# Многомерные адаптивные регрессионные сплайны*

***В. Р. Целых***
Celyh@inbox.ru
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

В работе рассматриваются многомерные адаптивные регрессионные сплайны. Метод позволяет получить модели, дающие достаточно точную аппроксимацию, даже в тех случаях, когда связи между предикторными и зависимыми переменными имеют немонотонный характер и сложны для приближения параметрическими моделями. Экспериментально исследуется зависимость ошибки аппроксимации от сложности модели. Для иллюстрации работы метода используются тестовые данные, данные ЭКГ и данные из области финансовой математики.

**Ключевые слова**: *непараметрическая регрессия, многомерные адаптивные регрессионные сплайны, метод наименьших квадратов, обобщенный скользящий контроль.*

# Multivariate adaptive regression splines*

***V. R. Tselykh***
Moscow Institute of Physics and Technology

The article describes multivariate adaptive regression splines, which are very useful for high dimensional problems and show a great promise for fitting nonlinear multivariate functions. This technique does not impose any particular class of relationship between the predictor variables and outcome variable of interest. The error of approximation in relation to the model complexity is investigated. To illustrate the method test data, ECG data and information from the area of financial mathematics are used.

**Keywords**: *nonparametric regression, multivariate adaptive regression splines, least-squares method, generalized cross-validation.*

## Введение

Многомерные адаптивные регрессионные сплайны были впервые предложены Фридманом в 1991 г. [1] для решения регрессионных задач и задач классификации, в которых требуется предсказать значения набора зависимых переменных по набору независимых переменных. Данный метод является непараметрической процедурой, не использующей в своей работе никаких предположений о виде функциональной зависимости между зависимыми и независимыми переменными. MAP-сплайны задаются базисными функциями и набором коэффициентов, полностью определяемых по данным.

MAP-сплайны находят свое применение во многих сферах науки и технологий, например, в предсказании видов распределений по имеющимся данным [2], кишечного поглощения лекарств [3], а также в воспроизведении речи [4] и поиске глобального оптимума в проектировании конструкций [5].

Метод MAP-сплайнов находит искомую зависимость за 2 стадии: “вперед” (forward stage) и “назад” (backward stage) [7]. Первая стадия заключается в добавлении базисных

---

Научный руководитель В. В. Стрижов

функций к набору, пока не будет достигнут максимальный уровень сложности. На второй стадии из набора удаляются функции, которые вносят наименьший вклад в ошибку.

В данной работе описывается алгоритм построения MAP-сплайнов, тестируется его работа на ряде данных, а также проводится анализ сложности (числа базисных функций) построенной модели.

## Описание работы алгоритма

Дана регрессионная выборка:

$$D = \{\mathbf{x}_i, y_i\}_{i=1}^{N},$$

где $\mathbf{x}_i \in \mathbb{R}^p, i = 1, \ldots, N$ — векторы независимой переменной, а $y_i$, $i = 1, \ldots, N$ — значения зависимой переменной (непрерывные или бинарные). Связь между $y_i$ и $\mathbf{x}_i$ $(i = 1, \ldots, N)$ может быть представлена в виде:

$$y_i = f(x_i^1, x_i^2, \ldots, x_i^p) + \varepsilon = f(\mathbf{x}_i) + \varepsilon,$$

где $f$ — неизвестная функция, а $\varepsilon$ — ошибка $(\varepsilon \sim N(0, \sigma^2))$.

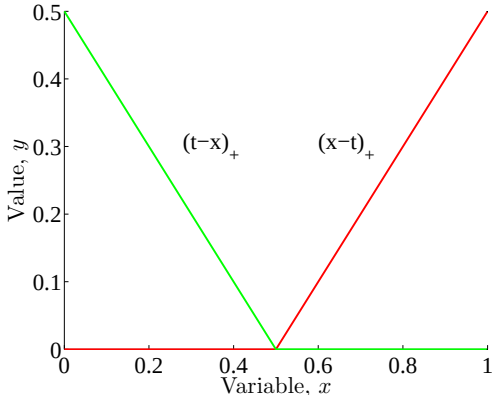


**Рис. 1.** Базисные функции $(x - t)_+$ и $(t - x)_+$

В одномерном случае MAP-сплайны выражаются через кусочно-линейные базисные функции, $(x-t)_+$ и $(t-x)_+$ с узлом в $t$. Данные функции являются усеченными линейными функциями (см. рис. 1), при $x \in \mathbb{R}$:

$$(x - t)_+ = \begin{cases} x - t, & \text{если } x > t; \\ 0, & \text{иначе,} \end{cases}$$

$$(t - x)_+ = \begin{cases} t - x, & \text{если } x < t; \\ 0, & \text{иначе.} \end{cases}$$

