Università Cattolica del Sacro Cuore · Università Bocconi Milano · Politecnico Milano 1863 · Università degli Studi di Milano Bicocca · Università degli Studi di Milano · mathesia

Smart Statistics for Smart Applications

Book of Short Papers SIS2019

SIS 2019 SMART STATISTICS FOR SMART APPLICATIONS

SIS Società Italiana di Statistica

Editors: Giuseppe Arbia, Stefano Peluso, Alessia Pini and Giulia Rivellini

PUBLISHED BY PEARSON

WWW.PEARSON.COM

Giugno 2019 ISBN 9788891915108

Prior specification in flexible models

Specificazione delle prior in modelli flessibili

Maria Franco-Villoria, Massimo Ventrucci and Haavard Rue

Abstract The linear predictor of generalized additive models is expressed as a sum of unspecified smooth functions. In a Bayesian hierarchical framework, smooth functions can be described by a vector of random effects distributed at prior as a Gaussian Markov Random Field. In this work, we present the use of Penalized Complexity Priors (PC priors) for flexible models, introducing a natural base model.
Abstract *Il predittore lineare nei modelli additivi generalizzati viene espresso come una somma di funzioni smooth (i.e. forma parametrica non specificata). Nel contesto dei modelli gerarchici Bayesiani, le funzioni smooth vengono descritte da un vettore di effetti casuali distribuiti a priori come un Gaussian Markov Random Field. In questo lavoro rivisitiamo i modelli flessibili attraverso l'uso della classe di distribuzioni a priori nota come Penalized Complexity Priors (PC priors).*

Key words: base model, Gaussian Markov random field, penalized complexity, random walk

1 Introduction

In some practical applications imposing a linear relationship between the response and explanatory variable might be too restrictive, and more flexible models, such as generalized additive models (GAM) [6], might be needed. In these flexible models, the linear predictor is specified as a sum of smooth functions of the explanatory vari-

Maria Franco-Villoria
University of Torino e-mail: maria.francovilloria@unito.it

Massimo Ventrucci
University of Bologna, e-mail: massimo.ventrucci@unibo.it

Haavard Rue
King Abdullah University of Science and Technology e-mail: haavard.rue@kaust.edu.sa

ables. We follow a Bayesian hierarchical framework where each smooth function is described by a vector of random effects distributed at prior as a Gaussian Markov Random Field (GMRF) [4]. A GMRF is a multivariate normal distribution with mean vector μ and a sparse precision $Q(\tau)$ that depends on some hyper-parameters τ and whose non zero pattern specifies conditional dependencies among neighbouring random effects.

Elicitation of priors for precision parameters is a long standing topic in the literature on hierarchical Bayesian models. Simpson et al. (2017) [5] recently introduced a new framework for building priors that avoid overfitting denoted as *Penalized Complexity (PC) priors*. PC priors are computed based on specific principles in which a model component is seen as a flexible parametrization of a base model. The idea is to penalize model complexity, defined in terms of distance from the base model, in such a way that the base model is favoured unless the available data support a more flexible one.

2 Penalized Complexity (PC) Priors

In this section we summarize the four main principles underpinning the construction of PC priors, namely: support to Occam's razor (parsimony), penalisation of model complexity, constant rate penalisation and user-defined scaling. For a more detailed presentation of these principles the reader is referred to [5].

Let f_1 denote the density of a model component w where τ is the parameter for which we need to specify a prior. The base model, corresponds to a fixed value of the parameter $\tau = \tau_0$ and is characterized by the density f_0.

1. The prior for τ should give proper shrinkage to τ_0 and decay with increasing complexity of f_1 in support of Occam's razor, ensuring parsimony; i.e. the simplest model is favoured unless there is evidence for a more flexible one.
2. The increased complexity of f_1 with respect to f_0 is measured using the Kullback-Leibler divergence (*KLD*) [2],

$$KLD(f_1||f_0) = \int f_1(w) \log\left(\frac{f_1(w)}{f_0(w)}\right) dw,$$

which, for zero mean multivariate normal densities is

$$KLD(f_1||f_0) = \frac{1}{2}\left(tr(\Sigma_0^{-1}\Sigma_1) - n - \ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right)\right)$$

where n is the dimension. For ease of interpretation, the KLD is transformed to a unidirectional distance measure

$$d(\tau) = d(f_1||f_0) = \sqrt{2KLD(f_1||f_0)} \tag{1}$$

that can be interpreted as the distance from the flexible model f_1 to the base model f_0.

