УДК 519.711.3

М.В. Потанина, доцент, канд. техн. наук
Севастопольский национальный технический университет
ул. Университетская, 33, г. Севастополь, Украина, 99053
E-mail: mem@sevgtu.sebastopol.ua

СВОЙСТВА СРЕДНЕКВАДРАТИЧЕСКОЙ ОШИБКИ ПРОГНОЗА РИДЖ-РЕГРЕССИИ ДЛЯ ИДЕНТИФИКАЦИИ МОДЕЛИ СЛОЖНОГО ОБЪЕКТА УПРАВЛЕНИЯ ПРИ МУЛЬТИКОЛЛИНЕАРНОСТИ ФАКТОРОВ

Произведено исследование функции среднеквадратической ошибки прогноза (СКОП) для ридж-регрессии на экстремум в зависимости от параметра регуляризации. Доказано наличие локального минимума СКОП, что позволяет решить задачу поиска оптимального параметра регуляризации для идентификации модели сложного объекта управления при мультиколлинеарности факторов.

Ключевые слова: *среднеквадратическая ошибка прогноза, ридж-регрессия, параметр регуляризации, идентификация, мультиколлинеарность.*

Постановка проблемы. Идентификация модели сложного объекта управления (СОУ) при мультиколлинеарности факторов является сложной задачей.

Одним из способов решения проблемы мультиколлинеарности является использование смещенных оценок, позволяющих устранить недостатки традиционного метода наименьших квадратов (МНК). К таким оценкам относятся «гребневые» или ридж-оценки. Но при этом необходимо учитывать, что в настоящее время не существует однозначно рекомендуемого способа выбора параметра регуляризации ридж-регрессии.

Особенностью выбора данного параметра является то, что и смещение и СКОП зависят от неизвестных истинных значений параметров модели.

Предлагается методика исследования СКОП, как критерия оптимальности качества идентификации модели СОУ в заданной области в случае ридж-регрессии, на экстремум в зависимости от параметра регуляризации, что доказало бы возможность автоматизированного вычисления оптимального параметра регуляризации.

Анализ последних исследований и публикаций. В случае обработки данных заранее не спланированного эксперимента при исследовании функционирования СОУ часто встречается явление, называемое мультиколлинеарностью [1,4]. Отрицательные последствия данного явления проявляются в следующем:

– неустойчивость оценок регрессионных коэффициентов. При добавлении или исключении совсем малого количества информации (например, одного наблюдения) может произойти очень сильное изменение оценок коэффициентов. При этом резко уменьшается и точность предсказания по модели;

– численная неустойчивость процедуры оценивания, вызванная ошибками машинного округления и накопления этих ошибок;

– коэффициенты регрессионной модели оказываются сильно коррелированными между собой, что лишает смысла их интерпретацию;

– резко увеличиваются дисперсии оценок коэффициентов.

Все это затрудняет построение прогнозирующих моделей СОУ и делает невозможным использования оценок типа МНК, так как точность их при мультиколлинеарности становится недостаточной [2].

Однако в этом случае можно повысить эффективность оценивания за счет привлечения априорной информации об оцениваемых параметрах уравнений регрессии вероятностного или детерминированного характера. Но это в свою очередь приводит к использованию смещенных оценок, позволяющих уменьшить СКОП по сравнению с МНК-оценками и тем самым получить значения параметров, наиболее близких к истинным.

Смещенные, например, ридж-оценки являются более устойчивыми, чем оценки МНК и имеют меньшее значение СКОП при коррелированности факторов, вызванной мультиколлинеарностью [1].

Главной трудностью при использования ридж-оценок является выбор параметра регуляризации r.

Процедура регуляризации представляет собой попытку преодолеть последствия плохой обусловленности информационной матрицы, как следствия мультиколлинеарности. Эта идея была обоснована в целом ряде работ как зарубежных, так и отечественных авторов [1, 3].

К сожалению, во всех подходах, рассматриваемых в указанных работах, обычно требуется дополнительная информация, которая редко имеется в распоряжении исследователя и поэтому ряд

вопросов, связанных с идентификацией модели объекта управления в условиях мультиколлинеарности факторов, остаются нерешенными.

Цель статьи. В настоящей работе приводится методика исследования СКОП как критерия оптимальности в случае ридж-регрессии на экстремум в зависимости от параметра регуляризации. Доказано наличие локального минимума СКОП, что позволяет решить задачу автоматического вычисления оптимального параметра регуляризации при смещенном оценивании для идентификации модели СОУ при мультиколлинеарности факторов.

