PhD Examination

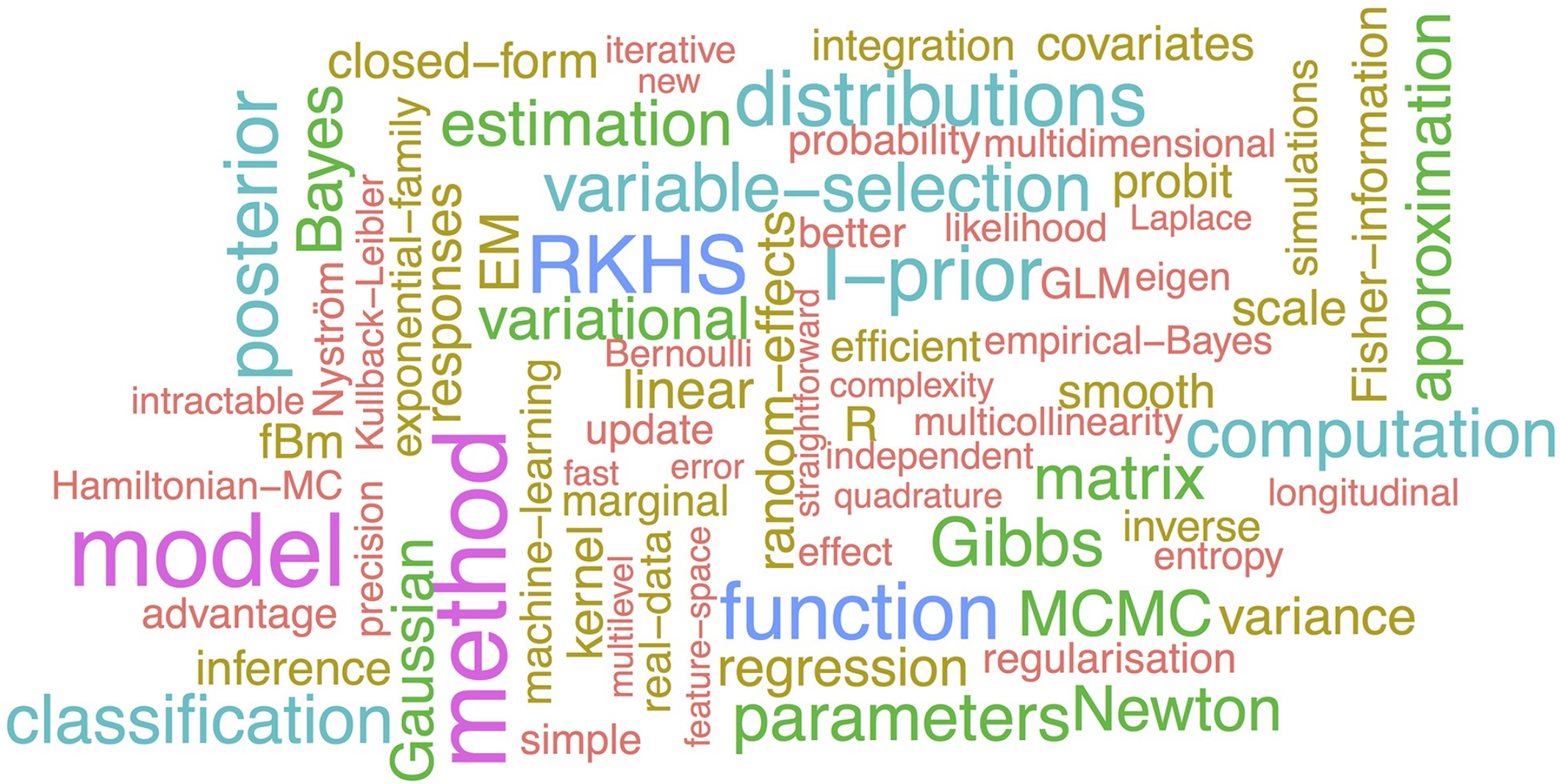
Regression modelling using priors depending on Fisher information covariance kernels (I-priors)

Haziq Jamil

24 September 2018

Executive summary

> Development of a novel methodology—theoretical and computational—for regression, classification and variable selection.



