DESIGN AND ANALYSIS OF COMPUTER EXPERIMENTS

PROPOSED 5000-LEVEL ST MODULE

by Wei-Liem Loh
1.1 Introduction

Various sciences use mathematical/statistical models to describe physical phenomena that are often so complex that physical experimentation is too time consuming or too expensive. Some examples of such phenomena being

- combustion [Mitchell and Morris (1992)],
- automobile manufacture [Mitchell and Morris (1992), Ye, Li and Sudjianto (2000)],
- VLSI-circuit design [Koehler and Owen (1996), Sacks, Schiller and Welch (1989)],
- radiological protection [Kennedy and O’Hagan (2001)],
- hydrology [McKay, Beckman and Conover (1979), Kennedy and O’Hagan (2001)],
- cooling systems [Ye (1998)].

These models are typically implemented in computer codes. Due to the complexity of the physical phenomenon, the resulting computer code is large and may be expensive in terms of computer time for a single run. Nevertheless it is much cheaper running a computer model than physical experimentation.
In these applications, the output $y$ from the computer code is often deterministic, i.e., running the code twice with the same inputs or explanatory variables $x$ would result in the same output. This is one of the crucial differences between computer experiments and statistical experiments. There is no measurement error here and hence replications or repeated measurements are unnecessary.

Berk, et. al. (2002) reports on a recent workshop on statistical approaches for the evaluation of complex computer models and Wu and Hamada (2000) is a very readable text on statistical experiments in general giving the state of the art techniques with lots of real data and case studies.

For definiteness, we shall denote the computer program as $f$ and the output $y \in \mathbb{R}^p$ is related to the input $x \in [0, 1)^d$, the $d$-dimensional unit hypercube, via

$$y = f(x).$$
This module will initially follow the approach of Koehler and Owen (1996). In particular, we shall present two statistical approaches to computer experiments. In both approaches, randomness is introduced in order to obtain error bounds (such as confidence intervals) for the relevant estimates.

The first approach will be loosely called the frequentist approach. Here \( f \) is a fixed (non-random) function and randomness is introduced by choosing random input points (with specified balance or uniformity properties).

The second approach will also be loosely called the Bayesian approach where the randomness is introduced by modeling the response \( y \) as a realization of a (Gaussian) random field.

Besides Koehler and Owen (1996), other expository articles on computer experiments include

- Sacks, Schiller and Welch (1989),
- Sacks, Welch, Mitchell and Wynn (1989),
- Currin, Mitchell, Morris and Ylvisaker (1991),
- Bates, Buck, Riccomagno and Wynn (1996),
1.2 Objectives in Computer Experiments

There are many different objectives in computer experiments. The following are just some of them.

**Optimization**
Find $x_0 \in [0, 1]^d$ such that

$$f(x_0) = \max \{ f(x) : x \in [0, 1]^d \},$$

using a fixed number, say $n$, of computer runs. The field of optimization is very large (possibly as large as the field of statistics). There are lots of very powerful optimization algorithms available. However these generally converge to a local maximum (as opposed to a global maximum) and the need for a good starting or initial value is crucial. Computer experiments can be used with advantage to obtain such initial values.

**Visualization**
As Diaconis (1988) stated, being able to compute a function at any given point does not necessarily mean that one “understands” the function. Computer experimentation can serve as a “crude” data analytic method of visualizing the function and its behavior.
Prediction and Approximation
The original computer program $f(.)$ may be exceedingly expensive to evaluate. The aim here is to approximate or proxy $f(.)$ by a much simpler function, say, $\hat{f}(.)$. $\hat{f}(x)$ can then be used as a cheap predictor for other values of $f(x)$.

Numerical Integration
The goal here is to estimate

$$\mu = \int_{[0,1]^d} f(x) dx$$

using a fixed number of computer runs evaluated at the inputs, say, $x_1, \ldots, x_n$. $\mu$ is then estimated via the “sample mean”

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} f(x_i).$$

This is a very active research area in the field of numerical analysis and it is often beneficial to draw on their results.
1.3 More Detailed Course Outline

1.3.1 Frequentist approach

- Regular grid
- Latin hypercube sampling
- Stein’s method of probability approximations
- Good lattice points or uniform design
- Randomized orthogonal arrays and other related designs
- Scrambled nets

1.3.2 Bayesian approach

- Gaussian random fields and prediction
- Exponential and Matérn-type correlation functions
- Fixed-domain or infill asymptotics
- Consistency of maximum likelihood estimates
2.1 Frequentist Approach

The frequentist approach to computer experiments can be thought of being based on numerical integration. For example consider the linear model:

\[ Y = f(X) \approx Z(X)\beta, \]

where \( Y \in R, X \in [0, 1]^d, Z(.) \) is a row vector of known predictor functions and \( \beta \) is a (column) vector of unknown parameters. Recall that \( f(.) \) denotes a computer program and \( Z(.) \) may be low order polynomials, trigonometric polynomials, wavelets, etc.

Assuming that \( X \) is distributed according to the probability distribution \( F \) on \([0, 1]^d\), the least squares estimate for \( \beta \) is given by

\[
\hat{\beta}_{LS} = \left[ \int_{[0,1]^d} Z(X)' Z(X) dF \right]^{-1} \int_{[0,1]^d} Z(X)' f(x) dF.
\]

So in order to evaluate \( \hat{\beta}_{LS} \), it suffices only to evaluate the two integrals in the r.h.s. of the above equation.

The precision of the estimate \( \hat{\beta}_{LS} \) can be assessed via the integrated mean squared error

\[
IMSE = \int_{[0,1]^d} [Y - Z(X)\beta]^2 dF.
\]
This is also a problem in numerical integration. For simplicity, we assume for the remainder of Section 2 that $F$ is the uniform distribution on $[0, 1]^d$ and that integrals of the form

$$
\mu = \int_{[0, 1]^d} f(X) dF
$$

be estimated by

$$(1) \quad \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} f(x_i),$$

where $x_1, \cdots, x_n$ are suitably chosen input design points.

This naturally gives rise to two questions:

- Given a fixed number of function evaluations, say $n$, how do we choose the design points $x_1, \cdots, x_n$ so that the integral estimate $\hat{\mu}$ in (1) is as accurate as possible?

- How do we compute error bounds for such an estimate?
2.2 Other Design Criteria

Before we review the various common designs used by computer experimenters, we shall first, for the sake of completeness, discuss other design criteria besides accuracy in numerical integration.

2.2.1 Integrated Mean Squared Error

Box and Draper (1959) proposed minimizing the integrated mean squared prediction error (IMSE) of \( \hat{f}(x) \) over \([0, 1]^d\). Also see Welch (1983), Sacks and Schiller (1988) and Sacks, Schiller and Welch (1989).

A design \( D_I = \{x_1, \cdots, x_n\} \) is an IMSE design if

\[
J(D_I) = \min\{J(D) : D\},
\]

where

\[
J(D) = \int_{[0,1]^d} E[f(x) - \hat{f}(x)]^2 dx,
\]

and \( \hat{f} \) is the predictor constructed from the generic design set \( D \) of \( n \) points in \([0, 1]^d\).
2.2.2 Maximum Mean Squared Error

Instead of integrating the mean squared prediction error, maximum mean squared error (MMSE) is a minimax criterion which chooses the design to minimize

\[ \max \{ E[f(x) - \hat{f}(x)]^2 : x \in [0, 1]^d \}. \]

Sacks and Schiller (1988) did a comparison between the IMSE and MMSE criteria. MMSE criterion in general is computationally more complex as it involves a \(d\)-dimensional optimization of \(f\).
2.2.3 Maximin and Minimax Designs

These designs go back at least to Johnson, Moore and Ylvisaker (1990). The designs are dependent on a distance measure or metric. Let $d(\cdot,\cdot)$ denote a metric on $[0,1)^d$. Then

$$d(x_1, x_2) = d(x_2, x_1),$$
$$d(x_1, x_1) \geq 0,$$
$$d(x_1, x_2) = 0 \text{ iff } x_1 = x_2,$$
$$d(x_1, x_2) \leq d(x_1, x_3) + d(x_3, x_2).$$

A design $D_{MI}$ is a minimax distance design if

$$\min_D \max_{x \in [0,1)^d} d(x, D) = \max_{x \in [0,1)^d} d(x, D_{MI}),$$

where

$$d(x, D) = \min_{x_0 \in D} d(x, x_0).$$

Minimax distance designs ensure that all points in $[0,1)^d$ are not too far away from a design point.

A design $D_{MA}$ is a maximin distance design if

$$\max_D \min_{x_1, x_2 \in D} d(x_1, x_2) = \min_{x_1, x_2 \in D_{MA}} d(x_1, x_2).$$

A maximin distance design ensures that no two points of the design are too close together.
2.2.4 Maximum Entropy Designs

Lindley (1956) introduced a measure, based on Shannon’s entropy [see Shannon (1948)], of the amount of information provided by an experiment. Maximum entropy designs are Bayesian measures that uses the expected reduction in entropy as a design criterion.

A design $D_E$ is a maximum entropy design if

$$E[- \log p(Y_{D_E})] = \min_D E[- \log p(Y_D)],$$

where $p(Y_D)$ is the posterior density of $Y$ given the observations at the design points $D$. Other uses of maximum entropy designs can be found in Box and Hill (1967) for model discrimination, Shewry and Wynn (1987) for spatial sampling and Currin, Mitchell, Morris and Ylvisaker (1988) for the design of computer experiments.

More of maximum entropy designs later in the course when we address the Bayesian approach to computer experiments.
2.2.5 Sequential Designs

Thus far, the design criteria described above are for the case of fixed design size \( n \). There does not seem to be too much work done on sequential designs for computer experiments. See Sacks, Welch, Mitchell and Wynn (1989), page 415, for a rather brief discussion and also Williams, Santner and Notz (2000).

**Remark.** Although the design criteria being discussed are quite intuitive and simple to state, except for the smallest values of \( n \), the exact optimal design for each criterion is unlikely to be found as each of these gives rise to extremely difficult (probably intractable) high dimensional optimization problems. A more realistic aim is to construct approximately optimal designs.
2.3 I.I.D. Sample

For the rest of Section 2, we shall for simplicity take $p = 1$, i.e., $f(.)$ is a real-valued function. It is also convenient to define

$$
\hat{\mu}_{IID} = \frac{1}{n} \sum_{i=1}^{n} f(X_i),
$$

if $X_1, \ldots, X_n$ is an i.i.d. sample where $X_1$ is uniformly distributed on $[0, 1)^d$.

If $E[f(X_1)^2] < \infty$ and Lindeberg’s condition holds, e.g. see Chung (1974) page 205, the central limit theorem states that $\hat{\mu}_{IID}$ is asymptotically normally distributed and that

$$
|\hat{\mu}_{IID} - \mu| = O_p(n^{-1/2}).
$$

This is usually the standard in which the performance of alternative designs are benchmarked against.

