УДК 111.11

ИДЕНТИФИКАЦИЯ ПАРАМЕТРОВ ЛИНЕЙНОЙ РЕГРЕССИИ ПРИ ПРОИЗВОЛЬНЫХ ВНЕШНИХ ПОМЕХАХ В НАБЛЮДЕНИЯХ

А.А. Сенов
Санкт-Петербургский государственный университет
Россия, 199034, Санкт-Петербург, Университетская наб., 7-9
E-mail: sensanek@gmail.com

О.Н. Граничин
Санкт-Петербургский государственный университет
Россия, 199034, Санкт-Петербург, Университетская наб., 7-9
E-mail: oleg_granichin@mail.com

Ключевые слова: рандомизация, линейная регрессия, идентификация параметров

Аннотация: В работе рассматривается задача построения не асимптотического доверительного множества для параметров множественной линейной регрессии. Вводится рандомизированный алгоритм *модифицированных знако-возмущенных сумм*, являющийся расширением метода *знако-возмущенных сумм* на случай почти произвольных помех. Оцениваются статистические свойства полученного доверительного множества. Приводится иллюстрация и сравнения методов на модельных данных.

1. Введение

Идентификация систем состоит из двух важных задач: определения математической модели динамической системы и оценки параметров установленной модели или их границ на основе зашумленных измерений [1]. Шум может породжаться как различием между системой из реального мира и выбранной математической моделью, так и поступать извне модели, в том числе и умышленно внедрятся оппонентом. Помехи последнего типа, порожденные источником находящимся вне модели, мы будет называть *почти произвольными помехами*. Важно заметить, что почти произвольные помехи могут не только быть распределены по любому закону, но и иметь детерменированную природу. Как бы то ни было, зачастую сложно определить реальную природу и закон распределения помех.Поэтому, развитие методов оценки качечства модели является еще однойважной проблемой идентификации систем [1]. Оценка параметров модели может быть проведена двумя способами. Первый заключается в построении оценок точных значений параметров. Второй состоит в построении множества, содержащего истинное значение параметра с заданной вероятностью. Мы сфокусируемся на задаче построения доверительного множества.

В этой работе мы сфокусируемся на идентификации параметров множественной линейной регрессии вида

$$y = x^T\theta + \varepsilon,$$

где y — рейтинг данного объекта для данного пользователя, x — вектор характеристик данного объекта, θ — неизвестный вектор предпочтений характеристик данного пользователя, ε — «шум», неизвестные, произвольные помехи.

Одной из основных задач линейной регрессии заключается в нахождении наилучшей в среднеквадратичном смысле аппроксимации параметра θ по обучающему множеству — заданному множеству значений пар (x, y) [2] [3] [4], а так же определения доверительного множества для θ. Способы определения параметра θ в условиях различных предположений о природе x, y и ε хорошо изучены, в том числе и при почти произвольных помехах (смотри, например, [5, 6, 9]). С другой стороны, до недавнего времени были известны способы определения лишь асимптотических доверительных множеств для параметра θ, которые применимы при определенных предположениях о распределении помех [1]. Значительным недостатком этих методов является то, что они гарантируют результат лишь при количестве измерений стремящимся к бесконечности, в то время как для небольшого количества измерений результаты могут быть крайне неудовлетворительны. К сожалению, на практике распространены случаи, когда обучающее множество мало из-за чего использование асимптотических методов оказывается крайне неэффективным и возникает необходимость в методах, которые могут обеспечить гарантированную точность результата при конечном числе измерений.

В случае произвольных помех любые методы, основанные на априорных предположениях о природе помех, оказываются неэффективными. Распространенным приемом для решения задач в условиях произвольных помех является *рандомизация*, заключающаяся в добавлении в алгоритм дополнительных, случайных, но контролируемых экспериментатором, влзмущений. Рандомизированные алгоритмы успешно используется как в методах оценивания параметров при почти произвольных помехах [5–8, 10], так и в некоторых алгоритмах построения доверительного множества для параметров системы (смотри [12–14]). Естественно ожидать, что и для задачи построения точного доверительного множества параметра линейной регрессии рандомизация может принести свои плоды.

В работе авторов Marco Campi и Erik Weyer [15] был предложен метод Знако-Возмущенных Сумм (SPS, Sign-Perturbated Sums) для определения точного доверительного множества параметра линейной регрессии при центрированных и симметрично распределенных помехах. Одним из основных ограничений этого метода является условие симметричности распределения помехи ε относительно нуля. Это ограничение затрудняет применение данного метода, так как во многих практических задачах может быть не только смещенной, но и иметь неслучайную природу. Более того, некоторые важные характеристики полученных методом SPS множеств, такие как его размер, вариация границ и др., остаются неизвестными, что затрудняет их применение на практике.

Таким образом, задача определения так называемых «точных», не асимптотического доверительного множества параметра θ припочти произвольных параметрах остается открытой, как и их характеристика. Интерес представляют другие подходы в из области рандомизированных алгоритмов, описанных в том числе в [2], позволяющие отказаться от ограничений на распределение помехи.

