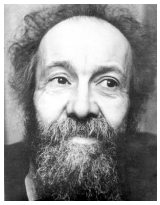

ТРУДЫ КОНФЕРЕНЦИИ,
ПОСВЯЩЕННОЙ 90-ЛЕТИЮ СО ДНЯ РОЖДЕНИЯ
Алексея Андреевича Ляпунова
Новосибирск, 8-12 октября 2001 г.

Об идентификации многозначных характеристик в задачах стохастического моделирования *

Кирик Е.С.
Институт Вычислительного Моделирования СО РАН,
Красноярск, Россия

Аннотация

The paper deals with the problem of identification and restoration of ambiguous dependencies in the tasks of stochastic modelling. Nonparametrical approach is proposed. With some restriction on the learning sample decision of the task is described. Some ideas is presented for the common case of the absence of any addition information about view of function of restoration.

1 Введение

Задача восстановления (идентификации) неизвестных зависимостей по наблюдениям известна давно. Среди основных подходов к ее решению можно выделить параметрический, предполагающий выбор класса моделей на основе априорной информации об объекте, и непараметрический, где в качестве модели принимается регрессия, а вероятностные характеристики заменяются их ядерными оценками типа Розенблата-Парзена. Это обстоятельство позволяет решать задачи в условиях малого объема априорной информации, что весьма привлекательно при решении практических задач. Однако нельзя не заметить, что наблюдается явный недостаток в методах, позволяющих восстанавливать многозначные зависимости [3]. Основная проблема непараметрических алгоритмов заключается в том, что в их основе лежит идея локальной аппроксимации значения функции в некоторой точке. Искомое значение ищется в виде взвешенного среднего близлежащих к нему элементов обучающей выборки, близкими элементами считаются те, у которых значения аргументов попадают в некоторую окрестность рассматриваемой точки. Как следствие алгоритм не чувствителен к возможному (в случае многозначности) значительному разбросу измеренных значений функции среди элементов выборки, попавших в окрестность данной точки. В работе излагается подход к восстановлению многозначных зависимостей

*Работа поддержана Красноярским краевым фондом науки, №10F123N и №12G188.

при определенных ограничениях на характер накопления обучающей выборки. Такой подход может быть реализован, например, для восстановления траектории движения объекта или для восстановления переходных характеристик в линейных динамических системах. Также обсуждается вопрос о возможных путях решения данной задачи без привлечения дополнительной информации.

2 Постановка задачи

Пусть дана обучающая выборка (ОВ) $V = \{x_i, y_i\}$, $i = \overline{1,n}$ — статистическая выборка независимых наблюдений случайной величины (x, y) , распределенной с неизвестной плотностью вероятности $p(x.y)$, $p(x) > 0 \forall x \in X$. Априори вид нелинейной стохастической зависимости $y = f(x)$ не задан. Требуется восстановить неизвестную зависимость по наблюдениям; предполагается, что она многозначная.

3 Идентификация при ограничениях на ОВ

В общем случае ($x_i = (x_i^1, ..., x_i^l)$) сходящаяся [4] непараметрическая оценка условного математического ожидания $\hat{y} = \hat{f}(x) = M(y/x) = \int yp(y/x)\,dy$ (или регрессии) имеет вид

$$y_n(x) = \frac{\sum\limits_{i=1}^{n} y_i \prod\limits_{j=1}^{l} \Phi\left(\frac{x^j - x_i^j}{C_n}\right)}{\sum\limits_{i=1}^{n} \prod\limits_{j=1}^{l} \Phi\left(\frac{x^j - x_i^j}{C_n}\right)}. \tag{1}$$

Оценка (1) получается из уравнения регрессии подстановкой в него оценок плотностей типа Розенблата-Парзена ([6], [5]) с учетом условия самовоспроизводимости $\Phi(\cdot)$

$$\frac{1}{C_n} \int\limits_{\Omega(y)} y\Phi\left(\frac{y^j - y_i^j}{C_n}\right) dy = y_i, i = \overline{1,n}, j = \overline{1,l}. \tag{2}$$