Эти функции также называются отраженной парой (reflected pair). В многомерном случае для каждой компоненты $x^j$ вектора $\mathbf{x} = (x^1, \ldots, x^j, \ldots, x^p)^T$ строятся отраженные пары с узлами в каждой наблюдаемой переменной $x_i^j$ $(i = 1, 2, \ldots, N; j = 1, 2, \ldots, p)$. Таким образом, набор построенных функций может быть представлен в виде:

$$C := \{(x^j - t)_+, (t - x^j)_+ | \, t \in \{x_1^j, x_2^j, \ldots, x_N^j\}, j \in \{1, 2, \ldots, p\}\}.$$

Если все входные данные различны, то в наборе $2Np$ функций, причем каждая из них зависит только от одной переменной $x^j$.

Используемые для аппроксимации базисные функции выглядят следующим образом:

$$B_m(\mathbf{x}) = \prod_{k=1}^{K_m} [s_{km} \cdot (x^{v(km)} - t^{km})]_+,$$

где $K_m$ — общее число усеченных линейных функций в $m$-ой базисной функции, $x^{v(km)}$ — компонента вектора $\mathbf{x}$, относящаяся к $k$-ой усеченной линейной функции в $m$-ой базисной функции, $t^{km}$ — соответствующий узел, а $s_{km} \in \{\pm 1\}$.

Построенная модель, как и в линейной регрессии, представляет собой линейную комбинацию, отличие состоит в том, что кроме входных переменных разрешается использовать функции из набора $C$ и их производные функции. Таким образом, модель имеет вид:

$$y = \hat{f}(\mathbf{x}) + \varepsilon = c_0 + \sum_{m=1}^{M} c_m B_m(\mathbf{x}) + \varepsilon,$$

где $M$ — число базисных функций в рассматриваемой модели, а $c_0$ — общий коэффициент. Как и в линейной регрессии, задав $B_m$, коэффициенты $c_m$ могут быть найдены по методу наименьших квадратов. Самое главное в данной модели — это выбор базисных функций. В начале модель содержит единственную функцию $B_0(\mathbf{x}) = 1$, а все функции из набора $C$ являются возможными кандидатами для включения в модель.

