Научный вестник НГТУ. – 2013. – № 2(51)

ОБРАБОТКА ИНФОРМАЦИИ

УДК 519.23

Исследование алгоритмов выбора оптимальных координат узловых точек в полупараметрических моделях штрафных сплайнов *

В.И. ДЕНИСОВ, В.С. ТИМОФЕЕВ, А.В. ФАДДЕЕНКОВ

Предложены модификации методов построения регрессионных зависимостей, базирующихся на полупараметрических моделях. Разработаны новые алгоритмы выбора оптимальных координат узловых точек, основанные на критериях точности и индивидуальной информативности наблюдений. Приведены результаты сравнительных исследований разработанных алгоритмов при различных вариантах засорения исходных данных, проведенных с использованием вычислительных экспериментов.

Ключевые слова: параметрические и непараметрические методы, полупараметрическая регрессия, модели штрафных сплайнов, базисные функции, планирование экспериментов, метод наименьших квадратов.

ВВЕДЕНИЕ

В последнее время большую популярность в статистическом анализе данных стало приобретать полупараметрическое моделирование. Главная причина такого повышенного внимания объясняется тем, что подобная модель является компромиссным решением между двумя крайностями: полностью параметрическим и полностью непараметрическим моделированием.

В первом случае для объяснения выборочных данных используется параметрическая модель, при этом априори известно, что число параметров конечно и распределение ошибки также принадлежит семействам с конечным числом параметров. В качестве оценки может быть использована, например, оценка максимального правдоподобия. Однако неверная спецификация некоторых компонент модели может привести к смещению оценок, и выводы, полученные на основе оцененной модели, могут быть ошибочными.

В противоположность этому в полностью непараметрических моделях заранее ничего не известно о существующих взаимосвязях в данных и ошибках, за исключением, возможно, некоторых свойств регулярности и формы, таких как непрерывная дифференцируемость или вогнутость. Непараметрические модели дают максимальную гибкость, сводя к минимуму вероятность неправильно специфицировать модель [1–3]. С другой стороны, непараметрическое оценивание требует большого количества исходных данных, и в малых выборках получаются довольно неточные оценки. Особенно ярко это проявляется в моделях большой размерности, где точность оценок падает по мере добавления новых переменных.

Компромиссным решением между непараметрическим и параметрическим подходами являются полупараметрические модели. Они сохраняют до некоторой степени гибкость непараметрической модели и гораздо менее подвержены неправильной спецификации по сравнению с полностью параметрическими моделями [1, 3]. В то же время параметрическую компоненту полупараметрической модели можно оценить с точностью, сравнимой с достигаемой при использовании верной полностью параметрической модели.

В данной работе авторами сделана попытка модификации некоторых полупараметрических алгоритмов построения регрессионных зависимостей на основе идей, заимствованных из теории планирования экспериментов.

---

* *Статья получена 27 марта 2013 г.*
Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта №13-07-00299 а

1. СПЛАЙНОВАЯ ПОЛУПАРАМЕТРИЧЕСКАЯ РЕГРЕССИЯ

Рассмотрим одну из известных полупараметрических регрессионных моделей следующего вида [3]:

$$y_i = \theta_1 x_{i1} + \ldots + \theta_m x_{im} + \beta_1 f_{i1} + \beta_2 f_{i2} + \ldots + \beta_k f_{ik} + \varepsilon_i, \quad (1)$$

где y_i – значение отклика в i-м наблюдении ($i = 1,2,\ldots,N$); x_{ij} – значение j-го регрессора в i-м наблюдении ($j = 1,2,\ldots,m$); $\theta_1,\ldots,\theta_m$, $\beta_1,\ldots,\beta_k$ – неизвестные параметры; $\theta_1 x_{i1} + \ldots + \theta_m x_{im}$ – параметрическая часть модели; $\beta_1 f_{i1} + \beta_2 f_{i2} + \ldots + \beta_k f_{ik}$ – непараметрическая часть; $f_{i1}, f_{i2},\ldots, f_{ik}$ – значения базисных функций в i-м наблюдении; ε_i – случайная ошибка в i-м наблюдении (предполагается, что все ошибки независимы и имеют одинаковое распределение с нулевым средним и дисперсией σ_ε^2: $\varepsilon_i \sim (0, \sigma_\varepsilon^2)$, $i = 1,2,\ldots,N$).

