УДК 519.711.3

ПРОЦЕДУРА СЕЛЕКЦИИ РИДЖ-РЕГРЕССИОННЫХ МОДЕЛЕЙ

М.В. Потанина

Севастопольский национальный технический университет

Serg_potanin@mail.ru

Досліджена ефективність процедури селекції регресійних моделей в разі невизначеності структури моделі, що апроксимується рідж-регресією
Ключові слова: селекція регресійних моделей, рідж-регресія, процедура «бутстреп»

Efficiency of procedure of selection of regressive models is investigational in the case of vagueness of model structure, by the approximated ridge-regression
Keywords: selection of regressive models, ridge-regression, procedure of «bootstrap»

Исследована эффективность процедуры селекции регрессионных моделей в случае неопределённости структуры модели, аппроксимируемой ридж-регрессией
Ключевые слова: селекция регрессионных моделей, ридж-регрессия, процедура «бутстреп»

1. Постановка проблемы и ее связь с научными заданиями

Обычно идентификация сложных объектов управления осуществляется при помощи настраиваемой модели определенной структуры, параметры которой могут изменяться.

Пусть объект исследования описывается выходной переменной (функцией отклика) y и вектором независимых входных переменных $x^T = [x_1, x_2, ..., x_k]$, где k – число входных переменных. Предположим также, что существует функциональная зависимость:

$$y = \eta(x) + \varepsilon, \qquad (1)$$

где $\eta(x)$ - функция, вид который неизвестен, ε - случайная ошибка, относительно которой, как правило, предполагается, что она имеет нулевое среднее, дисперсию σ^2 и нормальный закон распределения.

Пусть также имеется априорная информация о возможных значениях переменных вида $a \le \eta(x) \le b$ и $x \in W$, где W – множество возможных значений вектора x (область планирования).

Будем рассматривать случай, когда на классе моделей задана структура, позволяющая ввести частичный порядок.

При таком порядке классы моделей как бы вложены один в другой: $S_1 \subset S_2 \subset ... \subset S_q$, S_j - модель j класса, q - максимальный порядок модели. Так, для линейных по параметрам моделей можно задать структуру в зависимости

от количества членов модели. В этом случае каждый класс S_j может быть задан следующей моделью:

$$\eta_j(\vec{x},\vec{\alpha}_j)=\vec{f}_j^{\,T}(\vec{x})\vec{\alpha}_j\ (j=1,2,...,q), \tag{2}$$

где $\vec{x}$ - вектор входных переменных, $\vec{f}_j^{\,T}(\vec{x})=[f_1(\vec{x}),f_2(\vec{x}),.....f_{nj}(\vec{x})]-$ вектор известных функций от вектора $\vec{x}$, n_j- число параметров модели j, $\vec{\alpha}_j^{\,T}=[\alpha_1,\alpha_2,...,\alpha_{n_j}]-$ вектор неизвестных параметров. Задача состоит в том, чтобы выбрать модель $\eta_j(x)$ на основе экспериментальных данных.

Пусть t - номер «наилучшей» модели, т.е. все старшие коэффициенты $\alpha_j>\alpha_m$ равны нулю. Пусть η_t - модель из класса t, который или включает в себя «наилучшую» модель $\eta(x)$ или наиболее близок к ней. Тогда все модели из классов $1,2,....,t-1$ будут иметь значимое смещение относительно модели $\eta(x)$, а модели для классов $t+1,t+2,...,q$ будут для данного объема выборки n переопределены.

Пусть имеются следующие экспериментальные данные: матрица X - значения вектора $\vec{x}$ в n опытах и вектор столбец Y - измеренные значения функции отклика в n опытах. Кроме того, известна следующая дополнительная информация: пределы изменения входных переменных $x_i\subseteq[a,b]$, пределы значения функции отклика $-c\le E\{y\}\le+c$, максимальная дисперсия ошибки измерений σ^2 и закон распределения ошибки.

