数理解析研究所講究録
第 380 巻 1980 年 32-43

密度関数の推定について

東工大 理　柴田 里程
1980.1.24　於京大数解研

Introduction

確率密度関数 $f(x)$ の推定については，多くの論文があるが，その推定法は，大きく分けて2通りになる (Rosenblatt 1971 (総合報告), Leonard 1978).

1) Kernel estimate (Rosenblatt 1956, Whittle 1958, Parzen 1962, Bartlett 1963, Loftsgaarden & Quesenberry 1965, Craswell 1965, Epanechnikov 1969, Moore & Henrichon 1969, Shorack 1969, Rosenblatt 1975, Silverman 1976, 1978a, 1978b, Taylor & Cheng 1978)

Sample $X_1, \cdots, X_n$ が与えられたとき，一般的な Kernel estimate は

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_n(X_i, x)$$

で定義されるが，ほとんどの場合 K_n は

$$K_n(X_i, x) = \frac{1}{b(n)} K\left(\frac{x - X_i}{b(n)}\right)$$

にとられる。ここで

$b(n)$; *band width* , $b(n) \to 0$

$K(x)$; *kernel function* , $K(x) \geqq 0$ 。

これは Histogram

$$\hat{f}_n(x) = \sum_j \frac{1}{\mu(A_j)} \hat{p}_j \chi_{A_j}(x)$$

の一般化である。

2) Orthogonal series estimate (Schwartz 1967, Kronmal & Tarter 1968, Watson 1969, Brunk 1978, Ahmad 1979)

$\{\phi_i\}$: orthogonal functions (Hermite, Fourier etc.)

(A) $\hat{f}_n(x) = \sum_{j=1}^{q(n)} \hat{a}_{jn} \phi_j(x)$

$$\hat{a}_{jn} = \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i) ,$$

$q(n)$: trunction point , $q(n) \to \infty$

または

(B) $\hat{f}_n(x) = \sum_{j=0}^{\infty} \lambda_j(n) \hat{a}_{jn} \phi_j(x)$ (Watson 1969)

$\lambda_j(n)$: weighting function (c.f. window)

しかし，これだと一般に $\hat{f}_n(x) \geqq 0$ とならないので

(c) $\hat{f}_n(x) = \exp\left(\sum_{j=1}^{q(n)} \hat{\beta}_{jn}\phi_j(x)\right) / M(\hat{\beta})$

$$\frac{\partial}{\partial \beta_j} \log M(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\phi_j(X_i), \quad j=1,\cdots,q(n)$$

$$M(\beta) = \int \exp\left(\sum_{j=1}^{q(n)} \beta_j \phi_j(x)\right) dx$$

なる形も提案されている（Leonard 1978）。

Loss function としては，Maximum squared error

$$\sup_x \left(\hat{f}_n(x) - f(x)\right)^2$$

または，Integrated squared error

$$\int \left(\hat{f}_n(x) - f(x)\right)^2 dx$$

を採用して，$f(x)$，$b(n)$，$K(x)$ または $q(n)$ について，様々な条件をつけて，consistency，strong consistency あるいは asymptotic normality を論じているのがほとんどである。しかし，1) に関しては Epanechnikov (1969) が，漸近的に optimal な $K(x)$ と $b(n)$ を求めている。Loss function として Integrated squared error ととれば

$$0 \leqq K(x) < c < \infty, \quad K(x) = K(-x), \quad \int_{-\infty}^{\infty} K(x)dx = 1,$$

$$\int_{-\infty}^{\infty} x^2 K(x)\,dx = 1\,,\quad \int_{-\infty}^{\infty} x^m K(x)\,dx < \infty\,,\quad 0 \leqq m < \infty$$

の条件のもとで

$$E\int(\hat{f}_n(x) - f(x))^2 dx \sim \frac{L}{nb(n)} + \frac{1}{4} b(n)^4 M$$

$$L = \int_{-\infty}^{\infty} K^2(y)\,dy\,,\quad M = \int_{-\infty}^{\infty} \left(\frac{\partial^2 f(x)}{\partial x^2}\right)^2 dx$$

