УДК 535.12

ИДЕНТИФИКАЦИЯ МОДЕЛЕЙ С ОТБОРОМ ДАННЫХ ПО КРИТЕРИЮ ОБОБЩЕННОЙ ФРАКТАЛЬНОЙ РАЗМЕРНОСТИ

©2004 А.В. Никоноров

Самарский государственный аэрокосмический университет

Проведено исследование такой характеристики процедуры адаптивного оценивания параметров мделей, как плотность множества получаемых оценок. Обосновано применение этой характеристики в качестве критерия взаимной близости при согласованном оценивании по малому числу наблюдений. Для определения плотности рассчитывается показатель обобщенной фрактальной размерности. Эксперементально показана эффективность применения показателя фрактальной размерности в качестве критерия взаимной близости.

Введение

В работах [1, 2] сформулирован общий подход к решению задач идентификации по малому числу наблюдений, основанный на так называемом принципе согласованности оценок. Основная идея подхода заключается в использовании критериев, не требующих задания априорных вероятностных моделей, которые, как известно, при малом числе наблюдений ненадежны. Общая схема построения согласованных оценок сводится к формированию из исходной системы множества подсистем меньшей размерности, среди которых отбирается та, для которой значение заданного критерия согласованности (взаимной близости) минимально.

Применение различных критериев близости в рамках данного подхода, порождает целый класс различных адаптивных алгоритмов оценивания. В работе [2] были рассмотрены критерии близости оценок в пространстве параметров моделей. В работах [3, 4] продемонстрировано успешное использование согласованности в пространстве отклика модели. Однако можно показать, что при определенных предположениях о природе оценок истинное значение оцениваемого вектора параметров принадлежит наиболее плотной области множества оценок. Поэтому наиболее естественным представляется использование в качестве критерия близости оценку пространственной плотности распределения оценок в пространстве параметров модели.

Кроме функции взаимной близости определяющим для алгоритма является генерации подсистем исходной системы, на которых выполняется оценивание. Базовым методом генерации подсистем является метод, предложенный в работе [2]. Для построения множества оценок используются так называемые подсистемы верхнего и нижнего уровней. Для подсистемы верхнего строится множество подсистем нижнего уровня по принципу скользящего окна, на этих подсистемах выполняется оценивание. Рассчитывается взаимная близость оценок на подсистемах нижнего уровня принадлежащих некоторой подсистеме верхнего уровня, из всех подсистем верхнего выбирается система, оценки на которой наиболее близки. Окончательная оценка параметров модели выполняется по этой, наиболее согласованной подсистеме.

Нахождение множества оценок с наилучшим значением взаимной близости оптимизационная задача. В базовом алгоритме применяется переборный метод решения этой задачи. Как альтернатива в работе [3] предлагается методика оптимизации критерия взаимной близости с использованием генетических алгоритмов поиска. В развитие этого подхода в данной работе предлагается модифицированный алгоритм. Основное отличие предлагаемой схемы от базовой в способе выбора множества оценок, по которым выполняется рассчет согласованности. Если в базовом алгоритме оценки рассчитываются по множеству наборов строк некоторой

подсистем, то в предлагаемом методе ищется наиболее согласованное подмножество на множестве всех возможных оценок.

Множество оценок распределяется в пространстве параметров моделей неравномерно и не регулярно, для определения характеристик этого множества можно использовать методы фрактальной геометрии. В частности для оценки плотности этого множества в данной работе рассчитывается показатель так называемой дробной размерности. Эффективность предлагаемой методики проверяется на задаче линейной аппроксимации спектров отражения красочных смесей.

Постановка задачи и предположения

Предполагается, что на основе физических законов и/или эмпирических соотношений задана структура модели:

$$r_k^* = F_k\left(\mathbf{x}_k^*, c\right), \quad (1)$$

где r_k^*, $\mathbf{r}_k^*$ - известные или доступные для непосредственного измерения скаляр и вектор, удовлетворяющие точной параметрической модели заданного вида (1) (в предположении, что такая модель существует), а **c** - $M\times 1$-вектор неизвестных параметров, подлежащий определению.