Эта работа является продолжением статьи [16], в которой предлагается модификация метода, предложенного в [15], расширяющая его на случай несимметрично и не центрированно распределенных помех при дополнительном условии на симметричности и известности среднего распределение векторов характеристик объектов x. Актуальность этой модификации заключается в том, что значения шума не являются центрированными и симметричными, в то время как среднее векторов характеристик объектов с легкостью поддается оценке, а симметричность их распределения следует из закона больших чисел.

Далее в Разделе 2 описывается формальная постановка задачи, в Разделе 3 приводится теоретическое обоснование и формулировка алгоритма Модифицированных Знако-Возмущенных Сумм, в Разделе 4 проводится сравнение алгоритмов MSPS, SPS и классического метода построения доверительного множества на модельных данных. В Приложении приведены доказательства теоретических рузельтатов.

2. Постановка задачи

Модель линейной регрессии (а точнее, ее решение методом наименьших квадратов) была получена независимо Карлом Фридрихом Гауссом и Адриеном Мари Лежандром еще в начале 19-го века и и нашла широкое применения во многих областях науки благодаря своей простоте и эффективности. Для успешного ее применения необходимо учесть некоторые особенности распределения параметров модели. В некоторых практических задачах можно предположить симметричность распределения переменной входа относительно своего среднего. Примером могут служить задачи, где переменные входа либо формируются экспериментатором, либо же он может влиять на их формирование. Сформулируем соответствующую модель множественной линейной регрессии

$$y_i = \phi_i^T \theta^\star + \varepsilon_i, i = 1..N, \tag{1}$$

где N — число измерений, $\phi_i \in \mathrm{R}^p$ — наблюдаемые значения входа, $\varepsilon_i \in \mathrm{R}$ — неизвестная помеха, $y_i \in \mathrm{R}$ — наблюдаемые значения выхода, а $\theta^\star$ — неизвестный параметр модели. При этом предполагается, что $\{\phi_i\}_{i=1}^N$ имеют вероятностную природу, а помехи $\{\phi_i\}_{i=1}^N$ могут быть как случайными, так и детерменироваными, а также выполнены следующие предположения

- $\phi_1, \ldots, \phi_N$ — независимы, симметрично и одинаково распределены вокруг известного математического ожидания m_ϕ
- $\varepsilon_1, \ldots, \varepsilon_N$ неопределенные внешние помехи, независимые от $\{\phi_i\}_{i=1}^N$

Классическим решением в задаче нахождения оценки $\hat{\theta}$ параметра $\theta^\star$ является метод наименьших квадратов, который заключается в поиске минимума функционала $\sum_{i=1}^N ||y_i - \phi_i\theta||_2^2$. Полученная оценка называется оценкой наименьших квадратов и записывается следующим образом

$$\hat{\theta}_{LSE} = \arg\min \sum_{i=1}^N ||y_i - \phi_i^T\theta||_2^2, \tag{2}$$

Указанный минимум достигается в точках, где производная соответствующего функционала равна нулю

$$\sum_{i=1}^{N} \phi_i (y_i - \phi_i^T \theta) = 0$$

В этой работе ставится иная задача: вместо поиска оценки истинного значения параметра $\theta^\star$, необходимо найти множество, которому это значение принадлежит с заданной вероятностью. Иными словами, требуется по данным наблюдениями $\{(\phi_i, y_i)\}_{i=1}^N$ и параметру $\alpha \in [0,1]$ определить множество $\Theta_\alpha(\{(\phi_i, y_i)\}_{i=1}^N)$, такое что

$$P\left(\theta^\star \in \Theta_\alpha(\{(\phi_i, y_i\}_{i=1}^N)\right) = \alpha. \tag{3}$$

Стоит отметить, что равенство достигается не асимптотически, а выполняется при любом N.

3. Теоретическое решение

Для метода SPS, предложенного в статье [15], обязательным условием является центрированность ошибок, что делает его неприменимым по отношению к модели (1). Однако, можно провести доказательство, аналогичное приведенному в статье [15], использующее симметричность и известность среднего распределения входов, вместо помех, как это сделано в некоторых метода оценки при почти произвольных помехах (смотри [2]).

Введем обозначения

$$S_k(\theta) = \sum_{i=1}^{N} a_{i,k} \Delta_i (y_i - \phi_i \theta), \tag{4}$$

$$S_0(\theta) = \sum_{i=1}^{N} \Delta_i (y_i - \phi_i \theta), \tag{5}$$

где $\Delta_i = \phi_i - m_\phi$ — центрированные значения входа, $a_{i,k}$ — так называемые *случайные знаки*, определяемые по формуле

$$a_{k,i} = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

По аналогии с *знако-возмущенными суммами*, введенными в [15], $S_k(\theta)$ называются *модифицированными знако-возмущенными суммами*, где модифкация заключается в введении *рандомизированных* множителей $\{\Delta_i\}$.

