**Multivariate Extensions of Nonlinear Control Variates and Concomitants of Order Statistics**

Tim Hesterberg

**Abstract**

We develop multivariate extensions to two variance reduction techniques – nonlinear control variates for estimating expected values, and concomitants of order statistics for estimating a distribution (including expected values, probabilities, and percentiles). We allow the relationship between $Y$ and the $X$ variables to be nonlinear, and consider two cases – where the relationship is additive, and the general case. We use additive regression models and univariate numerical integration to implement the additive case, and numerical integration or Quasi-Monte Carlo methods as part of the implementation of Monte Carlo variance reduction methods.

1991 *Mathematics Subject Classification: 65C05*

**Keywords:** Quasi-Monte Carlo, Nonlinear Control Variates

## 1. Introduction

This article considers estimation of $\mathrm{E}[Y]$ and (the distribution function) $F_Y$ in Monte Carlo simulation when there exist one or more covariates with known distributions. The number of input variables in a simulation may be very large, but we assume that $Y$ can be well approximated by a function of a small number $d$ of covariates (which may be functions of the other input variables). That is, we assume that

$$Y = S(X_1, X_2, \ldots, X_d) + \epsilon \tag{1.1}$$

where $(X_1, X_2, \ldots, X_d)$ have a known joint distribution (or at least that an accurate approximation is available) and $\epsilon$ has variance which is small relative to the variance of $Y$. The $X$'s and $\epsilon$ need not be independent. A useful special case of (1.1) is the "additive model" where

$$Y = s_1(X_1) + s_2(X_2) + \ldots + s_d(X_d) + \epsilon. \tag{1.2}$$

The $S$ and/or $s$ functions may be chosen prior to a Monte Carlo simulation, or may be estimated by a linear or nonlinear regression on $n$ observations from a Monte Carlo simulation.

We discuss control variate procedures for estimating $\mathrm{E}[Y]$ in Section 2, and the method of concomitants of order statistics for estimating $F_Y$ in Section 3. Quasi-Monte Carlo methods play a role in implementation of both methods in higher-dimensional problems.

## 2. Control Variates

We progress in this section through increasingly general cases for the relationship between $Y$ and the $X$'s, beginning with the linear case, continuing with the additive case with $s$ functions chosen prior to a simulation, following with the additive case with $s$ functions estimated from the data, and conclude with the general $S$ case.

Suppose that

$$Y = c_0 + \sum_{j=1}^{d} c_j X_j + \epsilon \tag{2.1}$$

where the $X_j$ are random variables with known means. The classical control variates estimate for $\mu := \mathrm{E}[Y]$ is

$$\hat{\mu}_{\mathrm{CV}} = c_0 + \sum_{j=1}^{d} c_j \mathrm{E}[X_j] + \overline{\epsilon}. \tag{2.2}$$

The variance of the estimate is $n^{-1}\mathrm{Var}(\epsilon)$, compared to the variance $n^{-1}\mathrm{Var}(Y)$ of the simple Monte Carlo estimate $\overline{Y}$. The optimal coefficients are the regression coefficients for $Y$ against the $X$'s, which in practice are estimated using linear least-squares regression (with asymptotically negligible effects on the mean square error).

Consider next the additive model (1.2) in place of (2.1). If the $s$'s are chosen prior to the simulation, a trivial extension of (2.2) yields $\hat{\mu} = \sum_{j=1}^{d} \mathrm{E}[s_j(X_j)] + \overline{\epsilon}$. To implement this requires the calculation of the expected values $\mathrm{E}[s_j(X_j)]$, which can be done analytically or using univariate numerical integration.

Lewis, Ressler, and Wood (1989) considered the additive model with $s$'s estimated from the data using the ACE algorithm of Breiman and Friedman (1985). Unfortunately, they conclude (p. 658) that "the transformations ACE selects cannot be used to develop control variables for variance reduction since the

transformations are non-parametric and the true means of the transformed variables cannot be determined." We disagree. The true means can be calculated to any desired level of accuracy using numerical integration, using the capability of ACE to produce values of the estimated $s$ functions at any point (not just at the observed values of the $x$'s), which suffices for most numerical integration methods. We propose the following algorithm:

`Algorithm:`

`Generate` $n$ `Monte Carlo observations of` $Y$ `and the` $X$`'s.`

`Fit` $Y = \sum_{j=1}^{d} s_j(X_j) + \epsilon$.

