

A. A. Novosyolov

CONSTRUCTION OF MULTIDIMENSIONAL DISCRETE DISTRIBUTIONS WITH PREASSIGNED CORRELATION STRUCTURE

The paper is devoted to methods of construction of multidimensional discrete distributions with preassigned correlation structure and marginal distributions. The methods are based on mixture of the basic distributions and on decision of some optimization problems.

Keywords: discrete distribution, correlation, copula, mixture.

© Новоселов А. А., 2010

УДК 62-50:519.224

А. А. Новоселов

ПАРАМЕТРИЗАЦИЯ МОДЕЛЕЙ УПРАВЛЯЕМЫХ СИСТЕМ

Описывается применение метода ортогональных рядов для построения моделей управляемых систем параметрического вида в условиях непараметрической неопределенности. Ключевым элементом метода является выбор длины ортогонального ряда по данным наблюдений, т. е. определение параметрической структуры модели. Продемонстрировано применение метода к оцениванию плотности распределения и функции регрессии. Предложены пути обобщения оценок на многомерный случай.

Ключевые слова: плотность распределения, функция регрессии, ортогональный ряд, непараметрическая оценка.

В классической теории управления [1] обычно используются параметрические модели, которые строятся в условиях параметрической неопределенности. Здесь предполагается известным факт принадлежности модели к некоторому заранее определенному конечномерному классу моделей. Процедура идентификации (оценивания) модели сводится при этом к оцениванию по данным наблюдений некоторого конечного набора параметров.

Столь полную априорную информацию можно считать доступной далеко не всегда, поэтому параллельно развивается и непараметрическая теория управления [2; 3], в которой априорный класс моделей предполагается бесконечномерным; при этом говорят о непараметрической неопределенности. Малое количество априорной информации приходится компенсировать при этом большим количеством статистических наблюдений.

Основными строительными блоками для создания математических моделей систем управления по наблюдениям являются оценки плотности распределения и функции регрессии [1–3]. Непараметрическая оценка плотности распределения ядерного типа была впервые рассмотрена в [4; 5], а соответствующая непараметрическая оценка функции регрессии – в работах [6; 7].

Классические параметрические оценки этих функций гораздо проще для человеческого восприятия и лучше приспособлены для анализа, чем непараметрические оценки. Поэтому постоянно предпринимаются попытки построения оценок параметрического вида в

условиях непараметрической неопределенности. Это направление можно назвать параметризацией моделей.

Одной из ветвей этого направления является использование оценок в виде отрезков ортогональных рядов, называемых еще проекционными оценками ввиду прозрачной геометрической аналогии. Оценка такого типа для плотности распределения была предложена Н. Н. Ченцовым в [8], а оценка для функции регрессии – в [9].

Проекционные оценки были подвергнуты тщательному изучению. Так, в статьях [10–13] исследовались вопросы сходимости оценок в различных смыслах, в публикациях [14; 15] проводилось сравнение различных непараметрических оценок. В [16–19] исследовались проекционные оценки в конкретных базах. Во всех этих исследованиях длина отрезка ортогонального ряда выбиралась произвольно. В работах [20–24] предложен некий принцип оптимального выбора длины отрезка ортогонального ряда в проекционных оценках, исследованы свойства получаемых оптимальных проекционных оценок. В настоящей работе приведен обзор результатов [20–24], и представлены соображения по поводу обобщения оптимальных проекционных оценок на многомерный случай.

Проекционное приближение. Пусть L_r^2 – гильбертово пространство функций f , заданных на подмножестве U вещественной оси и интегрируемых с квадратом с весом r в смысле

$$\int_U f^2(x)r(x)dx < \infty,$$

где r – неотрицательная весовая функция на U . Скалярное произведение и норма задаются в L_r^2 выражениями:

$$(f, g) = \int_U f(x)g(x)r(x)dx, \quad \|f\| = \sqrt{(f, f)}.$$

Любая функция f из L_r^2 представима в виде ряда [25]

$$f(x) = \sum_{j=0}^{\infty} a_j g_j(x),$$

в смысле сходимости последовательности его частичных сумм

$$f_N(x) = \sum_{j=0}^N a_j g_j(x) \quad (1)$$

по норме пространства L_r^2 , что означает $\|f - f_N\| \rightarrow 0$ при $N \rightarrow \infty$. Здесь $G = \{g_j, j = 0, 1, \dots\}$ обозначена произвольная полная ортонормированная система в L_r^2 , а $a_j = (f, g_j), j = 0, 1, \dots$ представляют собой коэффициенты Фурье функции f относительно системы G .

