УДК 681.5.01

Н. С. ЛЕСНАЯ, В. Б. РЕПКА, Т. Б. ШАТОВСКАЯ

ИТЕРАЦИОННАЯ ГРЕБНЕВАЯ ПРОЦЕДУРА ОЦЕНКИ ПАРАМЕТРОВ МОДЕЛИ

Одной из сложных проблем при обработке данных регрессионным анализом является наличие мультиколлинеарности между независимыми переменными. Наиболее эффективными в этом случае являются методы смещенного оценивания, и, в частности, гребневые методы, позволяющие получить оценки параметров модели с меньшей среднеквадратической ошибкой по сравнению с обычным методом наименьших квадратов (МНК). При обработке данных с большим факторным пространством наиболее эффективными из гребневых методов, в смысле точности решения, являются итерационные. Однако при наличии вырожденности в информационной матрице, точность вычислений при применении этих методов значительно ухудшается, а также наблюдается значительное замедление их сходимости на множестве итераций. Метод погружается в "болота" – области замедления сходимости [3,4]. Значительный интерес представляет определение близости такого "болота". В этом случае появляется возможность управлять процессом сходимости метода. Одним из важных факторов сходимости метода является определение начальных значений итерационного процесса.

В статье рассматривается процедура улучшения сходимости итерационного гребневого метода Harshman [2], основанная на применении оценок вырожденности. Будет показано, что наличие "болот" связано с наличием линейной связи между входными векторами.

1. Моделирование исходных данных. Определим трехмерный массив данных $A=(a_{ijk})$ как:

$$A_{ijk} = \sum_{r=1}^{R} x_{ir} y_{jr} z_{kr} + n_{ijk} , \qquad (1)$$

для $i = 1,2, \ldots, I$; $j = 1,2, \ldots, J$; и $k = 1,2, .., K$;
где $X_I r_R = (x_{ir}) = [X_1, \ldots, X_R]$; $Y_J r_R = (y_{jr}) = [Y_1, \ldots, Y_R]$; $Z_K r_R = (z_{kr}) = [Z_1, \ldots, Z_R]$ – факторные матрицы.

Генерация исходных данных проводилась согласно модели (1). При моделировании были использованы 64 различных массива данных, имеющих $I = 40$, $J = 40$, $K = 4$, и $R = 4$. В каждом испытании четыре столбца в наборе данных были линейно независимы, что является достаточным условием гарантии того, что ранг массива равен 4 для каждого набора. Наложение шума проводилось пропорционально к сигналу, что является более предпочтительным, чем использование аддитивного белого шума, для которого процедуры расчета МНК являются оптимальными. Поскольку шум, с которым сталкиваются на практике, является некой комбинацией пропорционального шума и аддитивного белого шума, был выбран подход, позволяющий получить общие результаты.

При проведении имитационного эксперимента использовались только случайные старты, восемь для каждого из 64 массивов данных. Для оценки связи между вырождением и сходимостью исследуемого метода, тест выполнялся для каждой пары коэффициентов, x_n – наименьший из всех возможных коэффициентов корреляции для векторов; n – соответствующий номер итерации. Для оценки степени близости итераций, использовались оценки Dn и $Y_{(n,n-1)}$: если $Y_{(n,n-1)}$ близко к 1, то два разложения на множители оценены подобно, малое значение $D_n=\log_{10}[1-Y_{(n,n-1)}]$ указывает, что текущая итерация метода произвела малые изменения, потому что последовательные разрешающие способности почти идентичны. Таким образом, логически следует завершить итерации гребневого метода в случае, когда D_n становится меньше, чем некоторое предварительно установленное значение D_0.

Типичная реализация исследуемого метода начинается с выбора начальных значений для X и Y коэффициентов модели, пусть это будут X_0 и Y_0. Используя эти коэффициенты, данные и обычный МНК – оцениваем Z, например с Z_1, Z_1, по очереди. Затем, используя X_0 получаем Y_1. Возникающая в результате тройка, (X_1,Y_1,Z_1) выражает конец первой итерации метода Harshman.

2. Оценки гребневой регрессии. Известно, что один из методов стабилизации дисперсии оценок параметров модели, в случае наличия мультиколлинеарности, гребневая регрессия. Однако, к сожалению, не всегда удается управлять оценкой гребневых параметров, как хотелось бы. То есть, если выбрать некоторую оценочную гребневую функцию, основанную на уточнении в числовой стабильности, то не всегда понятно, какими статистическими свойствами эта оценочная функция обладает [1]. В этой связи, предлагается заменить $T^{1}XA$ на более устойчивую оценочную функцию, которая сохраняет некоторые тождественные статистические свойства.

