Haziq Jamil
Department of Statistics
London School of Economics and Political Science
PhD thesis: 'Regression modelling using priors depending on Fisher information covariance kernels'
29 October 2018 (`v1.2@dae7107`)

Chapter 1

Introduction

Regression analysis is undoubtedly one of the most important tools available at a practitioner's disposal to understand the relationship between one or more explanatory variables x, and the independent variable of interest, y. This relationship is usually expressed as $y \approx f(x|\theta)$, where f is called the *regression function*, and this is dependent on one or more parameters denoted by θ. Regression analysis concerns the estimation of said regression function, and once a suitable estimate $\hat{f}$ has been found, post-estimation procedures such as prediction and inference surrounding f or θ, may be performed.

Estimation of the regression function may be done in many ways. This thesis concerns the use of *I-priors* (Bergsma, 2018), in a semi-Bayesian manner, for regression modelling. The I-prior is an objective, data-dependent prior for the regression function which makes use of its Fisher information and is based on the principle of maximum entropy (Jaynes, 1957a, 1957b, 2003). Entropy-maximising priors are "uninformative" in the sense that it minimises the amount of prior information encoded into prior distributions, and thus should be advocated in the absence of any prior knowledge.

The essence of regression modelling using I-priors is introduced briefly in this chapter, but as the development of I-priors is fairly recent, we dedicate two full chapters (Chapters 2 and 3) to describe the concept fully, including a fairly comprehensive review of functional analysis (Sections 2.1 to 2.3) germane to our discussions. These two chapters constitutes the theoretical basis for the I-prior methodology.

Subsequently, this thesis has three main chapters which we hope to present as methodological innovations surrounding the use of I-priors for modelling. Chapter 4 describes the I-prior modelling framework and computational methods relating to the estimation of I-prior models. Chapter 5 extends the I-prior methodology to fit categorical outcome models. Chapter 6 discusses the use of I-priors in variable selection for linear models. In addition to introducing the statistical model of interest and motivating the use of I-priors, this introductory chapter ultimately provides a summary outline of the thesis.

1.1 Regression models

For subject $i \in \{1, \ldots, n\}$, assume a real-valued response y_i has been observed, as well as a row vector of p covariates $x_i = (x_{i1}, \ldots, x_{ip})$, where each x_{ik} belongs to some set $\mathcal{X}_k$, for $k = 1, \ldots, p$. Let $\mathcal{S} = \{(y_1, x_1), \ldots, (y_n, x_n)\}$ denote this observed sample of size n. Consider then the following regression model, which stipulates the dependence of the y_i's on the x_i's:

$$y_i = \alpha + f(x_i) + \epsilon_i. \tag{1.1}$$

Here, f is a regression function to be estimated, and α is an intercept. Additionally, it is assumed that the errors ϵ_i are zero-meaned and normally distributed according to

$$(\epsilon_1, \ldots, \epsilon_n)^\top \sim \mathrm{N}_n(\mathbf{0}, \mathbf{\Psi}^{-1}), \tag{1.2}$$

where $\mathbf{\Psi} = (\psi_{ij})_{i,j=1}^n$ is the precision matrix. We shall often refer to model (1.1) subject to (1.2) as the *normal regression model*. The choice of multivariate normal errors is not only a convenient one (as far as distributional assumptions go), but one that is also motivated by the principle of maximum entropy (Jaynes, 1957a, 1957b, 2003).

Interestingly, a wide variety of statistical models can be captured by the seemingly humble normal regression model, simply by varying the form of the regression function f. For instance, when f can be parameterised linearly as $f(x_i) = x_i^\top \beta$, $\beta \in \mathbb{R}^p$, we then have the ordinary linear regression—a staple problem in statistics and other quantitative fields.

