A Bayesian approach to estimate the number and position of knots for linear regression splines

Gioia Di Credico, Francesco Pauli and Nicola Torelli

UNIVERSITÀ DEGLI STUDI DI TRIESTE

Department of Economics, Business,
Mathematics and Statistics "Bruno de Finetti"

November 22, 2019

Framework

Assumptions

- the relationship between a response variable and some continuous covariates might be piecewice linear
- we are interested in the estimate of the number and position of the points of departure from linearity

Linear model :

$$y = z^{\mathsf{T}}\alpha + f(x) + \epsilon$$

where $f(x)$ is a regression spline

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

- ▷ $(x - \xi_k)_+ = \max(0, x - \xi_k)$
- ▷ ξ_k position of the k^{th} knot
- ▷ K total number of knots

!! Truncated linear basis : knot locations represent changing points for the slope

→ low number of knots, basis is not orthogonal

Knots : number and location

- fix number and location of knots
- **fix the number of knots and estimate knot locations**
- estimate both number and location of knots

In the first two settings it is possible to compare models throught information criteria *or* using variable selection techniques.

In the third setting, transdimensional techniques (RJMCMC) have to be applied.

Knots : number and location

- fix number and location of knots
- **fix the number of knots and estimate knot locations**
- estimate both number and location of knots

In the first two settings it is possible to compare models throught information criteria *or* using variable selection techniques.

In the third setting, transdimensional techniques (RJMCMC) have to be applied.

Free knots : knots location estimated with the regression coefficients

! The knots **estimation problem** is a non-linear optimization problem.

Bayesian approach :

Computational and methodological flexibility

Constraints on the free-knots locations may be expressed through an appropriate definition of the prior distribution

Knots : number and location

NVS : Estimate several models with free knot locations and with increasing but fixed number of knots and compare them through information criteria.

Prior distributions and constraints :

- ▷ $\boldsymbol{\alpha}$,$\boldsymbol{\beta}$, γ weakly informative prior distribution
- ▷ $\xi \sim Uniform(min(X), max(X))$, subject to $\xi_k \leq \xi_{k+1}$, for $k = 1, \ldots, K$

Note that each knot location is uniquely linked to a spline coefficient
$\Rightarrow$ the presence of a knot can be evaluated on the analysis of the associated coefficient posterior distribution.

Perform variable selection on the basis functions.

A two-step methodology

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

Note that in the overparameterized model the posterior of some knot locations concentrate at the limits of the predictor range.

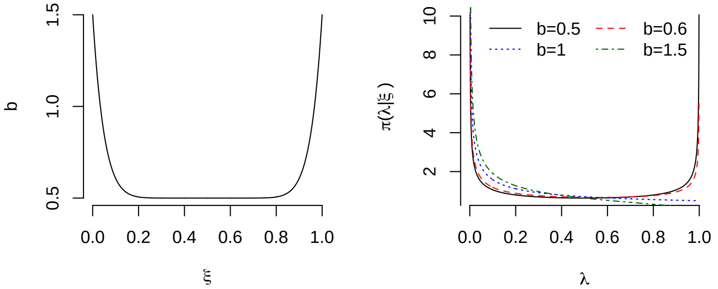
Stochastic search variable selection (SSVS$_\xi$)

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

and the mixing proportion

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

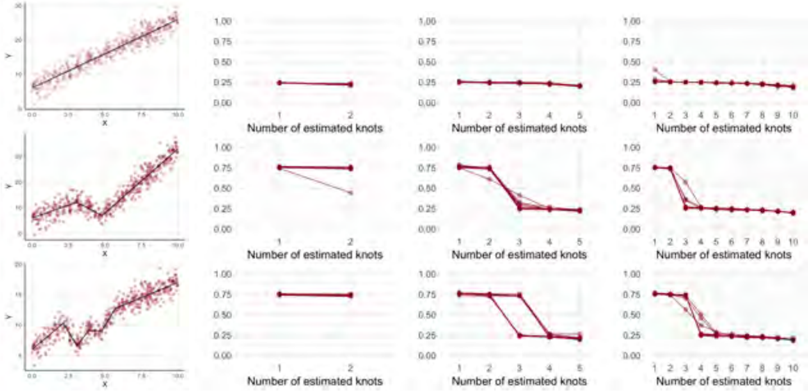
where $a = 0.5$ and $b_k : [\min(X); \max(X)] \rightarrow [a; 1 + a]$ is a U-shaped even function of the knot location.



