Известия НАН Армении, Математика, том 55, н. 3, 2020, стр. 68 – 74

СОСТОЯТЕЛЬНОСТЬ ИЗБЫТОЧНОГО РИСКА В РОБАСТНОМ ОЦЕНИВАНИИ ГАУССОВСКОГО СРЕДНЕГО

А. Г. МИНАСЯН

Ереванский Государственный Университет
E-mail: *arsh.minasyan@gmail.com*

Аннотация. В данной работе мы вводим понятие избыточного риска для оценки среднего вектора Гауссовского распределения, когда наблюдения искажены выбросами. Известно, что выборочное средство теряет свои хорошие свойства в присутствии выбросов [5, 6]. Кроме того, даже выборочная медиана неоптимальна в миминимаксном смысле скорость оптимальная в многомерном случае. Оценка достигающяя оптимальной минимаксной скорости был установлен в [1]. Однако даже эти результаты оптимальности минимаксной скорости не дают количественной оценки того, насколько быстро загрязненная модель приближается к риску в незагрязненной модели, когда скорость загрязнения стремится к нулю. Данная статья делает первый шаг в заполнении этого пробела, показывая, что существует оценка с избыточным риском, стремящимся к нулю, когда пропорция выбросов приближается к нулю.

MSC2010 number: 62F35.

Ключевые слова: избыточный риск; робастное оценивание; среднее нормального распределения.

1. Введение

В последние годы мы стали свидетелями возрождения интереса к статистическим методам, которые могут эффективно работать с наборами данных, содержащие выбросы. В частности, в модели загрязнения Хубера в задаче оценки среднего вектора нормального распределения [1] установил оптимальную минимаксную скорость и показал, что она достигается в случае многомерной медианы, известная как медиана Тьюки. Более того, [2] разработал общую теорию для получения минимаксной скорости (верхняя и нижняя границы) в широком классе статистических моделей. Эти работы ориентированы на статистическую сложность оценок, не обращая внимания на вычислительную сложность. Вычислительная сложность оценок был адресован в [4] и [3], которые проанализировали риск вычислимых оценок. Интересно, что результаты, доказанные в этих работах, обеспечивают только порядок минимаксной скорости, но ничего не говорят о том, насколько быстро риск в загрязненной выбросами модели стремится к риску без выбросов.

В этой статье мы вводим понятие избыточного риска, которое определяется как разница между рисками в моделях с выбросами и без. Затем мы представляем анализ этого риска для процедуры, которую мы назвали “group hard thresholding”. Это также можно рассматривать как версию усеченной средней оценки. Наш главный результат показывает что этот избыточный риск стремится к нулю, когда уровень загрязнения стремится к нулю.

Более формально, давайте предположим, что мы наблюдаем n случайных векторов $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ in $\mathbb{R}^p$, которые удовлетворяют

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{\theta}_i + \boldsymbol{\xi}_i, \quad \boldsymbol{\xi}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, I_p). \tag{1.1}$$

В верхней формуле, $\boldsymbol{\mu}$ неизвестный параметр, которую мы хотим оценить, $\{\boldsymbol{\theta}_i\}$ это произвольные детерминированные векторы, указывающие какие из наблюдений являются выбросами и $\boldsymbol{\xi}_i$ случайный шум. В этой статье, мы предполагаем, что $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1 \ldots \boldsymbol{\theta}_n]$ усеченная по столбцам матрица. Все наблюдения с индексами $i \in \mathcal{O} = \{\ell : \|\boldsymbol{\theta}_\ell\|_2 > 0\}$ являются выбросами, в то время, как остальные приходят из $\mathcal{N}(\boldsymbol{\mu}, I_p)$. Пусть

$$o = \operatorname{Card}(\mathcal{O}), \quad \text{и} \quad \varepsilon = \frac{o}{n}.$$

Предполагается, что параметр ε меньше $1/2$, который играет важную роль в робастном оценивании. В астности, известно что минимаксная скорость оценки в модели (1.1) имеет порядок $\frac{p}{n} + \varepsilon^2$.