В качестве базисных будем использовать функции

$$f_{ij} = f_j(x_i) = (x_i - b_j)_+^p = \begin{cases} (x_i - b_j)^p, при\ (x_i - b_j) > 0 \\ 0, \qquad при\ (x_i - b_j) \le 0 \end{cases}, \quad (2)$$

где $b_j \in [a,b]$ – узловые точки ($j = 1,2,\ldots,k$), p – некоторая положительная целая константа. Непараметрическая часть модели (1) с базисными функциями (2) представляет собой сплайн порядка p, а саму модель в этом случае называют сплайновой регрессией.

В матричном виде уравнение (1) может быть представлено следующим образом:

$$Y = \tilde{X}\tilde{\Theta} + \mathrm{E}, \quad (3)$$

где

$$Y = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix}^T,$$

$$\tilde{X} = \begin{bmatrix} x_{11} & \cdots & x_{1m} & (x_1 - b_1)_+^p & \cdots & (x_1 - b_k)_+^p \\ x_{21} & \cdots & x_{2m} & (x_2 - b_1)_+^p & \cdots & (x_2 - b_k)_+^p \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nm} & (x_N - b_1)_+^p & \cdots & (x_N - b_k)_+^p \end{bmatrix}.$$

$$\tilde{\Theta} = \begin{bmatrix} \theta_1 & \cdots & \theta_m & \beta_1 & \cdots & \beta_k \end{bmatrix}^T, \ \mathrm{E} = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_N \end{bmatrix}^T.$$

Оценивание неизвестных параметров этой модели проводится по классической методике регрессионного анализа, например, с использованием метода наименьших квадратов:

$$\hat{\tilde{\Theta}} = \left(\tilde{X}^T \tilde{X}\right)^{-1} \tilde{X}^T Y. \quad (4)$$

Естественно, что качество получаемой модели зависит от количества базисных точек и их координат. Однако рост числа узлов может приводить к излишней подгонке линии регрессии под исходные данные. Традиционным решением этой проблемы считается переход к так называемым «штрафным сплайнам» [3]. Идея этого метода заключается в том, что для снижения излишнего влияния непараметрической части, на ее параметры налагается ограничение (штраф) и вектор оценок параметров вычисляется следующим образом:

$$\hat{\tilde{\Theta}} = \left(\tilde{X}^T \tilde{X} + \lambda^2 D\right)^{-1} \tilde{X}^T Y, \quad (5)$$

где λ^2 – параметр сглаживания, D – $(m+k)\times(m+k)$-матрица штрафа:

$$D=\begin{bmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \cdots & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0_{(m\times m)} & 0_{(m\times k)} \\ 0_{(k\times m)} & I_{(k\times k)} \end{bmatrix}.$$

При $\lambda^2 = 0$ сглаживание непараметрической части не проводится и оценка (5) совпадает с обычной МНК-оценкой (4). Чрезмерное же увеличение параметра сглаживания ($\lambda^2 \to +\infty$) приводит к тому, что регрессионная модель (3) вырождается в модель, состоящую только из параметрической части. В связи с этим выбору величины параметра сглаживания следует уделять особое внимание.

Использование для этой цели метода максимального правдоподобия приводит к необходимости фиксации закона распределения случайных компонент модели (3) [7]. Наиболее часто предполагают нормальность распределения, однако на практике такое предположение далеко не всегда будет выполняться. В связи с этим в данной работе рассматриваются альтернативные способы оценивания, основанные на критериях кросс-валидации, обобщенной кросс-валидации, а также критерий Акаике.

