Bayesian estimation of number and position of knots in regression splines

Stima Bayesiana del numero e della posizione dei nodi in spline di regressione

Gioia Di Credico, Francesco Pauli and Nicola Torelli

Abstract Regression splines, based on piecewise polynomials, are useful tools to model departures from linearity in the regression context. The number and location of the knots can be of interest in many contexts since they can detect possible change points in the relationship between the variables. This work is focused on the estimate of both number and location of knots in the simple case where linear truncated splines are chosen to represent the relationship, in this case, the position of the knot detects a change in the slope. In a Bayesian context, we propose a two-step procedure, to first determine the true number of knots and then to fit the final model estimating simultaneously location of knots and regression and spline coefficients.
Sommario *Le spline applicate a modelli di regressione con polinomi a tratti possono essere utili al fine di descrivere relazioni non lineari. In alcuni contesti e immaginando di limitare l'attenzione a splines con componenti lineari, può essere interessante conoscere il numero e la posizione dei nodi che sono quindi i cambi di pendenza della retta di regressione. Tuttavia, stimare numero e posizione dei nodi aggiunge una componente di stima non lineare al problema di ottimizzazione. Proponiamo una procedura in due passi che prima determina il numero ottimale dei nodi e poi ne stima la posizione, congiuntamente agli altri coefficienti del modello. La metodologia viene applicata qui ai modelli lineari adottando un approccio bayesiano.*

Key words: Regression splines, Stan, spike-and-slab priors, knot location

Gioia Di Credico
Department of Statistics, Padua University, Padua, Italy, e-mail: gioia.dicredico@studenti.unipd.it

Francesco Pauli and Nicola Torelli
Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti", University of Trieste, Trieste, Italy e-mail: francesco.pauli@deams.units.it and e-mail: nicola.torelli@deams.units.it

1 Introduction

When modelling the relationship between a response and some (continuous) covariates the linearity assumption turns out to be too restrictive in many contexts. Naive solutions to overcome this limitation such as categorization of the predictor or its polynomial representation have well-known drawbacks.

A viable alternative is represented by spline functions. They are defined as piecewise polynomials with a fixed degree whose joint points are called knots. Splines are highly flexible, in fact, varying the number and position of knots may lead to extremely different shapes and a major risk is to overfit the data. A classical approach consists in using an optimizing criterion with a suitable penalization to control the roughness of the function. Other techniques proposed in the literature include the use of variable selection to choose basis function [6], or employing samplers that allow for varying dimension of the parameter [2, 3].

Assuming that the number and position of knots may have an important and substantial interpretation, here we consider their estimation following one of the most recent approaches to variable selection in a Bayesian context. Estimating the positions of the knots is not an easy task and, for a fixed degree, regression coefficients and locations of knots have to be estimated simultaneously, turning the standard estimation procedure into a nonlinear optimization problem. In the sequel, we propose a method to estimate the number and position of knots with a two-step procedure.

2 Methods

Consider the model

$$y_i = z_i^\intercal \alpha + f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

where z is the covariates vector that enters linearly in the model, α is the vector of regression coefficients, x is a continuous variable evaluated through a smooth function $f : \mathbb{R} \to \mathbb{R}$, described with a spline with few knots and ε is an i.i.d. Gaussian random error component.

We restrict our analysis to those situations in which a low number of knots can be adequate and their positions are directly interpretable and of specific interest for the analysis. This is the case, for example, when truncated power basis (TPB) of order one is used since in this case positions of knots represent changing points for the slope. One of the main drawbacks of truncated power basis representation is that the basis is not orthogonal, which can lead to numerical instability and slow convergence of the optimization algorithm. Keeping a low number of knots alleviates the issue [5].

Let then

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} \gamma_k (x - \xi_k)_+, \tag{1}$$

where ξ_k is the position of the k-th knot and K is the total number of knots, and

$$(x-\xi)_+ = \begin{cases} x-\xi, & \text{if } x \geq \xi \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

is the truncated linear function. Given the number of knots, parameters estimation reduces to maximum likelihood estimate. A usual approach is to choose the knot locations using standard criteria (such as quantiles of the predictor distribution, uniformly distributed knots on the range of the independent variables and user-defined knots following a priori information [5]), estimate models with a different number and location of knots and compare them through standard criteria, such as AIC, BIC or GCV. This procedure often results in a not clear discrimination among all competing models.

In order to enhance the fit of the model, a possible extension is to consider locations of knots as parameters to be estimated along with other regression coefficients. In such a case, within a maximum likelihood approach, exploration of the objective function surface could locate local maxima leading to apparent solutions strongly dependent on starting values. A Bayesian specification of the model and exploring of the posterior distribution, possibly by using MCMC simulations, could prove in this case much more effective.