$\Phi(\cdot)$ – финитная колоколообразная интегрируемая с квадратом функция, удовлетворяющая условиям

$$0 < \Phi(z) < \infty, \forall z \in \Omega(z); \quad \frac{1}{C_n} \int\limits_{\Omega(x)} \Phi\left(\frac{x - x_i}{C_n}\right) dx = 1;$$

$$\lim_{n\to\infty} \frac{1}{C_n} \Phi\left(\frac{x - x_i}{C_n}\right) = \delta(x - x_i); \tag{3}$$

C_n – параметр размытости такой, что

$$C_n > 0; \quad \lim_{n\to\infty} C_n = 0; \quad \lim_{n\to\infty} nC_n^k = \infty. \tag{4}$$

Последний является неизвестным параметром в (1), подлежащим определению. Оптимальный параметр размытости C_n соответствует минимуму квадратичного критерия оптимальности

$$\omega^2(C_n) = \sum_{i=1}^{n} \big(y(x_i) - \tilde{y}_n(x_i, C_n)\big)^2 \to \min_{C_n} \tag{5}$$

и находится в ходе скользящего экзамена на обучающей выборке.

По указанным выше причинам алгоритм (1) не может быть использован для качественного восстановления многозначных зависимостей. Однако если предположить, что неизвестная зависимость — траектория "движения"некоторого объекта, а обучающая выборка — последовательно измеренные характеристики положения объекта, то в такой ситуации представляется возможным средствами непараметрического регрессионного анализа восстановить искомую многозначную функцию ([1]). Кроме того свойство локальной аппроксимации непараметрической оценки регрессии — удобный инструмент, позволяющий "отслеживать"реальное количество значений функции и восстанавливать их в каждой интересующей точке пространства аргументов.

При данных ограничениях на характер поступления элементов в обучающую выборку становится очевидным тот факт, что каждому значению функции в точке будет соответствовать свой набор элементов из обучающей выборки, объединенных не только близостью значений аргументов, но и кучностью времени поступления в выборку, что при данных ограничениях соответствует близости выборочных точек в пространстве значений функции. Таким образом время, в роли которого выступает порядковый номер элемента, становится параметром, позволяющим однозначно определять, сколько значений имеет восстанавливаемая функция в данной точке, и каждое значение функции восстанавливать по соответствующим ему элементам обучающей выборки. Таким образом задача восстановления многозначной зависимости в точке сводится к последовательному решению нескольких задач восстановления однозначных зависимостей.

Назовем "ненулевым множеством"точки x такой набор элементов обучающей выборки $I_{\bar{0}}(x) = \{x_i\}, i = \overline{b, e}$, для которых выполняются следующие условия:

$$\prod_{j=1}^{l} \Phi\Big(\frac{x^j - x^j_{b-1}}{C_n}\Big) = 0; \prod_{j=1}^{l} \Phi\Big(\frac{x^j - x^j_i}{C_n}\Big) \neq 0, i = \overline{b, e}; \prod_{j=1}^{l} \Phi\Big(\frac{x^j - x^j_{e+1}}{C_n}\Big) = 0, \tag{6}$$

где x_b, x_e - нижняя и верхняя грани "ненулевого множества"точки x соответственно. Очевидно, что элементы $I_{\bar{0}}(x)$ объединены не только про принципу близости аргументов, но и по временному принципу. Поэтому если искомая функция однозначная, то каждая точка пространства аргументов имеет только одно "ненулевое множество", если многозначная - число $I_{\bar{0}}(x)$ в каждой точке равно числу значений функции в этой точке. Выпишем формулу для подсчета числа значений функции в точке x

$$a(x) = \frac{sgn\big(\prod_{j=1}^{l} \Phi\big(\frac{x^j - x^j_1}{C_n}\big)\big) + \sum_{i=2}^{n} sgn\big(\prod_{j=1}^{l} \Phi\big(\frac{x^j - x^j_i}{C_n}\big)\big)\big) I(x, x_{i-1}, x_{i+1}) + sgn\big(\prod_{j=1}^{l} \Phi\big(\frac{x^j - x^j_n}{C_n}\big)\big)}{2}, \tag{7}$$