Рассмотрим ситуацию, когда имеет место сильная мультиколлинеарность экспериментальных данных. Она типична для случая обработки данных заранее не спланированного эксперимента и имеет весьма отрицательные последствия для оценивания регрессионных коэффициентов. Пусть матрица X имеет полный ранг, но информационная матрица X^TX является плохо обусловленной, вследствие сильной коррелированности независимых переменных. Вектор-столбцы матрицы X при мультиколлинеарности становятся почти линейно зависимы, определитель $|X^TX|$ стремится к нулю. Так как обращение информационной матрицы связано с делением на ее определитель, который при сильной мультиколлинеарности будет очень малым числом, незначительные ошибки в вычислениях могут вызвать существенные различия в оцениваемых коэффициентах методом наименьших квадратов (МНК). Поэтому прогнозирующие свойства регрессионных уравнений ухудшаются, то есть возрастает дисперсия прогноза среднего значения переменной. Плохая обусловленность матрицы X^TX приводит к неустойчивости получаемых оценок, искажению их «физического» смысла.

2. Анализ последних исследований и публикаций

Для решения проблемы мультиколлинеарности можно перейти к смещенным оценкам, позволяющим устранить недостатки МНК-оценок в этой ситуации. К таким оценкам относятся «гребневые» или ридж-оценки [1,3]. Они являются более устойчивыми, чем оценки метода наименьших квадратов (МНК) и имеют меньшее значение среднеквадратической ошибки прогноза.

Процедура регуляризации была обоснована в целом ряде работ А.Н. Тихонова и В.Я. Арсенина. Независимо от этих работ появились результаты Херла и Кеннарда [1], а также еще целый ряд исследований в данной области.

Запишем регрессионную модель класса S_j в виде:

$$Y = F_j\vec{\alpha}_j + \vec{\varepsilon}\,, \tag{3}$$

где F_j – матрица значений функций от входных переменных в n опытах, $\vec{\varepsilon}$ - вектор-столбец случайных ошибок.

Оценки МНК в этом случае имеют вид:

$$\hat{\alpha}_{j(MÍÊ)} = \left(F_j{}^T F_j\right)^{-1} F_j{}^T Y\ , \tag{4}$$

где $F_j^T = [f_j(x_1), f_j(x_2), ...f_j(x_n)]$ - матрица значений вектора функций f_j в n экспериментальных точках для j-ой модели. Предполагается, что матрица F_j имеет полный ранг. Модель максимального размера с номером q будем считать истинной моделью.

Для стабилизации оценок МНК при решении системы нормальных уравнений в методе ридж-регрессия вместо матрицы $F_j^T F_j$ предлагается использовать матрицу $[F_j^T F_j + rI_j]$, где r – малое положительное число, добавленное к диагональным элементам матрицы $F_j^T F_j$, так называемый коэффициент регуляризации, а I_j - единичная матрица размера n_j [1]. Такие оценки могут быть записаны в виде:

$$\hat{\alpha}_{j(РИДЖ)} = (F_j^T F_j + rI_j)^{-1} F_j^T Y\ . \tag{5}$$

При смещенном оценивании параметров центральной проблемой является выбор параметра регуляризации. Главная трудность заключается в том, что смещение коэффициентов и квадратическая ошибка прогноза зависят от неизвестных истинных значений параметров.

Задача состоит в том, как выбрать «наилучшую» модель на основе экспериментальных данных.

В настоящее время предложен целый ряд критериев выбора регрессионной модели. Часть из них прямо учитывает сложность класса моделей путем введения специальных коэффициентов для величины среднеквадратического отклонения. Другие критерии используют принцип регуляризации или устойчивости оценок параметров модели при изменении объема или состава выборки.

Для линейных по параметрам классов моделей большинство квадратических критериев выбора модели объекта по эмпирическим данным можно представить в форме [2]:

$$Cr(j) = \beta_j Y^T L_j Y + \gamma_j, \ \ (j=1,2,....q), \tag{6}$$

где Cr – рассматриваемый критерий, $Y^T = [y_1, y_2, ..., y_n]$ – вектор значений функции отклика в n опытах, L_j – матрица канонической формы критерия для модели класса S_j , β_j и γ_j - константы. Обычно выбирается та модель, для которой значение критерия минимально.

Для селекции моделей, аппроксимируемых ридж-регрессией, в работе предлагается использовать критерий *Ср Маллоуза*. Этот критерий является достаточно распространенным, несложным для расчетов и регулируемым [3].