となるので

$$K(x) = \begin{cases} \dfrac{3}{4\sqrt{5}} - \dfrac{3x^2}{20\sqrt{5}} & |x| \leqq 5 \\ 0 & |x| > 5 \end{cases}$$

$$b(n) \sim \left(\frac{L}{nM}\right)^{\frac{1}{5}}$$

が一つの optimal な，kernel と band width になる。しかし，M は一般に未知な量であるので，実際には適用できない。また，$K(x)$ の形は，結果にそれほど sensitive ではなく，むしろ $b(n)$ のとり方が問題である（Silverman 1978）。2) に関しては，これまでに consistency だけしか議論されていないが，勿論，$q(n)$ のとり方が大きな問題である。ここでは，これらの $b(n)$，$q(n)$ の選択を中心に話をすすめる。Loss function としては，Kullback-Leibler Information number

$$I_n(f, \hat{f}_n) = n\int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{\hat{f}_n(x)}\, dx$$

を採用する。squared error や integrated squared error よりも山や谷に対する追従性はよく表わされる。また，これを離散化したものは漸近的に

$$\sum_{i=1}^{k} \frac{n(\hat{p}_i - p_i)^2}{p_i}$$

であり，χ^2 適合度検定の統計量と一致する。

Histogram

簡単のために，$f(x)$，$0 \leqq x \leqq 1$ は continuous，bounded and bounded away from zero であるとする。各 k に対応して $[0, 1]$ の分割

$$A_{ik} = \left[\frac{i-1}{k}, \frac{i}{k}\right),\ 1 \leqq i \leqq k-1,\ A_{kk} = \left[\frac{k-1}{k}, 1\right]$$

を考えると，Histogram は A_{ik} に入ったSample 数を n_i とすれば

$$\hat{f}_{n,k}(x) = k\sum_{i=1}^{k} \hat{p}_{ik}\, \chi_{A_i}(x),\quad \hat{p}_{ik} = \frac{n_i}{n}$$

で与えられる。これはモデル

$$f_k(x) = k\sum_{i=1}^{k} p_i\, \chi_{A_i}(x)$$

のもとでのパラメーター $p_1, \cdots, p_k$ の制限つき（$\sum_i p_i = 1$, $p_i \geqq 0$）m.l.e.である。

$\frac{1}{k}$ が 1) の $b(n)$ に対応する。このとき Loss は

$$I_n(f, \hat{f}_{n,k}) = I_n(\mathbb{p}_k, \hat{\mathbb{p}}_k) + H_n(\mathbb{p}_k) - H_n(f)$$

となる。ただし，

$$\mathbb{p}_k = (p_{1k}, p_{2k}, \cdots, p_{kk})',$$

$$\hat{\mathbb{p}}_k = (\hat{p}_{1k}, \hat{p}_{2k}, \cdots, \hat{p}_{kk})',$$

$$H_n(\mathbb{p}_k) = n\left(-\sum_{i=1}^{k} p_{ik} \log p_{ik} - \log k\right),$$

$$H_n(f) = -n\int_{-\infty}^{\infty} f \log f \, dx,$$

$$p_{ik} = \int_{A_{ik}} f(x)\, dx.$$

さて，

$$L_n(k) = H_n(\mathbb{p}_k) - H_n(f) + \frac{k-1}{2}$$

とおき，$K_n \to \infty$ なる sequence に対して

$$k_n^* : L_n(k_n^*) = \min_{1 \leqq k < K_n} L_n(k)$$

とすれば

Lemma. $K_n = o(n^{\frac{3}{4}})$, $k_n^* \to \infty$ ならば

$$p\text{-}\lim_{n\to\infty} \max_{1\leqq k\leqq K_n} \left| \frac{I_n(p_k, \hat{p}_k) - \frac{k-1}{2}}{L_n(k)} \right| = 0 .$$

proof) $\left\{\frac{p_{ik}}{p_{jk}}\right\}$ が bounded であることより

$$E\left(\frac{n\sum_{i=1}^{k} p_{ik}\left(\frac{\hat{p}_{ik}-p_{ik}}{p_{ik}}\right)^2 - (k-1)}{L_n(k)} \right)^4$$

$$\leqq \frac{48k + 12k^2 + k^4 \cdot O\left(\left(\frac{k}{n}\right)^3\right)}{L_n(k)^4} .$$

ここで，

$$\sum_{k=1}^{K_n} \frac{k^2}{(2L_n(k))^4} \leqq \frac{k_n^*}{(2L_n(k_n^*))^2} + \sum_{k=k_n^*+1}^{K_n} \frac{1}{k^2}$$