We might also have data that is separated naturally into groups or levels by design, for example, data from stratified sampling, students within schools, or longitudinal measurements over time. In such cases, we might want to consider a regression function with additive components

$$f(x_i^{(j)}, j) = f_1(x_i^{(j)}) + f_2(j) + f_{12}(x_i^{(j)}, j)$$

where $x_i^{(j)}$ denotes the p-dimensional i'th observation for group $j \in \{1, \ldots, m\}$. Again, assuming a linear parameterisation, this is recognisable as the standard multilevel or random-effects linear model (Rabe-Hesketh and Skrondal, 2012), with f_2 representing the varying intercept via $f_2(j) = \alpha_j$, f_{12} representing the varying slopes via $f_{12}(x_i^{(j)}, j) = x_i^{(j)\top} u_j$, $u_j \in \mathbb{R}^p$, and f_1 representing the fixed-effects linear component $x_i^{(j)\top} \beta$ as in the linear model above.

Moving on from linear models, smoothing models may be of interest as well. A myriad of models exist for this type of problem, with most classed as nonparametric regression (Wassermann, 2006), and the more popular ones include LOcal regrESSion (LOESS),

kernel regression, and smoothing splines (Wahba, 1990). Semiparametric regression models, on the other hand, combines the linear component of a regression model with a nonparameteric component.

Further, the regression problem is made more intriguing when the set of covariates $\mathcal{X}$ is functional—in which case the linear regression model aims to estimate coefficient functions $\alpha, \beta : \mathcal{T} \to \mathbb{R}$ from the model

$$y_i = \int_{\mathcal{T}} \Big\{ \alpha(t) + x_i(t)\beta(t) \Big\} \, \mathrm{d}t + \epsilon_i.$$

Nonparametric and semiparametric regression with functional covariates have also been widely explored (Ramsay and Silverman, 2005). Models of this nature still fall under the remit of the normal regression model by selecting a regression functional with domain over the functional covariates.

1.2 Vector space of functions

It would be beneficial to prescribe some sort of structure for which estimation of the regression function can be carried out methodically and reliably. This needed structure is given to us by assuming that our regression function f for the normal model lies in some topological vector space, namely, a reproducing kernel Hilbert or Kreĭn space (RKHS/RKKS) $\mathcal{F}$ equipped with the reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Often, the reproducing kernel (or simply kernel, for short) is shaped by one or more parameters which we shall denote by η. Correspondingly, the kernel is rightfully denoted h_η to indicate the dependence of the parameters on the kernels, though where this is seemingly obvious, might be omitted. For I-prior modelling, which is the focus of this thesis, we make the assumption that our regression function lies in an RKKS $\mathcal{F}$.

RKKSs, and more popularly RKHSs, provide a geometrical advantage to learning algorithms: projections of the inputs to a richer and more informative (and usually higher dimensional) *feature space*, where learning is more likely to be successful, need not be figured out explicitly. Instead, *feature maps* are implicitly calculated by the use of kernel functions. This is known as the "kernel trick" in the machine learning literature (Hofmann et al., 2008), and it has facilitated the success of kernel methods for learning, particularly in algorithms with inner products involving the transformed inputs.

Due to the one-to-one mapping between the set of kernel functions and the set of RKHSs, choosing the space in which the regression function lies is equivalent to choosing a particular kernel function, and this is chosen according to the desired effects of the covariates on the regression function. RKKSs on the other hand also possess unique

kernels, but every (generalised) kernel[1] is associated to *at least* one RKKS. An in-depth discussion (including the motivation for their use) on kernels, RKHSs and RKKSs will be provided later in Chapter 2, but for now, it suffices to say that kernels which invoke either a linear, smooth or categorical dependence, or any combinations thereof, are of interest. This would allow us to fit the various models described earlier within this RKHS/RKKS framework.

1.3 Estimating the regression function

Having decided on a vector space $\mathcal{F}$, we now turn to the task of choosing the best $f \in \mathcal{F}$ that fits the data sample $\mathcal{S}$. 'Best' here could mean a great deal of things, such as choosing f which minimises an empirical risk measure[2] defined by

$$\hat{\mathrm{R}}(f) = \frac{1}{n}\sum_{i=1}^{n} \Lambda\big(y_i, f(x_i)\big)$$

for some loss function $\Lambda : \mathbb{R}^2 \to [0, \infty)$. A common choice for the loss function is the *squared loss function*

$$\Lambda\big(y_i, f(x_i)\big) = \sum_{j=1}^{n} \psi_{ij}\big(y_i - f(x_i)\big)\big(y_j - f(x_j)\big),$$

and when used, defines the *(generalised) least squares regression*. For the normal model, the minimiser of the empirical risk measure under the squared loss function is also the maximum likelihood (ML) estimate of f, since $\hat{\mathrm{R}}(f)$ would be twice the negative log-likelihood of f, up to a constant.