В этой статье мы рассмотриваем более точную меру точности оценки, избыточный риск. Напомним, что риск оценки [1] $\widehat{\boldsymbol{\mu}}$ определяется как

$$R[\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}; \boldsymbol{\Theta}] = [\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Theta}} \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2]^{1/2}.$$

Здесь и далее в статье обозначение $\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Theta}}[h]$ означает математическое ожидание по распределению $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$ как определено в (1.1) (мы неявно предполагаем, что h зависит от наблюдений $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$). Это хорошо известный факт, что в случае, где нет выбросов, т.е. когда $\boldsymbol{\Theta} \equiv \mathbf{0}_{p\times n}$ риск удовлетворяет равняется $\sqrt{p/n}$. Формально,

$$\inf_{\widehat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathbb{R}^p} R[\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}; \mathbf{0}] = \sup_{\boldsymbol{\mu} \in \mathbb{R}^p} R[\overline{\boldsymbol{Y}}_n, \boldsymbol{\mu}; \mathbf{0}] = \sqrt{\frac{p}{n}}, \tag{1.2}$$

[1] Оценка является любой измеримой функция от $(\mathbb{R}^p)^n$ до $\mathbb{R}^p$

где $\overline{\boldsymbol{Y}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{Y}_i$ является выборочным средним наблюдаемых данных. Определим

$$\|\boldsymbol{\Theta}\|_{0,2} := \sum_{i=1}^n \mathbf{1}(\|\boldsymbol{\theta}_i\|_2 > 0).$$

Базируясь на (1.2), определим избыточный риск оценки $\widehat{\boldsymbol{\mu}}$ в худшем случае следующим образом

$$\mathcal{E}(\widehat{\boldsymbol{\mu}}; n, p, \varepsilon) = \sup_{\boldsymbol{\mu}\in\mathbb{R}^p; \|\boldsymbol{\Theta}\|_{0,2}\le\varepsilon n} R[\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}; \boldsymbol{\Theta}] - \sqrt{\frac{p}{n}}.$$

Минимаксный избыточный риск определяется таким образом

$$\mathcal{E}(n, p, \varepsilon) = \inf_{\widehat{\boldsymbol{\mu}}} \mathcal{E}(\widehat{\boldsymbol{\mu}}, n, p, \varepsilon),$$

где инфимум берется по всевозможным оценкам $\widehat{\boldsymbol{\mu}}$. Заметим, что согласно определению, рассматривающиеся оценки выше могут зависеть от n, p и $\varepsilon = o/n$. Основным результатом этой статьи является то, что избыточный риск нашей оценки, введенный в следующем разделе стремится к нулю когда $\varepsilon = \varepsilon_n \to 0$ и $n \to \infty$, при таком $p = p_n$, что p_n/n ограничена сверху константой.

2. Групповой жесткий порог

В этом разделе мы определяем оценку $\widehat{\boldsymbol{\mu}}_{\mathrm{GHT}}$, называемую group hard thresholding, и доказываем, что при этой оценке избыточный риск стремится к 0, когда доля выбросов ε стремится к 0. Грубо говоря, $\widehat{\boldsymbol{\mu}}_{\mathrm{GHT}}$ является средним арифмитическим векторов $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ заминяя все вектора с большим расстоянием от покоординатной медианы ею же.

Более формально, пусть $\widehat{\boldsymbol{\mu}}_{\mathrm{Med}} := \mathtt{Med}(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ есть покоординатная медиана выборки $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$. Для фиксированного порога $\lambda > 0$ и каждого $i \in \{1, \ldots, n\}$, положим

$$\widehat{\boldsymbol{\theta}}_i = HT_\lambda(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_{\mathrm{Med}}) := (\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_{\mathrm{Med}})\mathbf{1}(\|\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_{\mathrm{Med}}\|_2 > \lambda) \tag{2.1}$$

$$\widehat{\boldsymbol{\mu}}_{\mathrm{GHT}} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{Y}_i - \widehat{\boldsymbol{\theta}}_i) := L_n(\boldsymbol{Y} - \widehat{\boldsymbol{\Theta}}). \tag{2.2}$$

Далее мы сформулируем основную теорему статьи, показывающую, что избыточный риск нашей оценки стремится к нулю, если доля выбросов $\varepsilon = \varepsilon_n$ стремится к 0 так, что $\varepsilon_n p_n^{1/4}$ также стремится к 0. Заметим, что это условие выполнено при фиксированном p, но модель также позволяет размерность быть бесконечностью, т.е. $p = p_n \to \infty$ при ограничении $\varepsilon_n p_n^{1/4} \log^{1/2} \varepsilon_n^{-1} = o(1)$ когда размер выборки n стремится к бесконечности.

