

НЕПАРАМЕТРИЧЕСКОЕ РЕГРЕССИОННОЕ ПРОГНОЗИРОВАНИЕ НА ОСНОВЕ ЯДЕРНОЙ ОЦЕНКИ МНОГОМЕРНОЙ ПЛОТНОСТИ

В. И. Малюгин

Белорусский государственный университет

Минск, Беларусь

E-mail: Malugin@bsu.by

Для решения задач непараметрического регрессионного прогнозирования предлагаются оценки многомерной нелинейной функции регрессии, основанные на непараметрической оценке многомерной плотности распределения с адаптивным гауссовским ядром: оценка на основе моды условного распределения и ядерная оценка функции регрессии с адаптивным выбором параметра сглаживания. Проводится экспериментальное исследование точности прогнозирования на основе предлагаемых оценок, а также сравнительный анализ с известной ядерной оценкой функции регрессии.

Ключевые слова: многомерная плотность, нелинейная функция регрессии, ядерные оценки, адаптация, прогнозирование.

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ НАБЛЮДЕНИЙ

Пусть результатом наблюдения над объектом исследования в i -м эксперименте является случайный вектор $y_i \in \mathbb{R}^p$, допускающий известное разбиение на подвекторы $x_i = (x_{i1}, \dots, x_{iN})' \in \mathbb{R}^N$, $z_i = (z_{i1}, \dots, z_{iM})' \in \mathbb{Z} \subset \mathbb{R}^M$ эндогенных и экзогенных переменных соответственно:

$$y'_i = (x'_i, z'_i) \in \mathbb{R}^p, \quad (1)$$

где $p = N + M$, $i = 1, \dots, n$, $\mathbb{Z} \subset \mathbb{R}^M$ – ограниченная область в $\mathbb{R}^M$.

Компоненты вектора $y_i \in \mathfrak{R}^{N+M}$ связаны моделью многомерной нелинейной регрессии вида

$$x_i = f(z_i) = \xi_i, \quad i = 1, \dots, n, \quad (2)$$

где $f(\cdot)$ – неизвестная достаточно гладкая векторная функция; $\xi_i = (\xi_{i1}, \dots, \xi_{iN})' \in \mathfrak{R}^N$ – случайный вектор ошибок, с нулевым математическим ожиданием и невырожденной ковариационной матрицей $\Sigma = (\sigma_{ij})$; случайные векторы $z_i \in \mathbf{Z}$ и $\xi_i \in \mathfrak{R}^N$ являются статистически независимыми и имеют плотности распределения $p_z(\cdot)$ и $p_\xi(\cdot)$ соответственно. С учетом сделанных предположений плотность распределения $p(y)$ случайного вектора $y_i \in \mathfrak{R}^p$ ($i = 1, \dots, n$) имеет вид

$$p(y) = p_\xi(x - f(z)) p_z(z), \quad x \in \mathfrak{R}^N, \quad z \in \mathbf{Z} \subset \mathfrak{R}^M, \quad y \in \mathfrak{R}^{N+M}. \quad (3)$$

Актуальными для практики предположениями относительно плотностей $p_z(\cdot)$ и $p_\xi(\cdot)$ является предположение о том, что случайные векторы $z_i \in \mathbf{Z}$ и $\xi_i \in \mathfrak{R}^N$ имеют соответственно равномерное в области $\mathbf{Z} \subset \mathfrak{R}^M$ и N -мерное нормальное распределение с плотностями:

$$p_z(z) = \frac{1}{\text{mes}\{\mathbf{Z}\}} \mathbf{I}_{\mathbf{Z}}(z), \quad p_\xi(\xi) = n_N(\xi | 0_N, \Sigma), \quad z \in \mathbf{Z}, \quad \xi \in \mathfrak{R}^N, \quad (4)$$