Пусть производится серия из N измерений r_k^*, $\mathbf{r}_k^*$ в условиях, при которых вектор искомых параметров остается неизменным. Тогда можно записать систему уравнений вида:

$$y_k = F_k(x_k, c) + \xi_k \quad k = \overline{1,N}, \quad (2)$$

где x_k и y_k - непосредственно наблюдаемые в эксперименте вектор и скаляр соответственно, а ξ_k k=1,N – ошибки. Предполагается, что ошибки измерений пренебрежимо малы по сравнению с ошибками аппроксимации. Поэтому можно считать, что ошибки входят в уравнения аддитивно. Таким образом, задача идентификации модели в данном случае заключается в определении $M\times 1$- вектора **c** по малому числу наблюдений r_k, $\mathbf{r}_k$, $k = \overline{1,N}$ в присутствии вектора ошибок $\xi = [\xi_1, \xi_2, ..., \xi_N]^T$. Нам потребуются следующие предположения относительно компонентов соотношения (2):

1. Скаляры y_k и вектора x_k $k = \overline{1,N}$ фиксированы, т. е. известны в результате измерений на одной реализации.

2. Число наблюдений мало, так что имеет место неопределенность статистических характеристик вектора **o**.

3. Норма вектора ошибок $\boldsymbol{\xi} = [\xi_1, \xi_2, ..., \xi_N]^T$ ограничена, а его направление в шаре

$$\boldsymbol{\Xi} = \left\{\boldsymbol{\xi} : \left(\boldsymbol{\xi}^T\boldsymbol{\xi}\right)^{1/2} = \|\boldsymbol{\xi}\|_2 \le R_\xi = Const\right\}$$

случайно.

Под случайностью направления вектора ошибок понимается отсутствие систематического смещения к какому либо значению.

Процедура построения оценок

В зависимости от вида зависимости F можно предложить различные методики нахождения оценки $\hat{c}$ вектора **c**, в частности в работе [5] оценивание проводится для нелинейных моделей цветовоспроизведения. Однако предлагаемые в этой статье методики оценки плотности оценок с использованием дробной размерности инвариантны к способу получения оценок, и их применение вполне может быть рассмотрено для модели типа линейной регрессии:

$$Xc = y + \xi. \quad (3)$$

Оценка параметров (3) рассчитывается по МНК как:

$$\hat{c} = [X^T X]^{-1} X^T y. \quad (4)$$

В [6] показано, что для улучшения оценки (4) оправдано применение следующего класса модифицированных МНК оценок:

$$\hat{c} = [X^T G_k^2 X]^{-1} X^T G_k^2 (y + \Delta y_k). \quad (5)$$

где G_k^2 так называемая весовая матрица, а Δy_k -– корректирующий вектор.

Базовый метод согласованного соответствует нулевому значению корректирующего вектора $\Delta y = 0$ и весовой матрице диагонального вида

$$G_k = diag(d) = diag(d_k^1, ..., d_k^N), \quad (6)$$

где $d_i = 1$ если k-тая исходной системы входит в подсистему по которой проводится оценивание, иначе $d_i = 0$, причем ранг $R = rank(G_k^2)$ равен количеству строк в k-той подсистеме, в дальнейшем эту величину будем называть рангом подсистемы. Для G_k соответствующей некоторой подсистеме верхнего уровня строятся все возможные матрицы $G_{k,q}$, $rank(G_{k,q}) = P$. Строятся оценки:

$$\hat{c}_{k,q} = [X^T G_{k,q}^2 X]^{-1} X^T G_{k,q}^2 . \qquad (7)$$

Для подсистемы верхнего уровня вычисляется значение критерия взаимной близости оценок на подсистемах нижнего уровня, вообще говоря, как функция следующего вида:

$$W(\hat{c}_{k,1},...,\hat{c}_{k,Q}), \qquad Q = C_R^P . \qquad (8)$$

По подсистеме с наилучшим значением W строится окончательная точечная оценка вектора параметров $\hat{c}$.