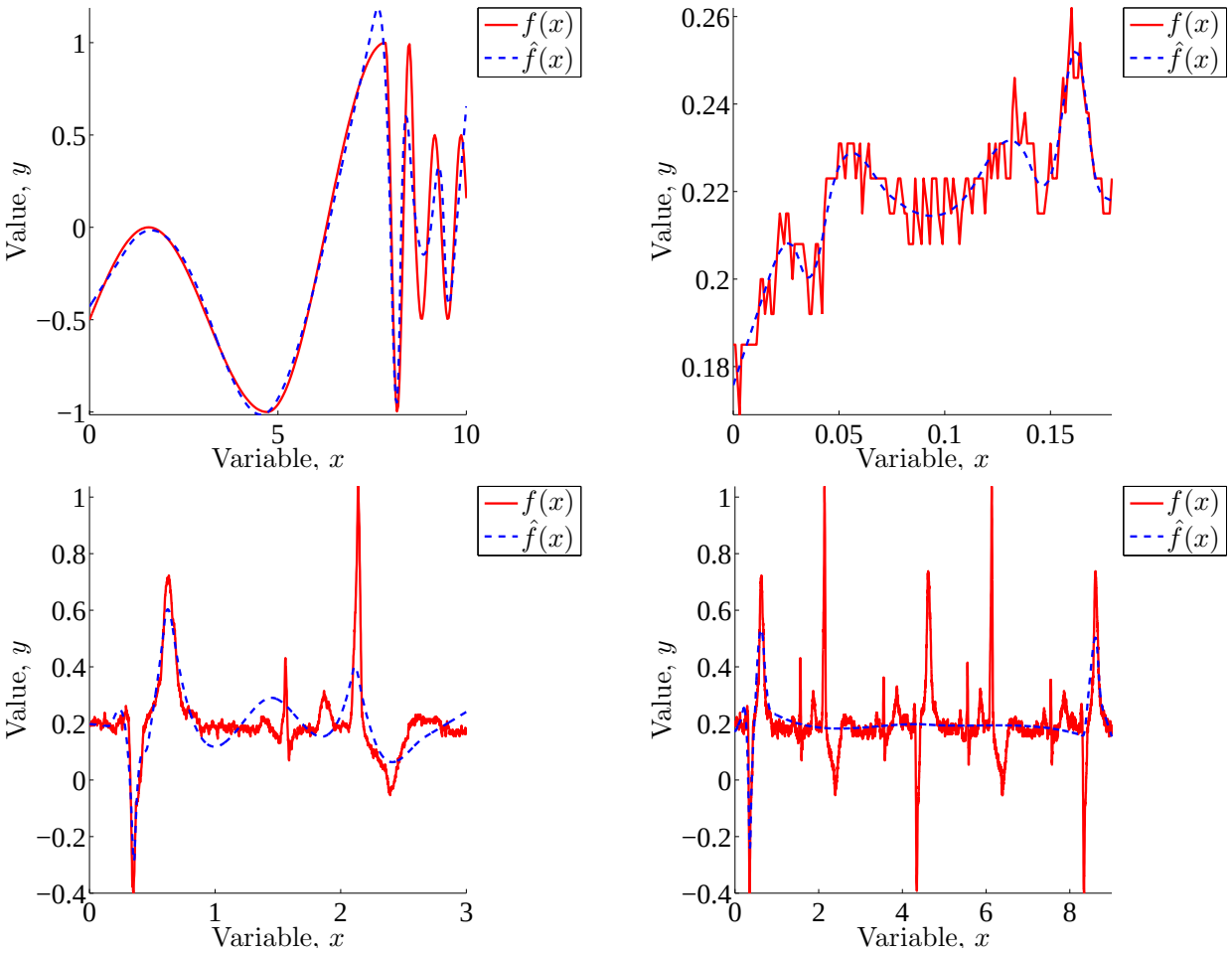


**Рис. 2.** Функции и построенные аппроксимации

К примеру, следующие функции могут быть базисными:

— $1$,
— $x^j$,
— $(x^j - t^k)_+$,
— $x^l x^j$,
— $(x^j - t^k)_+ x^l$,
— $(x^j - t^k)_+ (x^l - t^h)_+$.

В данном алгоритме каждая базисная функция зависит от разных переменных. Это означает, что $l \neq j$ в базисных функциях, указанных выше. На каждом шаге новая пара базисных функций является произведением функции $B_m(\mathbf{x})$ из множества моделей $\mathcal{M}$ на одну из отраженных пар множества $C$. Таким образом, в модель $\mathcal{M}$ будет добавлено:

$$\hat{C}_{M+1} B_l(\mathbf{x})(x^j - t)_+ + \hat{C}_{M+2} B_l(\mathbf{x})(t - x^j)_+;$$

что обеспечит наибольшее уменьшение ошибки. Коэффициенты $\hat{C}_{M+1}$ и $\hat{C}_{M+2}$ оцениваются методом наименьших квадратов, как и остальные $M + 1$ коэффициентов модели. Процедура добавления функций в модель продолжается до тех пор, пока множество $\mathcal{M}$ содержит менее заданного числа элементов.

Ниже предложены возможные базисные функции:

— $x^j \quad (j = 1, 2, \ldots, p)$,
— $(x^j - t^k)_+$, если $x^j$ уже в модели,
— $x^l x^j$, если $x^l$ и $x^j$ уже в модели,
— $(x^j - t^k)_+ x^l$, если $x^l x^j$ и $(x^j - t^k)_+$ уже в модели,
— $(x^j - t^k)_+ (x^l - t^h)_+$, если $(x^j - t^k)_+ x^l$ и $(x^l - t^h)_+ x^j$ уже в модели.