Выбор ортонормированной системы G зависит от конкретной задачи; в различных ситуациях могут быть использованы, например, классические ортогональные многочлены [26] или тригонометрическая система функций [27].

Оценка плотности распределения. Пусть $X, X_1, \dots, X_n, \dots$ – последовательность независимых, одинаково распределенных на $U \subseteq R$ случайных величин, и пусть распределение X имеет неизвестную плотность f . Рассмотрим задачу оценивания неизвестной плотности распределения f по выборке наблюдений $X_1, \dots, X_n$ объема n в ситуации, когда априори неизвестен какой-либо конечномерный класс плотностей, которому заведомо принадлежит f , т. е. в ситуации непараметрической неопределенности.

В [4; 5] для этого была предложена ядерная оценка плотности

$$\begin{aligned} \hat{f}_{n,h}(x) &= \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) = \\ &= \frac{1}{n} \sum_{i=1}^n K_h(x - X_i), \quad x \in U, \end{aligned} \quad (2)$$

где K – некоторая плотность распределения на R , называемая ядром оценки, $K_h(u) = K(u/h)/h$; h – параметр оценки, называемый параметром размытости или параметром ширины окна.

В [8] рассматривалась проекционная оценка плотности, основанная на отрезке ряда (1), в котором коэффициенты Фурье a_j заменены на оценки по выборке наблюдений вида

$$\hat{a}_j = \frac{1}{n} \sum_{i=1}^n g_j(X_i) r(X_i), \quad j = 0, 1, \dots$$

Таким образом, проекционная оценка плотности имеет вид

$$\hat{f}_{n,N}(x) = \sum_{j=0}^N \hat{a}_j g_j(x), \quad x \in U. \quad (3)$$

Отметим, что в результате несложных преобразований эту оценку можно представить следующим образом:

$$\hat{f}_{n,N}(x) = \frac{1}{n} \sum_{i=1}^n K_N(x, X_i),$$

$$K_N(x, u) = \sum_{j=0}^N g_j(x) g_j(u) r(x), \quad x, u \in U.$$

Это представление по виду напоминает второе представление в оценке (2), как среднее выборочное значение некоторого ядра. Существенное отличие состоит в том, что ядро в последнем выражении не является плотностью распределения, в частности, оно может принимать отрицательные значения. Таким же недостатком обладает и проекционная оценка (3), она не является, вообще говоря, плотностью распределения, поскольку может принимать отрицательные значения. В некоторых случаях избавиться от этого недостатка позволяет следующий прием. В качестве оценки плотности рассматривается среднее арифметическое первых частичных сумм ряда

$$\tilde{f}_{n,N}(x) = \frac{1}{N+1} \sum_{j=0}^N \hat{f}_{n,j}(x), \quad x \in U.$$

Такой прием позволяет обеспечить неотрицательность оценки плотности, например, при использовании тригонометрической ортонормированной системы [27]. При этом оценка оказывается построенной посредством так называемых сумм Фейера [27].

