УДК 519.234

Об одном методе сглаживания двумерной поверхности

П. Г. Любин

ФБГОУ ВО МГТУ «СТАНКИН», Москва, Россия

Регрессионный анализ ставит перед собой задачу отыскания функциональной зависимости между наблюдаемыми величинами изучаемого процесса. При этом исходные данные являются реализацией случайной величины, поэтому рассматривается зависимость математического ожидания. Такую задачу можно решать путём «сглаживания» исходных данных. Под сглаживанием понимается попытка удаления шума и несущественных фрагментов при сохранении наиболее важных свойств структуры данных, то есть результат подобен математическому ожиданию. Сглаживание данных, как правило, осуществляется путём параметрической или непараметрической регрессии. В случае параметрической регрессии необходимы априорные знания о форме уравнения регрессии. Большинство исследуемых данных, однако, невозможно параметризовать. С этой точки зрения непараметрическая и полупараметрическая регрессии представляются лучшим подходом к решению задачи сглаживания. Целью исследования ставилось разработка и реализация алгоритма быстрого сглаживания двумерных данных. Для достижения этой цели были проанализированы предыдущие работы в данной области и разработан свой подход, улучшающий предыдущие. В результате, в данной работе представлен алгоритм, который быстро и с минимальным потреблением памяти очищает данные от «шума» и «несущественных» частей. Для подтверждения «эффективности» алгоритма проведены сравнения с другими общепризнанными подходами на смоделированных и реальных данных. Результаты этих сравнений также приведены в статье.

Ключевые слова: непараметрическая регрессия, двумерное сглаживание, штрафные сплайны, сглаживающие сплайны, скользящий контроль, двумерное дискретное косинусное преобразование

1. Постановка задачи

При анализе некоторого реального процесса исходные данные зашумлены, из-за чего необходимо «сглаживание». Под сглаживанием понимается попытка фильтрации шума или несущественных фрагментов при сохранении наиболее важных свойств структуры данных. Рассмотрим следующую модель

$$y = \hat{y} + \varepsilon, \tag{1}$$

где ε — гауссов белый шум. Предполагается, что функция $f(x)$ должна быть гладкой, т.е. иметь непрерывные производные до некоторого порядка. Сглаживание данных, как правило, осуществляется путём параметрической или непараметрической регрессии. В случае параметрической регрессии необходимы некоторые априорные знания о форме уравнения регрессии, которое достаточно хорошо описывало бы исходные данные. Большинство наблюдаемых данных, однако, невозможно параметризовать с точки зрения задания функции $f(x)$ в аналитическом виде. С этой точки зрения непараметрическая и полупараметрическая регрессии являются лучшим подходом к решению задачи (1). Одним из классических подходов к сглаживанию является использование различных модификаций метода наименьших квадратов со штрафом. Он впервые был продемонстрирован вначале 1920-х в работе [1] и подробно разобран в книге [2]. Эта техника заключается в минимизации некоторого функционала, который уравновешивает «адекватность» и «гладкость» оценки

$$F(\hat{y}) = RSS + \lambda \cdot P(\hat{y}) = ||\hat{y} - y||^2 + \lambda \cdot P(\hat{y}), \tag{2}$$

где $||\cdot||$ — евклидова норма. Параметр λ является вещественным положительным числом, контролирующим гладкость решения: при его возрастании гладкость $\hat{y}$ также растёт. Когда штрафная функция записана в виде интеграла квадрата производной p-порядка, регрессия называется сглаживающим сплайном [1, 3, 4]. Другим простым и эффективным подходом к решению задачи (1)является квадратичный вид штрафной функции в следующей форме [5]:

$$P(\hat{y}) = ||D\hat{y}||^2, \tag{3}$$

где D — трёхдиагональная матрица следующего вида

$$\begin{bmatrix} -1 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -1 \end{bmatrix}.$$

Данная статья является продолжением исследований автора, затронутых в статьях [6, 7], а также некоторые идеи почерпнуты из статьи [8].