From a horseshoe shaped distribution to concentrate on values close to zero

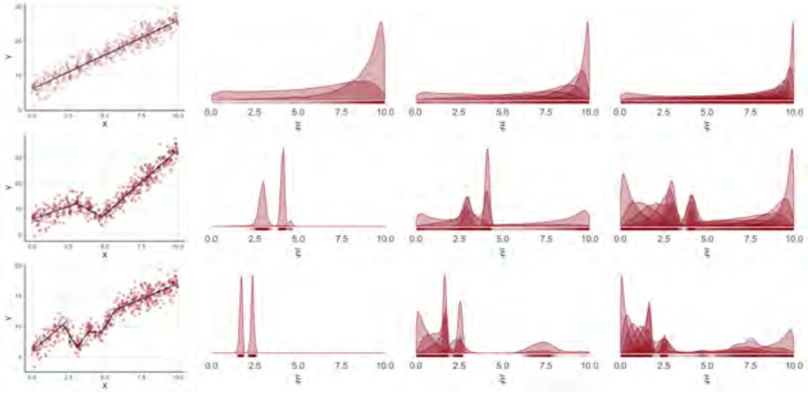
Mixing parameter posterior distributions

To test if the method is able to estimate the correct number of knots even if they are many and close together

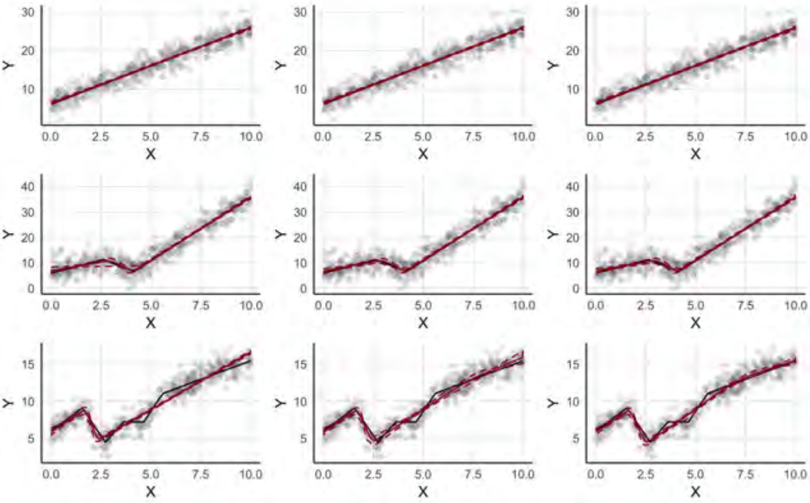


If a high number of knots is expected, this methodology may be not appropriate. . .

Knot locations posterior distributions

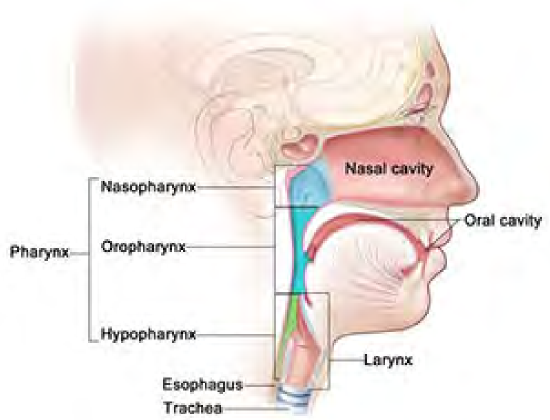


. . . but the knots corresponding to the most evident slope changes are correctly identified


Y
30
20
10
0.0
2.5
5.0
7.5
10.0
X
Y
40
30
20
10
0
0.0
2.5
5.0
7.5
10.0
X
Y
15
10
5
0.0
2.5
5.0
7.5
10.0
X

Head & Neck cancer - INHANCE consortium

Model the association between the risk factors and the outcome, adjusting for possible confounders



Current smokers - larynx : 24.642 subjects from 27 case-control studies collected worldwide
Exposures : intensity and duration of cigarettes consumption
Confounders : age, sex, race, education, study, drinking habits

Semiparametric logistic model and TLB expansion

$$Pr(Y=1|Z) = P(Z)$$
$$\text{logit}(P(Z)) = \log\left[\frac{P(Z)}{1-P(Z)}\right] = Z\alpha + f(x) \otimes f(w)$$