Our aim is to estimate both number and location of knots, thus a preliminary idea is to estimate several models with free knot locations and with increasing but fixed number of knots. Knots are constrained to be ordered and their prior distributions are Uniform on the range of the variable. Vague priors on the regression and spline coefficients are chosen (zero-centered normal with large variance). We will refer to this model as the no variable selection (NVS) model.

Models with an increasing number of knots were compared on the basis of diagnostic tools such as traceplots and $\hat{R}$ to check convergence of parameters and information criteria were used to choose the best model among the estimated ones. The main drawback of this procedure is a large number of models that one needs to consider and the implied computational effort in high dimensional problems. However, results obtained on simulated data on the basis of these diagnostic tools show that it can lead to reasonable estimates and that convergence of chains for knot related parameters is univocal only if the number of specified knots is lower or equal to the true one.

This prompted us to consider a two-step procedure:

- select the optimal number of knots considering a large, possibly, overparameterized model,
- fit the final model by simultaneously estimating locations of knots and regression and spline coefficients.

In the first step, we estimate a model having more knots than reasonably warranted. This leads to an overparameterized model where the posterior of some knot locations are expected to concentrate at the limits of the predictor range. To assess convergence of the spline parameters, our advice is to run several chains and look at

the results of each chain separately. Indeed, overparameterizing the model may lead to chains which converge at different points. Since each knot location is uniquely linked to a spline coefficient, we evaluate the presence of a knot based on the analysis of the associated coefficient posterior distribution.

The concept underlying the proposed methodology is to perform variable selection on the basis functions, for this purpose we employ one of the most common approaches in Bayesian literature: that based on the definition of spike-and-slab priors. Several versions have been proposed in the literature [4] but, generally speaking, prior distributions for the regression coefficients are defined with a spike component, usually highly concentrated around zero, and a diffused slab part. This is the case of the stochastic search variable selection approach (SSVS), that defines a mixture distribution for each parameter that has to be selected [4]. This type of methodology gives us the opportunity to evaluate the presence of a variable through the marginal posterior distribution of the mixing proportion. Starting from the NVS model specification, we set a prior distribution on each spline parameter γ_k such that

$$\pi(\gamma_k|\lambda_k) = \lambda_k N(0,\sigma_{sl}) + (1-\lambda_k)N(0,\sigma_{sp}),$$

where the mixing proportion $\lambda_k \sim Beta(a,b)$, with $a = b$. Standard deviations of the two mixture components, σ_{sl} and σ_{sp}, are chosen to be respectively large and small. Appropriate values have to be evaluated taking into account the unit of measurement of dependent and independent variables.

Our method adapts the SSVS approach by assuming λ_k to be dependent on the knot location ξ_k. The prior distributions of the ordered knots remain defined as Uniform on the support of the variable X and independent from both the mixing proportion λ and the coefficient γ. Each coefficient γ_k, conditioned on the mixing parameter λ_k follows the same mixture distribution of two components specified in the SSVS approach described above, while each element of the mixing proportion vector λ is now defined as:

$$\lambda_k|\xi_k \sim Beta(a,b_k),$$

where a is a positive but very small value and $b_k : [\min(X);\max(X)] \to [a;1+a]$ is a U-shaped even function of the knot location which returns values close to $1+a$ when the knot is near the boundaries of the variable, while it is almost uniform and close to a elsewhere. In practice, the prior for the mixing parameter swings between a beta U-shaped distribution when the knot location is on plausible values and a beta distribution highly concentrated on zero when the knot is close to the boundaries. All the other prior distributions remain defined as in the previous model specification.

In the next section, we compare results from (i) the proposed method, named later on SSVSξ, with (ii) the ones obtained from the SSVS approach and (iii) the same model without a variable selection procedure, NVS.

3 Preliminary results

We simulate data from the linear regression model

$$y_i = 6 + 2x_i - 5(x_i - 2.7)_+ + 8(x_i - 4.3)_+ + \varepsilon_i, \quad i = 1, \ldots, 500,$$

where $\varepsilon_i \sim$ i.i.d. $N(0,3)$ and the predictor X is defined on the interval $[0;10]$. Two knots are placed respectively in 2.67 and 4.33. We set the parameter a of the mixing proportions λ for the SSVS and the SSVSξ models equal to 0.5. Moreover, we chose σ_{sl} equal to 100 and σ_{sp} equal to 0.1. Standard deviations of the prior distributions on spline coefficients and intercept were chosen equal to 100.

