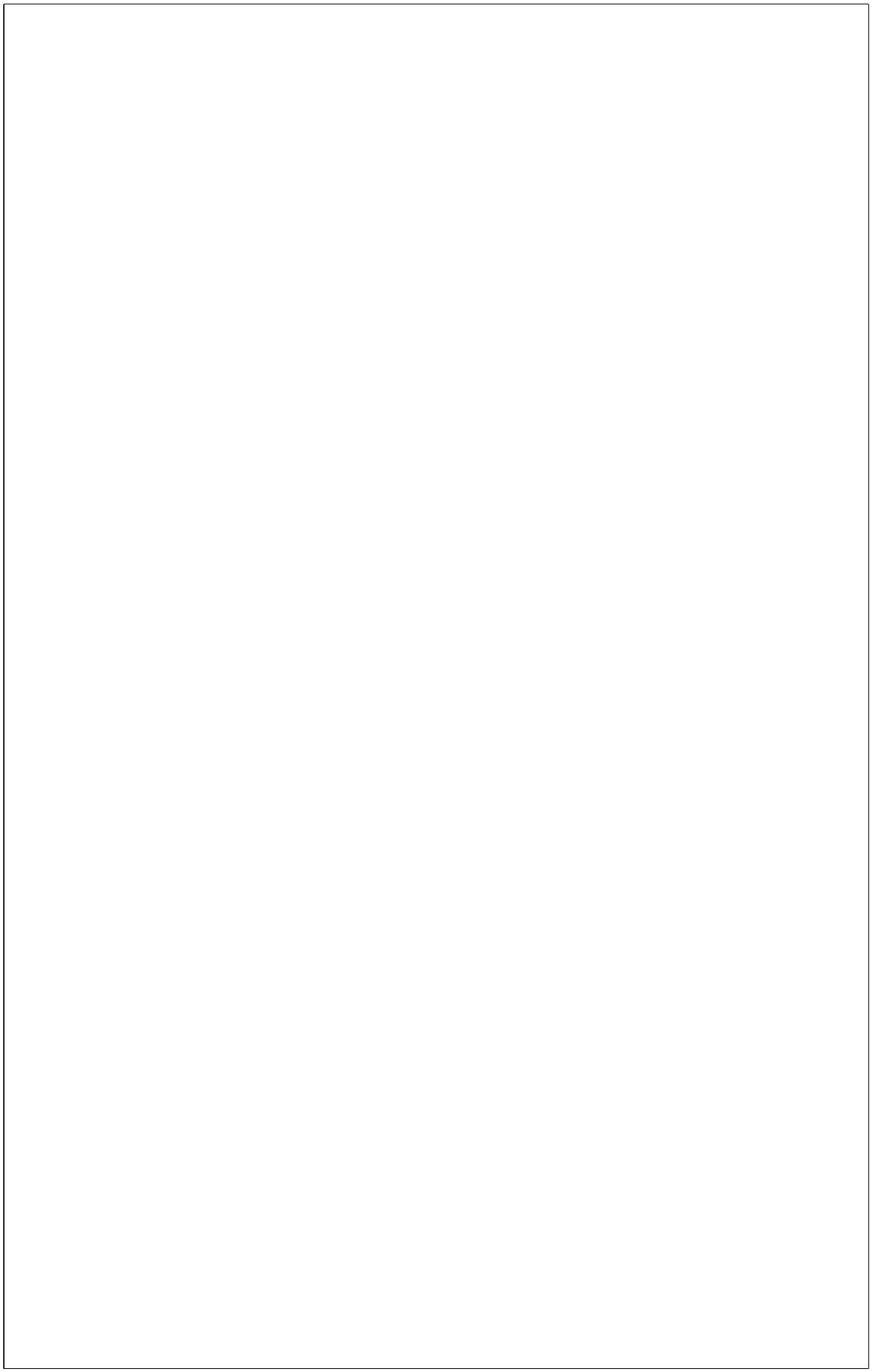
To-do list

Contents



Haziq Jamil
Department of Statistics
London School of Economics and Political Science
PhD thesis: 'Regression modelling using Fisher information covariance kernels (I-priors)'