`Calculate` $\mathrm{E}[s_j(X)]$ `for` $j = 1, \ldots, d$ `using numerical integration.`

`Let` $\hat{\mu} = \sum_{j=1}^{d} \mathrm{E}[s_j(X)] + tt$.

The fit can be performed with ACE or other additive regression models, such as those in Hastie and Tibshirani (1990)

Turn now to the general, non-additive model (1.2). As in the additive case, the function $S$ can be chosen a-priori or estimated from the data, using a nonlinear regression procedure. The difficulty in the general model is that $\mathrm{E}[S(X_1, \ldots, X_j)]$ cannot be evaluated using univariate numerical integration. If $d$ is relatively small the deterministic integration procedures in Davis and Rabinowitz (1984) may suffice. If $d$ is large either Monte Carlo integration or Quasi-Monte Carlo integration (Niederreiter 1978, 1988) may be used.

It seems odd to use Monte Carlo simulation within a Monte Carlo simulation. But if a single observation of $S(X_1, \ldots, X_j)$ is less expensive to obtain than an observation of $Y$ (because $S$ is less expensive to compute than is $Y$, or because $Y$ depends on many more input variables than does $S$), the optimal design would use relatively few Monte Carlo observations to estimate $\mathrm{E}[\epsilon]$, and relatively many to estimate $\mathrm{E}[S]$, particularly since $\mathrm{Var}(S) \gg \mathrm{Var}(\epsilon)$.

Similarly, the use of Quasi-Monte Carlo methods within a Monte Carlo simulation is justified if $S$ is less expensive than $Y$ to compute. The reduction in dimensionality (from many to $d$ variables) is especially helpful for Quasi-Monte Carlo methods.

## 3. Concomitants of Order Statistics

The concomitants of order statistics procedure is used to estimate the distribution $F_Y$ of $Y$. We begin with the case of one covariate (David 1973, Efron 1990, Do and Hall 1992), and continue with two proposals for the multivariate case. In this section it does not matter whether the relationship between $Y$ and the $X$'s is linear, additive, or general, so our notation assumes the general case.

In the univariate case we drop the subscript on $X_1$. Let $\pi$ be permutation of $(1, \ldots, n)$ such that $X_i = X_{(\pi_i)}$; $\pi_i$ is the rank of $X_i$. The concomitants of order statistics estimate of the distribution of $Y$ is

$$\hat{F}_Y(y) = n^{-1} \sum_{i=1}^{n} I(Y_i^* \leq y) \tag{3.1}$$

where

$$Y_i^* = S(X_i^*) + \epsilon_i. \tag{3.2}$$

and

$$X_i^* = F_X^{-1}(\frac{\pi_i - .5}{n}). \tag{3.3}$$

This is the empirical distribution formed by replacing the random $X$ in (1.1) with a deterministic quantile $X^*$ corresponding to the rank of $X$. Figures 1 and 2 show plots of $Y$ vs. $X$ and $Y^*$ vs. $X^*$, when $X \sim U(0,1)$, $\epsilon \sim N(0, 0.01^2)$ and $Y = X^2 + \epsilon$, for a set of 200 random observations. Note the perfectly even spacing in the marginal distribution of $X^*$, and that the distribution of $Y^*$ is more even than was that of $Y$.

We give the variance of the univariate concomitants procedure in the following theorem:

**Theorem**

Suppose that $X$ and $\epsilon$ have joint density $h(x, \epsilon)$, that $Y = S(X) + \epsilon$, that the marginal density of $X$ has convex support, and that $Q(a|x) := P(\epsilon \leq a | X = x)$ is twice differentiable with respect to the second argument. Then the asymptotic variance of the concomitants distribution function estimate (3.1) is

$$\lim_{n\to\infty} n\mathrm{Var}(\hat{F}_Y(a)) = \int Q(a - S(x)|x)(1 - Q(a - S(x)|x))f_X(x)dx + 2\iint_{x_1<x_2} Q_{01}(a - S(x_1)|x_1)Q_{01}(a - S(x_2)|x_2)F_X(x_1)(1 - F_X(x_2))dx_1dx_2 \quad (3.4)$$

where $Q_{01}(a|x) = \frac{d}{dx}Q(a|x)$. The proof is in the appendix.