где $n_N(\cdot | 0_N, \Sigma)$ – функция плотности N -мерного нормального распределения с нулевым вектором математического ожидания и ковариационной матрицей $\Sigma \in \mathfrak{S}_N$, где $\mathfrak{S}_N$ – семейство симметричных положительно определенных $(N \times N)$ -матриц; $\mathbf{I}_{\mathbf{Z}}(z)$ и $\text{mes}\{\mathbf{Z}\} < \infty$ – соответственно индикаторная функция и N -мерный объем (мера Лебега) области $\mathbf{Z}$. При этих предположениях

$$p(y) = \frac{1}{\text{mes}\{\mathbf{Z}\}} \mathbf{I}_{\mathbf{Z}}(z) n_N(x | f(z), \Sigma). \quad (5)$$

Особенностью рассматриваемой модели данных, определяемой плотностью (3) и ее частным случаем (5), является предположение о малости дисперсии компонент случайного вектора ошибок $\xi \in \mathfrak{R}^N$, которое в аналитических исследованиях формулируется следующим образом:

$$\text{tr}(\Sigma) \rightarrow 0 \quad \text{или} \quad \sigma^2 \equiv \max_{i,j=1,\dots,N} \{\sigma_{ii}\} \rightarrow 0. \quad (6)$$

Условие (6) означает, что наблюдения $\{y_i\} (i = 1, \dots, n)$ концентрируются в пространстве $\mathfrak{R}^p$ «вблизи» некоторой N -мерной ($N < p$) гиперповерхности (многообразия), определяемой тождеством $T(y) = x - f(z) = 0 \in \mathfrak{R}^N$. Таким образом, распределение вероятностей случайного вектора $y \in \mathfrak{R}^p$, описываемое соотношениями (1)–(3), при выполнении условия (6) близко к вырожденному.

Задача состоит в прогнозировании вектора эндогенных переменных $x \in \mathbb{R}^N$ по заданному значению вектора экзогенных переменных $z \in \mathbb{Z}$. Поскольку функциональный вид зависимости в (2) не известен, то для решения указанной задачи предлагается использовать непараметрические методы, основанные на многомерной ядерной оценке плотности вектора признаков $y \in \mathbb{R}^p$.

Предположение (6) существенно осложняет проблему оценивания плотности $p(y)$ с помощью непараметрических ядерных оценок, использующих фиксированное ядро, поскольку в данном случае приходится оценивать распределения, близкие к вырожденным. В [1] для данного случая предлагается использовать непараметрическую ядерную оценку многомерной плотности с адаптивным гауссовским ядром.

Предположение (4), а следовательно, и неизвестное на практике представление (5) при выборе функции ядра предлагаемой оценки плотности не используются. Данные предположения, а также асимптотика усиливающейся статистической зависимости компонент вектора признаков (6) и асимптотика растущего объема выборки ($n \rightarrow \infty$) используются в [2] для формирования альтернативных типов ситуаций, различающихся степенью статистической зависимости признаков и объемом выборки, в рамках аналитического исследования решающих правил, основанных на непараметрических оценках многомерной плотности с фиксированным и адаптивным ядром.

НЕПАРАМЕТРИЧЕСКАЯ ОЦЕНКА ПЛОТНОСТИ С АДАПТИВНЫМ ЯДРОМ

Поскольку параметрический вид функции $f(\cdot)$ не известен, то для оценивания плотности распределения $p(y)$ по случайной выборке $Y = (y_i) \in \mathbb{R}^{pn}$ используется непараметрическая оценка плотности Розенблатта – Парзена с многомерным гауссовским ядром вида [3]:

$$\hat{p}(y) = \frac{1}{n} \sum_{j=1}^n n_p(y|y_j, h^2 H), \quad (7)$$

где H, h – управляемые компоненты гауссовского ядра: $H \in \mathfrak{S}_p$ – матрица гауссовского ядра; $h \equiv h(n)$ – коэффициенты сглаживания, удовлетворяющие условиям асимптотической несмещенности и состоятельности оценки плотности:

$$h(n) \rightarrow 0, \quad nh(n) \rightarrow \infty, \quad n \rightarrow \infty.$$

При построении оценки (7), наряду с задачей вычисления коэффициентов сглаживания $h(n)$, приходится решать задачу выбора матрицы ядра H . Обычно в качестве H используется либо фиксированная (единичная) матрица $H^{(1)} \in \mathfrak{S}_p$, либо выборочная по всей выборке оценка ковариационной матрицы $H^{(2)} \in \mathfrak{S}_p$ [3].