We observe that the rate $n^{-1/2}$ does not depend on the dimension $d$ (it does not suffer from the so-called “curse of dimensionality”) even though, of course, $d$ does appear in the constant. Also the bound on the r.h.s. of the above equation is a probabilistic bound (not a deterministic one).
2.4 Regular Grid or Lattice

This design is probably the most intuitive. Given \( n = m^d \) runs of a computer experiment, choose input points \( x_1, \ldots, x_n \in [0, 1)^d \) such that

\[
\{x_1, \ldots, x_n\} = \left\{ \frac{1}{m^d} \prod_{k=1}^{d} (i_k - 0.5) : 1 \leq i_1, \ldots, i_d \leq m \right\}.
\]

There exist a number of serious drawbacks of this design.

- Firstly the number of design points must be some power of the dimension \( d \) which may be unnecessarily restrictive.

- It is well known that the accuracy based on the integral grid estimate is typically that of a univariate integral based on \( m = n^{1/d} \) inputs, e.g., see Davis and Rabinowitz (1984). Clearly for large \( d \), this is a severe disadvantage.
• If the computer program $f(.)$ is an additive function of only a few covariates, say,

$$f(t) = f_1(t_1) + f_2(t_2), \quad \forall t = (t_1, \cdots, t_d)' \in [0, 1)^d,$$

then the grid design gives rise to replications (assuming $d \geq 3$). This is wasteful and is clearly undesirable.
2.5 Latin Hypercube Sampling

Latin hypercube sampling was introduced by McKay, Beckman and Conover (1979) in what is widely considered to be the first statistics paper on computer experiments. A Latin hypercube sample $X_1, \cdots, X_n$ satisfies

$$X_i = (X_{i,1}, \cdots, X_{i,d})',$$

$$X_{i,j} = \frac{\pi_j(i) - U_{i,j}}{n}, \quad \forall 1 \leq i \leq n, 1 \leq j \leq d,$$

where $\pi_1, \cdots, \pi_d$ are independent permutations of the set $\{1, \cdots, n\}$, all $n!$ permutations equally likely, and $U_{i,j}, 1 \leq i \leq n, 1 \leq j \leq d$ are independent $U[0, 1]$ random variables independent of the $\pi_j$s.

The main point behind Latin hypercube sampling is that if the inputs, i.e., $X_1, \cdots, X_n$ are projected onto any one of the $d$ co-ordinate axes, then complete stratification is achieved. We write

$$\hat{\mu}_{LHS} = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

The following lists down some of the properties of Latin hypercube sampling found in the literature.
McKay, Beckman and Conover (1979) proved

**Theorem 1** If \( f(.) \) is monotone in each of its \( d \) arguments, then

\[
\text{Var}(\hat{\mu}_{LHS}) \leq \text{Var}(\hat{\mu}_{IID}).
\]

Note that Theorem 1 is a non-asymptotic result.

Stein (1987) finds an expression for the variance of \( \hat{\mu}_{LHS} \). Assuming that \( \int_{[0,1]^d} f(t)^2 dt < \infty \), he applies an ANOVA-type decomposition to \( f \) and obtains

\[
f(t) = \mu + \sum_{j=1}^{d} \alpha_j(t_j) + r(t),
\]

where

\[
t = (t_1, \cdots, t_d)',
\]

\[
\mu = \int_{[0,1]^d} f(t) dt,
\]

\[
\alpha_j(t_j) = \int_{[0,1]^{d-1}} [f(t) - \mu] \prod_{1 \leq k \leq d: k \neq j} dt_k.
\]

Here \( \mu \) can be thought of as the grand mean, \( \alpha_1, \cdots, \alpha_d \) the univariate main effects and \( r \) the residual from additivity.
Stein (1987) further shows that
\[
\text{Var}(\hat{\mu}_{\text{IID}}) = \frac{1}{n} \left[ \int_{[0,1]^d} r(t)^2 dt + \sum_{j=1}^{d} \int_{0}^{1} \alpha_j(t_j)^2 dt_j \right],
\]
\[
\text{Var}(\hat{\mu}_{LHS}) = \frac{1}{n} \int_{[0,1]^d} r(t)^2 dt + o\left(\frac{1}{n}\right).
\]
This clearly implies that for large \(n\),
\[
\text{Var}(\hat{\mu}_{LHS}) \leq \text{Var}(\hat{\mu}_{\text{IID}}),
\]
i.e., by balancing the univariate margins, Latin hypercube sampling has removed the main effects \(\alpha_j\) from the error variance. Note that generally, the improvement is in the constant and not the convergence rate which is \(n^{-1/2}\), i.e., \(|\hat{\mu}_{LHS} - \mu| = O_p(n^{-1/2})\).

Owen (1992a) proved the following central limit theorem for Latin hypercube sampling via method of moments. Such a result would be useful for constructing error bounds (e.g. confidence intervals).

**Theorem 2** Let \(f : [0,1]^d \rightarrow \mathbb{R}\) be a bounded function. Then \(n^{1/2}(\hat{\mu}_{LHS} - \mu)\) tends in distribution to \(N(0, \int_{[0,1]^d} r(t)^2 dt)\) as \(n \rightarrow \infty\).
It is noted that in the case $d = 2$, Owen’s result is exactly Hoeffding’s combinatorial central limit theorem [see Hoeffding (1951)]. Hoeffding’s CLT has been refined over the years and the sharpest result can be found in Bolthausen (1984) where he proved a Berry-Esseen type bound for the convergence rate to normality.

Owen (1997a), page 1905, proves that an $n$-point Latin hypercube sample never leads to a sample mean variance greater than that of simple Monte Carlo with $n - 1$ points. This result can be sharpened when $d = 1$.

Finally Loh (1996a) generalized Owen’s CLT and Bolthausen’s Berry-Esseen bound to functions $f : [0, 1)^d \rightarrow \mathbb{R}^p$ for Latin hypercube sampling. He also proved

**Theorem 3** Suppose $\int_{[0,1)^d} \| f(t) \|^2 dt < \infty$ where $\| \cdot \|$ denotes the Euclidean norm on $\mathbb{R}^p$. Then

$$\| \hat{\mu}_{LHS} - \mu \| \rightarrow 0,$$

almost surely as $n \rightarrow \infty$. 

21
He further showed that asymptotically valid empirical likelihood confidence regions for $\mu$ can be constructed using Latin hypercube samples. Roughly speaking, the shape of the confidence region is the same as if the sample is i.i.d. except that the region gets shrunk towards $\mu$. More precisely, let

$$S = \{w : (w_1, \cdots, w_n)' : \sum_{k=1}^{n} w_k \leq 1, w_k \geq 0, \forall k\}.$$  

Define for $0 < r < 1$ and $f : [0, 1)^d \to \mathbb{R}^p$,

$$\Theta_{n,r} = \left\{ \sum_{k=1}^{n} w_k f(X_k) : w \in S, \prod_{k=1}^{n} n w_k \geq r \right\}.$$

**Theorem 4** Let $X_1, \cdots, X_n$ be a Latin hypercube sample, $\mu = Ef(X_1)$ and $\int_{[0,1)^d} r(t)r(t)'dt$ is non-singular. Then $\Theta_{n,r}$ is a convex set.

If $E\|f(X_1)\|^4 < \infty$, we have

$$\lim_{n \to \infty} P(\mu \in \Theta_{n,r}) \geq P(\chi^2_{(p)} \leq -2 \log r),$$

where $\chi^2_{(p)}$ denotes a chi-square random variable with $p$ degrees of freedom.

The proof is given in Loh (1996a), page 2068.

The technique of empirical likelihood was introduced by Owen (1988), (1990). It is a very promising
alternative technique to the use of the CLT as well as the bootstrap in the construction of confidence regions. A very accessible up to date account of empirical likelihood can be found in Owen (2001). A point of interest is that if $X_1, \ldots, X_n$ were an i.i.d. sample then Owen (1990) showed that

$$\lim_{n \to \infty} P(\mu \in \Theta_{n,r}) = P(\chi^2_p \leq -2 \log r).$$

It appears that Latin hypercube sampling is rather similar to simple (i.i.d.) random sampling in that for any probabilistic/statistical result on simple random sampling, there probably exists an analogous result for Latin hypercube sampling. A consequence of this is that statistical inference based on Latin hypercube sampling is usually feasible.

A somewhat negative point to note is that the properties of Latin hypercube sampling and simple random sampling may be too similar and hence it is unlikely that dramatic improvements (over simple random sampling) be achieved with the use of Latin hypercube sample. Possibly because of this, a number of researchers [see Tang (1992), Owen (1992b) and Ye (1998)] have started looking for refinements of Latin hypercube sampling. More of this later.
2.5.1 Variance estimation

This section discusses the the issues in estimating the asymptotic variance of $\hat{\mu}_{LHS}$. Recall that

$$f(X_i) = \mu + \sum_{j=1}^{d} \alpha_j(X_{i,j}) + r(X_i),$$

$$\text{Var}(\hat{\mu}_{LHS}) = \frac{1}{n} \int_{[0,1]^d} r(t)^2 dt + o\left(\frac{1}{n}\right),$$

and

$$\int_0^1 \text{Var}_{LHS}[f(X_i)|X_{i,j} = s] ds$$

$$= \sum_{k \neq j} \int_0^1 \alpha_k^2(t_k) dt_k + \int_{[0,1]^d} r^2(t) dt$$

$$= \sigma_j^2, \text{ say.}$$

Also observe that

$$\sum_{j=1}^{d} \sigma_j^2 = (d - 1) \sum_{j=1}^{d} \int_0^1 \alpha_j^2(t_j) dt_j + d \int_{[0,1]^d} r^2(t) dt$$

$$= n(d - 1) \text{Var}\hat{\mu}_{IID} + \int_{[0,1]^d} r^2(t) dt.$$
Hence to estimate \( \int r^2(t)dt \), it suffices to estimate \( \sigma_j^2, \, j = 1, \cdots, d \), and \( \text{Var}(\hat{\mu}_{IID}) \).

Assuming a certain degree of smoothness on \( \alpha_j(.) \), Owen (1992a) observed that a “standard” nonparametric regression of \( f(X_i) \) on \( X_{i,j} \) provides a consistent estimate of the function \( \mu + \alpha_j(.) \) [see e.g. Eubank (1988)].

Next define

\[
N_k f(X_i) = \begin{cases} 
N_k f(X_i) & \pi_{j,k} = \pi_{i,k} + 1 \text{ if } \pi_{i,k} < n, \\
0, & \text{if } \pi_{i,k} = n,
\end{cases}
\]

and

\[
\hat{\sigma}_j^2 = \frac{1}{2n} \sum_{i=1}^{n} [f(X_i) - N_k f(X_i)]^2.
\]

\( N_k f(X_i) \) is the right nearest neighbor to \( f(X_i) \) along the \( k \)th coordinate axis.
Theorem 5 Let $f : [0,1)^d \rightarrow \mathbb{R}$ be a bounded function and that
\[ |\alpha_k(u) - \alpha_k(v)| \leq C|u - v|, \]
for some constant $C > 0$. Then
\[ \hat{\sigma}_j^2 = \sigma_j^2 + O_p(n^{-1/2}), \]
as $n \rightarrow \infty$.

The proof can be found in Owen (1992a), page 547.

