ИНТЕРВАЛЬНАЯ МАТЕМАТИКА И РАСПРОСТРАНЕНИЕ ОГРАНИЧЕНИЙ
МКВМ-2004 РАБОЧИЕ СОВЕЩАНИЯ С. 210–218

О построении зависимостей по эмпирическим данным с интервальной ошибкой

С.И. Жилин*

Аннотация. В работе рассматривается задача построения линейно параметризованных зависимостей типа "вход–выход" по эмпирическим данным с интервальной ошибкой измерения выходной переменной. Предложен метод выявления наблюдений–выбросов. По результатам имитационных экспериментов проведено сравнение точечных оценок максимального правдоподобия и наименьших квадратов с точечными оценками, получаемыми путем выбора срединной точки интервальных оценок, и получен вывод о конкурентоспособности подобного нестатистического подхода.

1. Введение

Рассматриваемая в работе постановка задачи построения и анализа зависимостей по эмпирическим данным с интервальной ошибкой восходит к идее Л. В. Канторовича [5] и исследуется в [1, 3, 8–11], совпадая в них с точностью до терминологии.

Суть задачи состоит в построении по эмпирическим данным линейно параметризованной зависимости

$$y = \sum_{i=1}^{n} \beta_i x_i, \tag{1}$$

где $x \in \mathbb{R}^n$ – вектор входных переменных, $\beta \in \mathbb{R}^n$ – вектор параметров, подлежащих оцениванию, y – скалярная выходная переменная.

Зависимость конструируется по эмпирической информации, главное место в которой занимает таблица экспериментальных данных, полученная в N наблюдениях:

$$T = \{(y_j, x_{1j}, \ldots, x_{nj}) \mid j = 1, \ldots, N\}.$$

При этом предполагается, что погрешностью измерения входных переменных x_i можно пренебречь, а выходная переменная y в j-м наблюдении измеряется с предельной абсолютной погрешностью ε_j.

Ограниченность ошибки измерения выходной переменной позволяет для каждого из наблюдений записать двустороннее неравенство

$$y_j - \varepsilon_j \le \sum_{i=1}^{n} \beta_i x_{ij} \le y_j + \varepsilon_j, \qquad j = 1, \ldots, N. \tag{2}$$

Неравенства (2) в совокупности определяют множество B допустимых значений параметров $\beta = (\beta_1, \ldots, \beta_n)$, именуемое множеством неопределенности.

*Алтайский государственный университет, Барнаул.

В качестве интервальных оценок параметров β_i часто используются проекции $[\underline{\beta}_i, \overline{\beta}_i]$ наименьшего из охватывающих множество B брусов, а в качестве точечных оценок – середины этих проекций

$$\hat{\beta}_i = (\underline{\beta}_i + \overline{\beta}_i)/2. \tag{3}$$

Границы проекций могут быть найдены решением задач линейного программирования:

$$\underline{\beta}_i = \min_{\beta \in B} \beta_i, \qquad \overline{\beta}_i = \max_{\beta \in B} \beta_i, \qquad i = 1, \ldots, n. \tag{4}$$

В отношении множества B может также ставиться задача интервального и точечного прогноза значения выходной переменной y в точке x. Границы интервальных оценок $[\underline{y}_i(x), \overline{y}_i(x)]$ могут быть найдены решением задач линейного программирования:

$$\underline{y}_i(x) = \min_{\beta \in B} \sum_{i=1}^{n} \beta_i x, \qquad \overline{y}_i(x) = \max_{\beta \in B} \sum_{i=1}^{n} \beta_i x, \qquad i = 1, \ldots, n. \tag{5}$$

Точечная оценка прогноза $\hat{y}(x)$ строится как середина интервальной оценки:

$$\hat{y}(x) = \frac{1}{2}(\underline{y}(x) + \overline{y}(x)). \tag{6}$$

Однако любая из указанных задач оценивания имеет смысл лишь в случае ограниченности и непустоты множества неопределенности B. Неограниченность множества B очевидным образом распознается в результате ранговых исследований матрицы наблюдений и содержательно может интепретироваться как недостаток эмпирической информации для построения зависимости. Пустота множества B говорит о противоречивости собранной информации, одной из возможных причин которой может служить наличие выбросов среди наблюдений. В работе предлагается метод выявления выбросов, позволяющий добиться непротиворечивости исходных данных и, соответственно, непустоты множества неопределенности.