Изложение основного материала исследования с полным обоснованием полученных научных результатов. В работе рассматривается случай, когда на классе возможных 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 – максимально возможный порядок модели. Так для линейных по параметрам моделей можно задать структуру в зависимости от количества членов модели. В этом случае каждый класс S_j задается следующей моделью

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

где $f_j^T(\vec{x}) = [f_1(\vec{x},\vec{z}), f_2(\vec{x},\vec{z}),..., f_{n_j}(\vec{x},\vec{z})]$ – вектор известных функций от векторов $\vec{x}$ и $\vec{z}$, n_j – число неизвестных параметров модели j, $\vec{\alpha}_j^T$ – вектор неизвестных параметров j-ой модели.

В качестве априорной информации в работе рассматриваются следующие данные: пределы среднего значения функции отклика ($\eta_{\min}$ и $\eta_{\max}$), область возможных значений факторов, максимальная дисперсия случайной составляющей модели σ^2 и закон распределения этой составляющей.

СКОП для любой модели в области прогноза W_1 можно оценить функционалом

$$EI(\vec{\alpha}) = \int_{W_1} (y - \eta(\vec{x},\vec{\alpha}))^2 P(\vec{x},\vec{y}) d\vec{x} dy, \tag{2}$$

где $P(\vec{x},\vec{y})$ – совместная плотность распределения переменных $\vec{x}$ и y.

Пусть матрица $F_j^T = [\vec{f}_j(\vec{x}_1), \vec{f}_j(\vec{x}_2),..., \vec{f}_j(\vec{x}_N)]$ – матрица функций от экспериментальных данных для модели j размером $N \times n_j$ и $Y^T = [y_1, y_2,..., y_N]$ – вектор N наблюдений над функцией отклика. Предполагается, что матрица F_j имеет полный ранг.

Если модель линейная по параметрам и используется ридж-регрессия, тогда оценка определится как

$$\vec{\hat{\alpha}}_{j(RIDG)} = (F_j^T F_j + rI_j)^{-1} F_j^T Y, \tag{3}$$

где r – параметр регуляризации.

Выражение для СКОП модели j в случае линейной параметризации при использовании ридж-регрессии, имеет вид

$$\begin{aligned} EI(j)_{RIDG} = {} & \vec{\alpha}_j^T (I_j - F_j^T F_j (F_j^T F_j + rI_j)^{-1}) J_j (I_j - F_j^T F_j (F_j^T F_j + rI_j)^{-1}) \vec{\alpha}_j + \\ & + \vec{\alpha}_{q-j}^T J_{q-j} \vec{\alpha}_{q-j} + \vec{\alpha}_{q-j}^T F_{q-j}^T F_j (F_j^T F_j + rI_j)^{-1} J_j (F_j^T F_j + rI_j)^{-1} F_j^T F_{q-j} \vec{\alpha}_{q-j} - \\ & - 2\vec{\alpha}_j^T (I_j - F_j^T F_j (F_j^T F_j + rI_j)^{-1}) J_j (F_j^T F_j + rI_j)^{-1} F_j F^T{}_{t-j} \vec{\alpha}_{t-j} + \\ & + \sigma^2 sp(F_j B_j J_j B_j F_j^T). \end{aligned} \tag{4}$$

где $B_j = (F_j{}^T F_j + rI_j)^{-1}$ и $J = \int_{W_1} \vec{f}(x) \vec{f}(x)^T d\vec{x} / \int_{W_1} d\vec{x}$.

Для полной модели q, включающей истинную модель, СКОП в случае ридж-регрессии имеет вид

$$EI(q)_{RIDG} = \vec{\alpha}^T (I_q - F_q{}^T F_q B_q) J (I_q - B_q F_q{}^T F_q) \vec{\alpha} + \sigma^2 sp F_q B_q J B_q F_q{}^T, \tag{5}$$

где $B_q = (F_q{}^T F_q + rI_q)^{-1}$.

Исследуем СКОП для полной модели q на экстремум в случае ридж-регрессии. Предполагается, что если функция СКОП в случае ридж-регрессии имеет локальный минимум, то он и будет соответствовать оптимальному значению параметра регуляризации.

Продифференцируем выражение (5) по *r*. Здесь и в дальнейшем будем считать, что $B = B_q$. Имеем

$$\frac{dEI}{dr} = \vec{\alpha}^T F^T F B^{-2} J\vec{\alpha} - \vec{\alpha}^T F^T F B^{-2} J B^{-1} F^T F\vec{\alpha} + \vec{\alpha}^T J B^{-2} F^T F\vec{\alpha} -$$

$$- \vec{\alpha}^T F^T F B^{-1} J B^{-2} F^T F\vec{\alpha} - \sigma^2 sp(F B^{-2} J B^{-1} F^T + F B^{-1} J B^{-2} F^T). \quad (6)$$

Экстремум функции EI в точке r_{OPT} будет при $\frac{dEI}{dr} = 0.$

Очевидно, что это выражение зависит от параметра регуляризации *r*, от конкретных собранных экспериментальных данных, от истинных коэффициентов и от дисперсии ошибки.