Длина разложения N выбирается в [8] произвольно, там также приводятся некоторые асимптотические соображения по поводу выбора N . В работах [20–22; 24] предложен и исследован оптимальный принцип выбора длины отрезка ряда, основанный на следующем наблюдении. Обозначим $t_j = g_j(X)r(X), j = 0, 1, \dots$ и вычислим среднеквадратичное отклонение проекционной оценки плотности (3) от неизвестной плотности f . Получаем выражение

$$T_n(N) = E \|f - \hat{f}_{n,N}\|^2 = \|f\|^2 - \frac{n+1}{n} \sum_{j=0}^N a_j^2 + \frac{1}{n} \sum_{j=0}^N E t_j^2. \quad (4)$$

Оказывается [20; 24], что величина $T_n(N)$ при фиксированном n достигает минимума по N , это означает, что при любом объеме выборки имеется оптимальная длина отрезка ряда в проекционной оценке (3), причем оптимальность понимается в смысле среднеквадратического критерия (4). Для приближенной минимизации критерия (4) с использованием выборки наблюдений заметим [20; 24], что слагаемое $\|f\|^2$ в правой части (4) не влияет на положение точки минимума, и его можно отбросить, а остальные члены оцениваются следующим образом:

$$\begin{aligned} \hat{T}_n(N) &= -\frac{n+1}{n} \sum_{j=0}^N \hat{a}_j^2 + \frac{2}{n} \sum_{j=0}^N \tilde{t}_j^2, \\ \tilde{t}_j^2 &= \frac{1}{n} \sum_{i=1}^n g_j^2(X_i) r^2(X_i). \end{aligned} \quad (5)$$

Минимизация $\hat{T}_n(N)$ по N дает оптимальное значение $\hat{N}$ длины отрезка ряда, подстановка которого в выражение (4) дает оптимальную проекционную оценку

$$\hat{f}_n(x) = \sum_{j=0}^{\hat{N}} \hat{a}_j g_j(x), \quad x \in U. \quad (6)$$

В качестве примера рассмотрим тригонометрическую ортонормированную систему на $[-1; 1]$, которая состоит из функций

$$\left\{ \frac{1}{2}, \sin(\pi j x), \cos(\pi j x), x \in [-1; 1], j = 1, 2, \dots \right\}.$$

При использовании этой ортонормированной системы удобно включать члены в ортогональное разложение парами, при этом оптимальная проекционная оценка приобретает вид

$$\hat{f}_n(x) = \frac{1}{2} + \sum_{j=1}^{\hat{N}} (\hat{c}_j \cos(\pi j x) + \hat{s}_j \sin(\pi j x)). \quad (7)$$

Здесь

$$\hat{c}_j = \frac{1}{n} \sum_{i=1}^n \cos(\pi j X_i), \quad \hat{s}_j = \frac{1}{n} \sum_{i=1}^n \sin(\pi j X_i), \quad j = 1, 2, \dots$$

Рассмотрим случайную величину X , имеющую распределение с треугольной плотностью на $[-1; 1]$, т. е. $f(x) = 1 - |x|$, $x \in [-1; 1]$. Методами статистического моделирования нетрудно получить выборку наблюдений такой случайной величины, для целей иллюстрации мы возьмем объем выборки равным $n = 50$. График оценки критериальной функции (5), построенный по полученной выборке, приведен на рис. 1. Видно, что в оптимальное разложение входит только первая пара тригонометрических функций ($N = 1$). Графики модельной плотности распределения f и ее оптимальной проекционной оценки показаны на рис. 2. С ростом объема выборки качество оценивания довольно быстро возрастает. Более подробный анализ оптимальных проекционных оценок плотности можно найти в работах [20–24].

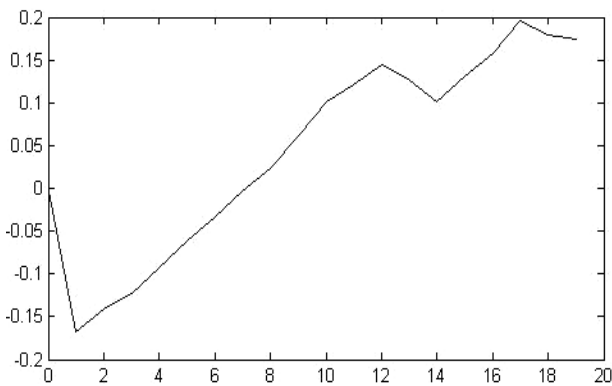


Рис. 1. График оценки критериальной функции (5), построенный по выборке наблюдений объема $n = 50$