The first term in (3.4) represents the expected value of the conditional variance of the Bernoulli variable $I(Y \leq a)$, $\mathrm{E}[\mathrm{Var}(I(Y \leq a|X))]$, and is the inherent variability induced by $\epsilon$. The second term follows from a hidden assumption in the procedure, that the conditional distribution of $\epsilon_i$ given $X = X_i$ is approximately the same as the conditional distribution given $X = F_X^{-1}((\pi_i - .5)/n)$. This assumption is met if the conditional distribution of $\epsilon$ given $X$ changes only slowly with respect to $X$ and if $X_i \approx X_i^*$. The theorem indicates that the estimate has a finite asymptotic variance as long as there are no intervals with zero probability within in the support of $X$. However, the variance may be large if the conditional distribution of $\epsilon$ given $X$ depends strongly on $X$. Figure 3 demonstrates what happens with a bad choice of $S$, in this case $S(x) = \sin(2\pi x)$; the conditional mean $\mathrm{E}[\epsilon|x]$ oscillates, and when the residuals $Y - S(X)$ are added to the curve at new $x$-values $X^*$ the result no longer approximates the true curve.

### 3.1 Multivariate Concomitants

In the multivariate case we consider two alternatives, replacing $X$ with $X^*$ independently for each variable, or a joint replacement.

Let $\mathbf{X}$ be the matrix with elements $X_{i,j}$, where $X_{i,j}$ is the $i$'th observation of the $j$'th covariate. For independent concomitants we let

$$Y_i^* = S(X_{i,1}^*, X_{i,2}^*, \ldots, X_{i,d}^*) + \epsilon_i \quad (3.5)$$

replace $Y_i$ in the empirical distribution function, where the $X_{i,j}^*$ are computed separately for each dimension $d$, as in (3.3). Figures 4 shows a scatterplot of unadjusted $X_2$ vs. $X_1$, and Figure 5 a plot of adjusted $X_2^*$ vs. $X_1^*$, for 200 random observations. The marginal distributions of $X_1^*$ and $X_2^*$ are perfectly spaced, but the joint distribution is still irregular.

We make the joint distribution regular by replacing the $d$-dimensional covariate observations with regularly-spaced points $\mathbf{X}_{i,\cdot}^{**}$, as shown in Figure 6. We begin by letting $\mathbf{U}_{i,\cdot}$ be a set of $n$ points generated by a Quasi-Monte Carlo procedure to be approximately uniformly distributed on the unit $d$-dimensional cube. In Figure 6 the points are based on a slightly modified version of the Hammersley sequence (Niederreiter 1978). Our first dimension is defined by $\mathbf{U}_{i,1} = (i - 0.5)/n$ (we subtract 0.5 to avoid observations on the boundary

of the cube). Subsequent dimensions are defined by $\mathbf{U}_{i,j} = \phi_{p_j}(i)$, where $p_j$ is the $(j-1)$'st prime number and $\phi$ is the radical inverse function defined by $\phi_p(i) = \sum_k a_k p^{-k-1}$ where $i = \sum_k a_k p^k$. These points are formed by reversing the base $p_j$ representation of $i$; for example, $\mathbf{U}_{\cdot,2} = (0.1, 0.01, 0.11, 0.001, 0.101, \ldots)$ in base 2 notation, or $(1/2, 1/4, 3/4, 1/8, 5/8, \ldots)$.

Now, let

$$\mathbf{X}^{**}_{i,\cdot} = \mathbf{G}(\mathbf{U}_{\pi(i),\cdot}),$$

where $\pi$ is some permutation and $\mathbf{G} : \mathbf{R}^d \mapsto \mathbf{R}^d$ is a transformation such that the vector-valued random variable $\mathbf{G}(U_1, \ldots, U_d)$ has the same distribution as $(X_1, \ldots, X_d)$ when $U_1, \ldots, U_d \sim U(0,1)$.

In choosing the permutation $\pi$, we note that (as in the one-dimensional case) it is important that $\mathbf{X}_{i,\cdot} \approx \mathbf{G}(\mathbf{U}_{\pi(i),\cdot})$, so that the distribution of $\epsilon$ given $\mathbf{X}_{i,\cdot}$ is approximately the same as given $\mathbf{X}^{**}_{i,\cdot}$. We choose $\pi$ so that sum of distances

$$\sum_{i=1}^{n} |\mathbf{X}_{i,\cdot} - \mathbf{G}(\mathbf{U}_{\pi(i),\cdot})|$$

is as small as can be achieved with reasonable computational efficiency. The global minimum could be achieved using the Hungarian Matching Algorithm (Nering and Tucker 1993) from the $n \times n$ matrix of pairwise distances, but that algorithm requires between $O(n^2)$ and $O(n^3)$ operations.

