В. И. Малюгин, М. Е. Васильков

Непараметрический анализ стохастических систем с нелинейной функциональной неоднородностью[1]

В данной статье рассматриваются задачи анализа стохастических систем, описываемых нелинейными статистическими моделями с неоднородной функциональной формой в пространстве существенно зависимых признаков. Предполагается, что функциональная неоднородность моделей обусловлена различными классами состояний системы. Приводятся результаты аналитического и экспериментального исследования алгоритма классификации состояний системы, а также алгоритма прогнозирования зависимых переменных, основанных на непараметрической оценке многомерной плотности вероятностей с адаптивным гауссовским ядром.

Ключевые слова: многомерные модели, существенно зависимые признаки, функциональная неоднородность, многомерная непараметрическая оценка плотности, адаптивное гауссовское ядро, непараметрическая классификация и прогнозирование.

JEL classification: C14, C38.

1. Введение

При решении задач анализа, прогнозирования и управления на основе эконометрических моделей экономических, технологических и производственных систем (далее — *сложных систем*) часто приходится сталкиваться со структурной неоднородностью моделей (Айвазян и др., 1989; Hardle, Simar, 2007; Hubler, Frohn, 2006). При описании сложных систем с помощью многомерных линейных по параметрам моделей регрессионного и авторегрессионного типа структурная неоднородность моделей вызывается скачкообразными изменениями параметров моделей, определяющих структуру зависимости между эндогенными и экзогенными переменными. Структурная неоднородность модели может быть обусловлена двумя основными причинами:

- наличием нескольких априорно предполагаемых режимов функционирования (классов состояния) сложных систем;
- структурными изменениями модели, вызванными внешними шоковыми воздействиями.

Задачам анализа и прогнозирования экономических систем в условиях *параметрической структурной* неоднородности моделей посвящены работы автора (Малюгин, Харин, 1986; Малюгин, 2008а, 2008б, 2009а).

[1] Данная статья является расширенной версией доклада на международной конференции «Computer Data Analysis and Modeling», Минск, 2010 (Malugin, Vasilkov, 2010).

На практике предположение о линейности модели по параметрам часто не подтверждается. В связи с этим возникает проблема построения многомерных нелинейных моделей. Параметрический вид таких моделей, как правило, точно не известен, и это приводит к необходимости использования непараметрических методов оценивания зависимостей, прогнозирования и классификации (Айвазян и др., 1985; Хардле, 1993). При построении многомерных эконометрических моделей приходится также сталкиваться с проблемой экзогенности факторов, т. е. проблемой разбиения совместно используемых экономических переменных на эндогенные (внутренние) и экзогенные (внешние по отношению к модели). Таким образом, в общем случае эндогенно-экзогенная структура модели может быть также точно не известна.

Рассматриваемая в работе многомерная статистическая модель имеет две существенные особенности. Первая состоит в том, что компоненты вектора признаков являются «существенно зависимыми» случайными величинами, совместное распределение вероятностей которых близко к вырожденному. Второй особенностью является неоднородность, обусловленная наличием нескольких режимов функционирования системы, каждому из которых соответствует своя модель зависимости признаков. Эта особенность интерпретируется как *функциональная структурная неоднородность* модели наблюдений. В общем случае допускается также неопределенность в разбиении вектора признаков на зависимые и независимые компоненты (эндогенные и экзогенные переменные). Если указанное разбиение известно, т. е. определена эндогенно-экзогенная структура модели, то рассматриваемые модели принимают вид многомерных нелинейных статистических зависимостей регрессионного типа.

В данной работе предполагается, что имеют место два режима функционирования сложной системы и, соответственно, два класса состояний, различающихся моделями статистических зависимостей для компонент случайного вектора признаков. Решаются две задачи анализа рассматриваемых систем: 1) задача прогнозирования (оценки) класса состояния системы в случае неизвестной эндогенно-экзогенной структуры вектора признаков; 2) задача прогнозирования эндогенных переменных по известному значению вектора экзогенных переменных для заданного класса состояния системы и известной эндогенно-экзогенной структуре вектора признаков.

Алгоритмы анализа сложных систем в обоих случаях основываются на использовании непараметрических ядерных оценок многомерных условных плотностей распределения наблюдений с адаптивным (переменным) гауссовским ядром (Малюгин, 1985). Для решения первой задачи используются непараметрические классификаторы, получающиеся подстановкой в оптимальное (в смысле минимума риска) байесовское решающее правило непараметрических ядерных оценок многомерных условных плотностей распределения наблюдений с адаптивным гауссовским ядром. При решении второй задачи прогнозное значение эндогенных переменных определяется как модальное значение непараметрической оценки условной плотности распределения вектора эндогенных переменных для заданного значения вектора экзогенных переменных и состояния системы. Приводятся результаты аналитического исследования алгоритмов, полученных в асимптотике усиливающейся статистической зависимости компонент вектора признаков и растущем объеме выборки. На тестовых модельных данных проводится сравнительный анализ предлагаемых алгоритмов с известными непараметрическими алгоритмами прогнозирования и классификации.

2. Математическая модель наблюдений и задачи исследования

Модель «существенной зависимости». Пусть характеристики сложной системы в i-ом эксперименте описываются случайным вектором $y_i \in \Re^p$, который допускает разбиение на подвекторы вида

$$y_i = \begin{pmatrix} x_i \\ z_i \end{pmatrix} \in \Re^p, \quad x_i = (x_{i1},...,x_{iN})' \in \Re^N, \quad z_i = (z_{i1},...,z_{iM})' \in \mathbf{Z} \subset \Re^M, \tag{1}$$

где $p = N + M$, $i = 1,...,n$, $\mathbf{Z} \subset \Re^M$ — ограниченная область в $\Re^M$.