Chapter 7

Summary

The work done in this thesis explores the concept of regression modelling using priors with Fisher information covariance kernels (I-priors, Bergsma, 2017). It is best seen as a flexible regression technique which is able to fit both parametric and nonparametric models, and bears similarity to Gaussian process regression. For the regression model (1.1) subject to (1.2), stated again here for convenience,

$$\begin{aligned} y_i &= \alpha + f(x_i) + \epsilon_i && \text{(from 1.1)} \\ (\epsilon_1, \dots, \epsilon_n) &\sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi}^{-1}) && \text{(from 1.2)} \\ i &= 1, \dots, n, \end{aligned}$$

and it is assumed that the regression function f lies in some reproducing kernel Hilbert or Krein space $\mathcal{F}$ with kernel h_η defined over the set of covariates $\mathcal{X}$. In Chapter 2, we built a primer on basic functional analysis, and described various interesting RKHS/RKKS for regression modelling.

We then ascertained the form of the Fisher information for f, treated as a parameter of the model to be estimated, and from Corollary 3.3.1, it is

$$\begin{aligned} \mathcal{I}\big(f(x), f(x')\big) &= \sum_{i,j=1}^{n} \psi_{ij} h_\eta(x, x_i) h_\eta(x', x_j) \\ &= \mathbf{h}_\eta(x)^\top \boldsymbol{\Psi} \mathbf{h}_\eta(x'), \end{aligned}$$

for any two points x, x' in the domain of f, obtained using appropriate calculus for topological spaces detailed in Chapter 3. An I-prior for f is defined as Gaussian with mean function f_0 chosen a priori, and covariance function equal to the Fisher information.

The I-prior for f has the simple representation

$$f(x_i) = f_0(x_i) + \sum_{k=1}^{n} h_\eta(x_i, x_k) w_k$$
$$(w_1, \ldots, w_n)^\top \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi})$$
$$i = 1, \ldots, n,$$

and is written equivalently as the Gaussian process prior

$$\big(f(x_1), \ldots, f(x_n)\big)^\top \sim \mathrm{N}_n(\mathbf{f}_0, \mathbf{H}_\eta \boldsymbol{\Psi} \mathbf{H}_\eta),$$

where $\mathbf{H}_\eta = \big(h_\eta(x_i, x_j)\big)_{i,j=1}^n$.

In Chapter 4, we looked how the I-prior model has wide-ranging applications, from multilevel modelling, to longitudinal modelling, and modelling with functional covariates. Estimation was conducted mainly using a simple EM algorithm, although direct optimisation and fully Bayesian estimation using MCMC is also possible. In the case of polytomous responses, we used a latent variable framework in Chapter 5 to assign I-priors to latent propensities which drive the outcomes under a probit-transform scheme. An extension of the EM algorithm was considered, in which the E-step was replaced with variational inference, so as to overcome the intractability brought about by the conditional distributions. For both continuous and categorical response I-prior models, we find advantages of using I-priors, namely that model building and estimation is simple, inference straightforward, and predictions comparable, if not better, to similar state-of-the-art techniques.

Finally, in Chapter 6, we dealt with the problem of model selection, specifically for linear regression models. There, we used a fully Bayesian approach for estimating model probabilities in which regression coefficients are assigned an I-prior. We devised a model that requires minimal tuning on the part of the user, yet performs well in simulated and real-data examples, especially if multicollinearity exists among the covariates.

7.1 Summary of contributions

We give a summary of the novel contributions of this thesis.