Критерий кросс-валидации:

$$CV = \sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{1 - S_{\lambda,ii}}\right)^2, \tag{6}$$

где $S_{\lambda,ii}$ – i-й диагональный элемент матрицы $S_\lambda = \tilde{X}\left(\tilde{X}^T\tilde{X} + \lambda^2 D\right)^{-1}\tilde{X}^T$.

Критерий обобщенной кросс-валидации:

$$GCV = \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\left(1 - N^{-1}tr(S_\lambda)\right)^2}, \tag{7}$$

Критерий Акаике:

$$AIC = \log\left\{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2\right\} + \frac{2tr(S_\lambda)}{N}. \tag{8}$$

Определение оптимального значения для параметра сглаживания сводится к задаче одномерной минимизации статистик (6), (7) или (8). Отдельные результаты исследования качества восстановления регрессионной зависимости с использованием модели (3), (5) при различных λ^2, k можно найти в [4].

Выбор координат узловых точек также оказывает существенное влияние на качество модели. Наиболее распространенным подходом для определения этих координат является следующий несложный алгоритм [3].

Алгоритм 1 – узловые точки b_j ($j = 1,2,...,k$) определяются как выборочные квантили порядка $\frac{j}{k+1}$ рассматриваемой независимой переменной x. В случае равномерного распределения наблюдений на отрезке $[a,b]$ и отсутствии повторных наблюдений координаты узловых точек могут быть определены следующим образом:

$$b_j = a + (b-a)\left(\frac{j}{k+1}\right).$$

Очевидно, что в этом случае при выборе координат узловых точек не учитываются структура модели и имеющиеся значения зависимой переменной. Вопросы использования этой дополнительной информации применительно к модели (3-4) рассматривались в работе [5]. Предложенные ниже три алгоритма являются логическим обобщением подходов, рассмотренных в [5].

Алгоритм 2 – выбор узловых точек с использованием информационной матрицы модели (3). Начальное приближение для координат узловых точек b_j ($j = 1,2,\ldots,k$) определяется последовательно за k шагов. На каждом j-м шаге определяются координаты одной узловой точки, как решение оптимизационной задачи

$$b_j = Arg \min_{b_j \in [a,b]} \det(M^{-1}(j)),$$

где

$$M(j) = \frac{1}{N}\tilde{X}^T(j)\tilde{X}(j),$$

$$\tilde{X}(j) = \begin{bmatrix} x_{11} & \cdots & x_{1m} & (x_1 - b_1)_+^p & \cdots & (x_1 - b_j)_+^p \\ x_{21} & \cdots & x_{2m} & (x_2 - b_1)_+^p & \cdots & (x_2 - b_j)_+^p \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nm} & (x_N - b_1)_+^p & \cdots & (x_N - b_j)_+^p \end{bmatrix}.$$

После определения начального приближения проводится дальнейшая оптимизация по координатам всех узловых точек, то есть решается задача

$$\min_{b_1,\cdots,b_k \in [a,b]} \det(M^{-1}), \tag{9}$$

где $M = \frac{1}{N}\tilde{X}^T\tilde{X}$ – информационная матрица.

Алгоритм 3 – выбор узловых точек на основе остаточной суммы квадратов. Начальное приближение для координат j-й узловой точки ($j = 1,2,\ldots,k$) определяются как решение оптимизационной задачи

$$b_j = Arg \min_{b_j \in [a,b]} ESS(j),$$

где $ESS(j) = e(j)^T e(j)$, $e(j) = Y - \tilde{X}(j)\left[\tilde{X}(j)^T\tilde{X}(j)\right]^{-1}\tilde{X}(j)^T Y$.

После определения начального приближения проводится дальнейшая оптимизация как решение задачи

$$\min_{b_1,\cdots,b_k\in[a,b]} ESS, \tag{10}$$

где $ESS = e^T e$, $e = Y - \tilde{X}\left[\tilde{X}^T\tilde{X}\right]^{-1}\tilde{X}^T Y$.

