

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, \dots, X_n が与えられたとき、一般的に 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, $b(n) \rightarrow 0$

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

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)

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

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

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

$q(n)$: truncation point, $q(n) \rightarrow \infty$

または

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

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

しかし、この仮定と一般に $\hat{f}_n(x) \geq 0$ と仮定する

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

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

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

この形も提案されている (Leonard 1978).

Loss function とするは、Maximum squared error

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

または、Integrated squared error

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

を採用して、 $f(x)$, $b(n)$, $K(x)$ については $q(n)$ に依り、様々な条件をつけて、consistency, strong consistency あるいは asymptotic normality と論じているも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 $b(n)$ を求めた。Loss function とする Integrated squared error とすれば

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

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

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

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

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

が一つの optimal な, kernel と band width となる。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 consistency だけしか議論されていなが, 勿論, $g(n)$ のとり方が大きな問題である。ここでは, これら $b(n)$, $g(n)$ の選択と中心に話をすすめる。Loss function としては, Kullback-Leibler Information number

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

と採用する。squared error と integrated squared error
より山と谷に対する追従性はよく表わされる。また、これ
と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、 $f(x)$, $0 \leq x \leq 1$ は continuous, bounded
and bounded away from zero とあるとする。各 k に対応
して $[0, 1]$ の分割

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

と考えると、Histogram は A_{ik} に λ なる Sample 数 n_i と
すれば

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

で与えられる。これはモデル

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

のあとでパラメータ $\theta = p_1, \dots, p_k$ の制限つき ($\sum_{i=1}^k p_i = 1$, $p_i \geq 0$) m.l.e. がある。

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

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

となる。ただし、

$$p_k = (p_{1k}, p_{2k}, \dots, p_{kk})',$$

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

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

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

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

さて、

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

と置き、 $K_n \rightarrow \infty$ となる sequence に対して

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

とすれば

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

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

proof) $\left\{ \frac{p_{ik}}{p_{jk}} \right\}$ は bounded \prec ある \succ とよ

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

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

$\therefore \prec$

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

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

\prec 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 $1 \leq k \leq K_n$ について

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

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

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

\prec あるので, $k_n^* \sim n^{\frac{1}{3}}$ \prec 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 の選択を考えると, 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 することを示す。 $S_n(k)$ 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

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

したがって, Lemma より

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

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, $p_{ik}, p_{ik_n^*} \in \{A_{ik}\}$ と $\{A_{ik_n^*}\}$ の細分の上で拡張したものを $\dot{p}_{ik}, \dot{p}_{ik_n^*}$ とし, \dot{p}_{ik} の細分の上でもとも定義されておいたものを $\dot{p}_i, \hat{\dot{p}}_i$ とすれば, 次の様に書きかえられ評価できる。

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

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

右辺の前半の部分は, 漸近 χ_m^2 であり, 後半は $\hat{k} \rightarrow \infty$ in prob. であることより, k が十分大きくなるとろで考えればよいため, そのとき

$$|H_n(p_k) - H_n(p_{k_n^*})| \frac{1}{2}$$

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) 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, $g(n)$ 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

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

$$f_k^*(x) = \exp\left(\sum_{j=1}^k \beta_j^*(k) \phi_j\right) / M(\beta^*(k)),$$

$$\beta^*(k); \frac{\partial}{\partial \beta_j} \log M(\beta^*(k)) = \int_{-\infty}^{\infty} \phi_j f d\alpha,$$

$$j = 1, \dots, k$$

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、 $(\beta_1, \beta_2, \dots)$ が infinitely many nonzero elements となる。

$$S_n(k) = -\sum_{i=1}^n \log \hat{f}_{n,k}(X_i) + k$$

を minimize する \hat{k} が漸近的な loss の下限を attain するようになる (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、かなりくわしい Bibliography が、最近、出版されていることを教えてくれた。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、感謝の意を表す。

(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.