- **Fisher information for infinite-dimensional parameters**. When the RKHS/ RKKS $\mathcal{F}$ is infinite-dimensional (e.g. covariates are themselves functions), then the Fisher information involves derivatives with respect to an infinite-dimensional vector. Finite-dimensional results using component-wise/partial derivatives may fail in infinite dimensions. The technology of Fréchet and Gâteaux differentials

accommodate for the fact that f may be infinite-dimensional, which, at minimum, requires $\mathcal{F}$ to be a normed vector space. We foresee the work of Section 3.2 being applicable elsewhere, such as learning in (reproducing kernel) Banach spaces (H. Zhang et al., 2009; H. Zhang and J. Zhang, 2012), or in the theory of parameter estimation for general exponential family type distributions of the form

$$p(X|\theta) = B(X) \exp\big(\langle \theta, T(X) \rangle_{\mathcal{H}} - A(\theta)\big),$$

for which θ lies in some inner-product space $\mathcal{H}$ which might be infinite-dimensional (Sriperumbudur et al., 2013).

- **Efficient estimation methods for normal I-prior models**. The preferred estimation method for normal I-prior models for stability is the EM algorithm. Implementing the EM algorithm can be computationally costly, due to the squaring and inversion of the kernel matrices in the Q function in (4.15). Unfortunately, not much can be done about the inversion part, but we explored some ways to perform the squaring methodically. Combining a 'front-loading method' of the kernel matrices (Section 4.3.2) and an exponential family ECM (expectation conditional maximisation) algorithm (Meng and Rubin, 1993), the estimation procedure is streamlined. Our computational work culminated in the publicly available and well-documented R package **iprior** (Jamil, 2017) published on CRAN.
- **Methodological extension of I-priors to categorical responses**. Extension of the I-prior methodology to fit categorical responses is of great interest. We proposed a latent variable framework, for which there corresponds latent propensities corresponding to each category of the observed response variable. Instead of modelling the responses directly, the latent propensities are modelled using an I-prior, and class probabilities obtained using a normal integral. We named this model the I-probit model. The challenge of estimation is overcoming said integral, and we used a variational EM algorithm in which the E-step uses a variational approximation to intractable conditional density. The variational EM algorithm was preferred over a fully Bayesian variational inference algorithm for two reasons: 1) the work done in the continuous case EM algorithm applies directly; and 2) prior specification for hyperparameter can be dispensed with. Classification, meta-analysis and spatio-temporal modelling are specific examples of the applications of the I-probit models.
- **Some distributional results for truncated normals**. In deriving the variational algorithm, some properties related to the conically truncated multivariate independent normal distribution (as defined in Appendix C.4) were required. A small contribution of ours was to derive the closed-form expressions for its first and second moments, and its entropy (Lemma C.5). We have only seen closed-

form expressions of the mean of such a distribution being used before (Girolami and Rogers, 2006) but not for the variance, nor an explicit derivation of these quantities.

- **Bayesian variable selection under collinearity**. Model comparison using likelihood ratio tests or Bayes factors is fine when the number of models under consideration is fairly small. Under a fully Bayesian scheme, we use MCMC to approximate posterior model probabilities of competing linear models. At the outset, we sought a model which required minimal intervention on the part of the user. The I-prior achieved this, with the added advantage of performing well under multicollinearity.

7.2 Open questions

In closing, we briefly discuss several questions which remain open during the course of completing this project.

- **Initialisation of EM or gradient-based methods**. Figure 4.1 indicates the impact that starting values can have on gradient-based optimisation. One can end up at a local optima on one of the two ridges. Usually, one of the ridges will have a higher maximum than the other, but it is not clear how to direct the algorithm in the direction of the 'correct' ridge.

  Importantly, the interpretation of a flat ridge in the likelihood is that there is insufficient information coming from the data to inform parameter estimation. In the EM algorithm, estimation is usually characterised by a fast increase in likelihood in the first few steps (as it climbs up the ridge), and then later iterations only improve the likelihood ever so slightly (as it moves along the ridge in search of the maximum). In some real-data cases (e.g. Tecator data set), we noticed that the EM sequence veers to the boundary of the parameter space, where the likelihood is infinite (e.g. $L(\psi) \to \infty$ as $\psi \to 0, \infty$).

  Ill-posed problems similar to this are resolved by adding penalty terms to the log-likelihood. As to what penalty terms are appropriate remains an open question.