В конце данной процедуры построена большая модель, которая включает в себя некоторые излишние переменные и обычно чрезмерно подгоняет данные. Необходимо проведение стадии “назад”, которая заключается в следующем: на каждом шаге удаляется функция, отсутствие которой вызывает наименьшее увеличение суммы квадратов невязок (RSS). Таким образом, для каждого размера $M$ строится наилучшая модель $\hat{f}_M$. Для оценки оптимальной величины $M$ используется процедура обобщенного скользящего контроля (generalized cross-validation). Данный критерий (также известный как lack-of-fit criterion) выглядит следующим образом [1]:

$$LOF\hat{f}_M = GCV(M) := \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{f}_M(x_i))^2 / (1 - C(M)/N)^2,$$

$$C(M) = \mathrm{trace}(\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T) + 1,$$

где $N$ — число исходных данных, $C(M)$ — оценка штрафов в модели, содержащей $M$ базисных функций, $B$ — матрица размером $M \times N$ ($B_{ij} = B_i(\mathbf{x}_j)$). $C(M)$ — число параметров, подлежащих определению. Один из способов задания числа параметров: $C(M) = r + cK$. Число линейно-независимых базисных функций в модели обозначено $r$, число узлов, выбранных в стадии “вперед” — $K$, а число $c$ показывает оценку оптимизации каждой из базисных функций. В общем случае, $c = 3$, но если используемая модель является аддитивной, то $c = 2$. Чем меньше $C(M)$, тем больше получаемая модель и больше число базисных функций, и наоборот соответственно. GCV представляет собой средний квадрат невязок умноженных на коэффициент, характеризующий сложность модели. Таким

образом, наилучшая модель состоит из $M^*$ базисных функций, где $M^*$ — решение задачи минимизации $LOF\hat{f}_M$ [8, 9]:

$$M^* = \arg\min_M LOF\hat{f}_M.$$

Особенность метода MAP-сплайнов заключается в использовании кусочно-линейных базисных функций и определенном способе построения модели. Главным свойством кусочно-линейных функций является их способность действовать локально, т. е. принимать ненулевые значения лишь на части их области определения. Результат умножения одной функции на другую отличен от нуля лишь в малой части пространства, где обе функции принимают ненулевые значения. Это и позволяет строить качественные модели, используя сплайны. Если же в качестве базисных функций использовать полиномы, то результат будет хуже по причине того, что полиномы отличны от нуля во всем пространстве.

## Вычислительный эксперимент

Для проверки работы алгоритма используется программное обеспечение ARESLab [6]. Сначала тестируется работа алгоритма MAP-сплайнов на простой зависимости $f(x)$, имеющей вид:

$$f(x) = \begin{cases} 0.5\sin x - 0.5, & \text{если } 0 \leqslant x < 1.5\pi; \\ \sin x, & \text{если } 1.5\pi \leqslant x < 2.5\pi; \\ -\cos(10x), & \text{если } 2.5\pi \leqslant x < 2.75\pi; \\ 0.5\cos(9x - 0.25\pi), & \text{если } 2.75\pi \leqslant x \leqslant 10; \end{cases}$$

Результат аппроксимации при мощности регрессионной выборки $N = 300$ изображен на рис. 2 в левом верхнем углу. На данном графике красным цветом обозначена исходная зависимость $f(x)$, а синим — ее аппроксимация.

Далее для иллюстрации работы алгоритма рассматривается электрокардиограмма. Результат работы алгоритма при $N = 130$ представлен на рис. 2 в правом верхнем углу. По горизонтальной оси откладывается время $t$, а по вертикальной — значение напряжения при получении ЭКГ. Красным цветом на графике обозначена исходная зависимость, а синим — ее аппроксимация.

Увеличив мощность регрессионной выборки до 2000, получим аппроксимацию, изображенную на рис. 2 в левом нижнем углу.

И наконец, рассмотрев $N = 6000$, получим аппроксимацию, изображенную на рис. 2 в правом нижнем углу.