Компоненты вектора $y_i \in \Re^{N+M}$ связаны моделью статистической зависимости

$$T(y_i) = x_i - f(z_i) = \xi_i, \; i = 1,\ldots,n, \tag{2}$$

где: $T(\cdot), f(\cdot)$ — неизвестные достаточно гладкие векторные функции; $\xi_i = (\xi_{i1,\ldots,}\xi_{iN})' \in \Re^N$ — случайный вектор ошибок с нулевым математическим ожиданием и невырожденной ковариационной матрицей $\Sigma = (\sigma_{ij}) \in \Im_N$, где $\Im_N$ — семейство положительно определенных симметричных матриц размерности $N \times N$; случайные векторы $z_i \in \mathbf{Z}$ и $\xi_i \in \Re^N$ являются статистически независимыми и имеют плотности распределения $p_z(z)$ и $p_\xi(\xi)$ соответственно; разбиение (1) в общем случае неизвестно. Для простоты записи случайный вектор и неслучайный аргумент соответствующей функции плотности распределения обозначаются одним и тем же символом.

С учетом сделанных предположений плотность распределения $p(y)$ случайного вектора $y_i \in \Re^p$ $(i = 1,\ldots,n)$ имеет вид:

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

Для пояснения сути рассматриваемой проблемы статистического оценивания плотности вида (3) введем конкретизирующие предположения относительно плотностей распределения $p_z(z)$ и $p_\xi(\xi)$. Будем считать, что случайные векторы $z_i \in \mathbf{Z}$ и $\xi_i \in \Re^N$ имеют, соответственно, равномерное в ограниченной области $\mathbf{Z} \subset \Re^M$ и N-мерное нормальное распределение с плотностями:

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

где $n_N(\xi | 0_N, \Sigma)$ — функция плотности N-мерного нормального распределения с нулевым вектором математического ожидания и ковариационной матрицей $\Sigma \in \Im_N$; $\mathbf{I}_{\mathbf{Z}}(z)$ и $mes\{\mathbf{Z}\} < \infty$ — соответственно индикаторная функция и N-мерный объем (мера Лебега) области $\mathbf{Z}$.

Согласно (3) и (4) плотность распределения случайного вектора $y \in \Re^p$ определяется выражением:

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

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

$$\mathrm{tr}(\Sigma) \to 0 \quad \text{или} \quad \sigma^2 \equiv \max_{i=1,\ldots,N} \{\sigma_{ii}\} \to 0. \tag{6}$$

Асимптотика (6) означает усиливающуюся статистическую зависимость компонент вектора $y_i \in \Re^p (p > 1)$ при уменьшении дисперсий компонент вектора случайных ошибок $\xi_i = \left(\xi_{i1,\ldots,}\xi_{iN}\right)' \in \Re^N$. Условие (6) означает, что наблюдения $\{y_i\}(i = 1,\ldots,n)$ концентрируются в пространстве $\Re^p$ «вблизи» некоторой N-мерной $(N < p)$ гиперповерхности (многообразия) Γ, определяемой тождеством $T(y) = 0_N$. Данное тождество можно интерпретировать как некоторое нелинейное ограничение, которому удовлетворяют анализируемые переменные в устойчивом состоянии системы для определенного режима функционирования. При этом вектор ξ_i рассматривается как случайное отклонение системы от этого состояния в i-ом эксперименте, обусловленное случайными и неконтролируемыми факторами. Таким образом, распределение вероятностей случайного вектора $y \in \Re^p$, описываемое соотношениями (1) – (3), при выполнении условия (6) близко к вырожденному. На этом основании в работе (Малюгин, 1985) вектор $y \in \Re^p$, удовлетворяющий условиям (1), (2), (6), называется *случайным вектором с «существенно зависимыми» компонентами*. Предположение о «существенной зависимости» компонент вектора признаков, очевидно, осложняет проблему оценивания плотности распределения данного вектора с помощью непараметрических ядерных оценок, использующих фиксированное ядро, поскольку в данном случае приходится оценивать распределения, близкие к вырожденным. Асимптотика (6), таким образом, означает не только усиливающуюся статистическую зависимость между компонентами вектора признаков, но и растущую сложность задачи оценивания плотности распределения вектора признаков.

Рассматриваемая модель данных допускает следующие интерпретации. В случае, когда разбиение вектора $y \in \Re^p$ на подвекторы эндогенных и экзогенных переменных известно, соотношение (2) представляет собой модель многомерной нелинейной регрессии. Поскольку функциональный вид зависимости в (2) не известен, то для анализа моделей типа (2) предлагается использовать непараметрические методы[2], основанные на многомерной ядерной оценке плотности вектора признаков $y \in \Re^p$. Предположение о малости дисперсии компонент случайного вектора ошибок в данном случае, как показывают проведенные исследования, делает предпочтительным (в смысле требуемого объема выборки) использование непараметрических ядерных оценок с адаптивным ядром.

Если разбиение (1) вектора $y \in \Re^p$ неизвестно, то описанная модель соответствует ситуации, когда вектор признаков $y \in \Re^p$ является избыточным, т. е. фактически необходимое

[2] Проблема выбора между параметрической и непараметрической спецификацией модели регрессионного типа комментируется в (Racine, 2008).

для описания состояния сложной системы число переменных (параметров) меньше размерности исходного пространства и совпадает с размерностью N многообразия Γ. Неизвестная величина N $(N < p)$ при этом называется *истинной размерностью* (*intrinsic dimensionality*) пространства признаков (Фукунага, 1979). Подобные модели данных возникают в различных приложениях (см., например, (Granlund, Knutsson, 1995; Camastra, Vinciarelli, 2002)).