Такая оценка может быть выбрана как наиболее близкая к среднему значению $\bar{c}$ множества Θ_k состоящего из всех оценок для k-той подсистемы верхнего уровня:

$$\hat{\mathbf{c}} = \arg\min_{c_q \in \Theta_k} (\bar{c} - c) ; \qquad (9)$$

в этом случае $\hat{c} \in \Theta_k$ и соответственно можно указать матрицу G_k из (6) и подмножество строк системы (3) по которым найдена оценка. Однако можно выбирать $\hat{c}$ равной среднему значению множества Θ_k:

$$\bar{c}_k = \frac{1}{Q_k} \sum_{\hat{c}_k^i \in \Theta_K} \hat{c}_k^i , \qquad (10)$$

$$\hat{\mathbf{c}}_k = \bar{c}_k . \qquad (11)$$

В этом случае $\hat{c}$ не обязательно принадлежит Θ_k. Если $\forall k: \ \hat{c} \notin \Theta_k$, то найденная оценка не удовлетворяет (7), однако удовлетворяет (5), в силу непрерывности этого соотношения. Т.е. этой оценке соответствует не только некоторая весовая матрица, но и не нулевой корректирующий вектор.

Можно предложить следующее обоснование метода. Рассмотрим геометрическую трактовку процесса формирования множества оценок и выбора среди них наилучшей. Каждой подсистеме верхнего уровня соответствует некоторое множество оценок Θ_k, обозначим диаметр этого множества за d_k. Тогда из требования (3) об отсутствии систематического смещения вектора ошибки следует, что, истинное значение вектора параметров находится внутри области покрываемой множеством Θ_k, по крайней мере, для подсистем с рангом $R_k > R_{пор}$. Таким образом, для большинства подсистем начиная с некоторого ранга истинное значение вектора параметров должно покрываться пересечением всех Θ_k. Если рассматривать все множества Θ_k соответствующие подсистемам одного ранга, то количество оценок в каждом из этих множеств будет одинаковое. Тогда наиболее согласованной системе будет соответствовать наиболее плотное множество оценок. В свою очередь оценки такого наиболее плотного множества оценок будут наименее удалены от истинного значения параметров. Таким образом, можно определить критерий взаимной близости для подсистемы как плотность соответствующего множества оценок.

Наиболее плотное множество оценок не обязательно принадлежит некоторой подсистеме верхнего уровня, это может быть произвольное подмножество оценок, причем возможно полученных для подсистем различного ранга. С учетом этого факта можно предложить алгоритм получения согласованных оценок не на подсистеме верхнего уровня, а непосредственно, а непосредственно на наиболее плотном подмножестве из пространства оценок.

Такой модифицированный метод получения согласованных оценок возможно реализовать при помощи генетического алгоритма поиска, определяемого следующим образом. Хромосома (особь) представляет собой вектор d_k^i и определяет весовую матрицу (6). Популяция соответствует множеству Θ_k. Селекция родителей проводиться по случайной схеме, чтобы увеличить пространство

поиска алгоритма. Оператор мутации стандартный, кроссовер одноточечный. Приспособленность особи считается как отклонение оценки соответствующей особи от среднего значения по популяции на данной итерации. Приспособленность рассчитывается в два этапа: сначала для всех особей рассчитываются оценки $\hat{c}_k^i(d_k^i)$ (5) и находится среднее по популяции по формуле (10), а потом рассчитывается приспособленность каждой особи как:

$$f(d_k^i) = \left\| \hat{c}_k^i(d_k^i) - \bar{c}_k \right\|_2^2. \quad (12)$$

Как было показано в [5] генетический поиск позволяет за счет некоторого увеличения ошибки оценивания значительно сократить вычислительные затраты на построение согласованных оценок.

Плотность множества оценок

В предыдущем разделе неявно использовалось среднее арифметическое приближение плотности множества Θ. Однако множество Θ не регулярно распределено в пространстве и поэтому такое приближение является очень грубым. Рассмотрим задачу более точного вычисления плотности точечного множества Θ. Плотность данного множества можно определить как отношение мощности множества к объему, покрываемому множеством в пространстве $\rho_\Theta = Q / V_\Theta$. Характеристикой того, как распределяется множество в пространстве, является размерность множества. Точечному множеству свойственна так называемая мультифрактальная размерность.

