogeneous Neumann boundary conditions;
- SOAP: quantile version of the soap film smoothing by Wood *et al.* (2008) and Fasiolo *et al.* (2021a);
- H-SOAP: quantile soap film smoothing with location-scale calibration for the heteroscedasticity by Fasiolo *et al.* (2021a);
- TPS: quantile version of thin plate spline smoothing by Wood (2003) and Fasiolo *et al.* (2021a);
- H-TPS: quantile thin plate spline smoothing with location-scale calibration for the heteroscedasticity by (Fasiolo *et al.*, 2021a);
- QSS: quantile smoothing splines with total variation regularization by Koenker and Mizera (2004).

SQR-PDE is implemented using the `R` parkage `fdaPDE` (Arnone *et al.*, 2022b); SOAP, H-SOAP, TPS and H-TPS are all implemented in the `R` package `qgam` (Fasiolo *et al.*, 2021b); QSS is implemented in the `R` package `quantreg` (Koenker, 2021).

The methods are compared in terms of root mean squared error (RMSE), computed as RMSE $= \left\{ \int_\Omega (f - \hat{f})^2 / |\Omega| \right\}^{1/2}$, where the integral is approximated by a sum over a fine regular grid that covers the domain Ω, and $|\Omega|$ is the area of Ω.

3.7.1 First simulation setup: field over irregular domain

In our first simulation setup we consider three scenarios, to which we refer as A, B, C, respectively; where the data are generated according to the heteroscedastic Gaussian

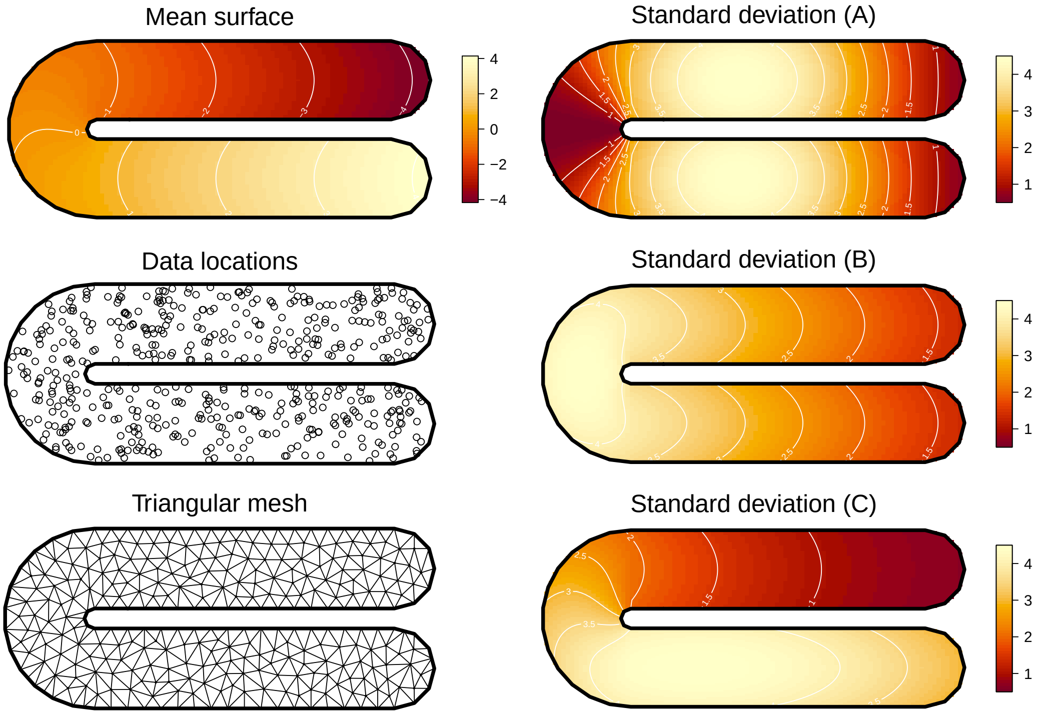

Figure 3.2: Horseshoe domain for the first simulation study. First row: mean surface, data locations, triangular discretization of the domain. Second row: standard deviation surfaces corresponding to three simulation settings A, B and C.

model (3.31), defined over the horseshoe domain proposed by Ramsay (2002). For the mean surface $\mu(\cdot)$ we consider the test function proposed by Wood *et al.* (2008), while a different standard deviation surface $\sigma(\cdot)$ is used for each scenario. The true mean and standard deviation fields are shown in Figure 3.2.

We estimate the quantile fields using the methods mentioned before. Since no anisotropy or flows are appreciable in the synthetic data, we consider an isotropic Laplacian regularization for the SQR-PDE model, i.e. $\mathbf{K} = \mathbf{I}$. The results for the three scenarios and the five quantile levels are shown in Figure 3.3 and 3.4. The proposed method (SQR-PDE) has comparable or better performances in terms of RMSE than SOAP in each simulation setting (Figure 3.3). Moreover, these methods outperform TPS and QSS. This effect is due to a proper management of the domain constraints of SQR-PDE and SOAP. TPS and QSS, instead, smooth the estimated quantile surface not only within the domain but also across the boundaries, causing undesired over-smoothing effects in a neighborhood of the interior boundary (Figure 3.4).

3.7.2 Second simulation setup: anisotropic field over regular domain

In the second simulation setup, we generate $\mu(\cdot)$ and $\sigma(\cdot)$ as Gaussian random fields with anisotropic Matérn covariance function over the square domain $\Omega = [0, 1]^2$, parametrized

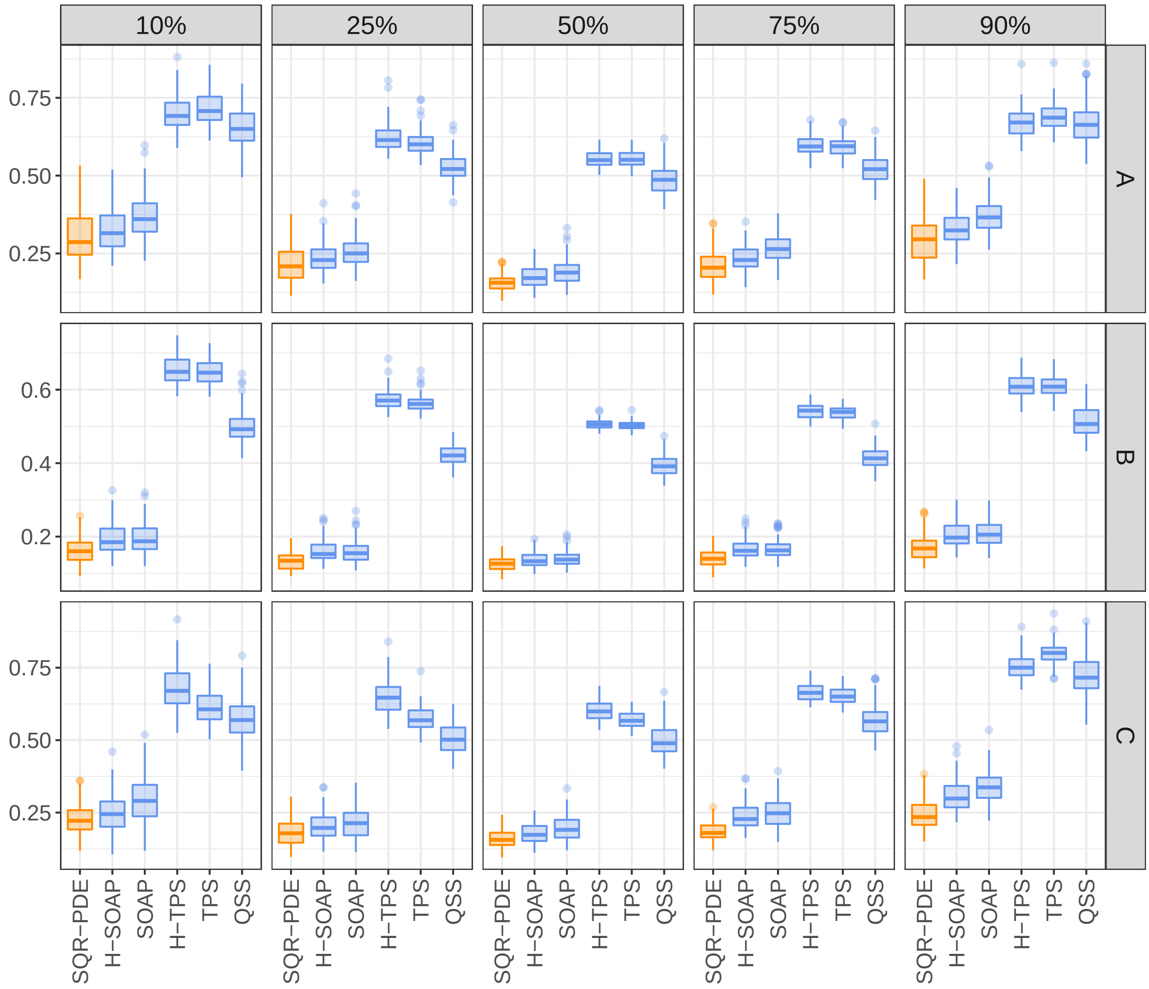

FIGURE 3.3: Boxplots of the RMSE over 100 replicates of the estimated spatial quantile fields in the first simulation setup. Each column corresponds to a quantile level (10%, 25%, 50%, 75%, 90%). Each row corresponds to a simulation setting (A, B, C).

as

$$C(d_{ij};\tau,\rho,\nu) = \tau^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{2\sqrt{\nu}\, d_{ij}}{\rho} \right)^{\nu} K_{\nu}\left(\frac{2\sqrt{\nu}\, d_{ij}}{\rho} \right),$$

where $d_{ij}^2 = (\mathbf{p}_i - \mathbf{p}_j)^\top \mathbf{A}\, (\mathbf{p}_i - \mathbf{p}_j)$ is the squared Mahalanobis distance between $\mathbf{p}_i$ and $\mathbf{p}_j$ calculated with respect to the anisotropy tensor $\mathbf{A} \in \mathbb{R}^{2\times 2}$, which determines the direction and intensity of the anisotropic effect; the parameters involved in the covariance function are, respectively, the marginal variance $\tau^2 > 0$, the covariance range $\rho > 0$, and the smoothing parameter $\nu > 0$. In particular, for the mean surface, we set $\tau^2 = 1$, $\rho = 0.3$, and $\nu = 2$. For the standard deviation surface, we set $\tau^2 = 0.3$, $\rho = 0.6$, and $\nu = 2$. The anisotropic tensor is specified according to the form: $A_{ij} = 1$ if $i = j$ and $A_{ij} = \phi$ if $i \neq j$, where the parameter $\phi \in (-1, +1)$ control the direction and intensity of the anisotropic effect. Specifically, we consider three scenarios, A, B and C, with an increasing level of anisotropy, namely $\phi_{\mathrm{A}} = 0$, $\phi_{\mathrm{B}} = -0.3$, $\phi_{\mathrm{C}} = -0.6$.

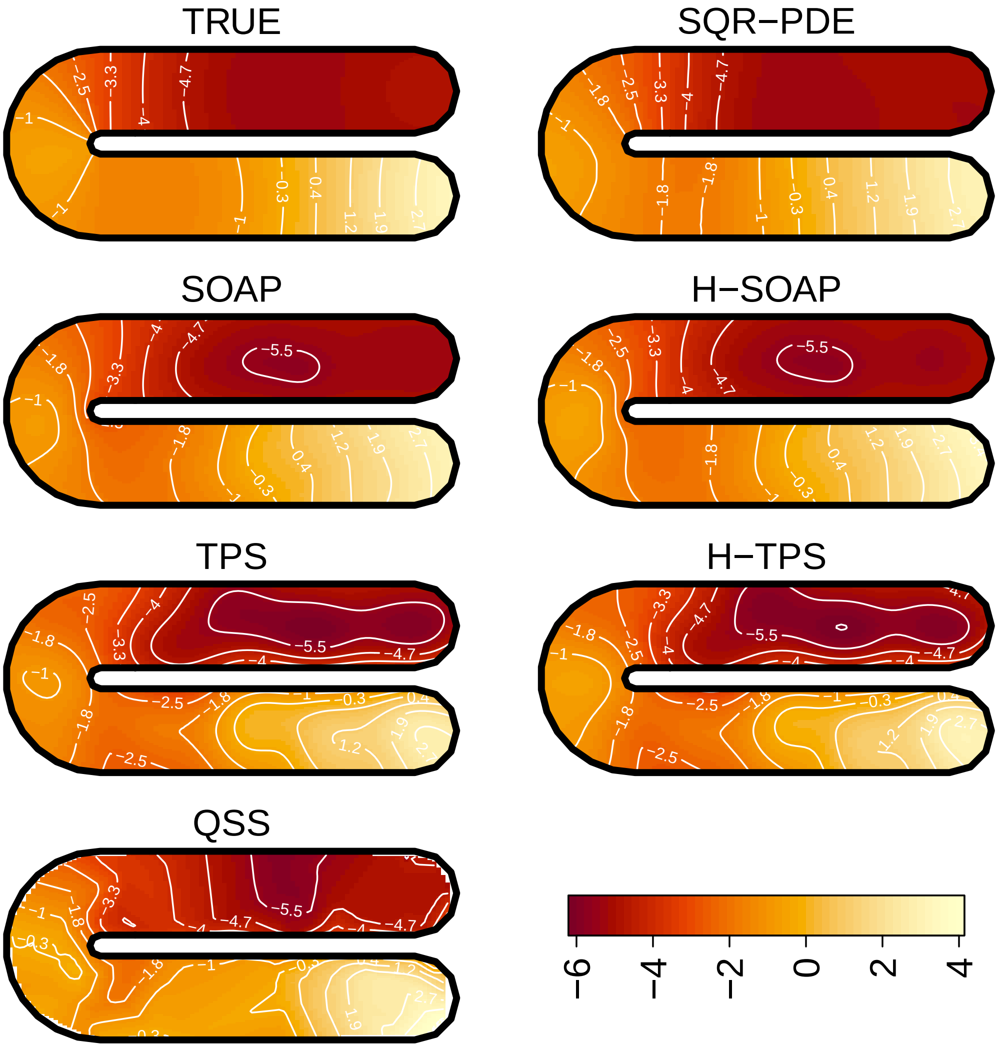

FIGURE 3.4: Estimated 10%-quantile surfaces for one dataset generated in the first simulation setup, setting C.

We use the function `RFsimulate` of the `R` package `RandomFields` (Schlather *et al.*, 2017) to simulate the mean and standard deviation fields on a fine grid covering Ω (see Figure 3.5). Estimation of quantile fields is performed with the same methods considered in the first simulation setup. In particular, for SQR-PDE, we consider two specifications for the differential regularization:

- SQR-PDE($\mathbf{I}$): spatial quantile regression with isotropic Laplacian regularization, i.e., $\mathbf{K} = \mathbf{I}$;
- SQR-PDE($\mathbf{K}$): spatial quantile regression with anisotropic diffusion regularization, i.e. $\mathbf{K} \neq \mathbf{I}$.

In order to determine the optimal diffusion tensor $\mathbf{K}$ for SQR-PDE($\mathbf{K}$), we use the parameter cascading algorithm proposed by Bernardi *et al.* (2018) in the context of spatial

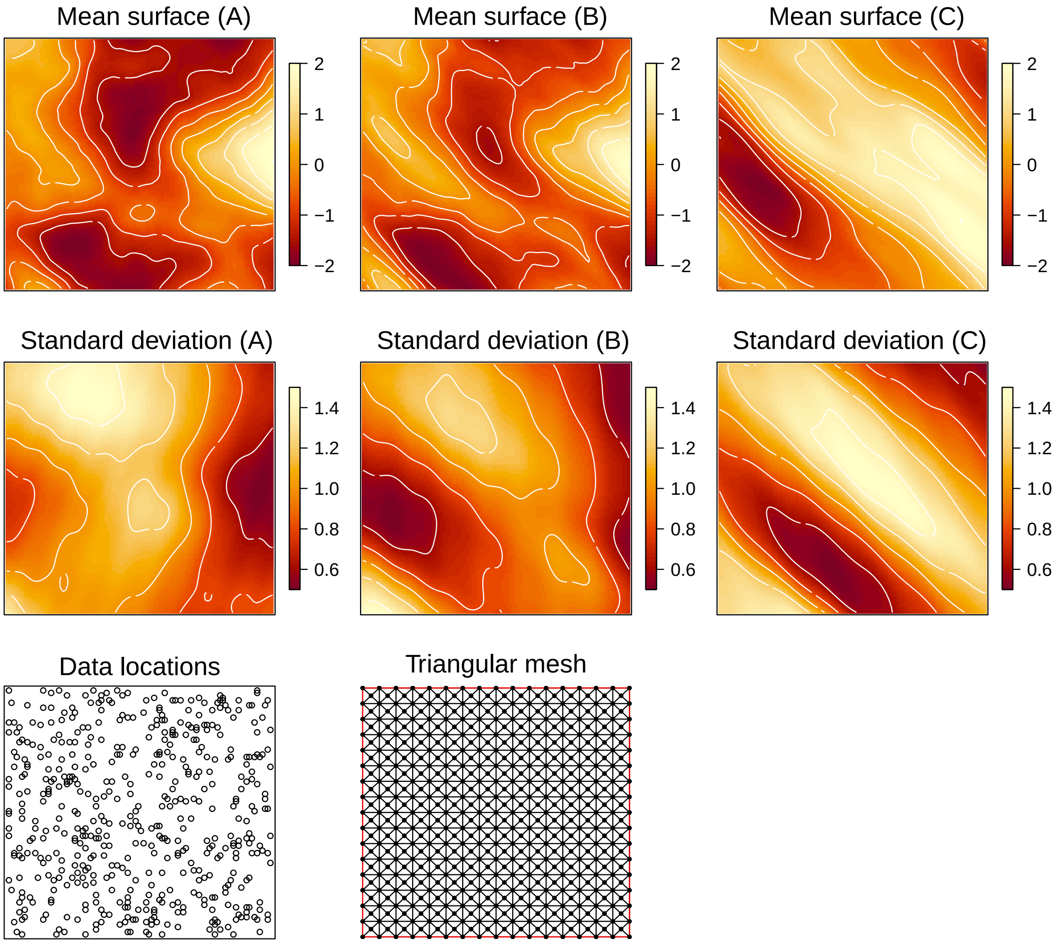

FIGURE 3.5: Square domain for the second simulation setup. First row: mean surfaces in settings A, B, C. Second row: standard deviation surfaces in settings A, B, C. Third row: data locations and triangular discretization of the domain.

smoothing with anisotropic PDE regularization. The complete estimation procedure for the anisotropic quantile model then involves two steps. In the first one, we estimate the diffusion tensor $\mathbf{K}$ by minimizing a squared error loss criterion (Bernardi *et al.*, 2018) using the R package `fdaPDE` (Arnone *et al.*, 2022b). At this stage of the procedure no quantile regression is considered, assuming that the anisotropic effect is homogeneous over the mean and all quantile surfaces. In the second step, we estimate the quantile field using the anisotropic penalty induced by the tensor $\mathbf{K}$, estimated in the first step.

The RMSE for the considered scenarios, quantile levels, and estimating methods are shown in Figure 3.6. As anisotropy increases, SQR-PDE models improve their performances compared with other methods. In scenario A, where no anisotropy is present, both anisotropic and isotropic SQR-PDE perform similarly to other methods, and slightly worse than SOAP and TPS. When some moderate anisotropic effects are present, like in scenario B, the RMSE of SQR-PDE improves and becomes comparable with SOAP and TPS. Moreover, in this scenario, anisotropic SQR-PDE always reaches a lower error than isotropic SQR-PDE. Such an effect is more evident when the anisotropy is stronger, like in scenario C. Here, anisotropic SQR-PDE performs significantly better

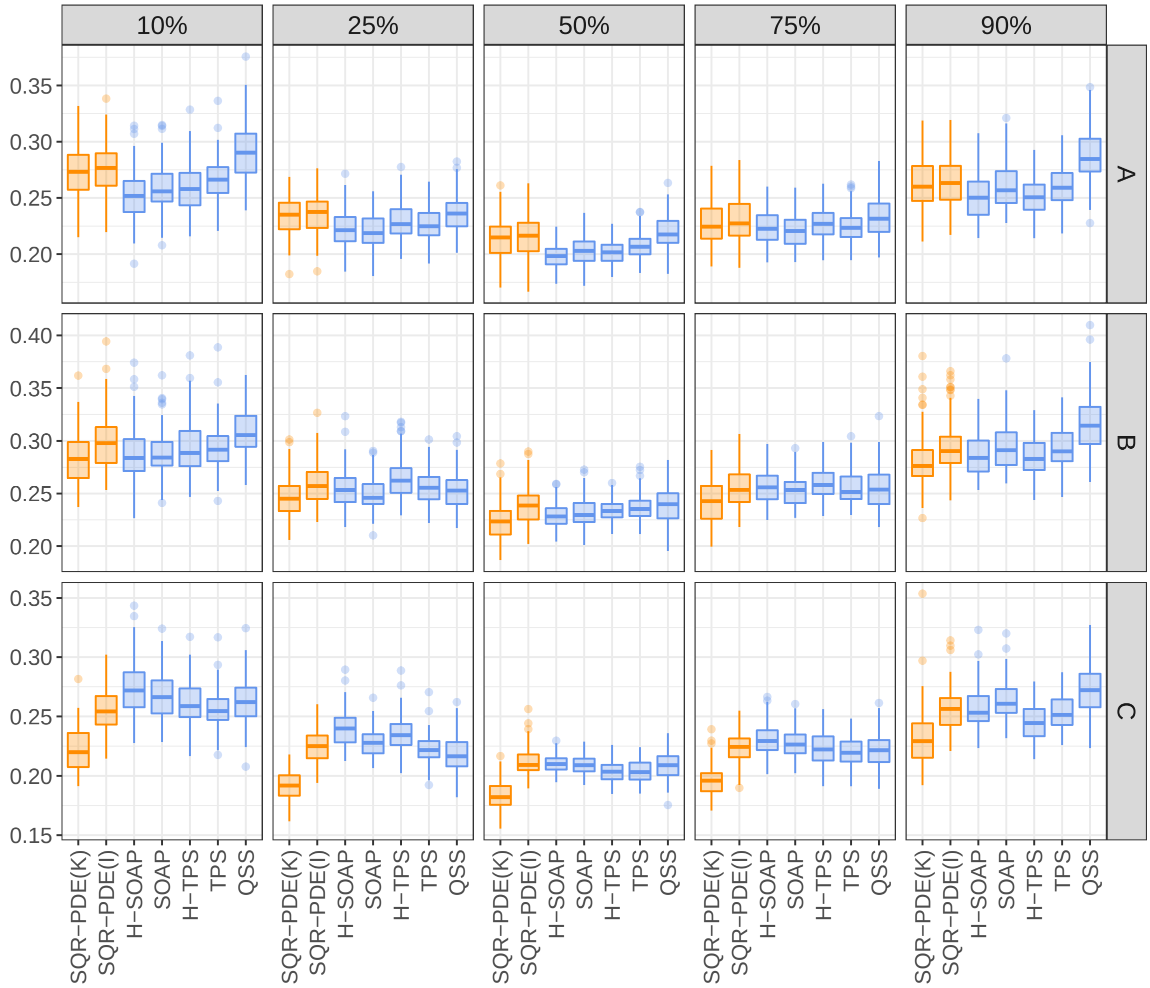

FIGURE 3.6: Boxplots of the RMSE over 100 replicates of the estimated spatial quantile fields in the second simulation setup. Each column corresponds to a quantile level (10%, 25%, 50%, 75%, 90%). Each row corresponds to a simulation setting (A, B, C). The colors correspond to different families of estimating methods.

than all other methods, confirming that a careful specification of the PDE regularization may have a relevant impact on model fit, especially when the phenomenon under study presents non-trivial spatial characteristics. Figure 3.7 shows the estimated quantile fields for one dataset generated under simulation setting C.

In Section 3.8 we adopt an anisotropic SQR-PDE model for the analysis of a real data example, which is characterized by a strong anisotropic effect and a heteroscedastic, locally skewed spatial distribution.