Следует также отметить, что предположения (4) используются в работах (Малюгин, 1985, 2009б), посвященных аналитическому асимптотическому анализу непараметрических оценок плотности вида (3), а также основанных на этих оценках решающих правил. Краткое описание указанных результатов приводится в разделе 4. Заметим, что предположения (4), а, следовательно, и неизвестное на практике представление (5), при выборе функции ядра предлагаемой оценки плотности не используются. В то же время они не противоречат традиционным предположениям относительно объясняющих переменных и случайных ошибок в моделях регрессионного типа, к числу которых может быть отнесена рассматриваемая модель зависимости вида (2) при известном разбиении (1) вектора признаков $y_i \in \Re^p (p > 1)$ на эндогенные и экзогенные переменные. Как и предположения (4), асимптотика (6) усиливающейся статистической зависимости компонент вектора признаков используется в разделе 4 наряду с асимптотикой растущего объема выборки для формирования различных типов ситуаций, различающихся степенью статистической зависимости признаков и объемом выборки, в рамках аналитического исследования непараметрических классификаторов с фиксированным и адаптивным ядром.

Модель структурной функциональной неоднородности. Пусть сложная система характеризуется случайным вектором $y \in \Re^p$, описываемым моделью (1), (2), (6), и имеют место два режима функционирования, которым соответствуют два класса состояний системы Ω_1 и Ω_2. Номер класса состояния системы в i-ом эксперименте описывается ненаблюдаемой случайной величиной $\nu_i = \nu(y_i) \in S = \{1,2\}$ $(i = 1,\ldots,n)$ с распределением вероятностей $P\{\nu_i = \alpha\} = \pi_\alpha > 0$ $(\alpha \in S)$, $\pi_1 + \pi_2 = 1$, параметры $\{\pi_\alpha, \alpha \in S\}$ называются *априорными вероятностями классов состояний системы*.

Классам $\{\Omega_\alpha\}$ соответствуют неизвестные функции $\{f_\alpha(z)\}$, удовлетворяющие *условию функциональной структурной неоднородности модели наблюдений*:

$$\mathbf{P}\left(f_1(z) = f_2(z)\right) = 0,\ z \in \mathbf{Z}, \tag{7}$$

которое означает, что для различных классов состояний модели статистических зависимостей признаков различны с точностью до множества меры нуль. Условные плотности распределения $p_\alpha(y)$ случайного вектора наблюдений $y \in \Re^p$ для классов состояний системы $\{\Omega_\alpha\}$ имеют вид (3) при $f(z) \equiv f_\alpha(z)$ $\left(T(y) \equiv T_\alpha(y)\right)$, $\alpha \in S$.

Априорная информация и задачи анализа. Относительно описания модели делаются следующие предположения:

- вероятностные характеристики классов $\left\{\pi_\alpha, p_\alpha(y), \alpha \in S\right\}$ неизвестны;
- эндогенно-экзогенная структура вектора признаков $y = \begin{pmatrix} x \\ z \end{pmatrix} \in \Re^{N+M}$ может быть известна, либо не известна;
- имеется классифицированная обучающая выборка наблюдений $Y = (y_i) \in \Re^{pn}$, допускающая разбиение на подвыборки наблюдений из классов $\{\Omega_\alpha\}$: $Y = Y_1 \cup Y_2$, где $Y_\alpha = (y_{\alpha i}) \in \Re^{pn_\alpha}$ — выборка наблюдений из класса $\Omega_\alpha (\alpha \in S,\ n = n_1 + n_2)$.

Актуальными являются следующие задачи анализа рассматриваемых систем.

1. *Задача прогнозирования (оценки) класса состояния системы.* Она заключается в определении класса состояния системы (режима функционирования) на основе классификации вновь поступающих наблюдений за системой к одному из заданных классов, т. е. в оценке номера класса состояния системы $\nu_i = \nu(y_i) \in S$ по наблюдаемому значению $y_i = \begin{pmatrix} x_i \\ z_i \end{pmatrix} \in \Re^{N+M}$ $(i = n+1,\ n+2,...)$. Эндогенно-экзогенная структура вектора признаков при этом может быть неизвестна.

2. *Задача прогнозирования эндогенных переменных для заданного класса состояния системы.* Предполагается, что известна эндогенно-экзогенная структура вектора признаков $y = \begin{pmatrix} x \\ z \end{pmatrix} \in \Re^{N+M}$. Задача состоит в прогнозировании вектора эндогенных переменных $x \in \Re^N$ по заданному значению вектора экзогенных переменных $z \in \mathbf{Z}$.

3. Методы и алгоритмы анализа состояния сложной системы

Приведем краткое описание непараметрической оценки плотности с адаптивным ядром и основанных на ней методов решения сформулированных выше задач.

Непараметрическая оценка плотности с адаптивным гауссовским ядром. Поскольку параметрический вид функций $T(\cdot), f(\cdot)$, а также само разбиение вектора $y \in \Re^p$ на подвекторы неизвестны, то для оценивания плотности распределения $p(y)$ по случайной выборке $Y = (y_i) \in \Re^{pn}$ используется непараметрическая оценка плотности Розенблатта – Парзена с многомерным гауссовским ядром, определяемая по формуле (Фукунага, 1979):

$$\hat{p}(y) = \frac{1}{n}\sum_{j=1}^{n} n_p\left(y \middle| y_j, h^2 H\right), \tag{8}$$

где *H, h* — управляемые компоненты гауссовского ядра: $H \in \Im_p$ — матрица гауссовского ядра; $h \equiv h(n)$ — коэффициенты сглаживания, удовлетворяющие условиям асимптотической несмещенности и состоятельности оценки плотности:

$$h(n) \to 0,\ nh(n) \to \infty,\ n \to \infty. \tag{9}$$

При построении оценки (8) наряду с задачей вычисления коэффициентов сглаживания $h(n)$ приходится решать задачу выбора матрицы ядра *H*. Обычно в качестве *H* используется либо фиксированная (единичная) матрица $H^{(1)} \in \Im_p$, либо выборочная (по всей выборке) оценка ковариационной матрицы $H^{(2)} \in \Im_p$ (Фукунага, 1979).