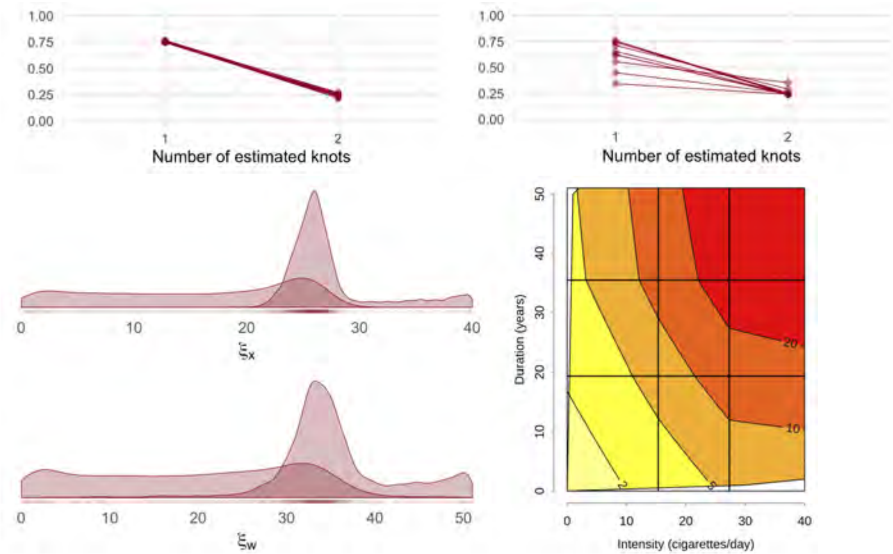
where

- ▷ $Y \sim Bernoulli(P(Z))$
- ▷ logit : $(0,1) \rightarrow \mathbb{R}$ canonical link function
- ▷ $Z = (Z_1, \dots, Z_{p-m})$
- ▷ $X = Z_{p-m+1}, \quad m = 1, 2$
- ▷ $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ an arbitrary smooth function
  → representing non-linear associations between continuous predictors and the log-odds of the binary outcome
  → spline functions

! Number of parameters : $4 + 2(K_x + K_w) + K_x K_w$

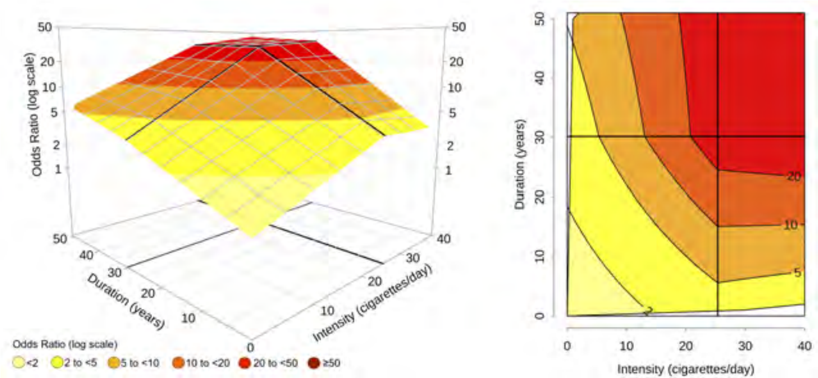
Meaningful knots that highlight cut-points in the risk pattern with biological interpretation

Current smokers - larynx



Current smokers - larynx

Parameter	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
Intensity	1	3,897	25.4	1.4	22.3	25.5	27.8
Duration	1	3,693	30.2	3.3	23.9	30.5	35.8



Iso pack-year points : OR $\sim$ 6 for 40 cigarettes/day and 10 years of duration, but 9 < OR < 10 for 10 cigarettes/day and 40 years of duration

- ✓ A well-known variable selection technique has been adapted in order to estimate the presence or absence of knots in possible overparameterised models. Once that the number of knots is selected, the appropriate model can be fitted with the preferred technique
- ✓ The method gives us a first guess on the knot locations → useful in the initialisation step of algorithms with difficulties in exploring entirely the parameter space
- ✓ SSVS$_\xi$ requires a higher number of parameters to be estimated if compared with one model as specified in the NVS, but only one model needs to be fitted to select the number on knots

- more complex models considering also higher degree splines
- comparing this procedure with alternative Bayesian approaches proposed in the literature

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)

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)

DiMatteo, I., Genovese, C. R., Kass, R. E. : Bayesian curve-fitting with free-knot splines, Biometrika, **88**, 1055–1071 (2001)

O'Hara, R.B., Sillanpää, M. J. : A review of Bayesian variable selection methods : what, how and which, Bayesian anal., **4**, 85–117 (2009)

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

Smith, M., Kohn, R. : Nonparametric regression using Bayesian variable selection, J. Econom., **75**, 317–343 (1996)

Di Credico, G., Edefonti, V., Polesel, J., Pauli, F., Torelli, N. et al. : Joint effects of intensity and duration of cigarette smoking on the risk of head and neck cancer : A bivariate spline model approach, Oral Oncology, **94**, 47–57 (2019)

Thank you for your attention !