한국통계학회논문집 제2권 2호
1995년 12월 pp. 243-248

Optimal Designs for Multivariate Nonparametric Kernel Regression with Binary Data

Dongryeon Park[1)]

Abstract

The problem of optimal design for a nonparametric regression with binary data is considered. The aim of the statistical analysis is the estimation of a quantal response surface in two dimensions. Bias, variance and IMSE of kernel estimates are derived. The optimal design density with respect to asymptotic IMSE is constructed.

Key words: multivariate nonparametric regression, design density, binary data

1. Introduction and the estimator

Let us consider a situation where the outcome of an experiment is dichtomous, response or nonresponse. In quantal bioassay, dose levels of a substance are selected, and experimental animals are administered the substance at each dose level. After a fixed period of time, the animal has responded in some way and the response can be classified into two categories, response or nonresponse, for example dead or alive. Suppose we want to test the strength of some material. The impact of testing material at various impact levels constitutes an experiment. The outcome of the experiment is either "break" or "not break". In educational testing, one may want to determine the difficulty level of the test item. The independent variable is the difficulty level of the test item in some scale, and the outcome of the experiment is right or wrong answer.

In the experiment, the observed reaction Y_i of the ith subject at stimulus level $\boldsymbol{x}_i$ $(i=1,\cdots,n)$ is encoded by $Y_i=0$ (if nonresponse) or $Y_i=1$ (if response) where $\boldsymbol{x}_i$ is a $k\times 1$ vector of independent variables, and we assume that Y_i is a Bernoulli random variable with parameter $p(\boldsymbol{x}_i)$, $i=1,\cdots,n$. Here, $p : \boldsymbol{R}^k \rightarrow (0,1)$ denotes the response surface. Therefore the distribution of Y_i is given by

$$\Pr(Y_i = 1) = p(\boldsymbol{x}_i), \quad \Pr(Y_i = 0) = 1 - p(\boldsymbol{x}_i), \quad i=1,\cdots,n.$$

1) Department of Applied Statistics, Yonsei University, Seoul, 120-749, KOREA

The specification of the stimulus levels x_i forms the design of the experiment. The aim of the statistical analysis is the estimation of the surface p. A further assumption is that p is continuous and $0<p(x)<1$ for *all* $x\in\Omega$

Sometimes, we have a fixed sample size available and must decide on the location of all the design points $x_1, x_2, \cdots,$ in advance. If one uses a kernel-based estimate of p, then Muller and Schmitt (1988) describe the asymptotically optimal design density in the one-dimensional case. In most experiments, we have more than one independent variables, so extension to the multivariate case is desirable. We consider only the two dimensional case in which the design points x_i lie in R^2, but the methods developed could be extended to higher dimensions.

As an extension of the estimate proposed in Muller and Schmitt (1988) to the two dimensional case, we define the kernel estimator

$$\hat{p}(x)=\frac{1}{b^2}\sum_{i=1}^{n}\int_{A_i}K\left(\frac{x-s}{b}\right)ds\,Y_i \tag{1}$$

where b is a sequence of a positive bandwidths depending on n such that

$$b\to 0,\quad nb^2\to\infty \text{ as } n\to\infty$$

and where K is a kernel function, and where A_i is a partition of Ω such that $x_i\in A_i$ and $\cup_i A_i=\Omega$ and $A_i\cap A_j=\varnothing$, for *all* $i\neq j$ where $x_1,\cdots, x_n$ are the design points. Assume $\Omega=[0,1]^2$. In addition we assume $K(u)$ is continuous and has a compact support.

As in the one dimensional case, we assume that there is a strictly positive design density $f(x)$ and we determine the design points using $f(x)$. We choose $x_1,\cdots, x_n$ such that

$$\int_{A_i}f(x)dx=\frac{1}{n}$$

Therefore,

$$\Delta A_i=\frac{1}{nf(a_i)},\quad \text{for } some\ a_i\in A_i \tag{2}$$

where ΔA_i is the area of A_i, so $\Delta A_i=O(n^{-1})$. Furthermore, we assume that

$$\sup_i\sup_{u,v\in A_i}||u-v||=O(n^{-1/2})$$

where $||\cdot||$ is the Euclidean distance. This assumption restricts the shape of A_i. For example, suppose A_i is a thin rectangle such that the height is 1 and the width is $1/n$.

Then $\Delta A_i = 1/n$, but $\sup_{u,v \in A_i} \| u - v \| = 1$, so this assumption is not satisfied.

2. Asymptotic IMSE and Optimal Design

In this section, we compute the asymptotic IMSE of $\hat{p}(x)$ and derive the optimal design density. Let

$$I_x = \{ i : \text{support } K\left(\frac{x - \cdot}{b} \right) \cap A_i \neq \varnothing \}.$$

Note that $\#(I_x) = O(nb^2)$ where $\#(\cdot)$ denotes the cardinality. Let

$$\widetilde{A_i} = A_i \cap x + bS$$

where S=support(K). We can derive the expectation of $\hat{p}(x)$ using an integral approximation.