Как показано в (Малюгин, 1985), если зависимость (2) является линейной $T(y) = \theta' y$ $(\theta \in \Re^P, p = M+1)$ 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 $y \in \Re^p$ имеет существенный 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 $\widehat{p}^{(2)}(y)$ и $\widehat{p}^{(1)}(y)$ требуются объемы выборок n_2 и n_1, связанные соотношением

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

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

Коэффициенты сглаживания для оценки $\widehat{p}^{(2)}(y)$, оптимальные в смысле рассматриваемой в (Епанечников, 1969) относительной глобальной ошибки аппроксимации плотности, вычисляются по следующей формуле (см. (Малюгин, 1985)):

$$h_0 = h_0(n) = \phi(p) n^{-\frac{1}{p+4}},\ \phi(p) = \left(\frac{4}{3} \left(\frac{3}{\pi} \right)^{\frac{p-1}{2}} \right)^{\frac{1}{p+4}}. \tag{10}$$

На этом основании для оценивания плотности распределения $p(y)$ в случае нелинейной зависимости компонент вектора $y \in \Re^p$, определяемой соотношениями (1) и (2), была предложена процедура адаптивного выбора матрицы ядра. Адаптация в указанной процедуре достигается за счет использования для каждого наблюдения $y_i\ (i = 1,...,n)$ своей матрицы ядра (локальной выборочной оценки ковариационной матрицы случайного вектора $y \in \Re^p$), 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 *V*-статистики (Андерсон, 1963), характеризующей степень множественной линейной зависимости компонент вектора $y \in \Re^p$ в окрестности точек $y_i\ (i = 1,...,n)$ (при этом само наблюдение y_i в число $m(i)$ наблюдений не входит):

$$m(i) = \arg \min_{K_0 \le k \le n-1} (V^{(i,k)})\ \ (p < K_0 < n),\quad V^{(i,k)} = \frac{|S^{(i,k)}|}{\prod_{l=1}^{N} s_{ll}^{(i,k)}}, \tag{11}$$

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

где $\mathrm{O}(i,k)$ — локальная окрестность точки y_i радиуса k, обеспечивающая наилучшую линейную аппроксимацию нелинейной зависимости (2) в точке $y_i\ (i = 1,...,n)$.

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

$$\widehat{p}^{(3)}(y) = \frac{1}{n} \sum_{i=1}^{n} n_p \left(y \middle| y_i, h_i^2 H^{(i,m(i))} \right), \tag{12}$$

$$H^{(i,m(i))} = \frac{1}{m(i)-1} \sum_{y_j \in O(i,m(i))} \left(y_j - \bar{y}^{(i,m(i))}\right)\left(y_j - \bar{y}^{(i,m(i))}\right)', \ \bar{y}^{(i,m(i))} = \frac{1}{m(i)} \sum_{y_j \in O(i,m(i))} y_j,$$

где коэффициенты сглаживания $h_i \equiv h_0(m(i))$, т. е. вычисляются по формуле (10) с заменой n на $m(i)$, $i = 1,...,n$.

Алгоритм прогнозирования (оценки) класса состояния сложной системы. Как известно (Айвазян и др., 1989; Харин, 1992), оптимальное в смысле минимума риска *байесовское решающее правило* (БРП) классификации наблюдений из классов $\{\Omega_\alpha\}$ с вероятностными характеристиками $\left\{\pi_\alpha, p_\alpha(y),\ \alpha \in S\right\}$ имеет вид:

$$d^{(0)}(y) = \mathbf{1}\left(G(y)\right) + 1 = \begin{cases} 1, \text{ если } G(y) < 0 \\ 2, \text{ если } G(y) \geq 0 \end{cases}, \text{ где } \mathbf{1}(u) = \begin{cases} 0, & \text{если } u < 0, \\ 1, & \text{если } u \geq 0 \end{cases}, \tag{13}$$

где $G(\cdot)$ — байесовская дискриминантная функция, определяемая соотношениями:

$$G(y) = c_2 p_2(y) - c_1 p_1(y),\ c_1 = \pi_1(w_{12} - w_{11}),\ c_2 = (1-\pi_1)(w_{21} - w_{22}), \tag{14}$$

а $W = \left(w_{\alpha\beta}\right)$ $(\alpha, \beta \in S)$ — заданная матрица потерь.

Для оценки класса состояния сложной системы в условиях параметрической неопределенности традиционно применяются *подстановочные байесовские решающие правила* (*подстановочные БРП*), использующие вместо неизвестных условных плотностей распределения $\left\{p_\alpha(y),\ \alpha \in S\right\}$ их непараметрические оценки. Согласно (13), подстановочные БРП $d^{(l)}(y)$, $l = 1,2,3$, использующие вышеописанные оценки $\{\hat{p}_\alpha^{(l)}\}$, $l = 1,2,3$, имеют вид:

$$d^{(l)}(y) = \mathbf{1}\left(\hat{G}^{(l)}(y)\right) + 1,\ \hat{G}^{(l)}(y) = \hat{c}_2 \hat{p}_2^{(l)}(y) - \hat{c}_1 \hat{p}_1^{(l)}(y), \tag{15}$$

где при вычислении $\{\hat{c}_\alpha\}$ по формуле (14) используются оценки априорных вероятностей классов $\pi_\alpha = n_\alpha / n$ $(\alpha = 1,2)$.