We suggest instead a recursive $O(n \log(n))$ method. Partition both sets of $d$-dimensional points into two equal groups based on the first dimension, so that the first $n/2$ observations have the smallest $n/2$ values of $X_1$. For both sets, partition each group on the second dimension (so the first $n/4$ values of $X_2$ are smaller than the second $n/4$ values). Continue partitioning on successive dimensions (return the the first dimension after the $d$'th dimension until the subgroups are reduced to size 1. For example, if (one set of) the unordered points are given by

$[X_1 = (3,5,2,4,1,8,7,6), X_2 = (4,2,6,1,3,7,5,8)]$, the successive orderings are:

$[X_1 = (3,2,4,1,\ 5,8,7,6), X_2 = (4,6,1,3,\ 2,7,5,8)]$ (after partitioning on $X_1$),

$[X_1 = (4,1,\ 3,2,\ 5,7,\ 8,6), X_2 = (1,3,\ 4,6,\ 2,5,\ 7,8)]$, (after partitioning on $X_2$),

$[X_1 = (1,4,\ 2,3,\ 5,7,\ 6,8), X_2 = (3,1,\ 6,4,\ 2,5,\ 8,7)]$, (after the final partition on $X_1$).

After both sets have been reordered the corresponding observations are matched. The result in shown in Figure 7, with line segments from the original data to the Quasi-Monte Carlo points, with filled circles at the latter. The matching is not optimal, but is reasonably good.

Note that the upper right and left corners of the scatterplot are devoid of (Monte Carlo) observations in Figure 3. In Figure 7 we observe that the nearest Monte Carlo points are matched with Quasi-Monte Carlo points in those corners, like tadpoles swimming to fill the corners.

For estimating a single probability $F_Y(y)$ for $y$ fixed, this concomitants algorithm for joint distributions should be used only for moderately-sized data sets. Asymptotically, the computational cost for $n$ observations is $O(n \log n)$ for an error variance of $O(n^{-1})$, whereas simple Monte Carlo estimation has cost $O(n)$ for error variance of $O(n^{-1})$; the advantage of the concomitants algorithm for moderately-sized data sets is that the constant term in the variance is smaller.

For estimating the complete distribution or quantiles of $Y$, the computational cost for simple Monte Carlo is also $O(n \log n)$ (from sorting the $Y$'s), so the concomitants algorithm is also asymptotically competitive.

## Appendix: Proof of Theorem (3.4)

By conditioning on $\mathbf{X} = (X_1, \ldots, X_n)$, we decompose the variance of $\hat{F}$ into two terms

$$\mathrm{Var}(\hat{F}) = \mathrm{E}[\mathrm{Var}(\hat{F}|\mathbf{X})] + \mathrm{Var}(\mathrm{E}[\hat{F}|\mathbf{X}]). \tag{A.1}$$

We will show that these two terms correspond to the single and double integrals in (3.4), respectively.

Let $\xi_i = F_X^{-1}((i-.5)/n)$, $X_{(i)}$ be the $i$'th order statistic of the $X$'s, and $\epsilon_{[i]}$ be the corresponding value of $\epsilon$. The conditional distribution of $\epsilon_{[i]}$ given $X_{(i)} = x$ is given by $Q(\cdot|x)$, and depends only on $X_{(i)}$. We rewrite (3.1) as

$$\hat{F}_Y(a) = n^{-1}\sum_{i=1}^{n} I(S(\xi_i) + \epsilon_{[i]} \le a) = n^{-1}\sum_{i=1}^{n} I_i$$

where $I$ is the logical indicator function and $I_i = I(S(\xi_i) + \epsilon_{[i]} \le a)$.