Оценка функции регрессии. Пусть теперь (X, Y) , $(X_1, Y_1), \dots, (X_n, Y_n), \dots$ — последовательность незави-

симых одинаково распределенных случайных векторов. Обозначим $f(x) = E(Y|X = x)$, $x \in U$ функцию регрессии Y на X . Для непараметрического оценивания этой функции по наблюдениям (X_i, Y_i) , $i = 1, \dots, n$ в [6; 7] была предложена ядерная оценка

$$\hat{f}_n(x) = \frac{\frac{1}{nh} \sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h}\right)}{\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)}, \quad x \in U.$$

Здесь объекты K, h имеют тот же смысл, что и для ядерной оценки плотности (2).

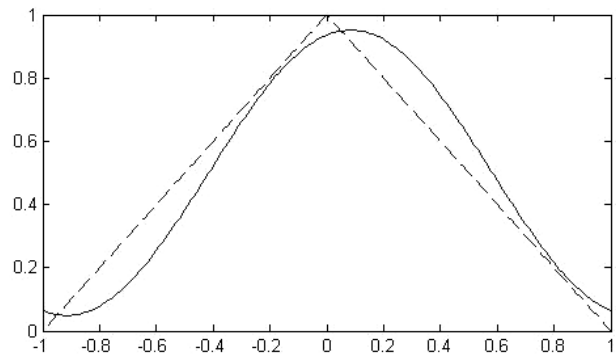


Рис. 2. График модельной треугольной плотности распределения и ее оптимальной проекционной оценки (7), объем выборки $n = 50$

Для простоты предположим, что $U = [0; 1]$ и случайная величина X имеет равномерное распределение на $[0; 1]$ так, что ее плотность равна $P(x) = 1$, $x \in [0; 1]$. Положим также, что весовая функция постоянна: $r(x) = 1$, $x \in [0; 1]$. При этом коэффициенты Фурье имеют вид

$$a_j = \int_0^1 f(x) g_j(x) dx = E(Y g_j(X)), \quad j = 0, 1, \dots$$

Поэтому проекционная оценка функции регрессии имеет вид (3), где оценки коэффициентов Фурье вычисляются по формулам:

$$\hat{a}_j = \frac{1}{n} \sum_{i=1}^n Y_i g_j(X_i), \quad j = 0, 1, \dots$$

Все свойства критерия выбора длины ортогонального разложения сохраняются и в случае оценивания функции регрессии, оптимальная проекционная оценка вычисляется по той же формуле (6), а качество приближения получается даже лучше, чем для плотности распределения, что связано со спецификой данной задачи.

Случай известной, но не равномерной плотности распределения X отличается лишь повышенной громоздкостью формулировок. В случае неизвестной плотности X можно использовать фейеровские оценки [23; 24].

Многомерное обобщение. Представляет интерес обобщение описанных оценок на многомерный случай, когда область распределения случайной вели-

ны X имеет размерность, превышающую 1. Для существования такого обобщения по существу нужно лишь уметь строить многомерные ортонормированные системы. Это легко достигается на основе одномерных ортонормированных систем, что мы и продемонстрируем для двумерного случая.

Пусть $\{g_j(x), j = 0, 1, \dots\}$, $x \in U$ и $\{h_k(y), k = 0, 1, \dots\}$, $y \in V$ – одномерные полные ортонормированные системы на соответствующих областях определения. Тогда полная ортонормированная система на $U \times V$ образуется из всевозможных произведений одномерных функций, а именно:

$$G_{jk}(x, y) = g_j(x)h_k(y), \\ j = 0, 1, \dots, k = 0, 1, \dots, (x, y) \in U \times V.$$

Следует иметь в виду, что с ростом размерности очень быстро нарастают комбинаторные сложности в вычислении ортонормированных функций, поэтому применимость описанного способа оценивания в пространствах высокой размерности сомнительна. Кроме того, метод проекционного оценивания не обладает свойствами локальных оценок, вследствие чего не может использовать «трубчатое» строение выборочных данных. По-видимому, возможно построение специальных ортонормированных систем, адаптированных под имеющуюся структуру данных, что позволит использовать такое специальное строение, однако, нам неизвестны исследования в этом направлении.