Алгоритм прогнозирования эндогенных переменных. В качестве прогнозного значения вектора эндогенных переменных $x \in \Re^N (N \geq 1)$ при заданных значениях $\alpha \in S$ и $z \in \mathbf{Z}$ в случае известной эндогенно-экзогенной структуры вектора признаков $y = \begin{pmatrix} x \\ z \end{pmatrix} \in \Re^{N+M}$ с плотностью распределения $p_\alpha(y) \equiv p_\alpha(x,z)$ предлагается использовать максимальное значение оценки условной плотности распределения $p_\alpha(x \mid z)$, т. е. *оценку моды условного распределения эндогенных переменных*:

$$\hat{x} = \arg\max_x \hat{p}_\alpha(x \mid z) = \arg\max_x \frac{\hat{p}_\alpha^{(3)}(x,z)}{\hat{p}(z)}, \tag{16}$$

где $\hat{p}_\alpha^{(3)}(x,z)$ — оценка совместной плотности распределения случайных векторов x, z вида (12), $\hat{p}(z)$ — непараметрическая оценка с фиксированным ядром частной плотности

распределения $p(z)$. Проблемы построения и применения оценок условной плотности распределения обсуждаются в (Hall et al., 2004).

В качестве альтернативного непараметрического алгоритма прогнозирования в случае одной эндогенной переменной $x \in \Re^1$ $(N=1, p=M+1)$ используется прогноз на основе ядерной оценки функции регрессии с некоторой функцией ядра $K(u)$ $(u \in \mathbf{Z})$ (Айвазян и др., 1985; Thomas, 1997). Для наиболее часто используемого на практике гауссовского ядра данная оценка определяется соотношениями:

$$\hat{x} = \hat{f}_\alpha(z) = \frac{\sum_{i=1}^{n} x_i K\left(\frac{z_i - z}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{z_i - z}{h}\right)}, \quad K(u) = \exp\left(-\frac{u'u}{2}\right), \quad u \in \mathbf{Z}. \tag{17}$$

Ключевой проблемой при использовании оценки (17), как известно (Айвазян и др., 1985), является выбор параметра масштаба h, определяющего размер локальной окрестности, по наблюдениям из которой оценивается функция регрессии в точке $z \in \mathbf{Z}$. В настоящей работе рассматриваются два варианта значений этого параметра: 1) фиксированное значение h_0 для всей области; 2) локальные значения $\{h_0(m(i))\}$, определяемые по формулам (10) и (11).

4. Результаты асимптотического анализа решающих правил

В качестве критериев оптимальности подстановочных решающих правил $d^{(l)}(y)$, $l=1,2,3$, с учетом вида (2) зависимости компонент вектора $y = \begin{pmatrix} x \\ z \end{pmatrix} \in \Re^{N+M}$, $\left(x \in \mathbf{X} = \Re^N,\ z \in \mathbf{Z} \subset \Re^M\right)$ используются следующие функционалы риска (средних потерь):

- условный риск (условное математическое ожидание потерь) в фиксированной точке $z \in \mathbf{Z}$:

$$r_n^{(l)}(z) = E_Y\left\{R_n^{(l)}(z,Y)\right\},\ R_n^{(l)}(z,Y) = \sum_{\alpha \in S} \pi_\alpha \int_X w\left(\alpha, \hat{d}^{(l)}(y)\right) p_\xi\left(x - f_\alpha(z)\right) dx; \tag{18}$$

- условный ε-риск в фиксированной точке $z \in \mathbf{Z}$:

$$r_n^{(l)}(\varepsilon,z) = E_Y\left\{R_n^{(l)}(\varepsilon,z,Y)\right\}(l=1,2),\ R_n^{(l)}(\varepsilon,z,Y) = \sum_{\alpha \in S} \pi_\alpha \int_{T(\varepsilon,z)} w\left(\alpha, \hat{d}^{(l)}(y)\right) p_\xi\left(x - f_\alpha(z)\right) dx,$$

где $T(\varepsilon,z) \subset \mathbf{Z}$ — ограниченная область, удовлетворяющая условию

$$\left|r_n^{(l)}(\varepsilon,z) - r_n^{(l)}(z)\right| \le \varepsilon, \tag{19}$$

здесь величина $\varepsilon > 0$ задает точность приближения $r_n^{(l)}(\varepsilon,z)$ к $r_n^{(l)}(z)$. Согласно (18), использование условного ε-риска обеспечивает возможность оценки условного риска клас-

сификации с заданной точностью, определяемой величиной ε $(0<\varepsilon<1)$. При $\varepsilon \to 0$ точность вычисления условного риска возрастает. В случае «антиединичной» матрицы потерь W (когда $w_{\alpha\beta}=1-\delta_{\alpha\beta}$, где $\delta_{\alpha\beta}$ — символ Кронекера) функционалы (17) и (18) имеют смысл условных вероятностей ошибок классификации.

В соответствии с описанной выше методологией сравнительного анализа непараметрических оценок плотности, рассматривается случай линейной зависимости компонент вектора $y \in \Re^p$:

$$T_\alpha(y_i)=x_i-B_\alpha z_i=\xi_i, \quad i=1,...,n \quad (\alpha \in S),$$

где B_α — неизвестная фиксированная $N \times M$ — матрица, удовлетворяющая условию функциональной неоднородности $\mathbf{P}\left(B_1 z=B_2 z\right)=0,\ z \in \mathbf{Z}$.