Scope

- What are I-priors?
    - Motivation: The normal regression model [CHAPTER 1]
    - Prerequisite: Functional analysis and RKHS/RKKS theory [CHAPTER 2]
    - Fisher information and the I-prior [CHAPTER 3]

Scope

- What are I-priors?
  - Motivation: The normal regression model [CHAPTER 1]
  - Prerequisite: Functional analysis and RKHS/RKKS theory [CHAPTER 2]
  - Fisher information and the I-prior [CHAPTER 3]
- My PhD work involving I-priors
  - Computational methods for estimation of I-prior models [CHAPTER 4]
  - Extensions to categorical responses [CHAPTER 5]
  - Bayesian variable selection for linear models [CHAPTER 6]

Scope

- What are I-priors?
  - Motivation: The normal regression model [CHAPTER 1]
  - Prerequisite: Functional analysis and RKHS/RKKS theory [CHAPTER 2]
  - Fisher information and the I-prior [CHAPTER 3]
- My PhD work involving I-priors
  - Computational methods for estimation of I-prior models [CHAPTER 4]
  - Extensions to categorical responses [CHAPTER 5]
  - Bayesian variable selection for linear models [CHAPTER 6]
- Supplementary material
  - Estimation concepts
  - EM algorithm and variational inference
  - Hamiltonian Monte Carlo

Chapters 2, 3 and 4 were jointly co-authored with Wicher Bergsma (main supervisor).

Regression modelling using I-priors

The normal regression model

- For $y_i \in \mathrm{R}$, $x_i \in \mathrm{X}$, and $i = 1, \ldots, n$,

$$y_i = f(x_i) + \epsilon_i$$

$$(\epsilon_1, \ldots, \epsilon_n)^\top \sim \mathrm{N}(0, \boldsymbol{\Psi}^{-1})$$

The normal regression model

- For $y_i \in \mathrm{R}$, $x_i \in \mathrm{X}$, and $i = 1, \ldots, n$,

$$y_i = f(x_i) + \epsilon_i$$

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

- Assume $f \in \mathrm{F}$ a reproducing kernel Hilbert or Krein space (RKHS/RKKS).

The normal regression model

- For $y_i \in \mathrm{R}$, $x_i \in \mathrm{X}$, and $i = 1, \ldots, n$,

$$y_i = f(x_i) + \epsilon_i$$

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

- Assume $f \in \mathrm{F}$ a reproducing kernel Hilbert or Krein space (RKHS/RKKS).
- The basis for various regression problems:
  - Linear models (canonical RKHS)
  - Multilevel models (canonical + Pearson = ANOVA RKKS)
  - Longitudinal models (fBm/canonical + Pearson = ANOVA RKKS)
  - Smoothing models (fBm RKHS)
  - etc.

The I-prior

- (Corollary 3.3.1, p. 93) The Fisher information for $f \in \mathrm{F}$, an RKHS/RKKS with kernel $h_\eta : \mathrm{X} \times \mathrm{X} \to \mathrm{R}$, is

$$\mathrm{I}\big(f(x), f(x')\big) = \sum_{i,j=1}^{n} \psi_{ij} h_\eta(x, x_i) h_\eta(x', x_j)$$

The I-prior

- (Corollary 3.3.1, p. 93) The Fisher information for $f \in \mathrm{F}$, an RKHS/RKKS with kernel $h_\eta : \mathrm{X} \times \mathrm{X} \to \mathrm{R}$, is

$$\mathrm{I}\big(f(x), f(x')\big) = \sum_{i,j=1}^{n} \psi_{ij} h_\eta(x, x_i) h_\eta(x', x_j)$$

- Define an I-prior for f to be

$$\mathbf{f} := \big(f(x_1), \ldots, f(x_n)\big) \sim \mathrm{N}_n(\mathbf{f}_0, \mathrm{I}[f])$$

where $\mathbf{f}_0$ is some prior mean, and $\mathrm{I}[f]$ is the $n \times n$ Fisher information matrix for $\mathbf{f}$.

The I-prior

- (Corollary 3.3.1, p. 93) The Fisher information for $f \in \mathrm{F}$, an RKHS/RKKS with kernel $h_\eta : \mathrm{X} \times \mathrm{X} \to \mathrm{R}$, is

$$\mathrm{I}\big(f(x), f(x')\big) = \sum_{i,j=1}^{n} \psi_{ij} h_\eta(x, x_i) h_\eta(x', x_j)$$

- Define an I-prior for f to be

$$\mathbf{f} := \big(f(x_1), \ldots, f(x_n)\big) \sim \mathrm{N}_n(\mathbf{f}_0, \mathrm{I}[f])$$

  where $\mathbf{f}_0$ is some prior mean, and $\mathrm{I}[f]$ is the $n \times n$ Fisher information matrix for $\mathbf{f}$.

- Objective and intuitive
  - Entropy-maximising prior (Theorem 3.6, p. 98).
  - More information about $f \Rightarrow$ less influence on prior mean choice (usually zero), and vice versa.