Теорема 2.1. *Для* $\widehat{\boldsymbol{\mu}}_{\mathrm{GHT}}$ *определенным в* (2.2) *и* $\lambda^2 = p + 8\sqrt{p\log\varepsilon^{-1}} + 16\log\varepsilon^{-1}$ *мы имеем*

$$\overline{\lim_{n\to\infty}}\, \mathcal{E}(\widehat{\boldsymbol{\mu}}_{\mathrm{GHT}}, n, p_n, \varepsilon_n) = 0,$$

если $\varepsilon_n p_n^{1/4} \log^{1/2}\varepsilon_n^{-1} = o(1)$ *и* $p_n = O(n)$ *когда* $n \to \infty$.

Доказательство. Пусть $s_i = \mathbf{1}(\|\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_{\mathrm{Med}}\|_2 \le \lambda)$ и $\boldsymbol{\delta} = \widehat{\boldsymbol{\mu}}_{\mathrm{Med}} - \boldsymbol{\mu}^*$. Используя тот факт, что $\boldsymbol{Y} = \boldsymbol{\mu}\mathbf{1}_n^\top + \boldsymbol{\Theta} + \boldsymbol{\Xi}$, можно написать следующее соотношение

$$\begin{aligned} \widehat{\boldsymbol{\mu}}_{\mathrm{GHT}} - \boldsymbol{\mu} &= L_n(\boldsymbol{\Theta} + \boldsymbol{\Xi} - \widehat{\boldsymbol{\Theta}}) \\ &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i + T_1(n) + T_2(n) + T_3(n), \end{aligned} \tag{2.3}$$

где

$$T_1(n) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i(s_i - 1), \quad T_2(n) = \frac{\boldsymbol{\delta}}{n}\sum_{i=1}^n (1 - s_i),$$

$$T_3(n) = \frac{1}{n}\sum_{i\in\mathcal{O}} \boldsymbol{\theta}_i s_i.$$

Для простоты обозначений обозначим $\mathbb{L}_2$ нормой вектора $\boldsymbol{V}$ следующим образом

$$\|\boldsymbol{V}\|_{\mathbb{L}_2} = \big(\mathbb{E}[\|\boldsymbol{V}\|_2^2]\big)^{1/2}.$$

Заметим, что достаточно показать, что $T_i(n) \to 0$ когда $\varepsilon \to 0$ для $i \in \{1, 2, 3\}$. В самом деле,

$$\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i\right\|_{\mathbb{L}_2}^2 = \frac{p}{n}.$$

Тогда,

$$\begin{aligned} \mathcal{E}(\widehat{\boldsymbol{\mu}}; n, p, \varepsilon) &\le \|T_1(n) + T_2(n) + T_3(n)\|_{\mathbb{L}_2} \\ &\le \|T_1(n)\|_{\mathbb{L}_2} + \|T_2(n)\|_{\mathbb{L}_2} + \|T_3(n)\|_{\mathbb{L}_2}. \end{aligned}$$

Можно проверить, что следующее отношение выполнено с некоторой константой C

$$\|\boldsymbol{\delta}\|_{\mathbb{L}_4} \le C\sqrt{p}\left(\frac{1}{\sqrt{n}} \vee \varepsilon\right) = o(p^{1/4}) \tag{2.4}$$

и

$$\|\boldsymbol{\xi}_i\|_{\mathbb{L}_4}^4 = \mathbb{E}\left(\sum_{j=1}^p \xi_{ij}^2\right)^2 = 3p + p(p-1) \le (p+1)^2. \tag{2.5}$$

Сначала мы ограничим сверху $\mathbb{E}[1-s_i]$ для всех $i \in \mathcal{O}^c$. Имея (2.4), мы получаем $\|\boldsymbol{\delta}\|_2 = o_{\mathbb{P}}\big(p^{1/4}\big)$ и, следовательно,

$$\begin{aligned}\mathbb{E}[1-s_i] &= \mathbb{P}(s_i = 0) = \mathbb{P}(\|\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_{\mathrm{Med}}\|_2^2 > \lambda^2) = \mathbb{P}(\|\boldsymbol{\delta} + \boldsymbol{\xi}_i\|_2^2 > \lambda^2) \\ &\le \mathbb{P}\big(\|\boldsymbol{\xi}_i\|_2^2 > \lambda^2(1 - o(1))\big) \lesssim \varepsilon^8, \qquad \forall i \in \mathcal{O}^c,\end{aligned} \tag{2.6}$$