Turn now to the first term of (A.1). We manipulate the inside of the first term to obtain

$$\begin{aligned}
\mathrm{Var}(\hat{F}_Y(a)|\mathbf{X}) &= n^{-2}\mathrm{Var}(\sum_{i=1}^{n} I_i|\mathbf{X}) \\
&= n^{-2}\sum_{i=1}^{n}\mathrm{Var}(I_i|X_{(i)}) \\
&= n^{-2}\sum_{i=1}^{n}\tilde{Q}(a - S(\xi_i)|X_{(i)}) \\
&= n^{-2}\sum_{i=1}^{n}\tilde{Q}(a - S(\xi_i)|\xi_i) + \tilde{Q}_{01}(a - S(\xi_i)|\xi_i)(X_{(i)} - \xi_i) + O((X_{(i)} - \xi_i)^2)
\end{aligned}$$

where $\tilde{Q} = Q(1-Q)$ and $\tilde{Q}_{01}(a|x) = \frac{d}{dx}\tilde{Q}(a|x)$. Without loss of generality, we assume that $X$ has bounded support (let $x' = t(x)$ where $t$ is an increasing twice-differentiable bounded transformation, and let $S'(\cdot) = S(t^{-1}(\cdot))$, then replace $S$ and $X$ in the theorem with $S'$ and $X'$), so that $\mathrm{E}[X_{(i)}] = \xi_i + o(1)$ and $(X_{(i)} - \xi_i)^2 = O_P(n^{-1})$ uniformly in $i$. Then

$$\begin{aligned}
n\mathrm{E}[\mathrm{Var}(\hat{F}|\mathbf{X})] &= n^{-1}\sum_{i=1}^{n}\tilde{Q}(a - S(\xi_i)|\xi_i) + o(1) \\
&\to \int \tilde{Q}(a - S(x)|x) f_X(x) dx.
\end{aligned}$$

Turn now to the second term of (A.1).

$$\begin{aligned}
n\mathrm{Var}(\mathrm{E}[\hat{F}|\mathbf{X}]) &= n^{-1}\mathrm{Var}(\sum_{i=1}^{n} Q(a - S(\xi_i)|X_{(i)})) \\
&= n^{-1}\mathrm{Var}(\sum_{i=1}^{n} Q(a - S(\xi_i)|\xi_i) + Q_{01}(a - S(\xi_i)|\xi_i)(X_{(i)} - \xi_i) + O_P(n^{-1})) \\
&= 2n^{-1}\sum_{i<j} Q_{01}(a - S(\xi_i)|\xi_i) Q_{01}(a - S(\xi_j)|\xi_j)\mathrm{Cov}(X_{(i)}, X_{(j)}) + O_P(n^{-1}) \\
&= 2n^{-2}\sum_{i<j} Q_{01}(a - S(\xi_i)|\xi_i) Q_{01}(a - S(\xi_j)|\xi_j)\frac{F_X(\xi_i)(1 - F_X(\xi_j))}{f_X(\xi_i) f_X(\xi_j)} + O_P(n^{-1}) \\
&\to 2\iint_{x_1 < x_2} Q_{01}(a - S(x_1)|x_1) Q_{01}(a - S(x_2)|x_2) F_X(x_1)(1 - F_X(x_2)) dx_1 dx_2.
\end{aligned}$$

Thus the two terms match, and the proof is complete. Similar calculations show that the bias of (3.1) is of order $O(1/n)$. Do and Hall (1992) give another formula for the variance, which they prove using a different method.

## References


Breiman, L. and Friedman, J.H. (1985) "Estimating Optimal Transformations for Multiple Regression and Correlation" (with discussion), *Journal of the American Statistical Association* **80**, 580-619.

David, H. A. (1973) "Concomitants of Order Statistics," *Bull. Int. Statist. Inst.* **45** 295-300.

Davis, P. J., and Rabinowitz, P. (1984) *Methods of Numerical integration*, 2nd ed., Academic Press, New York.

Do, K. & Hall, P. (1992). "Distribution Estimation using Concomitants of Order Statistics, with Application to Monte Carlo Simulation for the Bootstrap," *Journal of the Royal Statistical Society, Series B* **54**, 2, 595-607.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*, Chapman and Hall.

Lewis, P. A. W., Ressler, R. L., and Wood, R. K. (1989) "Variance Reduction Using Nonlinear Controls and Transformations," *Communications in Statistics B, Simulation and Computation* **18**, 655-672.

Nering E. D. and Tucker, A. W. (1993) *Linear Programs and Related Problems*, San Diego: Academic Press.

Niederreiter, H. (1978) "Quasi-Monte Carlo Methods and Pseudo-Random Numbers," *Bulletin of the American Mathematical Society* **84** 957-1041.