2. Одномерное сглаживание

Пусть имеются равноотстоящие точки $\{x_i\}_{1 \leqslant i \leqslant n}$ и значения функции отклика в этих точках следующего вида

$$y_i = f(x_i) + \varepsilon_i, \tag{4}$$

где $\varepsilon_i \sim N(0, \sigma^2)$. Пусть $\hat{y}$ является оценкой функции $f(x_i)$. Тогда минимизация выражения (2) приводит к следующему уравнению

$$\hat{y} = H(\lambda) \cdot y, \tag{5}$$

где $H(\lambda) = (I + \lambda \cdot D^T D)^{-1}$ представляет собой проекционную матрицу, а λ — это сглаживающий параметр. Сглаживающий параметр выбирается путём минимизации функционала следующего вида

$$GCV(\lambda) = \frac{RSS(\lambda)/n}{(1 - \mathrm{Tr}(H(\lambda))/n)^2}. \tag{6}$$

Такой подход называется методом скользящего контроля (crossvalidation). При равноотстоящих наблюдениях можно использовать свойства матрицы D, с помощью которых можно упростить вычисления GCV. Это свойство заключается в возможности разложить матрицу D в следующее произведение $U\Gamma U^T$, где матрица U является унитарной и представляет собой дискретное косинусное преобразование [9]. Тогда RSS можно переписать в следующем виде:

$$RSS = ||\hat{y} - y||^2 = ||H(\lambda) \cdot y - y||^2 = \left\| \left(\left(I + \lambda \cdot D^T D\right)^{-1} - I \right) \cdot y \right\|^2 =$$
$$= \left\| \left(U \cdot \left(I + \lambda \cdot \Gamma^2\right)^{-1} - I \right) \cdot U^T \cdot y \right\|^2 = \sum_i \left(\frac{1}{1 + \lambda\gamma_i^2} - 1 \right)^2 \cdot DCT_i^2(y).$$

В таком случае функционал (6) принимает более удобный для вычислений вид:

$$GCV(\lambda) = \frac{n \cdot \sum_i \left(\frac{1}{1+\lambda\gamma_i^2} - 1\right)^2 \cdot DCT_i^2(y)}{n - \sum_i \left(\frac{1}{1+\lambda\gamma_i^2}\right)^2}. \tag{7}$$

3. Двумерное сглаживание

Пусть имеются равномерная сетка $\{(x_{1,i}, x_{2,j})\}_{1\leqslant i\leqslant n_1, 1\leqslant j\leqslant n_2}$ и значения функции отклика в узлах этой сетки следующего вида

$$y_{i,j} = f(x_{1,i}, x_{2,j}) + \varepsilon_{i,j}, \tag{8}$$

где $\varepsilon_{i,j} \sim N(0, \sigma^2)$. В таком случае можно представить значения функции отклика в виде матрицы Y, в которой $y_{i,j}$ является элементом i-й строки j-го столбца. Тогда сглаженные значения будем обозначать как $\hat{Y}$. Введём операцию vec, которая заключается в записи матрицы в виде вектор-столбца путём выкладывания столбцов матрицы друг за другом. Очевидно, что оценка $vec(\hat{Y})$ имеет следующий вид:

$$vec(\hat{Y}) = (H_{x_2} \otimes H_{x_1}) \cdot vec(Y) = H_{x_2,x_1} \cdot vec(Y), \tag{9}$$

где H_{x_1}, H_{x_2} — проекционные матрицы соответствующего измерения. Очевидно, что эти проекционные матрицы имеют следующий вид

$$H_{x_i} = \left(I_{n_i} + \lambda_i D_{n_i}^T D_{n_i}\right)^{-1}, \quad i = 1, 2. \tag{10}$$

Применяя вышеизложенный подход и свойства тензорного произведения [10], выражение (9) можно упростить следующим образом

$$\begin{aligned}
\hat{y} = (H_{x_2} \otimes H_{x_1}) \cdot y &= \left(I_{n_2} + \lambda_2 D_{n_2}^T D_{n_2}\right)^{-1} \otimes \left(I_{n_1} + \lambda_1 D_{n_1}^T D_{n_1}\right)^{-1} \cdot y = \\
&= U_{x_2} \cdot \left(\frac{1}{1+\lambda_1\gamma_{x_1}^2}\right) \cdot U_{x_2}^T \otimes U_{x_2} \cdot \left(\frac{1}{1+\lambda_1\gamma_{x_1}^2}\right) \cdot U_{x_2}^T \cdot y = \\
= U_{x_2} \otimes U_{x_1} \cdot \left(\frac{1}{1+\lambda_1\gamma_{x_1}^2}\right) &\otimes \left(\frac{1}{1+\lambda_1\gamma_{x_1}^2}\right) \cdot U_{x_2}^T \otimes U_{x_1}^T \cdot y = U_{x_2,x_1} \cdot \Gamma_{x_2,x_1} \cdot U_{x_2,x_1}^T \cdot y.
\end{aligned}$$