であることに注意すれば $I(p_k, \hat{p}_k)$ の展開より，結果を得る。』

この Lemma より $I_n(f, \hat{f}_{n,k})$ は $L_n(k)$ と漸近的に同等であることがわかる。したがって，任意の $1\leqq \tilde{k} \leqq K_n$ について

$$\lim_{n\to\infty} P\left(\frac{I_n(f, \hat{f}_{n,\tilde{k}})}{L_n(k_n^*)} \geqq 1-\varepsilon \right) = 1$$

となり，$L_n(k_n^*)$ が Loss の漸近的な下限を与える。また，

$$L_n(k) \sim \frac{n}{8k^2}\int\left(\frac{f'}{f}\right)^2 f\,dx + \frac{k-1}{2}$$

であるので，$k_n^* \sim n^{\frac{1}{3}}$ であることもわかる。Epanechnikov

の結果と比較すると，Kernel estimate ならば対応するLoss は $O(n^{\frac{1}{5}})$ で，$L_n(k_n^*) = O(n^{\frac{1}{3}})$ であるので，もちろん Kernel estimate の方が優れている．k の選択と同じ様にMの選択を考えれば，同様の展開ができると思われる．

○ *Selection of k*

$$S_n(k) = nH(\hat{p}_k) + (k-1)$$

を最小にする $\hat{k}$ が漸近的に下限 $L_n(k_n^*)$ を attain することを示そう。$S_n(k)$ は

$$S_n(k) = L_n(k) + \left\{\frac{k-1}{2} - I_n(\hat{p}_k, p_k)\right\} + n\left\{\sum_{i=1}^{k}(p_{ik} - \hat{p}_{ik})\log p_{ik}\right\} + H(f)$$

と書きかえられ，右辺第2項は $L_n(k)$ に比べて一様に small order であることが，Lemma と同じように示せる。問題は第3，第4項だが，$S_n(k) - S_n(k_n^*)$ が $L_n(k)$ より一様に small order ならば

$$p\text{-}\lim_{n\to\infty} \frac{L_n(\hat{k})}{L_n(k_n^*)} = 1$$

したがって，Lemma より

$$p\text{-}\lim_{n\to\infty} \frac{I_n(f, \hat{f}_{n,\hat{k}})}{L_n(k_n^*)} = 1$$

であることがいえる。そのためには

$$n\left|\sum_{i=1}^{k}(p_{ik}-\hat{p}_{ik})\log p_{ik}-\sum_{i=1}^{k_n^*}(p_{ik_n^*}-\hat{p}_{ik_n^*})\log p_{ik_n^*}\right|$$

において，p_{ik}, $p_{ik_n^*}$ を $\{A_{ik}\}$ と $\{A_{ik_n^*}\}$ の細分の上に拡張したものを $\mathring{p}_{ik}$，$\mathring{p}_{ik_n^*}$ とし，その細分の上でもともと定義されていたものを $\mathring{p}_i$，$\hat{\mathring{p}}_i$ とすれば，次の様に書きかえられ評価できる。

$$n\left|\sum_{i=1}^{m}(\mathring{p}_i-\hat{\mathring{p}}_i)\log\frac{\mathring{p}_{ik}}{\mathring{p}_{ik_n^*}}\right|$$

$$\le\left(n\sum_{i=1}^{m}\frac{(\hat{\mathring{p}}_i-\mathring{p}_i)^2}{\mathring{p}_i}\right)^{\frac{1}{2}}\left(n\sum_{i=1}^{m}\mathring{p}_i\left(\log\frac{\mathring{p}_{ik}}{\mathring{p}_{ik_n^*}}\right)^2\right)^{\frac{1}{2}}.$$

右辺の前半の部分は，漸近 χ_m であり，後半は $\hat{k}\to\infty$ in prob. であることより，k が十分大きなところで考えればよいから，そのとき

$$|H_n(\mathbb{p}_k)-H_n(\mathbb{p}_{k_n^*})|^{\frac{1}{2}}$$

と漸近的に等しいことに注意すれば結果を得る（see Shibata 1978a, 1978b）.