Алгоритм *Модифицированных Знако-Возмущенных Сумм* (MSPS) построение не ассимптотического доверительного множества параметра линейной регрессии θ в предположениях 2. в базируется на суммах $S_k(\theta)$ и формулируется следующим образом.

Основная идея приведенного алгоритма MSPS заключается в том, что если Δ_i распределены симметрично вокруг нуля, то суммы $S_k(\theta^\star) = \sum_{i=1}^N a_{k,i}\Delta_i \varepsilon_i$ и $S_0(\theta^\star) = \sum_{i=1}^N \Delta_i \varepsilon_i$ одинаково распределены благодаря тому, что вектора $(a_{k,1}\Delta_1, \ldots, a_{k,N}\Delta_N)$

Алгоритм 1 MSPS algorithm

procedure MSPS($\{y_i\}_{i=1}^N$, $\{\phi\}_{i=1}^N$, m_ϕ,M, q, θ)
for i in $1 \dots N$ **do**
  $n_i \leftarrow y_i - \phi_i^T \cdot \theta$
  $\Delta_i \leftarrow \phi_i - m_\phi$
for k in $1 \dots M-1$ **do**
  $S_k \leftarrow 0 \in \mathbb{R}^p$
  for i in $1 \dots N$ **do**
    $S_k \leftarrow S_k + \Delta_i \cdot n_i \cdot RANDOMSIGN()$
  $Z_k \leftarrow ||S_k||_2^2$
$S_0 \leftarrow 0 \in \mathbb{R}^p$
for i in $1 \dots N$ **do**
  $S_0 \leftarrow S_0 + \Delta_i \cdot n_i$
$Z_0 \leftarrow ||S_0||_2^2$
$r \leftarrow 0$
for k in $1 \dots M-1$ **do**
  if $Z_0 > Z_k$ **then**
    $r \leftarrow r + 1$
if $r >= q$ **then**
  return $FALSE$
else
  return $TRUE$
end procedure

и $(\Delta_1, \dots, \Delta_N)$ распределены одинаково. Таким образом, если мы обозначим $Z_k = ||S_k(\theta^\star)||$ при $k = 0, \dots, M-1$ то Z_0 примет любую позицию в сортированном по возрастанию списке $Z_{(0)}, \dots, Z_{(M-1)}$ с вероятностью $\frac{1}{M}$. Следовательно, Z_0 *не* будет среди $q \leqslant M$ наибольших значений $\{Z_k\}_{k=0}^{M-1}$ с *точной* вероятностью $1 - \frac{q}{M}$.

Стоит отметить, что при $E\phi_i = 0$ MSPS алгоритм идентичен алгоритму SPS.

Следующая теорема утверждает, что значения θ для которых алгоритм 1 возвращает $TRUE$ формируют точный $(1 - \frac{q}{M})$-доверительное множество для параметра $\theta^\star$.

Теорема 1. *Рассмотрим модель линейной регрессии* (1). *Пусть* $M \geqslant q \geqslant 0$ *— два натуральных числа,* $\{S_k(\theta)\}_i = 1^{M-1}$ *—* $M-1$ *модифицированныз знаковозмущенных суммы,*

Тогда, обозначив $Z_k(\theta) = ||S_k(\theta)||_2^2$ *и*

$$\Theta_{M,q} = \{\theta \in \mathbb{R}^p : |\{k \in \{1, \dots, M-1\} : Z_0(\theta) < Z_k(\theta)\}| \geqslant 1\} \tag{6}$$

$\Theta_{M,q}$ *будет точным* $(1 - \frac{q}{M})$*-доверительное множества для параметра* $\theta^\star$

$$P_{\varepsilon_i, a_{k,i}}(\theta^\star \in \Theta_{M,q}) = 1 - \frac{q}{M}.$$

Для доказательства Теоремы 1 мы сперва приведем лемму, следующую из теоремы, доказанной авторами в статье [15], которая обосновывает методы SPS.

Лемма 1. *B.C. Csaji, M.C. Campi and E. Weyer [15] Пусть* g_0 *слуайная величина*

$$g_0 = ||\sum_{i=1}^N t_i x_i||_2^2,$$

где $\{t_i\}_{i=1}^N$ *— независимые случайные случайные величины симметричные относительно нуля и* $\{x_i\}_{i=1}^N$ *— некоторые величины, возможно детерменированные, независимые с* $\{t_i\}_{i=1}^N$. *Если* $a_{k,i}$ *— случайные знаки и* $\{g_k\}_{k=1}^{M-1}$ *определить как*

$$g_0 = ||\sum_{i=1}^N a_{k,i} t_i x_i||_2^2,$$

тогда $\{g_k\}_{k=0}^{M-1}$ *— независимы и одинаково распределены. Более того, вероятность того, что* g_0 *не будет среди* q *наибольших значений из* $\{g_k\}_{k=0}^{M-1}$ *равна* $1 - \frac{q}{M}$.