Для критерия *Ср Маллоуза* критерий (5) имеет вид:

$$C_p = C_j / \hat{\sigma}^2 - n, \tag{7}$$

где $C_j = RSS_j + \gamma k_j \hat{\sigma}^2$.

Здесь $\hat{\sigma}^2$ - оценка дисперсии ошибки случайной составляющей σ^2, полученная для модели наиболее высокого порядка, n_j - число неизвестных параметров в *j*-ой модели, γ - регулируемый коэффициент, который обычно принимается равным двум. Для классического критерия *Ср Маллоуза:* $L_j = I_n - F_j (F_j^T F_j)^{-1} F_j^T$, $\beta_j = 1$, $\gamma_j = \gamma k_j \hat{\sigma}^2$, I_n - единичная матрица размером $n \times n$.

Выбор модели из класса S_j, $j = 1,2,..,q$ будет происходить случайным образом. Пусть P_i - вероятность выбора модели *i,* тогда целью выбора метода селекции модели может случить максимизация величины P_t для минимальной истинной модели.

3. Выделение неразрешенной части общей проблемы. Формулировка цели статьи

Из анализа публикаций по данной тематике видно, что большинство исследователей подчеркивают важность устранения проблемы мультиколлинеарности и селекции полученных прогнозирующих моделей. Но,

несмотря на достаточную освещенность отдельных вопросов, связанных со смещенным оцениваем регрессионных коэффициентов, до сих пор имеется множество нерешенных вопросов, связанных с выбором прогнозирующей модели в подобной ситуации. Возникает проблема: когда при селекции лучше использовать МНК-оценки, а когда ридж-регрессию? Как выбирать оптимальный параметр регуляризации? С помощью какого критерия оценить само качество селекции? Какова будет вероятность выбора прогнозирующей модели правильной структуры?

Таким образом, целью данной работы является выбор метода селекции моделей в случае сильной коррелированности наблюдений.

4. Изложение основного материала исследований с полным обоснованием полученных научных результатов

Предлагается использовать следующую процедуру выбора модели.

На первом этапе определяется класс вложенных моделей.

Для каждой *j*-ой модели из вложенного класса находится оптимальный параметр регуляризации *r*. Так как этот параметр зависит от истинных значений коэффициентов уравнения регрессии, то оптимальное значение параметра регуляризации находится с помощью итерационной процедуры, которая состоит в следующем:

На основании априорной информации определяются границы возможных значений параметров модели.

Определяется предварительные оценки коэффициентов уравнения регрессии обычным методом наименьших квадратов (3).

Оценка дисперсии случайной составляющей определяется из модели *q* максимального размера:

$$Y = F_q \alpha_q + \varepsilon \,. \tag{8}$$

Находятся оптимальные параметры регуляризации для ридж-регрессии каждого класса моделей.

Находятся коэффициенты с помощью ридж-регрессии для каждой *j*-ой модели в соответствие с параметром регуляризации.

В работе [4,5] доказано, что для ридж–регрессии возможно применение критерия *Ср Маллоуза*. При этом происходит модификация данного критерия селекции за счет изменения матрицы канонической формы критерия для модели класса $S_{j.}$. Она примет вид:

$$L_j(РИДЖ) = I_n - F_j(F_j^T F_j + r_{OPT} I_n)^{-1} F_j^T \tag{9}$$

Для исследования поведения критерия селекции *Ср Маллоуза* в условиях ограниченной выборки была использована «*бутстреп*» процедура. Сущность

«*бутстреп*» процедуры заключается в генерации методом *Монте-Карло* на ЭВМ псевдоэкспериментальных данных с параметрами модели, вычисленными по реальным экспериментальным данным. По полученным псевдовыборкам оценивается распределение дисперсий и остаточных сумм квадратов (RSS_j) для класса моделей.

Критерий *Ср Маллоуза* является случайной величиной, поэтому необходимо знать параметры распределения индекса *j* – номера выбранной модели. Закон распределения случайной переменной *j* зависит от параметров истинной модели, выбранного критерия, дисперсии ошибки измерений и матрицы экспериментальных данных *X*.