Для автоматического поиска оптимальных значений λ_1 и λ_2 предлагаем также использовать GCV адаптированный для двумерного случая:

$$GCV(\lambda_1, \lambda_2) = \frac{RSS/n}{(1 - \mathrm{Tr}(H_{x_2,x_1})/n^2)}. \tag{11}$$

Принимая во внимание свойство следа тензорного произведения матриц [10], получаем $\mathrm{Tr}(H_{x_2,x_1}) = \sum \frac{1}{1+\lambda_1\gamma_{x_1}^2} \cdot \sum \frac{1}{1+\lambda_2\gamma_{x_2}^2}$. Очевидно, что основным затратным местом в части вычислений является расчёт RSS, так как требуется вычисление $\hat{y}$ для всех комбинаций λ_1 и λ_2. Данный расчёт можно упростить следующим образом:

$$\begin{aligned}
RSS = \|\hat{y} - y\|^2 = \|H_{x_2,x_1} \cdot y - y\|^2 &= \|(H_{x_2,x_1} - I_n) \cdot y\|^2 = \\
&= \left\|U_{x_2,x_1} \cdot (\Gamma_{x_2,x_1} - I_n) \cdot U_{x_2,x_1}^T \cdot y\right\|^2 = \\
= \left(U_{x_2,x_1} \cdot (\Gamma_{x_2,x_1} - I_n) \cdot U_{x_2,x_1}^T \cdot y\right)^T &\cdot \left(U_{x_2,x_1} \cdot (\Gamma_{x_2,x_1} - I_n) \cdot U_{x_2,x_1}^T \cdot y\right) =
\end{aligned}$$

$$= (DCT_2 \cdot y)^T \cdot (\Gamma_{x_2,x_1} - I_n)^2 \cdot DCT_2 \cdot y = \sum (\gamma_{x_2,x_1} - 1)^2 \cdot (DCT_2 \cdot y)^2 .$$

Здесь DCT_2 — это двумерный аналог дискретного косинусного преобразования. Из упрощённой формулы видно, что преобразование необходимо выполнить один раз, а меняются только значения γ_{x_2,x_1} в зависимости значений λ_1 и λ_2. Данный подход реализован на языке R. Для демонстрации преимуществ изложенного подхода выполнены численные эксперименты: на искусственно сгенерированных данных и на реальных данных.

4. Численный эксперимент

Для проведения эксперимента смоделирована следующая задача: взята функция $\sin(2\pi(x-0,5)^3)\cdot\cos(4\pi y)$ и зашумлена случайными значениями из нормального распределения $N(0,\ 0,2^2)$ (рис. 1). Сглаживание проводилось изложенным подходом и при помощи пакета MGCV [11], в котором реализовано сглаживание штрафными сплайнами, в том числе и для многомерного случая с использованием тензорного произведения базовых функций. Ниже приведена таблица с результатами сглаживания на различных сетках.

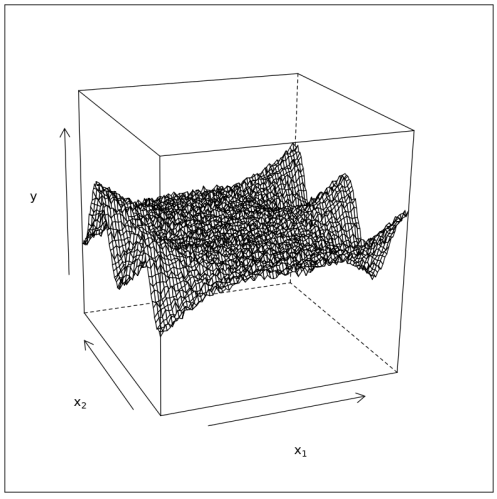

Рис. 1. Функция $\sin(2\pi(x-0,5)^3)\cdot\cos(4\pi y)$ **с шумом**

Таблица 1

Результаты сглаживания смоделированной задачи на разных сетках

Параметр	P-сплайны с ДКП	GAM с 10^2 узлами	GAM с 20^2 узлами
RSS	9,488243	11,72485	9,87163
MSE	0,001483	0,001832	0,00154
Корреляция с истин. значениями	0,9993394	0,996919	0,9991624
Время оцен. (с)	1,941	10,237	29,875

