Reliability Engineering and System Safety 148 (2016) 96–108

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

The generalization of Latin hypercube sampling

Michael D. Shields*, Jiaxin Zhang

Department of Civil Engineering, Johns Hopkins University, USA

ARTICLE INFO

Article history:
Received 1 July 2015
Received in revised form
29 October 2015
Accepted 7 December 2015
Available online 18 December 2015

Keywords:
Uncertainty quantification
Monte Carlo simulation
Stratified sampling
Latin hypercube sampling

ABSTRACT

Latin hypercube sampling (LHS) is generalized in terms of a spectrum of stratified sampling (SS) designs referred to as partially stratified sample (PSS) designs. True SS and LHS are shown to represent the extremes of the PSS spectrum. The variance of PSS estimates is derived along with some asymptotic properties. PSS designs are shown to reduce variance associated with variable interactions, whereas LHS reduces variance associated with main effects. Challenges associated with the use of PSS designs and their limitations are discussed. To overcome these challenges, the PSS method is coupled with a new method called Latinized stratified sampling (LSS) that produces sample sets that are simultaneously SS and LHS. The LSS method is equivalent to an Orthogonal Array based LHS under certain conditions but is easier to obtain. Utilizing an LSS on the subspaces of a PSS provides a sampling strategy that reduces variance associated with both main effects and variable interactions and can be designed specially to minimize variance for a given problem. Several high-dimensional numerical examples highlight the strengths and limitations of the method. The Latinized partially stratified sampling method is then applied to identify the best sample strategy for uncertainty quantification on a plate buckling problem.

1. Introduction

Monte Carlo simulation and its variants are used ubiquitously in assessing reliability probabilistically [1–5]. Latin hypercube sampling (LHS) [6,7] is perhaps the most widely used random sampling method for Monte Carlo-based uncertainty quantification and reliability analysis, employed in nearly every field of computational science, engineering, and mathematics [8–11]. The seminal work by McKay et al. [6] introducing Latin hypercube sampling is a classic in the field of design of computer experiments. LHS is an especially powerful and useful sampling method thanks primarily to the properties identified by Stein [12] who showed that LHS has the effect of filtering the variance associated with the additive components of a transformation (or main effects). This result, combined with the Hierarchical Ordering Principle [13] – which states that main effects and low order interactions are likely to govern most general transformations – causes LHS to reduce variance significantly for many applications.

The widespread popularity of LHS has led to the invention of numerous variants meant to improve space-filling [14–20], optimize projective properties [21], minimize least square error and maximize entropy [22], and reduce spurious correlations [23–26,16,18,27]. Meanwhile, LHS has been applied to nearly every type of probabilistic analysis one can imagine, ranging from the estimation of reliability (probability of failure) [28,4,29,30] to coefficient estimation for polynomial chaos, neural network, and other types of surrogate models [8,31]. The intention of this paper is not to present another variant of the LHS methodology or to apply it in a new or novel way. Rather, we present a broad generalization of the methodology in the context of stratified sampling – from which LHS is derived.

Stratified sampling (SS), the "parent" methodology of LHS has been widely used in the social sciences and financial mathematics owing to its ability to partition a population into strata (or categories) that can be weighted according to their conditional probabilities. Some recent developments have begun to encourage its use in Monte Carlo uncertainty and reliability analysis of computer models by adaptively stratifying the probability space of a random vector [32,5]. These methods rely on "true" stratified sampling wherein all dimensions of the space are stratified simultaneously allowing the analyst to concentrate samples in probabilistically weighted regions of the space that are important for the problem at hand. LHS, meanwhile, lies at the opposite end of the "spectrum" of stratified sampling methods (Fig. 1) such that each dimension of the random vector is stratified individually and the vector is constructed through random pairings.

In this work, the intermediate space on the spectrum of stratified sampling methods is explored such that stratification can occur on any set of N_i-dimensional orthogonal subspaces of the N-

* Corresponding author.
E-mail address: michael.shields@jhu.edu (M.D. Shields).

http://dx.doi.org/10.1016/j.ress.2015.12.002
0951-8320/

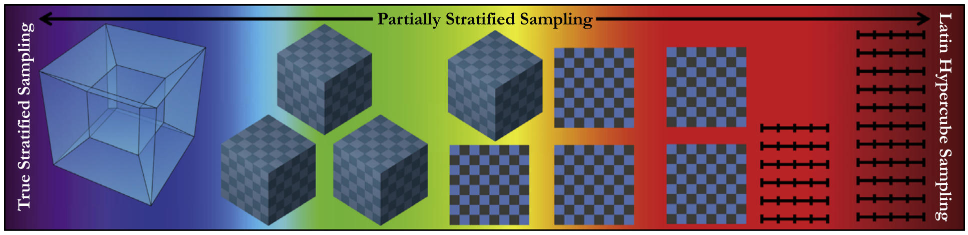

Fig. 1. Spectrum of stratified sampling methods.

dimensional sample space $\mathcal{S}$ subject to $\sum_i N_i = N$. This generalized sampling method is referred to as Partially Stratified Sampling (PSS) and the properties of their designs are explored in detail. In particular, the PSS method is shown to reduce variance associated with low-dimensional interactions within a high-dimensional transformation. Some discussion surrounding the appropriate use of a PSS design is provided and a powerful hybrid PSS-LHS method is proposed that simultaneously reduces variance associated with the main-effects and low-dimensional interactions. This method, referred to as "Latinized" Partially Stratified Sampling (LPSS), combines the variance reductions of both PSS and LHS to yield a major improvement in sample efficiency. Several high-dimensional demonstration problems are presented and the method is applied to the probabilistic assessment of plate buckling strength where the interaction of geometric and material variables are very important.