В работе дан краткий обзор оптимальных проекционных оценок плотности распределения и функции регрессии, которые применяются при построении моделей управляемых систем по наблюдениям. Описаны основные алгоритмы, участвующие в построении оценок, приведены примеры и иллюстрации. Отмечается возможность многомерного обобщения процедуры оценивания и имеющиеся ограничения по размерности.

Библиографические ссылки

1. Цыпкин Я. З. Адаптация и обучение в автоматических системах. М. : Наука, 1968.
2. Живоглазов В. П., Медведев А. В. Непараметрические алгоритмы адаптации. Фрунзе : Илим, 1974.
3. Медведев А. В. Непараметрические системы адаптации. Новосибирск : Наука, 1983.
4. Parzen E. On estimation of a probability density function and mode // Ann. Math. Statist. 1962. № 33. P. 1065–1076.
5. Rosenblatt M. Remarks on some nonparametric estimates of a density function // Ann. Math. Statist. 1956. № 27. P. 832–837.
6. Надарая Э. А. Об оценке регрессии // Теория вероятностей и ее применения. 1964. Т. 10. С. 157–159.
7. Watson G. S. Smooth Regression Analysis // Sankhya. Ser. A. 1964. № 26. P. 359–378.
8. Ченцов Н. Н. Оценка неизвестной плотности распределения по наблюдениям // ДАН СССР. 1962. № 147. С. 45–48.
9. Greblicky W. Nonparametric system identification by orthogonal series // Probl. Control Inform. Theory. 1979. № 8. P. 67–73.
10. Anderson G. L., Figueiredo R. J. P. An adaptive orthogonal series estimator for probability density functions // Ann. Statist. 1980. № 8. P. 347–376.
11. Bosq M. D. Sur l'estimation d'une densite multivariee par une serie de fonctions orthogonales // C. R. Acad. Sci. Ser. A. Paris, 1969. № 268. P. 555–557.
12. Crain B. R. A note on density estimation using orthogonal expansions // J. Amer. Statist. Assoc. 1973. № 68. P. 964–965.
13. Watson G. S. Density estimation by orthogonal series // Ann. Math. Statist. 1969. № 40. P. 1496–1498.
14. Fryer M. J. A review of some nonparametric methods of density estimation // J. Inst. Math. Appl. 1970. № 20. P. 335–354.
15. Wahba G. Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation // Ann. Statist. 1975. № 3. P. 15–29.
16. Hall P. Estimating a density on the positive half line by the method of orthogonal series // Ann. Inst. Statist. Math. 1980. № 32. Part A. P. 351–362.
17. Hall P. On trigonometric series estimates of densities // Ann. Statist. 1981. № 9. P. 683–685.
18. Kronmal R., Tarter M. The estimation of probability densities and cumulative by Fourier series methods // J. Amer. Statist. Assoc. 1968. № 63. P. 925–952.
19. Walter G. G. Properties of Hermite series estimation of probability density // Ann. Statist. 1977. № 5. P. 1258–1264.
20. Новоселов А. А. О выборе структуры моделей управляемых систем // Стохастические системы управления. Новосибирск : Наука, 1979. С. 72–77.
21. Новоселов А. А. Об оптимальных ортогональных оценках // Математическая статистика и ее приложения. Томск : Изд-во Том. ун-та, 1981. С. 151–155. Вып. 7.
22. Новоселов А. А. Оптимальная параметризация плотности вероятности и функции регрессии // Адаптация и обучение в системах управления и принятия решений. Новосибирск : Наука, 1982. С. 139–147.
23. Новоселов А. А. Оценивание кривой регрессии методом сумм Фейера // Математическая статистика и ее приложения. Томск : Изд-во Том. ун-та, 1986. С. 177–183. Вып. 10.
24. Новоселов А. А. Разработка и исследование методов параметризации структур адаптивных систем : автореф. дис. канд. физ.-мат. Наук. Томск : Изд-во Том. ун-та, 1987.
25. Канторович Л. В., Акилов Г. П. Функциональный анализ. М. : Наука, 1984.
26. Суетин П. К. Классические ортогональные многочлены. М. : Наука, 1979.
27. Бари Н. К. Тригонометрические ряды. М. : Физматгиз, 1961.