Алгоритм 4 – выбор узловых точек на основе ковариационной матрицы оценок параметров модели (3). Как и в предыдущих алгоритмах, на первом этапе проводится определение начального приближения:

$$b_j = Arg \min_{b_j\in[a,b]} \det(D(j)),$$

где $D(j) = \hat{\sigma}_\varepsilon^2(j)\left[\tilde{X}(j)^T\tilde{X}(j)\right]^{-1}$, $\hat{\sigma}_\varepsilon^2(j) = \dfrac{ESS(j)}{N-m-j}$.

Дальнейшая оптимизация проводится как решение задачи

$$\min_{b_1,\cdots,b_k\in[a,b]} \det D, \tag{11}$$

где $D = \hat{\sigma}_\varepsilon^2\left[\tilde{X}^T\tilde{X}\right]^{-1}$, $\hat{\sigma}_\varepsilon^2 = \dfrac{ESS}{N-m-k}$.

Комбинация различных методов выбора координат узловых точек и методов оценивания параметров модели (3) дает широкий спектр итоговых алгоритмов построения искомой линии регрессии.

2. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Для сравнительного исследования точности рассмотренных методов построения линии регрессии был проведен ряд вычислительных экспериментов. В качестве тестовой (истинной) была использована модель следующего вида:

$$y_i = y_i^0 + \varepsilon_i = \beta_0 + x_i - 1.8x_i^2 + x_i^3 + \beta_1(x_i - b_1)_+ + \beta_2(x_i - b_2)_+ + \varepsilon_i. \tag{12}$$

При моделировании отклика значения независимой переменной x равномерно варьировались на отрезке $[0,1]$. При каждой реализации набора исходных данных в функции (12) параметры, $j = 0,1,2$ определялись псевдослучайным датчиком как случайные величины, равномерно распределенные на отрезках $[1,3]$, $[0.25, 0.35]$, $[0.6, 0.7]$ соответственно. Случайная ошибка для каждого наблюдения генерировалась из предположений о нормальном распределении: $\varepsilon_i \sim N(0,\sigma_\varepsilon^2)$. Дисперсия ошибки σ_ε^2 выбиралась таким образом, чтобы величина уровня шума

$$\rho = \frac{\sigma_\varepsilon}{c}100\ \%,$$

была равна наперед заданному значению. Уровень шума ρ введен в [ ] и определяется как отношение «шум»/«сигнал» в процентах, здесь σ_ε^2 – дисперсия ошибки ε; $c^2 = \dfrac{1}{N-1}\sum_{i=1}^{N}(y_i^0 - \bar{y}^0)^2$ (y_i^0 – не зашумленные измерения отклика).

На рис. 1 представлены примеры различных серий наблюдений, в каждой из которых параметры модели определялись по описанным выше правилам с уровнем шума $\rho = 10\ \%$.

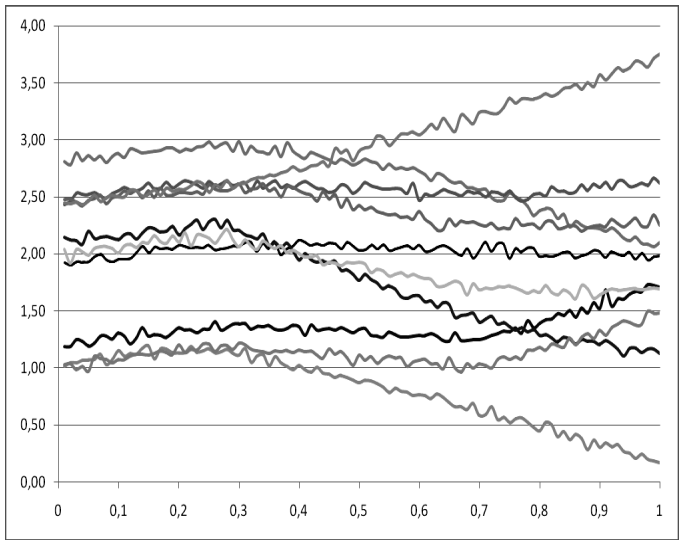

Рис. 1. Примеры различных серий наблюдений, сгенерированных для модели (12)

Оценивание точности построенной линии регрессии проводилось с двух точек зрения: с точки зрения точности воспроизведения исходных наблюдений и с точки зрения точности соответствия истинной модели.