The ML estimator of f typically interpolates the data if the dimension of $\mathcal{F}$ is at least n, so is of little use. The most common method to overcome this issue is *Tikhonov regularisation*, whereby a regularisation term is added to the risk function, with the aim of imposing a penalty on the complexity of f. In particular, smoothness assumptions on f can be represented by using its RKHS norm $\|\cdot\|_{\mathcal{F}} : \mathcal{F} \to \mathbb{R}$ as the regularisation term[3]. Therefore, the solution to the regularised least squares problem—call this f_{reg}—is the minimiser of the mapping from $\mathcal{F}$ to $\mathbb{R}$ defined by

$$f \mapsto \overbrace{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \psi_{ij}\big(y_i - f(x_i)\big)\big(y_j - f(x_j)\big)}^{\text{data fit term}} + \overbrace{\lambda^{-1}\|f - f_0\|_{\mathcal{F}}^2}^{\text{penalty term}}, \tag{1.3}$$

[1] By generalised kernels, we mean kernels that are not necessarily positive definite in nature.

[2] More appropriately, the risk functional $\mathrm{R}(f) = \int \Lambda(y, f(x))\,\mathrm{dP}(y, x)$, i.e. the expectation of the loss function under some probability measure of the observed sample, should be used. Often the true probability measure is not known, so the empirical risk measure is used instead.

which also happens to be the *penalised maximum likelihood* solution. Here, $f_0 \in \mathcal{F}$ can be thought of a prior "best guess" for the function f. The $\lambda^{-1} > 0$ parameter—known as the regularisation parameter—controls the trade-off between the data-fit term and the penalty term in (1.3), and is not usually known a priori and must be estimated.

An attractive consequence of the representer theorem (Kimeldorf and Wahba, 1970) for Tikhonov regularisation implies that f_{reg} admits the form

$$f_{\text{reg}} = f_0 + \sum_{i=1}^{n} h(\cdot, x_i) w_i, \qquad w_i \in \mathbb{R}, \ \forall i = 1, \dots, n, \tag{1.4}$$

even if $\mathcal{F}$ is infinite dimensional. This simplifies the original minimisation problem from a search for f over a possibly infinite-dimensional domain, to a search for the optimal coefficients w_i in n dimensions.

Tikhonov regularisation also has a well known Bayesian interpretation, whereby the regularisation term encodes prior information about the function f. For the normal regression model with $f \in \mathcal{F}$, an RKHS, it can be shown that f_{reg} is the posterior mean of f given a *Gaussian process prior* (Rasmussen and Williams, 2006) with mean f_0 and covariance kernel $\text{Cov}\big(f(x_i), f(x_j)\big) = \lambda\, h(x_i, x_j)$. The exact solution for the coefficients $\mathbf{w} := (w_1, \dots, w_n)^\top$ are in fact $\mathbf{w} = \big(\mathbf{H} + \mathbf{\Psi}^{-1}\big)^{-1}(\mathbf{y} - \mathbf{f}_0)$, where $\mathbf{H} = \big(h(x_i, x_j)\big)_{i,j=1}^{n}$ (often referred to as the Gram matrix or kernel matrix) and $(\mathbf{y} - \mathbf{f}_0) = (y_1 - f_0(x_1), \dots, y_n - f_0(x_n))^\top$.