Orthogonal series estimate

2) での (C) の形の推定量を考えると，$\hat{\beta}$ は m.l.e. であるので，話は AIC と同じになる。実際，$q(n)$ の代わりに k，$\hat{f}_n(x)$ の代わりに $\hat{f}_{n,k}(x)$ と書けば Loss は

$$I_n(f,\hat{f}_{n,k})=I_n(f,f_k^*)+n\int f\log\frac{f_k^*}{\hat{f}_{n,k}}dx$$

$$f_k^*(x) = \exp\left(\sum_{j=1}^{k} \beta_j^*(k)\phi_j\right) / M(\beta^*(k)),$$
$$\beta^*(k);\ \frac{\partial}{\partial \beta_j} \log M(\beta^*(k)) = \int_{-\infty}^{\infty} \phi_j f\,dx,$$
$$j = 1, \cdots, k$$

と書きかえられる。

結論としては
$$f(x) = \exp\left(\sum_{j=1}^{\infty} \beta_j \phi_j(x)\right) / M(\beta)$$
と展開できて，$(\beta_1, \beta_2, \cdots)$ が infinitely many nonzero elements をもてば
$$S_n(k) = -\sum_{i=1}^{n} \log \hat{f}_{n,k}(X_i) + k$$
を minimize する $\hat{k}$ が漸近的な Loss の下限を attain することになる（Shibata 1979）。

付記

研究集会において，渋谷政昭氏（日本アイ・ビー・エム）より Density Estimation についての，かなりくわしい Bibliography が，最近，出版されていることを教えていただいた。1956年～1978年に出版された400近い論文の List である。ここに記して，感謝の意を表したい。

（W. Wertz and B. Schneider）

REFERENCES

Ahmad, I. A. (1979). Strong consistency of density estimation by orthogonal series methods for dependent variables with applications. *Ann. Inst. Statist. Math.* 31 A 279-88.

Bartlett, M. S. (1963). Statistical estimation of density functions. *Sankhya* A 25 245-54.

Brunk, H. D. (1978). Univariate density estimation by orthogonal series. *Biometrika* 65 521-8.

Craswell, K. J. (1965). Density estimation in a topological group. *Ann. Math. Statist.* 36 1047-8.

Epanechnikov, V. A. (1969). Nonparametric estimation of a multivariate probability density. *Theor. Prob. Appl.* 14 153-8.

Kronmal, R. and Tarter, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *J. Amer. Statist. Assoc.* 63 925-52.

Leonard, T. (1978). Density estimation, stochastic process and prior information. *J. R. Statist.* B 40 113-46.

Loftsgaarden, D. O. and Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.* 36 1049-51.

Moore, D. S. and Henrichon, E. G. (1969). Uniform consistency of some estimate of a density function. *Ann. Math. Statist.* 40 1499-1502.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* 23 1065-76.

Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. *Ann. Math. Statist.* 27 832-7.

Rosenblatt, M. (1971). Curve estimates. *Ann. Math. Statist.* 42 1815-42.

Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann. Statist.* 3 1-14.

Schwartz, S. C. (1967). Estimation of probability density by a orthogonal series. *Ann. Math. Statist.* 38 1262-5.

Shibata, R. (1978a). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. to appear in *Ann. Statist.* 8.

Shibata, R. (1978b). An optimal selection of regression variables. submitted to *Biometrika.*

Shorack, G. R. (1969). Asymptotic normality of linear combinations of functions of order statistics. *Ann. Math. Statist.* 40 2041-50.

Silverman, B. W. (1976). On a Gaussian process related to multivariate probability density estimation. *Math. Proc. Camb. Phil. Soc.* 80 135-44.

Silverman, B. W. (1978a). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* 6 177-84.

Silverman, B. W. (1978b). Choosing the window width when estimating a density. *Biometrika* 65 1-11.

Steele, J. M. (1978). Invalidity of average squared error criterion in density function estimates. *Canadian J. Statist.* 6 193-200.

Taylor, R. L. and Cheng, K. F. (1978). On the uniform complete convergence of density function estimates. *Ann. Inst. Statist. Math.* 30 A 397-406.

Watson, G. S. (1969). Density estimation by orthogonal series. *Ann. Math. Statist.* 40 1496-8.

Wertz, W. and Schneider, B. (1979). Statistical density estimation, a Bibliography. *International Statistical Review* 47 155-75.

Whittle, P. (1958). On the smoothing of probability density functions. *J. R. Statist. Soc.* B 20 334-43.