УДК 519.68

А.А. Корнеева, Н.А. Сергеева, Е.А. Чжан

О НЕПАРАМЕТРИЧЕСКОМ АНАЛИЗЕ ДАННЫХ В ЗАДАЧЕ ИДЕНТИФИКАЦИИ

Исследуется задача восстановления матрицы наблюдений при оценивании функции регрессии по измерениям со случайными ошибками. Заполнение пропусков осуществляется с помощью непараметрической оценки кривой регрессии. Приводятся результаты численных исследований, иллюстрирующих эффективность работы предложенной методики. Рассматривается моделирование нового класса процессов стохастических объектов со статистической зависимостью компонент вектора входа.

Ключевые слова: *идентификация*, *непараметрические модели*, *«трубчатые» процессы.*

Проблема моделирования, идентификации, безусловно, надолго останется одной из центральных проблем кибернетики. При формулировке задач идентификации и управления особую роль играет уровень априорной информации. Он зависит как от априорных знаний о процессе, имеющихся средствах контроля, так и от самой технологии измерения переменных. Более того, отличие в средствах контроля неизбежно будет приводить к различным постановкам задач идентификации и моделирования даже для процессов одного и того же типа.

Приведем достаточно общую схему исследуемого процесса:

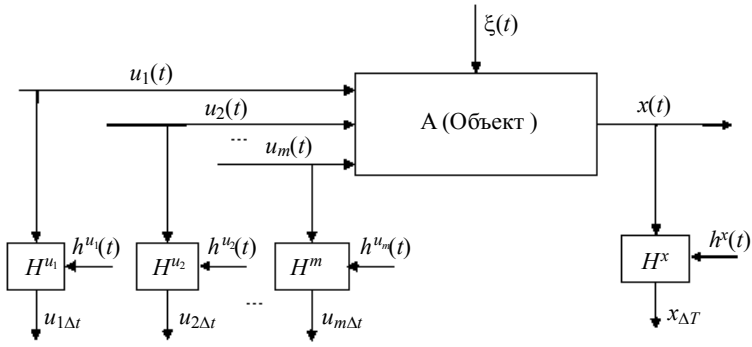

Рис. 1. Общая схема исследуемого процесса и контроля переменных

На рис. 1 приняты обозначения: A – неизвестный оператор объекта, $x(t) \in \Omega(x) \subset R^1$ – выходная переменная процесса,

$$u(t) = (u_1(t), u_2(t), ..., u_m(t)) \in \Omega(u) \subset R^m$$

– входное воздействие, $\xi(t)$ – векторное случайное воздействие, t – непрерывное время, H^u, H^x – каналы связи, соответствующие различным переменным, включающие в себя средства контроля, $h^u(t)$, $h^x(t)$ – случайные помехи измерений

соответствующих переменных процесса с нулевыми математическими ожиданиями и ограниченной дисперсией. Контроль $u(t)$ осуществляется через интервал времени Δt, контроль $x(t)$ – через ΔT, причем $\Delta t << \Delta T$. Пример матрицы измерений подобного рода объектов представлен в табл. 1 («–» – пропуск матрицы наблюдений).

Таблица 1

Матрица наблюдений исследуемого процесса

Входная переменная u				
u_1	u_2	...	u_m	x
u_{11}	u_{21}	...	u_{m1}	x_1
u_{12}	u_{22}	...	u_{m2}	–
u_{13}	u_{23}	...	u_{m3}	–
u_{14}	u_{24}	...	u_{m4}	x_4
...	...	...	...	...
u_{1s}	u_{2s}	...	u_{ms}	x_s

Отличие дискретности измерения переменных, характеризующих состояние исследуемого процесса, обусловлено средствами контроля. В частности, измерения некоторых переменных может осуществляться электрическими средствами и быть достаточно малой величиной. Измерения же других переменных может быть проведено только в результате химического анализа, который требует значительно больше времени.