- **Standard errors for variational approximation**. Under a variational scheme, the log-likelihood function $L(\theta)$ is replaced with the ELBO $\mathcal{L}_q(\theta)$ which serves as a conservative approximation to it. The question we have is whether the approximation degrades the asymptotic properties of the estimators obtained via variational inference? In particular, are the standard errors obtained from the information matrix involving $\mathcal{L}_q(\theta)$ reliable? This question has also been posed by Bickel et al. (2013), Chen et al. (2017), and Hall et al. (2011).

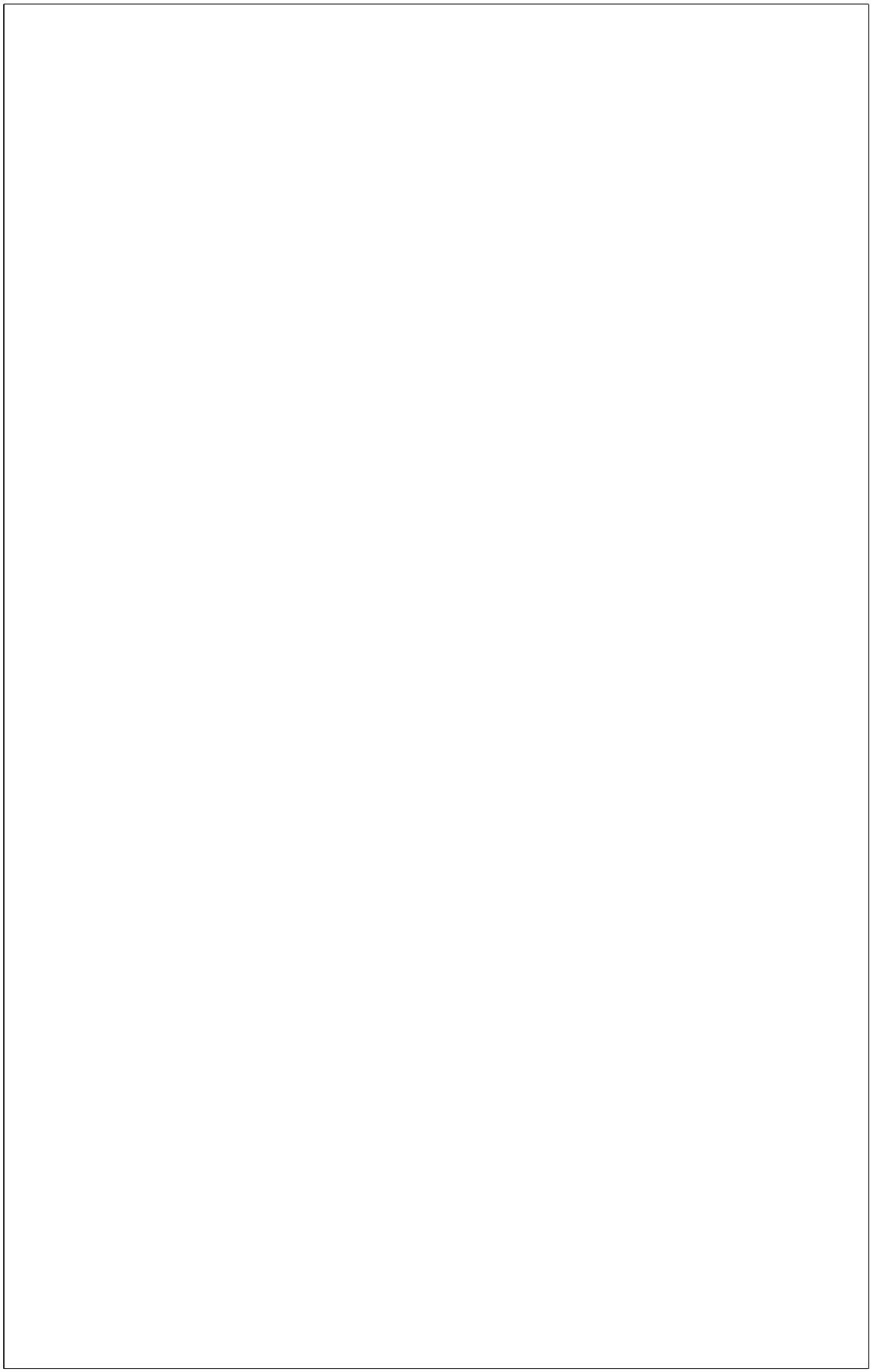
Variational methods for maximum likelihood learning can be seen as a deliberate misspecification of the model to achieve tractibility. As such, the variational EM has been referred to as obtaining pseudo- or quasi-ML estimates. The quasi-likelihood literature has results relating to efficiency of parameter estimates (adjustments to the information matrix is needed), and we wonder if these are applicable for variational inference.

