Версия, представленная в журнал

©2004 г. Г. Бушар,
(INRIA Rhône-Alpes, Гренобль, Франция)
С. Жирар, д-р философии,
А. Юдицкий, д-р философии,
(Университет Гренобль I, Франция)
А. В. Назин, д-р физ.-мат. наук
(Институт проблем управления им. В. А. Трапезникова РАН, Москва)

НЕПАРАМЕТРИЧЕСКОЕ ОЦЕНИВАНИЕ ГРАНИЦЫ НОСИТЕЛЯ ПОСРЕДСТВОМ ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ [1]

Предложен новый метод оценивания границы множества точек. Оценка определяется как ядерная функция, покрывающая все точки выборки и порождающая носитель минимальной поверхности. Она представляется линейной комбинацией ядерных функций, центрированных в точках выборки. Веса линейной комбинации вычисляются посредством решения задачи линейного программирования. В общем случае решение оптимизационной задачи разреженное, то есть содержит лишь небольшое число ненулевых коэффициентов. Соответствующие точки играют роль опорных векторов в статистической теории обучения. Показано, что ошибки оценивания в смысле L_1-нормы сходится к нулю с вероятностью 1, и получена оценка скорости сходимости.

1. Введение

Проблема оценивания множества S по конечной случайной выборке его точек широко представлена в литературе. Так, задача оценивания границы или самого носителя возникает в классификации [1], в кластерном и дискриминантном анализе [2, 3] и при детектировании больших выбросов. Приложения можно также найти в медицинской диагностике [4] и в области мониторинга машин [5]. При анализе изображений задача сегментации может рассматриваться с позиции оценки множества точек, выпуклого ограниченного множества в R^2 [6]. Укажем также на некоторые приложения в эконометрике [7]. В таких случаях неизвестный носитель может быть записан в виде

$$S \triangleq \{(x, y) \ : \ 0 \le x \le 1, \ 0 \le y \le f(x)\}, \tag{1}$$

где $f : [0, 1] \to (0, +\infty)$ — неизвестная функция. Таким образом, задача сводится к оцениванию f, называемой граничной функцией (см., например, [8]). Данные состоят из пар (X, Y), где X представляет собой производственные затраты (количество затраченного труда, энергии или капитала), порождающие доход Y данной фирмы. При таком

[1] Рекомендована к печати Программным комитетом Второй международной конференции по проблемам управления (Москва, Институт проблем управления, 17–19 июня 2003 г.).

рассмотрении значение $f(x)$ может интерпретироваться как максимальный уровень дохода, достигаемого при затратах x. В [9] функция $f(x)$ предполагалась возрастающей и вогнутой в соответствии с экономической трактовкой и предложен адаптивный метод оценивания, получивший название DEA (Data Envelopment Analysis) и состоящий в нахождении наименьшей вогнутой монотонно возрастающей функции, покрывающей все точки выборки. Поэтому эта оценка кусочно-линейна и, насколько нам известно, эта первая оценка границы, вычисляемая методом линейного программирования [10]. Ее асимптотическое распределение установлено в [11].

Ранняя статья Геффроя [12] посвящена независимым одинаково распределенным наблюдениям с плотностью ϕ, в которой предложена оценка гистограммного типа, основанная на экстремальных значениях выборки. Впоследствии эта работа была обобщена во многих направлениях.

С другой стороны, были предложены кусочно-полиномиальные оценки, которые определяются локально на данном слое как наименьший полином фиксированной степени, покрывающий все точки этого участка. Их оптимальность в асимптотически минимаксном смысле доказана в [6, 13] при слабых предположениях о скорости убывания α плотности ϕ к нулю. Методы экстремальных значений предлагались затем в [14, 15] для оценки параметра α.

Кроме того, были разработаны различные сглаженные оценки Геффроя для случая пуассоновского точечного процесса. Жирар и Жакоб [16] ввели оценку, основанную на ядерных регрессиях и методе ортогонального разложения [17, 18]. В том же ключе Жардес предложил в [19] оценку Фабера–Шаудера. Жирар и Меннето [20] разработали общий подход к изучению оценок этого типа и обобщили их на носители вида

$$S = \{(x, y) \ : \ x \in E,\ 0 \le y \le f(x)\}\,,$$

где $f : E \to (0, +\infty)$ — неизвестная функция, а E — произвольное множество. В каждом случае получено предельное распределение оценок.