Finally since $(n - 1)^{-1} \sum_{i=1}^{n} [f(X_i) - \hat{\mu}_{LHS}]^2$ is a $n^{-1/2}$-consistent estimate of $n \operatorname{Var}(\hat{\mu}_{IID})$,
\[
\sum_{j=1}^{d} \hat{\sigma}_j^2 - \frac{d - 1}{n - 1} \sum_{i=1}^{n} [f(X_i) - \hat{\mu}_{LHS}]^2
= \int_{[0,1)^d} r^2(t) dt + O_p(n^{-1/2}),
\]
as $n \rightarrow \infty$. The l.h.s. of the above equation gives a $n^{-1/2}$-consistent estimate for the asymptotic variance of $n^{1/2}\hat{\mu}_{LHS}$. 

26
**Remark.** There is another popular variant of Latin hypercube sampling: centered Latin hypercube sampling. A centered Latin hypercube sample satisfies

\[
X_i = (X_{i,1}, \ldots, X_{i,d})',
\]

\[
X_{i,j} = \frac{\pi_j(i) - 0.5}{n}, \quad \forall 1 \leq i \leq n, 1 \leq j \leq d,
\]

where \(\pi_1, \ldots, \pi_d\) are independent uniform permutations of the set \(\{1, \ldots, n\}\). Centered Latin hypercube sampling goes back at least to Patterson (1954) who called them lattice samples.

Because of the centering, the "sample mean" based on this design is slightly biased. However most of the results discussed previously for Latin hypercube sampling apply to the centered version as well.
2.6 Randomized orthogonal arrays

Let $d$, $n$ and $t$ be positive integers with $t \leq d$. An orthogonal array of strength $t$ is a matrix of $n$ rows and $d$ columns with elements (or symbols) taken from the set $\{0, 1, \ldots, q - 1\}$ such that in any $n \times t$ sub-matrix, each of the $q^t$ possible rows occurs the same number of times. The class of all such arrays is denoted by $OA(n, d, q, t)$. An immediate consequence is that $n = \lambda q^t$ for some integer $\lambda$.

Orthogonal arrays can be traced back fifty years to Rao (1949). Very comprehensive treatments of orthogonal arrays can be found in Raghavarao (1971) and Hedayat, Sloane and Stufken (1999). In the context of experimental design, $n$ denotes sample size and $d$ the number of co-variates or the dimension of the sample space. The main advantage here is that if the design points from $OA(n, d, q, t)$ were to be projected onto a $t$ or smaller dimension space of co-variates, stratification will result.
Randomized orthogonal arrays were recently and independently introduced by Owen (1992b), (1994b) and Tang (1993). Let \( A \in OA(n, d, q, t) \). \( X_1, \ldots, X_n \in [0, 1)^d \) are design points from a randomized orthogonal array (generated by \( A \)) if for all \( 1 \leq i \leq n \),

\[ X_i = (X_{i,1}, \ldots, X_{i,d})', \]

\[ X_{i,j} = \frac{\pi_j(A_{i,j}) + U_{i,j}}{q}, \quad 1 \leq j \leq d. \]

Here as in Latin hypercube sampling, \( \pi_1, \ldots, \pi_d \) are independent permutations of the set \( \{0, \ldots, q - 1\} \), all \( q! \) permutations equally probable, and \( U_{i,j} \) are independent \( U[0, 1) \) random variables (being also independent of the \( \pi_j \)s).

An immediate consequence is that \( X_i \) is uniformly distributed on \( [0, 1)^d \) and hence the sample mean

\[ \hat{\mu}_{OAS} = \frac{1}{n} \sum_{i=1}^{n} f(X_i), \]

is an unbiased estimator for \( \mu \).
Note that Latin hypercube sampling corresponds to a randomized orthogonal array with strength \( t = 1 \) and \( n = q \). Hence randomized orthogonal arrays can be thought of as a natural generalization of Latin hypercube sampling.

Owen (1992b), page 446, noted that arrays of strength \( t > 2 \) require quite large sample sizes for modest \( q \) and hence would seem to be of less practical use at present. We observe that \( f \) can be expressed via an ANOVA-type decomposition as follows.

\[
f(X_i) = \mu + \sum_{j=1}^{d} \alpha_j(X_{i,j}) + \sum_{j,k; j<k} \alpha_{j,k}(X_{i,j}, X_{i,k}) + \cdots + \alpha_{1,\ldots,d}(X_{i,1}, \ldots, X_{i,d}),
\]

where

\[
\mu = \int_{[0,1]^d} f(t) dt,
\]

\[
\alpha_j(t_j) = \int_{[0,1]^{d-1}} [f(t) - \mu] \prod_{k \neq j} dt_k,
\]

\[
\alpha_{j,k}(t_j, t_k) = \int_{[0,1]^{d-2}} [f(t) - \alpha_j(t_j) - \alpha_k(t_k)] \prod_{l \neq j,k} dt_l,
\]

etc.
For $A \in OA(q^2, d, q, 2)$, Owen (1992b), page 448, states that
\[
\text{Var}(\hat{\mu}_{OAS}) = \frac{1}{n} \sum_{|u| > 2} \int_{[0,1)^d} \alpha_u^2 \prod_{j \in u} dt_j \left[1 + O(n^{-1/2})\right].
\]

Here $u$ denotes a generic subset of $\{1, \cdots, d\}$ and $|u|$ its cardinality.

Possibly due to the complex combinatorial nature of randomized orthogonal arrays, not much theoretical work on them have appeared in the literature. The only other result that I know of is

**Theorem 6** Let $A \in OA(q^2, 3, q, 2)$ and
\[
Ef(X_1)^4 < \infty.
\]

Then
\[
n^{1/2}(\hat{\mu}_{OAS} - \mu) \to N(0, \sum_{|u| > 2} \int_{[0,1)^d} \alpha_u^2 \prod_{j \in u} dt_j)\]

in distribution as $q \to \infty$.

The proof and a more detailed statement can be found in Loh (1996b). It should be noted that the above theorem applies to $d = 3$ only. For $d > 3$, the asymptotic normality of $\hat{\mu}_{OAS}$ is still an open problem.
2.6.1 Centered randomized orthogonal arrays

Owen (1992b) also proposed another variant of randomized orthogonal array. Again let

\[ A \in OA(n, d, q, t). \]

Then \( X_1, \ldots, X_n \) are design points from a centered randomized orthogonal array if

\[ X_{i,j} = \frac{\pi_j(A_{i,j}) + 0.5}{q}, \]

where the \( \pi_j \)s are independent permutations of \( \{0, \ldots, q - 1\} \), all \( q! \) permutations equally likely.

Unfortunately, there are certain drawbacks to this design. Firstly the sample mean is a biased estimator for \( \mu \). Secondly (and more seriously) as observed in Tang, page 1392, this design if projected onto a subspace of fewer than \( t \) co-variates will give rise to replication of points and that is undesirable in the context of computer experiments.
2.6.2 Variance estimation

This section discusses the estimation of
\[ \sum_{|u|>2} \int_{[0,1]^d} \alpha_u^2 \prod_{j \in u} dt_j = \lim_{n \to \infty} n \text{Var}(\hat{\mu}_{OAS}). \]

based on a centered randomized orthogonal array design with \( A \in OA(q^2, d, q, 2). \)

The approach as suggested in Owen (1992b), page 450, is similar to that of Latin hypercube sampling (except that notation gets to be messier). First define
\[ \hat{\alpha}_j(X_{i,j}) = (2M)^{-1} \sum_{m=\pm1, \ldots, \pm M} N_j^m f(X_i), \]

where for \( m > 0 (< 0), N_j^m f(X_i) \) is the value of \( f(X_i) \) observed \(|m| \) observations to the right (left) of \( f(X_i) \) along axis \( j \). If \( \alpha_j \) is Lipschitz continuous, for suitable values of \( M \), \( \hat{\alpha}_j \) is a consistent estimate of \( \mu + \alpha_j \).

Next observe that
\[ \int_0^1 \text{Var}_{OAS}[f(X_i)|X_{i,j} = s]ds \]
\[ = \int_{[0,1]^d} [f(t) - \mu - \alpha_j(t_j)]^2 dt \]
\[ = \tau_j^2, \quad \text{say.} \]
Owen further showed that
\[
\hat{\tau}_j^2 = \frac{1}{(1 + 2M)n} \sum_{i=1}^{n}[f(X_i) - \hat{\alpha}_j(X_{i,j})]^2
\]
\[
= \tau_j^2 + O_p(n^{-1/2}).
\]
For the above result, \( M = 1 \) is adequate.

Hence as in Latin hypercube sampling
\[
MSE = \sum_{j=1}^{d} \hat{\tau}_j^2 - \frac{d-1}{n-1} \sum_{i=1}^{n}[f(X_i) - \hat{\mu}_{OAS}]^2
\]
\[
= \int [f(t) - \mu - \sum_{j=1}^{d} \alpha_j(t_j)]^2 dt
\]
\[
+ O_p(n^{-1/2})
\]
\[
= \sum_{|u| \geq 2} \int_{[0,1]^d} \alpha_u^2 \prod_{j \in u} dt_j + O_p(n^{-1/2}).
\]
Using a nearest type neighbor estimator \( \hat{\alpha}_{j,k} \) (analogous to \( \hat{\alpha}_j \)), Owen (1992b), page 451, further argued that

\[
\hat{\tau}_{j,k}^2 = \frac{4}{5n} \sum_{i=1}^{n} [f(X_i) - \hat{\alpha}_{j,k}(X_i)]^2
\]

is a \( n^{-1/2} \)-consistent estimator for

\[
\sum_{u \not\in \{j,k\}} \int_{[0,1)^d} \alpha_u^2(x)dx.
\]

Finally a \( n^{-1/2} \)-consistent estimator of

\[
\sum_{|u| > 2} \int_{[0,1)^d} \alpha_u^2(x)dx
\]

is given by a suitable linear combination of \( \sum_{j<k} \hat{\tau}_{j,k}^2 \), \( MSE \) and

\[
\frac{1}{n-1} \sum_{i=1}^{n} [f(X_i) - \hat{\mu}_{OAS}]^2,
\]

as the latter is a \( n^{-1/2} \)-consistent estimator for

\[
\sum_{|u| \geq 1} \int_{[0,1)^d} \alpha_u^2(x)dx.
\]
2.6.3 OA-based Latin hypercubes

Tang (1993) proposed the use of orthogonal arrays in the construction of Latin hypercubes. Such designs [based on $OA(n, d, q, t)$, say] while preserving the univariate stratification properties of Latin hypercubes also stratify each $t$-covariate subspace.

More precisely, for $A \in OA(n, d, q, t)$, randomize its rows, columns and symbols to obtain a randomized orthogonal array, say $A^*$. Note that $A^*$ is a $n \times d$ matrix and $n = \lambda q^t$ for some integer $\lambda$.

Then for each column of $A^*$, replace the $\lambda q^{t-1}$ positions with entry $k$ by a random permutation (each such permutation having an equal probability of being chosen) of

$$\{k\lambda q^{t-1} + 1, k\lambda q^{t-1} + 2, \ldots, (k + 1)\lambda q^{t-1}\},$$

for all $k = 0, \ldots, q - 1$. This procedure generates a centered (random) $OA$-based Latin hypercube, say $U = (u_{i,j})$. 