3. The PC prior is defined as an exponential distribution on the distance, $\pi(d(\tau)) = \lambda \exp(-\lambda d(\tau))$, with rate $\lambda > 0$, ensuring constant rate penalization. Therefore, the mode of a PC prior is always at the base model. The PC prior for τ follows by a change of variable transformation.
4. The user must select λ based on his prior knowledge on the parameter of interest (or an interpretable transformation of it $T(\tau)$). This knowledge can be expressed in terms of a probability statement, e.g. $P(T(\tau) > U) = a$, where U is an upper bound for $T(\tau)$ and a is a (generally small) probability.

One major advantage of PC priors is that they prevent overfitting by construction, as they guarantee shrinkage towards the base model. PC priors for the marginal variance of a Gaussian random effect have been shown to outperform other priors widely used in literature (such as Inverse Gamma priors) when data are weakly informative or the size of the effects is close to the base model [1]. Finally, prior information, if available, can be coded into an intuitive way by simply specifying U and a.

3 Flexible models

Consider the simple case where there are n observational units indexed by $i = 1,\ldots,n$ and one covariate x_i whose effect on the response y_i is not assumed to take any parametric shape. Assuming y_i belonging to the exponential family, the linear predictor of a generalized additive model is

$$\eta_i = \alpha + f(x_i) \qquad i = 1,\ldots,n. \tag{2}$$

The smooth function in (2) can be described by a vector of random effects $\beta = (\beta_1,\ldots,\beta_n)^T$, for which random walk priors are a very popular choice. A random walk is a particular intrinsic GMRF or order r, i.e. a process characterized by the following improper multivariate Gaussian distribution:

$$\pi(\beta|\tau) = (2\pi)^{-rank(R)/2}(|\tau R|^*)^{1/2} \exp\left\{-\frac{\tau}{2}\beta^\mathsf{T} R\beta\right\} \tag{3}$$

where $|\tau R|^*$ is the generalized determinant. Density (3) is improper as it is invariant to the addition of a polynomial of degree $r-1$. In the case of a random walk of order 2, a useful parametrization for the smooth function in Model (2) is

$$f(x_i) = \beta_0 x_i + \beta_i, \quad i = 1,\ldots,n \tag{4}$$

where β_i are subject to the linear constraints $\sum_{i=1}^n \beta_i = 0$, $\sum_{i=1}^n x_i\beta_i = 0$ and can be seen as deviations from the linear trend $\beta_0 x_i$. The model turns into a simple linear regression model when the smooth function $f(x_i)$ is linear over x_i, i.e. when $\beta_i = 0$

$\forall i$. The linear model can be regarded as a *base model*, while Model (2) can be seen as a flexible extension of it. The precision parameter τ controls how flexible the corresponding smooth function is, as shown in Fig. 1. The base model can be obtained setting the hyper-parameter $\tau = \infty$.

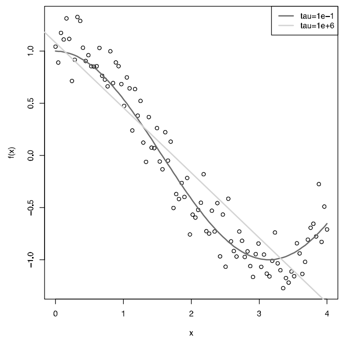

Fig. 1 Effect of the precision parameter on fitted smooth function assuming a RW2 prior.

4 Work in Progress

We consider rewriting Model (2) using an alternative and intuitive parametrization following the work by [3]. This way prior distributions for β_0 and β_i in (4) can be set jointly. The parametrization considered can be easily extended for models involving bivariate smooth functions.

References

1. Klein, N. and Kneib, T. (2016). Scale-Dependent Priors for Variance Parameters in Structured Additive Distributional Regression. *Bayesian Analysis*, 11(4):1071–1106.
2. Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.
3. Riebler, A., Srbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25(4):1145–1165. PMID: 27566770.
4. Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman and Hall/CRC.
5. Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Srbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.*, 32(1):1–28.
6. Wood, S.N (2006). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.