Следует также упомянуть работы Аббара [21] и Жакоба и Суке [22], которые использовали аналогичный метод сглаживания, хотя их оценки не основаны на экстремальных значения пуассоновского процесса. В настоящей статье предложенная оценка может рассматриваться как принадлежащая обоим указанным направлениям. Она определяется как ядерная оценка, полученная посредством сглаживания некоторого набора точек выборки. Эти точки выбираются автоматически решением задачи линейного программирования с целью получения оценки носителя, которая покрывает все эти точки и имеет наименьшую поверхность. Она имеет следующие преимущества: вычисления могут производиться с помощью стандартных алгоритмов оптимизации (см., например, гл. 4 в [23]), ее гладкость напрямую связана с гладкостью выбранного ядра и обладает полезными теоретическими свойствами. Например, мы доказываем, что она сходится почти наверное в L_1-норме. Оценка определяется в разделе 2. Ее теоретические свойства устанавливаются в разделе 3. Поведение этой оценки представлено в [24] на примерах с конечной выборкой. Проводится сравнение с аналогичным утверждением из [25]. Доказательства можно найти в [24].

2. Оценивание границы

Пусть все случайные величины определены на вероятностном пространстве $(\Omega, \mathcal{F}, P)$. Задача состоит в нахождении оценки неизвестной положительной функции $f : [0,1] \to (0, +\infty)$ по наблюдениям $Z_N = (X_i, Y_i)_{i=1,\ldots,N}$, представляющим собой последовательность независимых равномерно распределенных на S пар (X_i, Y_i), где множество S имеет вид (1). Для простоты, далее будем рассматривать функции f, определенные на всей числовой оси R, полагая $f(x) = 0$ для всех $x \notin [0,1]$. Введем

$$C_f \triangleq \int_0^1 f(u)\,du = \int_R f(u)\,du\,.$$

Случайные величины X_i распределены на отрезке $[0,\ 1]$ с плотностью $f(\cdot)/C_f$, а Y_i имеет равномерное условное распределение относительно X_i на отрезке $[0, f(X_i)]$. Рассматриваемая далее оценка выбирается из следующего семейства функций:

$$\text{(2)} \qquad \begin{cases} \widehat{f}_N(x) = \sum_{i=1}^N \alpha_i K_h(x - X_i)\,, & K_h(t) = h^{-1}K(t/h)\,, \\ \alpha_i \geq 0, \quad i = 1, \ldots, N, & h > 0\,, \end{cases}$$

где $K(\cdot)$ — нормированная ядерная функция $K : R \to [0, +\infty)$, т.е.

$$\int K(u)\,du = 1\,,$$

а параметр h — ширина окна. Каждый множитель α_i представляет важность точки X_i, Y_i в оценке $\widehat{f}_N$. В частности, если $\alpha_i \neq 0$, соответствующую точку X_i, Y_i можно называть опорным вектором по аналогии с Support Vector Mashines (SVM). Обзор по данной тематике см. в [26], а пример применения SVM для квантильного оценивания — в [27, гл. 8]. Ограничение $\alpha_i \geq 0$ для всех $i = 1, \ldots, N$ обеспечивает положительность оценки $\widehat{f}_N(x) \geq 0$ для всех $x \in R$ и предотвращает нерегулярность оценки. Заметим, что площадь оцененного носителя равна

$$\text{(3)} \qquad \int_R \widehat{f}_N(x)\,dx = \sum_{i=1}^N \alpha_i\,.$$

Таким образом приходим к определению вектора параметров $\alpha = (\alpha_1, \ldots, \alpha_N)^T$ как решения следующей задачи линейного программирования:

$$\text{(4)} \qquad J_P^* \triangleq \min_\alpha \mathbf{1}^T\alpha$$

$$\text{(5)} \qquad A\alpha \geq Y,$$

$$\text{(6)} \qquad \alpha \geq 0\,.$$

Здесь

$$\begin{aligned} \mathbf{1} &\triangleq (1, 1, \ldots, 1)^T \in R^N, \\ A &\triangleq \|K_h(X_i - X_j)\|_{i,j=1,\ldots,N}\,, \\ Y &\triangleq (Y_1, \ldots, Y_N)^T. \end{aligned}$$

Следовательно, $A\alpha = \left(\widehat{f}_N(X_1), \ldots, \widehat{f}_N(X_N)\right)^T$, и векторное ограничение (5) означает, что

$$\widehat{f}_N(X_i) \geq Y_i\,, \qquad i = 1, \ldots, N\,. \tag{7}$$

Другими словами, $\widehat{f}_N$ определяет ядерную оценку носителя, покрывающую все точки и имеющую наименьшую поверхность. На практике решение задачи линейного программирования разреженное в том смысле, что число ненулевых коэффициентов $n(\alpha) \triangleq \#\{\alpha_i \neq 0\}$ мало (для умеренных значений значений h) и, таким образом, получающаяся оценка вычисляется быстро даже при больших выборках.