3.8 Switzerland rainfall data

We apply the proposed spatial quantile regression method to the Switzerland rainfall data (Figure 3.8), which collect 467 rainfall measurements (in 1/10 mm units) recorded

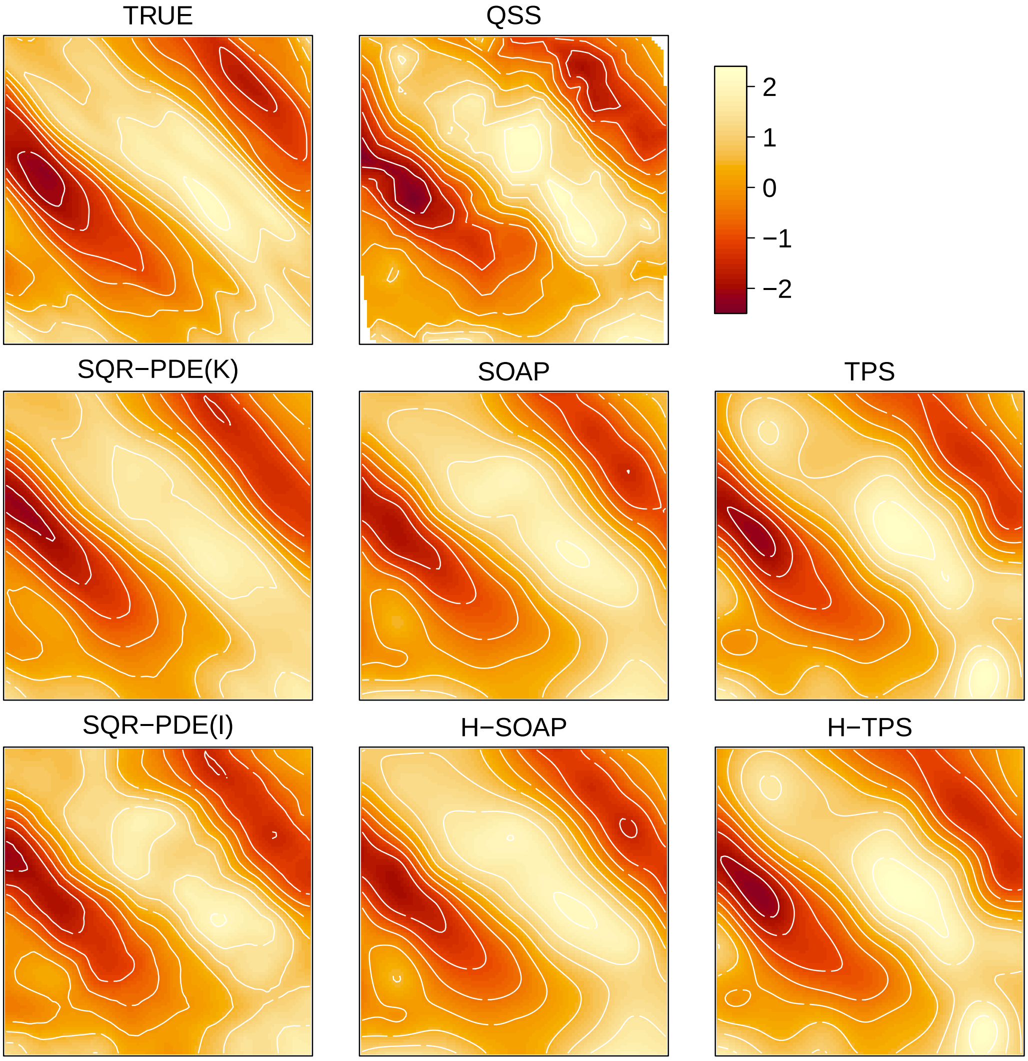

FIGURE 3.7: Estimated 90%-quantile surfaces for one dataset generated in the second simulation setup, setting C.

on May 8, 1986. This dataset has already been analyzed in, e.g., Dubois *et al.* (2003) and Bernardi *et al.* (2018). In their works the authors estimated the underlying spatial field employing several geostatistical approaches and taking into account the macroscopic south-west north-east directional anisotropy characterizing the spatial mean of the underlying data generating process. The cited works are restricted to mean field estimation. In our work, we instead explore quantile levels, therefore obtaining additional insights and exploring areas of exceedingly high or low precipitation.

Indeed, meteorological data are not solely characterized by a spatially varying mean. Precipitations, temperatures, pressure, humidity, and possibly other climatic variables, often present distributions manifesting heteroscedasticity and local skewness, as well as fat tails and extreme values. This is the case also for the Switzerland rainfall data, for which we can observe a highly heteroscedastic spatial distribution, where regions showing higher precipitation levels also have higher variability. For the sake of illustration, we replicate here the analysis proposed by Bernardi *et al.* (2018), which developed an

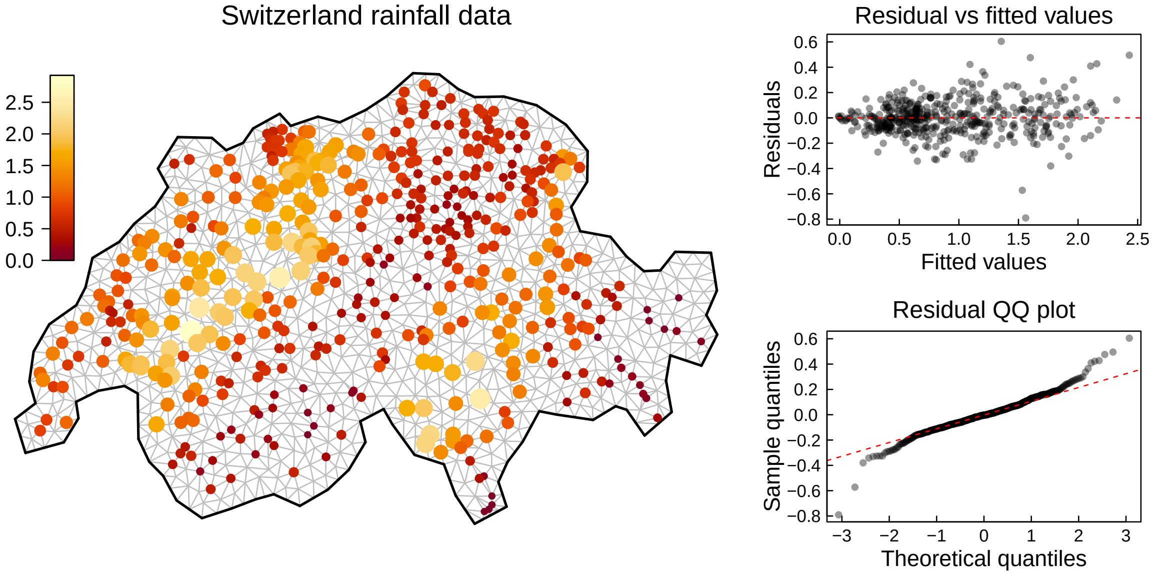

FIGURE 3.8: Left: Switzerland rainfall data and triangular discretization of the domain. Colored dots indicate the measurement locations. Color and dimension of the dots are proportional to the rainfall intensity. Right: residual versus fitted value plot and normal quantile-quantile plot of the residuals. The fitted values and the residuals are obtained by using the anisotropic spatial regression model by Bernardi *et al.* (2018).

anisotropic spatial regression model based on a PDE regularization term. The results are showed in Figure 3.8. We observe an increase of the residual variability as the estimated mean grows, which indeed provides evidence against the usual hypothesis of homoscedastic errors. Furthermore, the normal quantile-quantile plot of the residuals highlights strong deviations from the hypothesis of normal errors, since the distribution of the residuals has heavier tails than the Gaussian probability law. Similar interpretations arise also when adopting different models for the underlying spatial field, like spatial regression based on kriging, radial basis expansions, and neural networks, among others. See Dubois *et al.* (2003) for a detailed discussion on the usage of these and other methods in the analysis of the Switzerland rainfall data.

The local specification of quantile regression makes it appropriate in the presence of heteroscedasticity. Furthermore, in this data analysis it is clear that our interest lies more in the tails of the distribution, which might be severely different from shifted mean surfaces, due to heteroscedasticity.

The model we here propose for the Switzerland rainfall data is a spatial quantile regression with stationary anisotropic PDE regularization by $\int_\Omega \{\mathrm{div}(\mathbf{K}\nabla f)\}^2$. Specifically, as described in the second simulation setup (Section 3.7.2), we use a two stage procedure where we first estimate the anisotropic diffusion tensor $\mathbf{K}$, as in Bernardi *et al.* (2018), and then we estimate the spatial field f, for different quantile levels. We argue that an anisotropic diffusion tensor common to all quantile surfaces is appropriate. Indeed, there is no empirical evidence to suggest an heterogeneous effect of the spatial anisotropy over the probability distribution of the rainfall, i.e., over different quantiles.

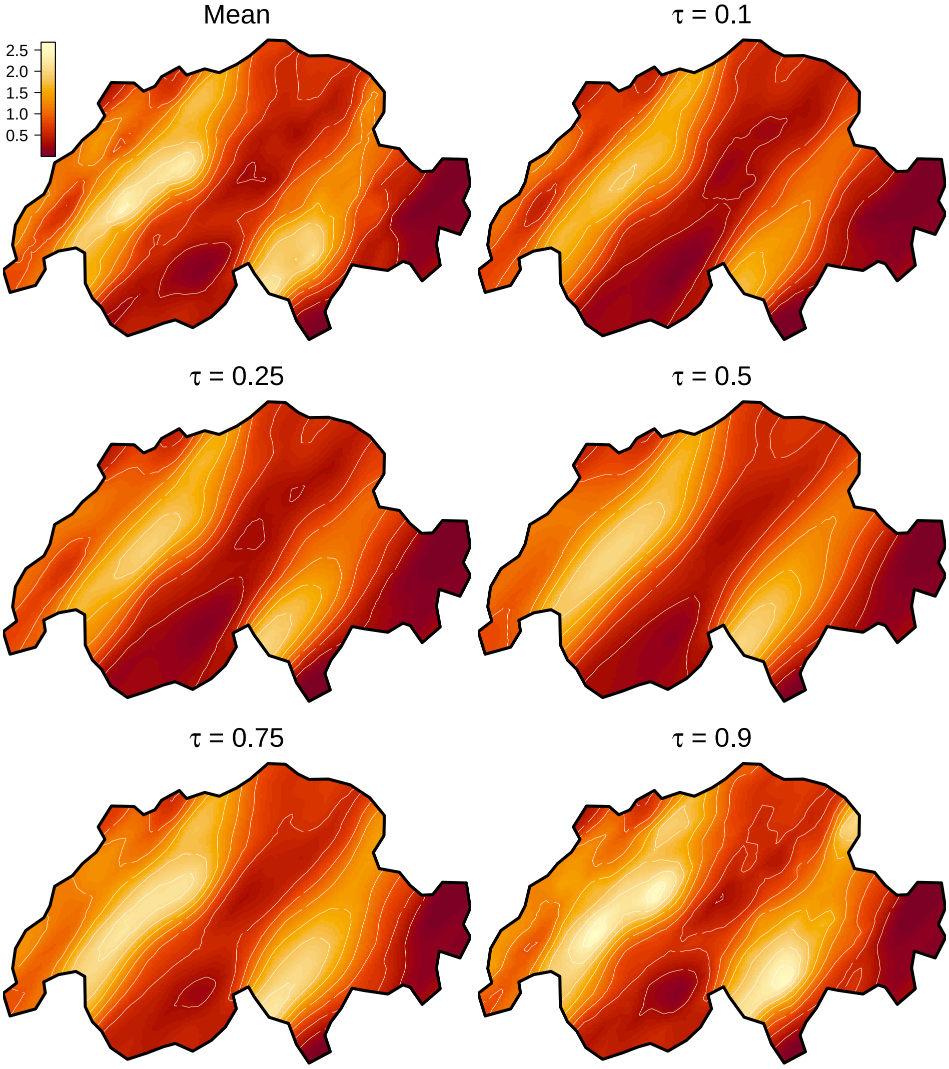

FIGURE 3.9: Estimated mean and τ-quantile fields for the Switzerland rainfall data.

Figure 3.9 shows the estimated mean and quantile surfaces at different quantile levels. In addition to a clear anisotropic effect on the mean, some interesting differences can be observed when comparing quantile and mean surfaces. The shape of the estimated quantile surfaces differs from that of the mean surface and across quantiles, with differences that are particularly evident over some (localized) regions. This indicates that some regions might expect more heterogenous patterns of rainfall than others, and more extreme events, even in the presence of similar mean rainfall. The median surface is very smooth when compared to the mean surface. This is an indication that there exist local outliers and skewness, that make the mean surface less smooth than it should be. The mean surface might also be misleading, for instance suggesting excessively low precipitations in certain areas. Surfaces corresponding to percentiles more in the tails of the distribution are less smooth, as could be reasonably expected, capturing local changes in the tail of the distribution and identifying areas where more extreme events shall be expected. The 90% percentile, for instance, is characterized by several spikes in correspondence to regions manifesting a high mean level of precipitations, a further indication of the fact that limiting the analysis to the mean surface might indeed be rather misleading.

Conclusions

Discussion and future research directions

This thesis has focused on two distinct contributions which are concerned with the estimation of misspecified robust statistical models combining the available prior information with a misfit criterion induced by a risk function. Both Bayesian and frequentist approaches have been explored, with a particular attention to data-augmentation and expectation-maximization methods.

Non-conjugate regression via variational belief updating

In Chapter 2, we developed a new variational approximation method for estimating risk-based regression models under a Bayesian belief updating setup (Bissiri *et al.*, 2016). Doing this, we built upon the works of Knowles and Minka (2011), Tan and Nott (2013) and Wand (2014), and we proposed a general variational coordinate ascent algorithm that applies for a broad range of statistical models, including generalized linear mixed models, quantile and expectile mixed models, and support vector machines.

The benefits of our approach are threefold. It is general and does not depend on model specific data-augmentation strategies. It allows for non-conjugate and structured prior distributions. It is a global approximation that preserves convexity and tail behavior of the original loss function, while leading to a natural regularization of non-smooth pseudo-likelihoods. Moreover, under suitable compatibility conditions, our marginal approach is guaranteed to dominate alternative data-augmented variational approximations in the Kullback-Leibler metric.

From an empirical point of view, we demonstrated the potentials of the proposed method by comparing it with alternative Markov chain Monte Carlo and conjugate mean field variational Bayes algorithms through extensive simulation studies and a real data application. In all the considered empirical experiments, our semiparametric variational Bayes approximation achieved good-to-excellent results in approximating the true target posterior. Moreover, it outperformed the accuracy of conjugate mean field variational Bayes over several dimensions, while keeping the same computational complexity and a similar execution time.

A final innovation of this work concerns with loss smoothing for non-regular minimization problems. This practice is often employed when it comes to optimizing non-smooth risk functions and consists of replacing an originally non-regular loss with a tilted one. The new objective function will be almost equal to the original one except for local corrections introduced to guarantee the uniform differentiability all over the function support. This way, many efficient routines are made available for the optimization, like Newton and quasi-Newton algorithms. Some examples in the quantile regression literature can be found, among others, in the works of Hunter and Lange (2000), Yue and Rue (2011), Oh *et al.* (2011) and Fasiolo *et al.* (2021a). As we discussed in Section 2.4, the variational loss averaging induced by our procedure actually provides a new recipe for constructing smooth majorizing objective functions starting from non-differentiable losses. However, differently from other existing smoothing methods based on geometric considerations, our strategy is based upon a statistical argument with a straightforward probabilistic interpretation, similarly to the expectation-maximization algorithm. Moreover, our proposal comes together with a practical rule for determining the local degree of smoothing induced by the approximation, which is in turn determined by the posterior variance of the i-th linear predictor. This leads to a different smoothing factor for each observation, allowing for adaptive calibration of the new loss.

As we already discussed in Section 2.5, many extensions to the basic approach are allowed, including models for multiple random effects, inducing shrinkage priors, dynamic and spatial processes. Other extensions not covered here, which might be of significant interest for practical applications, include nested random effects Nolan and Wand (2020); Nolan *et al.* (2020), heteroscedastic models Menictas and Wand (2015), multivariate generalized linear models Hughes *et al.* (2023).

We argue that frequentist mixed regression models can be dealt as well with our approach by a careful combination of expectation-maximization algorithm and Gaussian variational approximations, as in Ormerod and Wand (2010) and Ormerod and Wand (2012). This would be extremely useful for those models that do not enjoy the classical regularity conditions needed for implementing the Laplace approximation and would give an efficient alternative to cumbersome Monte Carlo integration (Geraci and Bottai, 2007) and multivariate quadrature methods (Geraci and Bottai, 2014). Along this line, it would be interesting to study the asymptotic property of Gaussian variational approximations for general risk-based mixed models by extending the works of Hall *et al.* (2011a) and Hall *et al.* (2011b) on variational random intercept Poisson models.

A final generalization worth to be investigated is concerned with the exploration of flexible parametric approximations alternative to the Gaussian variational approach. Indeed, it is nowadays well-recognized in literature that often posterior distributions of mixed regression models may strongly deviate from Gaussianity. This may happen, for instance, when the linear predictor is not well specified, when the number of random effects is relative to or even higher than the sample size, or when we are dealing with strongly unbalanced binary or categorical data; see, e.g., the works by Durante (2019), Fasano *et al.* (2022) and Fasano and Durante (2022). Hence, we are planning to extend our Gaussian variational approach considering the family of flexible skew-normal variational approximations proposed by Ormerod (2011).

Spatial quantile regression with differential regularization

In Chapter 3, we presented a new spatial quantile regression model with PDE regularization, which can be used to model the conditional distribution of complex spatial phenomena. A remarkable benefit of our model formulation is its generality, which permits us to naturally combine the capabilities of quantile regression (Koenker, 2005) with the flexibility of a structured PDE description of spatial phenomena (Sangalli, 2021). In this way, we can jointly handle distributional features, such as heteroscedasticity and local skewness, along with spatial features, such as boundary constraints, anisotropic diffusion effects, unidirectional flows and local shrinkage effects.

We then proposed an innovative functional expectation-maximization algorithm for parameter and field estimation, and we characterized its iterative solution by means of an exact variational constrained formulation. We then approximated the resulting functional estimator using finite elements methods. Doing so, we obtained a numerical routine which approach the optimal finite-dimensional quantile estimator by solving a sequence of high-dimensional sparse linear systems. Starting from the finite element discretization, we also developed an approximated Generalized Cross-Validation criterion to select the optimal smoothing parameter taking into account of the differential penalty and of the implicit constrained formulation of the estimation problem.

We then investigated the large-sample properties of our penalized estimator, proving the consistency of the infinite-dimensional estimator and determining the asymptotic distribution of the finite-dimensional estimator conditioned on a fixed triangularization of the domain. We are still working on a third asymptotic result, which studies the behavior of the finite-dimensional estimator for a discretization of the domain getting finer and finer at the growing of the sample size. Our conjecture is that there exists a rate of growth linking the mesh dimension and the sample size, which permits to achieving the L^2-consistency of the asymptotic spatial quantile estimator even in this scenario.

Finally, we showed how to generalize the pure nonparametric spatial quantile estimator to a semiparametric formulation, which jointly models the effect of space-varying covariates and a nonparametric spatial field. The estimation algorithm and selection procedure have been generalized as well.

The proposed methods have very good empirical performances and comparative advantages with respect to state-of-the-art alternatives, as shown by extensive simulation studies and the analysis of a real dataset. In particular, the considered scenarios highlight the importance of accounting for complex domain structures, for deviation from isotropic stationary fields, and for spatially heterogeneous characteristics of the distribution, such as heteroscedasticity or local skewness. The proposed method offers increasingly large advantages with respect to the available alternatives as the data complexity increases, in terms of domain shape or field characteristics.

An additional benefit of our model formulation is its generality, which permits to naturally accommodating for several extensions, different sampling schemes and PDE models. For instance, data recorded over areal regions can be handled by extending the proposal by Azzimonti *et al.* (2015) in the context of spatial linear models. Data observed in space and changing over time, as well as spatially varying functional data, could be handled by considering space-time PDEs, as shown by Bernardi *et al.* (2017) and Arnone *et al.* (2019) for spatial linear models.

Another fascinating possibility is to adapt our PDE-based spatial quantile regression model to data lying on a curved domain. In this context, the necessary theoretical background is provided by the works of Ettinger *et al.* (2016) and Lila *et al.* (2016) for spatial linear regression over two-dimensional Riemannian manifolds. This could be naturally integrated into our quantile regression framework, permitting, for instance, the application of our method to data observed over the globe. More generally, such an extension might be of significant interest for several applications, in mechanical and space engineering, as well as in life sciences.

Another relevant research direction is concerned with the joint modelling of multiple quantiles. Inspired by the recent works of Schnabel and Eilers (2013), Frumento and Bottai (2016) and Frumento *et al.* (2021) on varying coefficient modelling, we might extend our methodology in order to jointly estimate a whole family of quantile fields, as smooth functions of the confidence level τ. Such an approach, together with a careful management of the natural non-crossing constraints, could be beneficial to borrow information across quantiles and improve the approximation of the whole conditional distribution. In this context, the heteroscedastic quantile model by He *et al.* (1998) and the constrained quantile curves by Bondell *et al.* (2010) provide two remarkable examples of non-crossing approaches that are worth investigating and possibly extending to the proposed spatial regression with PDE regularization.

Appendix A

A.1 Integration results

The calculation of non-standard expected values is the main challenge in variational Bayesian inference. Here, we face this issue with a careful combination of integral simplification, analytic solution and efficient univariate numerical quadrature. The results in this section are then introduced in order to characterize and simplify the most complex integration tasks encountered in this article, but may also provide useful insights for future works not necessarily restricted to the variational approach.

In the rest of this section, we will make use of the following notation: $g : \mathbb{R} \to \mathbb{R}$ is a measurable function, having integrable n-th order weak derivative g_n up to the order N. We recall that g_n is defined as the measurable function satisfying the integral equation

$$\int_a^b \frac{\mathrm{d}^n}{\mathrm{d}x^n} \varphi(x)\, g(x)\, \mathrm{d}x = (-1)^n \int_a^b \varphi(x)\, g_n(x)\, \mathrm{d}x,$$

for any infinitely differentiable function $\varphi : [a, b] \to \mathbb{R}$ such that $\varphi(a) = \varphi(b) = 0$. If g is n times differentiable, then $g_n = \mathrm{d}^n g/\mathrm{d}x^n$. We denote with $H_n : \mathbb{R} \to \mathbb{R}$ the n-th order Hermite polynomial, which is the solution of the differential equation

$$\frac{\mathrm{d}^n}{\mathrm{d}z^n} \phi(z) = (-1)^n H_n(z) \phi(z), \quad z \in \mathbb{R}, \quad n \in \mathbb{N}.$$

In particular, we recall that $H_0(z) = 1$, $H_1(z) = z$ and $H_2(z) = z^2 - 1$. Finally, we introduce the notation $(\boldsymbol{x})^n$ for $n = 0, 1, 2$, that is $(\boldsymbol{x})^0 = 1$, $(\boldsymbol{x})^1 = \boldsymbol{x}$ and $(\boldsymbol{x})^2 = \boldsymbol{x}\boldsymbol{x}^\top$.

Most of the non-analytic integral encountered in our work take the following functional forms

$$\mathcal{F}_n(g, \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\mathbb{R}^d} g_n(\boldsymbol{a}^\top \boldsymbol{x})\, \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathrm{d}\boldsymbol{x}, \tag{A.1}$$

$$\mathcal{G}_n(g, \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\mathbb{R}^d} g(\boldsymbol{a}^\top \boldsymbol{x})\, (\boldsymbol{x})^n\, \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathrm{d}\boldsymbol{x}, \tag{A.2}$$

$$\mathcal{H}_n(g, \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{-\infty}^{+\infty} g\Big(\boldsymbol{a}^\top \boldsymbol{\mu} + \sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}\, z\Big)\, H_n(z)\, \phi(z)\, \mathrm{d}z, \tag{A.3}$$

where $\boldsymbol{a} \in \mathbb{R}^d$. Such integrals are assumed to be well-defined at least for $n = 0, 1, 2$. The main properties of functionals (A.1), (A.2), (A.3) are then provided in the follwing propositions.

