# MARS: Selecting Basis and Knots with the Empirical Bayes Method

Wataru SAKAMOTO [*]

*Abstract* — **The multivariate adaptive regression spline (MARS), proposed by Friedman (1991), estimates regression structure including interaction terms adaptively with truncated power spline basis functions. However, it adopts the generalized cross-validation criterion to add and prune basis functions, and hence it tends to choose such large numbers of basis functions that estimated regression structure may not be easily interpreted. On the other hand, some Bayesian approaches incorporated in MARS have been proposed, in which the reversible jump MCMC algorithm is adopted. However, they generate enormous combinations of basis functions, from which it would be difficult to obtain clear interpretation on regression structure.**

**An empirical Bayes method to select basis functions and knots in MARS is proposed, with taking both advantages of the frequentist model selection approach and the Bayesian approach. A penalized likelihood approach is used to estimate basis coefficients for given basis functions, and the Akaike Bayes information criterion (ABIC) is used to determine the number of basis functions. It is shown that the proposed method gives estimation of regression structure which is relatively simple and easy to interpret for some example data sets.**

## 1 Introduction

The multivariate adaptive regression spline (MARS), proposed by Friedman (1991), estimates regression structure including interaction terms adaptively with truncated power spline basis functions. The selected basis functions as well as the selected positions of knots give suggestive information for understanding regression structure including interaction terms.

However, Friedman's MARS adopts the generalized cross-validation criterion to add and prune basis functions. It has been pointed out that the cross-validation criterion tends to make estimated functions relatively rough, because it is intended to optimize prediction of responses. In some applications, Friedman's MARS often choose such large numbers of basis functions that estimated regression structure may not be easily interpreted. Moreover, the criterion based on the distance in the space of responses does not seem to match with the original objective of the MARS that is to explore regression structure. From these points of view, another criterion should be reconsidered.

Recently, some Bayesian approaches incorporated in MARS have been proposed (see Denison *et al.*, 2002). Priors are prescribed on the numbers of basis functions, the variables involved in each basis function, the position of knots, basis coefficients and so on, as well as some hyper-parameters, and then posterior sampling of the estimated function is conducted with the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm (Green, 1995). The Bayesian approaches provide good prediction performance in some applications. However, they generate enormous combinations of basis functions, from which it would be difficult to obtain clear interpretation on regression structure. Moreover, since the sampling with the RJMCMC is extremely computer-intensive, the analysis is not feasible on PC.

In this paper, we propose an empirical Bayes method to select basis functions and knots in MARS, with taking both advantages of the frequentist model selection approach and the Bayesian approach. We use a penalized likelihood approach to estimate basis coefficients for given basis functions, with a prior on the basis coefficients viewed as a penalty for complexity. We select the basis functions and the position of knots as well as the hyper-parameters by maximizing the Laplace approximation of the marginal likelihood, and determine the number of basis functions by minimizing the Akaike Bayes information criterion (ABIC) (Akaike, 1980). We show an application of our method to some example data.

## 2 The MARS model

We suppose that each response $y_i$, $i = 1, \ldots, n$, is observed with $r$ explanatory variables $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ir})^{\mathrm{T}}$. We would like to estimate a function $f$ that satisfies a regression model such as

$$\mathrm{E}(y_i) = f(\boldsymbol{x}_i), \quad i = 1, \ldots, n. \tag{1}$$

[*] Division of Mathematical Science, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama-cho, Toyonaka, Osaka 560-8531, JAPAN, E-mail: `sakamoto@sigmath.es.osaka-u.ac.jp`, Tel: +81-6-6850-6481

We assume that the regression function $f$ is represented as a linear combination of basis functions $\phi_k(\boldsymbol{x}), k = 1, \ldots, K$:

$$f(\boldsymbol{x}) = a_0 + \sum_{k=1}^{K} a_k \phi_k(\boldsymbol{x}), \tag{2}$$

and the coefficients $a_0, a_1, \ldots, a_K$ are estimated. Each basis function is the truncated power spline function

$$\phi_k(\boldsymbol{x}) = \prod_{l=1}^{L_k} [s_{kl}(x_{v(k,l)} - c_{kl})]_+^m, \tag{3}$$

where $[x]_+$ is the positive part of $x$, that is, $[x]_+ = x$ if $x > 0$ and $[x]_+ = 0$ if $x \leq 0$, and a given positive integer $m$ is the order of splines. The number of basis functions $K$, the degree of interaction $L_k$, the sign $s_{kl}$, which is $+1$ or $-1$, the number of the variable involved in the basis function $v(k, l)$, and the position of a knot $c_{kl}$ are all selected with some lack-of-fit criterion based on data.