Рассмотрим уравнение регрессии $Y_N r_1 = X_N r_R b_R r_1 + e_N r_1$ с заменой обычной оценки МНК – b, полученной из $(X'X)^{-1}X'Y$. более общей оценкой – $\hat{\beta}_H = H^{-1}X'Y$. Если числа обусловленности $\hat{\beta}_H$ определены как

$$k\hat{\beta}_H = \|H\|\,\|H^{-1}\|, \tag{2}$$

где $\|H\| \in \sup_{a'a} (a'H'Ha)^{1/2}$; k – коэффициент деформации,

то можно проверить, что $k(\hat{\beta}_{МНК}) = l_1/l_R$, где l_1 $il_2 i \ldots il_R$ – собственные значения информационной матрицы $X'X$. Однако, если H взято как обычная оценочная гребневая функция $X'X + kI$. для $k>0$, то очевидно, что

$$k(\hat{\beta}_{гребн.}) = (l1+k)/(l_R+k)\ J\ l_1/l_R, \tag{3}$$

и эффект стабилизации понятен.

Аналогично, если обобщенная оценочная гребневая функция определена как

$$\hat{\beta}_{об.гребн.} = P(Dl+K)^{-1}P', \tag{4}$$

где $K=\mathrm{diag}\,(k_1,k_2,\ldots,k_R)$, $P'TP=Dl=\mathrm{diag}\,(l_1,l_2,\ldots,l_R)$ и $P'P=I$,

то следует, что $k(\hat{\beta}_{об.гребн.}) = \max\limits_i \{l_i+k_i\} / \min\limits_i \{l_i+k_i\}$ и целью является выбор такого фактора деформации k, чтобы $k(\hat{\beta}_{гребн.})\ J\ k(\hat{\beta}_{МНК})$.

Используя значение минимаксити, получим условия, при которых оценочная функция минимакса может быть выбрана для уточнения обусловленности. В случае, если $\hat{\beta}_{МНК}$ уже является минимаксом относительно среднеквадратичной ошибки, предлагается использовать процедуру, которая позволит или улучшить обусловленность данных или уменьшить среднеквадратичную ошибку модели без ухудшения обусловленности. Даже, если известно, что выбор минимаксити как статистического критерия имеет ограничения, и использование оценочных гребневых функций приводит к некоторому смещению, данная процедура позволит сократить число итераций, требуемых для сходимости исследуемого метода. Однако не всегда удается найти условие, улучшающее оценочную функцию минимакса. Даже если оценочная функция существует, уточнение в обусловленности может быть не достаточным для сохранения метода от некорректности.

Анализируя вышеупомянутые факты, предлагается использование следующей процедуры:

1) если число обусловленности T ниже некоторого (низкого) предварительно установленного

порога, то оценочная функция будет найдена, используя результаты, полученные в работе [4], что позволит уменьшить среднеквадратичную ошибку модели. Если минимаксити возможно, то эта оценочная функция не будет ухудшать число обусловленности. Иначе – будет;

2) если число обусловленности T – выше некоторого (высокого) предварительно установленного порога, используется грубая стабилизация метода на этом шаге итерации;

3) если число обусловленности T находиться между первым и вторым порогами, то внимание будет сфокусировано на уменьшении числа обусловленности. Конечно, если CIME возможен, то минимаксити также сохраняется.

На рис. 2 представлены полученные результаты, когда метод Harshman стабилизирован. Очевидно, что число итераций, потраченных в "болоте" было значительно уменьшено. В частности, на рис. 1 показано уменьшение количества итераций метода от 6500 до 300 итераций (сплошная линия – сигнал, пунктирная линия – первоначальный результат, и точечная линия – стабилизированный результат).