где

$$I(x, x_{i-1}, x_{i+1}) = \begin{cases} 1, \left(\prod\limits_{j=1}^{l} \Phi\left(\frac{x^j - x_{i-1}^j}{C_n}\right)\right)\left(\prod\limits_{j=1}^{l} \Phi\left(\frac{x^j - x_{i+1}^j}{C_n}\right)\right) = 0, \\ 0, \left(\prod\limits_{j=1}^{l} \Phi\left(\frac{x^j - x_{i-1}^j}{C_n}\right)\right)\left(\prod\limits_{j=1}^{l} \Phi\left(\frac{x^j - x_{i+1}^j}{C_n}\right)\right) \neq 0. \end{cases} \tag{8}$$

Идея, лежащая в основе этой формулы, заключается в том, что находятся нижние и верхние грани всех "ненулевых множеств"точки x (для этого вводится индикатор (8)) и подсчитывается их количество. Половину от этого числа как раз и составляет искомое количество значений функции в точке x.

Тогда оценка регрессии будет повторять оценку (1) за одним исключением. В каждой точке восстанавливается не заведомо одно, а последовательно $a(x)$ значений функции, причем каждое значение восстанавливается по соответствующему "ненулевому множеству". Имеем следующую непараметрическую оценку

$$_p y_n(x) = \frac{\sum\limits_{i=b_p}^{e_p} y_i \prod\limits_{j=1}^{l} \Phi\left(\frac{x^j - x_i^j}{C_n}\right)}{\sum\limits_{i=b_p}^{e_p} \prod\limits_{j=1}^{l} \Phi\left(\frac{x^j - x_i^j}{C_n}\right)}, p = \overline{1, a(x)}, \tag{9}$$

где b_p, e_p — нижняя и верхняя границы соответственно p-ого ненулевого множества точки x.

Очевидно, что при восстановлении однозначной зависимости и выполнении наложенных условий на выборку оценки (1) и (9) совпадут.

Как и прежде одним настраиваемым параметром в непараметрической оценке (9) является параметр размытости C_n. Его следует находить в ходе скользящего экзамена на обучающей выборке. При этом необходимо учитывать следующую особенность задачи. Несмотря на то, что искомая зависимость, вообще говоря, является многозначной каждый элемент обучающей выборки является уникальным, то есть известно лишь одно значение $y_i, i = \overline{1, n}$ в каждой из точек $x_i, i = \overline{1, n}$ (хотя реально истинная функция может иметь в данной точке и не одно значение). И так как смысл скользящего экзамена заключается в том, чтобы как можно ближе восстановить все элементы выборки по другим, то восстанавливаться должно именно имеющееся выборочное значение функции в рассматриваемой точке выборки. Поэтому здесь нельзя пользоваться традиционной технологией, когда скользящий экзамен и восстановление характеристики осуществляются по одному алгоритму. И основное расхождение заключается в том, что из всех "ненулевых множеств"каждой точки выборки x_i необходимо выбрать именно то, которое соответствует выборочному значению y_i. Эта проблема решается следующим образом. Как уже было отмечено в "ненулевых множествах", элементы выборки объединены по принципу близости значений аргументов и принципу кучности времени поступления (то есть близости порядковых номеров), что соответствует близости в пространстве значений функции. Так как каждый элемент выборки имеет свой порядковый номер, то становится очевидным, что из числа всех $I_{\bar{0}}(x)$ следует использовать для восстановления выборочного значения то "ненулевое множество", порядковые номера элементов которого близки к номеру рас-

сматриваемого элемента. Таким образом оценка примет вид

$$y_n(x_k) = \frac{\sum\limits_{i=b_{k-1}}^{e_k} y_i \prod\limits_{j=1}^{l} \Phi\left(\frac{x^j - x_i^j}{C_n}\right)}{\sum\limits_{i=b_{k-1}}^{e_k} \prod\limits_{j=1}^{l} \Phi\left(\frac{x^j - x_i^j}{C_n}\right)}, k = \overline{1,n}, b_1 = 1; \tag{10}$$

Как и прежде, оптимальный параметр размытости C_n удовлетворяет минимуму квадратичного критерия оптимальности (5).