36
Now suppose that $X_{i,j} \sim U[(i - 1)/n, i/n], i = 1, \ldots, n, j = 1, \ldots, d$, are all generated independently. Then the OA-based Latin hypercube to be used is

$$X_i = (X_{u_{i,1},1}, \ldots, X_{u_{i,d},d})', \quad i = 1, \ldots, n.$$  

It is easy to check that $X_i$ is uniformly distributed on $[0, 1)^d$ and hence

$$\hat{\mu}_{OLH} = \frac{1}{n} \sum_{i=1}^{n} f(X_i),$$

is an unbiased estimate of $\mu = E[f(X_i)]$.

Using an ANOVA-type decomposition, Tang (1993) showed that

$$f(X_i)$$

$$= \mu + \sum_{j=1}^{d} \alpha_j(X_{i,j}) + \sum_{j<k} \alpha_{i,j}(X_{i,j}, X_{i,k}) + r(X_i),$$

where

$$\alpha_j(X_{i,j}) = E[f(X_i)|X_{i,j}] - \mu,$$

$$\alpha_{j,k}(X_{i,j}, X_{i,k}) = E[f(X_i)|X_{i,j}, X_{i,k}]$$

$$- \mu - \alpha_j(X_{i,j}) - \alpha_k(X_{i,k}).$$
In particular, Tang proved

**Theorem 7** If \( f : [0, 1)^d \rightarrow \mathbb{R} \) is continuous, then under OA-based Latin hypercube sampling (with \( A \in OA(q^2, d, q, 2) \)),

\[
\text{Var}(\hat{\mu}_{OLH}) = \frac{1}{n} \text{Var}[r(X_1)] + o\left(\frac{1}{n}\right),
\]

as \( n \rightarrow \infty \) where

\[
\hat{\mu}_{OLH} = \frac{1}{n} \sum_{i=1}^{n} f(X_i).
\]

Theorem 7 and Owen’s result show that for \( A \in OA(q^2, d, q, 2) \),

\[
\lim_{n \rightarrow \infty} n \text{Var}(\hat{\mu}_{OAS}) \leq \lim_{n \rightarrow \infty} n \text{Var}(\hat{\mu}_{LHS}),
\]

\[
\lim_{n \rightarrow \infty} n \text{Var}(\hat{\mu}_{OLH}) \leq \lim_{n \rightarrow \infty} n \text{Var}(\hat{\mu}_{LHS}).
\]

However in general, the improvements are in the constants and the convergence rate is still \( n^{-1/2} \) [i.e. \( |\hat{\mu}_{OAS} - \mu| = O_p(n^{-1/2}) \), etc.].
Finally as a comparison between $\hat{\mu}_{OAS}$ and $\hat{\mu}_{OLH}$, Tang (1993), page 1397, showed that if $f$ is additive, i.e.,

$$f(X_i) = \mu + \sum_{j=1}^{d} \alpha_j(X_{i,j}),$$

the variance of $\hat{\mu}_{OLH}$ is smaller than the variance of $\hat{\mu}_{OAS}$.

Other than Tang (1993), there does not seem to be any other work on $OA$-based Latin hypercubes in the literature. This design appears to be very promising and certainly deserves more attention. The asymptotic distribution of $\hat{\mu}_{OLH}$ is not known. The obvious conjecture is that the distribution of $\hat{\mu}_{OLH}$ is asymptotically normal, but that remains to be proven.
2.7 Better Latin hypercubes

For each sample size $n$ and dimension $d$, there are $(n!)^d$ possible centered Latin hypercubes. While most of these Latin hypercubes possess good balance/uniformity properties, some of them can be quite structured. Much effort has been made to construct especially good Latin hypercubes. This sections reports on some of these efforts.

2.7.1 Sample correlation minimization

One popular approach is to find Latin hypercubes in which the $d$ input variables or covariates have small correlations. The following motivation is due to Owen (1994a). Recall that

$$f(X_i) = \mu + \sum_{j=1}^{d} \alpha_j(X_{i,j}) + r(X_i),$$

where

$$\mu = \int_{[0,1]^d} f(t)dt,$$

$$\alpha_j(t_j) = \int_{[0,1]^{d-1}} [f(t) - \mu] \prod_{k \neq j} dt_k.$$
\( r(X_i) \) can be further decomposed as
\[
r(X_i) = \sum_{j<k} \gamma_{j,k}(X_{i,j} - 0.5)(X_{i,k} - 0.5) + r^*(X_i),
\]
where
\[
\gamma_{j,k} = \frac{\int_{[0,1]^d} r(t)(t_j - 0.5)(t_k - 0.5)dt}{\int_{[0,1]^d} [(t_j - 0.5)(t_k - 0.5)]^2 dt}.
\]
The \( d(d - 1)/2 + 1 \) terms on the r.h.s. of the above equation are mutually orthogonal. In particular,
\[
E[r^*(X_i)(X_{i,j} - 0.5)(X_{i,k} - 0.5)] = 0.
\]
This implies that
\[
\hat{\mu}_{LHS} = \mu + \frac{1}{n} \sum_{j=1}^{d} \sum_{i=1}^{n} \alpha_j(X_{i,j}) \\
+ \sum_{j<k} \gamma_{j,k} \sigma_{j,k} + \frac{1}{n} \sum_{i=1}^{n} r^*(X_i),
\]
where
\[
\sigma_{j,k} = \frac{1}{n} \sum_{i=1}^{n} (X_{i,j} - 0.5)(X_{i,k} - 0.5).
\]
\( \sigma_{j,k} \) is close to the sample covariance between \( X_{i,j} \) and \( X_{i,k} \). Thus minimizing the correlations between the input variables would lead to

\[
\hat{\mu}_{LHS} - \mu \approx E[r^*(X_i)],
\]

which implies that in addition to the additive components \( \alpha_j \)'s, a sum of bilinear components is filtered out of the error.
**Ranked Cholesky algorithm**

Iman and Conover (1982) proposed a ranked Cholesky (RC) algorithm that can be applied to the problem at hand. More precisely, let \( n \times d \) matrix \( X \) denote a Latin hypercube sample. Then an \( n \times d \) matrix \( Z \) is generated in which each column \( Z_j, j = 1, \ldots, d, \) is a random permutation of \( \Phi^{-1}(i/(n+1)), i = 1, \ldots, n, \) where \( \Phi \) is the standard normal c.d.f. Let \( C \) be the \( d \times d \) sample covariance matrix of \( Z \). If \( \det(C) = 0 \), then another matrix \( Z \) is generated. Using the Cholesky decomposition, there exists a lower triangular matrix \( Q \) such that \( QQ' = C \). Let \( Z^* = Z(Q')^{-1} \) and hence the sample covariance matrix of \( Z^* \) is the \( d \times d \) identity matrix. Finally the elements in the column \( X_j \) are permuted so that the vector of ranks of \( X_j \) matches the vector of ranks of \( Z_j^*, j = 1, \ldots, d. \)

The upshot here is that the transformed \( n \times d \) matrix \( X^* \) of \( X \) is also a Latin hypercube with a rank correlation matrix equal to that of \( Z^* \).
The van der Warden scores $\Phi^{-1}(i/(n + 1))$, $i = 1, \ldots, n$ are used because they resemble values of normal random variables and the algorithm should perform well as least for multivariate normal random vectors.

The ranked Cholesky algorithm is very general. In fact the prime motivation for them is to generate random vectors with known (possibly non-normal) marginal distributions with a specified correlation matrix.
**Ranked Gram-Schmidt algorithm**

Owen (1994a) proposed the following ranked Gram-Schmidt (RGS) algorithm to decrease the sample correlation between input variables in a centered Latin hypercube sample (expressed as an $n \times d$ matrix $X$). The basic idea is to apply Gram-Schmidt orthogonalization to each of the input variables with a modification to preserve their Latin hypercube structure. There are two steps to the algorithm. The forward step is

$$
\text{for } j = 2, \cdots, d \\
\text{for } k = 1, \cdots, j - 1 \\
X_k \leftarrow X_k - (X_j - 0.5)\rho_{j,k}\sigma_k/\sigma_j \\
X_k \leftarrow (\text{rank}(X_k) - 0.5)/n,
$$

where $X_k$ denotes the $k$th column of $X$, $\text{rank}(X_k)$ is the vector of ranks of $X_j$, $\leftarrow$ denotes assignment, $\rho_{j,k}$ is the sample correlation between $X_j$ and $X_k$ before the update and $\sigma_k$ the sample standard deviation of the elements of $X_k$. The third statement of the forward step replaces $X_k$ by the residual obtained from the linear regression (including the intercept) of the vector $X_k$ on the vector $X_j$.
The backward step is
\[
\text{for } j = d - 1, \cdots, 1 \\
\text{for } k = d, \cdots, j + 1 \\
X_k \leftarrow X_k - (X_j - 0.5)\rho_{j,k}\sigma_k/\sigma_j \\
X_k \leftarrow (\text{rank}(X_k) - 0.5)/n,
\]

The complete algorithm proceeds by alternating the forward and backward steps.

The basic idea behind the above algorithm is that for fixed $j, k$, we want to update $X_k$ such that the new $X_k$ has only a small correlation with $X_j$. Since the residual from a linear regression of $X_k$ on $X_j$ has zero correlation with $X_j$, we update $X_k$ with the vector of residuals. Then renormalize $X_k$ using $\text{rank}(X_k)$ to ensure that the final updated $X_k$ has the stratification of a Latin hypercube.
Owen reported a simulation study of the performance of the RC and RGS algorithms. As a performance measure, the root mean square correlation $\rho_{rms}^2(X)$ among the columns of $X$ is used where

$$\rho_{rms}^2(X) = \frac{\sum_{j=2}^{d} \sum_{k=1}^{j-1} \rho_{j,k}^2}{d(d-1)/2}. $$

The algorithms were applied with

$$n \in \{10, 20, 30, 100, 150, 250, 500\}$$

and $d = n - 1$. There were four replicates for each $n$.

The study concludes that there is very little variation in $\rho_{rms}$ for each method and value of $n$. The RC algorithm reduces $\rho_{rms}$ by roughly a factor of 3 (relative to Latin hypercube sampling) while $E_{RGS}\rho_{rms} \approx O(n^{-3/2})$ [relative to $E_{LHS}\rho_{rms} = O(n^{-1/2})$ ].

Owen through a heuristic argument conjectures that it may be possible by some other algorithm to achieve $\rho_{rms} = O(n^{-5/2})$. 

47
**Quadratic canonical correlation**

Tang (1998) gave the following motivation for minimizing the correlation between input variables for Latin hypercube samples. Consider the main effect model

\[
Y_i = \bar{Y} + \sum_{j=1}^{d} \beta_{j,1}(X_{i,j} - \bar{X}_j) + 
\cdots + \beta_{j,q}(X_{i,j}^q - \bar{X}_j^q)] + \varepsilon_i,
\]

for \( i = 1, \ldots, n \) where

\[
\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i,
\]

\[
\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{i,j},
\]

\[
\cdots \cdots
\]

\[
\bar{X}_j^q = \frac{1}{n} \sum_{i=1}^{n} X_{i,j}^q.
\]

The full design matrix is given by

\[
(1, D_1, \ldots, D_d),
\]

where \( 1 \) is the \( n \times 1 \) vector of 1s, and the \( s \)th column of matrix \( D_j \) is given by the vector \((X_{1,j}^s - \bar{X}_j^s, \ldots, X_{n,j}^s - \bar{X}_j^s)'\), \( s = 1, \ldots, q \).
It is desirable that the main effects \( \beta_1, \cdots, \beta_d \) can be uncorrelatedly estimated (this thus not exclude the possibility that the estimates of the components of each \( \beta_j \) be correlated). This can be achieved if \( D_1, \cdots, D_d \) are mutually orthogonal, i.e.,

\[
(D_j)'D_k = 0, \quad \text{for } 1 \leq j < k \leq d.
\]

However for a Latin hypercube design, the above equation can almost never be strictly satisfied. Thus there is a need to find a Latin hypercube sample such that the input variables are as uncorrelated as possible. To assess the degree of correlation between two vectors, Tang proposed the use of the polynomial canonical correlation coefficient.

