行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※

※ Orthogonal Latin Hypercube 抽樣方法 ※

※ 在電腦實驗中之應用 ※

※ (Application of Orthogonal Latin Hypercube ※

※ in Computer Experiments) ※

※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：X 個別型計畫 □整合型計畫

計畫編號：NSC89－2118－M－004－003－

執行期間：88 年 08 月 01 日至 89 年 07 月 31 日

計畫主持人：丁兆平

共同主持人：

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

□出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

執行單位：國立政治大學統計系

中 華 民 國 89 年 10 月 31 日

一、 中文摘要

從 1979 年開始以 Latin hypercube 設計來建構電腦實驗(computer experiment)的研究便開始展開,自此研究的方向大致是以改善 Latin hypercube 設計含蓋整體空間性以及改善其最適性為主要。本研究報告是以探究 Latin hypercube 設計之架構以及其基本採樣精神為主,並比較Latin hypercube抽樣與隨機抽樣(random sampling)和分層隨機抽樣(stratified random sampling)之差異。本研究亦探究 Latin hypercube 抽樣在電腦實驗上之應用,並研究當模型為迴歸模型時 Latin hypercube 抽樣如何優於隨機抽樣。

關鍵詞: Latin hypercube 設計,Latin hypercube 抽樣,隨機抽樣,分層隨機抽樣。

Abstract

Latin hypercube designs have been used in conducting computer experiments since 1979. After that, lots of efforts have been made to improve them either from a better-space filling perspective or form an optimality perspective. In this article we investigated thoroughly about the nature of Latin hypercube designs and compared Latin hypercube sampling with random sampling and stratified sampling. We also looked at the application of Latin hypercube sampling to computer experiments and found out that when a regression model is used, the regression may be more accurately estimated by Latin hypercube sampling than random sampling.

Keywords: Latin hypercube designs, Lain hypercube sampling, random sampling, stratified random sampling.

二、 研究報告

Latin Hypercube Designs

A. Notation

Input variable $= \bar{X} = (X_1, \cdots, X_k)' \in S \subset R^k$, where S is the 'sample'space.

Output variable $= Y = h(\bar{X}) \in R$, where 'h' is not explicitly known, it might be determined by computer code, for example.

We wish to estimate some quantity using estimator

$$T(Y_1, \cdots, Y_N) = \frac{1}{N} \sum_{i=1}^{N} g(Y_i), \text{ where } g \text{ is a known function,}$$

based on a sample $\bar{X}_1, \cdots, \bar{X}_N$ of the $\bar{X}'s$ of size N and then determining $Y_i = h(\bar{X})$ each of these $\bar{X}'s$.

Assume

(i) $\bar{X}$ is distributed in S according to some distribution $F(\bar{x})$.

(ii) $S_1, \cdots S_I$ are disjoint subsets of S such that

$S_1 \cup \cdots \cup S_I = S$, and

$p_i = P(\bar{X} \in S_i)$ = measure of the size of S_i,

B. Sample Selection Methods

(i) Random Sampling

Let $\bar{X}_1, \cdots, \bar{X}_N$ be a random sample from S according to the distribution $F(\bar{x})$ on S. For simplicity, assume F is the uniform distribution.

Note: If N is "large" we hope our sample will be spread nice and evenly throughout S. However, this need not be true if N is "small".

(ii) Stratified Samping

We attempt to "force" our sample to be more evenly spread throughout S using stratification.

Let $S_1, \cdots S_I$ be our "strata" and suppose we wish to sample n_i points from S_i with $\sum_{i=1}^{I} n_i = N$. Let $\bar{X}_{i1}, \cdots, \bar{X}_{in_i}$ be a random sample according to the conditional distribution of F on S_i, i.e.

$$F(\bar{X} \mid S_i) = \frac{1}{p_i} F(\bar{X}) \cdot I_{\bar{X}_i \in S_i}, \text{ where } I \text{ is an indicator function.}$$

The $\{\bar{X}_{ij}\}_{1 \le i \le I, 1 \le j \le n_i}$ is our stratified sample. For simplicity, we take F to be uniform on S. We assume samples in different strata are independent.

(iii) Latin Hypercube Sampling

Another attempt to "force" our sample to be more evenly spread throughout S than random sampling.

Let X_i be the i-th coordinate of $\bar{X} = (X_1, \cdots, X_k)'$. For each i, divide the range of X_i into N strata (intervals) of equal marginal probability $1/N$ under F. Sample once from each stratum and let these sample values be denoted $X_{i1}, \cdots, X_{iN}$. Form the $k \times N$ array as in the following.

$$\begin{array}{cccc} X_{11} & X_{12} & \cdots & X_{1N} \\ X_{21} & X_{22} & \cdots & X_{2N} \\ & \vdots & & \\ X_{k1} & X_{k2} & \cdots & X_{kN} \end{array}$$

Randomly permute each row, using independent permutations for each row. The N columns of the resulting array are our Latin Hypercube sample. For simplicity, we take F to be uniform on S.