1. Идентификация в «узком» и «широком» смыслах

При моделировании разнообразных дискретно-непрерывных процессов в настоящее время доминирует теория идентификации в «узком» смысле [1]. Ее содержание состоит в том, что на первом этапе, на основании имеющейся априорной информации, определяется параметрический класс операторов A^{α}, например:

$$\tilde{x}_{\alpha}(t) = A^{\alpha}(u(t), \alpha), \tag{1}$$

где A^{α} – параметрическая структура модели, а α – вектор параметров. На втором этапе осуществляется оценка параметров α на основе имеющейся выборки $\left\{x_i, u_i, i = \overline{1,s}\right\}$, s – объем выборки. Оценка параметров может осуществляться с помощью многочисленных рекуррентных процедур, в частности методом стохастических аппроксимаций либо методом наименьших квадратов. Успех решения задачи идентификации в этом случае существенно зависит от того, насколько «удачно» определен оператор (1). В настоящее время теория параметрической идентификации является наиболее развитой [1].

Идентификация в «широком» смысле предполагает отсутствие этапа выбора параметрического класса оператора [2]. Часто оказывается значительно проще определить класс операторов на основе сведений качественного характера, например линейности процесса или типа нелинейности, однозначности либо неоднозначности и др. В этом случае задача идентификации состоит в оценивании этого оператора на основе выборки $\left\{x_i, u_i, i = \overline{1,s}\right\}$ в форме

$$\tilde{x}_s(t) = A_s(u(t), \vec{x}_s, \vec{u}_s), \tag{2}$$

где $\vec{x}_s = (x_1, x_2, ..., x_s), \vec{u}_s = (u_1, u_2, ..., u_s)$ – временные векторы. Оценка оператора A_s может быть осуществлена средствами непараметрической статистики [3, 4]. Примечательным здесь является то, что при этом исключается этап выбора параметрической структуры. Тем самым можно утверждать, что идентификация в этом случае, а это вариант идентификации в «широком» смысле, является более адекватной для некоторых реальных задач.

2. Непараметрические оценки функции регрессии по наблюдениям

Пусть даны наблюдения $\left\{x_i, u_i, i = \overline{1,s}\right\}$ случайных величин x, u, распределенных с неизвестными плотностями вероятности $p(x,u), p(u) > 0 \; \forall u \in \Omega(u)$. Для восстановления $\tilde{x} = M\{x \mid u\}$ используются непараметрические оценки [2,4]:

$$x_s(u) = \sum_{i=1}^{s} x_i \prod_{j=1}^{m} \Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right) \Bigg/ \sum_{i=1}^{s} \prod_{j=1}^{m} \Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right), \tag{3}$$

где $\Phi(c_s^{-1}(u^j - u_i^j)), i = \overline{1,s}, \; j = \overline{1,m}, \;$ – ядерная колоколообразная функция и коэффициент размытости ядра c_s удовлетворяют следующим условиям сходимости [2, 4]:

$$c_s > 0; \qquad \Phi(c_s^{-1}(u^j - u_i^j)) \geq 0;$$

$$\lim_{s\to\infty} c_s = 0; \qquad \int_{\Omega(u)} \Phi(c_s^{-1}(u^j - u_i^j))du^j < \infty;$$

$$\lim_{s\to\infty} s c_s^m = \infty; \qquad \lim_{s\to\infty} c_s^{-1}\Phi(c_s^{-1}(u^j - u_i^j)) = \delta(u^j - u_i^j).$$

В данном случае в качестве колоколообразной функции $\Phi(c_s^{-1}(u^j - u_i^j))$ было использовано треугольное ядро:

$$\Phi(c_s^{-1}(u^j - u_i^j)) = \begin{cases} 1 - \left|c_s^{-1}(u^j - u_i^j)\right|, \text{если } \left|c_s^{-1}(u^j - u_i^j)\right| \leq 1, \\ 0, \text{если } \left|c_s^{-1}(u^j - u_i^j)\right| > 1. \end{cases}$$

Параметр размытости c_s определяется путем решения задачи минимизации квадратичного показателя соответствия выхода объекта и выхода модели, основанного на «методе скользящего экзамена», когда в модели (3) исключается *i*-я переменная, предъявляемая для экзамена:

$$R(c_s) = \sum_{k=1}^{s} \left(x_k - x_s(u_k, c_s)\right)^2 = \min_{c_s}, k \neq i.$$

В случае, если каждой компоненте вектора u соответствует компонента вектора c_s, то во многих практических задачах c_s можно принять скалярной величиной, если предварительно привести компоненты вектора u по выборке наблюдений к одному и тому же интервалу, например использовать операции центрирования и нормирования.