Найдем условие существования минимума для СКОП. Для этого необходимо найти вторую производную выражения СКОП $\frac{d^2EI}{dr^2}$.

$$\frac{d^2EI}{dr^2} = -2\vec{\alpha}^T B^{-3} J F^T F\vec{\alpha} + 2\vec{\alpha}^T F^T F B^{-3} J B^{-1} F^T F\vec{\alpha} + \vec{\alpha}^T F^T F B^{-2} J B^{-2} F^T F\vec{\alpha} +$$

$$+ \sigma^2 sp(2F B^{-3} J B^{-1} F^T + F B^{-2} J B^{-2} F^T) > 0 \ . \quad (7)$$

Формула (7) будет являться условием существования минимума для СКОП.

Найдем значение оптимального параметра регуляризации r_{OPT}. Для этого раскроем скобки в выражении (7). Имеем $\frac{dEI}{dr} = 0$, тогда, так как в формуле (7) данные слагаемые попарно равны, то

$$\vec{\alpha}^T F^T F B^{-1} J B^{-2} F^T F\vec{\alpha} = \vec{\alpha}^T F^T F B^{-2} J B^{-1} F^T F\vec{\alpha};$$

$$\vec{\alpha}^T F^T F J B^{-2}\vec{\alpha} = \vec{\alpha}^T B^{-2} J F^T F\vec{\alpha};$$

$$\sigma^2 sp(F B^{-1} J B^{-2} F^T) = \sigma^2 sp(F B^{-2} J B^{-1} F^T).$$

Формулу (7) можно переписать в виде

$$\frac{dEI}{dr} = 2*(\vec{\alpha}^T B^{-2} J F^T F\vec{\alpha} - \vec{\alpha}^T F^T F B^{-2} J B^{-1} F^T F\vec{\alpha} - \sigma^2 sp(F B^{-2} J B^{-1} F^T)) = 0.$$

Преобразуем это выражение, умножив на B^2 слева и разделив на 2:

$$\frac{dEI}{dr} = \vec{\alpha}^T J F^T F\vec{\alpha} - \vec{\alpha}^T F^T F J B^{-1} F^T F\vec{\alpha} - \sigma^2 sp(F J B^{-1} F^T) = 0.$$

Далее умножаем на *B* правую часть выражения.

$$\frac{dEI}{dr} = \vec{\alpha}^T F^T B J F\vec{\alpha} - \vec{\alpha}^T F^T F J F^T F\vec{\alpha} - \sigma^2 sp(F J F^T) = 0. \quad (8)$$

Подставим B в формулу (8), имеем

$$\vec{\alpha}^T J(F^T F + I_n r) F^T F\vec{\alpha} - \vec{\alpha}^T F^T F J F^T F\vec{\alpha} - \sigma^2 sp(F J F^T) = 0;$$

$$\vec{\alpha}^T J F^T F F^T F\vec{\alpha} + r\vec{\alpha}^T J F^T F\vec{\alpha} - \vec{\alpha}^T F^T F J F^T F\vec{\alpha} - \sigma^2 sp(F J F^T) = 0;$$

$$r\vec{\alpha}^T J F^T F\vec{\alpha} - \sigma^2 sp(F J F^T) = 0.$$

Окончательно имеем

$$r_{OPT} = \sigma^2 sp(F J F^T) / \vec{\alpha}^T J F^T F\vec{\alpha}. \quad (9)$$

Исследование свойств СКОП для ридж-регрессии с помощью имитационных экспериментов. В качестве тестовой рассматривается модель вида $y = \alpha_o + a_1 x_1 + \alpha_2 x_2 + \varepsilon$, где $\varepsilon \approx N(0,1)$ для трех различных вариантов матриц экспериментальных данных *X*1, *X*2, *X*3, отличающихся степенью корреляции.