Lemma 1

$$E\hat{p}(x) = \frac{1}{b^2} \int_\Omega K\left(\frac{x - s}{b} \right) p(s) ds + O(n^{-1/2}) \quad \textit{where } \Omega = [0,1]^2.$$

Proof

The exact expectation of $\hat{p}(x)$ is

$$E\hat{p}(x) = \frac{1}{b^2} \sum_{i=1}^{n} \int_{A_i} K\left(\frac{x - s}{b} \right) ds \cdot p(x_i)$$

Since $\sup_i \sup_{u,v \in \widetilde{A_i}} \| u - v \| = O(n^{-1/2})$ and $\sum_{i \in I_x} \Delta \widetilde{A_i} \leq \#(I_x) O(n^{-1}) = O(b^2)$,

$$\begin{aligned} & \left| \sum_{i=1}^{n} p(x_i) \int_{A_i} K\left(\frac{x - s}{b} \right) ds - \int_\Omega K\left(\frac{x - s}{b} \right) p(s) ds \right| \\ &= \left| \sum_{i=1}^{n} \int_{\widetilde{A_i}} (p(x_i) - p(s)) K\left(\frac{x - s}{b} \right) ds \right| \\ &\leq \sup_i \sup_{u,v \in \widetilde{A_i}} \| u - v \| \cdot \sup_u K(u) \cdot \sum_{i \in I_x} \Delta \widetilde{A_i} \\ &= O\left(\frac{b^2}{n^{1/2}} \right) \end{aligned}$$

Using a Taylor series expansion, we can derive the bias of $\hat{p}(x)$.

Lemma 2 *Suppose that* $K(\boldsymbol{u})$ *is such that* $\int u_i K(\boldsymbol{u}) d\boldsymbol{u} = 0$, $i=1,2$ *and* $\int |u_i||u_j| K(\boldsymbol{u}) d\boldsymbol{u} < \infty$, *for all* $i, j = 1, 2$, *where* $\boldsymbol{u}^T = (u_1, u_2)$. *Then for any* $\boldsymbol{x} \in \Omega$,

$$E\,\hat{p}(\boldsymbol{x}) - p(\boldsymbol{x}) = \frac{b^2}{2} Q(p)(\boldsymbol{x}) + o(b^2) + O(n^{-1/2})$$

where $Q(p)(\boldsymbol{x}) = \int \boldsymbol{u}^T \nabla^2 p(\boldsymbol{x})\, \boldsymbol{u} K(\boldsymbol{u})\, d\boldsymbol{u}$

and where $\nabla^2 p(\boldsymbol{x})$ *is the Hessian matrix of the mixed second partials of* p *at* $\boldsymbol{x}$.

Proof

Let $\Omega^* = \left[\frac{x_1 - 1}{b}, \frac{x_1}{b}\right] \times \left[\frac{x_2 - 1}{b}, \frac{x_2}{b}\right]$.

The asymptotic expectation of $\hat{p}(\boldsymbol{x})$ is

$$\begin{aligned}
&\frac{1}{b^2} \int_{\Omega} K\left(\frac{\boldsymbol{x} - \boldsymbol{s}}{b}\right) p(\boldsymbol{s}) d\boldsymbol{s} = \int_{\Omega^*} K(\boldsymbol{u}) p(\boldsymbol{x} - b\boldsymbol{u}) d\boldsymbol{u} \\
&= \int_{\Omega^*} K(\boldsymbol{u}) \left[p(\boldsymbol{x}) - b\, \boldsymbol{u}^T \nabla p(\boldsymbol{x}) + \frac{b^2}{2}\, \boldsymbol{u}^T \nabla^2 p(\boldsymbol{x} - \theta b \boldsymbol{u})\, \boldsymbol{u} \right] d\boldsymbol{u}, \quad \theta \in [0,1] \\
&= p(\boldsymbol{x}) + \frac{b^2}{2} \int \boldsymbol{u}^T \nabla^2 p(\boldsymbol{x})\, \boldsymbol{u} K(\boldsymbol{u}) d\boldsymbol{u} + o(b^2)
\end{aligned}$$

since $K(\boldsymbol{u})$ has compact support and p is continuous. Therefore,

$$E\,\hat{p}(\boldsymbol{x}) - p(\boldsymbol{x}) = \frac{b^2}{2} Q(p)(\boldsymbol{x}) + o(b^2) + O(n^{-1/2}).$$

Lemma 3

$$Var(\hat{p}(\boldsymbol{x})) = \frac{\sigma^2(\boldsymbol{x})}{nb^2 f(\boldsymbol{x})} \int K^2(\boldsymbol{u}) d\boldsymbol{u} + o\left(\frac{1}{nb^2}\right), \quad \text{where } \sigma^2(\boldsymbol{x}) = p(\boldsymbol{x})(1 - p(\boldsymbol{x})).$$