Proposition A.1. *Let $g : \mathbb{R} \to \mathbb{R}$ be a function such that g_n and $\mathcal{F}_n$ exist up to the order N. Then:*

(1) $\mathcal{F}_n$ has infinitely many continuous derivatives with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$;

(2) if g_n is continuous, then $\mathcal{F}_n(g, \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \to g_n(\boldsymbol{a}^\top \boldsymbol{\mu})$ as $\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a} \to 0$;

(3) if g is convex in x, then $\mathcal{F}_0$ is jointly convex with respect to $\boldsymbol{a}^\top \boldsymbol{\mu}$ and $\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}$;

(4) if g is convex in x, then $g(\boldsymbol{a}^\top \boldsymbol{\mu}) \leq \mathcal{F}_0(g, \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Proposition A.2. *Let $\boldsymbol{n} = (n_1, \dots, n_d) \in \mathbb{N}^d$ be a multi-index vector with rank $n = n_1 + \dots + n_d$. Then, if $\mathcal{F}_0, \dots, \mathcal{F}_n$ are well-defined, the $\boldsymbol{n}$-th order mixed derivative of $\mathcal{F}_0$ with respect to $\boldsymbol{\mu}$ is*

$$\frac{\partial^n \mathcal{F}_0}{\partial \mu_1^{n_1} \dots \partial \mu_d^{n_d}} = \left(\prod_{j=1}^{d} a_j^{n_j} \right) \mathcal{F}_n.$$

Proposition A.3. *Assuming that $\mathcal{F}_n$, $\mathcal{G}_n$ and $\mathcal{H}_n$ are well-defined for $n = 0, 1, 2$, then we have*

$$\mathcal{G}_0 = \mathcal{F}_0 = \mathcal{H}_0, \qquad \mathcal{G}_1 = \boldsymbol{\mu} \mathcal{H}_0 + \frac{\boldsymbol{\Sigma}\, \boldsymbol{a}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}} \mathcal{H}_1,$$
$$\mathcal{G}_2 = (\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}) \mathcal{H}_0 + \frac{\boldsymbol{\Sigma} \boldsymbol{a} \boldsymbol{\mu}^\top + \boldsymbol{\mu} \boldsymbol{a}^\top \boldsymbol{\Sigma}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}} \mathcal{H}_1 + \frac{\boldsymbol{\Sigma}\, \boldsymbol{a} \boldsymbol{a}^\top \boldsymbol{\Sigma}}{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}} \mathcal{H}_2.$$

Proposition A.4. *Let $\mathcal{H}_n = \mathcal{H}_n(\tilde{g}, \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\tilde{g}(\cdot) = \mathbb{I}_{(b,c]}(\cdot)$, where $b, c \in \mathbb{R}$ such that $b < c$. Then, $\mathcal{H}_n$ is well-defined and is equal to*

$$\mathcal{H}_n = \begin{cases} \Phi(z^+) - \Phi(z^-) & \text{if } n = 0, \\ (-1) \left[H_{n-1}(z^+) \phi(z^+) - H_{n-1}(z^-) \phi(z^-) \right] & \text{if } n > 0, \end{cases}$$

where $z^- = \dfrac{b - \boldsymbol{a}^\top \boldsymbol{\mu}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}}$ *and* $z^+ = \dfrac{b - \boldsymbol{a}^\top \boldsymbol{\mu}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}}$.

Proposition 2.4, introduces some fundamental properties of the functional $\mathcal{F}_n$ and provides mild regularity conditions for our variational problems to be well-defined. Those conditions essentially imply that $g(\cdot)$ is a measurable, convex, sub-exponential loss function. Actually, such properties are satisfied by almost all the most popular loss functions used for classification and regression tasks, going from the Hinge loss for support vector machine to the exponential family negative log-likelihoods for generalized linear models.

Proposition A.2 characterizes the recursive relation connecting the derivatives of $\mathcal{F}_0$ with $\mathcal{F}_n$. This allows to simplify the formulation of the first and second order necessary conditions needed for optimizing $\mathcal{F}_0$ with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Proposition A.3 permits to rephrase a wide class of complicate multidimensional integrals in terms of univariate expectations, that are way more feasible to solve than multivariate ones.

Finally, Proposition A.4, together with Proposition A.3, provides a closed form solution to the integrals $\mathcal{H}_n$, $n = 0, 1, 2$, for a class of functions $g(\cdot)$ that arises to be very useful when calculating the truncated moments of a multivariate Gaussian random vector subjected to linear constraints. Specifically, we use such results for deriving a variational approximation algorithm with closed form updating formulas for support vector machine, quantile and expectile regression. For more details refer to Section 2.4 and Appendix B.

A specialized version of Proposition A.3 has been used by Hall *et al.* (2020) for developing an expectation-propagation algorithm in the context of binary mixed regression. Previously, the works of Ormerod and Wand (2012) and Tan and Nott (2013) helped to clarify the potential of similar integral transformation techniques for calculating the marginal likelihood of possibly complicated generalized linear mixed models. Indeed, due to the curse of dimensionality, any multidimensional-to-unidimensional integral collapsing has the double advantage to allow for a more accurate numerical evaluation with an exponentially lower computation cost. Here, we extend this approach in order to make inference on a larger class of models that includes, as remarkable cases, generalized linear mixed models.

Remark A.5. The properties of the Ψ-functions described in Section 2.3 all derive from the observation that $\Psi_r(y, \boldsymbol{a}^\top\boldsymbol{\mu}, \boldsymbol{a}^\top\boldsymbol{\Sigma}\,\boldsymbol{a}) = \mathcal{F}_r(\psi(y,\cdot), \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, for $r = 0, 1, 2$ and for any $y \in \mathcal{Y}$. In particular, Theorem 2.4 directly follows from Proposition A.1, that is proved in the following section.

Proof of Proposition A.1

(1) The differentiability of $\mathcal{F}_n$ with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is guaranteed by the derivation under integral sign theorem and by the fact that $\phi(\cdot; \mu, \sigma^2)$ is an analytic function having infinitely many continuous derivatives with respect to μ and σ. In fact, defining the scalar variable $x = \boldsymbol{a}^\top\boldsymbol{x} \sim \mathrm{N}(\mu, \sigma^2)$, with $\mu = \boldsymbol{a}^\top\boldsymbol{\mu}$ and $\sigma^2 = \boldsymbol{a}^\top\boldsymbol{\Sigma}\,\boldsymbol{a}$, we have

$$\mathcal{F}_n = \int_{\mathbb{R}^d} g_n(\boldsymbol{a}^\top\boldsymbol{x})\, \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathrm{d}\boldsymbol{x} = \int_{-\infty}^{+\infty} g_n(x)\, \phi(x; \mu, \sigma^2)\, dx$$

and

$$\frac{\partial^r}{\partial\mu^r}\frac{\partial^s}{\partial\sigma^s}\int_{-\infty}^{+\infty} g_n(x)\, \phi(x; \mu, \sigma^2)\, \mathrm{d}x = \int_{-\infty}^{+\infty} g_n(x) \frac{\partial^r}{\partial\mu^r}\frac{\partial^s}{\partial\sigma^s}\, \phi(x; \mu, \sigma^2)\, \mathrm{d}x,$$

for any non-negative integer r and s.

(2) Let $\boldsymbol{x}_k \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}_k)$ be a sequence of random variables such that $\boldsymbol{a}^\top\boldsymbol{\Sigma}_k\,\boldsymbol{a} \to 0$ as $k \to \infty$. Then, we have

$$\lim_{k\to\infty} \mathbb{E}\,|\boldsymbol{a}^\top\boldsymbol{x}_k - \boldsymbol{a}^\top\boldsymbol{\mu}|^2 = \lim_{k\to\infty} (\boldsymbol{a}^\top\boldsymbol{\Sigma}_k\,\boldsymbol{a}) = 0,$$

which leads to $\boldsymbol{a}^\top\boldsymbol{x}_k \xrightarrow{L^2} \boldsymbol{a}^\top\boldsymbol{\mu}$ and $\boldsymbol{a}^\top\boldsymbol{x}_k \xrightarrow{\mathrm{p}} \boldsymbol{a}^\top\boldsymbol{\mu}$. Since the convergence in probability is closed with respect to continuous transformations, if g_n is continuous, we get

$g_n(\boldsymbol{a}^\top \boldsymbol{x}_k) \xrightarrow{\mathrm{p}} g_n(\boldsymbol{a}^\top \boldsymbol{\mu})$. Hence, $\mathbb{E}\{g_n(\boldsymbol{a}^\top \boldsymbol{x}_k)\} \to g_n(\boldsymbol{a}^\top \boldsymbol{\mu})$ as $k \to \infty$, which concludes the proof.

(3) For what regards the joint convexity of $\mathcal{F}_0$, we consider the transformation $\mu = \boldsymbol{a}^\top \boldsymbol{\mu}$ and $\sigma^2 = \boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}$. Recall that any smooth function is convex whenever its Hessian matrix is positive semidefinite, namely $\mathcal{F}_0$ is jointly convex in μ and σ if and only if

$$\nabla^2 \mathcal{F}_0 = \begin{bmatrix} \partial^2_{\mu\mu} \mathcal{F}_0 & \partial^2_{\mu\sigma} \mathcal{F}_0 \\ \partial^2_{\sigma\mu} \mathcal{F}_0 & \partial^2_{\sigma\sigma} \mathcal{F}_0 \end{bmatrix} \succeq 0.$$

Using the definition of weak derivatives of g together with a location-scale change of variable, we can write

$$\begin{aligned} \partial^2_{\mu\mu} \mathcal{F}_0 &= \frac{\partial^2 \mathcal{F}_0}{\partial\mu\partial\mu} = \int_{-\infty}^{+\infty} g_2(\mu + \sigma z)\, \phi(z)\, \mathrm{d}z, \\ \partial^2_{\mu\sigma} \mathcal{F}_0 &= \frac{\partial^2 \mathcal{F}_0}{\partial\mu\partial\sigma} = \int_{-\infty}^{+\infty} g_2(\mu + \sigma z)\, z\, \phi(z)\, \mathrm{d}z, \\ \partial^2_{\sigma\sigma} \mathcal{F}_0 &= \frac{\partial^2 \mathcal{F}_0}{\partial\sigma\partial\sigma} = \int_{-\infty}^{+\infty} g_2(\mu + \sigma z)\, z^2 \phi(z)\, \mathrm{d}z. \end{aligned}$$

Recall that $g_2(x) \geq 0$ for any $x \in \mathbb{R}$, because of the convexity of g. Next, let us define $f_1(z) = 1$, $f_2(z) = z$ and $h(z) = g_2(\mu + \sigma z)\phi(z)$. In this way, the above derivatives may be written as

$$\frac{\partial^2 \mathcal{F}_0}{\partial\mu\partial\mu} = \langle f_1, f_1 \rangle_{\mathbb{H}}, \quad \frac{\partial^2 \mathcal{F}_0}{\partial\mu\partial\sigma} = \langle f_1, f_2 \rangle_{\mathbb{H}}, \quad \frac{\partial^2 \mathcal{F}_0}{\partial\sigma\partial\sigma} = \langle f_2, f_2 \rangle_{\mathbb{H}}$$

where $\mathbb{H}$ is a positive measure having density function $h(\cdot)$, i.e. $\mathbb{H}(\mathrm{d}x) = h(x)\, \mathrm{d}x$, while $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and $\| \cdot \|_{\mathbb{H}}$ are, respectively, the inner product and the norm induced by $\mathbb{H}$:

$$\langle f_1, f_2 \rangle_{\mathbb{H}} = \int_{-\infty}^{+\infty} f_1(x)\, f_2(x)\, \mathbb{H}(\mathrm{d}x), \qquad \|f\|^2_{\mathbb{H}} = \int_{-\infty}^{+\infty} |f(x)|^2\, \mathbb{H}(\mathrm{d}x).$$

Then, thanks to the Holder inequality, we get $|\langle f_1, f_2 \rangle_{\mathbb{H}}| \leq \|f_1\|_{\mathbb{H}}\, \|f_2\|_{\mathbb{H}}$, namely $|\partial^2_{\mu\sigma} \mathcal{F}_0|^2 \leq (\partial^2_{\mu\mu} \mathcal{F}_0)(\partial^2_{\sigma\sigma} \mathcal{F}_0)$. Which implies that $\nabla^2 \mathcal{F}_0$ is a proper covariance matrix with positive semidefinite signature and, thereby, $\mathcal{F}_0$ is a convex function with respect to μ and σ.

(4) The lower bound $g(\boldsymbol{a}^\top \boldsymbol{\mu}) \leq \mathcal{F}_0(g, \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ immediately follows from the Jensen inequality: $g\{\boldsymbol{a}^\top \mathbb{E}(\boldsymbol{x})\} \leq \mathbb{E}\{g(\boldsymbol{a}^\top \boldsymbol{x})\}$. This concludes the proof. □

Proof of Proposition A.2

Let consider the Gaussian random variable $x = \boldsymbol{a}^\top \boldsymbol{x} \sim \mathrm{N}(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a})$, with $\boldsymbol{x} \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, so that the following d-dimensional integral collapses into a univariate integral:

$$\mathcal{F}_0 = \int_{\mathbb{R}^d} g(\boldsymbol{a}^\top \boldsymbol{x})\, \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathrm{d}\boldsymbol{x} = \int_{-\infty}^{+\infty} g(x)\, \phi(x; \boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a})\, \mathrm{d}x.$$

In order to prove the statement in Proposition A.2 we use an induction argument. Let us start from the initial step deriving under integral sign with respect to μ_j:

$$\begin{aligned}\frac{\partial \mathcal{F}_0}{\partial \mu_j} &= \int_{-\infty}^{+\infty} \frac{g(x)}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}} \frac{\partial}{\partial \mu_j} \phi\left(\frac{x - \boldsymbol{a}^\top \boldsymbol{\mu}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}}\right) \mathrm{d}x \\ &= -\frac{a_j}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}} \int_{-\infty}^{+\infty} \frac{g(x)}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}} \frac{\mathrm{d}\phi}{\mathrm{d}z} \left(\frac{x - \boldsymbol{a}^\top \boldsymbol{\mu}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}}\right) \mathrm{d}x.\end{aligned}$$

Because of the location-scale representation of the Gaussian distribution, we can write $x = \boldsymbol{a}^\top \boldsymbol{\mu} + \sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}\, z$, where $z \sim \mathrm{N}(0,1)$ and $\mathrm{d}y = \sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}\, \mathrm{d}z$; in this way, we have

$$\frac{\partial \mathcal{F}_0}{\partial \mu_j} = -\frac{a_j}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}} \int_{-\infty}^{+\infty} g\left(\boldsymbol{a}^\top \boldsymbol{\mu} + \sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}\, z\right) \frac{\mathrm{d}}{\mathrm{d}z} \phi(z)\, \mathrm{d}z.$$

Observing that $\mathrm{d}^n\phi/\mathrm{d}z^n$ vanishes in the limit as $|z| \to \infty$ for any $n \in \mathbb{N}$, we are allowed to integrate by parts with respect to z and apply the definition of weak derivative of g, obtaining

$$\frac{\partial \mathcal{F}_0}{\partial \mu_j} = a_j \int_{-\infty}^{+\infty} g_1\left(\boldsymbol{a}^\top \boldsymbol{\mu} + \sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}\, z\right) \phi(z)\, \mathrm{d}z,$$

where

$$\mathrm{d}g\left(\boldsymbol{a}^\top \boldsymbol{\mu} + \sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}\, z\right) = \sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}\, g_1\left(\boldsymbol{a}^\top \boldsymbol{\mu} + \sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}\, z\right) \mathrm{d}z.$$

For concluding the first step of the proof, we just need to back-transform z to x:

$$\frac{\partial \mathcal{F}_0}{\partial \mu_j} = a_j \int_{-\infty}^{+\infty} g_1(x)\, \phi(x; \boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a})\, \mathrm{d}x = a_j\, \mathbb{E}\big\{g_1(\boldsymbol{a}^\top \boldsymbol{x})\big\},$$

which satisfies the formula in Proposition A.2 for $\boldsymbol{n} = \boldsymbol{e}_j$.

Let us move to the induction step. We consider the $\boldsymbol{n}$-th order mixed derivative of $\mathcal{F}_0$ and we derive again under integral sign with respect to μ_j:

$$\frac{\partial}{\partial \mu_j} \frac{\partial^n \mathcal{F}_0}{\partial \mu_1^{n_1} \ldots \partial \mu_d^{n_d}} = \left(\prod_{j=1}^{d} a_j^{n_j}\right) \int_{-\infty}^{+\infty} g_n(x) \frac{\partial}{\partial \mu_j} \phi(x; \boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a})\, \mathrm{d}y.$$

Following the same arguments used for the initial step, sequentially applying derivation, transformation, integration by parts and back-transformation, we get that the right hand side integral becomes

$$
\begin{aligned}
\int_{-\infty}^{+\infty} \frac{g_n(x)}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}} \frac{\partial}{\partial \mu_j} \phi\left(\frac{x-\boldsymbol{a}^\top \boldsymbol{\mu}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}}\right) \mathrm{d}x =& \\
&= -\frac{a_j}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}} \int_{-\infty}^{+\infty} \frac{g_n(x)}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}} \frac{\mathrm{d}}{\mathrm{d}z} \phi\left(\frac{x-\boldsymbol{a}^\top \boldsymbol{\mu}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}}\right) \mathrm{d}x \\
&= -\frac{a_j}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}} \int_{-\infty}^{+\infty} g_n\left(\boldsymbol{a}^\top \boldsymbol{\mu} + \sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}\, z\right) \frac{\mathrm{d}}{\mathrm{d}z} \phi(z)\, \mathrm{d}z \\
&= a_j \int_{-\infty}^{+\infty} g_{n+1}\left(\boldsymbol{a}^\top \boldsymbol{\mu} + \sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}\, z\right) \phi(z)\, \mathrm{d}z \\
&= a_j \int_{-\infty}^{+\infty} \frac{g_{n+1}(x)}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}} \phi\left(\frac{x-\boldsymbol{a}^\top \boldsymbol{\mu}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}}\right) \mathrm{d}x.
\end{aligned}
$$

This leads to

$$
\frac{\partial^{n+1} \mathcal{F}_0}{\partial \mu_1^{n_1} \cdots \mu_1^{n_j+1} \cdots \partial \mu_d^{n_d}} = \left(a_j^{n_j+1} \prod_{k \neq j} a_k^{n_k} \right) \int_{-\infty}^{+\infty} \frac{g_{n+1}(x)}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}} \phi\left(\frac{x-\boldsymbol{a}^\top \boldsymbol{\mu}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}\, \boldsymbol{a}}}\right) \mathrm{d}y.
$$

This concludes the induction step and thus the proof. ☐

Proof of Proposition A.3

In order to prove Proposition A.3, we need to introduce some intermediate results.

Lemma A.6 (Lemma 1 from Hall *et al.* (2020))**.** *Let* $\boldsymbol{x} \sim \mathrm{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ *and* $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3 \in \mathbb{R}^d$*, then*

$$
\int_{\mathbb{R}^d} g(\boldsymbol{a}_1^\top \boldsymbol{x})\, \phi_d(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \mathcal{H}_0, \qquad \int_{\mathbb{R}^d} g(\boldsymbol{a}_1^\top \boldsymbol{x})\, (\boldsymbol{a}_2^\top \boldsymbol{x})\, \phi_d(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \frac{(\boldsymbol{a}_1^\top \boldsymbol{a}_2)}{\|\boldsymbol{a}_1\|} \mathcal{H}_1,
$$
$$
\int_{\mathbb{R}^d} g(\boldsymbol{a}_1^\top \boldsymbol{x})(\boldsymbol{a}_2^\top \boldsymbol{x})(\boldsymbol{a}_3^\top \boldsymbol{x})\, \phi_d(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = (\boldsymbol{a}_2^\top \boldsymbol{a}_3)\, \mathcal{H}_0 + \frac{(\boldsymbol{a}_1^\top \boldsymbol{a}_2)(\boldsymbol{a}_1^\top \boldsymbol{a}_3)}{\|\boldsymbol{a}_1\|^2} \mathcal{H}_2.
$$

where $\mathcal{H}_n = \mathcal{H}_n(g, \boldsymbol{a}_1, \mathbf{0}_d, \mathbf{I}_d)$*.*

Proof. See the proof of Lemma 1 in the supplement material of Hall *et al.* (2020). ☐

Lemma A.7. *Let* $\mathcal{G}_n = \mathcal{G}_n(g, \boldsymbol{a}, \mathbf{0}_d, \mathbf{I}_d)$ *and* $\mathcal{H}_n = \mathcal{H}_n(g, \boldsymbol{a}, \mathbf{0}_d, \mathbf{I}_d)$ *for* $n = 0, 1, 2$*, then*

$$
\mathcal{G}_0 = \mathcal{H}_0, \qquad \mathcal{G}_1 = \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} \mathcal{H}_1, \qquad \mathcal{G}_2 = \mathbf{I}_d\, \mathcal{H}_0 + \frac{\boldsymbol{a}\boldsymbol{a}^\top}{\|\boldsymbol{a}\|^2} \mathcal{H}_2.
$$

Proof. Let us consider a particular application of Lemma A.6, assuming $\boldsymbol{a}_1 = \boldsymbol{a}$, $\boldsymbol{a}_2 = \boldsymbol{e}_i$, $\boldsymbol{a}_3 = \boldsymbol{e}_j$, where $\boldsymbol{e}_i$ is the i–th column of an identity matrix. Then, for $\mathcal{G}_n = \mathcal{G}_n(g, \boldsymbol{a}, \mathbf{0}_d, \mathbf{I}_d)$, we have

$$
\mathcal{G}_0 = \int_{\mathbb{R}^d} g(\boldsymbol{a}^\top \boldsymbol{x})\, \phi_d(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \int_{-\infty}^{+\infty} g(\|\boldsymbol{a}\| z)\, \phi(z)\, \mathrm{d}z = \mathcal{H}_0(g, \boldsymbol{a}, \mathbf{0}_d, \mathbf{I}_d).
$$

The same result applies for the j-th component of the vector $\mathcal{G}_1$, being

$$\begin{aligned}[\mathcal{G}_1]_j &= \int_{\mathbb{R}^d} g(\boldsymbol{a}^\top \boldsymbol{x})\,(\boldsymbol{e}_j^\top \boldsymbol{x})\,\phi_d(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\ &= \frac{\boldsymbol{a}^\top \boldsymbol{e}_j}{\|\boldsymbol{a}\|} \int_{-\infty}^{+\infty} g(\|\boldsymbol{a}\|z)\,z\,\phi(z)\,dz = \frac{\boldsymbol{a}^\top \boldsymbol{e}_j}{\|\boldsymbol{a}\|}\,\mathcal{H}_1(g, \boldsymbol{a}, \mathbf{0}_d, \mathbf{I}_d).\end{aligned}$$

Finally, let us consider the (i, j)-th element of the matrix $\mathcal{G}_2$, that is

$$\begin{aligned}[\mathcal{G}_2]_{ij} &= \int_{\mathbb{R}^d} g(\boldsymbol{a}^\top \boldsymbol{x})\,(\boldsymbol{e}_i^\top \boldsymbol{x})\,(\boldsymbol{e}_j^\top \boldsymbol{x})\,\phi_d(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\ &= (\boldsymbol{e}_i^\top \boldsymbol{e}_j) \int_{-\infty}^{+\infty} g(\|\boldsymbol{a}\|z)\,\phi(z)\,dz + \frac{(\boldsymbol{a}^\top \boldsymbol{e}_i)(\boldsymbol{a}^\top \boldsymbol{e}_j)}{\|\boldsymbol{a}\|^2} \int_{-\infty}^{+\infty} g(\|\boldsymbol{a}\|z)(z^2-1)\,\phi(z)\,dz \\ &= (\boldsymbol{e}_i^\top \boldsymbol{e}_j)\,\mathcal{H}_0(g, \boldsymbol{a}, \mathbf{0}_d, \mathbf{I}_d) + \frac{(\boldsymbol{a}^\top \boldsymbol{e}_i)(\boldsymbol{a}^\top \boldsymbol{e}_j)}{\|\boldsymbol{a}\|^2}\,\mathcal{H}_2(g, \boldsymbol{a}, \mathbf{0}_d, \mathbf{I}_d).\end{aligned}$$

This concludes the proof of Lemma A.7. $\square$

Now, we have all the ingredients to complete the proof of Proposition A.3. Firstly, let us consider the change of variable $\boldsymbol{x} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{z}$ and its inverse map $\boldsymbol{z} = \mathbf{L}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$, where $\mathbf{L}$ is the lower-triangular Cholesky factor of $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$, so that the volume element transforms as $\mathrm{d}\boldsymbol{x} = |\boldsymbol{\Sigma}|^{1/2}\mathrm{d}\boldsymbol{z}$. Therefore,

$$\mathcal{G}_n = \int_{\mathbb{R}^d} (\boldsymbol{x})^n\, g(\boldsymbol{a}^\top \boldsymbol{x})\,\phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^d} (\boldsymbol{\mu} + \mathbf{L}\boldsymbol{z})^n\, g(\boldsymbol{a}^\top \boldsymbol{\mu} + \boldsymbol{a}^\top \mathbf{L}\boldsymbol{z})\,\phi_d(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}.$$

