PRIOR SELECTION BY AN ADAPTIVE SMOOTHING SPLINES APPROACH

Ronaldo Dias

Universidade Estadual de Campinas *

Abstract

Under the context of empirical bayes a prior density estimate is obtained by using B-splines. In this approach, there are two smoothing parameters, the number of basis functions and usual regularization parameter found in the context of penalized least squares problem. An algorithm is provided to get estimates of the two parameter as well as the prior density.

1 Prior selection

Usually, a lot of attention is paid on a particular parametric classes of priors using the data to assist in estimation of the prior hyperparameters or various functions thereof. The nonparametric empirical Bayes approach supposes that a large amount of data is available to estimate the prior, and hence places no (or minimal) restrictions on the form of the prior. One can try direct estimation of the prior by the following scheme.

Let $\Pi(\theta)$ be prior distribution and let $f(\cdot|\theta)$ be a likelihood density of the data $X_1, \ldots, X_n$ given θ. Suppose we do not know the form of the prior distribution Π and

*Postal address: Departamento de Estatística, IMECC, Cidade Universitária "Zeferino Vaz", Caixa Postal 6065, 13.081-970 - Campinas, SP - BRAZIL

there is considerable information available in data such that it is possible to obtain m the marginal density of the data. Then, we can estimate m by nonparametric methods of curve estimation such as, Logspline ((Kooperberg and Stone, 1991)), H-spline ((Dias, 1997)), Smoothing spline density estimation ((Gu, 1993)), etc. Thus, making use of the integral relationship

$$m^{\Pi} = \int_{\Theta} f(\cdot|\theta) d\Pi(\theta) \tag{1.1}$$

we can obtain a smooth estimate of the prior Π solving the inverse problem given in (1.1) substituting m by the nonparametric estimate $\hat{m}_n$. The difficult encountered in this approach is that the problem (1.1) may not have solution. For this, one may look for a distribution Π for which

$$||m^{\Pi} - \int_{\Theta} f(\cdot|\theta) d\Pi(\theta)|| \tag{1.2}$$

is minimum (in some sense) for a reasonable measure of distance denoted by $||\cdot||$. For Kullback-Leibler divergent, see Berger (1985).

Another approach can be provided by a similar method considering $\hat{m}_n$ a nonparametric density estimate based on data $X_1, \ldots, X_n$, then we search for the minimizer Π of

$$||\hat{m}_n(x) - \int_{\Theta} f(x|\theta) d\Pi(\theta)|| \tag{1.3}$$

Unfortunately, using the minimum distance criteria we will not have, in general, a solution which itself is a probability distribution.

Let's assume that Π is absolutely continuous with respect to a Lesbegue measure μ so that $\frac{d\Pi}{d\mu} = \pi$ and suppose that π is a smooth function in a sense that $\pi \in W_2^2 = \{\pi : \pi' \text{ is abs. continuous and} \int (\ddot{\pi})^2 < \infty\}$. Then, one idea is combine a measure of closeness and smoothness of the prior density. For this, consider penalized least squares equation for a fixed real number $\lambda > 0$ large,

$$A(\pi, \lambda) = ||\hat{m}_n - L^{\pi} f_{\theta}||_2^2 + \frac{\lambda}{2} J(\pi) \tag{1.4}$$

where $L^{\pi} f_{\theta} = \int f(x,\theta)\pi(\theta)d\theta$, $J(\pi)$ is the penalty term generally taken to be $J(\pi) = \int(\ddot{\pi})^2$ and λ is the smoothing parameter that controls the trade-off between smoothness and the $L_2 - distance||\cdot||_2^2$. The goal is to find π that minimizes (1.4) subject to the non linear constraints, $\int \pi = 1$ and $\pi > 0$, since π is a probability density function.

2 Computing the solution

The aim of this section is to show how to approximate π by a finite dimensional subspace of $\mathcal{W}$ spanned by the well known basis functions B-splines (de Boor, 1978). The choice of B-splines is based on computational reasons and their local properties that introduce banded matrices and consequently the computational scheme becomes faster and reliable. Assume $\pi \approx \pi_K = \sum_{j=1}^{K} c_j B_j$, such that B_j are the normalized B-splines, that is, $\int B_j = 1$ and c_j are parameters to be estimated subject to the constraint $\sum_{j=1}^{K} c_j = 1$, since for fixed K, $1 = \int \pi \approx \int \pi_K = \sum_{j=1}^{K} c_j \int B_j = \sum_{j=1}^{K} c_j$. Hence (1.4) becomes,