Niederreiter, H. (1988) "Quasi-Monte Carlo Methods for Multidimensional Numerical Integration," *Numerical Integration III, International Series of Numerical Mathematics*, H. Brass and G. Haemmerlin eds., Vol 85, Birkhaeuser-Verlag, Basel, 157-171.

Figure 1: Simple Monte Carlo

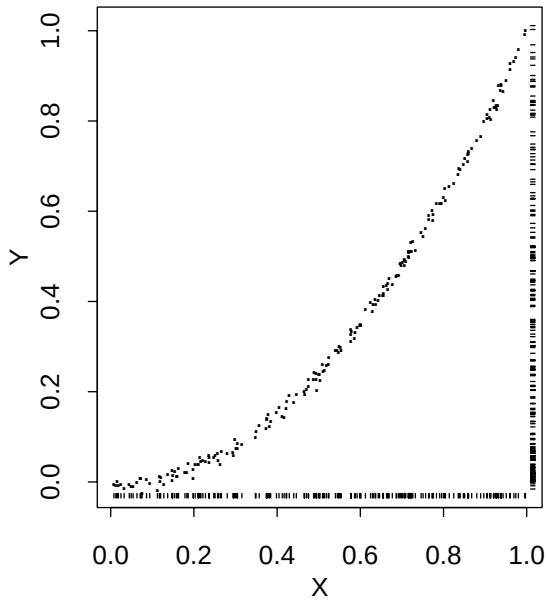



Figure 2: One-Dimensional Concomitants

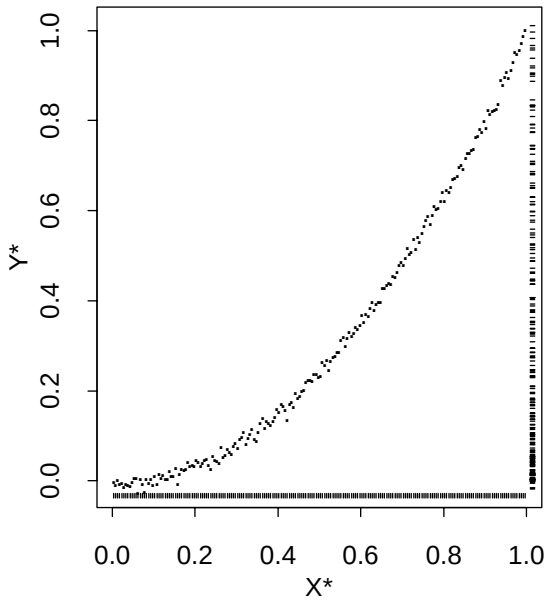
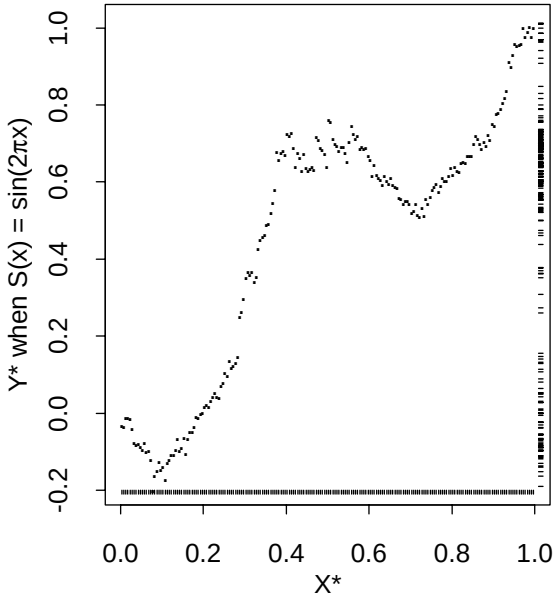



Figure 3: Bad Choice of Function S




Marginal distributions are at the right and bottom of each plot.