1.4 Regression using I-priors

Building upon the Bayesian interpretation of regularisation, Bergsma (2018) proposes an original prior distribution for the regression function such that its realisations admit the form for the solution given in the representer theorem. The *I-prior* for the regression function f in (1.1) subject to (1.2) and $f \in \mathcal{F}$, an RKKS with kernel h_η, is defined as the distribution of a random function of the form (1.4) when the w_i are distributed according to

$$(w_1, \dots, w_n)^\top \sim \text{N}_n(\mathbf{0}, \mathbf{\Psi}),$$

where $\mathbf{0}$ is a length n vector of zeroes, and $\mathbf{\Psi}$ is the error precision matrix. As a result, we may view the I-prior for f as having the Gaussian process distribution

$$\mathbf{f} := \big(f(x_1), \dots, f(x_n)\big)^\top \sim \text{N}_n(\mathbf{f}_0, \mathbf{H}_\eta \mathbf{\Psi} \mathbf{H}_\eta), \tag{1.5}$$

[3] Concrete notions of complexity penalties can be introduced if $\mathcal{F}$ is a normed space, though RKHSs are typically used as it gives great conveniences.

with $\mathbf{H}_\eta$ an $n \times n$ matrix with (i, j) entries equal to $h_\eta(x_i, x_j)$, and $\mathbf{f}_0$ a vector containing the $f_0(x_i)$'s, $i = 1, \ldots, n$. The covariance matrix of this multivariate normal prior is related to the Fisher information for f, and hence the name I-prior—the 'I' stands for information. Furthermore, the I-prior happens to be an entropy-maximising prior, subject to certain constraints. Chapter 3 contains details of the derivation of I-priors for the normal regression model.

As with Gaussian process regression (GPR), the function f is estimated by its posterior mean. For the normal model, the posterior distribution for the regression function conditional on the responses $\mathbf{y} = (y_1, \ldots, y_n)$,

$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}}, \tag{1.6}$$

can easily be found, and it is in fact normally distributed. The posterior mean for f evaluated at a point $x \in \mathcal{X}$ is given by

$$\mathrm{E}\left(f(x)\middle|\mathbf{y}\right) = f_0(x) + \mathbf{h}_\eta^\top(x)\overbrace{\boldsymbol{\Psi}\mathbf{H}_\eta\left(\mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{H}_\eta + \boldsymbol{\Psi}^{-1}\right)^{-1}(\mathbf{y} - \mathbf{f}_0)}^{\tilde{\mathbf{w}}} \tag{1.7}$$

where we have defined $\mathbf{h}_\eta(x)$ to be the vector of length n with entries $h_\eta(x, x_i)$ for $i = 1, \ldots, n$. Incidentally, the elements of the n-vector $\tilde{\mathbf{w}}$ defined in (1.7) are the posterior means of the random variables w_i in the formulation (1.4). The point-evaluation posterior variance for f is given by

$$\mathrm{Var}\left(f(x)\middle|\mathbf{y}\right) = \mathbf{h}_\eta^\top(x)\left(\mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{H}_\eta + \boldsymbol{\Psi}^{-1}\right)^{-1}\mathbf{h}_\eta(x). \tag{1.8}$$

Prediction for a new data point $x_{\text{new}} \in \mathcal{X}$ then concerns obtaining the *posterior predictive distribution*

$$p(y_{\text{new}}|\mathbf{y}) = \int p(y_{\text{new}}|f_{\text{new}}, \mathbf{y})p(f_{\text{new}}|\mathbf{y})\,\mathrm{d}f_{\text{new}},$$

where we had defined $f_{\text{new}} := f(x_{\text{new}})$. This is again a normal distribution in the case of the normal model, with similar mean and variance as in (1.7). For a derivation, see Section 4.2 (p. 109) in Chapter 4 for details.

There is also the matter of optimising model parameters θ, which in our case, collectively refers to the kernel parameters η and the precision matrix of the errors $\boldsymbol{\Psi}$. Model parameters θ may be estimated in several ways, either by likelihood-based methods or fully Bayesian methods. The former includes methods such as direct maximisation of the (marginal) likelihood, $L(\theta) = \int p(\mathbf{y}|\theta, \mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}$, and the expectation-maximisation (EM) algorithm. Both are seen as a form of *empirical Bayes* estimation, or a type-II ML estimation (Bishop, 2006), as it is known in machine learning. In a fully Bayesian setting on the other hand, Markov chain Monte Carlo (MCMC) may be employed, assuming prior distributions on the model parameters.