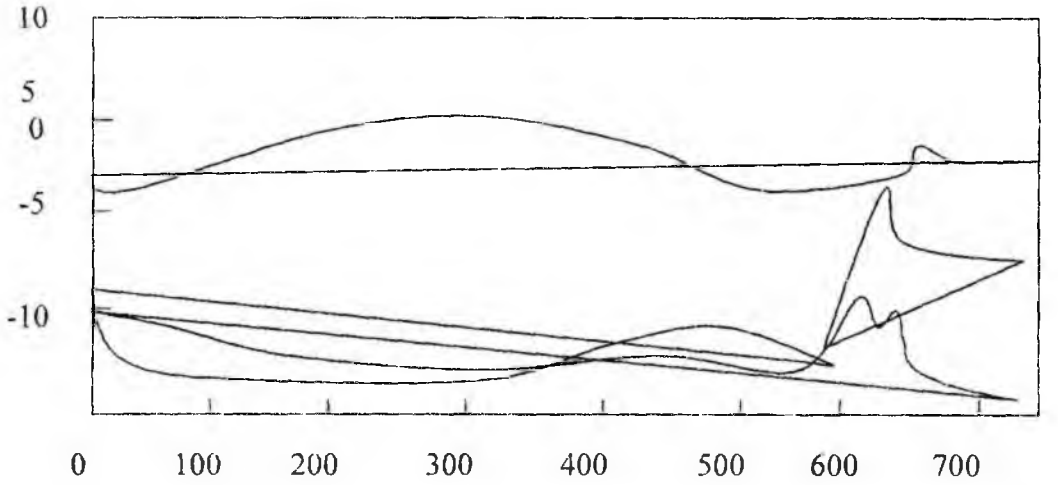

Рис. 1

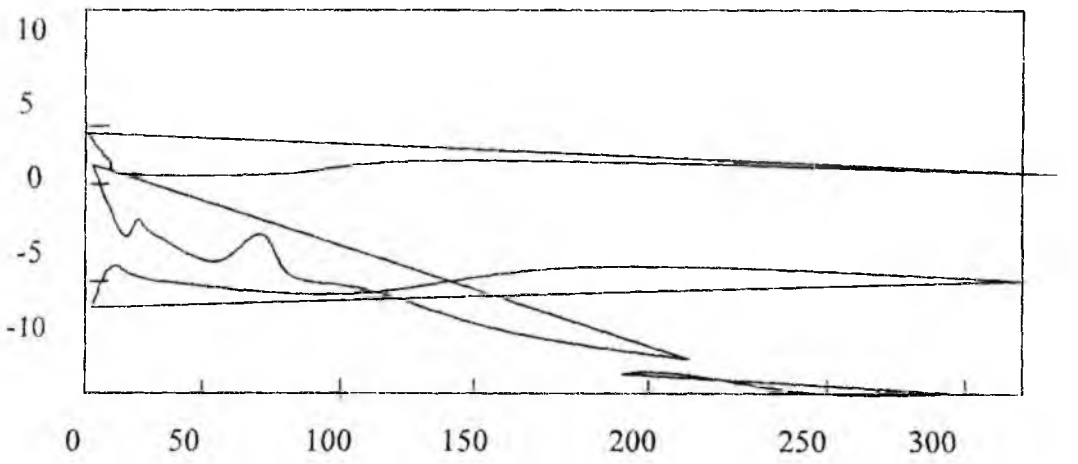

Рис. 2

Однако существует также множество других способов стабилизации МНК в случае наличия мультиколлинеарности. При этом можно ожидать, что накопление смещения оценок модели это часть цены, которую придется заплатить за использование процедуры гребневой регрессии.

3. Оценим адекватность статистических моделей, полученных на основе указанного итерационного метода. Заметим, что адекватность регрессионных моделей, полученных на базе МНК, зависит

от $\hat{\sigma}^2_{ост} = \sum_{i=1}^{n}(y_i - \hat{y})^2/(n-p)$ и дисперсии случайных возмущений σ_e^2.

Случайные величины $\hat{\sigma}^2_{ост}$ и σ_e^2 имеет χ^2 – распределение. Отношение двух случайных величин с распределением χ^2 имеет распределение Фишера (*F*). Поэтому величина $F=\hat{\sigma}^2_{ост}/\sigma_e^2$ имеет *F*-распределение с $n_{ост}$ и n_e степенями свободы.

Оценку σ_e^2 получаем из параллельных опытов. На практике получение величины σ_e^2 сопряжено с трудностями, связанными с проведением планируемых опытов и получением независимых данных при фиксированном значении остальных независимых переменных.

В случае отсутствия возможности получения σ_e^2 используют дисперсию относительно регрессии $\sigma_R^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2/\nu_R$. Для ответа на вопрос насколько полученные точечные оценки отличаются от соответствующих им истинных значений используют доверительные интервалы. Ширина доверительных интервалов зависит от величины отношения

$$t_j = \left|\beta_j^* - b_j\right| / \sigma_e\sqrt{c_{jj}}\,, \tag{5}$$

где b_j – истинное значение коэффициентов модели; c_{jj} – диагональный элемент матрицы $C=(X'X)^{-1}$.

Для оценки адекватности модели целесообразно использовать проверку гипотез о значимости регрессионных коэффициентов, провести анализ доверительных интервалов для предсказанных значений зависимой переменной и исследовать остатки.