Доказательство теоремы 1. Достаточно показать, что $\theta^\star \in \Theta_{M,q}$ с вероятностью $1 - \frac{q}{M}$, что эквивалентно условию, что $Z_0(\theta^\star)$ меньше как минимум q различных $Z_k(\theta^\star)$ из $\{Z_k(\theta^\star)\}_{k=1}^{M-1}$ с этой вероятностью. Так как $Z_0(\theta^\star)$ и $Z_k(\theta^\star)$ могут быть записаны в виде

$$Z_k(\theta^\star) = ||\sum_{i=1}^N a_{k,i}\Delta_i\varepsilon_i||, \quad Z_0(\theta^\star) = ||\sum_{i=1}^N \Delta_i\varepsilon_i||,$$

то, обозначив $x_i = \varepsilon_i$ и $t_i = \Delta_i$ для $i = 1, \ldots, N$ мы можем применить лемму 1 которая утверждает, что все случайные (в следствии случайности Δ_i) значения $\{Z_k(\theta^\star)\}_{k=0}^{M-1}$ независимы и одинаково распределены и $Z_0(\theta^\star)$ не среди q наибольших значений из $\{Z_k(\theta^\star)\}_{k=1}^{M-1}$ с точной вероятностью $1 - \frac{q}{M}$.

Стоит отметить, что множество $Theta_{M,q}$ является решением поставленной задачи 3 при $\alpha = 1 - \frac{q}{M}$.

Замечание 1. *Приведенные в данном разделе теорема 1 и алгоритм 1 основаны на теореме и алгоритме для метода SPS, приведенных в статье [15] и в значительной степени заимствует ее приемы. Несмотря на это, полученные результаты позволяют значительно расширить область применения алгоритма.*

4. Свойства доверительного множества

Далее, рассмотрим задачу простой (одномерной) линейной регрессии

$$y_i = \phi_i\theta + \varepsilon, \tag{7}$$

где все величины — скаляры, а ε — некоторая помеха. Хотя данный случай достаточно тривиален, он дает нам возможность продемонстрировать некоторые примечательные свойства доверительного множества.

Сперва, для определения характеристик доверительного множества, получим аналитическое выражение его границ.

4.1. Вычисление доверительного множества

Формула доверительного множества, полученная в [15], дает описание доверительного множества в виде обратной от выражения

$$\theta : \exists\{i_k\}_{k=1}^q \subset \{1, .., M-1\} : Z_0(\theta) < Z_{i_k}(\theta)\forall \quad k = 1..q,$$

что не очень удобно и вычисление которой далеко не тривиально (предполагается использование приближенно-численных процедур). Кроме того, отсутствие аналитического выражения границ доверительного множества соответственно лишает возможности определить его характеристики аналитически. Логичным представляется получить это выражения, путем разбиения данного выражения на более простые. Опишем $\Theta_{M,q}$ тщательней. Простыми словами, $\Theta_{M,q}$ состоит из точек $\theta \in \mathbb{R}^d$, для которых $Z_0(\theta)$ превосходит не более чем $M-q-1$ различных $Z_k(\theta)$. Что эквивалентно, $\Theta_{M,q}$ состоит из точек θ, для которых $Z_0(\theta)$ меньше как минимум q различных $Z_k(\theta)$ — так что по меньшей мере q различных неравенств $Z_0(\theta) < Z_k(\theta)$ выполняется. Используя это интерпретацию, $\Theta_{M,q}$ может быть записан в следующем образом

Замечание 2.

$$\Theta_{M,q} = \bigcup_{\mathbb{I}\subset\{1,..,M-1\}:|\mathbb{I}|=q} \left(\bigcap_{k\in\mathbb{I}} \{\theta : Z_0(\theta) < Z_k(\theta)\} \right) \neq \varnothing$$

Важность подобного представления заключается в том, что вычисление $\Theta_{M,q}$ может быть сведено к вычислению $\{\theta : Z_0(\theta) < Z_k(\theta)\}$.

Лемма 2. *Пусть $q = 1$. Тогда*

$$\{\theta : Z_0(\theta) < Z_k(\theta)\} = (B_k^{min}, B_k^{max}), \quad where$$

$$B_k^{(1)} = \frac{\sum_{i=1}^N (1-a_{k,i})\phi_i y_i}{\sum_{i=1}^N (1-a_{k,i})\phi_i^2}, \quad B_k^{(2)} = \frac{\sum_{i=1}^N (1+a_{k,i})\phi_i y_i}{\sum_{i=1}^N (1+a_{k,i})\phi_i^2}$$

$$B_k^{min} = \min(B_k^{(1)}, B_k^{(2)}), \qquad B_k^{max} = \max\left(B_k^{(1)}, B_k^{(2)}\right)$$