The algorithm of Friedman's (1991) MARS is composed of the following two sections:

- Forward stepwise:
  Set the basis for a constant term as $\phi_0 \equiv 1$. Supposing that we have $K + 1$ basis functions $\phi_0, \phi_1(\boldsymbol{x}), \ldots, \phi_K(\boldsymbol{x})$, add new two basis functions

  $$\phi_{K+1}(\boldsymbol{x}) = \phi_k(\boldsymbol{x})[+(x_{v(k,l)} - c_{kl})]_+^m,$$
  $$\phi_{K+2}(\boldsymbol{x}) = \phi_k(\boldsymbol{x})[-(x_{v(k,l)} - c_{kl})]_+^m,$$

  where $\phi_k(\boldsymbol{x})$ is a parent basis function, $x_{v(k,l)}$ is a variable that is not included in $\phi_k(\boldsymbol{x})$ and $c_{kl}$ is a knot position ($c_{kl} \in \{x_{iv(k,l)}\}_{i=1,\ldots,n}$), all of which are selected so that they minimize a lack-of-fit criterion.
  The addition of basis functions is continued until $K$ becomes greater than $K_{\max}$.

- Backward stepwise:
  Select one basis function (except $\phi_0$) and prune it so that the lack-of-fit criterion is minimized.
  The pruning is continued until the lack-of-fit criterion does not become decreasing even if a basis function is deleted.

Friedman (1991) uses a version of the generalized cross-validation score, originally proposed by Craven and Wahba (1979), in which he also suggests using a cost complexity function proposed by Friedman and Silverman (1989).

For given basis functions $\phi_k(\boldsymbol{x})$, $k = 0, 1, \ldots, K$, we obtain the least square estimate of $\boldsymbol{a} = (a_0, a_1, \ldots, a_K)^{\mathrm{T}}$, $\hat{\boldsymbol{a}}_{\mathrm{LS}} = (\Phi_K^{\mathrm{T}}\Phi_K)^{-1}\Phi_K^{\mathrm{T}}\boldsymbol{y}$, where $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $\Phi_K$ is the matrix with the $(i, k)$ component $\phi_k(\boldsymbol{x}_i)$. However, the problem of multicolinearity must occur in the matrix $\Phi_K$ as $K$ increases. The fact motivates the use of some regularization method such as the ridge regression.

## 3 Empirical Bayes method for MARS

### 3.1 Estimation of basis coefficients

We assume (2) and (3) as in Section 2. Supposing that the basis functions $\phi_k(\boldsymbol{x})$, $k = 0, 1, \ldots, K$, are given, we consider a normal prior on the coefficient $\boldsymbol{a}$,

$$p(\boldsymbol{a}|\Phi_K, \lambda) = (\lambda/2\pi)^{(K+1)/2} \exp\left(-\frac{\lambda}{2}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}\right), \tag{4}$$

where the positive number $\lambda$ is a hyper-parameter, which controls the complexity of the function. The prior is also adopted in some Bayesian approaches (for example, Denison *et al.*, 2002).

Under the prior (4), the posterior density of $\boldsymbol{a}$ is obtained, from Bayes' theorem, as

$$p(\boldsymbol{a}|\boldsymbol{y}, \Phi_K, \sigma^2, \lambda) \propto p(\boldsymbol{y}|\Phi_K, \boldsymbol{a}, \sigma^2)p(\boldsymbol{a}|\Phi_K, \lambda), \tag{5}$$

where $p(\boldsymbol{y}|\Phi_K, \boldsymbol{a}, \sigma^2)$ is the density of $\boldsymbol{y}$ for given $\Phi_K$ and $\boldsymbol{a}$, and $\sigma^2$ is the variance component of the distribution of $\boldsymbol{y}$. In the case of normal responses with the error variance $\sigma^2$, the posterior distribution of $\boldsymbol{a}$ becomes normal with the mean

$$\hat{\boldsymbol{a}} = (\Phi_K^{\mathrm{T}}\Phi_K + \lambda^*\mathrm{I})^{-1}\Phi_K^{\mathrm{T}}\boldsymbol{y} \tag{6}$$

and the variance matrix $\sigma^2(\Phi_K^{\mathrm{T}}\Phi_K + \lambda^*\mathrm{I})^{-1}$, where $\lambda^* = \lambda\sigma^2$, and I is the identical matrix.

Finding the mode, $\hat{\boldsymbol{a}}$, of the posterior density of $\boldsymbol{a}$, is equivalent to maximizing the penalized log-likelihood

$$\begin{aligned} &l_{\mathrm{P}}(\boldsymbol{a}|\boldsymbol{y}, \Phi_K, \sigma^2, \lambda) \\ &= \log p(\boldsymbol{y}|\Phi_K, \boldsymbol{a}, \sigma^2) - \tfrac{\lambda}{2}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a} + \tfrac{K+1}{2}\log\lambda + \text{const.} \end{aligned} \tag{7}$$

with respect to $\boldsymbol{a}$. In the normal distribution case, we can compute the maximum penalized likelihood estimate (MPLE) (6) directly. In non-normal distribution case, we can obtain the MPLE via an iterative algorithm such as the Fisher scoring method (see, for example, McCullagh and Nelder, 1989; Green and Silverman, 1994).