В качестве критерия точности воспроизведения исходных наблюдений использовалась сумма квадратов остатков

$$ESS = (Y - \hat{Y})^T (Y - \hat{Y}), \tag{13}$$

где Y – вектор исходных данных, $\hat{Y}$ – вектор оценок наблюдений, построенных по модели.

В качестве критерия точности соответствия истинной модели использовалась аналогичная сумма квадратов:

$$ESS_{\text{ист}} = (Y_{\text{ист}} - \hat{Y})^T (Y_{\text{ист}} - \hat{Y}), \tag{14}$$

где $Y_{\text{ист}}$ – вектор наблюдений, построенный по истинной модели (12) при полном отсутствии случайных ошибок.

Оценивание отклика проводилось по модели (1-2) при $p = 1$:

$$y_i = \theta_0 + \theta_1 x_i + \sum_{j=1}^{k} \beta_j (x_i - b_j)_+ + \varepsilon_i . \tag{15}$$

Рассматривались модели с числом базисных функций k от 1 до 10.

Для каждого значения уровня шума ρ проводилась серия генераций наборов данных по модели (12). Далее по моделям (15) с количествами базисных функций $k = 1,2,\ldots,10$ проводилось оценивание значений отклика с использованием различных алгоритмов. По результатам оценивания определялись значения сумм квадратов (13) и (14). После окончания серии экспериментов значения сумм квадратов, соответствующих каждой комбинации «число узлов – алгоритм оценивания» усреднялись. Для удобства дальнейшего изложения введем ряд

обозначений для усредненных сумм квадратов: $MS1$, $MS2$, $MS3$ – усредненные по серии наблюдений суммы квадратов (13), полученные при использовании оценок параметров (4) и алгоритмов 2, 3, 4 выбора координат узловых точек соответственно; $MS4$, $MS5$, $MS6$, $MS7$ – усредненные по серии наблюдений суммы квадратов (13), полученные при использовании оценок параметров (5) и алгоритмов 1, 2, 3, 4 соответственно. Аналогичным образом обозначим усредненные суммы квадратов (14): $MST1$, $MST2$, $MST3$, $MST4$, $MST5$, $MST6$, $MST7$.

В ходе вычислительных экспериментов генерация исходных данных проводилась при дисперсиях ошибок, соответствующих уровням шума ρ от 5 % до 25 %. Усреднение результатов проводилось по сериям из 300 наборов данных. Итоги этих экспериментов позволили сделать вывод, что при малом уровне шума (рис. 2) точность алгоритмов как с точки зрения воспроизведения исходных наблюдений, так и с точки зрения соответствия истинной модели меняется одинаковым образом. В частности, для всех алгоритмов справедливо повышение точности с увеличением количества базисных функций, при этом в среднем наилучшие результаты наблюдаются у алгоритмов, основанных на использовании обычных МНК-оценок параметров (4). Введение параметра сглаживания при малом уровне шума не влечет за собой улучшения качества модели.

С ростом уровня шума картина принципиально меняется (рис. 3). В этом случае с точки зрения точности воспроизведения исходных данных лидирующие позиции занимают модели, основанные на оценках (4) и алгоритмах 3 и 4 выбора координат узловых точек.