Доказательство леммы 2. Рассмотрим неравенство $Z_0(\theta^\star) < Z_k(\theta^\star)$

$$(\sum_{i=1}^N \Delta_i(y_i - \phi_i\theta))^2 < (\sum_{i=1}^N a_{k,i}\Delta_i(y_i - \phi_i\theta))^2$$

$$(\sum_{i=1}^N \Delta_i(y_i - \phi_i\theta))^2 - (\sum_{i=1}^N a_{k,i}\phi_i(y_i - \phi_i\theta))^2 < 0$$

$$(\sum_{i=1}^N (1-a_{k,i})\Delta_i(y_i - \phi_i\theta))(\sum_{i=1}^N (1+a_{k,i})\Delta_i(y_i - \phi_i\theta)) < 0$$

$$\left(\frac{\sum_{i=1}^N (1-a_{k,i})\Delta_i(\varepsilon_i + \Delta_i\theta^\star)}{\sum_{i=1}^N (1-a_{k,i})\Delta_i^2} - \theta \right) \cdot$$

$$\cdot \left(\frac{\sum_{i=1}^N (1+a_{k,i})\Delta_i(\varepsilon_i + \Delta_i\theta^\star)}{\sum_{i=1}^N (1+a_{k,i})\Delta_i^2} - \theta \right) < 0.$$

Стоит отметить, что данная лемма дает аналитическое выражения границ множества $\{\theta : Z_0(\theta) < Z_k(\theta)\}$ для одномерного случая, и его вычисление значительно более эффективно, чем поточечное вычисления с помощью алгоритма 1.

Утверждение 1. *Пусть B_k является $B_k^{(1)}$ или $B_k^{(2)}$ и в дополнении к 2. выполняются предположения*

- $\exists$ *второй и четвертый моменты для* $\{\phi_i\}$
- $\{\varepsilon_i\}$ *является либо случайной величиной с вторым нецентрированным моментом ограниченным константой* $C < \infty$, *либо детерменированной последовательностью, равномерно ограниченной константой* $\sqrt{C}$.

Тогда

$$E_{\varepsilon,\phi,a}[B_k] = \theta^\star,$$
$$E_{\varepsilon,\phi,a}[B_k - E_{\varepsilon,\phi,a}[B_k]]^2 \xrightarrow[N\to\infty]{} 0,$$

где $E_{\varepsilon,\phi,a}$ *— математическое ожидание совместного распределения* $\{\varepsilon_i\}_{i=1}^N, \{\phi_i\}_{i=1}^N$ *и* $\{a_i\}_{i=1}^N$

Доказательство утверждения 1.

$$E_{\varepsilon,\phi,a}B_{\{a_{k,i}\}} = E_{\varepsilon,\phi,a}\left[\frac{\sum_{i=1}^N(1\pm a_{k,i})\Delta_i(\varepsilon_i+\Delta_i\theta^\star)}{\sum_{i=1}^N(1\pm a_{k,i})\Delta_i^2}\right] =$$

$$= \theta^\star + E_{\varepsilon,\phi,a}\left[\frac{\sum_{i=1}^N(1\pm a_{k,i})\Delta_i\varepsilon_i}{\sum_{i=1}^N(1\pm a_{k,i})\Delta_i^2}\right]$$

Так как Δ_i независимы с ε_i и $E\Delta_i = 0$

$$E_{\varepsilon,\phi,a}\left[\frac{\sum_{i=1}^N(1\pm a_{k,i})\Delta_i\varepsilon_i}{\sum_{i=1}^N(1\pm a_{k,i})\Delta_i^2}\right] = \sum_{i=1}^N|E_{\varepsilon,\phi,a}\left[\frac{(1\pm a_{k,i})\Delta_i\varepsilon_i}{\sum_{j=1}^N(1\pm a_{k,j})\Delta_j^2}\right]| \leqslant$$
$$\leqslant \sum_{i=1}^N|E_{\varepsilon,\phi,a}\left[\frac{(1\pm a_{k,i})\Delta_i\varepsilon_i}{\sum_{j\neq i}^N(1\pm a_{k,j})\Delta_j^2}\right]| = 0.$$

Таким образом первая часть утверждения доказана. Для доказательства второй части, заметим что $\forall\{x_i\}_1^N$ неравенство $2\sum_1^N x_i^2 \geqslant (\sum_1^N x_i)^2$ выполняется. Следовательно,

$$E_{\varepsilon,\phi,a}[B_{\{a_{k,i}\}} - E_{\varepsilon,\phi,a}B_{\{a_{k,i}\}}]^2 \leqslant E_{\varepsilon,\phi,a}\left[\frac{(\sum_{i=1}^N(1\pm a_{k,i})\Delta_i\varepsilon_i)^2}{(\sum_{i=1}^N(1\pm a_{k,i})\Delta_i^2)^2}\right] \leqslant$$

$$\leqslant CE_{\varepsilon,\phi,a}\left[\frac{(\sum_{i=1}^N(1\pm a_{k,i})\Delta_i\varepsilon_i)^2}{(\sum_{i=1}^N(1\pm a_{k,i})\Delta_i)^4}\right] \xrightarrow[N\to\infty]{} 0,$$