3. Методика заполнения матрицы наблюдений

На практике часто возникают случаи, как было замечено ранее, когда дискретность измерения «входных-выходных» переменных исследуемого процесса может не совпадать. В результате матрица наблюдений состоит из не полностью заполненных строк (табл. 1).

Для решения задач идентификации предпочтительно иметь выборки большего объема. Отсюда возникает проблема восстановления пропусков в незаполненных строках матрицы наблюдений. Конечно, при решении задачи идентификации можно использовать только заполненные строки матрицы наблюдений. Но в этом случае объем выборки становится существенно меньше. В настоящей работе предлагается дать оценки x в незаполненных строках матрицы наблюдений при известных значениях входных переменных u. Таким образом, используется выборка, состоящая из результатов заполненных строк матрицы наблюдений (табл. 1). В этом случае получим заполненную матрицу, представленную в табл. 2, и оценку $x(u)$ класса (1) или (2) будем осуществлять уже на основании заполненной матрицы наблюдений.

Таблица 2

Матрица наблюдений с заполненными строками

Входная переменная u				x
u_1	u_2	...	u_m	
u_{11}	u_{21}	...	u_{m1}	x_1
u_{12}	u_{22}	...	u_{m2}	x_{s2}
u_{13}	u_{23}	...	u_{m3}	x_{s3}
u_{14}	u_{24}	...	u_{m4}	x_4
...	...	...	...	...
u_{1s}	u_{2s}	...	u_{ms}	x_s

В качестве оценки $\tilde{x} = M\{x \mid u\}$ можно использовать как параметрические оценки функции регрессии [1], так и непараметрические [2–4]. Такой прием, как это будет показано ниже, оказывается вполне оправданным, так как задача идентификации в последнем случае (табл. 2) решается более точно, чем в случае, когда мы исключаем строки с пропусками из матрицы наблюдений (табл. 1), тем самым уменьшая объем выборки.

4. Этапы восстановления пропусков матрицы наблюдений

Методику восстановления пропусков матрицы наблюдений можно разделить на три этапа [5].

На первом этапе восстанавливается функция регрессии x_s по наблюдениям u, полностью представленным в исходной матрице измерений, то есть по полностью заполненным строкам в результате эксперимента (табл. 1). Подбирается оптимальное значение коэффициента размытости c_s.

На втором этапе происходит заполнение пропусков матрицы с использованием оценки x_s и оптимального значения коэффициента размытости ядра c_s, полученных на предыдущем этапе. Там, где наблюдения x пропущены, в оценку $x_s(u_1, u_2, ..., u_m)$ подставляем значения измеренных $u = (u_1, u_2, ..., u_m)$ и вычисляем

соответствующую оценку x_s, которой восполняем недостающее наблюдение x (например, недостающее значение x_2 в представленной выше матрице наблюдений заполняется значением x_{s2}). После этого этапа матрица наблюдений принимает вид, представленный в табл. 2.

Заключительный этап восстановления зависимости x от $u=(u_1,u_2,...,u_m)$ состоит в построении модели. В данном случае была использована непараметрическая оценка функции регрессии (3) по всей имеющейся (заполненной) матрице наблюдений (табл. 2). При этом коэффициент размытости c_s подбирается по всей имеющейся выборке еще раз.

5. Вычислительный эксперимент

На исследуемый объект действует векторное входное воздействие $u=(u_1,u_2,u_3)\in[0;3]$, x – выходная переменная, скаляр. Имеются выборки статистически независимых наблюдений $(u_i,x_i,i=\overline{1,s})$, где s – объем выборки. Пусть матрица наблюдений объекта имеет вид, представленный в табл. 1, т.е. $\Delta T=3\Delta t$. Зададим структуру объекта следующим уравнением:

$$x=0,5u_1^2+\sin u_2+2\sqrt{u_3}\ . \qquad (4)$$

Вычислительный эксперимент состоял из двух этапов. На первом этапе оценка (3) строилась по исходной матрице наблюдений с пропусками (табл. 1). Затем данная матрица заполнялась при помощи предложенной выше методики. На втором этапе оценка строилась уже по восстановленной матрице наблюдений без пропусков (табл. 2).

