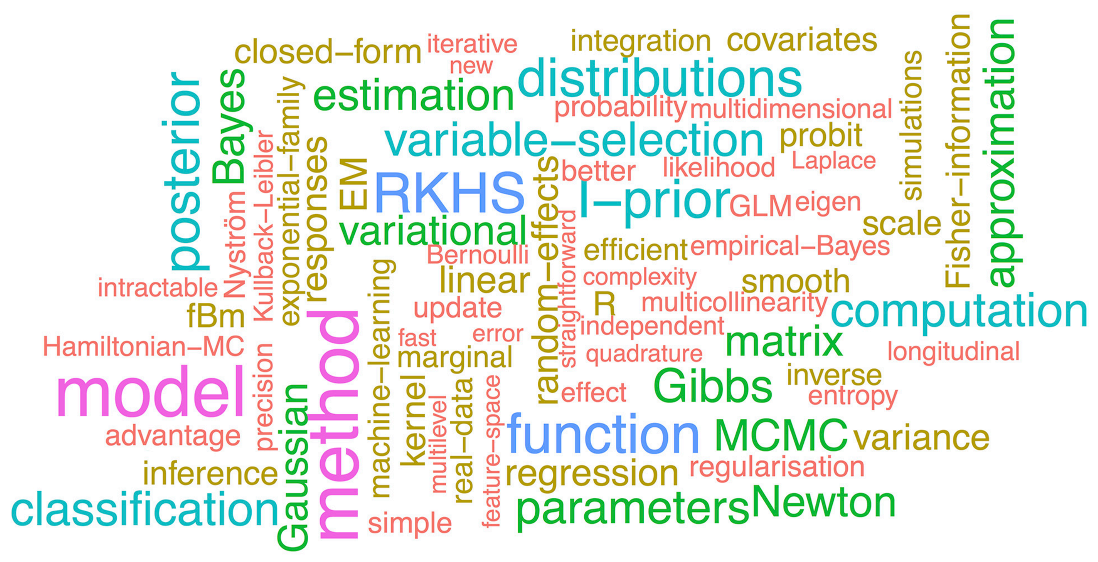
Haziq Jamil &
Wicher Bergsma

Department of Statistics
London School of Economics and
Political Science

`http://phd.haziqj.ml`

Binary and Multinomial Regression using Fisher Information Covariance Kernels (I-priors)



Introduction

Consider the regression model for $i = 1, \dots, n$:

$$\begin{gathered} y_i = \alpha + f(x_i) + \epsilon_i \\ (\epsilon_1, \dots, \epsilon_n)^\top \sim \mathrm{N}_n(0, \Psi^{-1}) \end{gathered} \tag{1}$$

where $y_i \in \mathbb{R}$, $x \in \mathcal{X}$, $f \in \mathcal{F}$ and $\alpha \in \mathbb{R}$ is an intercept. Let $\mathcal{F}$ be a reproducing kernel Hilbert space (RKHS) with kernel $h_\lambda : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The Fisher information for f evaluated at x and x' is

$$\mathcal{I}\big(f(x), f(x')\big) = \sum_{k=1}^n \sum_{l=1}^n \Psi_{k,l} h_\lambda(x, x_k) h_\lambda(x', x_l). \tag{2}$$

The I-prior

The entropy maximising prior distribution for f, subject to identifiability constraints, is

$$\mathbf{f} = \big(f(x_1), \dots, f(x_n)\big)^\top \sim \mathrm{N}_n\left(\mathbf{f}_0, \mathcal{I}[f]\right).$$

Equivalently, $f(x) = f_0(x) + \sum_{i=1}^n h_\lambda(x, x_i) w_i$, with

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

Of interest are

- the posterior distribution for the regression function
$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{y}}; \text{ and}$$
- the posterior predictive distribution for new data points
$$p(y_{\text{new}}|\mathbf{y}) = \int p(y_{\text{new}}|f_{\text{new}}, \mathbf{y}) p(f_{\text{new}}|\mathbf{y})\,\mathrm{d}f_{\text{new}}.$$

Model parameters (error precision Ψ, RKHS scale parameters λ, and any other kernel parameters) may need to be estimated.