Определим следующую индикаторную функцию:

$$\nu_j(Y) = \begin{cases} 1, \text{ если выбрана модель из класса } S_j; \\ 0 \text{ во всех остальных случаях.} \end{cases}$$

Очевидно, что функция $v_j(Y)$ будет также случайной величиной. Отношение частоты выбора модели из класса S_j к общему числу проведенных выборов модели даст оценку математического ожидания величины $v_j(Y)$ или вероятности выбора модели из класса S_j для данного критерия *Ср Маллоуза.*

В данной работе целью моделирования является идентификация «наилучшей» модели, поэтому эффективность процедуры селекции предлагается оценивать по $E\{\nu_t(Y)\}$ - вероятности выбора «наилучшей» модели.

Существует два подхода к исследованию эффективности процедуры селекции регрессионных моделей. Первый заключается в том, что предполагается $\sigma^2 = 1$ и исследуется поведение функции потерь в зависимости от значения вектора α_t и коэффициентов, позволяющих модифицировать критерий. При втором подходе выбираются «характерные» значения вектора α_t и исследуется поведение функции потерь в зависимости от значения σ^2. Так как в работе рассматривается класс объектов, для которого имеется априорная информация о пределах изменения вектора параметров $\vec{\alpha}_q$, то более эффективно исследовать поведение критерия селекции в зависимости от фиксированного вектора параметров при различных значениях дисперсии ошибки σ^2, пересчитывая, если необходимо, пределы изменения параметров.

Задавался следующий «класс» вложенных моделей S_i:

Модель 1 (недоопределенная): $y_1 = \alpha_0 + a_1 x_1 + \varepsilon_1$;

Модель 2 (истинная, «наилучшая»): $y_2 = \alpha_0 + a_1 x_1 + \alpha_2 x_2 + \varepsilon_2$;

Модель 3 (переопределенная): $y_3 = \alpha_0 + a_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \varepsilon_3$,

где ε_i - нормально распределенный вектор с нулевым математическим ожиданием и единичной дисперсией. P_i – вероятность выбора модели *i*. Проведем моделирование. Экспериментальные данные ограничены и задаются в самой программе моделирования, написанной на языке m-файлов *MATLAB6p1*. Пусть $a_i = b_i = 1$ для всех i, *c* = 1, $\sigma^2 = 1$, тогда можно вычислить

пределы изменения вектора параметров $\vec{\alpha}_q$. В данном примере пределы области допустимых значений параметров будут $0 \le \alpha_i \le 1$.

Использовались следующие матрицы экспериментальных данных X_i:

Вариант 1 – сильная корреляция.

$$X(VAR1)^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 0 & -0.8 & -1 & 0.8 & 1 & -0.8 & -1 & 0.8 & 1 & -0.8 & -1 & 0.8 \\ 1 & 0.8 & -1 & 1 & 1 & 0.8 & -1 & 1 & 1 & 0.8 & -1 & 1 \end{bmatrix}$$

Матрица коэффициентов корреляции для Варианта 1:

$$KK(VAR1) = \begin{bmatrix} 1 & 0 & 0.53 \\ 0 & 1 & 0.71 \\ 0.53 & 0.71 & 1 \end{bmatrix}.$$

Вариант 2 – слабая корреляция.

$$X(VAR2)^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 0 & -1 & 0.5 & 1 & 0 & -1 & 0 & -1 & 0 & -1 & 0.5 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Матрица коэффициентов корреляции для Варианта 2:

$$KK(VAR2) = \begin{bmatrix} 1 & -0.34 & 0 \\ -0.34 & 1 & 0.4 \\ 0 & 0.4 & 1 \end{bmatrix}.$$

Коэффициенты модели $\alpha_0 = \alpha_1 = 1$, $\alpha 2$, $\alpha 3$ изменялись от 0 до 1 с шагом 0, 1. Функция отклика Y задавалась по модели, случайная составляющая генерировалась с использованием датчика случайных чисел. Коэффициент регуляризации r менялся от 0 до 1 с шагом 0.1 в цикле. Будем считать, что для разработанной автором методики [4,5], оптимальным считается такое значение параметра регуляризации, когда вероятность выбора «истинной» модели наибольшая.

Для имитационного моделирования использовалась «*бутстреп*» процедура.