Обобщенная фрактальная размерность (размерность Реньи) согласно [7] определяется следующим образом. Разобьем область L M-мерного евклидова пространства, содержащую точечное множество Θ, на кубические ячейки со стороной ε*e* и объемом ε^M. Пусть номер *i* занятых ячеек в которых находится хотя бы одна точка множества Θ изменяется в пределах $i = 1,2...N(\varepsilon)$, здесь *N(*ε*e)* суммарное количество занятых ячеек.

Пусть $n_i(\varepsilon)$ представляет собой количество точек в ячейке с номером *i*, тогда предел

$$p_i(\varepsilon) = \lim_{\varepsilon \to 0} \frac{n_i(\varepsilon)}{N} \quad (13)$$

представляет собой вероятность того, что некоторая точка находится в *i*-той ячейке. Введем в рассмотрение обобщенную вероятностную сумму Z(q, ε), с показателем степени $-\infty < q < +\infty$

$$Z(q,\varepsilon) = \sum_{i=1}^{N(\varepsilon)} p_i^q(\varepsilon). \quad (14)$$

Спектр обобщенных фрактальных размерностей D_q характеризующий распределение точек множества определяется как:

$$D_q = \frac{\tau(q)}{q-1}, \quad (15)$$

где $\tau(q)$ имеет вид:

$$\tau(q) = \lim_{\varepsilon \to 0} \frac{\ln Z(q,\varepsilon)}{\ln \varepsilon}. \quad (16)$$

Если $D_q = const$, то множество представляет собой регулярный фрактал, если D_q переменная величина, то множество представляет собой мультифрактал. Величина $D_0 \geq D_q \quad \forall q$ равна хаусдорфовой размерности множества. Для оценки плотности точечного множества можно воспользоваться данным значением размерности в силу его экстремальных свойств.

Если предположить гиперкубическое строение множества И, то его объем можно оценить как

$$V_\Theta = r^{D_0}. \quad (17)$$

Если предположить, что множество оценок структурой близкой к радиально симметричной, то можно воспользоваться определением массовой размерности, что позволит существенно уменьшить вычислительную сложность.

Для радиально симметричного множества оценок объем рассчитывается как:

$$V_\Theta = \gamma(D_0) r^{D_0}, \quad (18)$$

где

$$\gamma(D_0) = \Gamma(1/2)^{D_0} / \Gamma(1 + D_0/2), \quad (19)$$

Γ(x) – Гамма-функция:

$$\Gamma(X) = \int_0^\infty e^{-t} t^{x-1} dt, \; x > 0 . \qquad (20)$$

Окончательно плотность множества Θ оценок на подсистеме определяется как:

$$\rho_\Theta = \frac{N_\Theta}{\gamma(D_0) r^{D_0}} . \qquad (21)$$

Вычисленная таким образом плотность множества оценок может быть использована в качестве критерия взаимной близости. Для множеств с одинаковым диаметром и количеством элементов в качестве критерия близости можно использовать непосредственно значение размерности Хаусдорфа. В частности такое использование эффективно для одинаковых подсистем в базовом алгоритме построения согласованных оценок или для вычисления значения согласованности популяции при генетическом построении оценок. При этом наиболее согласованному множеству оценок соответствует наименьшая размерность.

Численное исследование

При практическом вычислении размерности (15) возникает ряд проблем. Две основные заключаются в необходимости выбора шага по ее и в конечном размере множества, размерность которого оценивается. Вычисленное значение размерности точечного множества смещается относительно истинного значения тем сильнее, чем меньше количество точек в множестве. Значение не значительно при десятках тысяч точек в множестве. При расчете размерности подсистемы небольшого ранга приходится иметь дело с менее чем с тысячей точек, а для генетического алгоритма использовалась популяция в 400 точек, при таких количествах точек смешения вычисленной размерности становиться не приемлемым.