### 3.2 Selection of basis functions and knots

To estimate the knot position $\boldsymbol{c} = \{c_{kl}\}$, the variance component $\sigma^2$ and the hyper parameter $\lambda$, we consider the marginal likelihood as a lack-of-fit criterion. The marginal likelihood, in which $\boldsymbol{a}$ is integrated-out from the posterior density of $\boldsymbol{a}$ (5), is written as

$$\begin{aligned} p(\boldsymbol{y}|\Phi_K, \sigma^2, \lambda) &= \int p(\boldsymbol{y}|\Phi_K, \boldsymbol{a}, \sigma^2) p(\boldsymbol{a}|\Phi_K, \lambda) d\boldsymbol{a} \\ &= \int \exp\{l_{\mathrm{P}}(\boldsymbol{a}|\boldsymbol{y}, \Phi_K, \sigma^2, \lambda)\} d\boldsymbol{a}. \quad (8) \end{aligned}$$

The method of maximizing the marginal likelihood (8) can be regarded as a kind of the empirical Bayes method in which non-informative priors are assumed to $\sigma^2$ and $\lambda$,

The exact computation of the integral included in (8) is not generally feasible except in the normal response case. Using the Laplace approximation (Tierney and Kadane, 1986; Davison, 1986), that is, the Taylor expansion of the penalized log-likelihood (7) around its maximum $\tilde{\boldsymbol{a}}$, we obtain an approximated marginal log-likelihood

$$\begin{aligned} & l_{\mathrm{M}}(\boldsymbol{c}, \sigma^2, \lambda|\boldsymbol{y}) = \log p(\boldsymbol{y}|\Phi_K, \sigma^2, \lambda) \\ & \quad \approx l_{\mathrm{P}}(\hat{\boldsymbol{a}}|\boldsymbol{y}, \Phi_K, \sigma^2, \lambda) - \tfrac{1}{2} \log |H_{\mathrm{P}}(\hat{\boldsymbol{a}})| + \text{const.}, \quad (9) \end{aligned}$$

where $H_{\mathrm{P}}(\hat{\boldsymbol{a}})$ is the negative Hessian of the penalized log-likelihood

$$H_{\mathrm{P}}(\hat{\boldsymbol{a}}) = \left( - \frac{\partial^2 l_{\mathrm{P}}}{\partial \boldsymbol{a} \partial \boldsymbol{a}^{\mathrm{T}}} \right)_{\hat{\boldsymbol{a}}}.$$

In the normal response case, the expression (9) is exact, and $H_{\mathrm{P}}(\hat{\boldsymbol{a}}) = \sigma^{-2}(\Phi_K^{\mathrm{T}}\Phi_K + \lambda^* \mathrm{I})$. We can obtain a similar expression in non-normal response case.

We maximize the approximated marginal log-likelihood (9) with respect to the knot position $\boldsymbol{c} = \{c_{kl}\}$, the variance component $\sigma^2$ and the hyper parameter $\lambda$, as well as the variables $\{v(k, l)\}$ involved in basis functions and even the combination of basis functions. However, the marginal log-likelihood (9) seems to increase as the number of basis functions $K$ increase. Our idea is to select $K$ and the combination of basis functions that minimize Akaike's Bayes information criterion (ABIC) (Akaike, 1980)

$$\mathrm{ABIC} = -2 l_{\mathrm{M}}(\boldsymbol{c}, \sigma^2, \lambda|\boldsymbol{y}) + 2(q + 2)$$

in the forward stepwise section, where $q$ is the number of knots $\boldsymbol{c} = \{c_{kl}\}$ involved in the basis functions. In the backward stepwise section, we only maximize the marginal log-likelihood $l_{\mathrm{M}}(\boldsymbol{c}, \sigma^2, \lambda|\boldsymbol{y})$, as the knots $\boldsymbol{c}$ are already estimated.

## 4 Example

We applied our empirical Bayes method to the kyphosis data set, which is also analyzed by Hastie and Tibshirani (1990) and Chambers and Hastie (1992). We fitted the logistic regression model, instead of (1),

$$\log \frac{p_i}{1 - p_i} = f(\boldsymbol{x}_i), \quad i = 1, \ldots, n,$$

where $p_i = \Pr(y_i = 1)$, $y_i$ is the binary variable for the $i$-th patient that indicates whether the kyphosis remains after the surgery, and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})^{\mathrm{T}}$ are [1] the age (in month), [2] the starting vertebrae level of the surgery and [3] the number of the vertebrae levels involved, respectively. The order of the spline was set to $m = 1$ for simplicity.