**Definition.** Given two sets of vectors \( U = ((U_1)', \cdots, (U_p)')' \) and \( V = ((V_1)', \cdots, (V_q)')' \), where \( U_i \) and \( V_j \) are \( n \times 1 \) vectors, the canonical correlation is the maximal correlation \( \rho(U, V) \) between linear combinations of \( U \) and linear combinations of \( V \), i.e.,

\[
\rho(U, V) = \max_{a, b \in \mathbb{R}^n} \frac{\text{Cov}(a'U, b'V)}{\sqrt{\text{Var}(a'U)}\sqrt{\text{Var}(b'V)}}.
\]
Now for two \( n \times 1 \) vectors \( U^{(1)} = (u_1, \cdots, u_n)' \) and \( V^{(1)} = (v_1, \cdots, v_n)' \), let
\[
U^{(j)} = (u_1^j, \cdots, u_n^j)',
\]
\[
V^{(j)} = (v_1^j, \cdots, v_n^j)'.
\]

The canonical correlation between \( ((U^{(1)})', \cdots, (U^{(q)})')' \) and \( ((V^{(1)})', \cdots, (V^{(q)})')' \), is called the polynomial canonical correlation of order \( q \) between the vectors \( U^{(1)} \) and \( V^{(1)} \). It can be shown that if the polynomial canonical correlation of order \( q \) between the two vectors \( (X_{1,j}, \cdots, X_{n,j})' \) and \( (X_{1,k}, \cdots, X_{n,k})' \) is zero, then \( (D_j)'D_k = 0 \). Thus it is desirable to find a Latin hypercube \( X \) (in the form of an \( n \times d \) matrix) in which for any two columns, the polynomial canonical correlation of order \( q \) is minimized.
Tang (1998) proposed the following algorithm for reducing the quadratic (i.e., \( q = 2 \)) canonical correlations of a centered Latin hypercube sample. Following the notation of Tang, let a Latin hypercube be denoted by an \( n \times d \) matrix \( X \), \( X_k \) the \( k \)th column of \( X \) and \( X_k \) is a permutation of the set \( \{1, \cdots, n\} \). As in Owen’s RGS algorithm, there are two steps in Tang’s algorithm. The forward step is

\[
\text{for } j = 1, \cdots, d - 1 \\
\quad \text{for } k = j + 1, \cdots, d \\
\quad \quad X_k \leftarrow \text{rank}\left(\text{Res}\left( X_k, X_j \right)\right),
\]

where \( \text{Res}(X_k, X_j) \) denotes the vector of residuals from the regression

\[
X_k = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \varepsilon.
\]

The backward step is

\[
\text{for } j = d, \cdots, 2 \\
\quad \text{for } k = j - 1, \cdots, 1 \\
\quad \quad X_k \leftarrow \text{rank}\left(\text{Res}\left( X_k, X_j \right)\right).
\]

The full algorithm alternates between the forward and backward steps.
Tang reported that after a number of iterations, one of the following three situations must occur.

- the same design always shows up (esp. when $d$ is small relative to $n$),
- the $\rho$ values stabilize and fluctuate around a certain value,
- two designs show up alternating between the forward and backward steps. However this situation is quite rare.

In any case, $\rho$ decreases rapidly in the first few iterations and almost always stabilize within ten iterations.

For small $n$ and $d$, initial designs have a large effect on the final result. As $n$ and $d$ get larger, the effect of using different initial designs tends to diminish.

A comparison with Owen’s RGS algorithm was done. It appears that the RGS algorithm does not have too much control on the quadratic canonical correlations. However in reducing linear canonical correlations, Tang reported that the RGS algorithm generally does better.
2.7.2 Orthogonal Latin hypercubes

Let the $n \times d$ matrix $X$ denote a Latin hypercube. An orthogonal Latin hypercube is a Latin hypercube in which the columns of $X$ are uncorrelated. Ye (1998) recently exhibited a way of constructing orthogonal Latin hypercubes when $n = 2^m + 1, d = 2m - 2$ and $n = 2^m, d = 2m - 2$ for $m \geq 2$.

We shall illustrate Ye’s construction for $n = 2^m + 1, d = 2m - 2, m \geq 2$. For simplicity and clarity, denote the $n$ levels of each column of $X$ by

$$\{-2^{m-1}, \ldots, -1, 0, 1, \ldots, 2^{m-1}\}.$$  

This is valid because the sample correlation is invariant to linear transformations. Hence two columns, $X_j, X_k$ have zero correlation if they are orthogonal (i.e., $(X_j)'X_k = 0$) since elements from each column sum to zero.

Denote the top half of the Latin hypercube by $T$, which is an $2^{m-1} \times (2m - 2)$ matrix. The bottom half of $X$ is $-T$ and the center row consists of 0s.
**Definition** The $M$ matrix is the matrix of which each entry is the absolute value of the corresponding entry in $T$. The $S$ matrix is the matrix of which each entry is taken as 1 or $-1$ according to whether the corresponding entry in $T$ is positive or negative.

Thus, $T$ is the elementwise product (or Hadamard product) of $M$ and $S$. To construct $M$, define

$$A_k = \underbrace{I \otimes \cdots \otimes I}_{m-1-k} \otimes R \otimes \cdots \otimes R,$$

where $I$ is the $2 \times 2$ identity matrix, $\otimes$ the Kronecker product and

$$R = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Writing $e = (1, \cdots, 2^{m-1})'$, the columns of $M$ are permutations of $\{1, \cdots, 2^{m-1}\}$. They consist of

$$\{e, A_i e, A_{m-1} A_j e : i = 1, \cdots, m-1, j = 1, \cdots, m-2\}.$$
Next to construct $S$, we define the $2^{m-1} \times 1$ vector $a_k$ as

$$a_k = B_1 \otimes \cdots \otimes B_{m-1},$$

where

$$B_k = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad B_i = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

for $i \neq k$. The columns of $S$ are chosen to be

$$\{1, a_i, a_1a_{j+1} : i = 1, \cdots, m-1, j = 1, \cdots, m-2\}.$$

Finally $T$ is the Hadamard product of $T$ and $S$ and the columns of $T$ consist of

$$\{1 \times e, a_i \times A_i e, (a_1a_j) \times (A_jA_{m-1}e) : i = 1, \cdots, m-1, j = 1, \cdots, m-2\},$$

where ‘$\times$’ is the elementwise product. This completes the construction of an $2^m + 1 \times 2m - 2$ orthogonal Latin hypercube.

**Remark** The proof that the above construction indeed gives an orthogonal Latin hypercube can be found in Ye (1998).
For the case $n = 2^m$ and $d = 2m - 2$, Ye noted that an $2^m + 1 \times 2m - 2$ orthogonal Latin hypercube be first constructed, remove its center row of 0s and then rescale the levels to be equidistant. The resulting $2^m \times 2m - 2$ matrix is the desired orthogonal Latin hypercube.
2.7.2 Symmetric Latin hypercubes

Ye, Li and Sudjianto (2000) propose a class of symmetric Latin hypercube designs. A Latin hypercube is said to be a symmetric Latin hypercube (SLH) if it has the following property: in an $n \times d$ Latin hypercube with levels 1 to $n$, if $(a_1, \cdots, a_d)$ is one of its rows, then $(n + 1 - a_1, \cdots, n + 1 - a_d)$ must be another one of its rows. Intuitively, this implies that if $X_i$ is a design point in a SLH, then there exists another point $X_j$ in the design which is a reflection of $X_i$ through the center of the design space. Note that the definition of a Latin hypercube used here is an $n \times d$ matrix where each column is a permutation of $\{1, \cdots, n\}$.

Ye, Li and Sudjianto (2000) reported that

• the SLHs forms a good subset of Latin hypercubes with respect to both entropy and maximin distance criteria.
• SLHs generalize orthogonal Latin hypercubes in that they retain some orthogonality properties, e.g., the estimation of quadratic effects and bilinear interactions is uncorrelated with the estimation of linear effects.

• the \( CP \) algorithm described in Ye, et. al. (2000) greatly reduces the search time for optimal SLHs.

• the global optimal Latin hypercube is not always an SLH. Morris and Mitchell (1995) conducted an exhaustive search to find optimal Latin hypercubes and not all optimal Latin hypercubes are SLHs.
2.7.3 Optimal LHDs via IMSE and entropy

Park (1994) considered optimal Latin hypercube designs that minimize IMSE as well as Latin hypercube designs that maximize entropy. The model assumed is

\[ Y(x) = \beta + Z(x), \]

where \( Z(.) \) is a Gaussian random field with mean 0 and covariance function

\[ \text{Cov}(t, u) = \sigma^2 \exp[-\theta \sum_{j=1}^{d} (t_j - u_j)^2]. \]

\( \beta \) is unknown for the IMSE criteria and is assumed known for the entropy criteria.

Park reported that the (approximately) optimal designs turn out to be well spread out and frequently almost symmetric.
2.7.4 Optimal LHDs via a maximin criterion

Morris and Mitchell (1995) considered the construction of optimal Latin hypercube with the aim of minimizing a certain maximin distance criterion. The motivation for doing so is to come up with designs that are good in situations of ‘effect sparsity’ as well as in situations where all or most of the input variables are important.

The maximin distance criterion used is as follows. Given a distance measure and design, define a distance list \((d_1, \ldots, d_m)\) in which the elements are the distinct values of inter-site distances, sorted from smallest to largest. Also define the index list \((J_1, \ldots, J_m)\) in which \(J_i\) is the number of pairs of sites in the design separated by distance \(d_i\).
Morris and Mitchell defined a design $D$ to be a maximin design if among available designs, it

(1a) maximizes $d_1$, and among designs for which this is true,

(1b) minimizes $J_1$, and among designs for which this is true,

(2a) maximizes $d_2$, and among designs for which this is true,

(2b) minimizes $J_2$, and among designs for which this is true,

::

(ma) maximizes $d_m$, and among designs for which this is true,

(mb) minimizes $J_m$.

Statements (1a) and (1b) give the ‘usual definition’ of maximin criterion. The more elaborate statements 2(a), 2(b), $\cdots$, (mb), help break ties (if any) among optimal designs.
A real-valued design criterion function $\phi_p$ is used as a proxy for the above maximin criterion where for a given design $D$, 

$$\phi_p(D) = \left[ \sum_{i=1}^{d} J_i d_i^{-p} \right]^{1/p}.$$ 

Note that for sufficiently large $p$, each term in the sum on the r.h.s. of the above equation dominates all subsequent terms. Hence a design that minimize $\phi_p$ is also a maximin design.