Еще одним вопросом, рассматриваемым в работе, является выяснение соотношения оценок, получаемых с помощью изложенного выше подхода, с традиционно используемыми статистическими оценками метода максимального правдоподобия (ММП) и метода наименьших квадратов (МНК) на основе имитационного эксперимента. О необходимости проведения подобного эксперимента как единственного средства сравнения методов оценивания, опирающихся на различные системы гипотез, говорилось в заметке [2]. Для определенности и краткости оценки, получаемые посредством выражений (3)–(6), далее будем называть нестатистическими.

2. Выявление выбросов

Одним из наиболее значимых с практической точки зрения свойством описанного во введении подхода является его потенциальная способность выявлять ситуации, в которых собранные для построения зависимости совокупности данных противоречивы. Индикатором наличия противоречий в данных

является пустота множества неопределенности. Основными источниками противоречий являются либо нарушение гипотезы о структуре конструируемой зависимости, либо наличие выбросов в данных. Выбор способа разрешения противоречий в конечном итоге определяется исследователем по результатам всестороннего анализа. Однако результаты такого анализа во многом зависят и от того, какой информацией располагает для этого исследователь. Настоящий раздел посвящен описанию одного из возможных подходов к получению информации, позволяющей разрешать противоречия, возникающие в случае наблюдений с выбросами.

Выброс представляет собой определенную особенность, нетипичное наблюдение по отношению к остальным данным. Это означает, что выбросы должны подвергаться особенно тщательному рассмотрению с целью выяснения причин их возникновения. Иногда выброс дает такую информацию, которую не могут дать другие наблюдения, и является результатом измерений при необычной комбинации условий. В этом случае требуется дальнейшее углубленное исследование. Однако чаще выбросы вызваны грубыми промахами при регистрации значений наблюдаемых величин. В этом случае производится исключение или целенаправленное ослабление веса наблюдения–выброса в общей информационной совокупности.

Выброс, обусловленный грубым промахом при регистрации результатов измерений, можно трактовать как наблюдение, предельная погрешность которого занижена по отношению к реальной ошибке, имевшей место при измерении. Чтобы такое наблюдение стало “правильным”, необходимо найти нижнюю границу реальной ошибки, при которой наблюдение не будет вступать в противоречие с остальными. Сравнение значения этой нижней границы и приписанной наблюдению ошибки, позволяет строить некоторые суждения относительно степени несоответствия наблюдения–выброса общей картине.

Нижние границы предельных ошибок наблюдений, при которых множество неопределенности становится непустым, можно отыскивать, решая задачу

$$\min_{\beta,w} \sum_{j=1}^{N} w_j, \tag{7}$$

$$y_j - w_j\varepsilon_j \le \sum_{j=1}^{N} \beta_j x_j \le y_j + w_j\varepsilon_j, \qquad w_j \ge 1, \qquad j = 1,\dots,N, \tag{8}$$

где w_j – масштабирующие коэффициенты, указывающие, во сколько раз необходимо растянуть исходную предельную ошибку ε_j для того, чтобы j-е наблюдение не вступало в противоречие с общей совокупностью данных. Полученные в результате решения задачи (7), (8) значения масштабирующих коэффициентов, превосходящие единицу, соответствуют наблюдениям–выбросам. Если у исследователя есть основания считать, что надежность некоторых наблюдений одинакова, то система ограничений (8) может быть пополнена равенствами вида $w_{j_1} = w_{j_2} = \cdots = w_{j_K}$. В случае, когда в надежности каких-либо наблюдений исследователь уверен полностью, при решении задачи (7), (8) соответствующие им величины w_j можно положить равными единице.