На нижеследующем рисунке приведены результаты вычислительного эксперимента для объекта, описываемого (4). На выход объекта ξ накладывалась помеха по формуле

$$\xi=0.05\cdot\varsigma\cdot x,$$

где ς – случайная величина, нормально распределенная в интервале $[-1;1]$.

Объем выборки s варьировался от 100 до 1200. На рис. 2 представлено два графика, соответствующих оцениванию по исходной матрице наблюдений с пропусками (табл. 1) и по восстановленной (табл. 2). Графики показывают зависимость относительной ошибки моделирования σ от объема выборки s. Относительная ошибка моделирования σ находится по формуле

$$\sigma=\sqrt{\sum_{i=1}^{s}(x_i-x_s(u_i))^2\Big/\sum_{i=1}^{s}(\hat{m}-x_i)^2},$$

где $\hat{m}$ – оценка математического ожидания.

Так как мы имеем дело со случайными величинами, то необходимо провести усреднение полученных результатов. В данном случае производилось усреднение по результатам десяти экспериментов.

Как видно из приведенного рис. 2, ошибка моделирования, полученная при оценке по восстановленной матрице наблюдений, меньше, чем ошибка – по матрице с пропусками. В среднем качество оценивания повышается на 5–10 %. Данный численный эксперимент демонстрирует, что задача идентификации по вос-

становленной (заполненной) матрице наблюдений решается более точно, чем по матрице с пропусками.

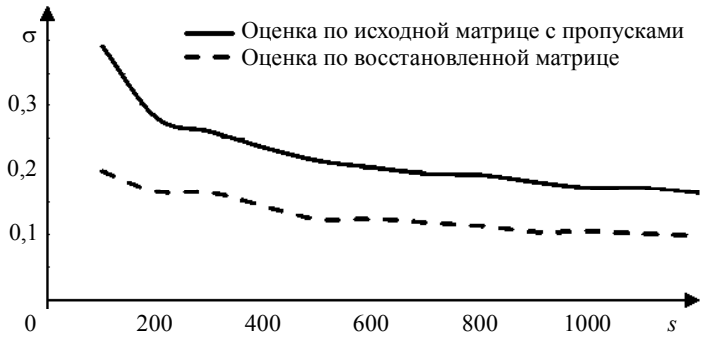

Рис. 2. Зависимость ошибки моделирования σ от объема выборки *s*

6. H-модели

На практике достаточно часто встречаются процессы, имеющие стохастическую зависимость компонент вектора входных переменных. Будем говорить, что объекты, обладающие подобной особенностью, имеют «трубчатую» структуру [6].

Рассмотрим в качестве примера процесса с «трубчатой» структурой объект, представленный на рис. 3.

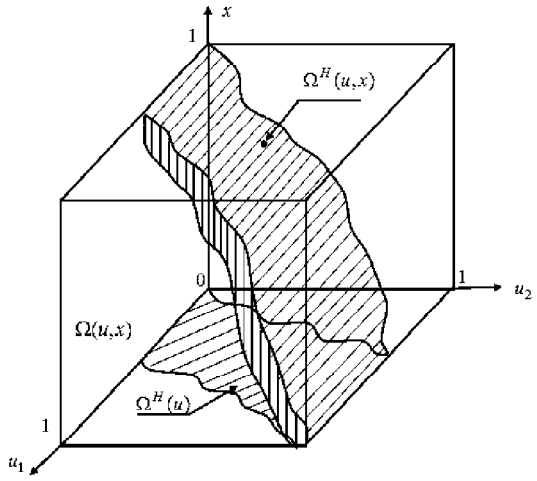

Рис. 3. Объект с «трубчатой структурой»

Как видно из рисунка, область протекания процесса $\Omega(u,x) \in R^3$ представляет собой, без нарушения общности, единичный гиперкуб, где $u \in R^2$, $x \in R^1$. Однако если исследуемый процесс имеет «трубчатую» структуру, то область его протекания ограничивается не всем объемом гиперкуба $\Omega(u,x)$, а его подобластью $\Omega^H(u,x) \in \Omega(u,x)$, которая нам никогда не известна. Поскольку подобласть $\Omega^H(u,x)$ никогда не известна, то и вид самой «трубчатой» структуры нам неиз-

вестен. При этом заметим, что объем гиперкуба, как это видно из вышеприведенного рисунка, может значительно превышать объем «трубки».