Заметим, что описанную выше оценку (2)–(6) можно было бы получить из метода максимального правдоподобия с использованием аппроксимирующего семейства (2). Действительно, совместное распределение наблюдений Z_N при заданной функции f имеет функцию плотности следующего вида:

$$p(Z_N \mid f) = \prod_{i=1}^{N} \frac{f(X_i)}{C_f} \cdot \frac{1}{f(X_i)} \mathbf{1}\{0 \leq Y_i \leq f(X_i)\}\,, \tag{8}$$

где $I\{\cdot\}$ — индикаторная функция. Более того,

$$C_f\Big|_{f=\widehat{f}_N} = \sum_{i=1}^{N} \alpha_i\,, \tag{9}$$

и, следовательно, логарифмическая функция правдоподобия равна

$$L(\alpha) \triangleq \log p(Z_N \mid \widehat{f}_N) = -N \log \sum_{i=1}^{N} \alpha_i + \sum_{i=1}^{N} \log \mathbf{1}\{Y_i \leq \widehat{f}_N(X_i)\}\,, \tag{10}$$

и ее максимизация на множестве неотрицательных параметров α эквивалентна задаче (4)–(6).

Отметим, что уже предлагались другие подходы к оцениванию α. Жирар и Меннето в [20] рассматривают разбиение интервалами $\{I_r : 1 \leq r \leq k\}$ отрезка $[0, 1]$ при $k \to \infty$. Для каждого $r = 1, \ldots, k$ они вводят $D_r = \{(x, y) : x \in I_r,\ 0 \leq y \leq f(x)\}$ — слой множества S, построенный на I_r, — определяют $Y_r^* = \max\{Y_i : (X_i, Y_i) \in D_r\}$ и рассматривают оценки

$$\widehat{\alpha}_i = \begin{cases} \lambda(I_r) Y_r^*, & \text{если } \exists\, r \in \{1, \ldots, k\} : Y_i = Y_r^*, \\ 0 & \text{иначе,} \end{cases}$$

где λ — мера Лебега. Предложенная ими оценка границы имеет вид

$$\check{f}_N(x) = \sum_{r=1}^{k} K_h(x - x_r) \lambda(I_r) Y_r^*,$$

где x_r — центр интервала I_r. Этот подход сталкивается с практической трудностью выбора k и конкретного разбиения отрезка $[0, 1]$. В нашем же подходе решение задачи линейного программирования непосредственно приводит к опорным векторам. В

этом смысле, предложенная в [25] оценка аналогична $\widehat{f}_N$. Она определяется Фурье-разложением

$$\widehat{g}_N(x) = c_0 + \sum_{k=1}^{M} a_k \cos(2\pi kx) + \sum_{k=1}^{M} b_k \sin(2\pi kx), \tag{11}$$

где вектор параметров $\beta = (c_0, a_1, \ldots, a_M, b_1, \ldots, b_M)^T$ задается решением следующей задачи линейного программирования:

$$\min c_0 \quad \left(= \int_0^1 \widehat{g}_N(x)dx \right) \tag{12}$$

при ограничениях

$$\widehat{g}_N(X_i) \;\geq\; Y_i, \quad i = 1, \ldots, N \tag{13}$$

$$\sum_{k=1}^{M} k\ (|a_k| + |b_k|) \;\leq\; \frac{L}{2\pi}\,. \tag{14}$$

Следовательно, $\widehat{g}_N$ определяет Фурье-оценку носителя, которая покрывает все точки (уравнение (13)), липшицева с константой L (уравнение (14)) и имеет наименьшую поверхность (уравнение (12)). С теоретической точки зрения, эта оценка обладает свойством минимаксной оптимальности.

3. Основные результаты

В этом разделе устанавливается сходимость почти наверное оценки $\widehat{f}_N$ для L_1 нормы на отрезке $[0, 1]$. В этой связи, введем следующие основные предположения на неизвестную граничную функцию:

A1. $0 < f_{\min} \leq f(x) \leq f_{\max} < \infty$ для всех $x \in [0, 1]$,

A2. $|f(x) - f(y)| \leq L_f|x - y|$ для всех $x, y \in [0, 1]$; $\quad L_f < \infty$.

Рассмотрим следующие предположения о свойствах ядерной функции:

B1. $K(t) = K(-t) \geq 0$,

B2. $\int_R K(t)\,dt = 1$,

B3. $|K(s) - K(t)| \leq L_K|s - t|,\ L_K < \infty$,

B4. $C_0(K) \triangleq \int_R K^2(t)\,dt < \infty$ и $C_2(K) \triangleq \int_R t^2\,K(t)\,dt < \infty$.