Приведем результаты сравнительного анализа риска двух подстановочных решающих правил $d^{(1)}(y)$ и $d^{(2)}(y)$, использующих соответственно фиксированную матрицу ядра $H^{(1)} \in \Im_p$ и вычисленную по всей выборке матрицу $H^{(2)} \in \Im_p$ на основе характеристики

$$\Delta r_n(\varepsilon,z)=r_n^{(1)}(\varepsilon,z)-r_n^{(2)}(\varepsilon,z) \ \ (z \in \mathbf{Z}), \tag{20}$$

где $r_n^{(1)}(\varepsilon,z)$, $r_n^{(2)}(\varepsilon,z)$ — условные ε-риски решающих правил $d^{(1)}(y)$ и $d^{(2)}(y)$ в точке $z \in \mathbf{Z}$. Очевидно, чем больше $\Delta r_n(\varepsilon,z)$, тем больший выигрыш дает решающее правило $d^{(2)}(y)$.

В (Малюгин, 2009б) получены асимптотические разложения условного ε-риска рассматриваемых решающих правил при одновременном использовании следующих трех асимптотик:

А) $h(n) \to 0,\ nh(n) \to \infty,\ n \to \infty$ (растущий объем обучающей выборки);

Б) $\bar{\sigma}^2 \equiv \min\limits_{i=1,...,N}\{\bar{\sigma}_{ii}\} \to \infty$, где $\Sigma^{-1}=(\bar{\sigma}_{ij})$ (усиливающаяся статистическая зависимость компонент вектора признаков);

В) $\varepsilon \to 0$ (повышающаяся точность приближения $r_n^{(l)}(\varepsilon,z)$ к $r_n^{(l)}(z)$).

Введем величину $\beta(n,\bar{\sigma})=\bar{\sigma}h(n)$ и рассмотрим последовательности однотипных ситуаций, которые могут иметь место в асимптотиках А и Б. В зависимости от соотношения между величиной $\bar{\sigma}$, характеризующей степень зависимости признаков, и величиной $h(n)$, зависящей от объема выборки, эти ситуации можно интерпретировать следующим образом:

С1) большой объем выборки, если $\beta(n,\bar{\sigma}) \to 0$;

С2) малый объем выборки, если $\beta(n,\bar{\sigma}) \to \lambda$ $(0<\lambda<\infty)$;

С3) очень малый объем выборки, если $\beta(n,\bar{\sigma}) \to \infty$.

В (Малюгин, 2009б) показано, что преимущество решающего правила $d^{(2)}(y)$ над правилом $d^{(1)}(y)$ по точности, характеризуемое величиной $\Delta r_n(\varepsilon,z)$, для ситуаций С1, С2 и С3 определяется соотношениями:

С1) $\Delta r_n(\varepsilon,z) \to 0$ (нет преимущества);

С2) $\Delta r_n(\varepsilon,z) \to \lambda > 0$ (есть некоторое преимущество);

С3) $\Delta r_n(\varepsilon,z) \to \infty$ (есть значительное преимущество).

Таким образом, подстановочное решающее правило $d^{(1)}(y)$ в ситуациях С2 и С3, возникающих в случае малого объема выборки, проигрывает по точности решающему прави-

лу $d^{(2)}(y)$, и этот проигрыш обусловлен увеличением смещения оценок плотности с фиксированным ядром. Аналогичным образом можно объяснить и более высокую точность прогнозов в виде моды условного распределения (16) по сравнению с прогнозами, полученными с помощью ядерной оценки функции регрессии (17) с постоянным параметром масштаба.

5. Результаты экспериментального исследования алгоритмов

Численные эксперименты преследуют две цели: 1) иллюстрация работоспособности и эффективности алгоритмов, основанных на непараметрической ядерной оценке многомерной плотности распределения наблюдений с адаптивным гауссовским ядром; 2) иллюстрация корректности результатов аналитических исследований указанных алгоритмов. В качестве альтернативы предлагаемым алгоритмам классификации и прогнозирования рассматриваются алгоритмы, использующие ядерные оценки плотности и функции регрессии с фиксированным гауссовским ядром.

Прогнозирование (оценка) класса состояния сложной системы. В рамках описанной выше модели зависимости признаков полагается: $L = 2$, $N=1$, $M=2$; объемы обучающих выборок классов $n_1 = n_2 = n/2$; объем экзаменационной выборки равен n_3; априорные вероятности классов $\pi_1 = \pi_2 = 0.5$; $y'_{\alpha,i} = (x'_{\alpha,i},\, z'_{\alpha,i,1},\, z'_{\alpha,i,2}) \in \Re^3$ — составной вектор признаков для класса $\alpha \in \{1,2\}$, компоненты которого связаны статистической моделью зависимости вида $x_{\alpha,i} = f_\alpha\left(z_{\alpha,i,1}, z_{\alpha,i,2}\right) + \xi_{\alpha,i}$, где $(z_{\alpha,i,1},\, z_{\alpha,i,2}) \in \mathbf{Z} = [a,b]\times[a,b]$ и $\xi_{\alpha,i}$ — взаимно независимые случайные величины, имеющие соответственно равномерный в области $\mathbf{Z} = [a,b]\times[a,b]$ и нормальный $N_1(0,\sigma^2)$ законы распределения.

Полагается, что для фиксированных $(z_1,\, z_2) \in \mathbf{Z}$ условие функциональной неоднородности имеет вид: $f_2\left(z_1, z_2\right) = f_1\left(z_1, z_2\right) + \delta$, где $\delta \in \Re^1$ — величина, характеризующая степень разделимости классов состояний: при $\delta \to 0$ классы состояний системы становятся трудно различимыми.