Итак, задача восстановления многозначной функции при указанных ограничениях на обучающую выборку была сведена к последовательному восстановлению всех значений функции в заданных точках. Для этого была использована обычная непараметрическая оценка регрессии с той лишь разницей, что в силу особенностей задачи каждое значение функции оценивалось не по всей выборке, а по заранее определенной локальной области. Известно [4], что оценка (5) является сходящейся в среднеквадратическом и асимптотически несмещенной:

$$\lim_{n\to\infty} E\{(y(x) - y_n(x))^2\} = 0, \forall x \in X \tag{11}$$

$$\lim_{n\to\infty} E\{y_n(x)\} = y(x), \forall x \in X. \tag{12}$$

Как следствие, оценка (9) обладает теми же свойствами.

О качестве работы представленного алгоритма можно судить по графикам восстановленных функций (Рис.1, 2).

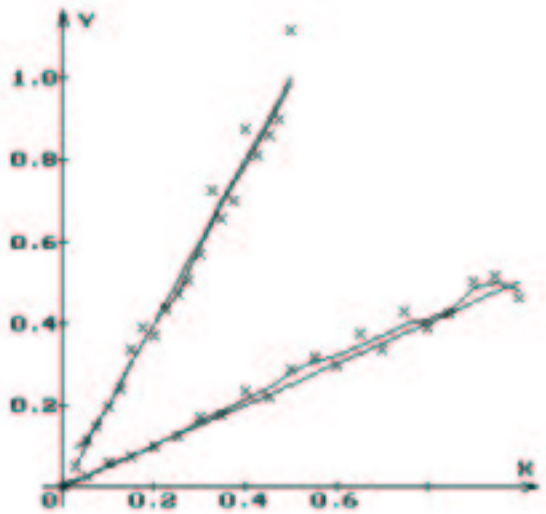

Рис. 1. Неизвестная функция, подлежащая восстановлению, ее оценка и ОВ.

Следует заметить, что алгоритм позволяет восстанавливать не только траектории движения объектов в прямом смысле этого слова. Это могут быть любые зависимости при условии, что элементы обучающей выборки представляют измерения, характеризующие последовательные стадии протекания исследуемого физического процесса во времени. И

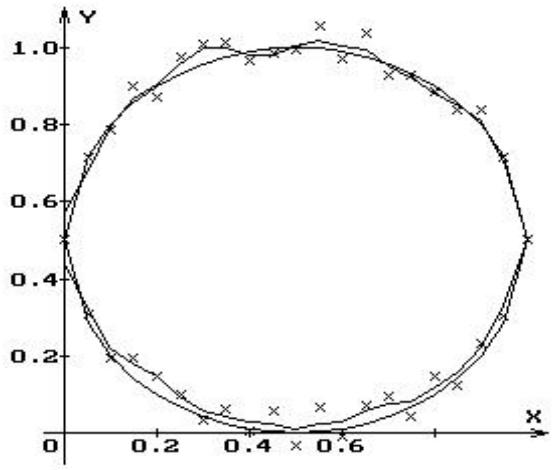

Рис. 2. Неизвестная функция, подлежащая восстановлению, ее оценка и ОВ.

в этом смысле их можно тоже назвать траекторными. В частности, алгоритм может быть использован в задаче идентификации динамической системы, опысывающей циклический процесс. В линейном случае, когда система описывается интегралом Дюамеля [4], алгоритм можно использовать для восстановления переходных характеристик. В общем случае, не предполагающем линейности системы, ее можно аппроксимировать регрессией путем замыкания выхода системы на вход.

4 Идентификация без ограничений на ОВ

Решение задачи идентификации многозначных зависимостей в отсутствии такой мощной дополнительной информации, как наличие временной упорядоченности элементов ОВ, является гораздо более сложным и трудоемким. Ввиду огромного разнообразия функций создание универсального алгоритма не представляется возможным. Скорее всего алгоритм следует разбить на два этапа. Первый будет заключаться в предварительном анализе ОВ, цель которого — определение типа неоднозначности, и затем на основе полученной информации непосредственно восстановление функции.