Morris and Mitchell described a simulated annealing algorithm for finding the optimal design among Latin hypercubes.
2.7.5 Resolution-type Latin hypercube designs

In this section, it is convenient to define a Latin hypercube to be an $n \times d$ matrix $X$ where the columns are permutations of $\{1, \cdots, n\}$. Butler (2001) constructed centered Latin hypercube designs under three different models, namely a full second-order model, a first-order model and a second-order main effects model. The full second-order model usually considered is the second-order polynomial model

$$f(x_i) = \mu + \sum_{j=1}^{d} \alpha_j x_{i,j} + \sum_{j=1}^{d} \gamma_j x_{i,j}^2 + \sum_{j=1}^{d-1} \sum_{k=j+1}^{d} \beta_{j,k} x_{i,j} x_{i,k} + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$. Note that, although observations in computer experiments are observed without error, the error term $\varepsilon$ is required in order to model higher-order systematic effects.
In the context of Latin hypercube sampling, Butler observed that it is more natural to consider the following closely related second-order Fourier polynomial model

\[
f(x_i) = \mu - \sqrt{2} \sum_{j=1}^{d} \alpha_j \cos\left\{ \frac{\pi(x_{i,j} - 0.5)}{n} \right\} + \sqrt{2} \sum_{j=1}^{d} \gamma_i \cos\left\{ \frac{2\pi(x_{i,j} - 0.5)}{n} \right\} + 2 \sum_{j=1}^{d-1} \sum_{k=j+1}^{d} \beta_{j,k} \cos\left\{ \frac{\pi(x_{i,j} - 0.5)}{n} \right\} \times \cos\left\{ \frac{\pi(x_{i,k} - 0.5)}{n} \right\} + \varepsilon.
\]

To see the similarities between the above two models, define \( p_1(t) = t - n/2 \) as the linear function of \( t \) that is orthogonal to a constant function on \([0, n]\) and \( p_2(t) = t^2 - nt + n^2/6 \) as the quadratic function of \( t \) orthogonal on \([0, n]\) to both a constant and \( p_1(.) \).
Then
\[
\int_0^n -p_1(t) \cos \left( \frac{\pi t}{n} \right) dt
\]
\[
= 0.993 \left\{ \int_0^n p_1(t)^2 dt \right\}^{1/2} \left\{ \int_0^n \cos^2 \left( \frac{\pi t}{n} \right) dt \right\}^{1/2},
\]
\[
\int_0^n p_2(t) \cos \left( \frac{2\pi t}{n} \right) dt
\]
\[
= 0.961 \left\{ \int_0^n p_2(t)^2 dt \right\}^{1/2} \left\{ \int_0^n \cos^2 \left( \frac{2\pi t}{n} \right) dt \right\}^{1/2}.
\]
In addition the inner product of the interactions of the linear effects equals \((0.993)^2 = 0.986\) under normalization.
We observe that the full second-order Fourier polynomial model can be more compactly expressed in linear model form as

\[ Y = Z\theta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n), \]

where \( Y \) is the \( n \times 1 \) vector of observations and

\[
\theta = (\mu, \alpha_1, \cdots, \alpha_d, \gamma_1, \cdots, \gamma_d, \\
\beta_{1,2}, \cdots, \beta_{d-1,d})'
\]

is the vector of model parameters.

\[ Z = (1_n, Z_L, Z_Q, Z_I) \]

is the corresponding design matrix with

\[
(Z_L)_{i,j} = -\sqrt{2} \cos \left\{ \frac{\pi (X_{i,j} - 0.5)}{n} \right\},
\]

\[
(Z_Q)_{i,j} = \sqrt{2} \cos \left\{ \frac{2\pi (X_{i,j} - 0.5)}{n} \right\},
\]

\[
(Z_I)_{i,(j,k)} = 2 \cos \left\{ \frac{\pi (X_{i,j} - 0.5)}{n} \right\} \cos \left\{ \frac{\pi (X_{i,k} - 0.5)}{n} \right\},
\]

for \( i = 1, \cdots, n, \, 1 \leq j, k \leq d \) and \( (j, k) = k + d(j - 1) - j(j + 1)/2 \) for \( j < k \).
We observe that $Z_L$ is the $n \times d$ design matrix of the Fourier linear effects, $Z_Q$ is the $n \times d$ design matrix of the Fourier quadratic effects, and $Z_I$ is the $n \times d(d - 1)/2$ design matrix of the interactions between the Fourier linear effects.

An important point to note is that for a Latin hypercube design, $Z_L$ and $Z_Q$ automatically satisfy the constraints

$$1_n' Z_L = 0_d',$$
$$1_n' Z_Q = 0_d'.$$

Hence all Fourier linear effects and quadratic effects are orthogonal to the overall main effect.

Other Fourier polynomial models

A first-order Fourier polynomial model is similarly defined by (2) with

$$\theta = (\mu, \alpha_1, \cdots, \alpha_d)',$$
$$Z = (1_n, Z_L).$$

This model is also closely related to the (more customary) first-order polynomial model.
Finally a second-order main effects Fourier polynomial model is as in (2) with
\[
\theta = (\mu, \alpha_1, \cdots, \alpha_d, \gamma_1, \cdots, \gamma_d)',
\]
\[
Z = (1_n, Z_L, Z_Q).
\]
This model is useful when there are relatively few active interactions compared with the main effects.

**Resolution-type Latin hypercube construction**

The designs will be constructed using Williams transformation [Williams (1949)] which is a one-to-one transformation between \( n \) one-dimensional locations \( x = 1, \cdots, n \) and \( n \) codes \( w = 0, \cdots, n - 1 \). The transformation is as follows

\[
x(w) = \begin{cases} 
2w + 1 & \text{if } w < n/2, \\
2(n - w) & \text{if } w \geq n/2.
\end{cases}
\]
Let \( n_0 = (n - 1)/2 \). For \( d \leq n_0 \), the \( n \times d \) design matrix \( X \) of the design \( D_n(g_1, \cdots, g_d) \) has elements \( X_{i,j} = x(w_{i,j}) \) for \( i = 1, \cdots, n \) and \( j = 1, \cdots, d \), where \( w_{i,j} \in \{0, \cdots, n - 1\} \) and

\[
  w_{i,j} = \begin{cases} 
    ig_j + (n - 1)/4 \mod n, & \text{for } n = 1 \mod 4, \\
    ig_j + (3n - 1)/4 \mod n, & \text{for } n = 3 \mod 4.
  \end{cases}
\]

**Remarks**

- The generators \( g_j \) are required to be distinct elements in the set \( \{1, \cdots, n_0\} \).

- The constants, \((n - 1)/4\) for \( n = 1 \mod 4 \) and \((3n - 1)/4\) for \( n = 3 \mod 4 \), ensure that the design contains the center point of the design space.

- It is assumed that \( n \) is a prime number as this guarantees that the resulting design is a Latin hypercube.
For \( d = n_0 + r \) with \( 1 \leq r \leq n_0 \), another type of design \( E_n(g_1, \cdots, g_r) \) has \( n \times d \) design matrix \( X = (X_0, X_1) \), where \( X_0 \) is the \( n \times n_0 \) design matrix \( D_n(1, 2, \cdots, n_0) \) and \( X_1 \) is an \( n \times r \) design matrix with elements

\[
(X_1)_{i,j} = x(w_{i,j}),
\]

\[
w_{i,j} = ig_j \text{ mod } n, \quad w_{i,j} \in \{0, 1, \cdots, n - 1\},
\]

for \( i = 1, \cdots, n \) and \( j = 1, \cdots, r \). Again, it is assumed that the generators \( g_1, \cdots, g_r \) of \( X_1 \) are distinct elements in \( \{1, \cdots, n_0\} \) and that \( n \) is a prime number.

**Further definitions**

A Latin hypercube design is said to have resolution III if the Fourier linear effects are mutually orthogonal, i.e., \( Z_L'Z_L = nI_d \).
A Latin hypercube design is said to have resolution IV if the Fourier linear effects are mutually orthogonal and orthogonal to all second order effects and the overall mean is orthogonal to all second order interactions, i.e.,

\[ Z'_L Z_L = n I_d, \]
\[ Z'_L Z_Q = 0, \]
\[ Z'_L Z_I = 0, \]
\[ 1'_n Z_I = 0. \]

Finally, a Latin hypercube design is said to have resolution V if all the effects in the full second order Fourier polynomial model are orthogonal, i.e., \( Z'Z \) is a diagonal matrix.

The above definitions are motivated by analogy with two-level factorial designs [see for example Chapter 4 of Wu and Hamada (2000)].
Some theoretical results

**Theorem 8** For $n$ prime, $D_n(1, 2, \cdots, n_0)$ is a resolution IV Latin hypercube design.

Butler (2001), page 853, noted that $D_n(g_1, \cdots, g_d)$ performs extremely well under the linear correlation criterion of Iman and Conover (1982) and the quadratic canonical correlation criterion of Tang (1998).

**Theorem 9** For $n$ prime, $D_n(g_1, \cdots, g_d)$ is a resolution V Latin hypercube design if all the elements

\[
\min(2g_j, n - 2g_j), \quad 1 \leq j \leq d,
\]

\[
|g_j - g_k|, \quad 1 \leq j < k \leq d,
\]

\[
\min(g_j + g_k, n - g_j - g_k), \quad 1 \leq j < k \leq d,
\]

are distinct.
Theorem 10 For the design $E_n(1, 2, \cdots, n_0)$,

$$(Z_L)_{i,j} = (-1)^k \sqrt{2} \sin \left( \frac{2\pi ij}{n} \right),$$

$$(Z_L)_{i,n_0+j} = -\sqrt{2} \cos \left( \frac{2\pi ij}{n} + \frac{\pi}{2n} \right),$$

for $i = 1, 2, \cdots, n$ and $1 \leq j \leq n_0$, where

$$k = \begin{cases} 
0, & \text{if } n = 1 \mod 4, \\
1, & \text{if } n = 3 \mod 4.
\end{cases}$$

Consequently, $n^{-1}Z'_LZ_L = I_{n-1} + P$, where $P$ is an $(n - 1) \times (n - 1)$ matrix with nonzero elements

$$P_{j,n_0+j} = P_{n_0+j,j} = (-1)^k \sin \left( \frac{\pi}{2n} \right), \quad 1 \leq j \leq n_0.$$ \hspace{1cm} \hspace{1cm} \hspace{1cm}

When $n$ is large, the elements of $P$ are close to zero. Hence $E_n(1, 2, \cdots, n_0)$ is said to be a near resolution III Latin hypercube design.

The proofs of Theorems 8, 9 and 10 are found in Butler (2001).
2.8 Nets and Scrambled Nets

Numerical analysts and number theorists have spent an enormous amount of effort in the design of points that are equidistributed in space, in particular in the $s$-dimensional unit hypercube $[0, 1)^s$. One of the most successful and popular is the class of designs known as $(t, m, s)$-nets. A comprehensive account of $(t, m, s)$-nets is given in Niederreiter (1992). Note that in this section, we shall follow the customary notation of nets in that the dimension of the unit hypercube is taken to be $s$ (and not $d$).