Merits

- A unifying methodology for regression
    - Choose appropriate RKHS/RKKS depending on problem

Merits

- A unifying methodology for regression
    - Choose appropriate RKHS/RKKS depending on problem
- Parsimoniuos specification
    - Often less number of parameters required to fit compared to the classical way

Merits

- A unifying methodology for regression
    - Choose appropriate RKHS/RKKS depending on problem
- Parsimoniuos specification
    - Often less number of parameters required to fit compared to the classical way
- Prevents overfitting and undersmoothing

Merits

- A unifying methodology for regression
    - Choose appropriate RKHS/RKKS depending on problem
- Parsimoniuos specification
    - Often less number of parameters required to fit compared to the classical way
- Prevents overfitting and undersmoothing
- Better prediction

Merits

- A unifying methodology for regression
    - Choose appropriate RKHS/RKKS depending on problem
- Parsimoniuos specification
    - Often less number of parameters required to fit compared to the classical way
- Prevents overfitting and undersmoothing
- Better prediction
- Straightforward inference
    - Model comparison using marginal likelihood
    - Bayesian post-estimation procedures possible, e.g. credibility intervals and posterior predictive checks

Main contributions

PROBLEM 1: Storage is $O(n^2)$ while estimation is $O(n^3)$ due to matrix inversion in posterior, specifically, assuming $f_0(x) = 0 \, \forall x$,

$$\mathrm{E}(f(x) \,|\, \mathbf{y}) = \mathbf{h}_\eta^\top(x) \cdot \mathbf{\Psi}(\mathbf{H}_\eta \mathbf{\Psi} \mathbf{H}_\eta + \mathbf{\Psi}^{-1})^{-1} \cdot \mathbf{y}$$

$$\mathrm{Var}(f(x) \,|\, \mathbf{y}) = \mathbf{h}_\eta^\top(x) \cdot (\mathbf{H}_\eta \mathbf{\Psi} \mathbf{H}_\eta + \mathbf{\Psi}^{-1})^{-1} \cdot \mathbf{h}_\eta(x)$$

PROBLEM 1: Storage is $O(n^2)$ while estimation is $O(n^3)$ due to matrix inversion in posterior, specifically, assuming $f_0(x) = 0 \, \forall x$,

$$\mathrm{E}(f(x) \,|\, \mathbf{y}) = \mathbf{h}_\eta^\top(x) \cdot \mathbf{\Psi}(\mathbf{H}_\eta \mathbf{\Psi} \mathbf{H}_\eta + \mathbf{\Psi}^{-1})^{-1} \cdot \mathbf{y}$$

$$\mathrm{Var}(f(x) \,|\, \mathbf{y}) = \mathbf{h}_\eta^\top(x) \cdot (\mathbf{H}_\eta \mathbf{\Psi} \mathbf{H}_\eta + \mathbf{\Psi}^{-1})^{-1} \cdot \mathbf{h}_\eta(x)$$

Chapter 4: An efficient estimation procedure

- Nyström approximation of kernel matrix to reduce storage and time requirements to $O(nm)$ and $O(nm^2)$, $m \ll n$.
- Multiple $O(n^3)$ calls in the EM algorithm, so pre-calculate and store for later use (front-loading).
- Exploit normality and exponential families and use ECM algorithm to avoid maximisation in M-step.
- Implemented in R package .
- Practical applications: Multilevel modelling (IGF-I data), longitudinal modelling (cow growth data), and smoothing models (Tecator data).

PROBLEM 2: Violations of modelling assumptions when responses are categorical, i.e. $y_i \in \{1, \ldots, m\}$.

PROBLEM 2: Violations of modelling assumptions when responses are categorical, i.e. $y_i \in \{1, \ldots, m\}$.

Chapter 5: Probit link on (latent) regression functions

- (Section 5.1, pp. 149–151) "Squash" the regression functions through a sigmoid function, via

$$\mathrm{P}(y_i = j) = g^{-1}\big(f_1(x_i), \ldots, f_m(x_i)\big)$$

  where g^{-1} is an integral involving a truncation of an m-variate normal density.

  - When $m = 2$ (binary data), the model simplifies to

$$y_i \sim \mathrm{Bern}(\Phi(f(x_i)))$$

- Model is estimated using a variational EM algorithm, because the E-step cannot be found in closed-form.

- Practical applications: Binary and multiclass classification, meta-analysis (smoking cessation data), and spatio-temporal modelling (BTB data).

PROBLEM 3: Model selection via pairwise marginal likelihood comparisons is intractable for large p.

PROBLEM 3: Model selection via pairwise marginal likelihood comparisons is intractable for large p.

Chapter 6: Gibbs-based variable selection for linear models

- Focusing on iid normal linear models only, i.e.

$$y_i \sim \mathrm{N}\Big(\alpha + \sum_{k=1}^{p} x_{ik}\gamma_k\beta_k,\, \psi^{-1}\Big)\gamma_k \sim \mathrm{Bern}(\pi_k), k = 1, \ldots, p$$

  with the I-prior $\boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{0},\, \kappa\psi^{-1}\mathbf{X}^\top\mathbf{X})$.

- Estimates of posterior model probabilities, inclusion probabilities, and regression coefficients obtained simultaneously using Gibbs sampling.

- Simulation study and real-data analysis show favourable results for I-prior (vs. sparse prior, g-prior, and Lasso).

- Practical applications: aerobic data, mortality and air pollution data, and ozone data.

End