Количество наблюдений, для которых масштабирующие коэффициенты w_j, полученные в результате решения задачи (7), (8), превосходят единицу, позво-

ляет судить о доле выбросов в совокупности данных. Большая доля выбросов может говорить либо о неверно выбранной структуре зависимости, либо о том, что предельные ошибки измерения занижены во многих наблюдениях (например, в результате неверной оценки точности измерительного прибора).

Любопытным представляется тот факт, что предложенный подход к разрешению противоречий в совокупности экспериментальных данных укладывается в рамки теории коррекции несобственных задач линейного программирования [4] и может рассматриваться как один из возможных способов праметризации несобственной задачи линейного программирования с целью поиска ее аппроксимации собственной задачей и путей коррекции с минимальными затратами.

3. Экспериментальное сравнение статистических и нестатистических оценок

Главным отличием в системах гипотез, лежащих в основании статистического и нестатистического подходов к построению и анализу зависимостей, является гипотеза о структуре ошибки.

В статистическом подходе ошибка полагается случайной величиной, описываемой некоторым законом распределения, выбираемым исследователем. На практике часто, но как показывает ряд исследований [6, 7], далеко не всегда обоснованно, закон распределения ошибки выбирается нормальным. В этом случае наиболее качественные (состоятельные и эффективные) оценки обеспечивает МНК, являющийся частной формой ММП.

Одним же из главных принципов нестатистической обработки наблюдений, определяющим все последующие алгоритмы и получаемые выводы, является равновозможность всех элементов интервала ошибки, а следовательно, и множества неопределенности B.

Описательные способности рассматриваемых методов при построении зависимости по эмпирическим данным предлагается выяснить по результатам вычислительного эксперимента, состоящего в многократном решении каждым из сравниваемых методов задачи точечного прогноза по модельным данным и выяснении стандартных отклонений прогнозных оценок от истинных модельных значений. Модельные данные предлагается генерировать путем добавления ошибки с заданным распределением к точным значениям выходной переменной при фиксированных значениях входных переменных для некоторой заранее известной зависимости. Выбор именно точечных оценок в качестве сравниваемых показателей объясняется их особой ролью для исследователей–практиков и возможностью одинаковой интерпретации, что не вполне реализуемо в отношении статистических нестатистических интервальных оценок.

Что касается выбора распределения ошибки при генерировании модельных данных, то интерес представляют ситуации “наилучшие” для каждого из сравниваемых методов, а также некоторые близкие к ним варианты. Наилучшими условиями для статистических методов являются ситуации, когда ошибка распределения подчиняется некоторому унимодальному распределению, в частности, для МНК таковым является нормальное распределение ошибки. Базовому для МЦН предположению о равновозможности всех элементов мно-

жества неопределенности в вероятностных терминах наиболее адекватно соответствует равномерное распределение. Таким образом, сравнительный эксперимент предлагается провести для унимодального и равномерного распределений ошибки, а также некоторых промежуточных распределений.

3.1. Нестатистические оценки и оценки максимума правдоподобия

Сравнение нестатистического метода построения оценок с ММП проведено для семейства распределений с плотностью

$$p_\alpha(x) = \begin{cases} \dfrac{1-2\varepsilon\alpha}{\varepsilon^2}x + \dfrac{1-\varepsilon\alpha}{\varepsilon}, & -\varepsilon \le x < 0; \\ \dfrac{2\varepsilon\alpha-1}{\varepsilon^2}x + \dfrac{1-\varepsilon\alpha}{\varepsilon}, & 0 \le x \le \varepsilon; \end{cases}$$

где ε – абсолютное значение предельной ошибки, а $\alpha \in \left[0, \dfrac{1}{2\varepsilon}\right]$ – параметр, определяющий степень близости распределения к треугольному. При $\varepsilon = 1$ графики функции $p_\alpha(x)$ для граничных и двух промежуточных значений параметра α приведены на рис. 1.