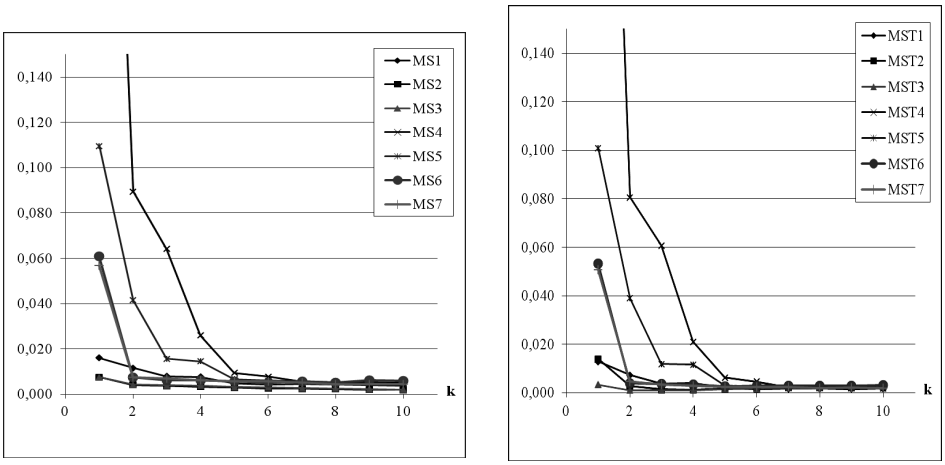

Рис. 2. Усредненные суммы квадратов при уровне шума $\rho = 5$ %

Однако с точки зрения соответствия истинной зависимости, качество этих моделей с ростом числа базисных функций постепенно ухудшается. С этой точки зрения, при достаточном количестве базисных функций (в данном случае при $k \geq 5$), наилучшие результаты демонстрируют модели, основанные на использовании оценок (5).

Следует обратить особое внимание на модель, основанную на оценках (4) и алгоритме выбора координат узловых точек номер 2 (на графиках этой модели соответствует номер 1). Данная модель с точки зрения точности занимает промежуточное значение между моделями, использующими и не использующими сглаживание непараметрической части. С позиции точности воспроизведения исходных данных эта модель уступает моделям с номерами 2 и 3, однако в среднем оказывается точнее всех моделей, основанных на оценках (5). С позиции точности соответствия истинной зависимости в среднем эта модель хоть и уступает моделям, основанным на оценках (5), но показывает значительно лучшие результаты, чем все остальные модели, основанные на оценках (4).

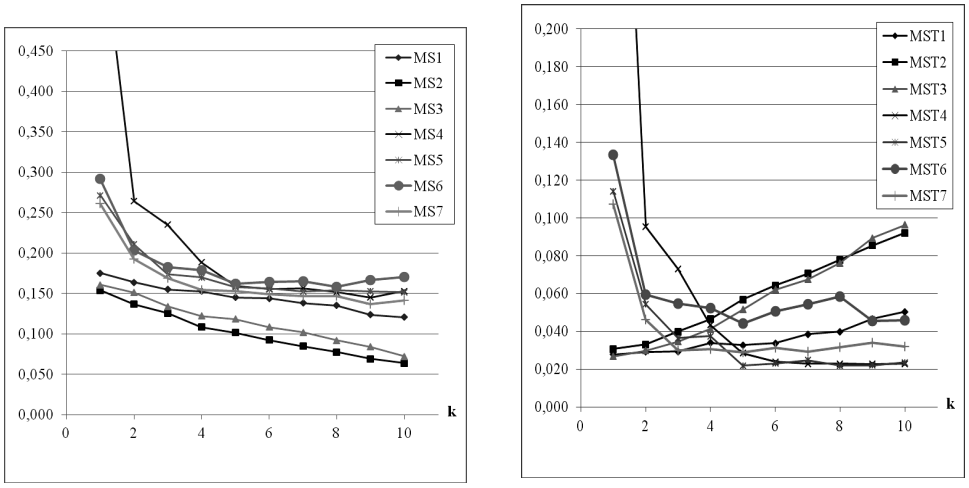

Рис. 3. Усредненные суммы квадратов при уровне шума $\rho = 20\ \%$

Для анализа устойчивости рассматриваемых методов построения линии регрессии был проведен ряд вычислительных экспериментов с искусственными нарушениями предположений модели (12). В частности, в ходе генерации случайной ошибки в выборку включались аномальные наблюдения – «выбросы». Доля выбросов в исходных данных составляла 3 %. Роль выбросов исполняли нормально распределенные случайные величины с десятикратно увеличенной дисперсией ($10\sigma_\varepsilon^2$). Координаты точек, содержащих выбросы, при каждой генерации исходных данных определялись случайно. Усредненные результаты испытаний представлены на рис. 4 и 5.