Наибольший интерес и трудность представляет первый этап, так как состоит в классификации восстанавливаемой функции в отсутствии дополнительных сведений о ее виде (например сведений о ее многозначности на всей области определения, количества значений функции). Предварительный специальный анализ обучающей выборки может значительно восполнить этот недостаток. Очевидно, если восстановить многозначную функцию обычной оценкой регрессии (1), то качество результата будет весьма далеко от желаемого, кроме того эта оценка будет однозначной и будет представлять некое локальное среднее на всей области определения функции. Но весьма информативным представляется последующее исследование невязок $\epsilon_i = y_i - y_n(x_i), i = \overline{1,n}$, где $y_n(x_i), i = \overline{1,n}$ — оценки (1) каждого элемента обучающей выборки, полученные в ходе скользящего экзамена. Это так потому, что, как видно из рисунков 3, 4, вид восстанавливаемой функции определяет распределение $\epsilon_i, i = \overline{1,n}$.

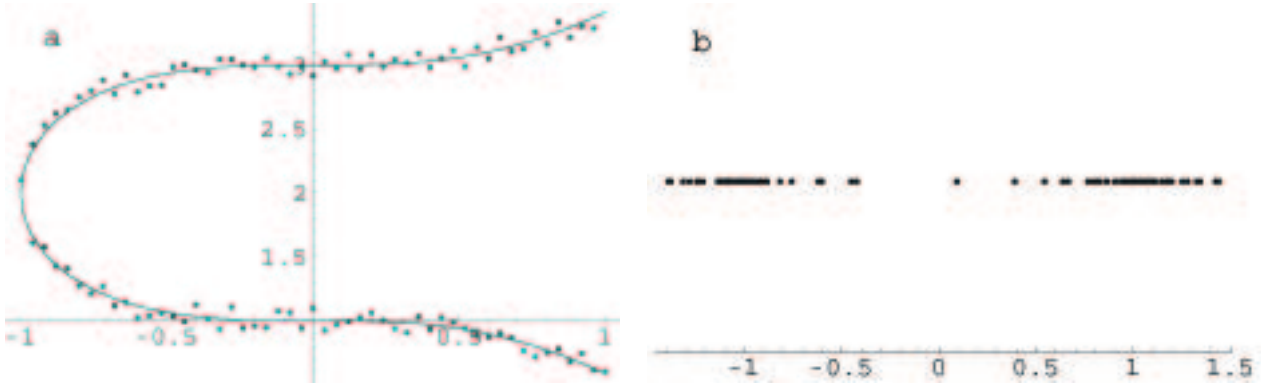

Рис. 3. a – неизвестная функция, подлежащая восстановлению и ОВ, b – распределение $\epsilon_i, i = \overline{1,n}$.

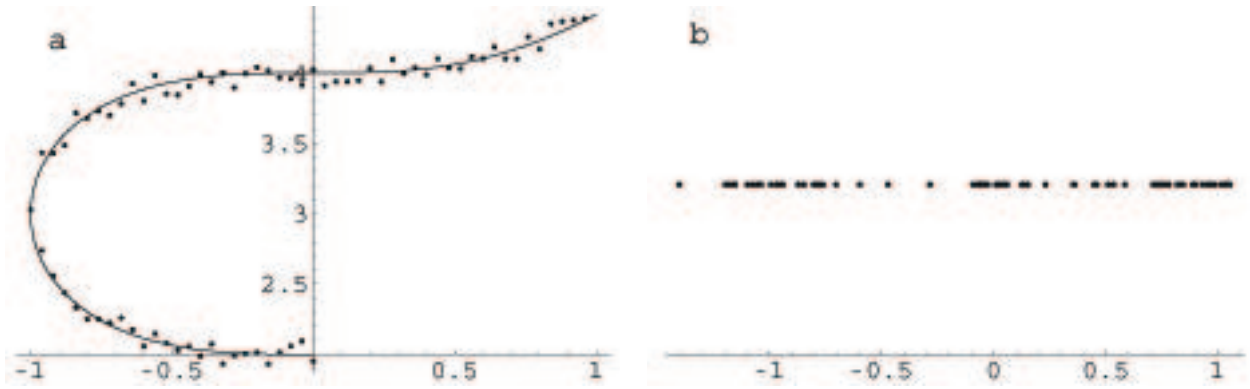

Рис. 4. a – неизвестная функция, подлежащая восстановлению, и ОВ; b – распределение $\epsilon_i, i = \overline{1,n}$.