$$A(\pi,\lambda) = ||\hat{m}_n - \sum_{j=1}^{K} c_j L_j f(\cdot|\theta)||_2^2 + \lambda c^t \Omega c \tag{2.1}$$

where Ω is a n $\times$ K matrix with entries $\Omega_{l,j} = \int \ddot{B}_l \ddot{B}_j$, $c = (c_1, \ldots, c_K)^t$ and the linear operator $L_j f_{\theta} = \int f(x|\theta)\pi(\theta)d\theta$. Let $\{t_i\}_{i=1}^{N}$ be a partition mesh where $f(t_i|\theta)$ is going to be evaluated, then the discretization of (2.1) is given by,

$$A_{\lambda}^{*}(c) = \frac{1}{n}\sum_{i=1}^{N}(\hat{m}_n(t_i) - \sum_{j=1}^{K} c_j L_{i,j})^2 + \lambda c^t \Omega c \tag{2.2}$$

with $L_{i,j} = \int f(t_i,\theta)B_j(\theta)d\theta$. It's not difficult to show that $A_{\lambda}^{*}(c)$ is a convex function on c. Moreover, the functions B_j are computed in the following way. Let $\xi = (\xi_1, \ldots, \xi_K)$ be a sequence of equally spaced knots such that $\xi_1 < \xi_2 < \ldots \xi_K$. Then, $B_j^1(\theta) = (\xi_{j+1} - \xi_j)^{-1}$, if $\xi_j \le \theta < \xi_{j+1}$ and zero otherwise. For $m > 1$, we have

$$B_j^m(\theta) = \frac{m[(\theta - \xi_j)B_j^{m-1}(\theta) + (\xi_{j+m} - \theta)B_{j+1}^{m-1}(\theta)]}{(m-1)(\xi_{j+m} - x_j)}$$

The fact that $B_j(\theta) > 0$ only when $\xi_j \leq \theta < \xi_{j+1}$ and zero otherwise is a very important property from the standpoint of controlling rounding errors in the computations. Since $B_j > 0$ and $\int B_j = 1$, any random variable Ψ with probability density function B_j has moments, $E[\Psi|B_j^m] = (\xi_j + \ldots + \xi_{j+m})/(m+1)$ and $Var[\Psi|B_j] = \sum_{i=l+1}^{j+m} \sum_{l=j}^{j+m} (\xi_i - \xi_l)^2/[(m+1)^2(m+2)]$. When $m = 4$, we denote $B_j^m = B_j$ for simplicity. Hence, for a fixed λ, the vector c_λ is the minimizer of the optimization problem (2.2) with the constraint $\sum_{j=1}^{K} c_j = 1$. Note that given K the prior density estimate depends on λ since $\pi_{K\lambda} = \sum_{j=1}^{K} c_{j,\lambda} B_j$. In order to estimate λ, consider the following distance, $d(\pi, \pi_{K\lambda}) = \int (\pi - \pi_{K\lambda})^2$. Thus,

$$\begin{aligned} d(\pi, \pi_{K\lambda}) - \int \pi^2 &= \int \pi_{K\lambda}^2 - 2\int \pi \pi_{K\lambda} \\ &= \int \pi_{K\lambda}^2 - 2E_\pi[\pi_{K\lambda}(\Theta)]. \end{aligned} \tag{2.3}$$

Define a cross-validation function of λ as,

$$CV(\lambda) = \int \pi_{K\lambda}^2 - \frac{2}{n}\sum_{i=1}^{p} \pi_{K\lambda}(\theta_i) \tag{2.4}$$

Since θ_i are not observable (Bennett and Martz, 1972) suggest to replacing θ_i by its corresponding estimate $\hat{\theta}_i$. Thus, instead to optimize (2.4) one may want to use the following criteria,

$$\widehat{CV}(\lambda) = \int \pi_{K\lambda}^2 - \frac{2}{n}\sum_{i=1}^{p} \pi_{K\lambda}(\hat{\theta}_i).$$

Note that the number of basis functions K is also a smoothing parameter since when K increases less smooth is the estimate $\pi_{K\lambda}$. Hence it is necessary to develop a criteria to find an optimal K. For this, let us define the following metric. Given any two density functions, the square of the Hellinger distance is define to be

$$H^2(f, g) = \int (\sqrt{f} - \sqrt{g})^2 = 2(1 - \rho(f, g)),$$

where

$$\rho(f, g) = \int \sqrt{fg} = \int \sqrt{\frac{f^2 g^2}{\int f^2 \int g^2}} = \int \frac{|fg|}{\sqrt{\int f^2 \int g^2}}.$$

It is not difficult to see that $0 \leq \rho(f,g) \leq 1$, $\forall f,g$. Note that $H^2(f,g)$ is minimum when $\rho(f,g) = 1$, i.e., $(\int f^2 g^2)^{1/2} = \int |fg|$ only if $\alpha f + g = 0$ for some α.