Проведем исследование устойчивости критериев проверки гипотез на адекватность моделей, полученных методами смещенного оценивания.

Введение параметра деформации *k* в выражение $\beta^* = (X'X + kI)^{-1}X'Y$ приводит к смещению оценок и смещение их равно $E(\beta^*) - \beta = -k(X'X + kI)^{-1}\beta$ [4]. Смещение зависит как от параметра *k*, так и от значений неизвестных коэффициентов b. Квадратичную ошибку оценок коэффициентов представим в виде

$$E[L_1^2] = \sum_{j=1}^{p}(\beta_j^* - \beta_j)^2 = \sigma^2\sum_{j=1}^{p}(\lambda_j/(\lambda_j + k)^2) + \sum_{j=1}^{p}[\alpha_j^2/(\lambda_j/k + I)^2], \tag{6}$$

где λ_j – собственные числа матрицы $X'X$; a=*V*b, а *V* – ортогональная матрица, столбцы которой – собственные вектора $X'X$.

Заметим, что $V'X'XV = \Lambda$, где Λ – диагональная матрица с собственными числами на диагонали $\lambda_1, \lambda_2, ..., \lambda_p$. Из (6) видно, что с ростом параметра *k* сумма дисперсий оценок коэффициентов первого слагаемого убывает, а квадрат смещения оценок в виде второго слагаемого возрастает. Увеличение смещения приводит к увеличению значений σ_R^2 и $\sigma_{ост}^2$. Однако, корректность использования

F-отношения не нарушается, т.к. и числитель и знаменатель распределены по закону χ^2 при условии нормальности распределения разностей $(y_j - \hat{y})^2$ и $(\hat{y}_j - \bar{y})^2$. Таким образом, теоретически для смещенных оценок возможно использовать *F*-отношение для оценки адекватности модели. Однако это не означает, что модель можно использовать для прогнозирования и управления. Если размах величин, предсказуемых моделью, не слишком значительно превосходит величину случайной ошибки, модель не будет иметь никакой ценности, хотя и была получена удовлетворительная величина *F*, т.к. уравнение будет описывать только ошибки, связанные со смещением.

Изменение дисперсии ошибки модели за счет смещения оценок коэффициентов можно уменьшить за счет выбора оптимального значения *k*. Квадратичная ошибка $E(L_1^2)$ будет иметь минимум при равенстве первого и второго слагаемых в (6). Исследования расчетного значения *F*-отношения в зависимости от параметра деформации *k*, показало, что с увеличением *k* значение *F*-критерия уменьшается. Это объясняется тем, что размах предсказываемых значений отклика растет быстрее, чем стандартная ошибка отклика. Исходя из анализа исследований, можно сделать вывод, что расчетное значение *F*-отношения среднего квадрата, обусловленного регрессией, и остаточной дисперсией должно не просто превышать выбранную процентную точку *F*-распределения, а превосходить в несколько раз, т.к. выбор оптимального значения *k* не всегда осуществим.

Если наблюдения независимы и имеют некоторое отличие от нормального распределения с коэффициентом асимметрии и эксцессом не равным нулю, необходимо аппроксимировать *F*-распределение распределением со скорректированными степенями свободы ν_R и ν_e [3]. В таких случаях необходимо иметь в виду, что такая коррекция возможна только при наличии достоверного эксцесса. Оценки эксцесса требуют большого числа наблюдений и крайне чувствительны к грубым наблюдениям и выбросам. Поэтому очень важен анализ остатков модели для оценки ее адекватности.

Результаты, описанные в данной статье, предполагают потенциальную переносимость методом Harshman серьезных проблем, связанных с тем, что истинные сигналы проявляют высоко некорректные коэффициенты корреляции.

Список литературы: 1. *Casella G.* Condition Numbers and Minimax Estimators. Journal of the American Statistical Association, 1995, V. 80, N. 391, 753-758. 2. *Harshman R.A* How can I know if it"s "real"? // A Catalog of Diagnostics for Use with Three-Mode Factor Analysis and Multidimensional Scaling, Research Methods for Multilinear Data Analysis, Edited by Law Snyder, Hattie and McDonald, New York: Praeger 1984. P. 566-591. 3. *Mitchell B.C., Burdick D.S.* An Empirical Comparison of Resolution Methods for Three-Way Arrays, Chemometrics and Intelligent Laboratory Systems, 1993, V. 20. P.149 – 161. 4. *Mitchell B.C., D. S. Burdick D.S.* Slowly Converging Parafac Sequences: Swamps and Two-Factor Degeneracies, Journal of Chemometrics, 1994, V. 8. P. 155-168.

Поступила в редколлегию 31.07.2000