Для демонстрации практического применения вышеизложенного подхода взяты данные о смертности в России. Данные взяты из открытого источника [12] и содержат наблюдения за 1959–2010 годы по возрастам 0–110. Оценивание проводилось на части данных, которая относится к старшим возрастам (50–101, рис. 2). К этим годам выборка достаточно сильно уменьшается, в результате чего наблюдения содержат ошибки и выбросы, из-за которых не видно общей картины происходящих процессов. Таким образом, анализируемые данные представляют собой равномерно расположенные значения коэффициентов смертности на сетке размерности 52×52. Сглаживание проводилось изложенным подходом, пакетом MGCV и параметрической моделью Ли-Картера, которая стала классической при оценивании двумерной поверхности смертности. Далее приведена таблица с результатами оценивания.

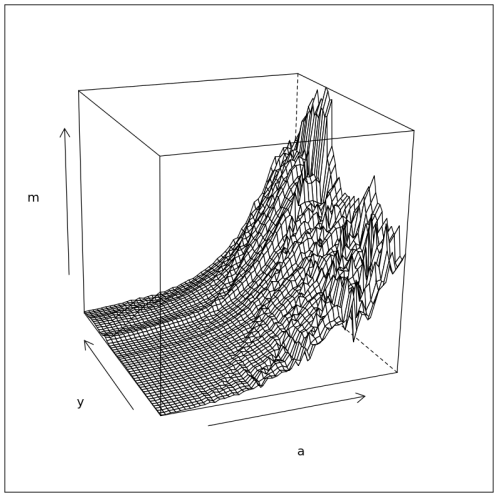

Рис. 2. Коэффициенты смертности населения России старше 50 лет с 1959 г. по 2010 г.

Таблица 2

Результаты сглаживания двумерной поверхности смертности России

Параметр	P-сплайны с ДКП	Модель Ли-Картера	GAM с 12^2 узлами
RSS	$0,21637$	$18,5092$	$0,41395$
MSE	$0,0000905$	$0,0077379$	$0,0001731$
Время оцен. (с)	$0,49$	$1,194$	$4,185$

5. Заключение

Из приведённых результатов очевидно, что подход описанный в данной статье более эффективен, так как имеет большую скорость сглаживания и малое потребление памяти. Также хочется отметить, что при росте выборки скорость оценивания растёт незначительно при сохранении качества оценки. Результатами проделанной работы являются:

1. получены выражения для двумерного случая с учётом двух параметров сглаживания;
2. полученный подход реализован на языке R;
3. выполнено сравнение подхода с другими аналогичными подходами и моделями.

В последующих работах предполагается рассмотрение следующих возможностей:

– расширение метода на многомерный случай;
– использование других распространённых критериев выбора сглаживающих параметров, например, BIC или AIC;
– использование более быстрого метода поиска минимального значения функционала GCV.