В работе [8] предлагается способ уменьшения этого смещения для клеточного алгоритма вычисления размерности. Он заключается в сравнении межточечных расстояний на исследуемом множестве и на сгенерированном фрактальном множестве. С использованием предлагаемого алгоритма удалось вычислить размерности для множеств из нескольких множеств точек. Однако для множеств, в которых количество точек различается на порядок и более значения получаемых размерностей не сравнимы. Поэтому, фрактальная размерность, вычисленная таким образом, может использоваться в качестве функции взаимной близости только для множеств, с примерно одинаковым количеством точек.

В качестве примера рассмотрим оценивание параметров линейной зависимости (3) спектра отражения красочной смеси от компонент входящих в смесь. Пример модельный, спектр смеси был получен расчетным путем, а потом на него было наложены шумы. Используется 17 отсчетов спектра, графики спектров смеси до (гладкая кривая) и после зашумления приведены на рисунке 1.

Спектры компонент, входящих в смесь, приведены на рисунке 2. Таким образом, исходная система имеет матрицу 17x5.

Остаточная сумма квадратов для оцен-

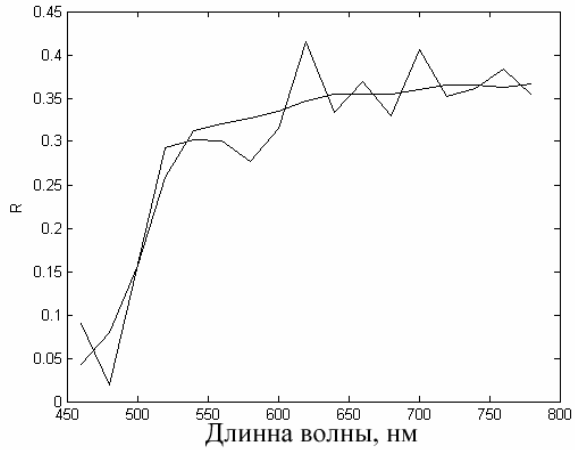

Рис. 1. Спектр смеси

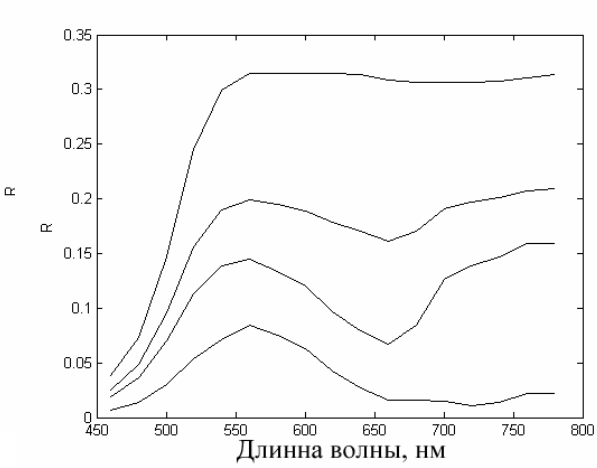

Рис. 2. Спектры компонент смеси

ки (4) по исходным данным составляет 2.785. В дальнейшем ошибкой будем называть СКО между рассчитанным и не зашумленным спектром. Для МНК оценки по исходной системе ошибка составляет 2.7115. Для базового алгоритма использовались подсистемы нижнего уровня ранга 7 и подсистемы верхнего уровня ранга 12. Исходной системе соответствует $C_{17}^{7} = 19448$ оценок на подсистемах нижнего уровня, это множество имеет размерность Хаусдорфа, рассчитанную клеточным методом, равную $D_0 = 1.831$. Каждой подсистеме верхнего уровня соответствует 792 оценки, для наиболее согласованной подсистемы оценка рассчитывается по (11). Ошибка для такой оценки составила 0.112, размерность $D_0 = 0.976$, что значительно меньше, чем для исходной системы. Для произвольно взятой подсистемы верхнего уровня ошибка составила 1.753 и $D_0 = 1.128$.