Proof

From (1),

$$Var(\hat{p}(\boldsymbol{x})) = \frac{1}{b^4} \sum_{i=1}^{n} \left[\int_{A_i} K\left(\frac{\boldsymbol{x} - \boldsymbol{s}}{b}\right) d\boldsymbol{s} \right]^2 \sigma^2(\boldsymbol{x}_i)$$

By (2), for sufficiently large n,

$$\left| Var(\hat{p}(x)) - \frac{1}{nb^4}\int_\Omega K^2\left(\frac{x-s}{b}\right)\frac{\sigma^2(s)}{f(s)}ds \right|$$

$$= \frac{1}{b^4}\left| \sum_{i\in I_x}(\Delta\widetilde{A}_i)^2 K^2\left(\frac{x-w_i}{b}\right)\sigma^2(x_i) - \sum_{i\in I_x}\Delta\widetilde{A}_i K^2\left(\frac{x-z_i}{b}\right)\frac{\sigma^2(z_i)}{nf(z_i)} \right|$$

$$\le \frac{1}{b^4}\sum_{i\in I_x}\Delta\widetilde{A}_i \left| K^2\left(\frac{x-w_i}{b}\right)\frac{\sigma^2(x_i)}{nf(a_i)} - K^2\left(\frac{x-z_i}{b}\right)\frac{\sigma^2(z_i)}{nf(z_i)} \right|$$

where $w_i, z_i \in \widetilde{A}_i$.

Since $K(x)$ and $p(x)$ are continuous on the compact set, we can get

$$Var(\hat{p}(x)) = \frac{1}{nb^2}\left[\frac{\sigma^2(x)}{f(x)}V + o(1)\right], \quad \text{where } V = \int K^2(u)du .$$

Therefore we arrive at

$$E(\hat{p}(x) - p(x))^2 = \frac{1}{nb^2}\left[\frac{\sigma^2(x)}{f(x)}V + o(1)\right] + b^4\left[\frac{Q^2(p)(x)}{4} + o(1)\right] \tag{3}$$

We may integrate the MSE over $[0,1]^2$ to get the IMSE.

$$E\int_\Omega(\hat{p}(x) - p(x))^2 dx = \frac{1}{nb^2}\left[V\int_\Omega\frac{\sigma^2(x)}{f(x)}dx + o(1)\right] + \frac{b^4}{4}\left[\int_\Omega Q^2(p)(x)dx + o(1)\right] \tag{4}$$

The optimal global bandwidth b^* is obtained by minimizing (4) w. r. t. b:

$$b^* = \left[\frac{2V\int_\Omega\frac{\sigma^2(x)}{f(x)}dx}{n\int_\Omega Q^2(p)(x)dx}\right]^{1/6} \tag{5}$$

and inserting this bandwidth into (4) yields

$$E\int_\Omega(\hat{p}(x) - p(x))^2 dx = n^{-2/3}\left(c\cdot\left[V\int_\Omega\frac{\sigma^2(x)}{f(x)}dx\right]^{2/3} \times \left[\int_\Omega Q^2(p)(x)dx\right]^{1/3} + o(1)\right) \tag{6}$$

here $c = 2^{-1/3} + 2^{2/3}$. The optimal design density based on (6) is given in the following.

Theorem 1 *The optimal design density w. r. t. asymptotic IMSE is given by*

$$f^*(x) = \frac{\sqrt{p(x)(1-p(x))}}{\int_\Omega \sqrt{p(y)(1-p(y))}\,dy} \tag{7}$$

Proof

By (6), f^* is the solution of the following variational problem:

Find $f(x)$ such that $f(x)$ minimizes

$$\int_\Omega \frac{\sigma^2(x)}{f(x)}\,dx \quad s.t. \quad \int_\Omega f(x)\,dx = 1 \quad \text{and} \quad f(x)>0 \text{ for } all \ x\in\Omega \ .$$

Therefore, we just can follow the proof of the theorem in Muller(1984), even though the response is not binary in that article.

3. Discussion

According to (7), the asymptotic optimal design density $f^*(x)$ for the two dimensional case has a similar functional form to the one dimensional case. In the one dimensional case, when we select the optimal design points, we use the quantile function. However, there is no clear definition of the percentile in two dimensions. Therefore, optimal design idea for the univariate case can not be applied directly to the two dimensional case.

As in section 1, we wish to choose optimal design points such that

$$\int_{A_i} f^*(x)dx = \frac{1}{n} \tag{8}$$

However, there is no unique way to select optimal design points such that (8) is satisfied.

We may convert this problem into an optimization problem using a Voronoi Tessellation, and this is done by Park(1995), but further investigation is needed.

References

[1] Muller, H. (1984), "Optimal Designs for Nonparametric Kernel Regression", *Statistics & Probability Letters*, 2, 285-290.

[2] Muller, H. and Schmitt, T. (1988), "Kernel and Probit Estimates in Quantal Bioassay", Journal of the American Statistical Association, 83, 750-759.

[3] Park, D. (1995), "Sequential Design for Nonparametric Regression with Binary Data," Ph.D dissertation, The University of Michigan, Ann Arbor.