Figure 4: Random Covariates

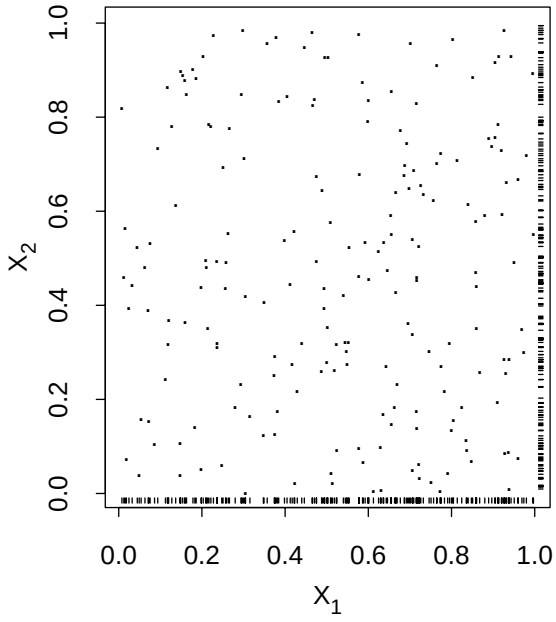



Figure 5: Independent Concomitants

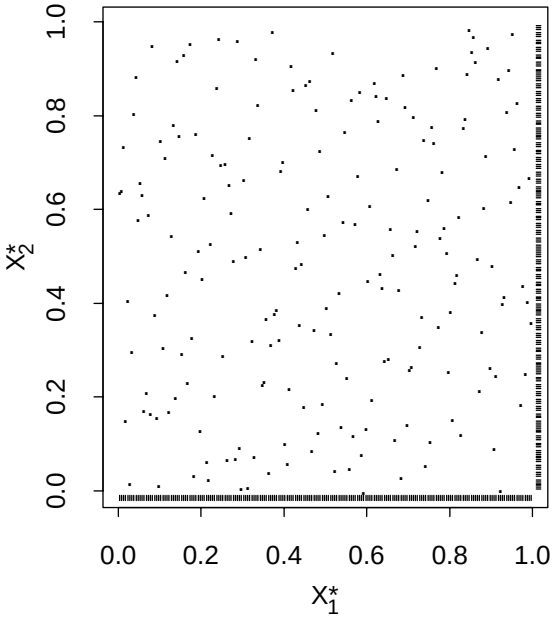



Figure 6: Quasi-Monte Carlo Points

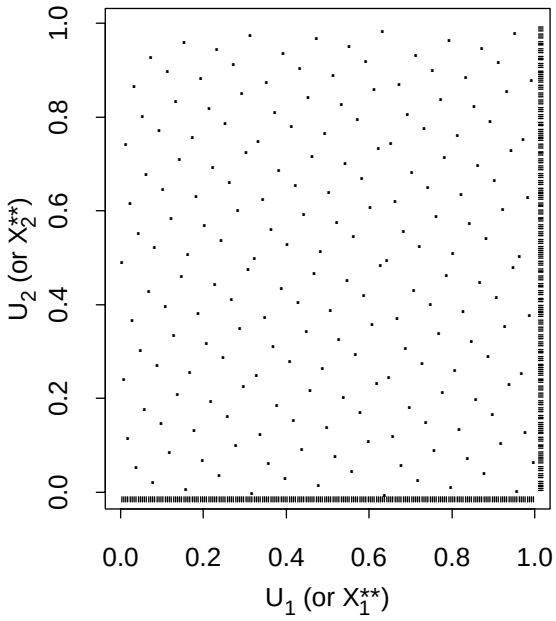
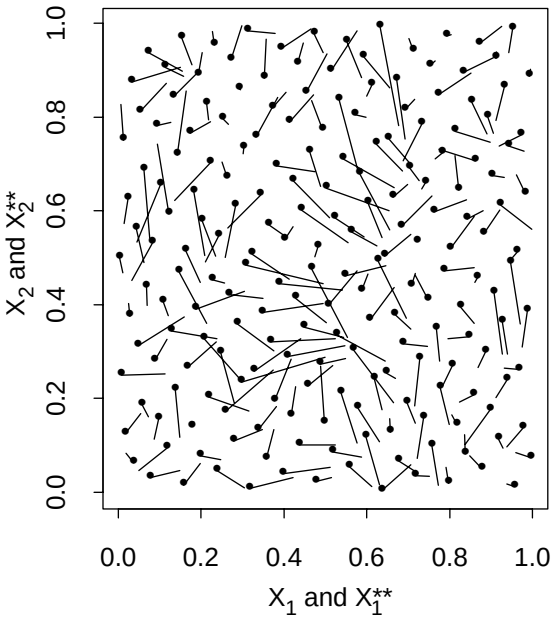



Figure 7: Matching Q-MC points to Data




Marginal distributions are at the right and bottom of each plot.