In particular, we have

$$\begin{aligned}\mathcal{G}_0 &= \int_{\mathbb{R}^d} g(\boldsymbol{a}^\top \boldsymbol{\mu} + \boldsymbol{a}^\top \mathbf{L}\boldsymbol{z})\,\phi_d(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = \tilde{\mathcal{H}}_0, \\ \mathcal{G}_1 &= \int_{\mathbb{R}^d} (\boldsymbol{\mu} + \mathbf{L}\boldsymbol{z})\, g(\boldsymbol{a}^\top \boldsymbol{\mu} + \boldsymbol{a}^\top \mathbf{L}\boldsymbol{z})\,\phi_d(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = \boldsymbol{\mu}\,\tilde{\mathcal{H}}_0 + \mathbf{L}\,\tilde{\mathcal{H}}_1, \\ \mathcal{G}_2 &= \int_{\mathbb{R}^d} (\boldsymbol{\mu} + \mathbf{L}\boldsymbol{z})(\boldsymbol{\mu} + \mathbf{L}\boldsymbol{z})^\top\, g(\boldsymbol{a}^\top \boldsymbol{\mu} + \boldsymbol{a}^\top \mathbf{L}\boldsymbol{z})\,\phi_d(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} \\ &= \boldsymbol{\mu}\boldsymbol{\mu}^\top \tilde{\mathcal{H}}_0 + \boldsymbol{\mu}\,\tilde{\mathcal{H}}_1^\top \mathbf{L}^\top + \mathbf{L}\,\tilde{\mathcal{H}}_1\,\boldsymbol{\mu}^\top + \mathbf{L}\,\tilde{\mathcal{H}}_2\,\mathbf{L}^\top,\end{aligned}$$

where the functional $\tilde{\mathcal{H}}_n = \int_{\mathbb{R}^d} (\boldsymbol{z})^n \, g(\boldsymbol{a}^\top \boldsymbol{\mu} + \boldsymbol{a}^\top \mathbf{L} \boldsymbol{z}) \, \phi_d(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{x}$ can be obtained by means of Lemma A.7. Hence, due to the identity $\|\mathbf{L}^\top \boldsymbol{a}\| = \sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma} \, \boldsymbol{a}}$, we get

$$
\begin{aligned}
\tilde{\mathcal{H}}_0 &= \int_{-\infty}^{+\infty} g\big(\boldsymbol{a}^\top \boldsymbol{\mu} + \|\mathbf{L}^\top \boldsymbol{a}\| \, z\big) \, \phi(z) \, \mathrm{d}z = \mathcal{H}_0, \\
\tilde{\mathcal{H}}_1 &= \frac{\mathbf{L}^\top \boldsymbol{a}}{\|\mathbf{L}^\top \boldsymbol{a}\|} \int_{-\infty}^{+\infty} g\big(\boldsymbol{a}^\top \boldsymbol{\mu} + \|\mathbf{L}^\top \boldsymbol{a}\| \, z\big) \, z \, \phi(z) \, \mathrm{d}z = \frac{\mathbf{L}^\top \boldsymbol{a}}{\|\mathbf{L}^\top \boldsymbol{a}\|} \, \mathcal{H}_1, \\
\tilde{\mathcal{H}}_2 &= \mathbf{I}_d \int_{-\infty}^{+\infty} g\big(\boldsymbol{a}^\top \boldsymbol{\mu} + \|\mathbf{L}^\top \boldsymbol{a}\| \, z\big) \, H_n(z) \, \phi(z) \, \mathrm{d}z \\
&\quad + \frac{\mathbf{L}^\top \boldsymbol{a} \boldsymbol{a}^\top \mathbf{L}}{\|\mathbf{L}^\top \boldsymbol{a}\|^2} \int_{-\infty}^{+\infty} g\big(\boldsymbol{a}^\top \boldsymbol{\mu} + \|\mathbf{L}^\top \boldsymbol{a}\| \, z\big) \, (z^2 - 1) \, \phi(z) \, \mathrm{d}z = \mathbf{I}_d \, \mathcal{H}_0 + \frac{\mathbf{L}^\top \boldsymbol{a} \boldsymbol{a}^\top \mathbf{L}}{\|\mathbf{L}^\top \boldsymbol{a}\|^2} \, \mathcal{H}_2.
\end{aligned}
$$

Finally, we note that

$$
\begin{aligned}
\mathbf{L} \, \tilde{\mathcal{H}}_1 &= \frac{\mathbf{L} \mathbf{L}^\top \boldsymbol{a}}{\|\mathbf{L}^\top \boldsymbol{a}\|} \, \mathcal{H}_1 = \frac{\boldsymbol{\Sigma} \boldsymbol{a}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma} \, \boldsymbol{a}}} \, \mathcal{H}_1, \\
\mathbf{L} \, \tilde{\mathcal{H}}_2 \, \mathbf{L}^\top &= \mathbf{L} \mathbf{L}^\top \, \mathcal{H}_0 + \frac{\mathbf{L} \mathbf{L}^\top \boldsymbol{a} \boldsymbol{a}^\top \mathbf{L} \mathbf{L}^\top}{\|\mathbf{L}^\top \boldsymbol{a}\|^2} \, \mathcal{H}_2 = \boldsymbol{\Sigma} \, \mathcal{H}_0 + \frac{\boldsymbol{\Sigma} \, \boldsymbol{a} \boldsymbol{a}^\top \boldsymbol{\Sigma}}{\boldsymbol{a}^\top \boldsymbol{\Sigma} \, \boldsymbol{a}} \, \mathcal{H}_2.
\end{aligned}
$$

This concludes the proof of Proposition A.3. □

Proof of Proposition A.4

Assuming $g(\cdot) = \mathbb{I}_{(b,c]}(\cdot)$, the functional $\mathcal{H}_n(g, \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ simplifies as

$$
\begin{aligned}
\mathcal{H}_n(g, \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \int_{-\infty}^{+\infty} g\Big(\boldsymbol{a}^\top \boldsymbol{\mu} + \sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma} \, \boldsymbol{a}} \, z\Big) \, H_n(z) \, \phi(z) \, \mathrm{d}z \\
&= \int_{-\infty}^{+\infty} \mathbb{I}_{(b,c)}\Big(\boldsymbol{a}^\top \boldsymbol{\mu} + \sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma} \, \boldsymbol{a}} \, z\Big) \, H_n(z) \, \phi(z) \, \mathrm{d}z \\
&= \int_{-\infty}^{+\infty} \mathbb{I}\bigg(\frac{b - \boldsymbol{a}^\top \boldsymbol{\mu}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma} \, \boldsymbol{a}}} < z < \frac{c - \boldsymbol{a}^\top \boldsymbol{\mu}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma} \, \boldsymbol{a}}}\bigg) \, H_n(z) \, \phi(z) \, \mathrm{d}z,
\end{aligned}
$$

that is a univariate definite integral with lower and upper bounds z^- and z^+. Then, using the recurrent identity

$$
H_0(z) = 1, \qquad H_n(z)\phi(z) = -\frac{\mathrm{d}}{\mathrm{d}z} H_{n-1}(z)\phi(z), \qquad n \geq 1,
$$

and integrating with respect to z, we obtain

$$
\int_{z^-}^{z^+} H_n(z) \, \phi(z) \, \mathrm{d}z = \begin{cases} \big[\Phi(z)\big]_{z^-}^{z^+} & \text{if } n = 0, \\ (-1) \big[H_{n-1}(z)\phi(z)\big]_{z^-}^{z^+} & \text{if } n \geq 1, \end{cases}
$$

which concludes the proof. □

A.2 Optimal distributions and evidence lower bound

In this section we derive the explicit solutions of the optimal variational distributions $q^*(\sigma_\varepsilon^2)$, $q^*(\sigma_u^2)$ and $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ presented in Section 2.3. Moreover, we prove the statements in Propositions 2.7–2.8, that are concerned with the calculation of the Ψ-functions for the model specifications considered in Section 2.4.

Optimal distribution of σ_ε^2

We consider the pseudo-likelihood model for the i-th observation

$$\log \pi(y_i|\boldsymbol{\theta}) = -\frac{1}{\alpha} \log \sigma_\varepsilon^2 - \frac{1}{\alpha} \psi(y_i, \eta_i)/\sigma_\varepsilon^2,$$

and we recall the prior law $\sigma_\varepsilon^2 \sim \mathrm{IG}(A_\varepsilon, B_\varepsilon)$. Then, the Gibbs full-conditional distribution of σ_ε^2 can be easily obtained by standard calculations and corresponds to an Inverse-Gamma distribution $\mathrm{IG}(\tilde{A}_\varepsilon, \tilde{B}_\varepsilon)$, with parameters $\tilde{A}_\varepsilon = A_\varepsilon + \frac{n}{\alpha}$ and $\tilde{B}_\varepsilon = B_\varepsilon + \frac{1}{\alpha} \sum_{i=1}^n \psi(y_i, \eta_i)$. Now, computing the partial expectation of the log-full-conditional distribution, we obtain the optimal density $q^*(\sigma_\varepsilon^2) \propto \exp\left\{\mathbb{E}_{-\sigma_\varepsilon^2}(\log \pi(\sigma_\varepsilon^2 \mid \text{rest}))\right\}$, that is

$$\begin{aligned} \log q^*(\sigma_\varepsilon^2) &= -(A_\varepsilon + n/\alpha + 1) \log \sigma_\varepsilon^2 - \left\{ B_\varepsilon + \frac{1}{\alpha} \sum_{i=1}^n \mathbb{E}_q(\psi(y_i, \eta_i)) \right\} / \sigma_\varepsilon^2 + \text{const} \\ &= -(A_\varepsilon + n/\alpha + 1) \log \sigma_\varepsilon^2 - \left\{ B_\varepsilon + \frac{1}{\alpha} \sum_{i=1}^n \Psi_0(y_i, \hat{\eta}_i, \hat{\nu}_i) \right\} / \sigma_\varepsilon^2 + \text{const}. \end{aligned}$$

The latter is the kernel of an Inverse-Gamma distribution $\mathrm{IG}(\hat{A}_\varepsilon, \hat{B}_\varepsilon)$ where $\Psi_0(y_i, \hat{\eta}_i, \hat{\nu}_i) = \mathbb{E}_q\{\psi(y_i, \eta_i)\}$. Then, the proof is concluded. ☐

Optimal distribution of σ_u^2

We consider the joint prior model for $(\boldsymbol{u}, \sigma_u^2)$, which is

$$\boldsymbol{u} \mid \sigma_u^2 \sim \mathrm{N}_d(\mathbf{0}_d, \sigma_u^2 \mathbf{R}^{-1}), \qquad \sigma_u^2 \sim \mathrm{IG}(A_u, B_u).$$

Standard calculations leads to an Inverse-Gamma full-conditional distribution of σ_u^2, that is $\mathrm{IG}(\tilde{A}_u, \tilde{B}_u)$ with parameters $\tilde{A}_u = A_u + \frac{d}{2}$ and $\tilde{B}_u = B_u + \frac{1}{2} \boldsymbol{u}^\top \mathbf{R}\, \boldsymbol{u}$. Compute then the partial variational expectation of the log-full-conditional density in order to find the optimal approximation $q^*(\sigma_u^2) \propto \exp\left\{\mathbb{E}_{-\sigma_u^2}(\log \pi(\sigma_u^2 \mid \text{rest}))\right\}$:

$$\log q(\sigma_u^2) = -(A_u + d/2 + 1) \log \sigma_u^2 - \left\{ B_u + \tfrac{1}{2} \mathbb{E}_q(\boldsymbol{u}^\top \mathbf{R}\, \boldsymbol{u}) \right\} / \sigma_u^2 + \text{const}$$

Notice that the latter is the kernel of an Inverse-Gamma random variable $\mathrm{IG}(\hat{A}_u, \hat{B}_u)$ and $\mathbb{E}_q(\boldsymbol{u}^\top \mathbf{R}\, \boldsymbol{u}) = \hat{\boldsymbol{\mu}}_u^\top \mathbf{R}\, \hat{\boldsymbol{\mu}}_u + \text{trace}(\mathbf{R} \hat{\boldsymbol{\Sigma}}_{uu})$. This concludes the proof. ☐

Optimal distribution of $(\boldsymbol{\beta}, \boldsymbol{u})$

The evidence lower bound (2.17) as a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be expressed as

$$\begin{aligned}\underline{\ell}(\mathbf{y}; q, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \mathbb{E}_q\{\log \pi(\mathbf{y}, \boldsymbol{\theta})\} - \mathbb{E}_q\{\log q(\boldsymbol{\theta})\} \\ &= f(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \tfrac{1}{2}\,\mathrm{logdet}(\boldsymbol{\Sigma}) + \text{const},\end{aligned}$$

where $f(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathbb{E}_q\{\log \pi(\mathbf{y}, \boldsymbol{\theta})\}$ is a smooth function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, strongly convex with respect to $\boldsymbol{\mu}$, because of Proposition 2.4 and the specification of model (2.12)–(2.13). Since also $\mathrm{logdet}(\boldsymbol{\Sigma})$ is smooth and concave, the lower bound is a well-behaved function whose maximum must satisfy the following first order equations $\partial f/\partial \boldsymbol{\mu} = 0$ and $\partial f / \partial \boldsymbol{\Sigma} = \frac{1}{2}\boldsymbol{\Sigma}^{-1}$, where $\partial f/\partial\boldsymbol{\Sigma}$ denotes the $m \times m$ matrix having (i, j) entry equal to $(\partial f/\partial \boldsymbol{\Sigma})_{ij} = \partial f/\partial \Sigma_{ij}$. By definition, $\mathbf{g}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \partial f/\partial\boldsymbol{\mu}$, while $\mathbf{H}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2}\,(\partial f/\partial\boldsymbol{\Sigma})$ (see, for instance, Rohde and Wand, 2016 and Ormerod, 2011). Therefore, if a good starting value is provided, any contractive fixed-point algorithm climbing the lower bound surface has a limiting point corresponding to a local maximum.

Now, recall that $\mathbf{g}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{H}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are the gradient and Hessian of $f(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ calculated with respect to $\boldsymbol{\mu}$ and define $\mathbf{K} = \mathrm{blockdiag}\big\{\sigma_\beta^{-2}\mathbf{I}_p, \hat{\mu}_{1/\sigma_u^2}\mathbf{R}\big\}$, thus we have

$$\begin{aligned}f(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\hat{\mu}_{1/\sigma_\varepsilon^2}\mathbf{1}_n^\top\hat{\boldsymbol{\Psi}}_0/\alpha - \tfrac{1}{2}\boldsymbol{\mu}^\top\mathbf{K}\boldsymbol{\mu} - \tfrac{1}{2}\mathrm{trace}(\mathbf{K}\boldsymbol{\Sigma}) + \text{const}, \\ \mathbf{g}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\hat{\mu}_{1/\sigma_\varepsilon^2}\nabla_{\boldsymbol{\mu}}\big\{\mathbf{1}_n^\top\hat{\boldsymbol{\Psi}}_0\big\}/\alpha - \mathbf{K}\boldsymbol{\mu}, \\ \mathbf{H}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\hat{\mu}_{1/\sigma_\varepsilon^2}\nabla^2_{\boldsymbol{\mu}}\big\{\mathbf{1}_n^\top\hat{\boldsymbol{\Psi}}_0\big\}/\alpha - \mathbf{K}.\end{aligned}$$

Defining $\Psi_r(y_i, \hat{\eta}_i, \hat{\nu}_i) = \mathcal{F}_r(\psi(y_i, \cdot), \mathbf{c}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and applying Proposition A.2 in Appendix A, we obtain the gradient and Hessian

$$\begin{aligned}\nabla_{\boldsymbol{\mu}}\big\{\mathbf{1}_n^\top\hat{\boldsymbol{\Psi}}_0\big\} &= \sum_{i=1}^n \frac{\partial\hat{\eta}_i}{\partial\boldsymbol{\mu}}\hat{\Psi}_{1,i} = \sum_{i=1}^n \mathbf{c}_i\hat{\Psi}_{1,i} = \mathbf{C}^\top\hat{\boldsymbol{\Psi}}_1, \\ \nabla^2_{\boldsymbol{\mu}}\big\{\mathbf{1}_n^\top\hat{\boldsymbol{\Psi}}_0\big\} &= \sum_{i=1}^n \frac{\partial\hat{\eta}_i}{\partial\boldsymbol{\mu}}\frac{\partial\hat{\eta}_i}{\partial\boldsymbol{\mu}^\top}\hat{\Psi}_{2,i} = \sum_{i=1}^n \mathbf{c}_i\mathbf{c}_i^\top\hat{\Psi}_{2,i} = \mathbf{C}^\top\mathrm{diag}\big[\hat{\boldsymbol{\Psi}}_2\big]\mathbf{C}.\end{aligned}$$

This concludes the proof. □

Derivation of the evidence lower bound

First, consider the definition of the evidence lower bound and notice that it can be written in terms of a sum of expected values calculated with respect to the q-density as:

$$\underline{\ell}(\mathbf{y}; q) = \int_\Theta q(\boldsymbol{\theta})\log\left\{\frac{\pi(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right\}\mathrm{d}\boldsymbol{\theta} = \mathbb{E}_q\{\log \pi(\mathbf{y}, \boldsymbol{\theta})\} - \mathbb{E}_q\{\log q(\boldsymbol{\theta})\}$$

where $\log \pi(\mathbf{y}, \boldsymbol{\theta}) = \log \pi(\mathbf{y}|\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta})$. From the model specification, we have

$$\log \pi(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \pi(y_i|\boldsymbol{\theta}) + \log \pi(\boldsymbol{\beta}, \boldsymbol{u}|\sigma_u^2) + \log \pi(\sigma_u^2) + \log p(\sigma_\varepsilon^2).$$

Similarly, for the variational density we have:

$$\log q(\boldsymbol{\theta}) = \log q(\boldsymbol{\beta}, \boldsymbol{u}) + \log q(\sigma_u^2) + \log q(\sigma_\varepsilon^2).$$

Therefore, the lower bound can be decomposed as a sum of terms associated to different parameter blocks:

$$\underline{\ell}(\mathbf{y}; q) = \underbrace{\sum_{i=1}^{n} \mathbb{E}_q\big[\log p(y_i|\boldsymbol{\theta})\big]}_{T_1} + \underbrace{\mathbb{E}_q\left[\log \frac{p(\boldsymbol{\beta}, \boldsymbol{u}|\sigma_u^2)}{q(\boldsymbol{\beta}, \boldsymbol{u})}\right]}_{T_2} + \underbrace{\mathbb{E}_q\left[\log \frac{p(\sigma_u^2)}{q(\sigma_u^2)}\right]}_{T_3} + \underbrace{\mathbb{E}_q\left[\log \frac{p(\sigma_\varepsilon^2)}{q(\sigma_\varepsilon^2)}\right]}_{T_4}.$$

We can thus evaluate each term separately and sum up the individual contributions T_k, $k = 1, \dots, 4$. The first term of the lower bound is the variational expectation of the pseudo log-likelihood function:

$$\begin{aligned} T_1 &= \mathbb{E}_q\left[-\frac{n}{\alpha} \log \sigma_\varepsilon^2 - \frac{1}{\alpha} \sum_{i=1}^{n} \psi(y_i, \eta_i)/\sigma_\varepsilon^2 \right] \\ &= -\frac{n}{\alpha} \hat{\mu}_{\log \sigma_\varepsilon^2} - \frac{1}{\alpha} \hat{\mu}_{1/\sigma_\varepsilon^2} \sum_{i=1}^{n} \Psi_0(y_i, \hat{\eta}_i, \hat{\nu}_i). \end{aligned}$$

The second term is the expected contribution of $\boldsymbol{\vartheta} = (\boldsymbol{\beta}, \boldsymbol{u})$ to the lower bound:

$$\begin{aligned} T_2 = \mathbb{E}_q\Big[&-\tfrac{p}{2}\log(2\pi) - \tfrac{p}{2}\log\sigma_\beta^2 - \tfrac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta}/\sigma_\beta^2 \\ &- \tfrac{p}{2}\log(2\pi) - \tfrac{d}{2}\log\sigma_u^2 + \tfrac{1}{2}\text{logdet}(\mathbf{R}) - \tfrac{1}{2}\boldsymbol{u}^\top\mathbf{R}\boldsymbol{u}/\sigma_u^2 \\ &+ \tfrac{p+d}{2}\log(2\pi) + \tfrac{1}{2}\text{logdet}(\hat{\boldsymbol{\Sigma}}) + \tfrac{1}{2}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\mu}})\Big] \\ = -&\tfrac{p}{2}\log\sigma_\beta^2 - \tfrac{d}{2}\hat{\mu}_{\log\sigma_u^2} + \tfrac{1}{2}\text{logdet}(\mathbf{R}) - \tfrac{1}{2}\mathbb{E}_q\big[\boldsymbol{\beta}^\top\boldsymbol{\beta}/\sigma_\beta^2 + \boldsymbol{u}^\top\mathbf{R}\boldsymbol{u}/\sigma_u^2\big] \\ &+ \tfrac{1}{2}\text{logdet}(\hat{\boldsymbol{\Sigma}}) + \mathbb{E}_q\big[(\boldsymbol{\vartheta} - \hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\mu}})\big] \\ = -&\tfrac{p}{2}\log\sigma_\beta^2 - \tfrac{d}{2}\hat{\mu}_{\log\sigma_u^2} + \tfrac{1}{2}\text{logdet}(\mathbf{R}) - \tfrac{1}{2}\hat{\boldsymbol{\mu}}^\top\mathbf{K}\hat{\boldsymbol{\mu}} \\ &- \tfrac{1}{2}\text{trace}(\mathbf{K}\hat{\boldsymbol{\Sigma}}) + \tfrac{1}{2}\text{logdet}(\hat{\boldsymbol{\Sigma}}) + \tfrac{p+d}{2}, \end{aligned}$$

where we used the following identities:

$$\begin{aligned} \mathbb{E}_q\big[\boldsymbol{\beta}^\top\boldsymbol{\beta}/\sigma_\beta^2 + \boldsymbol{u}^\top\mathbf{R}\boldsymbol{u}/\sigma_u^2\big] &= \hat{\boldsymbol{\mu}}^\top\mathbf{K}\,\hat{\boldsymbol{\mu}} + \text{trace}(\mathbf{K}\hat{\boldsymbol{\Sigma}}), \\ \mathbb{E}_q\big[(\boldsymbol{\vartheta} - \hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\mu}})\big] &= p + d. \end{aligned}$$

The third term is the expected contribution of σ_u^2 to the lower bound:

$$\begin{aligned}T_3 &= \mathbb{E}_q\Big[A_u \log B_u - \log\Gamma(A_u) - (A_u+1)\log\sigma_u^2 - B_u/\sigma_u^2 \\ &\quad - (A_u + \tfrac{d}{2})\log\hat{B}_u + \log\Gamma(A_u + \tfrac{d}{2}) + (A_u + \tfrac{d}{2} + 1)\log\sigma_u^2 + \hat{B}_u/\sigma_u^2\Big] \\ &= \mathbb{E}_q\Big[A_u\log(B_u/\hat{B}_u) - \log\big\{\Gamma(A_u)/\Gamma(A_u + \tfrac{d}{2})\big\} \\ &\quad - \tfrac{d}{2}\log\hat{B}_u + \tfrac{d}{2}\log\sigma_u^2 - \big\{B_u - \hat{B}_u\big\}/\sigma_u^2\Big] \\ &= A_u\log(B_u/\hat{B}_u) - \log\big\{\Gamma(A_u)/\Gamma(A_u + \tfrac{d}{2})\big\} \\ &\quad - \tfrac{d}{2}\big\{\log\hat{B}_u - \hat{\mu}_{\log\sigma_u^2}\big\} - \big\{B_u - \hat{B}_u\big\}\hat{\mu}_{1/\sigma_u^2}.\end{aligned}$$