Рассмотрим моделирование процессов, имеющих подобную структуру. Обычно в задаче идентификации безынерционных объектов предполагается наличие некоторой параметризованной модели, представляющей собою поверхность в пространстве «входных-выходных» переменных:

$$\hat{x}_s(u) = \hat{f}(u, \alpha_s), \tag{5}$$

где α_s – вектор параметров. В том случае, когда компоненты вектора входных переменных статистически зависимы, т.е. мы имеем дело с «трубчатой» структурой объекта, необходимо ввести индикатор $I(u)$. Модель вышеприведенного типа при этом должна быть скорректирована следующим образом:

$$\hat{x}_s(u) = \hat{f}(u, \alpha_s) I_s(u), \tag{6}$$

где в качестве оценки индикатора можно принять следующее приближение:

$$I_s(u) = \operatorname{sgn}(sc_s) \sum_{i=1}^{s} \prod_{j=1}^{m} \Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right), \tag{7}$$

Параметр размытости ядра c_s определяется так же, как и в (3), а колоколообразная функция $\Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right)$ имеет вид треугольного ядра.

Логика построения такого индикатора состоит в том, что при произвольно заданном значении текущей переменной $u = u'$ индикатор $I_s(u)$ примет значение единицы, если u' принадлежит «трубчатой» структуре, определяемой имеющейся выборкой $\left\{x_i, u_i, i = \overline{1,s}\right\}$, если же u' приняло значение за пределами «трубки», то индикатор равен нулю. Заметим, что если процесс описывается поверхностью в пространстве $\Omega(u, x)$, то модели (5) и (6) совпадают. Если же процесс имеет трубчатую структуру в этом пространстве, то необходимо использовать модель (6).

7. Численное исследование Н-модели

Рассмотрим результаты численного эксперимента. Пусть исследуемый объект описывается системой уравнений:

$$\begin{cases} x(t) = 0.5u_1(t) + 0.5u_2(t) + \xi; \\ u_2(t) = u_1(t) + \psi, \end{cases} \tag{8}$$

где ξ и ψ – случайные числа, распределенные по равномерному закону на интервале $[-0,05; 0,05]$, $u_1, u_2 \in [0;3]$. В данном случае уравнение объекта задано с целью получения выборок «входных-выходных» переменных для решения задачи идентификации. При построении модели на основе полученных выборок, структура зависимости выходной переменной x от входных переменных u принята с точностью до параметров. При оценивании параметров используется метод наименьших квадратов (МНК).

Итак, получена выборка статистически независимых наблюдений $\{x_i, u_i, i = \overline{1,s}\}$, где x – измеряемая выходная переменная, $u = (u_1, u_2)$ – векторное входное воздействие, s – объем выборки.

Построим параметрическую модель исследуемого объекта. Результаты моделирования показаны на рис. 4 ($s = 100$), где черными точками обозначен исследуемый объект, серыми – полученная параметрическая модель. Как видно из рис. 4, исследуемый объект является «трубкой». Модель (5) при этом представляет собой плоскость. Как известно, прямую, в данном случае «трубку», можно аппроксимировать бесконечным числом плоскостей. Пусть имеется шесть выборок статистически независимых измерений $\{x_i, u_{1i}, u_{2i}, i = \overline{1,s}\}$ объемом $s = 100$. Для каждого случая построим 6 моделей с помощью МНК. Результаты моделирования представлены на рис. 5.