Литература

1. *Whittaker E. T.* On a New Method of Graduation // Proceedings of the Edinburgh Mathematical Society. — 1923. — Vol. 41. — Pp. 62–75.
2. *Wahba G.* Spline Models for Observational Data. — Society for Industrial Mathematics, 1990. — ISBN 9780898712445.
3. *Schoenberg I. J.* Spline Functions and the Problem of Graduation // Proceedings of the National Academy of Sciences of the United States of America. — 1964. — Vol. 52. — Pp. 947–950.
4. *Takezawa K.* Introduction to Nonparametric Regression. — Wiley & Sons, Inc., 2005. — ISBN 9780471745839.
5. *Weinert H. L.* Efficient Computation for Whittaker–Henderson Smoothing // Computational Statistics & Data Analysis. — 2007. — Vol. 52. — Pp. 959–974.
6. *Щетинин Е. Ю., Любин П. Г.* Робастный алгоритм построения сглаживающих сплайнов // Научное обозрение. — 2015. — Т. 1. — С. 86–94.
7. *Любин П. Г., Щетинин Е. Ю.* Стохастические модели сглаживания и прогнозирования коэффициентов смертности // Научное обозрение. — 2015. — Т. 18. — С. 147–155.
8. *Xiao L., Li Y., Ruppert D.* Fast Bivariate P-splines: the Sandwich Smoother // Journal of the Royal Statistical Society. — 2013. — Vol. 75. — P. 577–599.
9. *Garcia D.* Robust Smoothing of Gridded Data in One and Higher Dimensions with Missing Values // Computational Statistics & Data Analysis. — 2010. — Vol. 54. — P. 1167–1178.
10. *Seber G.* A Matrix Handbook for Statisticians. — Wiley-Interscience, 2007. — ISBN 9780471748694.
11. *Wood S.* MGCV: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation. — R package version 1.8.10. `https://cran.r-project.org/web/packages/mgcv/index.html`, r package version 1.8.10.
12. The Human Mortality Database. — `http://www.mortality.org/`.

UDC 519.234

On a Method of Two-Dimensional Smoothing

P. G. Lyubin

Moscow State Technology University "STANKIN", Moscow, Russia

Regression analysis has the task of finding a functional relationship between the observed values the studied process. The raw data is the realization of a random variable, it is therefore considered dependent on the expectation. This problem can be solved by "smoothing" the raw data. Smoothing is the process of removing the noise and insignificant fragments while preserving the most important properties of the data structure. It is similar to finding the expectation of data. Data smoothing usually attained by parametric and nonparametric regression. The nonparametric regression requires a prior knowledge of the regression equation form. However,

most of the investigated data cannot be parameterized simply. From this point of view, nonparametric and semiparametric regression represents the best approach to smoothing data. The aim of the research is development and implementation of the fast smoothing algorithm of two-dimensional data. To achieve this aim previous works in this area have been analyzed and its own approach has been developed, improving the previous ones. As a result, this paper presents the algorithm that quickly and with minimal memory consumption cleanses the data from the "noise" and "insignificant" parts. To confirm the "efficiency" of the algorithm the comparisons with other generally accepted approaches were carried out on simulated and real data with other generally accepted approaches. The results of these comparisons are also shown in the paper.

Key words and phrases: nonparametric regression, two-dimensional estimation, penalized splines, smoothing splines, cross-validation, discrete cosine transform

References

1. E. T. Whittaker, On a New Method of Graduation, Proceedings of the Edinburgh Mathematical Society 41 (1923) 62–75.
2. G. Wahba, Spline Models for Observational Data, Society for Industrial Mathematics, 1990.
3. I. J. Schoenberg, Spline Functions and the Problem of Graduation, Proceedings of the National Academy of Sciences of the United States of America 52 (1964) 947–950.
4. K. Takezawa, Introduction to Nonparametric Regression, Wiley & Sons, Inc., 2005.
5. H. L. Weinert, Efficient Computation for Whittaker-Henderson Smoothing, Computational Statistics & Data Analysis 52 (2007) 959–974.
6. E. Y. Shetinin, P. G. Lyubin, Robust Smoothing with Splines, Science Review 1 (2015) 86–94, in Russian.
7. P. G. Lyubin, E. Y. Shetinin, Stochastic Models of Mortality Estimation, Science Review 18 (2015) 147–155, in Russian.
8. L. Xiao, Y. Li, D. Ruppert, Fast Bivariate P-splines: the Sandwich Smoother, Journal of the Royal Statistical Society 75 (2013) 577–599.
9. D. Garcia, Robust Smoothing of Gridded Data in One and Higher Dimensions with Missing Values, Computational Statistics & Data Analysis 54 (2010) 1167–1178.
10. G. Seber, A Matrix Handbook for Statisticians, Wiley-Interscience, 2007.
11. S. Wood, MGCV: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation, r package version 1.8.10.
URL `https://cran.r-project.org/web/packages/mgcv/index.html`
12. The human mortality database, last visited on 25.02.2016.
URL `http://www.mortality.org/`