Let $s \geq 1$ and $b \geq 2$ be integers. An elementary interval in base $b$ is a subset of $[0, 1)^s$ of the form

$$\mathcal{E} = \prod_{j=1}^{s} \left[ \frac{c_j}{b^{k_j}}, \frac{c_j + 1}{b^{k_j}} \right),$$

for integers $c_j, k_j$ with $k_j \geq 0$ and $0 \leq c_j \leq b^{k_j} - 1$. 
**Definition.** Let \( 0 \leq t \leq m \) be integers. A finite sequence \( \{ A_i : i = 1, \cdots, b^m \} \) of points from \([0, 1)^s\) is a \((t, m, s)\)-net in base \(b\) if every elementary interval \( \mathcal{E} \) in base \(b\) of \(s\)-dimensional Lebesgue measure \(b^{t-m}\) satisfies
\[
\sum_{i=1}^{b^m} I\{A_i \in \mathcal{E}\} = b^t.
\]

**Definition.** For \(t \geq 0\), an infinite sequence \( \{ A_i : i = 1, 2, \cdots \} \) of points from \([0, 1)^s\) is a \((t, s)\)-sequence in base \(b\) if for all integers \(k \geq 0\) and \(m \geq t\), the finite sequence \( \{ A_i : i = kb^m + 1, \cdots, (k+1)b^m \} \) is a \((t, m, s)\)-net in base \(b\).

The usual \((t, m, s)\)-net or \((t, s)\)-sequence estimate for \(\mu = \int_{[0,1)^s} f(x)dx\) is given by the “sample mean”
\[
\hat{\mu}_{tms} = \frac{1}{n} \sum_{i=1}^{n} f(A_i).
\]
Tables for the construction of \((t, m, s)\)-nets and \((t, s)\)-sequences can be found in Mullen, Mahalanabis and Niederreiter (1995) and Clayman, Lawrence, Mullen, Niederreiter and Sloan (1999).

Next we need a concept of total variation for functions of several variables. For a function \(f\) on \([0, 1]^s\) and a subinterval \(J\) of \([0, 1]^s\), let \(\Delta(f; J)\) be an alternating sum of the values of \(f\) at the vertices of \(J\) (i.e., function values at adjacent vertices have opposite signs). The variation of \(f\) on \([0, 1]^s\) in the sense of Vitali is defined by

\[
V^{(s)}(f) = \sup_{\mathcal{P}} \sum_{J \in \mathcal{P}} |\Delta(f; J)|,
\]

where the supremum is extended over all partitions \(\mathcal{P}\) of \([0, 1]^s\) into subintervals.

A more convenient formula

\[
V^{(s)}(f) = \int_0^1 \cdots \int_0^1 \left| \frac{\partial^s f}{\partial u_1 \cdots \partial u_s} \right| du_1 \cdots du_s
\]

holds whenever the indicated mixed partial derivative is continuous on \([0, 1]^s\).
For $1 \leq k \leq s$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq s$, let $V^{(k)}(f; i_1, \cdots, i_k)$ be the variation in the sense of Vitali of the restriction of $f$ to the $k$-dimensional face 
\[
\{(u_1, \cdots, u_s) \in [0, 1]^s : u_j = 1 \text{ for } j \neq i_1, \cdots, i_k\}.
\]
Then
\[
V(f) = \sum_{k=1}^{s} \sum_{1 \leq i_1 < \cdots < i_k \leq s} V^{(k)}(f; i_1, \cdots, i_k)
\]
is called the variation of $f$ on $[0, 1]^s$ in the sense of Hardy and Krause, and $f$ is of bounded variation in this sense if $V(f)$ is finite.

We observe that when $s = 1$, bounded variation in the sense of Hardy and Krause reduces to the usual bounded total variation (or equivalently $f$ can be expressed as a difference of two monotone functions).
The next result (very well known to numerical analysts) is of crucial importance in what follows. If $f$ is of finite total variation in the sense of Hardy and Krause, it follows from the Koksma-Hlawka inequality that as $n \to \infty$,

$$|\hat{\mu}_{tms} - \mu| = O((\log_b n)^{s-1}/n)$$

if \{A_i : i = 1, \cdots, n\} is a $(t, m, s)$-net in base $b$ and

$$|\hat{\mu}_{tms} - \mu| = O((\log_b n)^s/n)$$

if \{A_i : i \geq 1\} is a $(t, s)$-sequence in base $b$.

The proofs of the above statements are given in Chapters 2 and 4 of Niederreiter (1992).

**Remark.** Observe that the above error bounds, namely $O((\log_b n)^{s-1}/n)$ and $O((\log_b n)^s/n)$ are deterministic (or worst case scenario) error bounds.
Furthermore it is clear that asymptotically, $1/\sqrt{n}$ is bigger than $(\log n)^{s-1}/n$. However if $s$ is large, say $s = b = 10$, the asymptotics do not take effect until $n \geq 10^{26}$ as only then will

$$(\log_b n)^{s-1}/n \leq 1/\sqrt{n}.$$ 

As a result these error bounds may not have too much practical significance under such circumstances.
2.8.1 Scrambled nets and sequences

In 1995, Owen introduced the idea of scrambled \((t, m, s)\)-nets as follows. Suppose that

\[ \{ A_i = (A_{i,1}, \cdots, A_{i,s})' : i = 1, \cdots, b^m \} \]

is a \((t, m, s)\)-net in base \(b\). We observe that \(A_{i,j}\) can be expressed as

\[
A_{i,j} = \sum_{k=1}^{\infty} a_{i,j,k} b^{-k}
\]

for suitable integers \(0 \leq a_{i,j,k} \leq b - 1\). Let

\[
\{ \pi_j, \pi_{j;a_1}, \pi_{j;a_1,a_2}, \cdots : 1 \leq j \leq s, \quad 0 \leq a_k \leq b - 1, k = 1, 2, \cdots \}
\]

be a set of mutually independent random permutations of \(\{0, 1, \cdots, b - 1\}\) where each of these permutations is uniformly distributed over its \(b!\) possible values.
Now a scrambled \((t, m, s)\)-net in base \(b\) has the form
\[
\{X_i = (X_{i,1}, \ldots, X_{i,s})' : i = 1, \ldots, b^m\}
\]
where
\[
X_{i,j} = \sum_{k=1}^{\infty} x_{i,j,k} b^{-k},
\]
and for \(1 \leq i \leq b^m, 1 \leq j \leq s,\)
\[
x_{i,j,1} = \pi_j(a_{i,j,1}),
\]
\[
x_{i,j,k} = \pi_{j;a_{i,j,1},\ldots,a_{i,j,k-1}}(a_{i,j,k}), \quad \forall k \geq 2.
\]

Owen further showed that
\[
\{X_i : i = 1, \ldots, b^m\}
\]
is also a \((t, m, s)\)-net in base \(b\) with probability 1 and that for each \(i\), \(X_i\) has the uniform distribution on \([0, 1)^s\).

**Remark.** The same randomization algorithm can be applied to \((t, s)\)-sequences to obtain analogous scrambled \((t, s)\)-sequences.
It should be noted that Owen’s scrambling is done independently for each of the $s$ coordinates of $[0, 1)^s$. It is not clear whether that is excessive randomization (e.g., one can instead choose a coordinate at random and just scrambled that coordinate leaving the remaining coordinates unchanged).

The estimator for $\mu$ that we are concerned with based on the scrambled $(t, m, s)$-net

$$\{X_i : i = 1, \ldots, b^m\}$$

is the sample mean

$$\hat{\mu}_{t,m,s} = \frac{1}{b^m} \sum_{i=1}^{b^m} f(X_i).$$

Since $X_i$ is uniformly distributed over $[0, 1)^s$, $\hat{\mu}_{t,m,s}$ is an unbiased estimator for $\mu$.

We have noted that the deterministic Koksma-Hlawka inequality (worst case) bounds may be too large in some circumstances for practical purposes. One motivation for introducing scrambled nets is the hope that the average case bounds (e.g., standard deviation of $\hat{\mu}_{tms}$) that result may be significantly sharper.
2.8.2 Variance of scrambled nets

**Definition.** A real-valued function $f$ on $[0, 1)^s$ is smooth if there exist finite constants $B \geq 0$ and $\beta \in (0, 1]$ such that

$$\left| \frac{\partial^s}{\partial x_1 \cdots \partial x_s} f(x) - \frac{\partial^s}{\partial x_1 \cdots \partial x_s} f(y) \right| \leq B \|x - y\|_s^{\beta}, \quad \forall x, y \in [0, 1)^s.$$  

Writing $\text{Var}(\hat{\mu}_{0,m,s}) = \sigma_{0,m,s}^2$, Owen (1997a) essentially proved the following result.

**Theorem 11** Let $b \geq \max\{s, 2\}$ and $f : [0, 1)^s \to \mathcal{R}$ be smooth such that

$$\int_{[0,1)^s} \left[ \frac{\partial^s}{\partial x_1 \cdots \partial x_s} f(x) \right]^2 dx > 0.$$  

Then there exist positive constants $c, C$ such that

$$cm^{s-1}b^{-3m} \leq \sigma_{0,m,s}^2 \leq Cm^{s-1}b^{-3m},$$

as $m = \log_b n \to \infty$.

**Remark.** If $s = b = 10$, say 10, then

$$(\log_b n)^{(s-1)/2}/n^{3/2} \leq 1/n^{1/2}$$

whenever $n \geq 1$. 

---

83
The proof of Theorem 11 is given in Owen (1997a). One key feature of the proof is that it uses a nested ANOVA decomposition of $f$ via a $s$-dimensional base $b$ Haar multiresolution analysis.

The next lemma on $(0, m, s)$-nets and $(0, s)$-sequences complements the Theorem 11.

**Lemma 1** For $m \geq 2$, a $(0, m, s)$-net in base $b$ can only exist if $s \leq b + 1$. Furthermore a $(0, s)$-sequence in base $b$ can only exist if $s \leq b$.

The proof of the above lemma is given in Niederreiter (1992), page 62.
Finally Owen (1998), page 477, showed that

**Theorem 12** Let $X_1, \cdots, X_n$ be the points of a scrambled $(t, m, s)$-net in base $b$, and let $f$ be a function on $[0, 1)^s$ with integral $\mu$ and

$$\int_{[0,1)^s} [f(x) - \mu]^2 dx = \sigma^2 < \infty.$$ 

Then

$$\text{Var}(\hat{\mu}_{tms}) = o(1/n), \quad \text{as } n \to \infty,$$

$$\text{Var}(\hat{\mu}_{tms}) \leq \frac{b^t b + 1}{n[b - 1]^s} \sigma^2.$$

For $t = 0$,

$$\text{Var}(\hat{\mu}_{tms}) \leq \frac{1}{n} \left(\frac{b}{b - 1}\right)^{s-1} \sigma^2.$$

For $s = 1$,

$$\text{Var}(\hat{\mu}_{tms}) \leq \frac{b^t}{n} \sigma^2.$$

For $b = 2$,

$$\text{Var}(\hat{\mu}_{tms}) \leq \frac{2^{t3^s}}{n} \sigma^2.$$
2.8.3 Asymptotic distribution

Assuming that $\sigma_{0,m,s}^{2} > 0$, define

$$W = \sigma_{0,m,s}^{-1}(\hat{\mu}_{0,m,s} - \mu).$$

Then we have the following result.