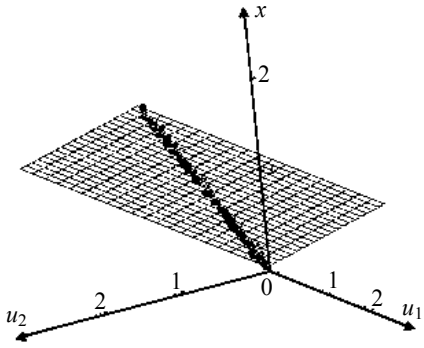

Рис. 4. Объект с «трубчатой структурой» и его параметрическая модель

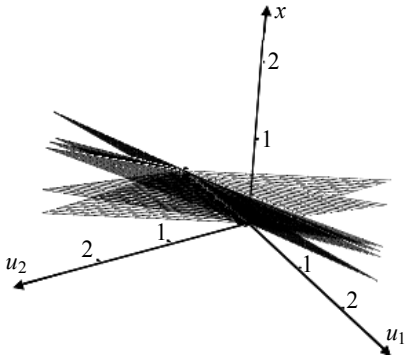

Рис. 5. Множество параметрических моделей объекта с «трубчатой» структурой

В результате моделирования было получено 6 моделей с различными оценками параметров, так как в каждом случае выборки были различны. Можно сделать вывод о том, что полученные модели не являются адекватными. Кроме того, для построения параметрической модели необходим большой объем выборки.

Теперь для рассматриваемого объекта (8) будем использовать модель (6), содержащую индикатор. Вид индикатора описан формулой (7). Результаты численного моделирования объекта (8) для объема выборки $s = 1000$ представлены на рис. 6.

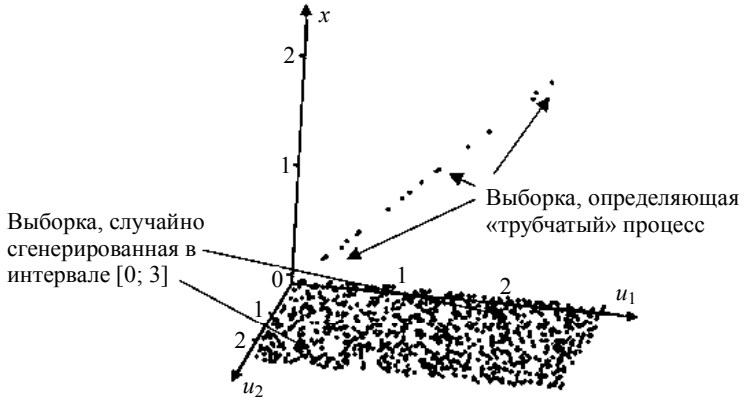

Рис. 6. Выборка измерений «входных-выходных» переменных

Как мы видим, индикаторная функция (7) при построении модели учитывала только те точки, которые принадлежат области протекания «трубчатого» процесса, т. е. $I_s(u)=1$. В остальных точках выборки значение выхода модели не восстанавливалось, т. е. $I_s(u)=0$. В частности, из общего объема выборки $s=1000$ объему трубки принадлежат лишь 20.

Изменим систему уравнений, которой описывается исследуемый объект, на следующую:

$$\begin{cases} x(t)=u_1^2(t)+u_2^2(t)+\xi; \\ u_2(t)=u_1(t)+\psi. \end{cases} \tag{9}$$

Результаты аналогичных экспериментов представлены на нижеследующих рисунках:

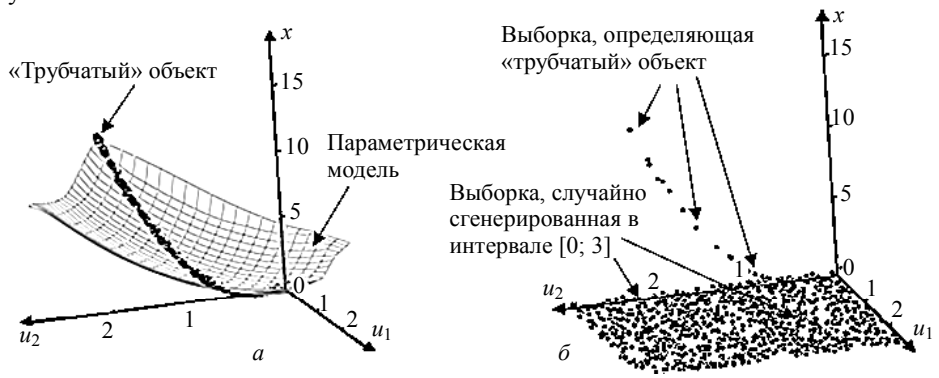

Рис. 7. Моделирование «трубчатого» процесса

На рис. 7, *а* показан объект, описываемый системой (9) и имеющий «трубчатую» структуру (черные точки на графике). Плоскость (рис. 7, *а*) показывает восстановленную с помощью МНК параметрическую модель объекта. Также построена (рис. 7, *б*) параметрическая модель с использованием индикаторной функции (6). Как и в предыдущем эксперименте показано, что индикатор учитывает только те точки, которые принадлежат объему «трубки» ($I_s(u)=1$), в остальных точках выборки индикатор становится равным нулю и в дальнейшем эти точки не участвуют в построении модели.