С точки зрения точности соответствия исходным данным картина при наличии выбросов принципиально не меняется. Лидерами, как и ранее, являются модели 2 и 3, однако разрыв между ними становится более существенным.

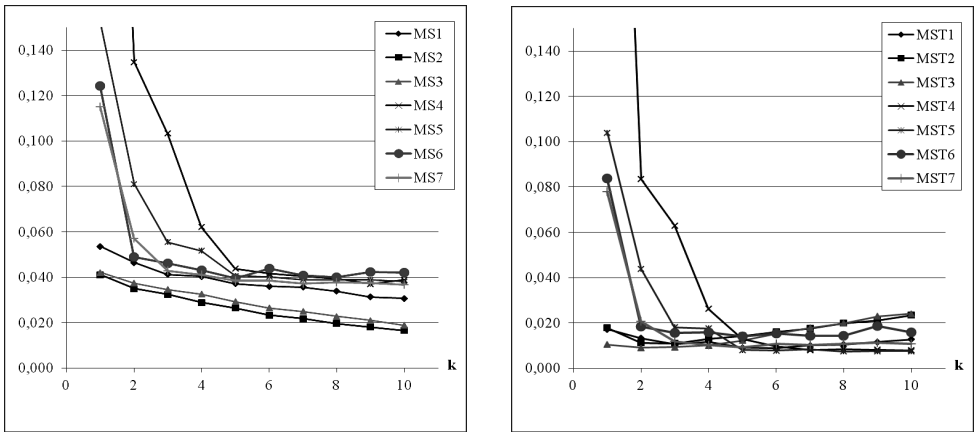

Рис. 4. Усредненные суммы квадратов при уровне шума $\rho = 5\ \%$ с выбросами

Среди всех моделей, использующих сглаживание непараметрической части, наиболее устойчивой оказалась модель, в которой координаты узловых точек базисных функций определялись по второму алгоритму. Наименее устойчивыми к выбросам оказались модели, полученные при использовании алгоритмов 3 и 4. Модели, построенные с использованием алгоритма 1, показывают наилучшие результаты при увеличении числа базисных функций.

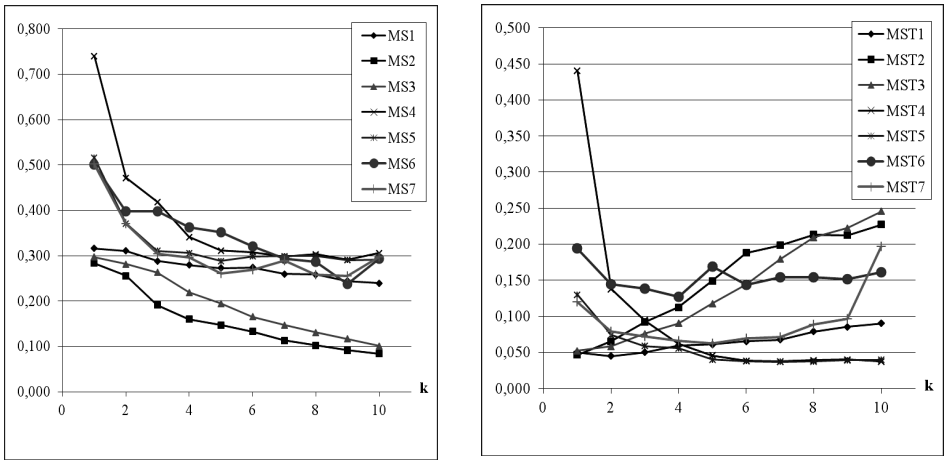

Рис. 5. Усредненные суммы квадратов при уровне шума $\rho = 20\ \%$ с выбросами

Аналогичные вычислительные эксперименты проводились и на других, отличных от (12), вариантах истинных зависимостей. При этом были получены результаты, качественно схожие с описанными выше.