1.5 Advantages and limitations of I-priors

The I-prior methodology has the following advantages:

1. **A unifying methodology for various regression models.**

   The I-prior methodology has the ability to fit a multitude of regression models simply by choosing the RKKS to which the regression function belongs. As such, it can be seen as a unifying methodology for various parametric and nonparametric regression models including additive models, multilevel models and models with one or more functional covariates.

2. **Simple estimation procedure.**

   Estimation of model parameters using the aforementioned methods are very simple to implement, barring any computational and numerical hurdles, which shall be discussed in Chapter 4.

3. **Parsimonious specification.**

   I-prior models are most typically specified using only RKHS scale parameters and the error precision. This encourages parsimony in model building; for example, smoothing models can be fitted using only two parameters, while linear multilevel models can be fitted with notably fewer parameters than the standard versions.

4. **Prevents overfitting and undersmoothing.**

   As alluded to earlier, any function f that passes through the data points is a least squares solution. Regularising the problem with the use of I-priors prevents overfitting, with the added advantage that the posterior solution under an I-prior does not tend to undersmooth as much as Tikhonov regularisation does (Bergsma, 2018). Undersmoothing can adversely impact the estimate of f, and in real terms might even show features and artefacts that are not really there.

5. **Better prediction.**

   Empirical studies and real-data examples show that predictive performance of I-priors are comparative to, and often better than, other leading state-of-the-art models, including the closely related GPR.

6. **Straightforward inference.**

   Marginal likelihoods after integrating out the I-prior are easily obtained, making model selection via likelihood comparison a viable option. This method of comparing marginal likelihood with maximum likelihood estimate plug-ins of the model parameters, is viewed as empirical Bayes factors comparison in the Bayesian literature (Casella, 1985; George and Foster, 2000).

The main drawback of using I-prior models is computational in nature, namely, the requirement of working with an $n \times n$ matrix and its inverse, as seen in equations (1.7) and (1.8), regardless of estimation method (ML or Bayes). Analysis of data sets that are not more than a few thousand in size can be considered feasible; anything more than this is debilitatingly slow to compute. In addition, care must be taken to avoid numerical instabilities when calculating the marginal log-likelihood during parameter estimation, which can affect gradient based optimisation or the EM algorithm.

Another issue when performing likelihood-based methods is that the optimisation objective may be non-convex such that multiple local optima may exist. In such cases, multiple restarts from different initialisations may ultimately lead to a global maximum, although difficulties may be faced if numerical instabilities occur.

Lastly, a remark on model assumptions, which are twofold: 1) the assumption of $f \in \mathcal{F}$ an RKKS; and 2) normality of errors. Of the two, the latter is more likely to be violated, especially when dealing with discrete responses, e.g. in classification. Deviating from the normality assumption would require approximation techniques to be implemented in order to obtain the posterior distributions of interest.

1.6 Outline of thesis

This thesis is structured as follows:

- Following this introductory chapter, **Chapter 2** provides an overview of functional analysis, and in particular, descriptions of interesting function spaces for regression. In **Chapter 3**, the concept of the Fisher information is extended to potentially infinite-dimensional parameters. This allows us to define the Fisher information for the regression function which parameterises the normal regression model, and we explain how this relates to the I-prior.
- The aforementioned computational methods relating to the estimation of I-prior models are explored in **Chapter 4**, namely the direct optimisation of the log-likelihood, the EM algorithm, and MCMC methods. The goal is to describe stable and efficient algorithms for estimating I-prior models. The R package **iprior** (Jamil, 2017) is the culmination of the effort put in towards completing this chapter, which has been made publicly available on the Comprehensive R Archive Network (CRAN).
- Many models of interest involve response variables of a categorical nature. A naïve implementation of the I-prior model is certainly possible, but proper ways do exist to handle non-normality of errors. **Chapter 5** extends the I-prior methodology to discrete outcomes. There, the non-Gaussian likelihood that arises in the posteriors

are approximated by way of variational inference. The advantages of the I-prior in normal regression models carry over into categorical response models.