Also, obtaining standard errors directly from an EM algorithm is of interest, especially under a variational EM setting. Though this is described in McLachlan and Krishnan (2007, Ch. 4), we have not seen this implemented widely.

- **Comparison of logistic and probit links**. For general binary and multinomial models, the logistic link function sees more prevalent use than its probit counterpart. Of course, we chose the probit as it has distributional advantages which we can exploit for estimation using variational inference. However, is there a difference between the behaviour of the probit and logistic model? We know that there is a difference between the logistic and normal distribution, especially in scaling and behaviour in the tails, but do these affect the outcome of I-prior models?
- **Consistency of I-prior Bayesian variable selection**. We wondered about model selection consistency for I-priors in Bayesian variable selection. That is, assuming that model M_{true} is behind the true data generative process, do

$$\lim_{n\to\infty} \mathrm{P}(M_{\text{true}}|\mathbf{y}) = 1 \quad \text{and} \quad \lim_{n\to\infty} \mathrm{P}(M_k|\mathbf{y}) = 0, \forall M_k \neq M_{\text{true}}$$

hold for the I-prior Bayesian variable selection methodology? In machine learning, this property is referred to as the *oracle property*. For the g-prior specifically, model consistency results were obtained by Fernandez et al. (2001) and Liang et al. (2008). Casella et al. (2009) also looks at consistency of Bayesian procedures for a wide class of prior distributions, but we have yet to examine whether the I-prior falls under the remit of their work.



Bibliography

bergsma2017 Bergsma, Wicher (2017). "Regression with I-priors". In: *Unpublished manuscript*.

bickel2013asymptotic Bickel, Peter, David Choi, Xiangyu Chang, and Hai Zhang (2013). "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels". In: *The Annals of Statistics*, pp. 1922–1943.

casella2009consistency Casella, George, F Javier Girón, M Lina Martínez, and Elias Moreno (2009). "Consistency of Bayesian procedures for variable selection". In: *The Annals of Statistics*, pp. 1207–1228.

chen2017use Chen, Yen-Chi, Y Samuel Wang, and Elena A Erosheva (2017). "On the use of bootstrap with variational inference: Theory, interpretation, and a two-sample test example". In: *arXiv preprint arXiv:1711.11057*.

fernandez2001benchmark Fernandez, Carmen, Eduardo Ley, and Mark FJ Steel (2001). "Benchmark priors for Bayesian model averaging". In: *Journal of Econometrics* 100.2, pp. 381–427.

girolami2006variational Girolami, Mark and Simon Rogers (2006). "Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors". In: *Neural Computation* 18.8, pp. 1790–1817.

hall2011asymptotic Hall, Peter, Tung Pham, Matt P Wand, Shen SJ Wang, et al. (2011). "Asymptotic normality and valid inference for Gaussian variational approximation". In: *The Annals of Statistics* 39.5, pp. 2502–2532.

jamil2017iprior Jamil, Haziq (2017). ***iprior**: Regression Modelling using I-Priors*. R package version 0.7.1. URL: `https://cran.r-project.org/web/packages/iprior`.

liang2008mixtures Liang, Feng, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger (2008). "Mixtures of g priors for Bayesian variable selection". In: *Journal of the American Statistical Association* 103.481, pp. 410–423.

mclachlan2007algorithm McLachlan, Geoffrey and Thriyambakam Krishnan (2007). *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons.