Table 1 shows the stepwise process of the Empirical Bayes MARS applied to the kyphosis data. The rightmost column indicates the numbers of the variable involved in the additional basis functions. In the forward stepwise section, the ABIC score is maximized when seven basis functions are used. In the backward stepwise section, two basis functions are pruned out of the seven. The resulting estimated function becomes

$$\begin{aligned} \hat{f}(\boldsymbol{x}) &= 1.02 - 0.369[x_2 - 6]_+ - 0.105[6 - x_2]_+ \\ & - 0.00285[96 - x_1]_+ - 0.00191[96 - x_1]_+[14 - x_3]_+, \end{aligned}$$

which suggests that the risk of kyphosis is the highest if the age is greater than 96 months, the starting level is around 6 and the number of the level is around 14. The result is similar to those obtained in some literature such as Chambers and Hastie (1992).

| $K+1$ | $l_{\mathrm{M}}$ | ABIC | $\log_{10}\lambda$ | $v(k,\cdot)$ |
|---|---|---|---|---|
| Forward stepwise: | | | | |
| 1 | −44.725 | 91.450 | | |
| 3 | −37.841 | 79.682 | −1.16 | 2 |
| 5 | −36.289 | 78.578 | −1.43 | 1 |
| 7 | −34.713 | 77.426 * | −1.68 | 1, 3 |
| 9 | −34.186 | 78.372 | −1.67 | 1, 3 |
| 11 | −33.918 | 79.836 | −1.76 | 1, 3 |
| Backward stepwise: | | | | |
| 6 | −34.713 | | −1.68 | |
| 5 | −34.136 | | −1.43 | |

Table 1: Empirical Bayes MARS applied to the kyphosis data: the stepwise process.

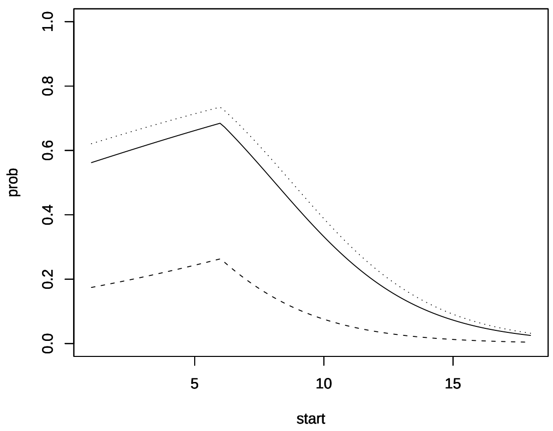

Figure 1: Empirical Bayes MARS applied to the kyphosis data: the estimated probability.

Figure 1 shows the estimated probability as functions of $x_2$. The solid line is drawn at $x_1 = 84.7$ and $x_3 = 4.2$ (both are the mean values), the dashed line is drawn at $x_1 = 1$ (minimum) and $x_3 = 4.2$, The dotted line is drawn at $x_1 \geq 96$ and $x_3 = 4.2$. The figure suggests that both the age and the start level are highly related to the risk of kyphosis.

## 5 Concluding Remarks

We have proposed the empirical Bayes approach for the MARS model. The proposed method provides estimation of regression structure which is relatively simple and easy to interpret in the case of the data set we applied.

We are studying the performance of the empirical Bayes MARS about how to extract regression structure including interaction terms rightly. Although we used a simple prior on basis coefficients in this paper, but we could use more elaborate priors in connection with the roughness penalty. We are considering the extension to apply to hierarchical data, longitudinal data, functional data and so on, and some application to classification and discrimination problems. Furthermore, we will be able to develop an empirical Bayes method for classification and regression tree (CART) (Breiman *et al.*, 1984).

## Acknowledgement

This research is supported by MEXT Grant-in-Aid for Young Scientists (B), No. 17700280 (2005).

## References

[1] Akaike, H. (1980). Likelihood and Bayes procedure. In *Bayesian Statistics* (Bernardo, J. M. et al., eds.), pp. 143–166, University Press, Valencia.

[2] Breiman, L., Friedman, J. H., Olshen, R. and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth, California.

[3] Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S.* Pacific Grove, California.

[4] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377–403.

[5] Davison, A. C. (1986). Approximate predictive likelihood. *Biometrika* **73**, 323–332.

[6] Denison, D. G. T., Holmes, C. C., Mallick, B. K. and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression.* John Wiley and Sons.

[7] Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19**, 1–141.

[8] Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics* **31**, 3–39.

[9] Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

[10] Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

[11] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

[12] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* (2nd edition). Chapman and Hall, London.

[13] Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81**, 82–86.