We run 10 chains with 2000 iterations each. Posterior inference is based on the last 1000 draws of each chain. To support the complete exploration of the posterior distribution, initial values for the location of the knots are chosen widely spread on the range of the predictor variable X. Spline coefficients and intercept are initialized at zero. The three models are fitted with a different number of knots (respectively with 2, 5 and 10 knots). The interest lies in the parameter estimates, both spline coefficients and knot locations, and in the analysis of the chains behavior.

The number of knots can be chosen in the SSVS and SSVSξ models looking at the plots in Fig 1. The x-axis represents the specified number of knots in the overparameterized models, while the y-axis represents the posterior mean of the mixing proportion. Vectors of posterior means are sorted in descending order and each line corresponds to one chain. In both models performing variable selection, the selected number of relevant knots is always equal to 2, even if the SSVSξ approach makes a slightly clearer distinction with respect to the classic SSVS method.

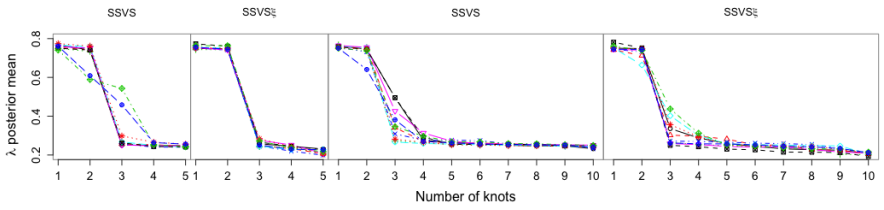

Fig. 1 Posterior means of the mixing parameters λ in the overparameterized SSVS and SSVSξ models with 5 and 10 knots.

The second step of the procedure is to estimate the models with the selected number of knots.

The three models are compared by means of diagnostic tools, such as traceplots, $\hat{R}$, effective sample size (n_{eff}) and analysis of marginal posterior distributions. Due to space constraints, in table 1 we report estimates only for the SSVSξ model, parameter estimates are close to the true parameter values. The greatest discrepancies among the model results are on the order of one decimal point. For the three models, $\hat{R}$ statistics equal to 1 suggest that the chains show good mixing, but differences in

the n_{eff} estimates highlight a lower estimate stability of SSVS model compared to the other two fitted models.

Table 1 Posterior distributions of the SSVSξ model parameters. The model is estimated with the true number of knots. $\hat{R}$ and n_{eff} statistics.

Parameter	true	mean	sd	2.5%	50%	97.5%	Rhat	$n_{eff}^{SSVS\xi}$	n_{eff}^{SSVS}	n_{eff}^{NVS}
β_0	6	6.6	0.6	5.4	6.6	7.6	1.0	4644	762	3313
β_1	2	2.1	0.4	1.3	2.1	3.1	1.0	3039	667	2171
γ_1	-5	-4.5	0.7	-5.8	-4.5	-3.3	1.0	2290	287	3468
γ_2	8	7.4	0.6	6.3	7.3	8.5	1.0	2704	410	2690
ξ_1	2.7	2.4	0.2	1.9	2.4	2.8	1.0	3183	3105	2066
ξ_2	4.3	4.4	0.1	4.2	4.4	4.5	1.0	4634	597	5196
λ_1		0.7	0.3	0.1	0.8	1.0	1.0	9387	3717	
λ_2		0.7	0.3	0.1	0.8	1.0	1.0	8299	7027	

According to this limited evidence, SSVSξ approach should be chosen to perform the proposed procedure to estimate the number and location of the knots. Among the three tested models, SSVSξ gives us the best results in terms of estimation of the parameters and in terms of convergence of the algorithm.

Future developments involve (i) fitting more complex models considering also higher degree splines or multidimensional spline representation and (ii) comparing this procedure with alternative Bayesian approaches proposed in the literature (such as those mentioned in the Sec.1.

References

1. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language, J. Stat. Softw., **76**, 1–32 (2017)
2. Denison, D. G. T., Mallick, B. K., Smith, A. F. M.: Automatic Bayesian curve fitting, J. R. Stat. Soc. Ser. B, **60**, 333–350, (1998)
3. DiMatteo, I., Genovese, C. R., Kass, R. E. : Bayesian curvefitting with freeknot splines, Biometrika, **88**, 1055–1071 (2001)
4. O'Hara, R.B., Sillanp, M. J.: A review of Bayesian variable selection methods: what, how and which, Bayesian anal., **4**, 85–117 (2009)
5. Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric Regression, Cambridge Series in Statistical and Probabilistic Mathematics, Camb. Univ. Press (2003) doi: 10.1017/CBO9780511755453
6. Smith, M., Kohn, R.: Nonparametric regression using Bayesian variable selection, J. Econom., **75**, 317–343 (1996)