A Unified Regression Framework

- Multiple linear regression (canonical RKHS)
- Smoothing models (fBm RKHS)
- Multilevel regression (ANOVA RKHS: canonical & Pearson)
$$f(x_i^{(j)}) = f_1(j) + f_2(x_i^{(j)}) + f_{12}(x_i^{(j)}, j)$$
- Longitudinal modelling (ANOVA RKHS: fBm & Pearson)
$$f(x_i, t_i) = f_1(t_i) + f_2(x_i) + f_{12}(x_i, t_i)$$
- Functional covariates ($\mathcal{X}$ a Hilbert-Sobolev space)

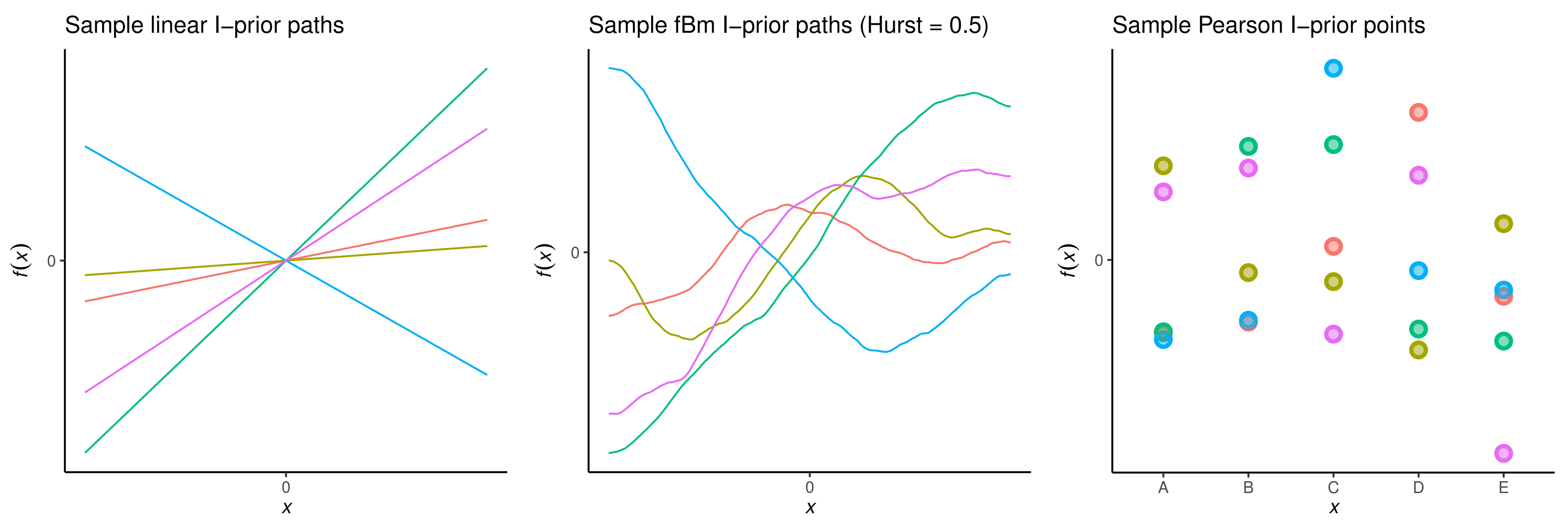


Figure 1: (L-R) Sample paths from the canonical (linear), fractional Brownian motion (fBm), and Pearson RKHS. The (reproducing) kernels corresponding to each RKHS are $h_\lambda(x, x') = \lambda\langle x, x'\rangle_{\mathcal{X}}$ (linear), $h_\lambda(x, x') = -\frac{\lambda}{2}\big(\|x - x'\|_{\mathcal{X}}^{2\gamma} - \|x\|_{\mathcal{X}}^{2\gamma} - \|x'\|_{\mathcal{X}}^{2\gamma}\big)$ (fBm), and $h_\lambda(x, x') = \lambda\big(\delta_{xx'}/\mathrm{P}[X = x] - 1\big)$ (Pearson).