C. Properties of the Estimator T Under These Plans

(i) Random Sampling

Recall

$$T(Y_1,\cdots,Y_N)=\frac{1}{N}\Sigma_{i=1}^{N}g(Y_i)$$

Let T_R denote this estimator when random sampling is used. Assume $F(\bar{x})$ has a density, say $f(\bar{x})$ and let

$$\tau = E(g(Y)) = E(g(h(\bar{X}))) = \int_S g(h(\bar{x}))f(\bar{x})d\bar{x},$$

$$\theta^2 = Var(g(Y)),$$

so that $E(T_R)=\tau$, $Var(T_R)=\theta^2/N$.

(ii) Stratified Sampling

General case – an unbiased estimator of τ. Recall $\{S_i;1\le i\le I\}$ is a partition of S and

$$p_i = P(\bar{X}\in S_i), 1\le i\le I,$$

we again assume F and density f. Recall that we select a random sample

$$f(\bar{x}\mid S_i)=\begin{cases}(1/p_i)f(\bar{x}) & if\ \bar{x}\in S_i\\ 0 & o.w.\end{cases}$$

of size n_i from $S_i, \Sigma_{i=1}^{I}n_i = N$. Let $\bar{X}_{i1},\cdots,\bar{X}_{in_i}$ be the random sample from S_i so that the X_{ij} are *i.i.d.* $f(\bar{x}\mid S_i)$, $1\le j\le n_i$.

Let $Y_{ij}=h(X_{ij})$ = observation corresponding to X_{ij}, then

$$\mu_i = E(g(Y_{ij})) = \int_{S_i} g(y)(1/p_i)f(\bar{x})d\bar{x},$$

$$\sigma_i^2 = Var(g(Y_{ij})) = \int_{S_i}[g(y)-\mu_i]^2(1/p_i)f(\bar{x})d\bar{x}.$$

Let $T_S=\Sigma_{i=1}^{I}[(p_i/n_i)\Sigma_{j=1}^{n_i}g(Y_{ij})]$, then $E(T_S)=\tau$. Thus T_S is an unbiased estimator of τ. Also since samples from different strata are independent $Var(T_S)=\Sigma_{i=1}^{I}(p_i^2/n_i)\sigma_i^2$.

(iii) Latin Hypercube Sampling

Here we assume F is such that the coordinates of $\bar{X}$ are independent. We also assume F has a density f. Recall that for each coordinate X_i of $\bar{X}=(X_1,\cdots,X_k)'$ we divide the range of X_i into N strata or intervals of equal marginal probability $1/N$ under F. The Cartesian product of these interval partitions S into N^k cells each of probability N^{-k}. Each of these N^k cells can be labelled by a set of k cell coordinates

$$\bar{m}_i=(m_{i1},\cdots,m_{ik})',\ 1\le i\le N^k,$$

where m_{ij} = interval number (between 1 and N) of coordinate X_i represented in cell i.

One way to select a Latin hypercube sample of size N is to take a random sample of N of the N^k cells, say $\bar{m}_{l_1},\cdots,\bar{m}_{l_N}$ subject to the condition that for each j, the set $\{m_{i_l j}\}_{l=1}^{N}$ is a permutation of the integers 1, 2, ... , N. Then a single random observation is made in each cell. We then have that the density of $\bar{X}$ given $\bar{X} \in$ cell i is

$$f(\bar{x} \mid \bar{x} \in \textit{cell } i) = \begin{cases} N^k f(\bar{x}), & \bar{x} \in \textit{cell } i \\ 0 & \textit{o.w.} \end{cases}$$

Thus the distribution of $Y = h(\bar{X})$ under Latin hypercube sampling is

$$P(Y \le y) = \sum_{i=1}^{N^k} P(Y \le y \mid \bar{X} \in \textit{cell } i) P(\bar{X} \in \textit{cell } i)$$

$$= \int_{h(\bar{x}) \le y} f(\bar{x}) d\bar{x},$$

which is the same as for random sampling. Thus if

$$T_L(Y_1,\cdots,Y_N) = (1/N)\sum_{i=1}^{N} g(Y_i)$$

is T under Latin hypercube sampling then $E(T_L) = \tau$ which is same as for random sampling.