- **Chapter 6** is a contribution to the area of variable selection. Specifically for linear models with p variables to select from, model comparison requires elucidation of 2^p marginal likelihoods, and this becomes infeasible when p is large. To circumvent this issue, we use a stochastic search method to choose models that have high posterior probabilities of occurring, equivalent to choosing models that have large Bayes factors. We experiment with the use of I-priors to improve false selections, especially in the presence of multicollinearity.

Chapters 4 to 6 contain R computer implementations of the statistical methodologies described therein, and the code for replication are made available at `http://myphdcode.haziqj.ml`.

Familiarity with basic estimation concepts (maximum likelihood, Bayes, empirical Bayes) and their corresponding estimation methods (gradient-based methods, Newton, quasi-Newton methods, MCMC, EM algorithm) are assumed throughout. Brief supplementary chapters are attached for readers who wish to familiarise themselves with topics such as variational inference and Hamiltonian Monte Carlo, which are used in Chapters 4 and 5. These brief readings are designed to be ancillary in nature, and are not strictly essential for the main chapters. Additionally, Appendices A to I contain references to several statistical distributions and their properties, proofs of various claims, and derivations of the algorithms described in this thesis.

On a closing note, a dedicated website for this PhD project has been created, and it can be viewed at `http://phd.haziqj.ml`.