Матрицы экспериментальных данных имеют вид

$$X1=\begin{bmatrix}-1 & -1\\ -1 & +1\\ +1 & -1\\ +1 & +1\end{bmatrix};\qquad X2=\begin{bmatrix}+1 & +1\\ -1 & +1\\ -1 & +1\\ -1 & -1\end{bmatrix};\qquad X3=\begin{bmatrix}-1 & -1\\ +1 & -1\\ +1 & +1\\ -1 & -1\end{bmatrix}.$$

Эти матрицы удовлетворяют условию (7). При проведении имитационных экспериментов: $a1=1$, $a2$ изменялся от 1 до 2. Параметр регуляризации r менялся на отрезке [0;2] с шагом 0,1. В тестовых планах присутствует сильная мультиколлинеарность. Значение r_{OPT} рассчитывалось по формуле (9), а также определялось с помощью имитационного эксперимента. Экспериментальные данные функции отклика Y вычислялись по модели, где случайная составляющая генерировалась с использованием датчика случайных чисел. Предполагается, что в каждой точке имитационного эксперимента было проведено по 100 опытов.

Для проведения имитационного эксперимента было разработано программное обеспечение на языке m-файлов MATLAB.

В таблице 1 представлены результаты вычисления оптимального значения параметра регуляризации r_{OPT} по формуле (9) и полученного экспериментальным путем, в зависимости от варианта экспериментальных данных.

Таблица 1– Результаты вычислений для полной модели

Вариант	Значение коэффициентов	Значение r_{OPT} (расчетное)	Значение r_{OPT} (экспериментальное)
X1	a2=1	1	0,99
	a2=2	0,4	0,41
X2	a2=1	1	0,99
	a2=2	0,4	0,39
X3	a2=1	0,4	0,40
	a2=2	0,2	0,21

Результаты теоретического вывода и имитационного эксперимента практически совпадают. Таким образом, функция СКОП имеет локальный минимум, которым и является r_{OPT}.

На рисунках 1 – 3 изображены графики зависимости СКОП EI от значения коэффициента регуляризации r для различных вариантов экспериментальных данных.

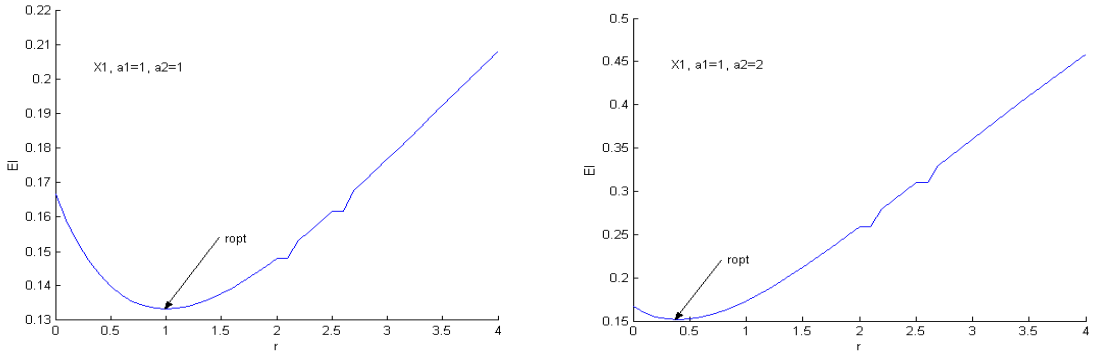

Рисунок 1 – График зависимости СКОП от значения коэффициента регуляризации r для варианта $X1$($a2$=1,2)

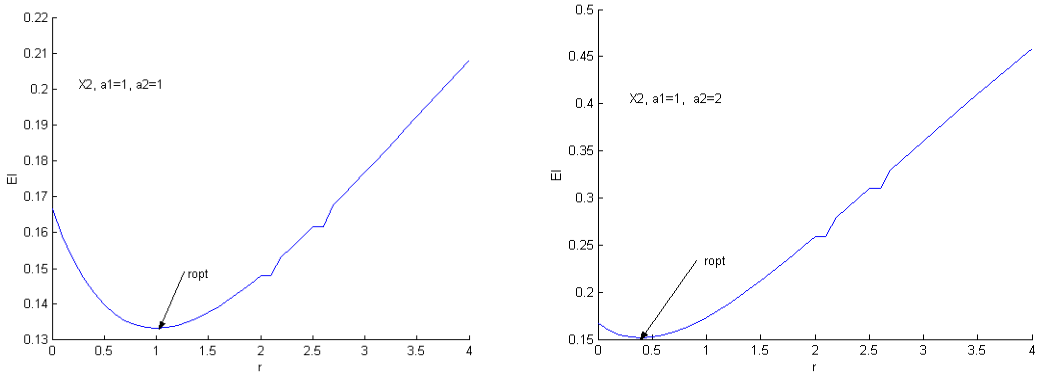

Рисунок 2 – График зависимости СКОП от значения коэффициента регуляризации r для варианта $X2$ ($a2$=1,2)