Let $\hat{\lambda}(K) = \hat{\lambda} = \arg\min_{\lambda>0} \widehat{CV}(\lambda)$ and take $g(\cdot) = \pi_{K,\hat{\lambda}}(\cdot)$, $f(\cdot) = \pi_{K+1,\hat{\lambda}}(\cdot)$. By increasing the number of basis functions K by one, the procedure will stop when $\pi_{K,\hat{\lambda}} \approx \pi_{K+1,\hat{\lambda}}$ in the sense of,

$$\frac{\int |\pi_{K,\hat{\lambda}} \pi_{K+1,\hat{\lambda}}|}{\sqrt{\int \pi^2_{K,\hat{\lambda}} \pi^2_{K+1,\hat{\lambda}}}} \approx 1$$

One may notice that both K and λ controls the trade-off between smoothness and fidelity to the data. The idea now is to estimate K and λ jointly. The algorithm below is an attempt to implement this idea.

Algorithm

(1) Let K_0 the initial number of basis functions and fix λ_0.

(2) Compute π_{K_0,λ_0} the minimizer of (2.2)

(3) Find $\hat{\lambda}$ which minimizes the cross validation criteria, $\widehat{CV}(\lambda)$.

(4) Compute $\pi_{K_0,\hat{\lambda}}$

(5) Increment the number of basis functions by one and repeat steps (2) to (4) in order to get $\pi_{K_0+1,\hat{\lambda}}$.

(6) Given a number $\delta > 0$, the procedure will stop if $\rho(\pi_{K,\hat{\lambda}}, \pi_{K+1,\hat{\lambda}}) < 1 - \delta$ or when K reaches n (sample size). The number δ is determined empirically.

Note that each time that the number of basis functions K is incremented by one the numerator of $\widehat{CV(\lambda)}$ changes and hence this procedure provides an optimal smoothing parameter λ for the estimate π_K based on K basis functions. Consequently, the estimate provided by this algorithm is given by $\pi_{K,\hat{\lambda}}$. One may notice that in theory there is no guarantee that the stop criterion will stop before K reaches n but numerical simulations in (Kooperberg and Stone, 1991; Dias, 1997; Dias, 1996) have shown that this would happen in general. In fact, (Dias, 1997; Dias, 1996) showed by using numerical simulations that the optimal K is around to $\log n + c$ where a constant c depends on the data. Intuitively speaking, K is the dimension of the finite space where the approximation is taking place, as K increases more details of the density is picked up by the density estimate from the data and then there exists a K_0 such that $K > K_0$, the coefficients of the expansion get close to zero and so they would not be need on the fitting. Thus the difference of two consecutive densities will be undistinguishable. This procedure should be used on the context of nonparametric empirical Bayes where there is considerable amount of information to estimate the marginal density of the data $m(\cdot)$. The optimization of (2.2) is known to be of difficult task but the author believes that the approach developed in this paper would make this task easier.

ACKNOWLEDGEMENT

I would like to thank the referee for the comments and suggestions that made this work better and clearer. This research was partially support by FAPESP grant no. 95/4996-3.

References

Bennett, G. K. and Martz, H. F. (1972). A continuous empirical bayes smoothing technique, *Biometrika* **59**: 361–368.

Berger, J. O. (1985). *Statistical Decision theory and Bayesian Analysis (Second Edition)*, Springer-Verlag.

de Boor, C. (1978). *A Practical Guide to Splines*, Springer Verlag, New York.

Dias, R. (1996). Sequential adaptive nonparametric regression via h-splines. Technical Report RP 43/96, University of Campinas, June 1996. Submitted.

Dias, R. (1997). Density estimation via h-splines, *Journal of Statistical Computation and Simulation*. To appear.

Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm, *J. of the Amer. Stat'l. Assn.* **88**: 495–504.

Kooperberg, C. and Stone, C. J. (1991). A study of logspline density estimation, *Computational Stat. and Data Analysis* **12**: 327–347.