To calculate $Var(T_L)$, we look at the sampling as follows. Select $\bar{X}_i$ independently and at random from each of the N^k cells and let

$$Y_i = h(\bar{X}_i),\ 1 \le i \le N^k.$$

We then independently select our sample of N cells as defined previously, letting

$$w_i = \begin{cases} 1 & \textit{if cell } i \textit{ is in our sample} \\ 0 & \textit{o.w.} \end{cases}$$

Then

$$Var(T_L) = \frac{1}{N^2}\left\{\sum_{i=1}^{N^k} Var(w_i g(Y_i)) + \sum_{i=1}^{N^k}\sum_{j \ne i}^{N^k} \mathrm{cov}(w_i g(Y_i), w_j g(Y_j))\right\}$$

Through some tedious calculation one can show that

$$Var(T_L) = Var(T_R) + \frac{N-1}{N}\left(\frac{1}{N^k (N-1)^k}\right)\sum_R \sum (\mu_i - \tau)(\mu_j - \tau)$$

$$\le Var(T_R)\ ,\ \text{provided the second term above is 0.}$$

Theorem.(McKay, Beckman, and Conover (1979)) If is monotonic in each of its arguments and if $g(Y)$ is monotonic in Y, then $Var(T_L) \le Var(T_R)$.

An Application to Computer Experiments

Let $\vec{Z}(\vec{X})$ be a vector valued function for which a linear model

$$Y \cong \vec{Z}'(\vec{X})\vec{\beta}$$

seems an appropriate approximate to $h(\vec{X})$. The "population" least squares value of $\vec{\beta}$ is

$$\vec{\beta}^{p} = \left(\int_S \vec{Z}(\vec{x})\vec{Z}'(\vec{x})dF(\vec{x}) \right)^{-1} \int_S \vec{Z}(\vec{x})Y(\vec{x})d\vec{x}.$$

Assuming $\int \vec{Z}(\vec{x})\vec{Z}'(\vec{x})dF(\vec{x})$ is known or easily computable, we might estimate $\vec{\beta}^{p}$ by

$$\hat{\vec{\beta}} = \left(\int_S \vec{Z}(\vec{x})\vec{Z}'(\vec{x})dF(\vec{x}) \right)^{-1} \frac{1}{N} \Sigma_{i=1}^{N} \vec{Z}(\vec{x}_i)Y_i .$$

The variance of $\hat{\vec{\beta}}$ is

$$\left(\int_S \vec{Z}(\vec{x})\vec{Z}'(\vec{x})dF(\vec{x}) \right)^{-1} \Sigma \left(\int_S \vec{Z}(\vec{x})\vec{Z}'(\vec{x})dF(\vec{x}) \right)^{-1},$$

where

$\Sigma = ((\Sigma_{ij})) = d \times d$ matrix,

$\Sigma_{ij} = \int r_i(\vec{x})r_j(\vec{x})d\vec{x}$,

$r_l(\vec{x})$ = residual from additivity for h_l,

$\vec{h} = (h_1, \cdots, h_d)' \in R^d$ = bounded function on S,

$\vec{Y}_i = \vec{h}(\vec{X}_i)$, $1 \le i \le N$

Owen argues that to the extent that $\vec{Z}(\vec{X})Y(\vec{X})$ is additive, the regression may be more accurately estimated by Latin hypercube sampling than random sampling.

三、 參考文獻

Hoehler, J. R. and A. B. Owen (1996). Computer Experiments. Handbook of Statistics, Vol. 13 (S. Ghosh and C. R. Rao, eds.), Elsevier Science B. V.

Loh, W. L. (1993). On Latin hypercube sampling. Tech. Report No. 93-52, Dept. of Statistics, Purdue University.

McKay, M. D., Beckman, R. J. and W. J. Conover (1979). A comparison of three methods for selecting values of input variables in the analysis of output from computer code. *Technometrics* **21**, 239-245.

Morris, M. D., Mitchell, T. J. and D. Ylvisaker (1993). Bayesian design and analysis of computer experiments: use of derivatives in surface prediction. *Technometrics* **35**,

243-255.

O'Hagan, A. (1989). Comment: Design and analysis of computer experiments. *Statist. Sci.* **4**(4), 430-432.

O'Hagan, A. (1992). Some Bayesian numerical analysis. *Bayesian Statist.* **4**, 345-363.

Owen, A. B. (1992a). A central limit theorem for Latin hypercube sampling. *J. Roy. Statist. Soc. Ser.* B **54**, 541-555.

Owen, A. B. (1992b). Orthogonal arrays for computer experiments, integration and visualization. *Statist. Sinica* **2**, 439-452.

Owen, A. B. (1994). Controlling correlations in Latin hypercube samples. *J. Amer. Statist. Assoc.* **89**, 1517-1522.

Sacks, J., Welch, W. J., Mitchell, T. J. and H. P. Wynn (1989). Design and analysis of computer experiments. *Statist. Sci.* **4**, 409-435.

Stein, M. (1987). Large sample properties of simulations using Latin hypercube sampling. *Technometrics* **29**, 143-151.