При использовании генетического алгоритма наилучшее значение ошибки составило 0.964. Размерность D_0 вычислялась для множеств из 200 точек, что соответствует использовавшемуся размеру популяции. Значения ошибки и размерности приведены в таблице 1.

Из таблицы видно, что меньшему значению размерности соответствует меньшее значение ошибки, что действительно позволяет использовать оценку размерности в качестве критерия взаимной близости. Однако для множеств с равным количеством точек это не так.

Поиск таких фрактальных характеристик множества оценок, которые были бы не зависели от количества точек в множестве, требует дальнейшего исследования.

Заключение

В данной работе проведено исследование такой характеристики адаптивного оценивания параметров мделей, как плотность множества получаемых оценок. Обосновано применение этой характеристики в качестве критерия взаимной близости при согласованном оценивании по малому числу наблюдений.

Для определения плотности точечного множества использовалась обобщенная фрактальная размерность. В ходе мсследования была установлена непосредственная связь значения размерности и погрешностью оценивания. Экспериментально показана возможность применения размерности Хаусдорфа в качестве критерия взаимной близости. Однако процедура вычислениия обобщенной фрактальной размерности с большими ограничениями применима к точечным множествам оценок. Разработка процедуры определения размерности свободной от этих ограничений, равно как исследование других фрактальных характеристик множества оценок чвляется перспективным направлением для дальнейшей работы.

Благодарности

Автор выражает глубочайшую признательность профессору В.А.Фурсову и доценту С.Б.Попову за помощь, без которой не была бы написана эта статья. Работа выполнена при поддержке Министерства образования РФ, Администрации Самарской области и Американского фонда гражданских исследований и развития (CRDF) в рамках российско-американской программы “Фундаментальные исследования и высшее образование” (BRHE) и РФФИ (гранты № 03-01-00109, 04-07-90149, 04-07-96500).

Таблица 1. Зависимость ошибки от размерности Хаусдорфа

Ошибка	Оценка размерности D_0
0.6756	0.8234
0.9611	0.8360
1.2950	0.8661
1.6054	0.9107
2.0753	0.9392

СПИСОК ЛИТЕРАТУРЫ

1. *Fursov V. A.* Theoretical and calculational aspects of constructing recognition algorithms using a small number of observations. Proc. of the All-Russian Conf. “Mathematical Methods of Pattern Recognition” (MMPR-10). Moscow, 19-23 November. 2001.
2. *Fursov V.A.* Conformity Principle in Problems of Identification, International Conference Melborne, Australia and St.Petersburg, Russia. Springer-Verlag/ 2003.
3. *Никоноров А.В., Попов С.Б., Фурсов В.А.* Вычислительные аспекты реализации идентификации моделей цветовоспрозведения. // Известия СНЦ РАН. Т 4. №1.
4. *Nikonorov A., Popov S., Fursov V.* Identifying Color Reproduction Models, Pattern Recognition and Image Analysis. 2003. Vol. 13. №2.
5. *Никоноров А.В, Попов С.Б., Фурсов В.А.* Идентификация нелинейных моделей цветовоспроизведения. Доклады 11 Всероссийской конференции “Математические методы распознавания образов” ММРО-11, Москва. 2003.
6. *Фурсов В.А.* Идентификация моделей систем формирования изображений по малому числу наблюдений. Самара: ИПО СГАУ. 1998.
7. *Федер Е.* Фракталы, М.: Мир, 1991.
8. *Roberts. A.~J.* Estimate generalised fractal dimensions of a set of points. Technical report, http://www.sci.usq.edu.au/staff/aroberts/fdim.sh, 1994.

MODEL IDENTIFICATION WITH DATA SELECTION BY CRITERIA OF GENERALIZED FRACTAL DIMENSION

©2004 A.V. Nikonorov

Samara State Aerospace University

In the given work the investigation of the density of set of received adaptive parameters estimations is carried out. Application of this characteristic is proved as criterion of mutual closenes during conforming estimayion using small number of observation. For definition of density the value of generalized fractal dimension is calculated. Efficiency of application of a fractal dimension as critera of mutual closenes is experimentally shown.