Сравнивались следующие два метода:

Метод 1 - Селекции регрессионных моделей с помощью классического критерия *Ср Маллоуза*. Оценки параметров получены с помощью метода МНК. Обозначим на рисунке (МНК);

Метод 2 - Модифицированный метод селекции моделей *Ср Маллоуза*. Оценки параметров получены с помощью ридж-регрессии. Обозначим на рисунке (РИДЖ).

Очевидно, что значение коэффициента a_1 не влияет на оценки, полученные с помощью МНК, но влияет на оценки, полученные с помощью ридж-регрессии. Значение параметра a_2 влияет на смещение оценок, а значение параметра $a_3 = 0$ для Модели 2.

Результаты имитационного эксперимента представлены на рисунке 1.

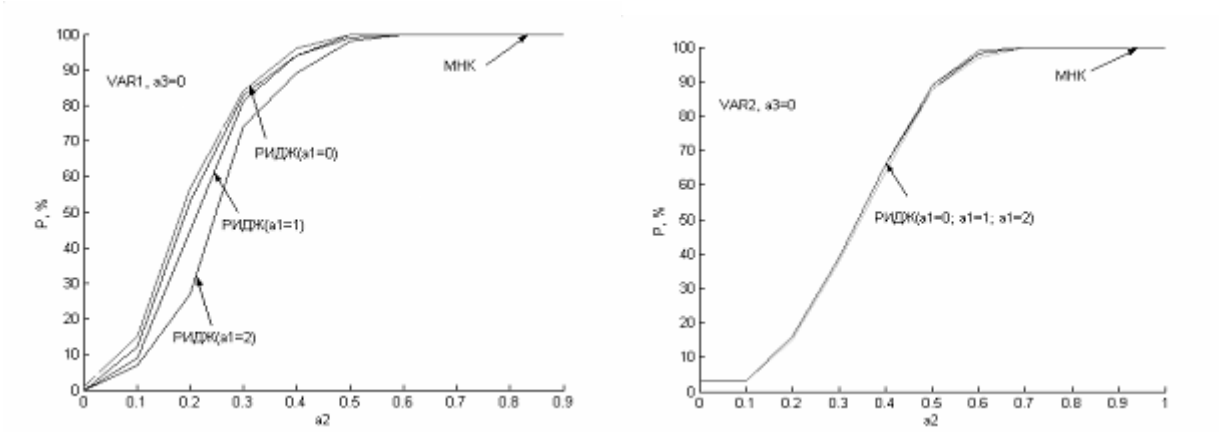

Рис. 1. Вероятность выбора «истинной» модели с помощью методов 1 и 2 для вариантов с сильной и слабой корреляцией (при $r=1$)

Было произведено исследование зависимости вероятности выбора «наилучшей» модели от параметра регуляризации r. Результаты эксперимента представлены на рисунке 2.