The fourth term is the expected contribution of σ_ε^2 to the lower bound:

$$\begin{aligned}T_4 &= \mathbb{E}_q\Big[A_\varepsilon \log B_\varepsilon - \log\Gamma(A_\varepsilon) - (A_\varepsilon+1)\log\sigma_\varepsilon^2 - B_\varepsilon/\sigma_\varepsilon^2 \\ &\quad - (A_\varepsilon + \tfrac{n}{\alpha})\log\hat{B}_\varepsilon + \log\Gamma(A_\varepsilon + \tfrac{n}{\alpha}) + (A_\varepsilon + \tfrac{n}{\alpha} + 1)\log\sigma_\varepsilon^2 + \hat{B}_\varepsilon/\sigma_\varepsilon^2\Big] \\ &= A_\varepsilon\log(B_\varepsilon/\hat{B}_\varepsilon) - \log\big\{\Gamma(A_\varepsilon)/\Gamma(A_\varepsilon + \tfrac{n}{\alpha})\big\} \\ &\quad - \tfrac{n}{\alpha}\big\{\log\hat{B}_\varepsilon - \hat{\mu}_{\log\sigma_\varepsilon^2}\big\} - \big\{B_\varepsilon - \hat{B}_\varepsilon\big\}\hat{\mu}_{1/\sigma_\varepsilon^2}.\end{aligned}$$

In the end, summing up the individual contributions T_1, T_2, T_3, T_4 and simplifying the redundant components we obtain the final result. ☐

Appendix B

B.1 EM algorithm and discretization

EM algorithm derivation

Let us consider the (unpenalized) log-likelihood for the i-th completed data pair (y_i, ω_i) induced by the augmented representation (3.10), that is

$$\begin{aligned}
\ell(\boldsymbol{\theta}; \omega_i, y_i) &= -\frac{3}{2}\log\sigma_\varepsilon^2 - \frac{1}{2}\log\omega_i - \frac{\omega_i}{\sigma_\varepsilon^2} - \frac{\{\varepsilon_i - a_1\omega_i\}^2}{2a_2^2\sigma_\varepsilon^2\omega_i} \\
&= -\frac{3}{2}n\log\sigma_\varepsilon^2 - \frac{1}{2}\log\omega_i - \frac{\omega_i}{\sigma_\varepsilon^2} - \frac{1}{2a_2^2\sigma_\varepsilon^2}\Big\{\varepsilon_i^2/\omega_i - 2a_1\varepsilon_i + a_1^2\omega_i\Big\} \\
&= -\frac{3}{2}n\log\sigma_\varepsilon^2 - \frac{1}{2}\log\omega_i - \frac{1}{2a_2^2\sigma_\varepsilon^2}\Big\{\varepsilon_i^2/\omega_i - 2a_1\varepsilon_i + (a_1^2 + 2a_2^2)\omega_i\Big\}
\end{aligned}$$

where $\varepsilon_i = y_i - f(\mathbf{p}_i)$ is the i-th residual, while $a_1 = \frac{1-2\tau}{\tau(1-\tau)}$ and $a_2^2 = \frac{2}{\tau(1-\tau)}$ denote non-stochastic constants, which only depend on the quantile level τ. The E-step of the EM algorithm thus prescribes to calculate $\mathbb{E}^{(k)}\{\ell(\boldsymbol{\theta}; \omega_i, y_i)\}$, that is the expectation of $\ell(\boldsymbol{\theta}; \omega_i, y_i)$ calculated with respect to the conditional distribution $\omega | y_i; \boldsymbol{\theta}$ evaluated in $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$. Doing this, using the linearity of the expectation, we obtain

$$\begin{aligned}
\mathbb{E}^{(k)}\{\ell(\boldsymbol{\theta}, \omega_i; y_i)\} &= -\frac{3}{2}\log\sigma_\varepsilon^2 - \frac{1}{2}\mu^{(k)}_{\log\omega_i} - \frac{1}{2a_2^2\sigma_\varepsilon^2}\Big\{\mu^{(k)}_{1/\omega_i}\varepsilon_i^2 - 2a_1\varepsilon_i + (a_1^2 + 2a_2^2)\mu^{(k)}_{\omega_i}\Big\} \\
&= -\frac{3}{2}\log\sigma_\varepsilon^2 - \frac{1}{2}\mu^{(k)}_{\log\omega_i} - \frac{1}{2}\frac{\mu^{(k)}_{1/\omega_i}}{a_2^2\sigma_\varepsilon^2}\varepsilon_i^2 + \frac{a_1}{a_2^2\sigma_\varepsilon^2}\varepsilon_i - \frac{1}{2}\frac{(a_1^2 + 2a_2^2)}{a_2^2\sigma_\varepsilon^2}\mu^{(k)}_{\omega_i},
\end{aligned}$$

where $\mu^{(k)}_{\omega_i} = \mathbb{E}^{(k)}(\omega_i)$, $\mu^{(k)}_{1/\omega_i} = \mathbb{E}^{(k)}(1/\omega_i)$ and $\mu^{(k)}_{\log\omega_i} = \mathbb{E}^{(k)}(\log\omega_i)$.

The explicit solution for such expectations may be obtained by exploiting the properties of the conditional law ω_i given y_i and $\boldsymbol{\theta}^{(k)}$, which is proportional to the Generalized-Inverse-Gaussian (GIG) distribution

$$\omega_i | y_i; \boldsymbol{\theta}^{(k)} \sim \mathrm{GIG}\bigg(\frac{1}{2}, \frac{a_1^2 + 2a_2^2}{a_2^2\sigma_\varepsilon^{2(k)}}, \frac{\{y_i - f^{(k)}(\mathbf{p}_i)\}^2}{a_2^2\sigma_\varepsilon^{2(k)}}\bigg).$$

This distribution can be easily obtained by discarding all the terms not depending on ω_i in $\ell(\boldsymbol{\theta}; \omega_i, y_i)$ and noting that the remainder is a function of ω_i proportional to a GIG log-density; refer to, e.g., Kozumi and Kobayashi (2011) and Tian *et al.* (2014) for more details. Now, thanks to standard properties of the GIG distribution (Jørgensen, 1982), we have

$$\mu^{(k)}_{\omega_i} = \{\mu^{(k)}_{1/\omega_i}\}^{-1} + \frac{a_2^2 \sigma_\varepsilon^{2(k)}}{a_1^2 + 2a_2^2}, \qquad \mu^{(k)}_{1/\omega_i} = \frac{(a_1^2 + 2a_2^2)^{1/2}}{|y_i - f^{(k)}(\mathbf{p}_i)|}.$$

Therefore, defining $w_i^{(k)} = \mu^{(k)}_{1/\omega_i}/a_2^2$ and $z_i^{(k)} = y_i - a_1/\mu^{(k)}_{1/\omega_i}$, summing all the individual contributions to the expected likelihood and the differential penalty in (3.6), we end up with

$$\begin{aligned}
\underline{\ell}^{(k)}_\lambda(\boldsymbol{\theta}; \mathbf{y}) &= \sum_{i=1}^n \mathbb{E}^{(k)}\{\ell(\boldsymbol{\theta}, \omega_i; y_i)\} - \frac{\lambda n}{2\sigma_\varepsilon^2} \int_\Omega (Lf - u)^2 \\
&= -\frac{3}{2} n \log \sigma_\varepsilon^2 - \frac{1}{2} \sum_{i=1}^n \mu^{(k)}_{\log \omega_i} - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^n \frac{(a_1^2 + 2a_2^2)}{a_2^2} \mu^{(k)}_{\omega_i} \\
&\quad - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^n w_i^{(k)} \{z_i^{(k)} - f(\mathbf{p}_i)\}^2 - \frac{\lambda n}{2\sigma_\varepsilon^2} \int_\Omega (Lf - u)^2.
\end{aligned}$$

Equivalently, we can write

$$\underline{\ell}^{(k)}_\lambda(\boldsymbol{\theta}; \mathbf{y}) = -\frac{3}{2} n \log \sigma_\varepsilon^2 - \frac{(a_1^2 + 2a_2^2)}{2a_2^2 \sigma_\varepsilon^2} \mathbf{1}_n^\top \boldsymbol{\mu}^{(k)}_\omega - \frac{n}{2\sigma_\varepsilon^2} \tilde{J}^{(k)}_\lambda(f),$$

where $\tilde{J}^{(k)}_\lambda(f)$ is a quadratic functional not depending on σ_ε^2:

$$\tilde{J}^{(k)}_\lambda(f) = \frac{1}{n} (\mathbf{z}^{(k)} - \mathbf{f}_n)^\top \mathbf{W}^{(k)} (\mathbf{z}^{(k)} - \mathbf{f}_n) + \lambda \int_\Omega (Lf - u)^2.$$

This concludes the derivation.

Finite element approximation

Derivation of Equation (3.20)

Let us start by considering the finite element system in Proposition 3.4 written in extended form:

$$\begin{cases} \frac{1}{n} \boldsymbol{\Psi}_{-s}^\top \mathbf{W}_{-s} \boldsymbol{\Psi}_{-s} \tilde{\mathbf{f}} + \lambda \mathbf{R}_1^\top \tilde{\mathbf{g}} + \boldsymbol{\Psi}_s^\top \tilde{\boldsymbol{\eta}} = \frac{1}{n} \boldsymbol{\Psi}_{-s}^\top \mathbf{W}_{-s} \mathbf{z}_{-s}, \\ \lambda \mathbf{R}_1 \tilde{\mathbf{f}} - \lambda \mathbf{R}_0 \tilde{\mathbf{g}} = \lambda(\mathbf{u} + \boldsymbol{\gamma}), \\ (\mathbf{z} - \boldsymbol{\Psi} \tilde{\mathbf{f}})_s = 0. \end{cases} \tag{B.1}$$

Solving the second equation with respect to $\tilde{\mathbf{g}}$, we have

$$\tilde{\mathbf{g}} = \mathbf{R}_0^{-1}(\mathbf{R}_1 \tilde{\mathbf{f}} - \mathbf{u} - \boldsymbol{\gamma}) = \mathbf{R}_0^{-1} \mathbf{R}_1 \tilde{\mathbf{f}} - \mathbf{R}_0^{-1}(\mathbf{u} + \boldsymbol{\gamma}).$$

If we substitute the expression for $\tilde{\mathbf{g}}$ in the first equation in (B.1), we get

$$(\boldsymbol{\Psi}_{-s}^\top \mathbf{W}_{-s}\boldsymbol{\Psi}_{-s} + \lambda\mathbf{R}_1^\top\mathbf{R}_0^{-1}\mathbf{R}_1)\,\tilde{\mathbf{f}} = \boldsymbol{\Psi}_{-s}^\top\mathbf{W}_{-s}\mathbf{z}_{-s} + \lambda\mathbf{R}_1^\top\mathbf{R}_0^{-1}(\mathbf{u}+\boldsymbol{\gamma}) - \boldsymbol{\Psi}_s^\top\tilde{\boldsymbol{\eta}}.$$

We then use the definition of $\mathbf{A} = \boldsymbol{\Psi}_{-s}^\top\mathbf{W}_{-s}\boldsymbol{\Psi}_{-s} + \lambda\mathbf{R}_1^\top\mathbf{R}_0^{-1}\mathbf{R}_1$, we denote $\mathbf{d} = \boldsymbol{\Psi}_{-s}^\top\mathbf{W}_{-s}\mathbf{z}_{-s} + \lambda\mathbf{R}_0^{-1}(\mathbf{u}+\boldsymbol{\gamma})$ and we solve for $\tilde{\mathbf{f}}$, obtaining $\tilde{\mathbf{f}} = \mathbf{A}^{-1}(\mathbf{d} - \boldsymbol{\Psi}_s^\top\tilde{\boldsymbol{\eta}})$. By plugging this expression for $\tilde{\mathbf{f}}$ in (B.1) and solving for $\tilde{\boldsymbol{\eta}}$, we get

$$\boldsymbol{\Psi}_s\mathbf{A}^{-1}(\mathbf{d} - \boldsymbol{\Psi}_s^\top\tilde{\boldsymbol{\eta}}) = \mathbf{z}_s \qquad \Rightarrow \qquad \tilde{\boldsymbol{\eta}} = (\boldsymbol{\Psi}_s\mathbf{A}^{-1}\boldsymbol{\Psi}_s^\top)^{-1}(\boldsymbol{\Psi}_s\mathbf{A}^{-1}\mathbf{d} - \mathbf{z}_s).$$

Finally, we use the notation $\mathbf{B} = \boldsymbol{\Psi}_s\mathbf{A}^{-1}\boldsymbol{\Psi}_s^\top$ and we substitute the above expression for $\tilde{\boldsymbol{\eta}}$ in $\tilde{\mathbf{f}} = \mathbf{A}^{-1}(\mathbf{d} - \boldsymbol{\Psi}_s^\top\tilde{\boldsymbol{\eta}})$:

$$\begin{aligned}
\tilde{\mathbf{f}} &= \mathbf{A}^{-1}(\mathbf{d} - \boldsymbol{\Psi}_s^\top\tilde{\boldsymbol{\eta}}) \\
&= \mathbf{A}^{-1}\big[\mathbf{d} - \boldsymbol{\Psi}_s^\top\mathbf{B}^{-1}(\boldsymbol{\Psi}_s\mathbf{A}^{-1}\mathbf{d} - \mathbf{z}_s)\big] \\
&= \mathbf{A}^{-1}\mathbf{d} + \mathbf{A}^{-1}\boldsymbol{\Psi}_s^\top\mathbf{B}^{-1}(\mathbf{z}_s - \boldsymbol{\Psi}_s\mathbf{A}^{-1}\mathbf{d}).
\end{aligned}$$

This concludes the proof.

Derivation of Equation (3.25)

The linear predictor is obtained by pre-multiplying the coefficient vector $\tilde{\mathbf{f}}$ by its design matrix $\boldsymbol{\Psi}$, namely

$$\begin{aligned}
\tilde{\mathbf{f}}_n &= \boldsymbol{\Psi}\tilde{\mathbf{f}} \\
&= \boldsymbol{\Psi}\mathbf{A}^{-1}\mathbf{d} + \boldsymbol{\Psi}\mathbf{A}^{-1}\boldsymbol{\Psi}_s^\top\mathbf{B}^{-1}(\mathbf{z}_s - \boldsymbol{\Psi}_s\mathbf{A}^{-1}\mathbf{d}) \\
&= \boldsymbol{\Psi}\mathbf{A}^{-1}\mathbf{d} - \boldsymbol{\Psi}\mathbf{A}^{-1}\boldsymbol{\Psi}_s^\top\mathbf{B}^{-1}\boldsymbol{\Psi}_s\mathbf{A}^{-1}\mathbf{d} + \boldsymbol{\Psi}\mathbf{A}^{-1}\boldsymbol{\Psi}_s^\top\mathbf{B}^{-1}\mathbf{z}_s \\
&= \boldsymbol{\Psi}\mathbf{A}^{-1}(\mathbf{I} - \boldsymbol{\Psi}_s^\top\mathbf{B}^{-1}\boldsymbol{\Psi}_s\mathbf{A}^{-1})\mathbf{d} + \boldsymbol{\Psi}\mathbf{A}^{-1}\boldsymbol{\Psi}_s^\top\mathbf{B}^{-1}\mathbf{z}_s.
\end{aligned}$$

If we let $\mathbf{C} = \mathbf{I} - \boldsymbol{\Psi}_s^\top\mathbf{B}^{-1}\boldsymbol{\Psi}_s\mathbf{A}^{-1}$, then

$$\begin{aligned}
\tilde{\mathbf{f}}_n &= \boldsymbol{\Psi}\mathbf{A}^{-1}\mathbf{C}\mathbf{d} + \boldsymbol{\Psi}\mathbf{A}^{-1}\boldsymbol{\Psi}_s^\top\mathbf{B}^{-1}\mathbf{z}_s \\
&= \boldsymbol{\Psi}\mathbf{A}^{-1}\mathbf{C}\{\boldsymbol{\Psi}_{-s}^\top\mathbf{W}_{-s}\mathbf{z}_{-s} + \lambda\mathbf{R}_0^{-1}(\mathbf{u}+\boldsymbol{\gamma})\} + \boldsymbol{\Psi}\mathbf{A}^{-1}\boldsymbol{\Psi}_s^\top\mathbf{B}^{-1}\mathbf{z}_s \\
&= \boldsymbol{\Psi}\mathbf{A}^{-1}\mathbf{C}\boldsymbol{\Psi}_{-s}^\top\mathbf{W}_{-s}\mathbf{z}_{-s} + \boldsymbol{\Psi}\mathbf{A}^{-1}\boldsymbol{\Psi}_s^\top\mathbf{B}^{-1}\mathbf{z}_s + \lambda\boldsymbol{\Psi}\mathbf{A}^{-1}\mathbf{C}\mathbf{R}_1^\top\mathbf{R}_0^{-1}(\mathbf{u}+\boldsymbol{\gamma}) \\
&= \mathbf{S}_{-s}\mathbf{z}_{-s} + \mathbf{S}_s\mathbf{z}_s + \mathbf{r},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{S}_{-s} &= \boldsymbol{\Psi}\mathbf{A}^{-1}\mathbf{C}\,\boldsymbol{\Psi}_{-s}^\top\mathbf{W}_{-s}, \\
\mathbf{S}_s &= \boldsymbol{\Psi}\mathbf{A}^{-1}\boldsymbol{\Psi}_s^\top\mathbf{B}^{-1}, \\
\mathbf{r} &= \lambda\boldsymbol{\Psi}\mathbf{A}^{-1}\mathbf{C}\mathbf{R}_1^\top\mathbf{R}_0^{-1}(\mathbf{u}+\boldsymbol{\gamma}).
\end{aligned}$$

Therefore, the smoothing matrix $\mathbf{S} = \left[\; \mathbf{S}_{-s},\, \mathbf{S}_s \;\right]$ is defined according to a block-structure, separating the constrained and unconstrained smoothing factors.

The effective degrees of freedom is defined as $\mathrm{df} = \mathrm{tr}(\mathbf{S})$. In particular, the trace of $\mathbf{S}$ is equal to

$$\begin{aligned}
\mathrm{tr}(\mathbf{S}) &= \mathrm{tr}\begin{bmatrix} \boldsymbol{\Psi}_{-s}\mathbf{A}^{-1}\mathbf{C}\,\boldsymbol{\Psi}_{-s}^{\top}\mathbf{W}_{-s} & \boldsymbol{\Psi}_{-s}\mathbf{A}^{-1}\boldsymbol{\Psi}_{s}^{\top}\mathbf{B}^{-1} \\ \boldsymbol{\Psi}_{s}\mathbf{A}^{-1}\mathbf{C}\,\boldsymbol{\Psi}_{-s}^{\top}\mathbf{W}_{-s} & \boldsymbol{\Psi}_{s}\mathbf{A}^{-1}\boldsymbol{\Psi}_{s}^{\top}\mathbf{B}^{-1} \end{bmatrix} \\
&= \mathrm{tr}\big(\boldsymbol{\Psi}_{-s}\mathbf{A}^{-1}\mathbf{C}\,\boldsymbol{\Psi}_{-s}^{\top}\mathbf{W}_{-s}\big) + \mathrm{tr}\big(\boldsymbol{\Psi}_{s}\mathbf{A}^{-1}\boldsymbol{\Psi}_{s}^{\top}\mathbf{B}^{-1}\big) \\
&= \mathrm{tr}\big(\mathbf{A}^{-1}\mathbf{C}\,\boldsymbol{\Psi}_{-s}^{\top}\mathbf{W}_{-s}\boldsymbol{\Psi}_{-s}\big) + \mathrm{tr}\big(\boldsymbol{\Psi}_{s}\mathbf{A}^{-1}\boldsymbol{\Psi}_{s}^{\top}\mathbf{B}^{-1}\big) \\
&= \mathrm{tr}\big(\mathbf{A}^{-1}\mathbf{C}\,\boldsymbol{\Psi}_{-s}^{\top}\mathbf{W}_{-s}\boldsymbol{\Psi}_{-s}\big) + \mathrm{tr}\big(\mathbf{B}\mathbf{B}^{-1}\big) \\
&= \mathrm{tr}\big(\mathbf{A}^{-1}\mathbf{C}\,\boldsymbol{\Psi}_{-s}^{\top}\mathbf{W}_{-s}\boldsymbol{\Psi}_{-s}\big) + |s|.
\end{aligned}$$

This concludes the derivation.

Derivation of Equation (3.30)

Let us recall the definition of $\mathbf{A}_c$, $\mathbf{B}_c$ and $\mathbf{d}_c$, that are

$$\mathbf{A}_c = \mathbf{C}_{-s}^{\top}\mathbf{W}_{-s}\mathbf{C}_{-s} + \lambda\mathbf{P}, \qquad \mathbf{B}_c = \mathbf{C}_s\mathbf{A}^{-1}\mathbf{C}_s^{\top}, \qquad \mathbf{d}_c = \mathbf{C}_{-s}\mathbf{W}_{-s}\mathbf{z}_{-s} + \lambda\mathbf{h}_c$$

where $\mathbf{C}$ is the completed design matrix, while $\mathbf{P}_c$ and $\mathbf{h}_c$ are defined in (3.29). Then, using $\tilde{\mathbf{g}} = \mathbf{R}_0^{-1}\mathbf{R}_1\tilde{\mathbf{f}} + \mathbf{R}_0^{-1}(\mathbf{u} + \boldsymbol{\gamma})$, the system in (3.30) can be formulated as

$$\begin{bmatrix} \mathbf{A}_c & \mathbf{C}_s^{\top} \\ \mathbf{C}_s^{\top} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\theta}}_c \\ \tilde{\boldsymbol{\eta}} \end{bmatrix} = \begin{bmatrix} \mathbf{d}_c \\ \mathbf{z}_s \end{bmatrix}$$

Solving the system with respect to $\tilde{\boldsymbol{\theta}}_c$, we get

$$\tilde{\boldsymbol{\theta}}_c = \mathbf{A}_c^{-1}(\mathbf{d}_c - \mathbf{C}_s^{\top}\tilde{\boldsymbol{\eta}}).$$

Thus, the solution for $\tilde{\boldsymbol{\eta}}$ is

$$\tilde{\boldsymbol{\eta}} = (\mathbf{C}_s\mathbf{A}_c^{-1}\mathbf{C}_s^{\top})^{-1}(\mathbf{z}_s - \mathbf{C}_s\mathbf{A}_c^{-1}\mathbf{d}_c),$$

which leads to

$$\tilde{\boldsymbol{\theta}}_c = \mathbf{A}_c^{-1}(\mathbf{d}_c - \mathbf{C}_s^{\top}\mathbf{B}_c^{-1}(\mathbf{z}_s - \mathbf{C}_s\mathbf{A}_c^{-1}\mathbf{d}_c)).$$

This concludes the proof.

B.2 Large sample properties

Proof of Theorem 3.7

In order to prove Theorems 3.7, we first need to rephrase Problem 1 in a more tractable way. The key idea is to represent the convex functional (3.6) as the sum of a quadratic functional plus a negligible term vanishing in the limit when n goes to infinity. Then, the asymptotic properties of $\hat{f}$ may be derived by adapting existing results for penalized nonparametric estimators based on quadratic variational problems. Similar approaches have been used by Pollard (1991) and Knight (1998) (least absolute regression), Knight and Fu (2000) (lasso type estimators), Kato (2009) (linear quantile regression and general argmin processes).

Let $f_i = f(\mathbf{p}_i)$ be the evaluation of f at $\mathbf{p}_i$, let $\varepsilon_i = y_i - f_0(\mathbf{p}_i)$ be the true i-th quantile residual, let $\pi_{\varepsilon_i|\mathbf{p}_i}(\varepsilon) = \pi_{y_i|\mathbf{p}_i}(f_{0,i} + \varepsilon)$ and $\Pi_{\varepsilon_i|\mathbf{p}_i}(\varepsilon) = \Pi_{y_i|\mathbf{p}_i}(f_{0,i} + \varepsilon)$ be, respectively, the probability and cumulative density functions of ε_i given $\mathbf{p}_i$. Furthermore, we denote $\pi_i = \pi_{\varepsilon_i|\mathbf{p}_i}(0) = \pi_{y_i|\mathbf{p}_i}(f_{0,i})$. Then, we introduce the reparametrization $\delta = f - f_0 \in \mathcal{F}_0$ and the relative estimator $\hat{\delta} = \hat{f} - f_0 \in \mathcal{F}_0$, which can be found by minimizing the functional

$$J_n^*(\delta) = \underbrace{\frac{1}{n}\sum_{i=1}^n \Big[\rho_\tau(\varepsilon_i - \delta_i) - \rho_\tau(\varepsilon_i)\Big]}_{= S_n(\delta)} + \frac{\lambda}{2} \underbrace{\int_\Omega \Big[(L\delta)^2 + 2(L\delta)g_0 + g_0^2\Big]}_{= P(\delta)} . \tag{B.2}$$

Here, $S_n(\delta)$ is a loss function that measures the misfit between ε_i and δ_i, and $P(\delta)$ is a regularization term which only depends on δ, on Ω and on the PDE specification.