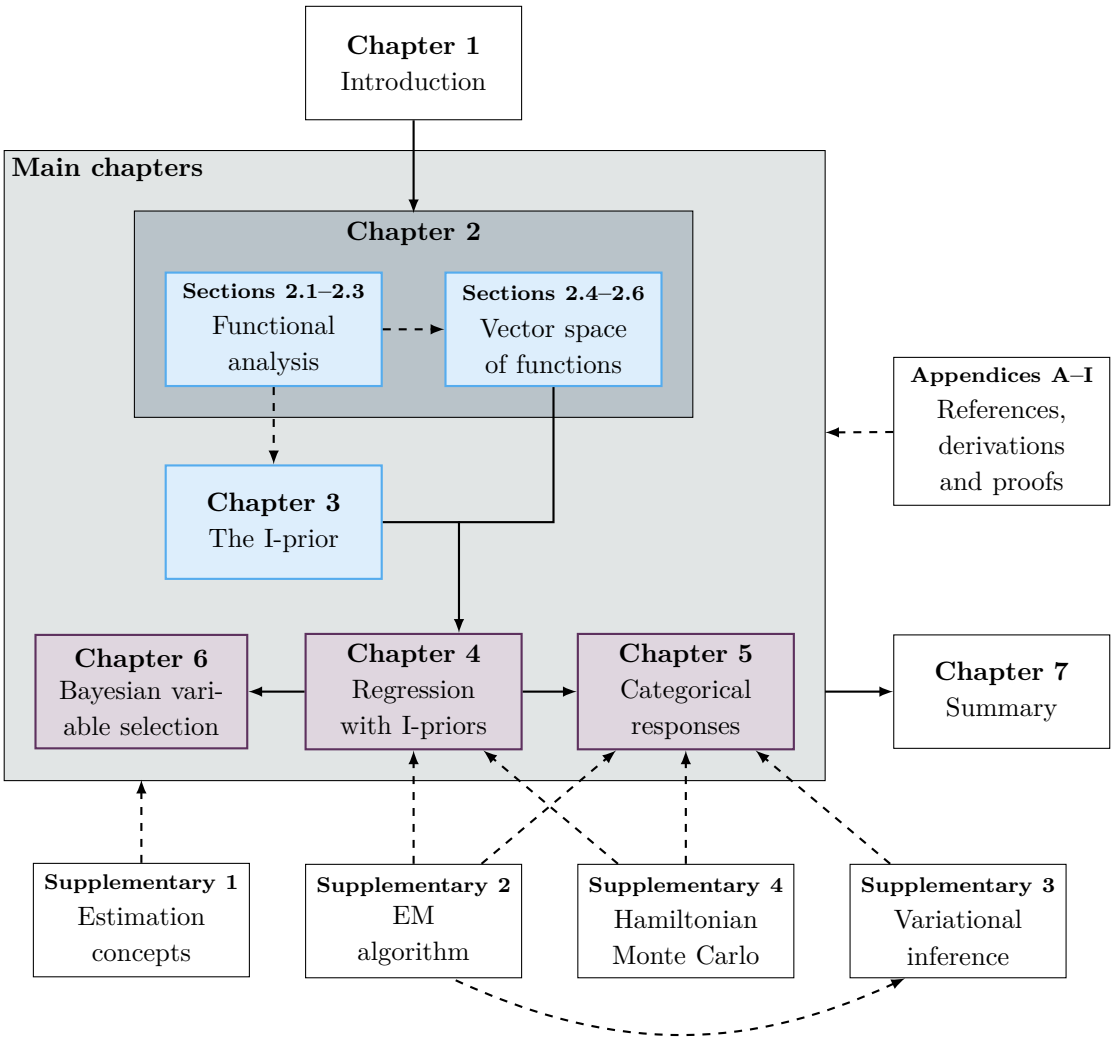

Figure 1.1: Schematic representation of the organisation of the chapters of this thesis. Solid lines indicate requisite relevances, while dashed lines indicate supporting and supplementary relevances. Chapters indicated by **blue** boxes are theoretical in nature, while those in **purple** are methodological.

Bibliography

Bergsma, Wicher (2018). *Regression and classification with I-priors*. Manuscript in submission. arXiv: 1707.00274 [math.ST].

Bishop, Christopher (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag. ISBN: 978-0-387-31073-2.

Casella, George (1985). "An Introduction to Empirical Bayes Data Analysis". In: *The American Statistician* 39.2, pp. 83–87. DOI: 10.2307/2682801.

George, Edward I. and Dean P. Foster (2000). "Calibration and Empirical Bayes Variable Selection". In: *Biometrika* 87.4, pp. 731–747. DOI: 10.1093/biomet/87.4.731.

Hofmann, Thomas, Bernhard Schölkopf, and Alexander J. Smola (2008). "Kernel Methods in Machine Learning". In: *The Annals of Statistics* 36.3, pp. 1171–1220. DOI: 10.1214/009053607000000677.

Jamil, Haziq (2017). ***iprior**: Regression Modelling using I-Priors*. R package version 0.7.1. URL: https://cran.r-project.org/web/packages/iprior.

Jaynes, Edwin Thompson (1957a). "Information Theory and Statistical Mechanics". In: *Physical Review* 106.4, p. 620. DOI: 10.1103/PhysRev.106.620.

— (1957b). "Information Theory and Statistical Mechanics II". In: *Physical Review* 108.2, p. 171. DOI: 10.1103/PhysRev.108.171.

— (2003). *Probability Theory: The Logic of Science*. Ed. by G. Larry Bretthorst. Cambridge University Press. ISBN: 978-0-521-59271-0.

Kimeldorf, George S and Grace Wahba (1970). "A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines". In: *The Annals of Mathematical Statistics* 41.2, pp. 495–502. DOI: 10.1214/aoms/1177697089.

Rabe-Hesketh, Sophia and Anders Skrondal (2012). *Multilevel and Longitudinal Modeling Using Stata*. 3rd ed. Stata Press. ISBN: 978-1-59718-108-2.

Ramsay, James and Bernard W. Silverman (2005). *Functional Data Analysis*. New York: Springer-Verlag. ISBN: 978-1-4757-7107-7. DOI: 10.1007/978-1-4757-7107-7.

Rasmussen, Carl Edward and Christopher K. I. Williams (2006). *Gaussian Processes for Machine Learning*. The MIT Press. ISBN: 0-262-18253-X. URL: http://www.gaussianprocess.org/gpml/.

Wahba, Grace (1990). *Spline Models for Observational Data*. SIAM. ISBN: 978-0-89871-244-5. DOI: 10.1137/1.9781611970128.

Wassermann, Larry (2006). *All of Nonparametric Statistics*. New York: Springer-Verlag. ISBN: 978-0-387-25145-5. DOI: 10.1007/0-387-30623-4.