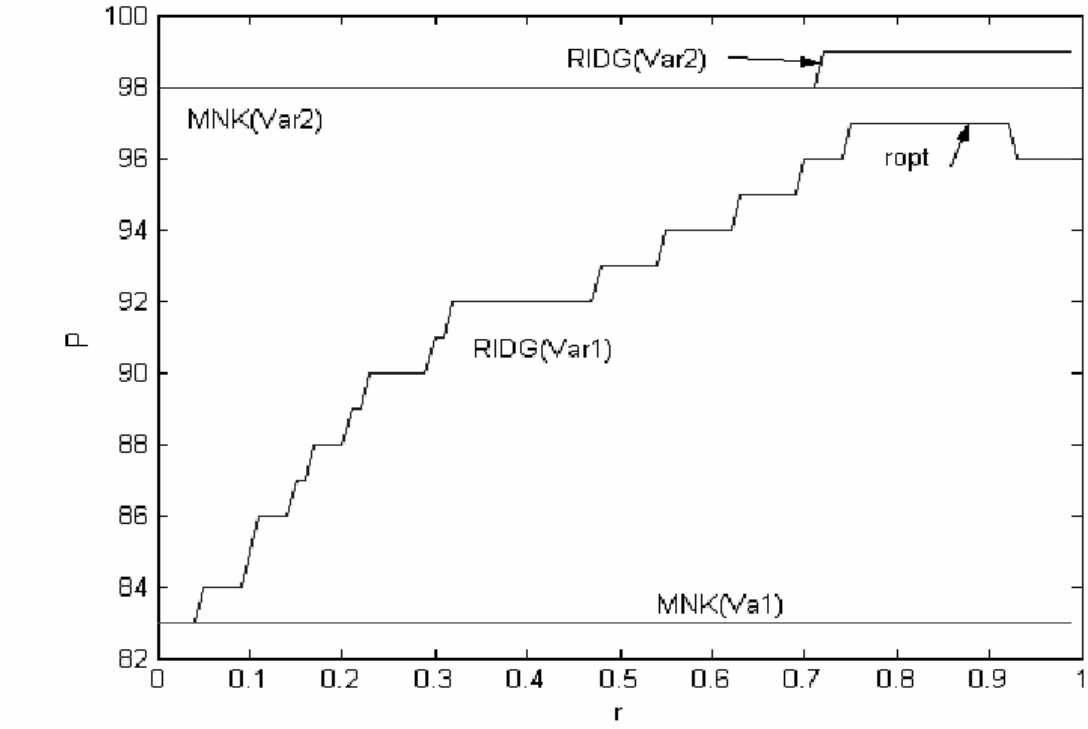

Рис. 2. Зависимость вероятности выбора «наилучшей» модели от параметра регуляризации для вариантов с разной корреляцией (при $\alpha_2 = 0.8$)

5. Выводы из данного исследования и перспективы последующих разработок в данном направлении

Результаты имитационного эксперимента показывают, что при сильной мультиколлинеарности исходных данных практически всегда лучшим

методом оценивания оказывается модифицированный метод селекции моделей *Ср Маллоуза* (РИДЖ). Вероятность выбора «наилучшей» модели с помощью ридж-регрессии будет значительно больше, чем при применении МНК.

При отсутствии мультиколлинеарности результаты, полученные с использованием модифицированного метода селекции моделей *Ср Маллоуза* и классического метода селекции моделей *Ср Маллоуза*, практически совпадают и составляют примерно 99 %. Это доказывает, что разработанную методику можно использовать для матриц экспериментальных данных с различной корреляцией. Также можно сделать вывод, что качество модели зависит от параметров.

В случае сильной мультиколлинеарности результаты эксперимента подтверждают существование оптимального параметра регуляризации, при котором вероятность выбора «наилучшей» модели оптимальна. При отсутствии корреляции результаты 1 и 2 Метода практически совпадают. Таким образом, доказано, что вероятность выбора «наилучшей» модели зависит от регрессионных коэффициентов и параметра регуляризации.

Предметом дальнейших исследований предполагается произвести сравнительный анализ эффективности методов селекции моделей для других видов регуляризованных оценок, в частности, для сжимающих оценок Джеймса–Стейна. А также рассмотреть задачу оптимизации нахождения координат точки оптимального параметра регуляризации для ридж-регрессии.

Литература

1. Hoerl A. Ridge regression: biased estimation for no orthogonal problems / A. Hoerl., R. Kennard. // Technometrics. — Vol. 12. — 1970. — P. 55 —67.

2. Herzberg A. M., Tsukanov A.V. The design of experiments for model selection with the Jackknife criterion// Utilitus Mathematica. — 1985. — Vol. 28. — P. 243-253.

3. Айвазян С.А. Прикладная статистика: Исследование зависимостей: Справ. Изд./ С.А Айвазян., И.С Енюков., Л.Д Мешалкин.; под ред. С.А. Айвазяна— М.: Финансы и статистика, 1985. — 487 с.

4. Потанина М.В. Построение модели оценки рентабельности автотранспортного средства методом смещенного оценивания параметров регрессионных уравнений / М.В. Потанина // Вестник СевНТУ. Экономика и финансы: сб. науч. тр., Севастополь, 2008. – Вып. 92. – С. 166 – 171.

5. Потанина М.В. Оценка качества идентификации сложного объекта управления в условиях мультиколлинеарности / М.В. Потанина // Вестник СевНТУ. Сер. Автоматизация процессов управления: сб. науч. тр. – Севастополь, 2009. — Вып. 95. —— С. 135 ——140.