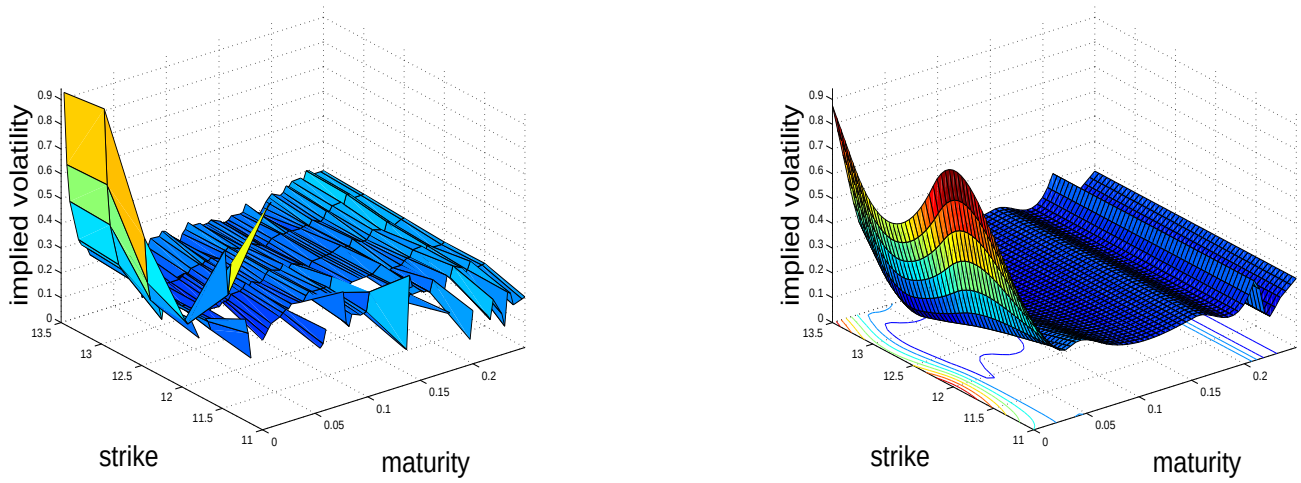


**Рис. 3.** Функция двух переменных (слева) и построенная аппроксимация (справа)

Данный метод эффективен не только при рассмотрении функций одной переменной, но и в многомерных пространствах. Рассматривается функция двух переменных. Данные взяты из области финансовой математики [10]. По оси $x$ откладывается время до исполнения опциона (maturity), по оси $y$ — цена исполнения опциона (strike), а по оси $z$ — волатильность (implied volatility) опциона [11]. Исходная зависимость представлена на рис. 3 слева. Аппроксимация данной зависимости при максимальном числе пересечений равном 2 и максимальном числе базисных функций на этапе добавления равном 21 изображена на рис. 3 справа.

Из представленных результатов следует, что метод MAP-сплайнов достаточно хорошо описывает любые зависимости.