Заключение

Рассмотрена задача восстановления матрицы наблюдений с пропусками для повышения эффективности решения задачи идентификации стохастических безынерционных объектов. Предложена методика восстановления пропусков матрицы наблюдений и приведены соответствующие алгоритмы, основанные на использовании непараметрической оценки функции регрессии. Результаты численных экспериментов иллюстрируют, что задача идентификации по заполненной матрице решается более точно, чем по незаполненной.

Рассмотрен класс объектов, имеющих «трубчатую» структуру. При моделировании объектов такого рода необходимо учитывать ряд его особенностей и использовать Н-модели. Предложенные модели процессов, имеющих «трубчатую» структуру, относятся к категории новых по отношению к своим предшественни-

кам, рассматриваемым в теории идентификации. Здесь важно иметь в виду, что область протекания такого процесса никогда не известна и при моделировании должна подлежать определению. Н-модели отличаются от общепринятых моделей безынерционных систем наличием индикаторной функции, которая, по существу, определяет область протекания «трубчатого» процесса.

ЛИТЕРАТУРА

1. *Эйкхоф П.* Основы идентификации систем управления. М.: Мир, 1975. 683 с.
2. *Медведев А.В*. Непараметрические системы адаптации. Новосибирск: Наука, 1983.174 с.
3. *Кошкин Г.М. Пивен И.Г.* Непараметрическая идентификация стохастических объектов. Хабаровск: Российская академия наук. Дальневосточное отделение, 2009. 336 с.
4. *Надарая Э.А.* Непараметрические оценки плотности вероятности и кривой регрессии. Тбилиси: Изд-во Тбил. ун-та, 1983. 194 с.
5. *Корнеева А.А.* О непараметрическом восстановлении матрицы наблюдений с пропусками в задаче идентификации с шумами / Молодой ученый. 2012. № 3(38). С. 51–60.
6. *Медведев А.В.* Анализ данных в задаче идентификации // Компьютерный анализ данных моделирования. Минск: БГУ, 1995. Т. 2. С. 201–206.
7. *Чжан Е.А.* О непараметрической идентификации стохастических систем с запаздыванием / Е.А. Чжан, Н.А. Сергеева // Кибернетика и высокие технологии XXI века: труды XIII Международной научно-технической конференции. Воронеж, 2012. Т. 1. С. 63–74.

Корнеева Анна Анатольевна
Сергеева Наталья Александровна
Чжан Екатерина Анатольевна
Сибирский федеральный университет,
E-mail: anna.korneeva.90@mail.ru; sergena@list.ru;
ekach@list.ru

Поступила в редакцию 30 апреля 2012 г.

Korneeva Anna A., *Sergeeva Natalya A.*, *Chzhan Ekaterina A.* (Siberian Federal University).
Nonparametric data analysis in identification problem.

Keywords: identification, nonparametric models, «tubular» processes.

By investigation of many processes it is necessary to solve the modeling and identification problem. The qualitatively constructed models help to simplify the control of the object as well as to predict its future behavior. This paper focuses on the identification of a new class of processes which have statistical relationship between the components of the input variables. Further, these objects will be called «tubular».

As it is known, the quality of the identification problem solution is determined by the quality of source data, so the stage of data preprocessing is an important part of the modeling process. In this paper some peculiarities of the samples, as blanks are described. There are proposed two nonparametric estimation algorithms using the regression function.

The use of parametric identification methods does not give satisfactory results in modeling «tubular» processes.

A modification of the parametric identification algorithm using the indicator function is suggested. The indicator shows whether the points belong to the true course of the process or not. The experimental results demonstrate the feasibility of the proposed algorithms.