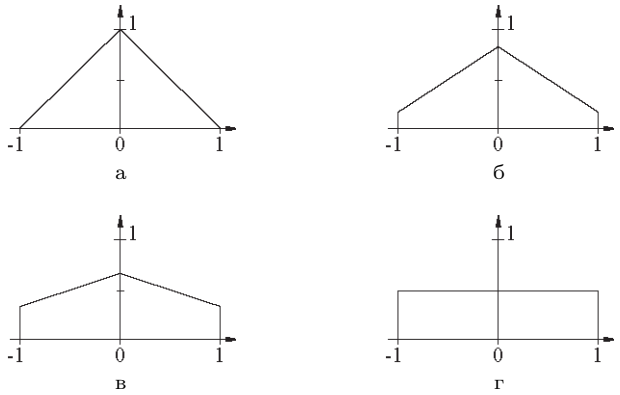

Рис. 1. Графики функции плотности $p_\alpha(x)$ при $\varepsilon = 1$ и (а) $\alpha = 0$, (б) $\alpha = 1/6$, (в) $\alpha = 1/3$, (г) $\alpha = 1/2$

При построении оценок ММП при распределении ошибки, близком к равномерному, возникают сложности в выборе оценки, обусловленные неединственностью максимума функции правдоподобия. Выход из этой ситуации видится в регуляризации задачи поиска максимума функции правдоподобия $L(\beta)$ путем добавления слагаемого $\delta|L(\beta)|(\beta-\beta^0)^2$, где $\delta < 0$ — постоянный весовой коэффициент (в эксперименте $\delta = -0.1$), а β^0 – известное модельное значение параметров.

В качестве модельной зависимости была выбрана функция $y = x + 1$, т. е. истинное модельное значение вектора параметров $\beta^0 = (1, 1)$. Совокупность точных значений модельной зависимости была получена путем вычисления значений выходной переменной в узлах регулярной сетки с шагом 1 на интервале $[1; 10]$. Для каждого из фиксированных значений $\alpha_m = \dfrac{m}{20}\dfrac{1}{2\varepsilon}$ $(m = 0, \ldots, 20)$

5000 раз генерировалась таблица наблюдений путем добавления к точным значениям выходной переменной случайной ошибки из интервала $[-0.5; 0.5]$ с плотностью распределения $p_\alpha(x)$ и каждым из сравниваемых методов строились точечные оценки параметров зависимости, на основе которых вычислялись прогнозные значения зависимости в точке $x = 5.5$. По результатам повторений эксперимента при фиксированном m вычислялось стандартное отклонение каждого из типов оценок от истинного значения модельной зависимости в этой точке. Зависимость стандартного отклонения нестатистических оценок и оценок ММП от m приведена на рис. 2.

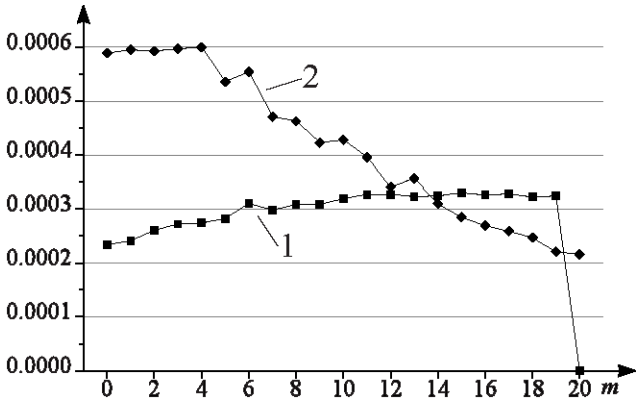

Рис. 2. Среднеквадратичные отклонения прогнозных значений от истинных для ММП (1) и нестатистического метода (2)