ЗАКЛЮЧЕНИЕ

Таким образом, в данной работе предложены модификации известного метода полупараметрического оценивания регрессионной зависимости, основанные на оптимизации размещения узловых точек в факторном пространстве. Проведенные посредством вычислительных экспериментов исследования подтвердили работоспособность предложенных алгоритмов и высокую точность воспроизведения искомой зависимости.

Результаты проведенных исследований позволяют сделать вывод о том, что при малом уровне шума следует отдавать предпочтение моделям, не предполагающих сглаживание непараметрической части, при построении которых используются алгоритмы 3 и 4. При больших уровнях шума как и при наличии некоторой доли аномальных наблюдений наилучших результатов можно добиться применяя алгоритмы, основанные на сглаживании и выборе координат узловых точек, позволяющем максимизировать информацию Фишера.

СПИСОК ЛИТЕРАТУРЫ

[1] **Horowitz J.L.** Semiparametric and Nonparametric Methods in Econometrics / J.L. Horowitz. –New York: Springer, 2009. – 286 p.

[2] **Ichimura H.** Implementing nonparametric and semiparametric estimators Handbook of Econometrics / H. Ichimura, P.E. Todd. – Vol. 6. – Part B. – Elsevier Science. – 2007. – P. 5369–5468.

[3] **Ruppert D.** Semiparametric Regression / D. Ruppert, M.P. Wand, R.J. Carroll. – New York: Cambridge university press, 2003. – 404 p.

[4] **Денисов В.И.** Штрафные сплайны в задаче идентификации полупараметрической регрессии / В.И. Денисов, В.С. Тимофеев, О.И. Бузмакова // Научн. вестник НГТУ. – Новосибирск: Изд-во СО РАН. – 2011. – № 4 (45). – С. 11–24.

[5] **Денисов В.И.** К вопросу выбора оптимальных координат узловых точек в моделях полупараметрической регрессии / В.И. Денисов, А.В. Фаддеенков // Научн. вестник НГТУ. – Новосибирск: Изд-во СО РАН. – 2012. – № 4 (49). – С. 3–11.

[6] **Ивахненко А.Г.** Помехоустойчивость моделирования / А.Г. Ивахненко, В.С. Степашко. – Киев: Наукова думка, 1985. – 216 с.

[7] **Кендалл М.** Статистические выводы и связи / М. Кендалл, А. Стьюарт. – М.: Наука, 1973. – 466 с.

Денисов Владимир Иванович, заслуженный деятель науки РФ, доктор технических наук, профессор, академик МАН ВШ, член-корреспондент АИН РФ, профессор кафедры прикладной математики факультета прикладной математики и информатики НГТУ. Основное направление научных исследований: разработка и исследование статистических методов анализа, планирования экспериментов и прогнозирования многофакторных статистических и динамических объектов. Имеет более 250 публикаций. E-mail: videnis@nstu.ru

Тимофеев Владимир Семенович, доктор технических наук, доцент, декан факультета прикладной математики и информатики НГТУ. Основное направление научных исследований: разработка и исследование устойчивых методов и алгоритмов анализа многофакторных объектов, в том числе с использованием непараметрической статистики. Имеет более 75 публикаций, в том числе один учебник. E-mail: netsc@fpm.ami.nstu.ru

Фаддеенков Андрей Владимирович, кандидат технических наук, доцент кафедры «Теория рынка» НГТУ. Основное направление научных исследований: разработка и исследование методов и алгоритмов анализа многофакторных объектов со структурированной ошибкой. Имеет более 30 публикаций, в том числе один учебник. E-mail: fadd@fb.nstu.ru

Denisov V.I., Timofeev V.S., Faddeenkov A.V.

Investigation of algorithms of selection of knots' optimal coordinates for semiparametric models with penalized splines

Modifications of the regression construction methods on the basis of the semi-parametric models are suggested. New algorithms choosing of knots' optimal coordinates on the basis of criteria of the accuracy and descriptiveness of individual observations are developed. The results of comparative studies of the developed algorithms for different variants of data contamination taken with the computational experiments presented.

Key words: parametric and non-parametric methods, semiparametric regression, penalized splines models, basis functions, least squares method, design of experiments.