We recall the definition of $L^2(\Omega)$ inner product and norm

$$\langle \psi, f\rangle_{L^2} = \int_\Omega \psi f, \qquad \|f\|_{L^2} = \sqrt{\langle f, f\rangle_{L^2}}, \qquad \forall\, \psi, f \in L^2(\Omega).$$

We recall the definition of $H^2(\Omega)$ norm

$$\|f\|_{H^2} = \left(\sum_{\alpha=0}^{2} \|D^\alpha f\|_{L^2}^2\right)^{1/2}, \qquad \forall f \in H^2(\Omega).$$

We define the discrete inner product $\langle \psi, f\rangle_n$ and the induced norm $\|f\|_n$ as

$$\langle \psi, f\rangle_n = \frac{1}{n}\sum_{i=1}^n \psi_i f_i = \int_\Omega \psi f \,\mathrm{d}\Pi_{\mathbf{p}}^n, \qquad \|f\|_n = \sqrt{\langle f, f\rangle_n}.$$

Furthermore, we define the dicretized $L^2(\Omega)$ space, say $L_n^2(\Omega)$, as the Hilbert space endowed with the inner product $\langle \cdot, \cdot\rangle_n$, the norm $\|\cdot\|_n$ and such that any function $f \in L_n^2(\Omega)$ satisfies $\|f\|_n < \infty$. The space $L_n^2(\Omega)$ is then the finite-sample approximation of $L^2(\Omega)$ induced by the empirical distribution $\Pi_{\mathbf{p}}^n$. Recall that, thanks to Assumption 3, $\Pi_{\mathbf{p}}^n$ uniformly converges to $\Pi_{\mathbf{p}}$, and $\big|\int_\Omega f^2 \,\mathrm{d}(\Pi_{\mathbf{p}} - \Pi_{\mathbf{p}}^n)\big| \xrightarrow{\mathrm{P}} 0$, meaning that $\|f\|_n \xrightarrow{\mathrm{P}} \int_\Omega f^2 \,\mathrm{d}\Pi_{\mathbf{p}}$.

Lemma B.1. *Under Assumptions 2 and 3, we have*

$$S_n(\delta) = \frac{1}{n}\sum_{i=1}^n \int_0^{\delta_i} \big[\,\Pi_{y_i|\mathbf{p}_i}(t) - \tau\,\big]\mathrm{d}t - \frac{1}{n}\sum_{i=1}^n \delta_i x_i, \tag{B.3}$$

where $x_1, \dots, x_n$ *is a sequence of independent, bounded random variables such that* $\mathbb{E}(x_i) = 0$ *and* $\mathbb{E}(x_i^2) < \infty$.

Proof. Let $S(\varepsilon_i, \delta_i) = \rho_\tau(\varepsilon_i - \delta_i) - \rho_\tau(\varepsilon_i)$ be the i-th data contribution to the loss function $S_n(\delta)$ and recall the identity $\rho_\tau(x) = \frac{1}{2}|x| + (\tau - \frac{1}{2})x$. Then, $S_n(\delta)$ may be written as

$$S_n(\delta) = \frac{1}{n}\sum_{i=1}^n S(\varepsilon_i, \delta_i) = \frac{1}{n}\sum_{i=1}^n \Big\{ \tfrac{1}{2}\big(|\varepsilon_i - \delta_i| - |\varepsilon_i|\big) - \big(\tau - \tfrac{1}{2}\big)\delta_i \Big\}.$$

From this representation, it is easy to show that $-|\delta_i| \leq |\varepsilon_i - \delta_i| - |\varepsilon_i| \leq +|\delta_i|$, hence, for any $|\delta_i| < \infty$, the random function $S(\varepsilon_i, \delta_i)$ is bounded by

$$|S(\varepsilon_i, \delta_i)| \leq \big(\tfrac{1}{2} + |\tau - \tfrac{1}{2}|\big)|\delta_i| \leq |\delta_i|,$$

and it thus has finite first and second moments.

In order to determine the mean of $S_n(\delta)$, we first recall the decomposition proposed by, e.g., Pollard (1991), Knight (1998) and Knight and Fu (2000), that is

$$\begin{aligned} S(\varepsilon_i, \delta_i) &= \tfrac{1}{2}\big(|\varepsilon_i - \delta_i| - |\varepsilon_i|\big) - \big(\tau - \tfrac{1}{2}\big)\delta_i \\ &= \tfrac{1}{2}(\varepsilon_i - \delta_i)\,\mathrm{sign}(\varepsilon_i - \delta_i) - \tfrac{1}{2}\varepsilon_i\,\mathrm{sign}(\varepsilon_i) - \big(\tau - \tfrac{1}{2}\big)\delta_i \\ &= \tfrac{1}{2}(\varepsilon_i - \delta_i)\big[\mathrm{sign}(\varepsilon_i - \delta_i) - \mathrm{sign}(\varepsilon_i)\big] + \tfrac{1}{2}\delta_i\,\mathrm{sign}(\varepsilon_i) - \big(\tau - \tfrac{1}{2}\big)\delta_i \\ &= \int_0^{\delta_i} \big[\mathbb{I}(\varepsilon_i < t) - \mathbb{I}(\varepsilon_i < 0)\big]\mathrm{d}t - \delta_i\big[\tau - \mathbb{I}(\varepsilon_i < 0)\big], \end{aligned}$$

where $\mathrm{sign}(\varepsilon_i)$ is the sign function, equal to 0 if $\varepsilon_i = 0$, -1 if $\varepsilon_i < 0$, and $+1$ if $\varepsilon_i > 0$. Taking the expectation, we get

$$\begin{aligned} \mathbb{E}\{S(\varepsilon_i, \delta_i)\} &= \int_0^{\delta_i} \big[\mathbb{P}(\varepsilon_i < t) - \mathbb{P}(\varepsilon_i < 0)\big]\mathrm{d}t - \delta_i\big[\tau - \mathbb{P}(\varepsilon_i < 0)\big] \\ &= \int_0^{\delta_i} \big[\mathbb{P}(\varepsilon_i < t) - \tau\big]\mathrm{d}t, \end{aligned}$$

since $\mathbb{P}(\varepsilon_i < 0) = \mathbb{P}(y_i < f_{0,i}) = \tau$. The variance is finite as well and can be bounded above by the inequality

$$\mathrm{Var}\{S(\varepsilon_i, \delta_i)\} \leq \mathbb{E}\big\{|S(\varepsilon_i, \delta_i)|^2\big\} \leq |\delta_i|^2.$$

Then, the finite-sample mean and variance of $S_n(\delta)$ are given by

$$\mathbb{E}\{S_n(\delta)\} = \frac{1}{n}\sum_{i=1}^n \int_0^{\delta_i} \big[\Pi_{\varepsilon_i|\mathbf{p}_i}(t) - \tau\big]\mathrm{d}t, \quad \mathrm{Var}\{S_n(\delta)\} \leq \frac{1}{n^2}\sum_{i=1}^n \delta_i^2 = \frac{\|\delta\|_n^2}{n},$$

where $\Pi_{\varepsilon_i|\mathbf{p}_i}(t) = \mathbb{P}(\varepsilon_i < t)$ and $\mathrm{Var}\{S_n(\delta)\} = O(n^{-1}\|\delta\|_n^2)$. As a consequence, we may write $S_n(\delta)$ as the sum of a deterministic term, $\mathbb{E}\{S_n(\delta)\}$, and a zero mean stochastic term of order $O_{\mathrm{p}}(n^{-1/2}\|\delta\|_n)$:

$$\begin{aligned}
S_n(\delta) &= \mathbb{E}\{S_n(\delta)\} + \big[S_n(\delta) - \mathbb{E}\{S_n(\delta)\}\big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\delta_i}\big[\Pi_{\varepsilon_i|\mathbf{p}_i}(t) - \tau\big]\mathrm{d}t - \frac{1}{n}\sum_{i=1}^{n}\delta_i x_i,
\end{aligned}$$

where $x_1,\dots,x_n$ is a sequence of independent random variables such that $\mathbb{E}(x_i) = 0$ and $\mathbb{E}(x_i^2) < \infty$. This concludes the proof. □

Let us denote by $\overset{n}{=}$ the equality over the $\mathbf{p}_i$ point locations, that is, for any pair of functions δ, ψ, $\delta \overset{n}{=} \psi$ if and only if $\delta(\mathbf{p}_i) = \psi(\mathbf{p}_i)$ for all $i = 1,\dots,n$.

Remark B.2. Let us denote $\bar{S}_n(\delta) = \mathbb{E}\{S_n(\delta)\}$, which is a non-negative, differentiable, strictly convex, coercive functional in $\delta \in L_n^2(\Omega)$ attaining its minimum at $\delta \overset{n}{=} 0$. Indeed, for any $\psi \in L_n^2(\Omega)$, and in particular for $\psi \in H^2(\Omega) \subset L^2(\Omega) \cap C(\bar{\Omega})$, we have

$$\begin{aligned}
\bar{S}_n(\delta)\Big|_{\delta=0} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\delta_i}\big[\,\Pi_{\varepsilon_i|\mathbf{p}_i}(t) - \tau\,\big]\,\mathrm{d}t\,\Big|_{\delta=0} = 0, \\
\partial_\psi \bar{S}_n(\delta)\Big|_{\delta=0} &= \frac{1}{n}\sum_{i=1}^{n}\psi_i\big[\,\Pi_{\varepsilon_i|\mathbf{p}_i}(\delta_i) - \tau\,\big]\,\Big|_{\delta=0} = 0, \\
\partial_\psi^2 \bar{S}_n(\delta)\Big|_{\delta=0} &= \frac{1}{n}\sum_{i=1}^{n}\psi_i^2\,\pi_{\varepsilon_i|\mathbf{p}_i}(\delta_i)\,\Big|_{\delta=0} = \frac{1}{n}\sum_{i=1}^{n}\psi_i^2\pi_i > 0,
\end{aligned}$$

where ∂_ψ and ∂_ψ^2 denote the first and second order directional derivatives along ψ. The last inequality follows by Assumption 2. Moreover, thanks to the law of large numbers, $S_n(\delta) \xrightarrow{\mathrm{p}} \bar{S}(\delta)$ for $n \to \infty$, where $\delta = 0$ is the unique minimizer of $\bar{S}(\delta)$ with $\bar{S}(0) = 0$, $\partial_\psi \bar{S}(0) = 0$ and $\partial_\psi^2 \bar{S}(0) = \int_\Omega \pi\psi^2\,\mathrm{d}\Pi_{\mathbf{p}} \geq h_1 k_1 \|\psi\|_{L^2}^2 > 0$ (see Assumptions 2 and 3).

Remark B.3. Because of Lemma B.1, for any $\delta \in L_n^2(\Omega)$, the second order Taylor expansion of $S_n(\delta)$ in a neighborhood of $\delta \overset{n}{=} 0$ is

$$S_n(\delta) = \left[\frac{1}{2n}\sum_{i=1}^{n}\pi_i\delta_i^2\right](1 + o(1)) - \frac{1}{n}\sum_{i=1}^{n}\delta_i x_i.$$

The reparametrized objective functional (B.2) can thus be expressed as

$$\begin{aligned}
J_n^*(\delta) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\delta_i}\big[\Pi_{\varepsilon_i|\mathbf{p}_i}(t) - \tau\big]\mathrm{d}t - \frac{1}{n}\sum_{i=1}^{n}\delta_i x_i + \frac{\lambda}{2}\int_\Omega\Big[(L\delta)^2 + 2(L\delta)g_0 + g_0^2\Big] \\
&= \left[\frac{1}{2n}\sum_{i=1}^{n}\pi_i\delta_i^2\right](1 + o(1)) - \frac{1}{n}\sum_{i=1}^{n}\delta_i x_i + \frac{\lambda}{2}\int_\Omega\Big[(L\delta)^2 + 2(L\delta)g_0 + g_0^2\Big].
\end{aligned}$$

Therefore, for any $\psi \in \mathcal{F}_0$, the estimator $\hat{\delta}$ must satisfy the first order condition

$$\frac{1}{n}\sum_{i=1}^{n}\psi_i\big[\,\Pi_{\varepsilon_i|\mathbf{p}_i}(\hat{\delta}_i)-\tau\,\big]-\frac{1}{n}\sum_{i=1}^{n}\psi_i x_i+\lambda\int_\Omega\Big[(L\psi)(L\hat{\delta})+(L\psi)g_0\Big]=0.$$

By expanding $\Pi_{\varepsilon_i|\mathbf{p}_i}(\delta_i)-\Pi_{\varepsilon_i|\mathbf{p}_i}(0)=\pi_{\varepsilon_i|\mathbf{p}_i}(0)\,\delta_i+o(|\delta_i|)$ with a first order Taylor approximation around $\delta_i=0$, we get

$$\left[\frac{1}{n}\sum_{i=1}^{n}\pi_i\psi_i\hat{\delta}_i\right](1+o(1))-\frac{1}{n}\sum_{i=1}^{n}\psi_i x_i+\lambda\int_\Omega\Big[(L\psi)(L\hat{\delta})+(L\psi)g_0\Big]=0,$$

which is equivalent to the following first order equation in the original parametrization $f=\delta+f_0$:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\pi_i\psi_i(\hat{f}_i-f_{0,i})\right](1+o(1))-\frac{1}{n}\sum_{i=1}^{n}\psi_i x_i+\lambda\int_\Omega(L\psi)(L\hat{f}-u)=0, \tag{B.4}$$

and, because of the linearity of (B.4), the quantile estimator $\hat{f}$ may be decomposed in the additive form $\hat{f}=\hat{f}^*+\hat{w}$, where $\hat{f}^*\in\mathcal{F}_\gamma$ and $\hat{w}\in\mathcal{F}_0$ solve, respectively,

$$\left[\frac{1}{n}\sum_{i=1}^{n}\pi_i\psi_i(\hat{f}_i^*-f_{0,i})\right](1+o(1))+\lambda\int_\Omega(L\psi)(L\hat{f}^*-u)=0, \tag{B.5}$$

$$\left[\frac{1}{n}\sum_{i=1}^{n}\pi_i\psi_i\hat{w}_i\right](1+o(1))-\frac{1}{n}\sum_{i=1}^{n}\psi_i x_i+\lambda\int_\Omega(L\psi)(L\hat{w})=0, \tag{B.6}$$

for all $\psi\in\mathcal{F}_0$. Equation (B.5) is purely deterministic and only involves the probability density function $\pi_{y|\mathbf{p}}(\cdot)$, the true quantile field $f_0(\cdot)$, and the non-homogeneous regularization terms u and γ. Equation (B.6) instead solves a noisy problem, which involves the random variables $x_1,\dots,x_n$ and a homogeneous regularization term. Therefore, $\hat{f}^*$ is purely deterministic, while $\hat{w}$ is such that $\mathbb{E}(\hat{w})=0$. As a consequence, $\mathbb{E}(\hat{f})=\hat{f}^*$ and $\mathrm{Var}(\hat{f})=\mathrm{Var}(\hat{w})$. Thanks to this fact, we can split the bias and variance analysis of the estimator. Actually, the bias only depends on (B.5), whereas the variance only depends on (B.6).

Asymptotic bias

Now we study the asymptotic behavior of the bias of the estimator

$$\mathcal{B}=f_0-\mathbb{E}(\hat{f})=f_0-\hat{f}^*\in\mathcal{F}_0(\Omega)$$

with respect to the number of observations n and the smoothing parameter λ. Doing this, we consider different Sobolev regularity cases for f_0, namely $f_0\in H^2(\Omega)$ and $f_0\in H^4(\Omega)$. We further make use of the of fractional Sobolev spaces $H^\theta(\Omega)$, with non-integer $\theta>0$, where $H^\theta(\Omega)$ is the interpolation between $H^k(\Omega)$ and $L^2(\Omega)$, with integer

$k > \theta$. We also consider the adjoint operator L^*, defined as

$$L^* g = -\mathrm{div}(\mathbf{K}\nabla g) - \mathbf{b} \cdot \nabla g + (c - \mathrm{div}(\mathbf{b}))g.$$

Finally, we recall the following result by Cox (1984).

Lemma B.4. *Under Assumption 3, if $\partial\Omega \in C^2(\mathbb{R})$, for all $h, g \in H^2(\Omega)$, there exists a constant $c > 0$ such that*

$$\left| \int_\Omega hg \,\mathrm{d}\Pi_{\mathbf{p}} - \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{p}_i)g(\mathbf{p}_i) \right| = \left| \int_\Omega hg \,\mathrm{d}(\Pi_{\mathbf{p}} - \Pi_{\mathbf{p}}^n) \right| \le c\, d_n \|h\|_{H^2} \|g\|_{H^2}$$

where $d_n = \sup_{\mathbf{p}\in\Omega} |\,\Pi_{\mathbf{p}}(\mathbf{p}) - \Pi_{\mathbf{p}}^n(\mathbf{p})|$.

Lemma B.5. *Under Assumptions 2–5, for n sufficiently large, if $f_0 \in H^2(\Omega)$ and $Bf_0 = \gamma$, then $\|\mathrm{Bias}(\hat{f})\|_{L^2} = O(\lambda^{1/2})$. Moreover, if $Bf_0 = \gamma$ and $g_0 \in H^2(\Omega)$, then $\|\mathrm{Bias}(\hat{f})\|_{L^2} = O(\lambda)$ and $\|\mathrm{Bias}(\hat{f})\|_{H^2} = O(\lambda^{1/2})$.*

Proof. As shown in the derivation of the system (B.5)–(B.6), finding $\hat{f}^*$ is equivalent to solving the first order equation

$$\lambda \int_\Omega (L\psi)(L\hat{f}^* - u) = \frac{1}{n}\sum_{i=1}^{n} \pi_i \psi_i \left[f_{0,i} - \hat{f}_i^* \right](1 + o(1)),$$

for all $\psi \in \mathcal{F}_0$. Let us express such an equation in terms of $\mathcal{B} = f_0 - \hat{f}^*$:

$$\lambda \int_\Omega (L\psi)(L\mathcal{B}) = \lambda \int_\Omega (L\psi) g_0 - \left[\frac{1}{n}\sum_{i=1}^{n} \pi_i \psi_i \mathcal{B}_i \right](1 + o(1)).$$

We add $\int_\Omega \pi\psi\mathcal{B}\,\mathrm{d}\Pi_{\mathbf{p}}$ on both sides, obtaining

$$\lambda \int_\Omega (L\psi)(L\mathcal{B}) + \int_\Omega \pi\psi\mathcal{B}\,\mathrm{d}\Pi_{\mathbf{p}} = \lambda \int_\Omega (L\psi) g_0 + \left[\int_\Omega \pi\psi\mathcal{B}\,\mathrm{d}(\Pi_{\mathbf{p}} - \Pi_{\mathbf{p}}^n) \right](1 + o(1)).$$

Such an equation must hold for any $\psi \in \mathcal{F}_0$ and, in particular, for $\psi = \mathcal{B}$. Hence, thanks to Assumptions 2 and 3, we get

$$\int_\Omega \pi\mathcal{B}^2 \,\mathrm{d}\Pi_{\mathbf{p}} \ge h_1 \int_\Omega \mathcal{B}^2 \,\mathrm{d}\Pi_{\mathbf{p}} \ge k_1 h_1 \int_\Omega \mathcal{B}^2 = k_1 h_1 \|\mathcal{B}\|_{L^2}^2,$$

which leads to

$$\lambda \|L\mathcal{B}\|_{L^2}^2 + c_1 \|\mathcal{B}\|_{L^2}^2 \le \lambda \int_\Omega (L\mathcal{B}) g_0 + \left[\int_\Omega \pi\mathcal{B}^2 \mathrm{d}(\Pi_{\mathbf{p}} - \Pi_{\mathbf{p}}^n) \right](1 + o(1)),$$

where $c_1 = k_1 h_1$. The second term on the right-hand side may be upper-bounded by

$$\left[\int_\Omega \pi\mathcal{B}^2 \mathrm{d}(\Pi_{\mathbf{p}} - \Pi_{\mathbf{p}}^n) \right](1 + o(1)) \le c\, h_2 d_n \|\mathcal{B}\|_{H^2}^2$$

for some positive constant c independent on $\mathcal{B}$ and n. Moreover, thanks to the H^2-regularity and the equivalence between the norms $\|L\psi\|_{L^2}$ and $\|\psi\|_{H^2}$ for any $\psi \in \mathcal{F}_0$, there exists a constant c_{L} only depending on Ω and L such that $c_{\mathrm{L}}\|\mathcal{B}\|^2_{H^2} \leq \|L\mathcal{B}\|^2_{L^2}$. Using these two inequalities, we obtain

$$\lambda c_{\mathrm{L}}\|\mathcal{B}\|^2_{H^2} + c_1\|\mathcal{B}\|^2_{L^2} \leq \lambda \int_\Omega (L\mathcal{B})g_0 + c\, h_2 d_n \|\mathcal{B}\|^2_{H^2}.$$

Because of Assumption 4, for n large enough that $d_n/\lambda \leq c_{\mathrm{L}}/(2ch_2)$, we can write

$$\lambda c_{\mathrm{L}}\|\mathcal{B}\|^2_{H^2} + c_1\|\mathcal{B}\|^2_{L^2} \leq \lambda \int_\Omega (L\mathcal{B})g_0 + \frac{\lambda c_{\mathrm{L}}}{2}\|\mathcal{B}\|^2_{H^2},$$

and hence

$$\frac{\lambda c_{\mathrm{L}}}{2}\|\mathcal{B}\|^2_{H^2} + c_1\|\mathcal{B}\|^2_{L^2} \leq \lambda \int_\Omega (L\mathcal{B})g_0.$$

Furthermore, using inequality (13) by Arnone *et al.* (2022a), we get

$$\frac{\lambda c_{\mathrm{L}}}{2}\|\mathcal{B}\|^2_{H^2} + c_1\|\mathcal{B}\|^2_{L^2} \leq \frac{\lambda}{c_{\mathrm{L}}}\|g_0\|^2_{L^2} + \frac{\lambda c_{\mathrm{L}}}{4}\|\mathcal{B}\|^2_{H^2},$$

which implies, for n sufficiently large, that $\|\mathcal{B}\|_{L^2} \leq C\lambda^{1/2}$ with C independent on n and λ.

Now, assuming $g_0 = Lf_0 - u \in H^2(\Omega)$, thanks to inequality (14) and (15) by Arnone *et al.* (2022a), we have

$$\frac{\lambda c_{\mathrm{L}}}{2}\|\mathcal{B}\|^2_{H^2} + c_1\|\mathcal{B}\|^2_{L^2} \leq \lambda \int_\Omega (L^* g_0)\mathcal{B} + \lambda \int_{\partial\Omega} g_0 (\mathbf{K}\nabla\mathcal{B})\cdot\boldsymbol{\nu}, \tag{B.7}$$

$$\lambda \int_\Omega (L^* g_0)\mathcal{B} \leq \frac{\lambda^2}{2c_1}\|L^* g_0\|^2_{L^2} + \frac{c_1}{2}\|\mathcal{B}\|^2_{L^2}. \tag{B.8}$$

Since $g_0 \in H^2(\Omega)$, both $\|L^* g_0\|^2_{L^2}$ and $\|g_0\|^2_{H^1}$ are finite. Then, thanks to the Neumann boundary conditions, $\mathbf{K}\nabla\mathcal{B}\cdot\boldsymbol{\nu} = 0$ on $\partial\Omega$, and inequalities (B.7) and (B.8), we obtain

$$\frac{\lambda c_{\mathrm{L}}}{2}\|\mathcal{B}\|^2_{H^2} + \frac{c_1}{2}\|\mathcal{B}\|^2_{L^2} \leq \frac{\lambda^2}{2c_1}\|L^* g_0\|^2_{L^2},$$

which implies $\|\mathcal{B}\|_{L^2} = O(\lambda)$ and $\|\mathcal{B}\|_{H^2} = O(\lambda^{1/2})$. This concludes the proof. □

Asymptotic variance

In the following lemma, we study the convergence rate of the variance of the estimator in Problem (1) as a function of n and λ.

Lemma B.6. *Under Assumptions 2–5, for all* $0 < \epsilon \leq 1/2$ *and* n *sufficiently large,* $\mathrm{Var}_{L^2}(\hat{f}) = O\big(n^{-1}\lambda^{-1/2-\epsilon}\big)$ *with a constant diving to* $+\infty$ *where* $\epsilon \to 0$.