Categorical Responses

When each $y_i \in \{1, \dots, m\}$, normality assumptions are violated. Model instead $y_i = \arg\max_k y_{ik}^*$, where

$$\begin{gathered} y_{ij}^* = \alpha_j + f_j(x_i) + \epsilon_{ij} \\ (\epsilon_{i1}, \dots, \epsilon_{im})^\top \sim \mathrm{N}_m(0, \Sigma) \end{gathered} \tag{3}$$

with $\mathrm{Cov}(\epsilon_{ij}, \epsilon_{kj}) = 0$, for all $i \neq k$, $j = 1, \dots, m$. In other words, $\Psi = I_n$ in (1) and (2). The I-prior is

$$\begin{gathered} \mathbf{f}_j = \big(f_j(x_1), \dots, f_j(x_n)\big)^\top \sim \mathrm{N}_n\left(\mathbf{f}_{0j}, \Sigma_{jj}^{-1} \cdot \mathcal{I}[f]\right) \\ \mathrm{Cov}(\mathbf{f}_j, \mathbf{f}_k) = \Sigma_{jk}^{-1} \cdot \mathcal{I}[f]. \end{gathered}$$

Class probabilities p_{ij} are obtained using a *conically truncated m-variate normal* density

$$p_{ij} = \int_{\{y_{ij}^* > y_{ik}^* \,|\, k \neq j\}} \mathrm{N}_m\left(\mathbf{y}_i^* \,|\, \mathbf{f}(x_i), \Sigma\right) \mathrm{d}\mathbf{y}_i^* =: g_j^{-1}\big(\mathbf{f}(x_i)\big).$$

where we had defined $\mathbf{f}(x_i) = \big(f_1(x_i), \dots, f_m(x_i)\big)^\top$. Now, the marginal, on which the posterior depends,

$$p(\mathbf{y}) = \int \prod_{i,j} \left\{ g_j^{-1}\big(\mathbf{f}(x_i)\big)\right\}^{[y_i = j]} \cdot \mathrm{N}_{nm}\left(\mathbf{f} \,|\, \mathbf{f}_0, \Sigma \otimes \mathcal{I}[f]\right) \mathrm{d}\mathbf{f},$$

cannot be found in closed form. By working in a fully Bayesian setting, we append model parameters and employ a *variational approximation*.

Spatio-Temporal Modelling of BTB[a]

Determine the existence of spatial segregation of the different spoligotypes of bovine tuberculosis (BTB) in Cornwall, UK, and whether the spatial distribution had changed over time.

❶ Constant model (constant RKHS)
$$p_{ij} = g_j^{-1}\big(\alpha_k\big)_{k=1}^m$$
❷ Spatial segregation (fBm RKHS)
$$p_{ij} = g_j^{-1}\big(\alpha_k + f_{1k}(x_i)\big)_{k=1}^m$$
❸ Spatio-temporal segregation (ANOVA RKHS)
$$p_{ij} = g_j^{-1}\big(\alpha_k + f_{1k}(x_i) + f_{2k}(t_i) + f_{12k}(x_i, t_i)\big)_{k=1}^m$$

Evidence Lower Bound (ELBO) values for the three models are -1197.4, -665.3, and -656.2 respectively.

Detecting Cardiac Arrhythmia[b]

Predict whether or not patients suffer from a cardiac disease based on various patient profiles such as age, height, weight and a myriad of electrocardiogram (ECG) readings ($p = 271, n = 451$).

Table 1: Mean out-of-sample misclassification rates and standard errors for 100 runs of various training (s) and test ($451 - s$) sizes for the cardiac arrhythmia binary classification task.