где последнее неравенство следует от концентрации случайной величины χ_p^2 и выбора λ^2. Для $T_1(n)$, используя (2.5), мы имеем

$$\begin{aligned}\|T_1(n)\|_{\mathbb{L}_2} &\le \Big\|\frac{1}{n}\sum_{i\in\mathcal{O}^c}\boldsymbol{\xi}_i(1-s_i)\Big\|_{\mathbb{L}_2} + \Big\|\frac{1}{n}\sum_{i\in\mathcal{O}}\boldsymbol{\xi}_i(1-s_i)\Big\|_{\mathbb{L}_2} \\ &= O(\sqrt{p}\varepsilon^2) + \Big\|\frac{1}{n}\sum_{i\in\mathcal{O}}\boldsymbol{\xi}_i(1-s_i)\Big\|_{\mathbb{L}_2}.\end{aligned}$$

С другой стороны, неравенство Коши-Шварца-Буняковского дает

$$\Big\|\sum_{i\in\mathcal{O}}\boldsymbol{\xi}_i(1-s_i)\Big\|_2^2 \le \sum_{i\in\mathcal{O}}(1-s_i)\Big\|\sum_{i\in\mathcal{O}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\Big\|_{\mathrm{op}} \le n\varepsilon\Big\|\sum_{i\in\mathcal{O}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\Big\|_{\mathrm{op}}.$$

Известная верхняя оценка для операторной нормы матрицы с нормальными элементами (см. Лемма 9 в [4]) дает

$$\Big\|\sum_{i\in\mathcal{O}}\boldsymbol{\xi}_i(1-s_i)\Big\|_{\mathbb{L}_2}^2 \le 3n\varepsilon(p + n\varepsilon + 4).$$

Следовательно,

$$\|T_1(n)\|_{\mathbb{L}_2} \lesssim \sqrt{p}\,\varepsilon^2 + \frac{\sqrt{n\varepsilon p} + n\varepsilon}{n} = \sqrt{p}\,\varepsilon^2 + \sqrt{\varepsilon p/n} + \varepsilon = o(1). \tag{2.7}$$

Для ограничения сверху $\|T_2(n)\|_{\mathbb{L}_2}$, мы используем неравенство Коши-Шварца-Буняковского вместе с (2.4) и (2.6), чтобы получить

$$\begin{aligned}\|T_2(n)\|_{\mathbb{L}_2} &= \Big\|\frac{\boldsymbol{\delta}}{n}\sum_{i=1}^n(1-s_i)\Big\|_{\mathbb{L}_2} \le \frac{\|\boldsymbol{\delta}\|_{\mathbb{L}_4}}{n}\cdot\sum_{i=1}^n\|1-s_i\|_{\mathbb{L}_4} \\ &= \frac{o(p^{1/4})}{n}(\varepsilon n + n(1-\varepsilon)\varepsilon^2) = \varepsilon o(p^{1/4}) = o(1),\end{aligned} \tag{2.8}$$

где на третьем шаге мы ограничили $\|1-s_i\|_{\mathbb{L}_4} = \mathbb{E}^{1/4}[1-s_i] \le 1$ для $i \in \mathcal{O}$ и $\|1-s_i\|_{\mathbb{L}_4} \le \varepsilon^2$ для $i \in \mathcal{O}^c$. Для получения верхней оценки $\mathbb{L}_2$ нормы $T_3(n)$ заметим, что когда $s_i = 1$, тогда у нас есть верхняя оценка для $\|\boldsymbol{\theta}_i\|_2$. Формально, неравенство $\|\boldsymbol{\theta}_i + \boldsymbol{\delta} + \boldsymbol{\xi}_i\|_2 \le \lambda$ эквивалентно $s_i = 1$, следовательно

$$\|\boldsymbol{\theta}_i\|_2 \le 2\|\boldsymbol{\delta}\|_2 + 2|\eta_i| + \big(\lambda^2 - \|\boldsymbol{\xi}_i\|_2^2\big)_+^{1/2} + (2|\boldsymbol{\delta}^\top\boldsymbol{\xi}_i|)^{1/2}, \tag{2.9}$$

для стандартных нормальный случайных величин η_i для $i \in \{1, \dots, n\}$.