meng1993maximum Meng, Xiao-Li and Donald B Rubin (1993). "Maximum likelihood estimation via the ECM algorithm: A general framework". In: *Biometrika* 80.2, pp. 267–278.

sriperumbudur2013density Sriperumbudur, Bharath, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar (2013). "Density estimation in infinite dimensional exponential families". In: *arXiv preprint arXiv:1312.3516*.

zhang2009reproducing Zhang, Haizhang, Yuesheng Xu, and Jun Zhang (2009). "Reproducing kernel Banach spaces for machine learning". In: *Journal of Machine Learning Research* 10.Dec, pp. 2741–2775.

zhang2012regularized Zhang, Haizhang and Jun Zhang (2012). "Regularized learning in Banach spaces as an optimization problem: representer theorems". In: *Journal of Global Optimization* 54.2, pp. 235–250.

Figures

Tables

Theorems

Definitions

Nomenclature

As much as possible, and unless otherwise stated, the following conventions are used throughout this thesis.

Conventions

$\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$	Boldface lower case letters denote real vectors
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$	Boldface upper case letters denote real matrices
$\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots$	Calligraphic upper case letters denote sets

Indexing

$\mathbf{A}_{ij}$, A_{ij}, a_{ij}	The (i, j)'th element of the matrix $\mathbf{A}$
$\mathbf{A}_{i\cdot}$	The i'th row of the matrix $\mathbf{A}$ as a tall vector (transposed row vector)
$\mathbf{A}_{\cdot j}$	The j'th column vector of the matrix $\mathbf{A}$

Symbols

$\mathbb{N}$	The set of natural numbers (excluding zero)
$\mathbb{Z}$	The set of integers
$\mathbb{R}$	The set of real numbers
$\mathbb{R}_{>0}$	The set of positive real numbers, $\{x \in \mathbb{R} \vert x > 0\}$
$\mathbb{R}^d$	The d-dimensional Euclidean space
x'	Primes are used to distinguish elements, rather than to denote derivatives
$\hat{\theta}$	Hats are used to denote estimators of a parameter θ
$\mathcal{A}^c$	The complement of a set $\mathcal{A}$
$\mathcal{P}(\mathcal{A})$	The power set of the set $\mathcal{A}$
$\{\}, \emptyset$	The empty set
$\mathbf{0}$	A vector of zeroes
$\mathbf{1}_n$	A length n vector of ones
$\mathbf{I}_n$	The $n \times n$ identity matrix
$\exists$	(short hand) There exists
$\forall$	(short hand) For all
$\lim_{n\to\infty}$	The limit as n tends to infinity
$\xrightarrow{\text{dist.}}$	Convergence in distribution
$O(n)$	Computational complexity (time or storage)
Δx	A quantity representing a change in x

Relations

$a \approx b$	a is approximately or almost equal to b
$a \propto b$	a is equivalent to b up to a constant of proportionality
$a \equiv b$	a is identical to b
$A \Rightarrow B$	The statement B being true is predicated on A being true
$A \Leftrightarrow B$	The statement A is true if and only if B is true
$a \in \mathcal{A}$	a is an element of the set $\mathcal{A}$
$\mathcal{A} \subseteq \mathcal{B}$	$\mathcal{A}$ is a subset of $\mathcal{B}$ which may include itself
$\mathcal{A} \subset \mathcal{B}$	$\mathcal{A}$ is a subset of $\mathcal{B}$ which does not include itself
$a := b$, $a \leftarrow b$	a is assigned the value b
$X \sim p(X)$	The random variable X is distributed according to the pdf $p(X)$
$X \sim D$	The random variable X is distributed according to the pdf specified by the distribution D, e.g. $D \equiv \mathrm{N}(0, 1)$
$X_1,...,X_n \overset{\text{iid}}{\sim} D$	Each random variable X_i, $i = 1, \ldots, n$ is independently and identically distributed according to the pdf specified by the distribution D
$X \vert Y$	The (random) variable X given/conditional on Y