2. Review of sampling methods

This section provides a brief review of the sampling methods used for analysis in this work. We will consider only uncorrelated random variables as it is common practice to maps correlated variables onto a set of uncorrelated ones using, for example, Principal Component Analysis.

2.1. Simple random sampling

Classical Monte Carlo methods rely on so-called Simple Random Sampling (SRS) or Monte Carlo Sampling in which realizations of the vector $\mathbf{x}$ (samples) are generated as independent and identically distributed (iid) realizations on $\mathcal{S}$ with marginal cumulative distribution functions (CDFs) $D_{X_i}(\cdot)$ by

$$x_i = D_{X_i}^{-1}(U_i); \quad i = 1, 2, \ldots, n \tag{1}$$

where U_i are iid uniformly distributed samples on [0, 1]. The realizations $\mathbf{x}$ are then applied to the system $\mathbf{y} = \mathbf{F}(\mathbf{x})$ and $\mathbf{y}$ is statistically evaluated.

2.2. Stratified sampling

Stratified Sampling begins by dividing the sample space $\mathcal{S}$ into a collection of M disjoint subsets (strata) $\mathbf{\Omega}_k; k = 1, 2, \ldots, M$ with $\cup_{k=1}^{M} \mathbf{\Omega}_k = \mathcal{S}$ and $\mathbf{\Omega}_p \cap \mathbf{\Omega}_q = \varnothing; p \neq q$. Sample realizations from a given stratum k, $\mathbf{x}_k = \{x_{1k}, x_{2k}, \ldots, x_{Nk}\}$, are generated by randomly sampling the vector components according to

$$x_{ik} = D_{X_i}^{-1}(U_{ik}); \quad i = 1, 2, \ldots, N \tag{2}$$

where U_{ik} are iid uniformly distributed samples on $[\xi_{ik}^l, \xi_{ik}^u]$ with $\xi_{ik}^l = D_{X_i}(\zeta_{ik}^l)$ and $\xi_{ik}^u = D_{X_i}(\zeta_{ik}^u)$, and ${\zeta_{ik}}^l$ and ${\zeta_{ik}}^u$ denote the lower and upper bounds respectively of the ith vector component of stratum $\mathbf{\Omega}_k$. Typically the stratification is performed directly in the probability space meaning that the strata are defined directly by the bounds ξ_{ik}^l and ξ_{ik}^u.

2.3. Latin Hypercube Sampling

Latin Hypercube Sampling (LHS) operates by dividing the subspace of each vector component $\mathcal{S}_i$; $i = 1, 2, \ldots, N$ into $M = n$ disjoint subsets (strata) of *equal probability* $\mathbf{\Omega}_{ik}$; $i = 1, 2, \ldots, N; k = 1, 2, \ldots, M$. Samples of each vector component are drawn from the respective strata according to

$$x_{ik} = D_{X_i}^{-1}(U_{ik}); \quad i = 1, 2, \ldots, N; \quad k = 1, 2, \ldots, M \tag{3}$$

where U_{ik} are iid uniformly distributed samples on $[\xi_k^l, \xi_k^u]$ with $\xi_k^l = (k-1)/M$ and $\xi_k^u = (k/M)$. The samples $\mathbf{x}$ are assembled by randomly grouping the terms of the generated vector components. That is, a term x_{ik} is randomly selected from each vector component (without replacement) and these terms are grouped to produce a sample. This process is repeated $M = n$ times.

Because the component samples are randomly paired, an LHS is not unique; there are $(M!)^{N-1}$ possible combinations. With this in mind, improved LHS algorithms iterate to determine optimal pairings according to some specified criteria – such as reduced correlation among the terms or enhanced space-filling properties (e.g. [14,24,26,33,21,18,19]).

2.4. Variance reduction in stratified sampling and Latin hypercube sampling

Consider the general statistical estimator defined by

$$T(y_1, \ldots, y_n) = \sum_{l=1}^{n} w_l g(y_l) \tag{4}$$

where $y_l = h(x_l)$ and x_l denotes a sample generated using SRS, SS, or LHS, w_l are sample weights, and $g(\cdot)$ is an arbitrary function. Note that when $g(y) = y^r$, T is an estimate of the rth moment and when $g(y) = I\{y \le Y\}$, where $I\{\cdot\}$ denotes the indicator function, T is the empirical CDF. For conventional Monte Carlo analysis, $w_l = (1/n)\ \forall l$ and the variance of the statistical estimator (T_R) is given by $\mathrm{Var}\,[T_R] = (\sigma^2/n)$ where $\sigma^2 = \mathrm{Var}[g(Y)]$. Classical Monte Carlo estimates generally serve as the measure by which variance reduction techniques are compared.

SS and LHS are both techniques to reduce the variance of statistical estimators when compared to classical Monte Carlo estimates although they do so through different statistical mechanisms. Stratified sampling has been proven to unconditionally reduce the variance of statistical estimators (denoted T_S) when compared to SRS such that the variance reduction depends on the differences between the strata means μ_k and the overall mean τ as [6]

$$\mathrm{Var}[T_S] = \mathrm{Var}[T_R] - \frac{1}{n} \sum_{k=1}^{M} p_k (\mu_k - \tau)^2 \tag{5}$$

when the strata are sampled proportionately such that the number of samples on stratum k, $n_k = p_k n$. LHS, on the other hand, reduces variance by creating negative covariance between sample cells such that the variance of a LHS estimator T_L can be expressed