Используя неравенство Гельдера можно показать, что

$$\sum_{i\in\mathcal{O}} |\boldsymbol{\delta}^\top \boldsymbol{\xi}_i|^{1/2} \le (n\varepsilon)^{3/4} \Big\{ \sum_{i\in\mathcal{O}} |\boldsymbol{\delta}^\top \boldsymbol{\xi}_i|^2 \Big\}^{1/4} = (n\varepsilon)^{3/4} \Big\{ \boldsymbol{\delta}^\top \sum_{i\in\mathcal{O}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top \boldsymbol{\delta} \Big\}^{1/4}$$
$$\le (n\varepsilon)^{3/4} \|\boldsymbol{\delta}\|_2^{1/2} \|\boldsymbol{\xi}_O\|_{\mathrm{op}}^{1/2},$$

где $\|\boldsymbol{\xi}_O\|_{\mathrm{op}}^{1/2}$ спектральная норма матрицы полученной из векторов $\boldsymbol{\xi}_i$ для $i \in \mathcal{O}$. Отсюда следует, что[2]

$$\Big\| \sum_{i\in\mathcal{O}} |\boldsymbol{\delta}^\top \boldsymbol{\xi}_i|^{1/2} \Big\|_{\mathbb{L}_2} \le (n\varepsilon)^{3/4} \|\boldsymbol{\delta}\|_{\mathbb{L}_2}^{1/2} \|\boldsymbol{\xi}_O\|_{\mathbb{L}_2}^{1/2}$$
$$= O\Big((n\varepsilon)^{3/4} (\varepsilon^{1/2} p^{1/4}) ((n\varepsilon)^{1/4} + p^{1/4}) \Big)$$
$$= O\Big(n\varepsilon p^{1/4} + (n\varepsilon)^{3/4} \sqrt{\varepsilon p} \Big).$$

Можно показать, что

$$\mathbb{E}[(\lambda^2 - \|\boldsymbol{\xi}_i\|_2^2)_+]) \lesssim \sqrt{p \log \varepsilon^{-1}} + \log \varepsilon^{-1} \lesssim \sqrt{p}\, \log \varepsilon^{-1}$$

Используя неравенство треугольника для $\mathbb{L}_2$ нормы, получаем

$$\|T_3(n)\|_{\mathbb{L}_2} \le \frac{1}{n} \sum_{i\in\mathcal{O}} \|\boldsymbol{\theta}_i \mathbf{1}(s_i = 1)\|_{\mathbb{L}_2}$$
$$\lesssim \varepsilon(\|\boldsymbol{\delta}\|_{\mathbb{L}_2} + 1) + \varepsilon p^{1/4} \log^{1/2} \varepsilon^{-1} + \varepsilon p^{1/4} + \varepsilon^{5/4} p^{1/2} n^{-1/4}$$
$$\lesssim \varepsilon p^{1/4} \log^{1/2} \varepsilon^{-1} + \varepsilon p^{1/4} + \varepsilon^{5/4} p^{1/4}.$$

Поскольку $\varepsilon_n = o(1)$ и $\varepsilon_n p_n^{1/4} \log^{1/2} \varepsilon_n^{-1} = o(1)$, получаем, что $\|T_3(n)\|_{\mathbb{L}_2} = o(1)$ и доказательство теоремы следует. □

Abstract. In this work we introduce the notion of the excess risk in the setup of estimation of the Gaussian mean when the observations are corrupted by outliers. It is known that the sample mean loses its good properties in the presence of outliers [5, 6]. In addition, even the sample median is not minimax-rate-optimal in the multivariate setting. The optimal rate of the minimax risk in this setting was established by [1]. However, even these minimax-rate-optimality results do not quantify how fast the risk in the contaminated model approaches the risk in the uncontaminated model when the rate of contamination goes to zero. The present paper does a first step in filling this gap by showing that the group hard thresholding estimator has an excess risk that goes to zero when the corruption rate approaches zero.

[2]Для простоты мы рассматриваем случай когда $n^{-1/2} \lesssim \varepsilon_n$.

Список литературы

[1] M. Chen, C. Gao and Z. Ren, "A general decision theory for Huber's ε-contamination model", Electron. J. Statist., 10:3752–3774 (2016).
[2] M. Chen, C. Gao and Z. Ren, "Robust covariance and scatter matrix estimation under Huber's contamination model", Annals of Statistics, **46**(5), 1932 – 1960 (2018).
[3] Y. Cheng, I. Diakonikolas and R. Ge, "High-dimensional robust mean estimation in nearly-optimal time", arXiv:1811.09380 (2018).
[4] O. Collier and A. S. Dalalyan, "Rate-optimal estimation of p-dimensional linear functionals in a sparse gaussian model", Electron. J. Statist., **13**(2), 2830 – 2864 (2019).
[5] P. J. Huber, "Robust estimation of a location parameter", The annals of mathematical statistics, **35**(1), 73 – 101 (1964).
[6] P. J. Huber, "A robust version of the probability ratio test", The annals of mathematical statistics, **36** (6), no. 1, 753 – 758 (1965).

Поступила 14 ноября 2019

После доработки 19 января 2020

Принята к публикации 06 февраля 2020