Functions

$\inf \mathcal{A}$	The infimum of a set $\mathcal{A}$
$\sup \mathcal{A}$	The supremum of a set $\mathcal{A}$
$\min \mathcal{A}$	The minimum value of a set $\mathcal{A}$
$\max \mathcal{A}$	The maximum value of a set $\mathcal{A}$
$\arg\min_x f(x)$	The value of x which minimises the function $f(x)$
$\arg\max_x f(x)$	The value of x which maximises the function $f(x)$
$\lvert a \rvert$ with $a \in \mathbb{R}$	The absolute value of a; $\lvert a \rvert = a$ if a is positive, and $-a$ if a is negative, and $\lvert 0 \rvert = 0$
$\delta_{xx'}$	The Kronecker delta; $\delta_{xx'} = 1$ if $x = x'$, and 0 otherwise
$[A]$	The Iverson bracket; $[A] = 1$ if the logical proposition A is true, and 0 otherwise
$\mathbb{1}_{\mathcal{A}}(x)$	The indicator function; $\mathbb{1}_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$, and 0 otherwise
e^x, $\exp(x)$	The natural exponential function
$\log(x)$	The natural logarithmic function
$\frac{\mathrm{d}}{\mathrm{d}x} f(x)$, $\dot{f}(x)$	The derivative of f with respect to x
$f \circ g$	Composition of functions, i.e. g following f

Abstract vector space operations and notations

$\mathcal{V}^{\perp}$	The orthogonal complement of the space $\mathcal{V}$
$\mathcal{V}^{\vee}$	The algebraic dual space of $\mathcal{V}$
$\mathcal{V}^{*}$	The continuous dual space of $\mathcal{V}$
$\overline{\mathcal{V}}$	The closure of the space $\mathcal{V}$
$\mathcal{B}(\mathcal{V})$	The Borel σ-algebra of $\mathcal{V}$
$\mathrm{L}^p(\mathcal{X}, \nu)$	The set of p-integrable functions over the measure space $\mathcal{X}$ with measure ν
$\mathrm{L}(\mathcal{V}; \mathcal{W})$	The set of bounded, linear operators from $\mathcal{V}$ to $\mathcal{W}$
$\dim(\mathcal{V})$	The dimensions of the vector space $\mathcal{V}$

$\langle x, y\rangle_{\mathcal{V}}$	The inner product between x and y in the vector space $\mathcal{V}$
$\|x\|_{\mathcal{V}}$	The norm of x in the vector space $\mathcal{V}$
$D(x, y)$	The distance between x and y
$x \otimes y$	The tensor product of x and y which are elements of a vector space
$\mathcal{F} \otimes \mathcal{G}$	The tensor product space of two vector spaces
$\mathcal{F} \oplus \mathcal{G}$	The direct sum (or tensor sum) of two vector spaces
$\mathrm{d}f(x)$, $\mathrm{d}^2 f(x)$	The first and second Fréchet differentials of f at x
$\partial_v f(x)$, $\partial_v^2 f(x)$	The first and second Gâteaux differentials of f at x in the direction v
$\nabla f(x)$, $\nabla^2 f(x)$	The gradient and Hessian of f at x in the direction v (f is a mapping of a Hilbert space)

Matrix and vector operations

$\mathbf{a}^\top$, $\mathbf{A}^\top$	The transpose of a vector $\mathbf{a}$ or matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	The inverse of a square matrix $\mathbf{A}$
$\|\mathbf{a}\|^2$	The squared 2-norm the vector $\mathbf{a}$, equivalent to $\mathbf{a}^\top\mathbf{a}$
$\|\mathbf{A}\|$	The determinant of a matrix $\mathbf{A}$
$\mathrm{tr}(\mathbf{A})$	The trace of a square matrix $\mathbf{A}$
$\mathrm{diag}(\mathbf{A})$	The diagonal elements of a square matrix $\mathbf{A}$
$\mathrm{rank}(\mathbf{A})$	The rank of a matrix $\mathbf{A}$
$\mathrm{vec}(\mathbf{A})$	The column-wise vectorisation of a matrix $\mathbf{A}$
$\mathbf{a} \otimes \mathbf{b}$	The outer product of two vectors $\mathbf{a}$ and $\mathbf{b}$
$\mathbf{A} \otimes \mathbf{B}$	The Kronecker product of matrix $\mathbf{A}$ with matrix $\mathbf{B}$
$\mathbf{A} \circ \mathbf{B}$	The Hadamard product two matrices $\mathbf{A}$ and $\mathbf{B}$