Так как ε_i независимы с Δ_i мы можем брать их математическое ожидания отдельно Беря математическое ожидание по ε_i и используя наложенные на них ограничения

$$E_{\varepsilon,\phi,a}[B_{\{a_{k,i}\}} - E_{\varepsilon,\phi,a}B_{\{a_{k,i}\}}]^2 \leqslant CE_{\varepsilon,\phi,a}\left[\frac{(\sum_{i=1}^N(1\pm a_{k,i})\Delta_i)^2}{(\sum_{i=1}^N(1\pm a_{k,i})\Delta_i)^4}\right] =$$

$$= CE_{\varepsilon,\phi,a}\left[\frac{1}{(\sum_{i=1}^N(1\pm a_{k,i})\Delta_i)^2}\right] \xrightarrow[N\to\infty]{} 0.$$

Таким образом, утверждение доказано.

Это предложение дает два важных следствия. Во-первых, границы интервала $\Theta_{M,q}$ B_k в среднем равны $\theta^\star$, колеблясь вокруг истинного значения параметра. Во-вторых, эти границы стремятся к $\theta^\star$ с увеличением N. Резюмируя, доверительный интервал сходится $\theta^\star$, когда $N \to \infty$.По аналогии с оценками параметров, доверительный интервал $\Theta_{M,q}$ можно назвать *несмещенным* и *состоятельным*.

5. Практическое сравнение

Опишем классический метод построения доверительног множества для случая ассимптотический номральных помех.

5.1. Ассимптотическое доверительное множество

Одним из стандартных методов построения оценки $\theta^\star$ параметра является метод наименьших квадратов, оценка которого $\hat{\theta}_{LSE}$ определяется по формуле 2 Известно, что при определенных условиях на распределение помех ε_i оценка $\hat{\theta}_{LSE}$ ассимптотически нормальная и $cov(\hat{\theta}_{LSE} - \theta^\star)$ сходится к $\Xi_N = \hat{\sigma}^2 \sum_{i=1}^{N} (\phi_i \phi_i^T)^{-1}$, где $\hat{\sigma}^2$ — оценка дисперсии помех [1]. Пользуясь этим, ассимптотическое доверительное множество для параметра $\theta^\star$ может быть построен по следующей формуле

$$\Theta_\alpha = \{\theta : (\theta - \theta_{LSE})^T \Xi_N^- 1 (\theta - \theta_{LSE}) \leqslant \chi_\alpha(1) \hat{\sigma}^2 / N\}, \tag{8}$$

где $\chi_\alpha(1)$ — α-квантиль распределения χ^2 с 1 степенью свободы. К сожалению, это ассимптотическое множество в условиях 2. является очень грубой оценкой и может привести к ошибкам и некорректным результатам.

5.2. Описание модельных данных

Мы заинтересованы в случайх, подпадающих под условия 2.. Кроме того, нам интересн случай малого количество измерений, так как он соответствует проблеме холодного старта и именно здесь преимущества алгоритма MSPS проявляются в особой степени. Мы рассматриваем сдучай $\phi_i \in \mathbb{R}^2$ из-за простоты визуализации. В этом разделе, ϕ_i моделируется следующим образом

$$\phi_i \sim N(\mu_\phi, \Sigma_\phi), \quad \mu_\phi = (1, -1)^T, \quad \Sigma_\phi = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$$

и истинное значение параметра $\theta^\star = (-1, 2)^T$ always. Далее в примерах изменяется лишь закон распределения помех $\{\varepsilon_i\}$.

Раздел разделен на две части: вначала рассматривается случай относительно большого количества измерений ($N = 50$), а затем иллюстрируется случай малогочисла измерений ($N = 15$). На каждом графике изображены три типа доверительных множества: полученное алгоритмом SPS, полученное алгоритмом MSPS и ассимптотическое доверительное множество, полученное по формуле (8). *Замечание по обозначениям*: под $N(0, 1)$ понимается стандартное нормальное распределение, а под $Exp(\lambda = 1)$ понимается экспоненциальное распределение с параметром масштаба равным 1.