Как показано в [1], если зависимость (2) является линейной $T(y) = \theta'y$ ($\theta \in \mathbb{R}^p$, $p = M+1$) и выполняются предположения (3)–(5), то оценка плотности $\hat{p}(y)$ с матрицей $H = H^{(2)}$ (обозначается далее $\hat{p}^{(2)}(y)$) в асимптотике усиливающейся статистической зависимости (6) компонент вектора $y \in \mathbb{R}^p$ имеет су-

щественный выигрыш по сравнению с оценкой $\hat{p}^{(1)}(y)$, использующей произвольную фиксированную матрицу $H^{(1)}$. Так, для достижения оценками одной и той же точности в смысле среднего квадрата относительного смещения оценкам $\hat{p}^{(2)}(y)$ и $\hat{p}^{(1)}(y)$ требуются объемы выборок n_2 и n_1 , связанные соотношением

$$n_2 = n_1 \delta^{-\frac{p}{\gamma}} (0 < \gamma < 1), \quad \delta = \left(\frac{\theta' H^{(1)} \theta}{\sigma^2} \right),$$

откуда следует, что в рассматриваемой асимптотике $n_2 \ll n_1$.

Оптимальные в смысле относительной глобальной ошибки аппроксимации плотности коэффициенты сглаживания для оценки $\hat{p}^{(2)}(y)$ получены в [1] и вычисляются по формуле

$$h_0 = h_0(n) = \varphi(p) n^{-\frac{1}{p+4}}, \quad \varphi(N) = \left(\frac{4}{3} \left(\frac{3}{\pi} \right)^{\frac{p-1}{2}} \right)^{\frac{1}{p+4}}. \quad (8)$$

На этом основании для оценивания плотности распределения $p(y)$ в случае нелинейной зависимости компонент вектора $y \in \mathfrak{R}^p$, определяемой соотношениями (1), (2), предлагается процедура адаптивного выбора матрицы ядра. Адаптация в указанной процедуре достигается за счет использования для каждого наблюдения $y_i (i = 1, \dots, n)$ своей матрицы ядра (локальной выборочной оценки ковариационной матрицы случайного вектора $y \in \mathfrak{R}^p$), вычисленной в некоторой окрестности наблюдения y_i , обеспечивающей наилучшую линейную аппроксимацию нелинейной зависимости (2) в данной точке. Оптимальный размер локальной окрестности для точки y_i , определяемый количеством попавших в нее точек $m(i)$ из выборки Y , находится из условия минимума, характеризующей степень множественной линейной зависимости компонент вектора $y \in \mathfrak{R}^p$ в окрестности точек $y_i (i = 1, \dots, n)$ (при этом само наблюдение y_i в число $m(i)$ наблюдений не входит):

$$m(i) = \arg \min_{K_0 \leq k \leq n-1} (V^{(i,k)})(p < K_0 < n), \quad V^{(i,k)} = \frac{|S^{(i,k)}|}{\prod_{l=1}^N s_{ll}^{(i,k)}}, \quad (9)$$

$$S^{(i,k)} = \frac{1}{k-1} \sum_{y_j \in O(i,k)} (y_j - \bar{y}^{(i,k)})(y_j - \bar{y}^{(i,k)})', \quad \bar{y}^{(i,k)} = \frac{1}{k} \sum_{y_j \in O(i,k)} y_j,$$

где $O(i, m(i))$ – локальная окрестность точки y_i радиуса $m(i)$, обеспечивающая наилучшую линейную аппроксимацию нелинейной зависимости (2) в точке $y_i (i = 1, \dots, n)$, K_0 – заданный минимальный размер локальной окрестности.