Сравнительный анализ стандартных отклонений нестатистического и ММП-прогноза показывает, что при распределениях погрешности, близких к "треугольным", характер поведения ошибки прогноза соответствует известным соотношениям и закономерностям, свойственным использованным методам оценивания. Действительно, нестатистическая процедура не учитывает дополнительную информацию, связанную с характером распределения, и, соответственно, имеет большую ошибку прогноза. Но по мере приближения распределения погрешности к равномерному ошибка нестатистического прогноза снижается. Это объясняется тем, что такая ситуация становится все более соответствующей базовому для метода построения нестатистических оценок предположению о равновозможности всех элементов множества неопределенности. В то же время при приближении распределения ошибки к равномерному стандартное отклонение ММП-прогноза возрастает и с некоторого момента начинает превосходить стандартное отклонение нестатистического прогноза. Резкое падение стандартного отклонения ММП-прогноза в точке $m = 20$ объясняется возросшим относительным весом регуляризующего слагаемого.

3.2. Нестатистические оценки и оценки наименьших квадратов

Схема и параметры экспериментов по сравнению нестатистических оценок с оценками наименьших квадратов в основном повторяет схему экспериментов, описанную в предыдущем разделе. Изменения касаются лишь количества

повторений, вида распределения ошибки и кратности наблюдений.

Количество повторений эксперимента при фиксированных параметрах распределения составляло 1000.

Семейство распределений ошибки $N_k(a, \sigma^2)$ в этом случае представляло собою нормальные распределения, усеченные на уровне k, т. е. ошибка принимает значения из интервала $[a - k\sigma, a + k\sigma]$, где a – математическое ожидание, σ – среднеквадратическое отклонение. В проведенной серии экспериментов математическое ожидание было нулевым, среднеквадратическое отклонение единичным, а k выбиралось из интервала $[0, 2; 3]$ с шагом $0, 2$. По мере роста k получаемые распределения принимали вид от почти равномерного до почти нормального.

По указанной схеме эксперимент проводился в полном объеме для каждого из фиксированных значений кратности наблюдений $Q = 1, 3, 9$.

Результаты эксперимента в графическом виде приведены на рис. 3.

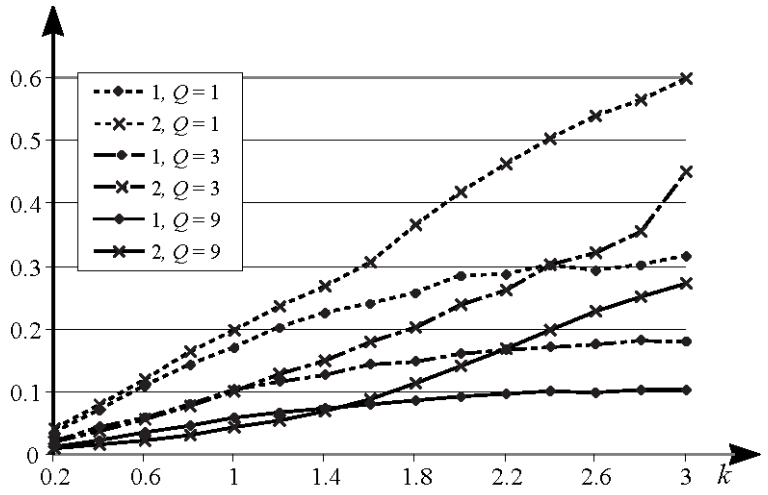

Рис. 3. Зависимость среднеквадратичных отклонений прогнозных значений от истинных для МНК (1) и нестатистического метода (2) от уровня усечения нормального распределения k и кратности наблюдений Q

Качественный анализ взаимосвязей среднеквадратичных отклонений нестатистического и МНК-прогнозов с уровнем усечения нормального распределения ошибки и кратностью измерений позволяет сделать следующие наблюдения:

1. По мере уменьшения уровня усечения нормального распределения ошибки измерений среднеквадратичные отклонения и нестатистического, и МНК-прогнозов также убывают. При этом для МНК-прогноза скорость убывания можно качественно охарактеризовать как логарифмическую или линейно-логарифмическую, в то время как для нестатистического прогноза – как полиномиальную.
2. При больших значениях k оценки МНК-прогноза более устойчивы, чем нестатистические оценки. Однако с уменьшением k их преимущество

утрачивается. Кроме того, с увеличением кратности измерений более устойчивыми становится нестатистические оценки. Объяснение этому факту, так же как и в случае сравнения с ММП, состоит в уменьшении степени соответствия распределения ошибки измерения гипотезе о нормальности, в рамках которой МНК дает наилучшие результаты. В то же время, приближение распределения ошибки к равномерному все более соответствует одному из базовых предположений нестатистического метода о равновозможности всех элементов интервала ошибки и множества неопределенности.

3. С увеличением кратности измерений устойчивость нестатистических оценок растет несколько быстрее. Тенденция усиливается по мере уменьшения уровня усечения нормального распределения ошибки измерений, то есть по мере приближения распределения к равномерному. Этот факт свидетельствует о способности нестатистического метода неявно накапливать информацию о распределении ошибки, незадействуемую явным образом в отличии от статистических процедур оценивания.

Таким образом, результаты сравнительного анализа точечных нестатистических оценок с оценками, получаемыми методами максимального правдоподобия и наименьших квадратов позволяют сделать вывод о конкурентоспособности нестатистического подхода к построению и анализу зависимостей в случае ограниченности ошибки наблюдений несмотря на то, что статистические методы, вообще говоря, задействуют больше информации, требуя указания явным образом структуры предпочтений на интервале ошибки в виде закона распределения.

Список литературы

[1] Белов В.М., Суханов В.А., Гузеев В.В., Унгер Ф.Г. Оценивание параметров линейных физико-химических зависимостей прямоугольником метода центра неопределенности // Изв. вузов. Физика. – 1991. – № 8. – С. 35–45.

[2] Бородюк В.П. Комментарий I к статье А.П. Вощинина, А.Ф. Бочкова, Г.Р. Сотирова “Метод анализа данных при интервальной нестатистической ошибке” // Заводская лаборатория. – 1990. – Т. 56, № 7. – С. 81–83.

[3] Вощинин А.П., Бочков А.Ф., Сотиров Г.Р. Метод анализа данных при интервальной нестатистической ошибке // Заводская лаборатория. – 1990. – Т. 56, № 7. – С. 76–81.

[4] Еремин И.И. Противоречивые модели оптимального планирования. – М.: Наука, 1988. – 160 с.

[5] Канторович Л.В. О некоторых новых подходах к вычислительным методам и обработке наблюдений // Сиб. мат. журнал. – 1962. – Т. 3, № 5. – С. 701–709.

[6] Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. – Л.: Энергоатомиздат, 1985. – 248 с.

[7] Орлов А.И. Часто ли распределение результатов наблюдений является нормальным? // Заводская лаборатория. – 1991. – Т. 57, № 7. – С. 64–66.

[8] Оскорбин Н.М., Максимов А.В., Жилин С.И. Построение и анализ эмпирических зависимостей методом центра неопределенности // Известия Алтайского государственного университета. – 1998. – № 1. – С. 35–38.

[9] Спивак С.И. Детальный анализ применения методов линейного программирования при определении параметров кинетической модели // Математические проблемы химии. – Новосибирск: ВЦ СО АН СССР, 1975. – Ч. 2. – С. 35–42.

[10] Milanese M., Belforte G. Estimation theory and uncertainty intervals evaluation in presence of unknown but bounded errors: linear families of models and estimators // IEEE Transactions on Automatic Control. – 1982. – Vol. 27, № 2. – P. 408–414.

[11] Rodionova O.Ye., Pomerantsev A.L. Antioxidants activity prediction using DSC measurements and SIC data processing // II Conference on Experimental Methods in Physics of Heterogeneous Condensed Media. – Barnaul, 2001. – P. 239–246.