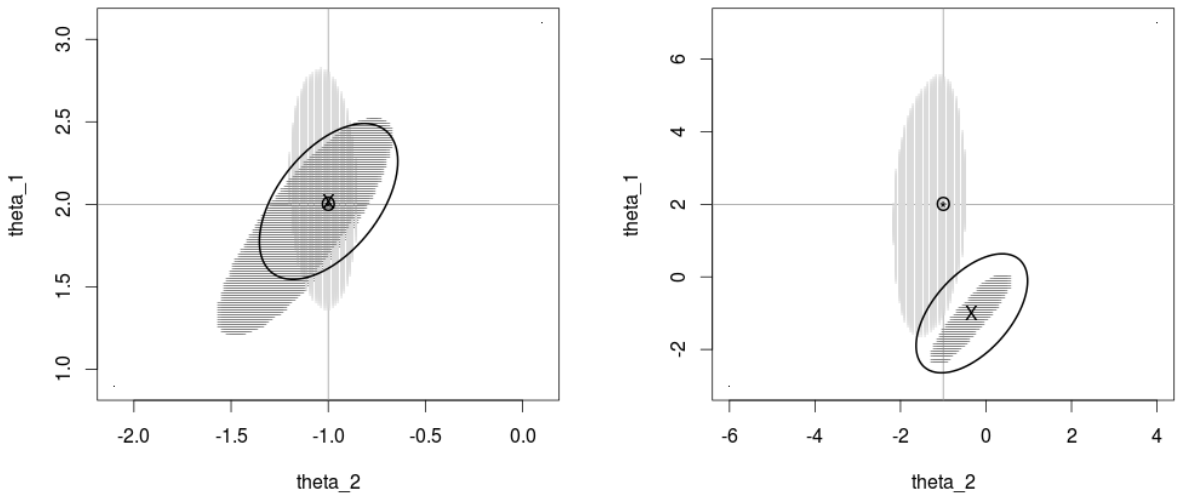
5.3. Случай большого числа измерений

Рассмотрим два типа помех: *несмещенные симметричные* помехи и *смещенные ассиметричные*.

1. $\varepsilon_i \sim N(0,1)$
2. $\varepsilon_i \sim Exp(\lambda = 1) + 4$

По рис. 1 (a), который соответствует помехам со стандартным нормальным распределением 1, можно заметить что все три доверительных множества содержат как истинное значение параметра так и оценку методом МНК. Примечательно, что MSPS and SPS дают значительно большее по размеру чем ассимптотическое доверительное множество.

Рис. 1 (b) соответствует смещенным помехам и показывает преимущества алгоритма MSPS. Несмотря на то, что доверительное множество, полученное алгоритмом MSPS увеличилось в размерах, оно единственное содержит истинное значение параметра, в то время как и ассимптотическое и основанное на SPS доверительные множества сместились из-за а ассимметричности и смещенности помех.



(a) Несмещенные симметричные помехи

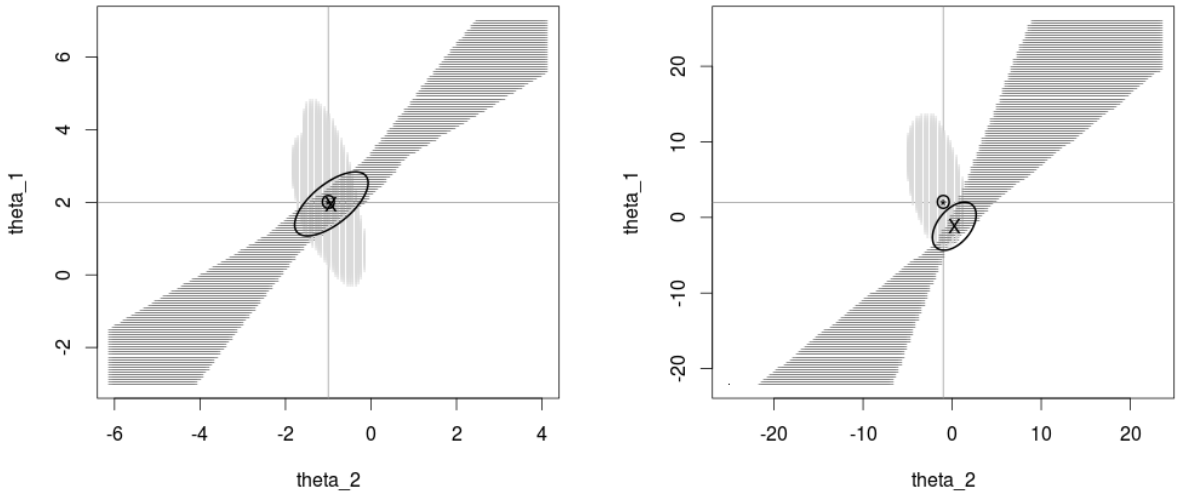
(b) Смещенные асимметричные помехи

Рис. 1: Доверительные множества, полученные SPS (темные горизонтальные линии), MSPS (светлые вертикальные линии) и ассимптотический (эллипсоид). Истинное значение параметра отмечено кругом с точкой, оценка МНК отмечена крестом

5.4. Случай малого числа измерений

Рассмотрим два типа шума: *несмещенный ассимметричный* и *смещенный ассимметричный*

1. $\varepsilon_i \sim Exp(\lambda = 1) - 1$



(a) Несмещенные асимметричные помехи

(b) Смещенные асимметричные помехи

Рис. 2: Доверительные множества, полученные SPS (темные горизонтальные линии), MSPS (светлые вертикальные линии) и ассимптотический (эллипсоид). Истинное значение параметра отмечено кругом с точкой, оценка МНК отмечена крестом

2. $\varepsilon_i \sim Exp(\lambda = 1) + 4$

Малое число измерений играет важную роль так как эффект ассимметрии шума усиливается. Так как при большом количестве имзерений помехи становятся ассимптотически нормальными в силу Центральной Предельной Теоремы. Как бы то ни было, рис. 2 (a) соответствующий случаю несмещенных помех 1 иллюстрирует, что доверительное множество полученное методом MSPS содержит окрестность точки $\theta^\star$, в то время как SPS дает неограниченное доверительное множество, что неприемлимо. Схожая ситуация изображена на рис. 2 (b) соответствуещем несимметричной смещенной помехе 1 и ассимптотичсекое доверительное множество смещено от истинного значения параметра $\theta^\star$.