На тестовых примерах с различными функциями $f_2\left(z_1, z_2\right)$, $f_1\left(z_1, z_2\right)$ исследуется точность следующих алгоритмов классификации наблюдений за системой с целью оценки (прогнозирования) класса состояния системы:

- БРП (13), использующие условные плотности распределений вида (5) (алгоритм А0);
- непараметрический классификатор $d^{(1)}(y)$, использующий единичную матрицу гауссовского ядра $H^{(1)}$ (алгоритм А1);
- непараметрический классификатор $d^{(2)}(y)$, использующий оценку матрицы гауссовского ядра $H^{(2)}$ (алгоритм А2);
- предлагаемый непараметрический классификатор $d^{(3)}(y)$, использующий локальные выборочные оценки матрицы ядра $\{H^{(i,m(i))}\}$ (алгоритм А3).

В качестве характеристик точности классификации используются оценки безусловной вероятности ошибок по обучающей выборке

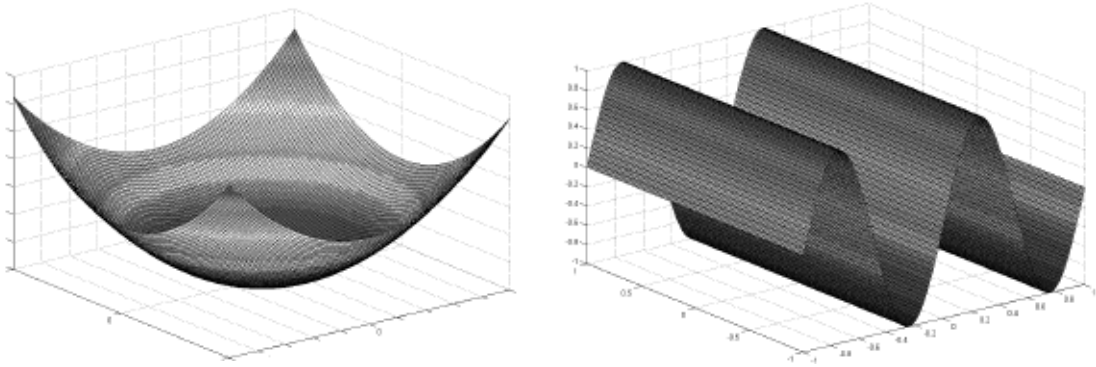
$$\widehat{P}^{(l)} = \frac{1}{2}(\widehat{P}_1^{(l)} + \widehat{P}_2^{(l)}),\ \ l = 0, 1, 2, 3,$$

где $\{\widehat{P}_\alpha^{(l)},\ \alpha = 1,2\}$ — оценки условных вероятностей ошибок по обучающей выборке для классов $\left\{\Omega_\alpha\right\}$ и аналогично определяемые оценки безусловной вероятности ошибок по экзаменационной выборке $\overline{P}^{(l)} = \frac{1}{2}(\overline{P}_1^{(l)} + \overline{P}_2^{(l)}),\ l = 0, 1, 2, 3.$

Приведем результаты экспериментального исследования рассматриваемых алгоритмов для двух тестовых примеров (рис. 1):

а) $f_1(z_1, z_2) = \frac{1}{4}(z_1^2 + z_2^2)$, $\delta = 1$, $\sigma^2 = 10^{-4}$, $[a;b] = [-5;5]$;

б) $f_1(z_1, z_2) = \sin(2\pi z_1)$, $\delta = \frac{1}{2}$, $\sigma^2 = 10^{-2}$, $[a;b] = [-1;1]$.



а) б)

Рис. 1. Графики функции $x = f_1(z_1, z_2)$ для тестовых примеров а) и б)

В таблицах 1 и 2 содержатся усредненные по 10 независимым прогонам оценки вероятностей ошибок альтернативных алгоритмов для тестовых примеров а) и б) соответственно.

Таблица 1. Оценки вероятностей ошибок для тестового примера а)

n	n_3	Алгоритмы	$\widehat{P}^{(l)}$ $(l = 0,1,2,3)$	$\overline{P}^{(l)}$ $(l = 0,1,2,3)$
200	400	A0	0.000	0.000
		A1	0.268	0.230
		A2	0.298	0.243
		A3	0.040	0.080
400	700	A0	0.000	0.000
		A1	0.134	0.144
		A2	0.186	0.171
		A3	0.010	0.030
600	1000	A0	0.000	0.000
		A1	0.113	0.097
		A2	0.158	0.136
		A3	0.008	0.016
800	1300	A0	0.000	0.000

Таблица 2. Оценки вероятностей ошибок для тестового примера б)

n	n_3	Алгоритмы	$\widehat{P}^{(l)}$ $(l=0,1,2,3)$	$\overline{P}^{(l)}$ $(l=0,1,2,3)$
400	500	A0	0.051	0.038
		A1	0.230	0.232
		A2	0.195	0.168
		A3	0.054	0.084
600	800	A0	0.059	0.073
		A1	0.202	0.218
		A2	0.152	0.173
		A3	0.028	0.085
800	1100	A0	0.064	0.069
		A1	0.178	0.178
		A2	0.136	0.132
		A3	0.021	0.046

Прогнозирование эндогенных переменных. Целью экспериментов является сравнительный анализ точности двух типов прогнозов: прогнозов в виде оценки моды условного распределения (16) (алгоритм B1) и прогнозов, полученных на основе оценки функции регрессии (17) с постоянным параметром масштаба (алгоритм B2). В качестве критерия точности прогнозов используется среднеквадратическая ошибка (*root mean square error — RMSE*):

$$RMSE = \sqrt{\frac{1}{n_3}\sum_{i=1}^{n_3}\left(\widehat{x}_i - x_i\right)^2}\ .$$