Statistical functions

$\mathrm{P}(A)$	The probability of event A occurring
$p(X\|\theta)$	The probability density function of X given parameters θ
$L(\theta\|X)$	The log-likelihood of θ given data X, sometimes simply $L(\theta)$
$\mathrm{BF}(M, M')$	Bayes factor for comparing two models M and M'
$\mathcal{I}(\theta)$	The Fisher information for θ
$\mathrm{E}[X]$, $\mathrm{E}\,X$	The expectation[1] of the random element X
$\mathrm{Var}[X]$, $\mathrm{Var}\,X$	The variance[1] of the random element X
$\mathrm{Cov}[X, Y]$	The covariance[1] between two random elements X and Y
$H(p)$	The entropy of the distribution $p(X)$
$\mathrm{D}_{\mathrm{KL}}\big(q(x)\|p(x)\big)$	The Kullback-Leibler divergence from $p(x)$ to $q(x)$, denoted also by $\mathrm{D}_{\mathrm{KL}}(q\|p)$ for short

Statistical distributions

$\mathrm{N}(\mu, \sigma^2)$	Univariate normal distribution with mean μ and variance σ^2
$\mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	d-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\phi(z)$	The standard normal pdf

foot:exp

[1] When there is ambiguity as to which random element the expectation or variance is taken under or what its distribution is, this is explicated by means of subscripting, e.g. $\mathrm{E}_{X\sim\mathrm{N}(0,1)}\,X$ to denote the expectation of a standard normal random variable.

$\Phi(z)$	The standard normal cdf
$\phi(x\|\mu, \sigma^2)$	The pdf of $\mathrm{N}(\mu, \sigma^2)$
$\phi(\mathbf{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	The pdf of $\mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
$\mathrm{MN}_{n,m}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$	Matrix normal distribution with mean $\boldsymbol{\mu}$ and row variances $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ and column variances $\boldsymbol{\Psi} \in \mathbb{R}^{m\times m}$
${}^{\mathrm{t}}\mathrm{N}(\mu, \sigma^2, a, b)$	Truncated univariate normal distribution with mean μ and variance σ^2 restricted to the interval (a, b)
$\mathrm{N}_+(0, 1)$	The half-normal distribution with variance σ^2
$\mathrm{N}_+(0, \sigma^2)$	The folded-normal distribution with variance σ^2
${}^{\mathrm{t}}\mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathcal{A})$	Truncated d-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ restricted to the set $\mathcal{A}$
$\Gamma(s, r)$	Gamma distribution with shape s and rate r parameters
$\Gamma^{-1}(s, \sigma)$	Inverse gamma distribution with shape s and scale σ parameters
χ^2_d	Chi-squared distribution with d degrees of freedom
$\mathrm{Bern}(p)$	Bernoulli distribution with probability of success p
$\mathrm{Cat}(p_1, \ldots, p_m)$	Categorical distribution with m categories, and each category has probability of success p_j

Abbreviations

ANOVA	Analysis of variance
cdf	cumulative distribution function
CRAN	Comprehensive R Archive Network
DAG	directed acyclic graph
EM	expectation-maximisation
fBm	Fractional Brownian motion
GPR	Gaussian process regression
HPM	highest probability model
IIA	independent of irrelevant alternatives
iid	Identical and independently distributed
Lasso	Least absolute shrinkage and selection operator
MCMC	Markov chain Monte Carlo
ML	maximum likelihood
OLS	ordinary least squares
pd/p.d.	positive definite
pdf	probability density function
PIP	posterior inclusion probability
pmf	probability mass function
PMP	posterior model probability
RKHS	Reproducing kernel Hilbert space
RKKS	Reproducing kernel Kreĭn space
RSS	residual sum of squares
SE	Squared exponential (kernel)