Proof. We recall that minimizing (B.6) is equivalent to finding $\hat{w} \in \mathcal{F}_0$ such that

$$\lambda \int_\Omega (L\psi)(L\hat{w}) + \left[\frac{1}{n}\sum_{i=1}^n \pi_i \psi_i \hat{w}_i\right](1+o(1)) = \frac{1}{n}\sum_{i=1}^n \psi_i x_i, \qquad \forall\, \psi \in \mathcal{F}_0,$$

or equivalently

$$\lambda \int_\Omega (L\psi)(L\hat{w}) + \int_\Omega \pi\psi\hat{w}\,\mathrm{d}\Pi_{\mathbf{p}} = \frac{1}{n}\sum_{i=1}^n \psi_i x_i + \left[\int_\Omega \pi\psi\hat{w}\,\mathrm{d}(\Pi_{\mathbf{p}} - \Pi_{\mathbf{p}}^n)\right](1+o(1)), \quad \text{(B.9)}$$

We define the following inner product on $\mathcal{F}_0$:

$$\langle \phi, \psi\rangle_\lambda = \lambda \int_\Omega (L\phi)(L\psi) + \int_\Omega \pi\phi\psi\,\mathrm{d}\Pi_{\mathbf{p}}, \qquad \forall\, \phi,\psi \in \mathcal{F}_0,$$

which is equivalent to the H^2-inner product. We denote by $\|\cdot\|_\lambda$ the norm induced by $\langle\cdot,\cdot\rangle_\lambda$. Since, thanks to Assumption 5, the norms $\|L\cdot\|_{L^2}$ and $\|\cdot\|_{H^2}$ are equivalent on $\mathcal{F}_0$, there exists a constant c_{L} such that

$$\|\psi\|_{H^2}^2 \le \frac{1}{c_{\mathrm{L}}}\|L\psi\|_{L^2}^2 \le \frac{1}{\lambda c_{\mathrm{L}}}\left(\lambda\|L\psi\|_{L^2}^2 + \int_\Omega \pi\psi^2\,\mathrm{d}\Pi_{\mathbf{p}}\right) = \frac{1}{\lambda c_{\mathrm{L}}}\|\psi\|_\lambda^2.$$

Let us define T_1 and T_2 as follows:

$$T_1(\psi) = \int_\Omega \pi\psi\hat{w}\,\mathrm{d}(\Pi_{\mathbf{p}} - \Pi_{\mathbf{p}}^n), \qquad T_2(\psi) = \frac{1}{n}\sum_{i=1}^n \psi_i x_i.$$

Thanks to the Sobolev embedding theorems, for each $\epsilon > 0$ and $\theta = 1 + 2\epsilon$, we have $T \in H^{\theta,*}(\Omega)$, where $H^{\theta,*}(\Omega)$ denotes the dual space of $H^\theta(\Omega)$. We also denote by $\langle\cdot,\cdot\rangle_{\theta,*}$ and $\|\cdot\|_{\theta,*}$ the natural inner product and norm on $H^{\theta,*}(\Omega)$, respectively. Therefore, we can rephrase equation (B.9) as

$$\langle \psi, \hat{w}\rangle_\lambda = T_1(\psi)(1+o(1)) + T_2(\psi), \qquad \forall \psi \in \mathcal{F}_0.$$

Then, thanks to inequality (19) and (20) by Arnone *et al.* (2022a), we have

$$\begin{aligned}\|\hat{w}\|_\lambda &\le \sup_{\psi\in\mathcal{F}_0}\frac{T_1(\psi)}{\|\psi\|_\lambda}(1+o(1)) + \sup_{\psi\in\mathcal{F}_0}\frac{T_2(\psi)}{\|\psi\|_\lambda}\\ &\le c_1 d_n \lambda^{-1}\|\hat{w}\|_\lambda(1+o(1)) + c_2\lambda^{-\theta/4}\|T_2\|_{\theta,*},\end{aligned}$$

for some positive constants c_1 and c_2 independent on n and λ. Thus, since $d_n\lambda^{-1} \to 0$, for n large enough, we obtain the upper bound

$$\|\hat{w}\|_\lambda^2 \le c\lambda^{-\theta/2}\|T_2\|_{\theta,*}^2 \quad \Rightarrow \quad \mathbb{E}\Big(\|\hat{w}\|_\lambda^2\Big) \le c\lambda^{-\theta/2}\,\mathbb{E}\Big(\|T_2\|_{\theta,*}^2\Big). \qquad \text{(B.10)}$$

Now, from the definition of T_2, we can write

$$T_2 = \frac{1}{n}\sum_{i=1}^{n} x_i \delta_{\mathbf{p}_i},$$

where $\delta_{\mathbf{p}_i}$ is the Dirac delta in $\mathbf{p}_i$, which belongs to $H^{\theta,*}(\Omega)$ because of the Sobolev embedding theorems. Recalling that $x_1, \dots, x_n$ are zero mean, uncorrelated, bounded random variables with variance $\mathbb{E}(x_i^2) < \infty$, we have

$$\begin{aligned}\mathbb{E}\Big(\|T_2\|_{\theta,*}^2\Big) = \mathbb{E}\Big(\langle T_2, T_2\rangle_{\theta,*}\Big) &= \mathbb{E}\bigg(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \langle \delta_{\mathbf{p}_i}\delta_{\mathbf{p}_j}\rangle_{\theta,*}\bigg)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}(x_i x_j)\|\delta_{\mathbf{p}_i}\|_{\theta,*}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}(x_i^2)\|\delta_{\mathbf{p}_i}\|_{\theta,*}^2 \leq \frac{c}{n}s_n^2\end{aligned}$$

where $c = \max_{i=1,\dots,n}\|\delta_{\mathbf{p}_i}\|_{\theta,*}^2 < \infty$ and $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i^2)$. Then, from inequality (B.10) and recalling that $\|\hat{w}\|_{L^2} \leq k_1^{-1}\|\hat{w}\|_\lambda$, we have

$$\mathrm{Var}\Big(\|\hat{w}\|_{L^2}\Big) \leq \mathbb{E}\Big(\|\hat{w}\|_{L^2}^2\Big) \leq \frac{1}{k_1^2}\mathbb{E}\Big(\|\hat{w}\|_\lambda^2\Big) \leq \frac{c\, s_n^2}{k_1^2 n\lambda^{\theta/2}} = O\bigg(\frac{s_n^2}{n\lambda^{\theta/2}}\bigg).$$

This concludes the proof. □

Asymptotic mean squared error

Finally, using Lemmas B.5 and B.6, we can study the behavior of the asymptotic Mean Squared Error (MSE) of the estimator in Problem 1, that is

$$\mathrm{MSE}_{L^2}(\hat{f}) = \|\mathrm{Bias}(\hat{f})\|_{L^2}^2 + \mathrm{Var}_{L^2}(\hat{f}).$$

Thanks to Lemmas B.5 and B.6, if $f_0 \in H^2(\Omega)$, we have

$$\mathrm{MSE}_{L^2}(\hat{f}) = O(\lambda) + O(n^{-1}) + O\big(n^{-1}\lambda^{-1/2-\epsilon}\big),$$

which is minimized when $\lambda = \lambda_n = n^{-2/3}$, leading to $\mathrm{MSE}_{L^2}(\hat{f}) = O\big(n^{-2/3+\epsilon}\big)$. Moreover, if $f_0 \in H^4(\Omega)$, i.e. $g_0 \in H^2(\Omega)$, and $Bf_0 = \gamma$, we have

$$\mathrm{MSE}_{L^2}(\hat{f}) = O(\lambda^2) + O(n^{-1}) + O\big(n^{-1}\lambda^{-1/2-\epsilon}\big),$$

which is minimized for $\lambda = \lambda_n = n^{-2/5}$, leading to $\mathrm{MSE}_{L^2}(\hat{f}) = O\big(n^{-4/5+\epsilon}\big)$. This concludes the proof of Theorem 3.7. □

Proof of Theorem 3.8

Let $J_h(\mathbf{f}) = J(f_h)$ be the finite element discretization of the objective functional (3.6) with $f_h = \mathbf{f}^\top \boldsymbol{\psi} \in \mathcal{F}_{\gamma,h}$, which is given by

$$J_h(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{f}^\top \boldsymbol{\psi}_i) + \frac{\lambda}{2} P_h(\mathbf{f}).$$

The function $P_h(\mathbf{f}) \equiv P(f_h)$ denotes the discretization of the differential penalization $P(f) \equiv \int_\Omega (Lf - u)$ and corresponds to

$$P_h(\mathbf{f}) = (\mathbf{R}_1\mathbf{f} - \mathbf{u} - \boldsymbol{\gamma})^\top \mathbf{R}_0^{-1} (\mathbf{R}_1\mathbf{f} - \mathbf{u} - \boldsymbol{\gamma}),$$

Following the convexity argument proposed by, e.g., Pollard (1991), Knight (1998) and Knight and Fu (2000) in the context of parametric linear quantile regression, we introduce the true quantile residual $\varepsilon_i = y_i - \boldsymbol{\psi}_i^\top \mathbf{f}_0$ and the reparametrization $\delta_h = \boldsymbol{\psi}^\top \boldsymbol{\delta} = \boldsymbol{\psi}^\top (\mathbf{f} - \mathbf{f}_0) \in \mathcal{F}_{0,h}$. Notice that $\boldsymbol{\psi}^\top \mathbf{f}_0$ is the finite element interpolation of f_0 on the mesh knots. Similarly, $\boldsymbol{\psi}^\top \mathbf{g}_0$ is the interpolation of g_0; moreover, thanks to the second identity in (3.20), we have $\mathbf{g}_0 = \mathbf{R}_0^{-1}(\mathbf{R}_1\mathbf{f}_0 - \mathbf{u} - \boldsymbol{\gamma})$. Furthermore, we define the reparametrized objective function

$$J_h^*(\boldsymbol{\delta}) = S_h(\boldsymbol{\delta}) + \frac{\lambda}{2} P_h^*(\boldsymbol{\delta}),$$

where

$$S_h(\boldsymbol{\delta}) = \frac{1}{n}\sum_{i=1}^{n} \Big[\rho_\tau(\varepsilon_i - \boldsymbol{\psi}_i^\top \boldsymbol{\delta}) - \rho_\tau(\varepsilon_i) \Big],$$

and

$$\begin{aligned} P_h^*(\boldsymbol{\delta}) &= \big[\mathbf{R}_1(\boldsymbol{\delta} + \mathbf{f}_0) - \mathbf{u} - \boldsymbol{\gamma}\big]^\top \mathbf{R}_0^{-1} \big[\mathbf{R}_1(\boldsymbol{\delta} + \mathbf{f}_0) - \mathbf{u} - \boldsymbol{\gamma}\big] \\ &= (\mathbf{R}_1\boldsymbol{\delta})^\top \mathbf{R}_0^{-1} (\mathbf{R}_1\boldsymbol{\delta}) + 2(\mathbf{R}_1\boldsymbol{\delta})^\top \mathbf{g}_0 + \mathbf{g}_0^\top \mathbf{R}_0 \mathbf{g}_0 \\ &= \boldsymbol{\delta}^\top \mathbf{P}\, \boldsymbol{\delta} + 2\, \boldsymbol{\delta}^\top \mathbf{R}_1^\top \mathbf{g}_0 + \mathbf{g}_0^\top \mathbf{R}_0 \mathbf{g}_0. \end{aligned}$$

It is easy to verify that minimizing $J_h^*(\cdot)$ with respect to $\boldsymbol{\delta}$ is equivalent to minimizing $J_h(\cdot)$ with respect to $\mathbf{f}$, that is $\hat{\boldsymbol{\delta}} = \hat{\mathbf{f}} - \mathbf{f}_0$. Therefore, we can use the asymptotic properties of $\hat{\boldsymbol{\delta}}$ to infer the limiting behavior of $\hat{\mathbf{f}}$.

Thanks to the fact that $\pi_{Y|\mathbf{p}}(f_0(\mathbf{p}))$ is absolutely continuous and bounded (Assumption 2), the matrices $\mathbf{D}_{0,n}$ and $\mathbf{D}_{1,n}$ are positive definite (Assumptions 6) and their asymptotic limits are finite (Assumptions 7), for n large enough, we can use Theorem 1 by Knight (1998), obtaining

$$S_h(\boldsymbol{\delta}) = \tfrac{1}{2}\, \boldsymbol{\delta}^\top \mathbf{D}_{1,n}\, \boldsymbol{\delta} - \tfrac{1}{\sqrt{n}} \boldsymbol{\delta}^\top \boldsymbol{x}_n + o_{\mathrm{P}}(1),$$

where $\boldsymbol{x}_n \sim \mathrm{N}_{N_h}(0, \tau(1-\tau)\mathbf{D}_{0,n})$. Combining such an asymptotic expansion of $S_h(\boldsymbol{\delta})$ with the penalty term $P_h^*(\boldsymbol{\delta})$, we can write

$$J_h^*(\boldsymbol{\delta}) = \tfrac{1}{2}\,\boldsymbol{\delta}^\top(\mathbf{D}_{1,n} + \lambda\mathbf{P})\,\boldsymbol{\delta} - \boldsymbol{\delta}^\top\Big(\tfrac{1}{\sqrt{n}}\boldsymbol{x}_n - \lambda\mathbf{R}_1^\top\mathbf{g}_0\Big) + \tfrac{\lambda}{2}\,\mathbf{g}_0^\top\mathbf{R}_0\mathbf{g}_0 + o_{\mathrm{p}}(1).$$

Thanks to the non-singularity of $\mathbf{D}_{1,n}$, $J_h^*(\boldsymbol{\delta})$ is strictly convex and its asymptotic minimizer is unique. Therefore, using Corollary 1 by Knight (1998), $\sqrt{n}\,\hat{\boldsymbol{\delta}} = \sqrt{n}(\hat{\mathbf{f}} - \mathbf{f}_0)$ converges in distribution to the solution of the asymptotic first order equation

$$\sqrt{n}(\mathbf{D}_{1,n} + \lambda\mathbf{P})\boldsymbol{\delta} = (\boldsymbol{x}_n - \sqrt{n}\,\lambda\mathbf{R}_1^\top\mathbf{g}_0).$$

Then, for n sufficiently large, we have

$$\sqrt{n}(\hat{\mathbf{f}} - \mathbf{f}_0) = (\mathbf{D}_{1,n} + \lambda\mathbf{P})^{-1}(\boldsymbol{x}_n - \sqrt{n}\,\lambda\mathbf{R}_1^\top\mathbf{g}_0) + o_{\mathrm{p}}(1), \qquad \text{(B.11)}$$

In order to obtain a non-exploding bias for (B.11), we require that $\sqrt{n}\,\lambda = \sqrt{n}\,\lambda_n \to \lambda_0$ for some finite value $\lambda_0 > 0$. Under this condition, the large-sample bias and variance of the estimator $\hat{\mathbf{f}}$ are

$$\begin{aligned}\mathrm{Bias}_n(\hat{\mathbf{f}}) &= -\lambda(\mathbf{D}_{1,n} + \lambda\mathbf{P})^{-1}\mathbf{R}_1^\top\mathbf{g}_0 + o(n^{-1/2}),\\ \mathrm{Var}_n(\hat{\mathbf{f}}) &= \tfrac{1}{n}\tau(1-\tau)(\mathbf{D}_{1,n} + \lambda\mathbf{P})^{-1}\mathbf{D}_{0,n}(\mathbf{D}_{1,n} + \lambda\mathbf{P})^{-1} + o(n^{-1}).\end{aligned}$$

Following Ferraccioli *et al.* (2022), we expand the bias term with a second order Taylor approximation for λ around 0, which leads to

$$\mathrm{Bias}_n(\hat{\mathbf{f}}) = -\lambda\big(\mathbf{D}_{1,n}^{-1} - \lambda\mathbf{D}_{1,n}^{-1}\mathbf{P}\mathbf{D}_{1,n}^{-1} + O(\lambda^2)\big)\mathbf{R}_1^\top\mathbf{g}_0 + o(n^{-1/2}),$$

that is $\mathrm{Bias}_n(\hat{\mathbf{f}}) = O(\lambda) + o(n^{-1/2})$. In the same way, the variance can be expanded as

$$\mathrm{Var}_n(\hat{\mathbf{f}}) = \tfrac{1}{n}\tau(1-\tau)\big(\mathbf{D}_{1,n}^{-1}\mathbf{D}_{0,n}\mathbf{D}_{1,n}^{-1} - 2\lambda\mathbf{D}_{1,n}^{-1}\mathbf{P}\,\mathbf{D}_{1,n}^{-1}\mathbf{D}_{0,n}\mathbf{D}_{1,n}^{-1} + O(\lambda^2)\big) + o(n^{-1}),$$

which leads to $\mathrm{Var}_n(\hat{\mathbf{f}}) = O(n^{-1}) + O(\lambda n^{-1})$. Considering the Taylor expansions of $\mathrm{Bias}_n(\hat{\mathbf{f}})$ and $\mathrm{Var}_n(\hat{\mathbf{f}})$, using the linearity of estimator (B.11), and setting $\sqrt{n}\,\lambda_n \to \lambda_0$, the asymptotic distribution of $\hat{\mathbf{f}}$ becomes

$$\sqrt{n}(\hat{\mathbf{f}} - \mathbf{f}_0) \xrightarrow{\mathrm{d}} \mathrm{N}\big(-\lambda_0\,\mathbf{D}_1^{-1}\mathbf{R}_1^\top\mathbf{g}_0,\ \tau(1-\tau)\mathbf{D}_1^{-1}\mathbf{D}_0\,\mathbf{D}_1^{-1}\big).$$

If we further assume that $\lambda = \lambda_n = o(n^{-1/2})$, that is $\lambda_0 = 0$, the estimator $\sqrt{n}(\hat{\mathbf{f}} - \mathbf{f}_0)$ is asymptotically unbiased.

Finally, we prove the consistency of the estimator $\hat{\mathbf{f}}$ by studying the limiting behavior of the mean squared error (MSE) and leveraging the results obtained for $\mathrm{Bias}_n(\hat{\mathbf{f}})$ and $\mathrm{Var}_n(\hat{\mathbf{f}})$:

$$\mathrm{MSE}_n(\hat{\mathbf{f}}) = \mathrm{Var}_n(\hat{\mathbf{f}}) + \mathrm{Bias}_n(\hat{\mathbf{f}})\,\mathrm{Bias}_n(\hat{\mathbf{f}})^\top = O(\lambda^2) + O(n^{-1}) + O(\lambda n^{-1}).$$

Then, the proof is concluded by noting that, for any $\lambda = \lambda_n = O(n^{-1/2})$, we obtain a convergent MSE with limiting rate $O(n^{-1})$. $\square$

Bibliography

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**(422), 669–679.

Alquier, P. and Ridgway, J. (2020) Concentration of tempered posteriors and of their variational approximations. *Ann. Statist.* **48**(3), 1475–1497.

Alquier, P., Ridgway, J. and Chopin, N. (2016) On the properties of variational approximations of Gibbs posteriors. *J. Mach. Learn. Res.* **17**(1), 8374–8414.

Armagan, A., Dunson, D. B. and Lee, J. (2013) Generalized double Pareto shrinkage. *Statist. Sinica* **23**(1), 119–143.

Arnone, E., Azzimonti, L., Nobile, F. and Sangalli, L. M. (2019) Modeling spatially dependent functional data via regression with differential regularization. *J. Multivariate Anal.* **170**, 275–295.

Arnone, E., Kneip, A., Nobile, F. and Sangalli, L. M. (2022a) Some first results on the consistency of spatial regression with partial differential equation regularization. *Statist. Sinica* **32**(1), 209–238.

Arnone, E., Sangalli, L. M., Lila, E., Ramsay, J., Formaggia, L., Ardenghi, G., Clemente, A., Colli, A., Colombo, A., Colombo, L., de Falco, C., Dall'Acqua, E., Ferla, G., Ghilotti, L., Kim, J., Massardi, M., Meretti, G., Perin, G., Pigolotti, C., Poiatti, A., Rinaldi, G. M., Spaziani, S. and Vicini, A. (2022b) `fdaPDE`*: Functional Data Analysis and Partial Differential Equations (PDE); Statistical Analysis of Functional and Spatial Data, Based on Regression with PDE Regularization.* R package version 1.1.8.

Azzimonti, L., Nobile, F., Sangalli, L. M. and Secchi, P. (2014) Mixed finite elements for spatial regression with PDE penalization. *SIAM/ASA J. Uncertain. Quantif.* **2**(1), 305–335.

Azzimonti, L., Sangalli, L. M., Secchi, P., Domanin, M. and Nobile, F. (2015) Blood flow velocity field estimation via spatial regression with PDE penalization. *J. Amer. Statist. Assoc.* **110**(511), 1057–1071.

Bellini, F. and Bignozzi, V. (2015) On elicitable risk measures. *Quant. Finance* **15**(5), 725–733.

Bernardi, M. S., Carey, M., Ramsay, J. O. and Sangalli, L. M. (2018) Modeling spatial anisotropy via regression with partial differential regularization. *J. Multivariate Anal.* **167**, 15–30.

Bernardi, M. S., Sangalli, L. M., Mazza, G. and Ramsay, J. O. (2017) A penalized regression model for spatial functional data with application to the analysis of the production of waste in venice province. *Stoch. Environ. Res. Risk Assess.* **31**(1), 23–38.

Bhattacharya, A., Pati, D., Pillai, N. S. and Dunson, D. B. (2015) Dirichlet-Laplace priors for optimal shrinkage. *J. Amer. Statist. Assoc.* **110**(512), 1479–1490.

Bishop, C. M. (2006) *Pattern recognition and machine learning.* Information Science and Statistics. Springer, New York.

Bissiri, P. G., Holmes, C. C. and Walker, S. G. (2016) A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78**(5), 1103–1130.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017) Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112**(518), 859–877.

Bondell, H. D., Reich, B. J. and Wang, H. (2010) Noncrossing quantile regression curve estimation. *Biometrika* **97**(4), 825–838.

Boos, D. D. and Stefanski, L. (2013) M-estimation (estimating equations). In *Essential Statistical Inference*, pp. 297–337. Springer.

Bosch, R. J., Ye, Y. and Woodworth, G. G. (1995) A convergent algorithm for quantile regression with smoothing splines. *Comput. Statist. Data Anal.* **19**(6), 613–630.

Breckling, J. and Chambers, R. (1988) M-quantiles. *Biometrika* **75**(4), 761–771.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**(421), 9–25.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010) The horseshoe estimator for sparse signals. *Biometrika* **97**(2), 465–480.

Casella, G. and George, E. I. (1992) Explaining the Gibbs sampler. *Amer. Statist.* **46**(3), 167–174.

Chan, J. C. C. and Jeliazkov, I. (2009) Efficient simulation and integrated likelihood estimation in state space models. *Int. J. Math. Model. Numer. Optim.* **1**(1-2), 101–120.

Cox, D. D. (1984) Multivariate smoothing spline functions. *SIAM J. Numer. Anal.* **21**(4), 789–813.

Cressie, N. A. C. (2015) *Statistics for spatial data.* Revised edition. Wiley Classics Library. John Wiley & Sons, Inc., New York. Paperback edition of the 1993 edition.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**(1), 1–38.

Duan, L. L., Johndrow, J. E. and Dunson, D. B. (2018) Scaling up data augmentation MCMC via calibration. *J. Mach. Learn. Res.* **19**, Paper No. 64, 34.

Dubois, G., Malczewski, J. and De Cort, M. (2003) *Mapping radioactivity in the environment: Spatial interpolation comparison 97.* Office for Official Publications of the European Communities, Luxembourg.

Durante, D. (2019) Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika* **106**(4), 765–779.

Durante, D. and Rigon, T. (2019) Conditionally conjugate mean-field variational Bayes for logistic models. *Statist. Sci.* **34**(3), 472–485.

Durbin, J. and Koopman, S. J. (2012) *Time series analysis by state space methods.* Second edition, volume 38 of *Oxford Statistical Science Series.* Oxford University Press, Oxford.

Efron, B. (1991) Regression percentiles using asymmetric squared error loss. *Statist. Sinica* **1**(1), 93–125.

Ettinger, B., Perotto, S. and Sangalli, L. M. (2016) Spatial regression models over two-dimensional manifolds. *Biometrika* **103**(1), 71–88.

Evans, L. C. (2010) *Partial differential equations.* Second edition, volume 19 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI.

Fasano, A. and Durante, D. (2022) A class of conjugate priors for multinomial probit models which includes the multivariate normal one. *J. Mach. Learn. Res.* **23**, Paper No. [30], 26.