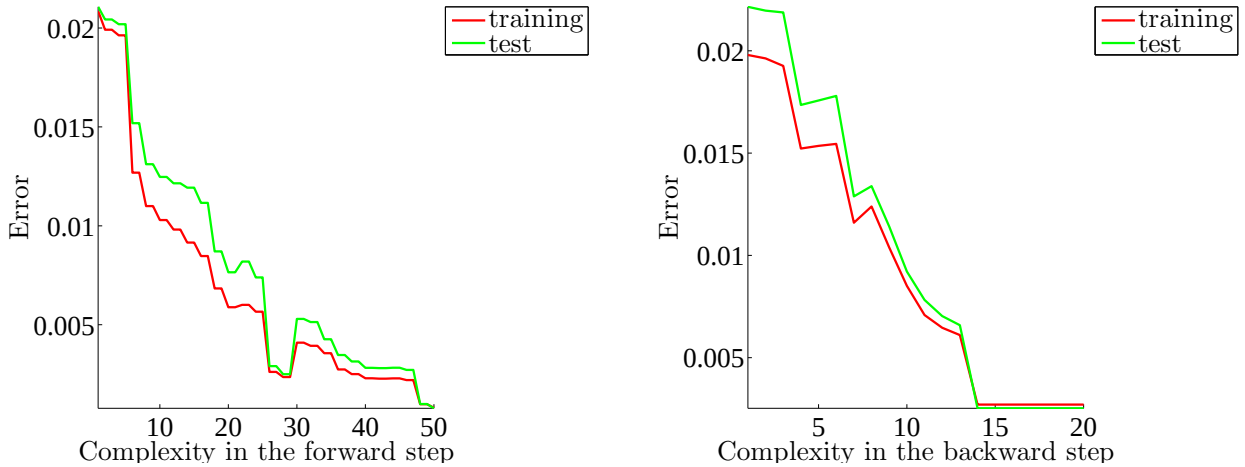


**Рис. 4.** Зависимость ошибки от числа базисных функций на этапе добавления (слева) и удаления (справа)

Исследуется анализ сложности модели, т. е. число базисных функций в модели. Рассматривается зависимость, которая была приближена на рис. 2 в левом нижнем углу. Выборка случайным образом разделяется на обучающую и проверяющую. Зависимость ошибки (суммы квадратов невязок) при обучении и контроле от числа базисных функций на этапе добавления изображена на рис. 4 слева.

Из данных графиков можно сделать вывод о том, что на стадии “вперед” при числе базисных функций порядка 27 ошибка и на обучающей, и на проверяющей выборке достигает локального минимума.

Зависимость ошибки при обучении и контроле от числа базисных функций на этапе удаления функций из модели (при числе функций равном 27 на этапе добавления) изображена на рис. 4 справа.

Из представленных графиков следует, что на предложенных данных оптимальное число базисных функций на этапе удаления функций из модели равно 14. Значит, во второй стадии эффективного алгоритма из модели удаляется примерно половина базисных функций.

## Заключение

В данной работе был описан метод MAP-сплайнов, используемый для нахождения функциональной зависимости между предикторными и зависимыми переменными. Алгоритм был протестирован на ряде данных и показал достаточно хороший результат. Была исследована зависимость ошибки на обучении и контроле от числа базисных функций (как на этапе добавления, так и на этапе удаления функций из модели). При этом оказалось, что число базисных функций в заключительной модели примерно вдвое меньше числа базисных функций в конце первой стадии работы алгоритма.

## Литература


[1] Friedman, J.H. *Multivariate adaptive regression splines*, The Annals of Statistics,19, 1 (1991) 1-141.

[2] Elith, J., and Leathwick, J. *Predicting species distribution from museum and herborium records using multiresponse models fitted with multivariate adaptive regression splines*, Diversity and Distributions, 13, 3 (2007) 265-275.

[3] Deconinck, E., Coomons, D., and Heyden, Y.V. *Explorations of linear modeling techniques and their combinations with multivariate adaptive regression splines to predict gastro-intestinal absorption of drugs*, Journal of Pharmaceutical and Biomedical Analysis, 43, 1 (2007) 119-130.

[4] Haas, H., and Kubin, G. *A multi-band nonlinear oscillator model for speech*, Conference Record of the Thirty- Second Asilomar Conference on Signals, Systems and Computers, 1 (1998) 338-342.

[5] Crino, S., and Brown, D.E. *Global optimization with multivariate adaptive regression splines*, IEEE Transactions on Systems Man and Cybernetics Part b — cybernetics, 37, 2 (2007) 333-340.

[6] Gints Jekabsons' webpage, *ARESLab: Adaptive Regression Splines toolbox for Matlab/Octave*, http://www.cs.rtu.lv/jekabsons/regression.html.

[7] Yerlikaya, F. *A new contribution to nonlinear robust regression and classification with MARS and its applications to data mining for quality control in manufacturing*, M.Sc., Department of Scientiic Computing (2008) 1-102.

[8] Di, W. *Long Term Fixed Mortgage Rate Prediction Using Multivariate Adaptive Regression Splines*, School of Computer Engineering, Nanyang Technological University, 2006.

[9] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning, Data Mining, Inference and Prediction*, Springer, 2001.

[10] Strijov, V., *Volatility smile modelling: two-dimensional linear regression demo*, http://strijov.com/sources/demo_linfit_options.php#1.

[11] Стрижов, В., и Сологуб, Р. *Индуктивное построение регрессионных моделей волатильности опционных торгов*, Вычислительные технологии, том 14, №5, 2009.