Ниже, запись вида "$a_N \asymp b_N$" означает "асимптотическую эквивалентность" двух последовательностей $\{a_N\}$ и $\{b_N\}$ положительных чисел, т.е. $0 < \liminf_{N\to\infty} a_N/b_N \leq \limsup_{N\to\infty} a_N/b_N < +\infty$.

Теорема 1 *Пусть* $h \to 0$ *и* $\log N/(Nh^2) \to 0$ *при* $N \to \infty$. *Пусть выполнены сформулированные выше предположения A и B. Тогда оценка (2)–(6) имеет следующие асимптотические свойства:*

$$\limsup_{N\to\infty} \varepsilon_1^{-1}(N)\|\hat{f}_N - f\|_1 \le C(\omega) < \infty \quad п.н., \tag{15}$$

$$\varepsilon_1(N) \triangleq \max\left\{h, \sqrt{\log N/(Nh^2)}\right\}. \tag{16}$$

Следствие 1 *Гарантируемая теоремой 1 максимальная скорость сходимости*

$$\|\hat{f}_N - f\|_1 = O\left((\log N/N)^{1/4}\right)$$

достигается при

$$h \asymp (\log N/N)^{1/4}. \tag{17}$$

Эта скорость сходимости может быть улучшена ценой некоторой модификации оценки. С этой целью введем дополнительное ограничение с тем, чтобы обеспечить для каждого коэффициента α_i порядок $1/N$. Обратная сторона этой модификации состоит в том, что новая оценка $\widetilde{f}_N$ будет использовать большее число опорных векторов, чем $\widehat{f}_N$.

Определим модифицированную оценку следующим образом:

$$\tilde{f}_N(x) = \sum_{i=1}^{N} K_h(x - X_i)\alpha_i \tag{18}$$

где вектор $\alpha = (\alpha_1, \ldots, \alpha_N)^T$ задается решением модифицированной задачи линейного программирования

$$J^*_{MP} \triangleq \min_{\alpha} \mathbf{1}^T\alpha \tag{19}$$

$$A\alpha \ge Y, \tag{20}$$

$$0 \le \alpha \le C_\alpha/N, \tag{21}$$

где

$$C_\alpha > f_{\max}. \tag{22}$$

Замечание. Фактически, нужно обеспечить $C_\alpha > C_f$, что гарантируется неравенством (22).

Модифицированная оценка (18)–(22) отличается от оценки (2)–(6) дополнительным ограничением сверху на каждый коэффициент α_i, см. (21).

Теорема 2 *Пусть* $h \to 0$ *и* $\log N/(Nh) \to 0$ *при* $N \to \infty$. *Пусть ядерная функция* $K(\cdot)$ *имеет конечный носитель, т.е.* $K(t) = 0\ \forall\, |t| \ge 1$, *и выполняются предположения A и B. Тогда оценка (18)–(22) имеет следующие асимптотические свойства:*

$$\limsup_{N\to\infty} \varepsilon_2^{-1}(N)\|\tilde{f}_N - f\|_1 \le C(\omega) < \infty \quad п.н., \tag{23}$$

$$\varepsilon_2(N) \triangleq \max\left\{h, \sqrt{\log N/(Nh)}\right\}. \tag{24}$$

Следствие 2 *Гарантируемая теоремой 2 максимальная скорость сходимости*

$$\|\tilde{f}_N - f\|_1 = O\left((\log N/N)^{1/3}\right)$$

достигается при

$$h \asymp (\log N/N)^{1/3}. \tag{25}$$

4. Заключение

Приведенные выше теоретические результаты показывают, что при использовании ядерных оценок (2)–(6) и (18)–(22) за счет "правильного" выбора ширины окна можно обеспечить скорость сходимости ошибки оценивания в L_1-норме, близкую по порядку к $O\left(N^{-1/4}\right)$ и $O\left(N^{-1/3}\right)$ соответственно. Вычисление этих оценок сводится к использованию процедуры линейного программирования, а получающиеся в результате решения довольно разрежены, т.е. учитывают сравнительно небольшую долю опорных точек. Таким образом, некоторая потеря в скорости сходимости по сравнению с оптимальной, равной для рассматриваемого класса липшицевых граничных функций $O\left(N^{-1/2}\right)$ [25], объясняется значительным уменьшением вычислительной сложности получаемых оценок.

СПИСОК ЛИТЕРАТУРЫ

1. *Hardy A., Rasson J.P.* Une nouvelle approche des problèmes de classification automatique// Statistique et Analyse des données. 1982. V. 7. P. 41–56.