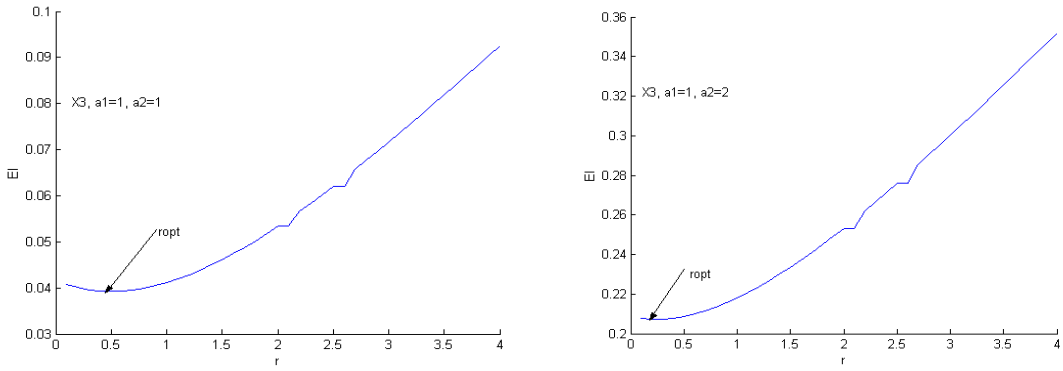

Рисунок 3 – График зависимости СКОП от значения коэффициента регуляризации r для варианта $X3$ ($a2$=1,2)

Формула (9) может быть использоваться только для ограниченного количества матриц экспериментальных данных F, для которых выполняется условие (7), например, для матриц с данными в точках границы для переменных. Для остальных вариантов матриц F определение r_{OPT} является оптимизационной поисковой задачей. Для таких матриц экспериментальных данных предлагается использовать итерационный метод поиска r .

Выводы и перспективы дальнейших исследований. В результате применения разработанной методики проведено исследование СКОП СОУ на предмет существования локального минимума, что доказало и позволило определить значение оптимального параметра регуляризации при смещенном оценивании параметров регрессионных уравнений. После проведенных имитационных экспериментов видно, что полученное в результате использования методики значение оптимального параметра регуляризации r_{OPT} практически не отличается от экспериментального. Это подтверждает правильность разработанной методики.

Таким образом, можно прийти к выводу, что в рассматриваемой задаче достигнута поставленная цель – повышено качество прогнозирующей модели СОУ в условиях мультиколлинеарности, с учетом неопределенности структуры модели и метода смещенного оценивания параметров модели, что свидетельствует о корректности использованных и разработанных методов управления, созданного программного обеспечения на языке m-файлов MATLAB, реализующее предложенную методику.

Предметом дальнейших исследований предполагается изучение итерационного метода нахождения оптимального значения параметра регуляризации r_{OPT} для ридж-регрессии.

Библиографический список использованной литературы

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

2. Бородич С.А. Эконометрика: учеб. пособие для вузов / С.А. Бородич. — Минск: Новое знание, 2001. — 416 с.

3. Тихонов А.Н. Методы решения некорректных задач / А.Н. Тихонов, В.Я. Арсенин. — М.: Наука, 1974. — 276 с.

4. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников / А.И. Кобзарь. — М.: Физматлит, 2006. — 816 с.

Поступила в редакцию 17.01.2014 г.

Потаніна М.В. Властивості середньоквадратичної помилки прогнозу рідж-регресії для ідентифікації моделі складного об'єкта управління при мультиколінеарності факторів

Проведено дослідження функції середньоквадратичної помилки прогнозу (СКПП) у разі рідж-регресії на екстремум в залежності від параметра регуляризації. Доведено наявність локального мінімуму СКПП, що дозволяє вирішити задачу пошуку оптимального параметра регуляризації для ідентифікації моделі складного об'єкта управління при мультиколінеарності факторів.

Ключові слова: середньоквадратична помилка прогнозу, рідж-регресія, параметр регуляризації, ідентифікація, мультиколінеарність.

Potanina M.V. The properties of the mean-squared prediction error ridge regression to identify models of complex control object by multicollinearity

The research on the functions of the mean-squared prediction error (MSPE) for ridge-regression to the extremum depending on the regularization parameter. Proved the existence of a local minimum of MSPE, that allows solving a problem of search of the optimal regularization parameter to identify models of complex control object by multicollinearity.

Keywords: mean-squared prediction error, ridge regression, the regularization parameter, identification, multicollinearity.