Fasano, A., Durante, D. and Zanella, G. (2022) Scalable and accurate variational Bayes for high-dimensional binary regression models. *Biometrika* **109**(4), 901–919.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R. and Goude, Y. (2021a) Fast calibrated additive quantile regression. *J. Amer. Statist. Assoc.* **116**(535), 1402–1412.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R. and Goude, Y. (2021b) qgam: Bayesian nonparametric quantile regression modeling in R. *J. Stat. Softw.* **100**(9), 1–31.

Ferraccioli, F., Sangalli, L. M. and Finos, L. (2022) Some first inferential tools for spatial regression with differential regularization. *J. Multivariate Anal.* **189**, Paper No. 104866, 12.

Frühwirth-Schnatter, S. (2006) *Finite mixture and Markov switching models.* Springer Series in Statistics. Springer, New York.

Frumento, P. and Bottai, M. (2016) Parametric modeling of quantile regression coefficient functions. *Biometrics* **72**(1), 74–84.

Frumento, P., Bottai, M. and Fernández-Val, I. (2021) Parametric modeling of quantile regression coefficient functions with longitudinal data. *J. Amer. Statist. Assoc.* **116**(534), 783–797.

Gaillard, P., Goude, Y. and Nedellec, R. (2016) Additive models and robust aggregation for gefcom2014 probabilistic electric load and electricity price forecasting. *Int. J. Forecast.* **32**(3), 1038–1050.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1**(3), 515–533.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian data analysis, third edition.* CRC press.

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

Geraci, M. (2014) Linear quantile mixed models: the lqmm package for Laplace quantile regression. *J. Stat. Softw.* **57**, 1–29.

Geraci, M. and Bottai, M. (2007) Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8**(1), 140–154.

Geraci, M. and Bottai, M. (2014) Linear quantile mixed models. *Stat. Comput.* **24**(3), 461–479.

Germain, P., Bach, F., Lacoste, A. and Lacoste-Julien, S. (2016) Pac-bayesian theory meets Bayesian inference. *Advances in Neural Information Processing Systems* **29**.

Greene, W. H. and Seaks, T. G. (1991) The restricted least squares estimator: a pedagogical note. *Rev. Econ. Stat.* pp. 563–567.

Griffin, J. E. and Brown, P. J. (2011) Bayesian hyper-lassos with non-convex penalization. *Aust. N. Z. J. Stat.* **53**(4), 423–442.

Hall, A. R. (2005) *Generalized method of moments.* Advanced Texts in Econometrics. Oxford University Press, Oxford.

Hall, P., Johnstone, I. M., Ormerod, J. T., Wand, M. P. and Yu, J. C. F. (2020) Fast and accurate binary response mixed model analysis via expectation propagation. *J. Amer. Statist. Assoc.* **115**(532), 1902–1916.

Hall, P., Ormerod, J. T. and Wand, M. P. (2011a) Theory of Gaussian variational approximation for a Poisson mixed model. *Statist. Sinica* **21**(1), 369–389.

Hall, P., Pham, T., Wand, M. P. and Wang, S. S. J. (2011b) Asymptotic normality and valid inference for Gaussian variational approximation. *Ann. Statist.* **39**(5), 2502–2532.

Hallin, M., Lu, Z. and Yu, K. (2009) Local linear spatial quantile regression. *Bernoulli* **15**(3), 659–686.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The elements of statistical learning. Data mining, inference, and prediction.* Second edition. Springer Series in Statistics. Springer, New York.

He, X., Ng, P. and Portnoy, S. (1998) Bivariate quantile smoothing splines. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60**(3), 537–550.

Hodges, J. S. (2014) *Richly parameterized linear models: Additive, time series, and spatial models using random effects.* Chapman & Hall/CRC Texts in Statistical Science Series. CRC Press, Boca Raton, FL.

Hoffman, M. D., Blei, D. M., Wang, C. and Paisley, J. (2013) Stochastic variational inference. *J. Mach. Learn. Res.* **14**, 1303–1347.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A. and Hyndman, R. J. (2016) Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *Int. J. Forecast.* **32**(3), 896–913.

Honkela, A., Raiko, T., Kuusela, M., Tornio, M. and Karhunen, J. (2010) Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *J. Mach. Learn. Res.* **11**, 3235–3268.

Hughes, D. M., García-Fiñana, M. and Wand, M. P. (2023) Fast approximate inference for multivariate longitudinal data. *Biostatistics* **24**(1), 177–192.

Hunter, D. R. and Lange, K. (2000) Quantile regression via an MM algorithm. *J. Comput. Graph. Statist.* **9**(1), 60–77.

Jaakkola, T. S. and Jordan, M. I. (2000) Bayesian parameter estimation via variational methods. *Stat. Comput.* **10**(1), 25–37.

Johndrow, J. E., Smith, A., Pillai, N. and Dunson, D. B. (2019) Mcmc for imbalanced categorical data. *J. Amer. Statist. Assoc.* **114**(527), 1394–1403.

Jørgensen, B. (1982) *Statistical properties of the generalized inverse Gaussian distribution.* Volume 9 of *Lecture Notes in Statistics.* Springer-Verlag, New York-Berlin.

Kato, K. (2009) Asymptotics for argmin processes: convexity arguments. *J. Multivariate Anal.* **100**(8), 1816–1829.

Khan, M. and Lin, W. (2017) Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. In *Artificial Intelligence and Statistics*, pp. 878–887.

Khan, M. E. and Rue, H. (2021) The bayesian learning rule. *arXiv preprint arXiv:2107.04562* .

Knight, K. (1998) Limiting distributions for L_1 regression estimators under general conditions. *Ann. Statist.* **26**(2), 755–770.

Knight, K. and Fu, W. (2000) Asymptotics for lasso-type estimators. *Ann. Statist.* **28**(5), 1356–1378.

Knowles, D. and Minka, T. (2011) Non-conjugate variational message passing for multinomial and binary regression. *Advances in Neural Information Processing Systems* **24**, 1701–1709.

Koenker, R. (2005) *Quantile regression.* Volume 38 of *Econometric Society Monographs.* Cambridge University Press, Cambridge.

Koenker, R. (2021) *quantreg: Quantile Regression.* R package version 5.86.

Koenker, R. and Bassett, Jr., G. (1978) Regression quantiles. *Econometrica* **46**(1), 33–50.

Koenker, R., Chernozhukov, V., He, X. and Peng, L. (eds) (2018) *Handbook of quantile regression.* Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL.

Koenker, R. and Mizera, I. (2004) Penalized triograms: total variation regularization for bivariate smoothing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66**(1), 145–163.

Koenker, R. and Ng, P. (2005) A Frisch-Newton algorithm for sparse quantile regression. *Acta Math. Appl. Sin. Engl. Ser.* **21**(2), 225–236.

Koenker, R., Ng, P. and Portnoy, S. (1994) Quantile smoothing splines. *Biometrika* **81**(4), 673–680.

Kotz, S., Kozubowski, T. J. and Podgórski, K. (2001) *The Laplace distribution and generalizations. A revisit with applications to communications, economics, engineering, and finance.* Birkhäuser Boston, Inc., Boston, MA.

Kozumi, H. and Kobayashi, G. (2011) Gibbs sampling methods for Bayesian quantile regression. *J. Stat. Comput. Simul.* **81**(11), 1565–1578.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. and Blei, D. M. (2017) Automatic differentiation variational inference. *J. Mach. Learn. Res.* **18**, Paper No. 14, 45.

Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *Ann. Math. Statistics* **22**, 79–86.

Lai, M.-J. and Wang, L. (2013) Bivariate penalized splines for regression. *Statist. Sinica* **23**(3), 1399–1417.

Lange, K. (2010) *Numerical analysis for statisticians.* Second edition. Statistics and Computing. Springer, New York.

Lange, K. (2013) *Optimization.* Second edition, volume 95 of *Springer Texts in Statistics.* Springer, New York.

Leng, C., Tran, M.-N. and Nott, D. (2014) Bayesian adaptive Lasso. *Ann. Inst. Statist. Math.* **66**(2), 221–244.

Lewandowski, A., Liu, C. and Vander Wiel, S. (2010) Parameter expansion and efficient inference. *Statist. Sci.* **25**(4), 533–544.

Li, Y., Liu, Y. and Zhu, J. (2007) Quantile regression in reproducing kernel Hilbert spaces. *J. Amer. Statist. Assoc.* **102**(477), 255–268.

Lila, E., Aston, J. A. D. and Sangalli, L. M. (2016) Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *Ann. Appl. Stat.* **10**(4), 1854–1879.

Lindgren, F., Bolin, D. and Rue, H. v. (2022) The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running. *Spat. Stat.* **50**, 100599.

Lindgren, F., Rue, H. v. and Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**(4), 423–498.

Liu, Q. and Pierce, D. A. (1994) A note on Gauss-Hermite quadrature. *Biometrika* **81**(3), 624–629.

Loaiza-Maya, R., Smith, M. S., Nott, D. J. and Danaher, P. J. (2022) Fast and accurate variational inference for models with many latent variables. *J. Econometrics* **230**(2), 339–362.

Luts, J. and Ormerod, J. T. (2014) Mean field variational Bayesian inference for support vector machine classification. *Comput. Statist. Data Anal.* **73**, 163–176.

Luts, J. and Wand, M. P. (2015) Variational inference for count response semiparametric regression. *Bayesian Anal.* **10**(4), 991–1023.

McCullagh, P. and Nelder, J. A. (1989) *Generalized linear models. Second edition.* Chapman & Hall, London.

McLachlan, G. J. and Krishnan, T. (2008) *The EM algorithm and extensions*. Second edition. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

McLean, M. W. and Wand, M. P. (2019) Variational message passing for elaborate response regression models. *Bayesian Anal.* **14**(2), 371–398.

Menictas, M. and Wand, M. P. (2015) Variational inference for heteroscedastic semiparametric regression. *Aust. N. Z. J. Stat.* **57**(1), 119–138.

Minka, T. (2005) Divergence measures and message passing. Technical report, Citeseer.

Minka, T. P. (2013) Expectation propagation for approximate bayesian inference. *arXiv preprint arXiv:1301.2294* .

Neville, S. E., Ormerod, J. T. and Wand, M. P. (2014) Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies. *Electron. J. Stat.* **8**(1), 1113–1151.

Newey, W. K. and Powell, J. L. (1987) Asymmetric least squares estimation and testing. *Econometrica* **55**(4), 819–847.

Ng, P. T. (1996) An algorithm for quantile smoothing splines. *Comput. Statist. Data Anal.* **22**(2), 99–118.

Nocedal, J. and Wright, S. J. (2006) *Numerical optimization*. Second edition. Springer Series in Operations Research and Financial Engineering. Springer, New York.

Nolan, T. H., Menictas, M. and Wand, M. P. (2020) Streamlined computing for variational inference with higher level random effects. *J. Mach. Learn. Res.* **21**, Paper No. 157, 62.

Nolan, T. H. and Wand, M. P. (2020) Streamlined solutions to multilevel sparse matrix problems. *ANZIAM J.* **62**(1), 18–41.

Norris, J. R. (1998) *Markov chains*. Volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. Reprint of 1997 original.

Nychka, D., Gray, G., Haaland, P., Martin, D. and O'connell, M. (1995) A nonparametric regression approach to syringe grading for quality improvement. *J. Amer. Statist. Assoc.* **90**(432), 1171–1178.

Oh, H.-S., Lee, T. C. M. and Nychka, D. W. (2011) Fast nonparametric quantile regression with arbitrary smoothing methods. *J. Comput. Graph. Statist.* **20**(2), 510–526.

Ong, V. M.-H., Nott, D. J. and Smith, M. S. (2018) Gaussian variational approximation with a factor covariance structure. *J. Comput. Graph. Statist.* **27**(3), 465–478.

Opper, M. and Archambeau, C. (2009) The variational Gaussian approximation revisited. *Neural Comput.* **21**(3), 786–792.

Ormerod, J. T. (2011) Skew-normal variational approximations for Bayesian inference. *Unpublished article* .

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

Ormerod, J. T. and Wand, M. P. (2012) Gaussian variational approximate inference for generalized linear mixed models. *J. Comput. Graph. Statist.* **21**(1), 2–17.

Park, T. and Casella, G. (2008) The Bayesian lasso. *J. Amer. Statist. Assoc.* **103**(482), 681–686.

Patterson, H. D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.

Pollard, D. (1991) Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7**(2), 186–199.

Polson, N. G., Scott, J. G. and Windle, J. (2013) Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Amer. Statist. Assoc.* **108**(504), 1339–1349.

Polson, N. G. and Scott, S. L. (2011) Data augmentation for support vector machines. *Bayesian Anal.* **6**(1), 1–23.

Portnoy, S. and Koenker, R. (1997) The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statist. Sci.* **12**(4), 279–300. With comments by Ronald A. Thisted and M. R. Osborne and a rejoinder by the authors.

Quarteroni, A. (2017) *Numerical models for differential problems.* Third edition, volume 16 of *MS&A. Modeling, Simulation and Applications.* Springer, Cham.

Ramsay, T. (2002) Spline smoothing over difficult regions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**(2), 307–319.

Ranganath, R., Gerrish, S. and Blei, D. (2014) Black box variational inference. In *Artificial intelligence and statistics*, pp. 814–822.

Rockafellar, R. T. (1997) *Convex analysis.* Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ. Reprint of the 1970 original, Princeton Paperbacks.

Rohde, D. and Wand, M. P. (2016) Semiparametric mean field variational Bayes: general principles and numerical issues. *J. Mach. Learn. Res.* **17**, Paper No. 172, 47.

Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71**(2), 319–392.

Rue, H. v. and Held, L. (2005) *Gaussian Markov random fields. Theory and applications.* Volume 104. Chapman & Hall/CRC, Boca Raton, FL.

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

Sangalli, L. M. (2021) Spatial regression with partial differential equation regularisation. *Int. Stat. Rev.* **89**(3), 505–531.

Sangalli, L. M., Ramsay, J. O. and Ramsay, T. O. (2013) Spatial spline regression models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75**(4), 681–703.

Schlather, M., Malinowski, A., Oesting, M., Boecker, D., Strokorb, K., Engelke, S., Martini, J., Ballani, F., Moreva, O., Auel, J., Menck, P. J., Gross, S., Ober, U., Berreth, C., Burmeister, K., Manitz, J., Ribeiro, P., Singleton, R., Ben, P. and R Core Team (2017) *RandomFields: Simulation and Analysis of Random Fields.* R package version 3.1.50.

Schnabel, S. K. and Eilers, P. H. C. (2013) Simultaneous estimation of quantile curves using quantile sheets. *AStA Adv. Stat. Anal.* **97**(1), 77–87.

Severini, T. A. (2000) *Likelihood methods in statistics.* Volume 22 of *Oxford Statistical Science Series.* Oxford University Press, Oxford.

Silverman, B. W. (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J. Roy. Statist. Soc. Ser. B* **47**(1), 1–52. With discussion.

Sobotka, F. and Kneib, T. (2012) Geoadditive expectile regression. *Comput. Statist. Data Anal.* **56**(4), 755–767.

Sriram, K., Ramamoorthi, R. V. and Ghosh, P. (2013) Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Bayesian Anal.* **8**(2), 479–504.

Stefanski, L. A. and Boos, D. D. (2002) The calculus of M-estimation. *Amer. Statist.* **56**(1), 29–38.

Tan, L. S. L. and Nott, D. J. (2013) Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statist. Sci.* **28**(2), 168–188.

Tian, Y., Tian, M. and Zhu, Q. (2014) Linear quantile regression based on EM algorithm. *Comm. Statist. Theory Methods* **43**(16), 3464–3484.

Triantafyllopoulos, K. (2021) *Bayesian inference of state space models: Kalman filtering and beyond.* Springer Series in Statistics. Springer, Cham.

Vapnik, V. N. (1998) *Statistical learning theory.* John Wiley & Sons, Inc., New York.

Wahba, G. (1990) *Spline models for observational data.* Volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Waldmann, E., Sobotka, F. and Kneib, T. (2017) Bayesian regularisation in geoadditive expectile regression. *Stat. Comput.* **27**(6), 1539–1553.

Wand, M. P. (2014) Fully simplified multivariate normal updates in non-conjugate variational message passing. *J. Mach. Learn. Res.* **15**, 1351–1369.

Wand, M. P. and Ormerod, J. T. (2008) On semiparametric regression with O'Sullivan penalized splines. *Aust. N. Z. J. Stat.* **50**(2), 179–198.

Wand, M. P., Ormerod, J. T., Padoan, S. A. and Frührwirth, R. (2011) Mean field variational Bayes for elaborate distributions. *Bayesian Anal.* **6**(4), 847–900.

Wang, L., Wang, G., Lai, M.-J. and Gao, L. (2020) Efficient estimation of partially linear models for data on complicated domains by bivariate penalized splines over triangulations. *Statist. Sinica* **30**(1), 347–369.

Wang, L., Wu, Y. and Li, R. (2012) Quantile regression for analyzing heterogeneity in ultra-high dimension. *J. Amer. Statist. Assoc.* **107**(497), 214–222.

Wang, Y. and Blei, D. (2019a) Variational bayes under model misspecification. *Advances in Neural Information Processing Systems* **32**.

Wang, Y. and Blei, D. (2019b) Variational bayes under model misspecification. *Advances in Neural Information Processing Systems* **32**.

Wang, Y. and Blei, D. M. (2019c) Frequentist consistency of variational Bayes. *J. Amer. Statist. Assoc.* **114**(527), 1147–1161.

Wilhelm, M. and Sangalli, L. M. (2016) Generalized spatial regression with differential regularization. *J. Stat. Comput. Simul.* **86**(13), 2497–2518.

Wolfinger, R. (1994) Laplace's approximation for nonlinear mixed models. *Biometrika* **80**(4), 791–795.

Wood, S. N. (2003) Thin plate regression splines. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65**(1), 95–114.

Wood, S. N. (2017) *Generalized additive models. An introduction with* R*, Second edition.* CRC Press, Boca Raton, FL.

Wood, S. N., Bravington, M. V. and Hedley, S. L. (2008) Soap film smoothing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70**(5), 931–955.

Xing, J.-J. and Qian, X.-Y. (2017) Bayesian expectile regression with asymmetric normal distribution. *Comm. Statist. Theory Methods* **46**(9), 4545–4555.

Yu, K. and Jones, M. C. (1998) Local linear quantile regression. *J. Amer. Statist. Assoc.* **93**(441), 228–237.

Yu, K. and Moyeed, R. A. (2001) Bayesian quantile regression. *Statist. Probab. Lett.* **54**(4), 437–447.

Yuan, M. (2006) GACV for quantile smoothing splines. *Comput. Statist. Data Anal.* **50**(3), 813–829.

Yue, Y. R. and Rue, H. (2011) Bayesian inference for additive mixed quantile regression models. *Comput. Statist. Data Anal.* **55**(1), 84–96.

Zhu, J., Ahmed, A. and Xing, E. P. (2012) MedLDA: maximum margin supervised topic models. *J. Mach. Learn. Res.* **13**, 2237–2278.

Zhu, J., Chen, N., Perkins, H. and Zhang, B. (2014) Gibbs max-margin topic models with data augmentation. *J. Mach. Learn. Res.* **15**, 1073–1110.

Ziegel, J. F. (2016) Coherence and elicitability. *Math. Finance* **26**(4), 901–918.

Cristian Castiglione

CURRICULUM VITAE

Contact Information

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

Tel. +39 049 827 4141
e-mail: `cristian.castiglione@phd.unipd.it` (official)
e-mail: `cristian_castiglione@libero.it` (personal)

Current Position

Since February 2023
Postdoctoral Research fellow
University of Padova, Department of Statistical Sciences
Project: *Statistical methods and models for the integration of multiomic data*
Supervisor: Prof. Davide Risso

Since October 2019 (expected completion: January 2023, expected final discussion: May 2023)
PhD Student in Statistical Sciences
University of Padova, Department of Statistical Sciences
Thesis title: *Approximate inference for misspecified additive and mixed regression models*
Supervisor: Prof. Mauro Bernardi
Co-supervisor: Prof. Laura M. Sangalli
Co-supervisor: Prof. Alessio Farcomeni

Research interests

- Bayesian statistics
- Computational statistics
- Mixed and additive models
- Spatial Statistics
- Variational approximations

Education

October 2016 – November 2018
Master degree (*laurea specialistica/magistrale*) in Statistical Sciences
University of Padova, Department of Statistical Sciences
Title of dissertation: *Dynamic quantile models for spatio-temporal data*
Supervisor: Prof. Mauro Bernardi
Final mark: 110/110 cum Laude

September 2013 – July 2016
Bachelor degree (*laurea triennale*) in Statistics, Economics and Finance.
University of Padova, Department of Statistical Sciences

Title of dissertation: *Multistate models for competing risks*
Supervisor: Prof. Giuliana Cortese
Final mark: 110/110

Work experience

January 2019 – September 2019
Junior consultant in business intelligence and analytics
Blue BI S.R.L.

Awards and Scholarship

June 2017
Best Report Prize at *Stats Under the Stars 3 (SuS3).*

Computer skills

- Advanced knowledge of `R`, `Python` and `Julia`
- Good knowledge of `Stan` and `Keras`
- Basic knowledge of `Matlab` and `C++`
- Good knowledge of LaTeX
- Basic knowledge of `MySQL`

Language skills

- Italian: native
- English: good (written/spoken)

Publications

Conference proceedings

Castiglione, C., Bernardi, M. (2022). Probabilistic load forecasting via dynamic quantile regression. In *Book of Short Papers IWSM 2022, Proceedings of the 36th International Workshop on Statistical Modeling* (Torelli, N., Bellio, R., Muggeo, V.), pp. 400–405.

Castiglione, C., Bernardi, M. (2022). Sparse signal extraction via variational SVM. In *Book of Short Papers SIS 2022, Proceedings of the 51th Scientific Meeting of the Italian Statistical Society* (Balzanella, A., Bini, M., Cavicchia, C., Verde, R.), pp. 864–870.

Castiglione, C., Bernardi, M. (2021). Semiparametric variational inference for Bayesian quantile regression. In *Book of Short Papers SIS 2021, Proceedings of the 50th Scientific Meeting of the Italian Statistical Society* (Perna, C., Salvati, N. and Schirripa Spagnolo, F.), pp. 683–688.

Working papers

Castiglione, C., Bernardi, M. (2022+). Bayesian non-conjugate regression via variational belief updating. Unpublished manuscript publicly available at `https://arxiv.org/abs/2206.09444`.

Castiglione, C., Arnone, E., Bernardi, M., Farcomeni, A., Sangalli, L.M. (2022+). Spatial quantile regression with PDE regularization. Unpublished manuscript.

Conference presentations

Castiglione, C. (2022). Approximate belief updating via semiparametric variational Bayes. (poster presentation) *Statistical Methods and Models for Complex Data 2022*, Padova, Italy, 21 – 21 September.

Castiglione, C., Bernardi, M. (2022). Approximate general Bayesian inference via semiparametric variational Bayes. (invited presentation) *24th Conference on Computational Statistics (COMPSTAT 2022)*, Bologna, Italy, 23 – 26 August.

Castiglione, C., Bernardi, M. (2022). Probabilistic load forecasting via dynamic quantile regression. (poster presentation) *36th International Workshop on Statistical Modelling (IWSM 2022)*, Trieste, Italy, 18 – 22 July.

Castiglione, C., Bernardi, M. (2022). Approximate general Bayesian inference via semiparametric variational Bayes. (oral presentation) *2022 World Meeting of the International Society for Bayesian Analysis (ISBA 2022)*, Montreal, Canada, 26 June – 1 July.

Castiglione, C., Bernardi, M. (2022). Sparse signal extraction via Variational SVM. (oral presentation) *51th Scientific Meeting of the Italian Statisrtical Society (SIS 2022)*, Caserta, Italy, 22 – 24 June.

Castiglione, C. (2021). Approximate variational inference based on data augmentation methods. (oral presentation) *14th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2021)*, London, UK, 18 – 20 December.

Castiglione, C., Bernardi, M. (2021). Variational inference for non-crossing quantile regression. (poster presentation) *2021 World Meeting of the International Society for Bayesian Analysis (ISBA 2021)*, Online, 28 June – 02 July.

Castiglione, C., Bernardi, M. (2022). Semiparametric variational inference for Bayesian quantile regression. (oral presentation) *50th Scientific Meeting of the Italian Statisrtical Society (SIS 2021)*, Cagliari, Italy, 22 – 24 June.

Teaching experience

September 2017 – September 2018
Calculus 1 and Advanced Statistics
Academic tutor, 50 hours
Department of Statistical Sciences, University of Padova
Instructor: Prof. Annalisa Cesaroni; Prof. Alessandra R. Brazzale