В таблице 3 содержатся усредненные по 10 независимым прогонам значения ошибки прогноза *RMSE* для альтернативных алгоритмов в условиях двух тестовых примеров:

в) $f(z) = \exp\{z\},\ \ \sigma^2 = 10^{-2},\ \ [a;b] = [0;2];$

г) $f(z) = \sin(\pi z),\ \ \sigma^2 = 10^{-2},\ [a;b] = [0;2],$

где в обоих случаях $n_3 = 200,\ \ h = n^{-\gamma},\ \gamma = 0.3.$

Таблица 3. Значения ошибки прогноза *RMSE*

n	Алгоритмы	Пример в)	Пример г)
25	B1	0.7513	0.5716
	B2	0.9549	1.4024
50	B1	0.2849	0.3486
	B2	0.5629	0.7523
100	B1	0.2090	0.2330
	B2	0.3211	0.4365
200	B1	0.1412	0.1433
	B2	0.2175	0.1897
400	B1	0.1339	0.1361
	B2	0.1159	0.1108

6. Заключение

Сравнительный анализ точности рассматриваемых алгоритмов в условиях существенно зависимых признаков позволяет сделать следующие выводы:

1) алгоритм, реализующий непараметрический классификатор с адаптивным гауссовским ядром, имеет более высокую точность классификации по сравнению с непараметрическими классификаторами, использующими фиксированное ядро, при этом выигрыш в точности классификации увеличивается с возрастанием степени статистической зависимости признаков; с ростом объема обучающей выборки непараметрический классификатор с адаптивным ядром демонстрирует свойство состоятельности (точность классификации приближается к точности БРП);

2) предлагаемый алгоритм прогнозирования на основе непараметрической оценки моды условной плотности допускает построение многомерных прогнозов $(N > 1)$ и в одномерном случае $(N = 1)$ может иметь преимущество по точности прогнозирования перед непараметрической ядерной оценкой функции регрессии в условиях малой обучающей выборки;

3) результаты экспериментов согласуются с результатами аналитических исследований предлагаемых алгоритмов оценивания и классификации при растущем объеме обучающей выборки и усиливающейся статистической зависимости компонент вектора признаков.

В силу известного недостатка непараметрических ядерных оценок плотности (так называемого «проклятья размерности» (Silverman, 1986)), заключающегося в значительном росте требуемого объема выборки при увеличении размерности пространства признаков, их практическое применение для больших значений p может быть невозможно. В то же время, в рамках рассматриваемой модели данных использование предлагаемых непараметрических ядерных оценок многомерных плотностей с адаптивным ядром позволяет в некоторой степени сократить требуемый объем данных.

Список литературы

Айвазян С. А., Енюков И. С., Мешалкин Л. Д. (1985). *Прикладная статистика. Исследование зависимостей*. М.: Финансы и статистика.

Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. (1989). *Прикладная статистика. Классификация и снижение размерности*. М.: Финансы и статистика.

Андерсон Т. (1963). *Введение в многомерный статистический анализ*. М.: Физматгиз.

Епанечников В. А. (1969). Непараметрическая оценка многомерной плотности вероятностей, *Теория вероятностей и ее применения,* 14 (1), 156–161.

Малюгин В. И. (1985). Об оценивании плотности случайных векторов с существенно зависимыми компонентами. *Вестник БГУ,* Сер. 1, 2, 41–44.

Малюгин В. И., Харин Ю. С. (1986). Об оптимальности классификации случайных наблюдений, различающихся уравнениями регрессии. *Автоматика и телемеханика*, 7, 35–46.

Малюгин В. И. (2008а). Дискриминантный анализ многомерных зависимых регрессионных наблюдений в условиях структурной параметрической неоднородности моделей. *Информатика*, 3, 17–28.

Малюгин В. И. (2008б). Статистический анализ смесей распределений регрессионных наблюдений. *Информатика*, 4, 79–88.

Малюгин В. И. (2009а). Методы анализа эконометрических моделей со структурной неоднородностью. *В кн.: Теория вероятностей, случайные процессы, математическая статистика и приложения*. Минск. БГУ, 87 – 95.

Малюгин В. И. (2009б). Асимптотический анализ риска непараметрической классификации в случае существенно зависимых признаков. *Известия НАН Беларуси*, *Сер. 1*.: Физ. Мат. Информ., 3, 10–23.

Фукунага К. (1979). *Введение в статистическую теорию распознавания образов*. М.: Наука.

Хардле В. (1993). *Прикладная непараметрическая регрессия*. М.: Мир.

Харин Ю. С. (1992). Робастность в статистическом распознавании образов. Минск: Университетское.

Camastra F., Vinciarelli A. (2002) Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24 (10), 1404 – 1407.

Granlund G. H., Knutsson H. (1995). *Signal processing in computer vision*. Kluwer Academic Publishers.

Hall P., Racine J. S., Li Q. (2004). Cross-validation and the estimation of conditional densities. *Journal of American Statistical Association*, 99 (468), 1015 – 1026.

Hardle W., Simar L. (2007). *Applied multivariate statistical analysis*. Springer.

Hubler O., Frohn J. (2006). *Modern econometric analysis: surveys on recent development*. Springer.

Malugin V. I., Vasilkov M. E. (2010). Nonparametric analysis of stochastic systems with nonlinear functional heterogeneity. *Proceedings of the 9th International Conference «Computer Data Analysis and Modeling», Minsk*. Vol. 1, 81 – 84.

Racine J. S. (2008). Nonparametric econometrics: A primer. *Foundations and Trends in Econometrics*, 3 (1), 1 – 88.

Silverman B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Thomas P. R. (1997). *Modern regression methods.* John Wiley.