Распределение $\epsilon_i, i = \overline{1,n}$ на рисунке 3б говорит о четном количестве значений функции, причем это число постоянно на всей области определения. Это наиболее простой случай. И тогда второй этап в общих чертах состоит в отдельном восстановлении ветвей функции по элементам соответствующим положительным и отрицательным $\epsilon_i, i = \overline{1,n}$ соответственно. Если разброс невязок достаточно велик, то это может свидетельствовать о 4-,6- и т.д значности функции и тогда следует повторить первый этап для элементов соответствующих $\epsilon_i > 0$ и $\epsilon_i < 0$ отдельно.

Распределение невязок, соответствующих следующей функции (Рис.4), свидетельствует о том, что функция принимает несколько значений лишь на некотором участке области определения. Такое распределение так же возможно, если функция имеет нечетное число значений на всей области, или разное число значений. Восстановление функции в этом случае представляет гораздо более трудоемкий процесс. Здесь существенную помощь может оказать и анализ функции распределения $y_i, i = \overline{1,n}$.

Приведенные примеры свидетельствуют о трудности задачи и неоднозначности ее решения. Тем не менее о возможности такового говорить можно. Еще одно доказательство информативности предлагаемого анализа представлено на рис.5. Хорошо видно, что картина распределения невязок в случае однозначной зависимости резко отличается от пре-

дыдущих и является характерной для аналогичных случаев (конечно, если в выборке не содержатся выбросы [2]).

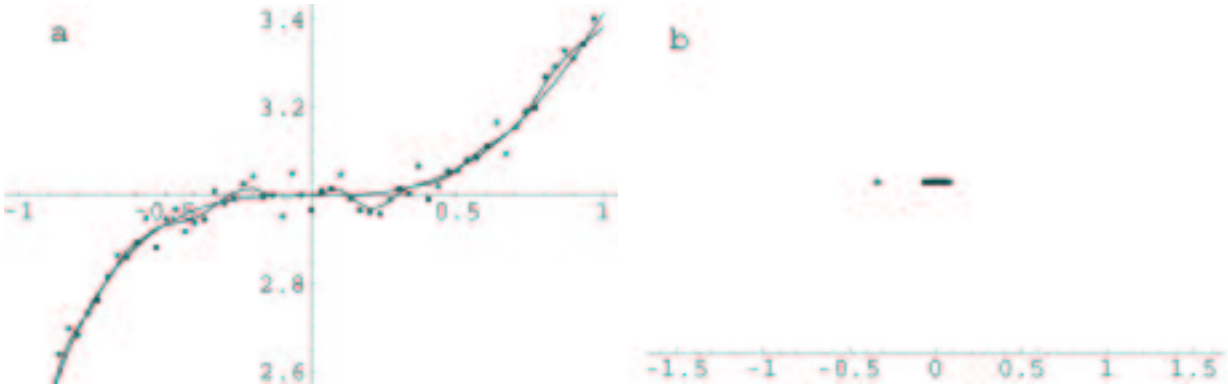

Рис. 5. a – неизвестная функция, подлежащая восстановлению, ее оценка (1) и ОВ; b – распределение $\epsilon_i, i = \overline{1,n}$.

5 Заключение

Представленные результаты свидетельствуют, что несмотря на сложность задачи подходы к ее решению существуют. При определенных ограничениях на выборку задача решена полностью. Для более общего случая, наименее информативного, и следовательно более трудного, проведенный предварительный анализ дает основания полагать, что решение может быть получено.

Список литературы

[1] *Кирик Е.С.* О непараметрическом восстановлении многозначных зависимостей по экспериментальным данным. //Вестник НИИ СУВПТ, В.5, Красноярск: НИИ СУВПТ, 2000. С.25-33.

[2] *Кирик Е.С.* Моделирование и оптимизация робастных оценок функций по наблюдениям. //Вычислительные технологии, Vol.6, P.2, Spesial Issue, 2001. С.351-355.

[3] *Катковник В.Я* Непараметрическая идентификация и сглаживание данных. Москва: Наука, 1985, 336с.

[4] *Медведев А.В* Непараметрические системы адаптации. Новосибирск: Наука, 1983, 174с.

[5] *Parzen E.* On estimation of probability density function and mode. //Ann. Math. Stat. 1962. V. 33. P. 1065–1076.

[6] *Rozenblatt M.* Remarks on some nonparametric estimates of density function. //Ann.Math. Stat., 1956, Vol.27, P.832-837.