2. *Hartigan J.A.* Clustering Algorithms. Wiley, Chichester, 1975.

3. *Baufays P., Rasson J.P.* A new geometric discriminant rule// *Computational Statistics Quaterly.* 1985. V. 2. P. 15–30.

4. *Tarassenko L., Hayton P., Cerneaz N. and Brady M.* Novelty detection for the identification of masses in mammograms// Proc. fourth IEE International Conference on Artificial Neural Networks. Cambridge, 1995. P. 442–447.

5. Devroye L.P., Wise G.L. Detection of abnormal behavior via non parametric estimation of the support// SIAM J. Applied Math. 1980. V. 38. P. 448–480.

6. *Korostelev A.P., Tsybakov A.B.* Minimax theory of image reconstruction// Lecture Notes in Statistics. Springer-Verlag, New York, 1993. V. 82.

7. *Deprins D., Simar L. and Tulkens H.* Measuring Labor Efficiency in Post Offices// The Performance of Public Enterprises: Concepts and Measurements Ed. by M. Marchand, P. Pestieau and H. Tulkens, North Holland ed., Amsterdam, 1984.

8. *Härdle W., Hall P. and Simar L.* Iterated bootstrap with application to frontier models// J. Productivity Anal. 1995. V. 6. P. 63–76.

9. *Korostelev A., Simar L. and Tsybakov A. B.* Efficient estimation of monotone boundaries// The Annals of Statistics. 1995. V. 23. P. 476–489.

10. *Charnes A., Cooper W.W. and Rhodes E.* Measuring the inefficiency of decision making units// European Journal of Operational Research. 1978. V. 2. P. 429–444.

11. *Gijbels I., Mammen E., Park B.U. and Simar L.* On estimation of monotone and concave frontier functions// Journal of the American Statistical Association. 1999. V. 94. P. 220–228.

12. *Geffroy J.* Sur un problème d'estimation géométrique// Publications de l'Institut de Statistique de l'Université de Paris. 1964. V. XIII. P. 191–200.

13. *Härdle W., Park B. U. and Tsybakov A. B.* Estimation of a non sharp support boundaries// J. Multivariate Analysis. 1995. V. 43. P. 205–218.

14. *Hall P., Nussbaum M. and Stern S.E.* On the estimation of a support curve of indeterminate sharpness// J. Multivariate Analysis. 1997. V. 62. P. 204–232.

15. *Gijbels I. and Peng L.* Estimation of a support curve via order statistics// Discussion Paper **9905**, Institut de Statistique, Université Catholique de Louvain, 1999.

16. *Girard S. and Jacob P.* Extreme values and kernel estimates of point processes boundaries// Technical report ENSAM-INRA-UM2, **01-02**. 2001.

17. *Girard S. and Jacob P.* Extreme values and Haar series estimates of point processes boundaries// Scandinavian Journal of Statistics. 2002. V. 30. P. 369–384.

18. *Girard S. and Jacob P.* Projection estimates of point processes boundaries// Journal of Statistical Planning and Inference. 2003. V. 116. *№* 1. P. 1–15.

19. *Gardes L.* Estimating the support of a Poisson process via the Faber-Shauder basis and extreme values// Publications de l'Institut de Statistique de l'Université de Paris. 2002. V. XXXXVI. P. 43–72.

20. *Girard S. and Menneteau L.* Cental limit theorems for smoothed extreme values estimates of point processes boundaries// Journal of Statistical Planning and Inference. 2003.

21. *Abbar H.* Un estimateur spline du contour d'une répartition ponctuelle aléatoire// Statistique et analyse des données. 1990. V. 15. *№* 3. P. 1–19.

22. *Jacob P. and Suquet P.* Estimating the edge of a Poisson process by orthogonal series// Journal of Statistical Planning and Inference. 1995. V. 46. P. 215–234.

23. *Bonnans F., Gilbert J.C., Lemaréchal C. and Sagastizábal, C.* Optimisation numérique. Aspects théoriques et pratiques// Mathématiques & Applications. V. 27. Springer, Paris, 1997.

24. *Bouchard G., Girard S., Iouditski A. and Nazin A.* Linear programming problems for frontier estimation// Technical report INRIA, RR-4717, 2003.

25. *Barron A.R., Birgé L. and Massart P.* Risk Bounds for model selection via penalization// Probab. Theory Relat. Fields. 1999. V. 113. P. 301–413.

26. *Cristianini N. and Shawe-Taylor J.* An introduction to support vector machines. Cambridge University Press, 2000.

27. *Schölkopf B. and Smola A.* Learning with kernels. MIT University Press, Cambridge, 2002.