	Misclassification rate (%)		
Method	$s = 50$	$s = 100$	$s = 200$
I-probit (linear)	34.5 (0.4)	31.4 (0.4)	29.7 (0.4)
I-probit (fBm)	34.7 (0.6)	27.3 (0.3)	24.5 (0.3)
GP (Gaussian)	37.3 (0.4)	33.8 (0.4)	29.3 (0.4)
L-1 logistic	34.9 (0.4)	30.5 (0.3)	26.1 (0.3)
SVM (linear)	36.2 (0.5)	35.6 (0.4)	35.2 (0.4)
SVM (Gaussian)	48.4 (0.5)	47.2 (0.5)	46.9 (0.4)
RF	31.7 (0.4)	26.7 (0.3)	22.4 (0.3)
k-NN	40.6 (0.3)	38.9 (0.3)	35.8 (0.4)

Conclusions

- Simple estimation of various categorical models:
  - Choice models (with or without IIA);
  - Random-effects models;
  - Binary and multiclass classification.
- Inference is straightforward (e.g. model comparison or (transformed) credibility intervals).
- Often gives better predictions.

References

[1] Wicher Bergsma.
Regression and classification with I-priors.
arXiv: 1707.00274, July 2017.

[2] Mark Girolami and Simon Rogers.
Variational Bayesian multinomial probit regression with Gaussian process priors.
Neural Computation, 18(8), 2006.

[3] Robert E McCulloch, Nicholas G Polson, and Peter E Rossi.
A Bayesian analysis of the multinomial probit model with fully identified parameters.
Journal of Econometrics, 99(1):173–193, 2000.

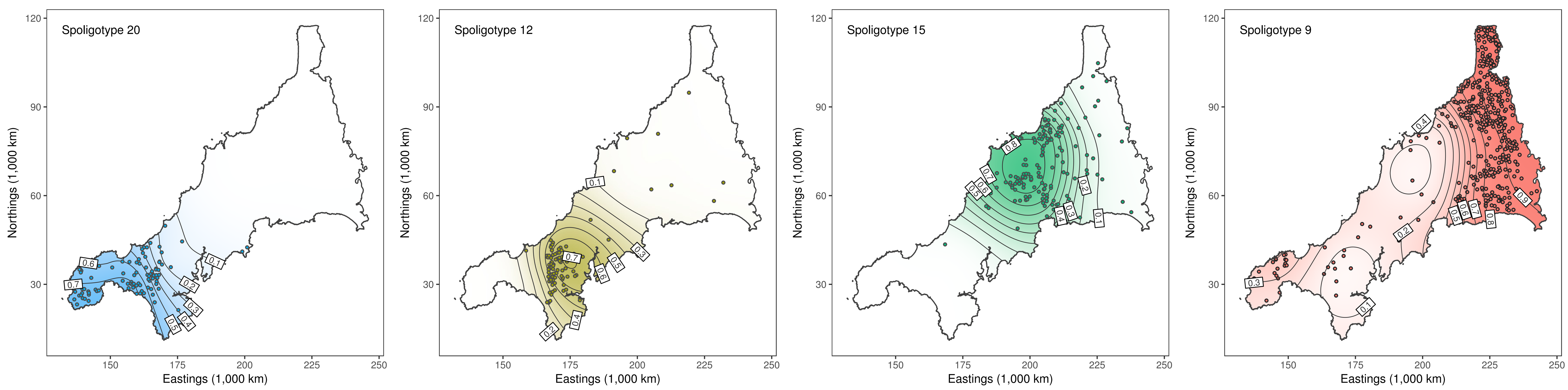


Figure 2: Predicted probability surfaces for BTB contraction in Cornwall for the four largest spoligotypes of the bacterium *Mycobacterium bovis* over the entire time period 1989–2002 using Model 2.

Data sources: [a]Peter Diggle, Pingping Zheng, and Peter Durr. Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK. *J. Royal Stat. Soc. Series C (Appl. Statist.)*, 54(3):645–658, 2005. [b]Timothy I Cannings and Richard J Samworth. Random projection ensemble classification. *J. Royal Stat. Soc. Series B (Stat. Methodol.)*, 79(4):959–1035, 2017.