Таким образом, непараметрическая оценка плотности $p(y)$ с адаптивным гауссовским ядром по случайной выборке $Y = \{y_i\} (i = 1, \dots, n)$, удовлетворяющей предположениям (1), (2), определяется соотношениями:

$$\begin{aligned}\hat{p}^{(3)}(y) &= \frac{1}{n} \sum_{i=1}^n n_p \left(y \middle| y_i, h_i^2 H^{(i,m(i))} \right), \\ H^{(i,m(i))} &= \frac{1}{m(i)-1} \sum_{y_j \in O(i,m(i))} (y_j - \bar{y}^{(i,m(i))}) (y_j - \bar{y}^{(i,m(i))})', \\ \bar{y}^{(i,m(i))} &= \frac{1}{m(i)} \sum_{y_j \in O(i,m(i))} y_j,\end{aligned}\tag{10}$$

где коэффициенты сглаживания $h_i \equiv h_0(m(i))$, т. е. вычисляются по формуле (8) с заменой n на $m(i)$, $i = 1, \dots, n$.

НЕПАРАМЕТРИЧЕСКИЕ ОЦЕНКИ ФУНКЦИИ РЕГРЕССИИ И АЛГОРИТМЫ ПРОГНОЗИРОВАНИЯ

В работе предлагаются два алгоритма прогнозирования вектора эндогенных переменных $x \in \mathfrak{R}^N (N \geq 1)$ для заданного значения $z \in \mathbf{Z}$ с использованием оценки (9), (10) многомерной плотности $p_\alpha(y) \equiv p_\alpha(x, z)$:

1) алгоритм прогнозирования на основе моды условного распределения (алгоритм 1):

$$\hat{x} = \arg \max_x \hat{p}_\alpha(x|z) = \arg \max_x \frac{\hat{p}_\alpha^{(3)}(x, z)}{\hat{p}(z)};$$

2) алгоритм прогнозирования на основе ядерной оценки функции регрессии с адаптивным ядром, т.е. адаптивным выбором параметра сглаживания $\{h_i = h_0(m_i)\}$ по формулам (8), (9) (алгоритм 2):

$$\hat{x} = \hat{f}_\alpha(z) = \frac{\sum_{i=1}^n x_i K\left(\frac{z_i - z}{h_i}\right)}{\sum_{i=1}^n K\left(\frac{z_i - z}{h_i}\right)}, K(u) = \exp\left(-\frac{u'u}{2}\right), u \in \mathbf{Z}.$$

Прогнозы на основе предлагаемых алгоритмов при $N=1$ сравниваются на тестовых данных по точности с прогнозом на основе ядерной оценки функции регрессии с фиксированным $h_i = h_0 (i = 1, \dots, n)$ [4, 5] (алгоритм 3). В качестве критерия точности прогнозов используется среднеквадратическая ошибка прогноза (*root mean square error – RMSE*). Показывается, что предлагаемые алгоритмы 1, 2 допускают возможность построения многомерных прогнозов, а в одномерном случае ($N=1$) могут иметь преимущество по точности прогнозирования перед алгоритмом 3 в условиях малой обучающей выборки. Результаты экспериментов согласуются с результатами аналитических исследований предлагаемых алгоритмов в асимптотике растущего объема обучающей выборки и усиливающейся статистической зависимости компонент вектора признаков [2].

ЛИТЕРАТУРА

1. *Малюгин, В. И.* Об оценивании плотности случайных векторов с существенно зависимыми компонентами / В. И. Малюгин // Вест. БГУ. Сер. 1. 1985. № 2. С. 41–44.
 2. *Малюгин, В. И.* Асимптотический анализ риска непараметрической классификации в случае существенно зависимых признаков / В. И. Малюгин // Изв. НАН Беларуси. Сер.1. Физ. Мат. Информ. 2009. № 3. С. 10–23.
 3. *Фукунага, К.* Введение в статистическую теорию распознавания образов / К. Фукунага. М. : Наука, 1979. 367 с.
 4. *Thomas, P. R.* Modern Regression Methods / P. R. Thomas. John Wiley, 1997. 462 p.
 5. *Айвазян, С. А.* Прикладная статистика. Исследование зависимостей / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. М. : Финансы и статистика, 1985. 487 с.
-