6. Заключение

Задача определения не асимптотического доверительного множества для параметра множественной линейной регрессии еще недавно оставалась нерешенной задачей. Благодаря рандомизированной процедуре Знако-Возмущенных Сумм, введенной Марко Кампи и Эриком Вейером в статье [15], стало возможным определение точных доверительных множества заданной вероятности с относительно слабыми ограничениями на центрированность и симметричность помехи. К сожалению, в статье не описывается эффективного способа вычисления доверительного множества, а его вычисление напрямую крайне неэффективно.

В статье введена процедура Модифицированных Знако-Возмущенных Сумм, позволяющая расширить область применимости на случай почти произвольных помех

с неизвестным, но существующим вторым моментом. Получено аналитическое выражение границ доверительного множества для одномерного случая, а так же выведены их основные статистические свойства (первый и второй моменты). Приведены практические примеры, показывающие превосходство модифицированной процедуры над оригинальной в случае почти произвольных помех.

Работа выполнена при частичной финансовой поддержке РФФИ (грант №13-07-00250-а).

Список литературы

1. Льюнг Л. Идентификация систем: теория для пользователя. М.: Наука. 1991. 431 с.
2. Граничин О.Н., Поляк Б.Т. Рандомизированные алгоритмы оценивания и оптимизации при почти произвольных помехах. М.: Наука. 2003. 293 с.
3. Montgomery D.C., Peck E.A., Vining G.G. Introduction to Linear Regression Analysis. Wiley-Interscience, 2007.
4. Фомин В.Н. Элементы регрессионного анализа. М.: Наука, 1984. С. 36–45.
5. Граничин О.Н. Оценивание параметров линейной регрессии при произвольных помехах // Автоматика и телемеханика. 2002. № 1. С. 30-41.
6. Granichin O.N. Linear Regression and Filtering under Nonstandard Assumptions (Arbitrary noise) // IEEE Trans. on Automat. Contr. 2004. Vol. 49, No. 10. P. 1830-1835.
7. Goldenshluger A.V., Polyak B.T. Estimation of regression parameters with arbitrary noise // Mathematical Methods of Statistics. 1993. Vol. 2, No. 1. P. 18-29.
8. Граничин О.Н. Алгоритм стохастической аппроксимации с возмущением на входе для идентификации статического нестационарного дискретного объекта // Вестник Ленинградского университета. 1988. Сер. 1. Вып. 3. № 15. С. 92-93.
9. Фомин В.Н. Рекуррентное оценивание и адаптивная фильтрация. М.: Наука. 1984. 288 с.
10. Граничин О.Н., Фомин В.Н. Адаптивное управление с использованием пробных сигналов // Автоматика и телемеханика. 1986. № 2. С. 100-112.
11. Граничин О.Н. Неасимптотическое доверительное множество для параметров линейного объекта управления при почти произвольных помехах // Автоматика и телемеханика. 2012. № 1. С. 24-35.
12. Amelin K., Amelina N., Granichin O., Granichina O. Two procedures with randomized controls for the parameters' confidence region of linear plant under external arbitrary noise // Proc. of the IEEE Int. Symposium on Intelligent Control ISIC '2012. P. 1226-1231.
13. Amelin K., Amelina N., Granichin O., Granichina O. Combined procedure with randomized controls for the parameters' confidence region of linear plant under external arbitrary noise // Proc. of the 51st Conference on Decision and Control CDC '2012. P. 2134-2139.
14. Amelin K., Granichin O. Randomized Controls for Linear Plants and Confidence Regions for Parameters under External Arbitrary Noise // Proc. of the American Control Conference ACC '2012. P. 743-1618.
15. Csaji B.C., Campi M.C., Weyer E. Non-asymptotic conidence regions for the least-squares estimate // Proceedings of the 16th IFAC Symposium on System Identification SYSID '2012. P. 227–232.
16. Сенов А.А. Доверительные множества при почти произвольных помехах в контексте линейных моделей рекомендательных систем // Стохастическая оптимизация в информатике. 2013. Т. 9, Вып. 1. С. 68-86.
17. Wei C.-P., Shaw M.J., Easley R.F. A survey of recommendation systems in electronic commerce // E-Service: New Directions in Theory and Practice. 2001. P. 12.