УДК 519.68

О МЕТОДИКЕ ИСКЛЮЧЕНИЯ ВЫБРОСОВ ИЗ ИСХОДНОЙ ВЫБОРКИ

Корнеева А.А.
научный руководитель профессор, д.т.н. А.В.Медведев
Сибирский Федеральный Университет

В задаче идентификации статических, стохастических объектов при наблюдении «входны-выходных» переменных могут возникать ситуации, когда измерение той или иной переменной осуществляется с многократной ошибкой типа «промах». Наличие в матрице наблюдений подобных значений естественно затрудняет процесс построения модели. Для борьбы с ошибками типа «промах», как известно, можно использовать методологию робастной статистики. Ниже предлагается методика анализа имеющейся выборки наблюдений, позволяющая исключить подобные аномальные измерения.

Пусть входная переменная объекта представляет собой вектор $u=(u_1,u_2,\ldots,u_m)$, а выход x, без нарушения общности, примем скалярным. В результате измерения «входных-выходных» переменных получаем выборку $(u_i,x_i,i=\overline{1,s})$, содержащую промахи. Непосредственное применение или использование этой выборки в задаче идентификации, восстановления регрессионных характеристик при некоторых значениях u может приводить к существенным ошибкам. В этой связи для исключения аномальных наблюдений предлагается следующая методика.

Предложенная методика включает в себя несколько этапов. На первом – по исходной выборке в режиме скользящего экзамена (не учитывается i-я пара измерений) строится непараметрическая оценка регрессии x_s:

$$x_s(u_i)=\sum_{i=1}^{s}x_i\prod_{j=1}^{m}\Phi\left(\frac{u_j-u_{ij}}{c_s(j)}\right)\Bigg/\sum_{i=1}^{s}\prod_{j=1}^{m}\Phi\left(\frac{u_j-u_{ij}}{c_s(j)}\right), \qquad (1)$$

где $\Phi(\cdot)$ – ядерная колоколообразная функция и коэффициент размытости ядра c_s удовлетворяют некоторым условиям сходимости. Находится оптимальное значение коэффициента c_s в соответствии с критерием $I=\sum_{j\neq i,\,j=1}^{s}\left(x_j-x_s\left(u_j,c_s\right)\right)^2=\min_{c_s}$. На втором этапе вводится число ε_s, которое показывает близость оценки x_s к истинному значению величины x. Данный показатель вычисляется по формуле: $\varepsilon_s=\frac{1}{s}\left(\sum_{i=1}^{s}|x_s(u_i)-x_i|\right)$, $i=\overline{1,s}$, где s- объем выборки. Затем, для каждого x_i проверяется выполнение условия:

$$|x_s(u_i)-x_i|\geq\alpha\varepsilon_s, \qquad (2)$$

где α - коэффициент, определяемый экспериментальным путем. Если условие выполняется, то точка x_i становится кандидатом на исключение из выборки. Формируется выборка, состоящая из точек, удовлетворяющих условию (2) и из них находится измерение с максимальным отклонением $|x_s(u_i)-x_i|$, $i=\overline{1,s}$. Эта точка объявляется выбросом и удаляется из исходной выборки. После чего, по полученной выборке, заново находится оптимальный коэффициент размытости c_s и строится

оценка x_s. Затем выполняется проверка условия (2). Если выборка содержит аномальные, измерения методика повторяется.

Рассмотрим результаты вычислительного эксперимента. Пусть исследуемый объект описывается уравнением вида:

$$x(t) = f(u_1(t), u_2(t), \xi(t)) = 0.5u_1(t) + 0.5u_2(t). \qquad (3)$$

Дана выборка статистически независимых измерений $(x_i, u_{1i}, u_{2i}, i = 1,2,\ldots,s)$, где s - объем выборки равен 300, $u_1, u_2 \in [0;3]$, α экспериментальным путем выбрано равным 4. Помеха, наложенная на выход x_i, составляет ξ=5%. Внесем в выборку выброс: $x[39] = 4$. Построим оценку (1) x_s по выборке, содержащей выброс. Результат моделирования представлен на рисунке 1.

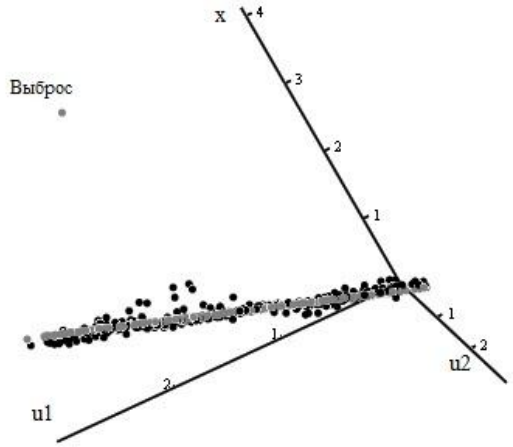

Рисунок 1 – Результаты оценивания по выборке с выбросом

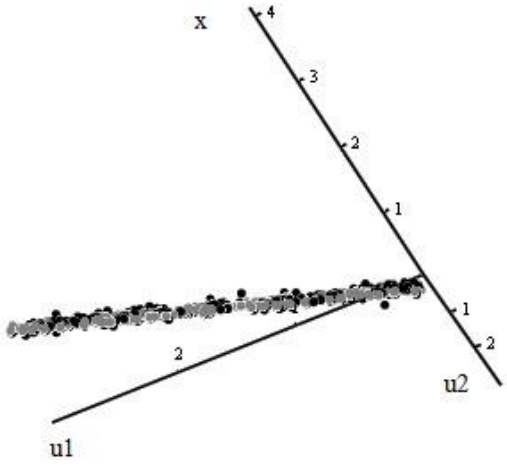

Рисунок 2 – Результаты оценивания по выборке, после исключения выброса

Как видно из рисунка 1, в области выброса оценка x_s достаточно груба. Относительная ошибка моделирования при этом составляет σ_1 = 10%. Используя предложенную выше методику, исключим выброс из исходной выборки. Результаты моделирования представлены на рисунке 2. Относительная ошибка моделирования в этом случае уменьшилась и составила σ_2 = 0.8%. Полученные оценки x_s достаточно точны. При наличии в исходной выборке двух выбросов, расположенных друг от друга на некотором расстоянии (в эксперименте принято $x[20] = 5$ и $x[39] = 4$), ошибка моделирования по исходной выборке с выбросами составила σ_1 = 15%, после исключения выбросов σ_2 =0.9%. При наличии двух выбросов, следующих друг за другом ($x[39] = 5$ и $x[40] = 4$), ошибка составила σ_1 = 24%, после исключения выбросов уменьшилась до σ_2 = 0.9%.

Выше была изложена методика, позволяющая исключать из исходной выборки наблюдения, содержащие многократные ошибки при измерении. Проведенные вычислительные эксперименты показали достаточно высокую эффективность анализа исходной обучающей выборки в соответствии с приведенной методикой. В итоге, как показало численное исследование, наблюдения, содержащие промахи, исключались из обучающей выборки, которая и использовалась для оценки регрессионных характеристик исследуемого процесса. При этом их точность существенно возрастала.