PARAMETRIZATION OF MODELS OF CONTROLLED SYSTEMS

The paper describes application of orthogonal series method for construction of controlled systems models under non-parametric uncertainty. A key element of the method is draw of orthogonal expansion length based on observations, in other words, defining parametric structure of the model. The method is demonstrated for estimation of distribution density and regression function. Directions for generalizing onto multi-dimensional case are also presented.

Keywords: distribution density, regression function, orthogonal series, non-parametric estimate.

© Новоселов А. А., 2010

УДК 004.932.2

И. А. Пестунов, В. Б. Бериков, Ю. Н. Синявский

СЕГМЕНТАЦИЯ МНОГОСПЕКТРАЛЬНЫХ ИЗОБРАЖЕНИЙ НА ОСНОВЕ АНСАМБЛЯ НЕПАРАМЕТРИЧЕСКИХ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ*

Предложен метод сегментации многоспектральных изображений на основе ансамбля непараметрических алгоритмов; дано теоретическое обоснование метода. Результаты статистического моделирования на модельных данных и реальных изображениях подтверждают эффективность предложенного метода.

Ключевые слова: сегментация многоспектральных изображений, непараметрические алгоритмы кластеризации, ансамблевый подход.

Сегментация является одним из важнейших этапов анализа цифровых изображений [1]. Она заключается в разбиении изображения на сегменты на основе подобия спектральных, текстурных и других характеристик пикселей. Методы сегментации нашли широкое применение во многих прикладных областях, в том числе в дистанционном зондировании Земли (ДЗЗ) [2; 3], интерес к которому в последние годы непрерывно возрастает.

Один из наиболее распространенных подходов к сегментации многоспектральных изображений основан на статистических методах кластеризации [4; 5]. В этом случае задачу кластеризации, как правило, приходится решать при отсутствии каких-либо априорных сведений о числе классов и их вероятностных характеристиках. Для этого наиболее подходящими являются непараметрические алгоритмы, позволяющие получить хорошие результаты при минимальной априорной информации. Их общим недостатком является высокая чувствительность к входным параметрам, что существенно усложняет процесс настройки алгоритма для решения конкретной задачи.

Известно [6–8], что устойчивость решений в задачах кластеризации может быть повышена благодаря формированию ансамбля алгоритмов и построению на его основе коллективного решения. При этом используются результаты, полученные различными алгоритмами либо одним алгоритмом с различными

значениями параметров, по разным подсистемам переменных и т. д. В настоящее время ансамблевый подход является одним из наиболее перспективных направлений в кластерном анализе [9].

В данной работе предложен алгоритм сегментации многоспектральных изображений с использованием ансамбля непараметрических алгоритмов кластеризации, основанных на оценках плотности Розенблатта–Парзена [10; 11]. Для формирования ансамбля используются результаты выполнения непараметрического алгоритма MeanSC (представляющего собой оптимизацию предложенного ранее алгоритма [12]) с различными значениями параметра сглаживания h . Итоговое коллективное решение строится на основе попарной классификации объектов. Дано теоретическое обоснование предложенного алгоритма, приведены результаты статистического моделирования на модельных данных и реальных изображениях, подтверждающие его эффективность.

Непараметрический подход к задаче кластеризации данных ДЗЗ. Предположим, что произведена k -спектральная съемка участка местности, содержащего N элементов разрешения, тогда результат съемки можно представить в виде множества $X = \{x^{(i)} = (x_1^{(i)}, \dots, x_k^{(i)}) \in R^k, i = \overline{1, N}\}$, где $x_j^{(i)}$ – значение яркости i -го элемента разрешения в j -м диапазоне спектра ($j = \overline{1, k}$).

* Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (код проекта 09-07-12087-офи_м).