**Theorem 13** Let $b \geq \max\{s,2\}$ and $f: [0,1)^{s} \to \mathcal{R}$ be smooth such that

$$\int_{[0,1)^{s}} \left[ \frac{\partial^{s}}{\partial x_{1} \cdots \partial x_{s}} f(x) \right]^{2} dx > 0.$$

Then $W$ converges weakly to the standard normal distribution as $m = \log_{b} n \to \infty$.

**Remark.** Recently, Hong, Hickernell and Wei (2001) reported additional empirical evidence that the central limit effect for scrambled nets can take place for reasonable sample sizes.
2.8.4 Asymptotic distributions in general

Before we discuss the proof of Theorem 13, we shall briefly describe three other related problems on asymptotic distributions (just to illustrate that interest in asymptotic distributions exists across many disciplines).

2.8.4.1 Ulam’s problem

The following problem was raised by Ulam in the early sixties [Ulam, (1961)]. Let $S_n$ be the group of permutations of $\{1, 2, \cdots, n\}$. If $\pi \in S_n$, we say that $\pi(i_1), \cdots, \pi(i_k)$ is an increasing subsequence in $\pi$ if $i_1 < i_2 < \cdots < i_k$ and $\pi(i_1) < \pi(i_2) < \cdots < \pi(i_k)$. Let $l_n(\pi)$ be the length of the longest increasing subsequence.

For example, if $n = 5$ and $\pi$ is the permutation $5 \ 1 \ 3 \ 2 \ 4$ (in one-line notation: $\pi(1) = 5, \pi(2) = 1, \cdots$) then the longest increasing subsequences are $1 \ 2 \ 4$ and $1 \ 3 \ 4$. Hence $l_n(\pi) = 3$. 

87
If $S_n$ is equipped with the uniform distribution, what is the distribution of $l_n$ as $n \to \infty$? The motivation for this question has its roots in mathematical physics. A great deal of work has been invested into this problem in the last forty years and in 1999, Baik, Deift and Johansson proved

**Theorem 14** With the above notation,

$$\lim_{n \to \infty} P \left( n^{-1/6}(l_n - 2n^{1/2}) \leq t \right) = F(t), \quad \forall t \in R,$$

where $F(t)$ is the c.d.f. of the Tracy-Widom distribution.

This is a pretty intriguing result as the Tracy-Widom distribution is associated with the law of the largest eigenvalue of a random (Wishart) matrix [see Johnstone (2001)]. More of this and related issues can be found in the very readable account by Aldous and Diaconis (1999).
2.8.4.2 Biological sequence matching

Consider two finite sequences $x$ and $y$ from a finite alphabet $\mathcal{A}$. Thus $x = x_1x_2 \cdots x_m$ and $y = y_1y_2 \cdots y_n$ with $x_i, y_j \in \mathcal{A}$. Assume that $x_1, \cdots, x_m$ are independently distributed with

$$P_0\{x_i = \alpha\} = \mu_\alpha, \ \forall i,$$

and similarly $y_1, \cdots, y_n$ are independently distributed with

$$P_0(y_j = \beta) = \nu_j, \ \forall j.$$

These two sequences are to be aligned. A candidate alignment $z = \{(i_t, j_t) : 1 \leq t \leq k\}$ for some $1 \leq i_1 < i_2 < \cdots < i_k \leq m$ and $1 \leq j_1 < j_2 < \cdots < j_k \leq n$, specifies that $x_{i_t}$ and $y_{j_t}$ are aligned for all $t = 1, \cdots, k$. The other $x$’s with subscripts between $i_1$ and $i_k$ and other $y$’s with subscripts between $j_1$ and $j_k$ are said to be unaligned.

With each candidate alignment $z$, we associate a score $S_z(x, y)$. Aligned letters $x_i$ and $y_j$ are scored according to a similarity matrix $K(x_i, y_j)$. In addition, a penalty $\delta$ is imposed for each unaligned letter as well as a ‘gap’ penalty $\Delta$ for each nonempty interval of unaligned letters.
Given a collection \( Z \) of candidate alignments, one can identify the alignment with the highest score. Under the null hypothesis that the sequences \( x \) and \( y \) are independent, what is the distribution of

\[
\max_{z \in Z} S_z(x, y),
\]
as \( m, n \) tends to infinity.

Thus far, work on this problem has been largely confined to Poisson approximations [e.g., Neuhauser (1994)] and to large deviations of \( \max_{z \in Z} S_z(x, y) \), i.e.,

\[
P(\max_{z \in Z} S_z(x, y) \geq b)
\]
as \( n, m, b \to \infty \). [see, e.g., Siegmund and Yakir (2000)].

This problem is especially important in DNA and protein sequence alignments as the above probability is needed in the BLAST algorithm


Very readable accounts of this and related problems can be found in Waterman (1995) and Grant and Ewens (2001).
2.8.4.3 Medical imaging

In his review paper, Adler (2000) described the following problem in biostatistics, related to brain imaging. A small number of subjects are injected with a positron emitting radio isotope, which enables a positron emission tomography (PET) machine to follow its flow through the bloodstream, and in particular the brain. Images of the brain are taken while the subjects are at rest, and then when performing a task, such as silent reading of words projected onto a screen. The underlying principle is that those parts of the brain involved in performing the task will require oxygen, and hence blood, so that there will be heavier blood flow to, and hence positron emission from, these regions.

The image is modeled as a (Gaussian) random field $X(.)$ which leads to interest in the probability

$$P(\sup_{t \in T} X(t) \geq \lambda)$$

where $T$ is the parameter space (e.g., the brain region).
Adler observed that a promising approach to approximating the above probability is the Euler characteristic method and he posed the following question: what is the asymptotic distribution of the empirical Euler characteristic? It should also be noted that he thinks [Adler (2000), page 68] it is unlikely that this is a nut we shall ever be able to crack.
2.8.5 Stein’s method

Stein (1972) introduced a powerful and general method for obtaining an explicit bound for the error in the normal approximation to the distribution of a sum of dependent random variables. This method was extended from the normal distribution to the Poisson distribution by Chen (1975). Since then, Stein’s method has found considerable applications in combinatorics, probability and statistics, e.g.,

- Stein (1986): binomial approximation.

Stein (1986) gives an excellent account of this method.
Central to Stein’s normal approximation technique is the following lemma.

**Lemma 2 (Stein)** Let $z \in \mathcal{R}$ and $\Phi$, $\phi$ denote the cumulative distribution function, probability density function of the standard normal distribution respectively. The unique bounded solution $g_z : \mathcal{R} \to \mathcal{R}$ of the differential equation

$$g^{(1)}(w) - wg(w) = \mathcal{I}\{w \leq z\} - \Phi(z),$$

for all $w \in \mathcal{R}$, is given by

$$g_z(w) = \begin{cases} 
\Phi(w)[1 - \Phi(z)]/\phi(w), & \text{if } w \leq z, \\
\Phi(z)[1 - \Phi(w)]/\phi(w), & \text{if } w > z.
\end{cases}$$

Furthermore,

$$0 \leq g_z(w) \leq 1, \quad |g_z^{(1)}(w)| \leq 1,$$

for all $w \in \mathcal{R}$.

**Remark.** Equation (3) is often called the normal identity. There exist analogous identities for other distributions too (e.g., binomial, Poisson, compound Poisson, multinomial, etc.).
The basic idea behind Stein’s method is that equation (3) is an identity in \( w \) and hence is valid for any random variable \( W \). Now replacing \( w \) by \( W \), and then taking expectation, we have

\[
P(W \leq z) - \Phi(z) = E[g_z^{(1)}(W) - W g_z(W)].
\]

If \( W \) is standard normal, then the right hand side of the above equation must be zero. Hence to show that \( W \) is asymptotically standard normal, it suffices to bound the terms on the right hand side and show that they tend to zero.
We shall now give a sketch of the proof of Theorem 13 as an illustration of Stein’s method.

**Proof of Theorem 13**

**Step 1.** We first need some notation. Define for integers $k \geq 0$, $0 \leq t \leq b^k - 1$, $0 \leq c \leq b - 1$,

$$ \psi_{k,t,c}(x) = b^{(k+1)/2} \mathcal{I}\{[b^{k+1}x] = bt + c\} - b^{(k-1)/2} \mathcal{I}\{[b^kx] = t\}, \quad x \in [0,1). $$

**Remark.** For arbitrary (but fixed) integers $t_1$ and $t_2$, $\psi_{k,t,c}$ is constant on $[t_1 b^{-k-1}, (t_1 + 1)b^{-k-1})$ and integrate to zero over $[t_2 b^{-k}, (t_2 + 1)b^{-k})$.

Owen (1997a), page 1897, observed that a typical basis function for $L^2([0,1]^s)$ is of the form

$$ \prod_{r=1}^{l} \psi_{k_{j_{r}}t_{j_{r}},c_{j_{r}}}(x_{j_{r}}), \quad (x_1, \cdots, x_s)' \in [0,1)^s, $$

where $1 \leq j_1 < \cdots < j_l \leq s$. 

96
Next let

\[
\{U[\tilde{c}_{j,1}, \ldots, \tilde{c}_{j,u_j} : 1 \leq j \leq s] : \\
0 \leq \tilde{c}_{j,1}, \ldots, \tilde{c}_{j,u_j} \leq b - 1, u_j \geq 0, 1 \leq j \leq s\}
\]

be a set of independent random vectors where each \(U[\tilde{c}_{j,1}, \ldots, \tilde{c}_{j,u_j} : 1 \leq j \leq s]\) has the uniform distribution on the elementary interval

\[
\prod_{j=1}^{s} \left[ \sum_{k=1}^{u_j} \tilde{c}_{j,k} b^{-k}, b^{-u_j} + \sum_{k=1}^{u_j} \tilde{c}_{j,k} b^{-k} \right].
\]
Finally define

$$\nu_{0, \ldots, 0[\cdot]} = \mu, \quad |u| = 0,$$

and

$$\nu_{u_1, \ldots, u_s}[^{\tilde{c}_{j,1}, \cdots, \tilde{c}_{j,u_j} : 1 \leq j \leq s}],$$

$$= \sum_{t_{j_1} = 0}^{u_{j_1} - 1} \sum_{c_{j_1} = 0}^{b - 1} \cdots \sum_{t_{|u|} = 0}^{u_{|u|} - 1} \sum_{c_{|u|} = 0}^{b - 1} \langle f, \prod_{l=1}^{|u|} \psi_{u_{j_l} - 1, t_{j_l}, c_{j_l}} \rangle$$

$$\times E\{\prod_{l=1}^{|u|} \psi_{u_{j_l} - 1, t_{j_l}, c_{j_l}} \circ U_{j_l}[^{\tilde{c}_{j,1}, \cdots, \tilde{c}_{j,u_j} : 1 \leq j \leq s}]\},$$

if $|u| \geq 1$, where

$$|u| = \sum_{j=1}^{s} \mathcal{I}\{u_j \geq 1\},$$

$U_{j_l}$ denotes the $j$th coordinate of $U$ and

$$\langle f, \prod_{r=1}^{l} \psi_{k_{j_r}, t_{j_r}, c_{j_r}} \rangle$$

$$= \int_{[0,1]^s} f(x)[\prod_{r=1}^{l} \psi_{k_{j_r}, t_{j_r}, c_{j_r}}(x_{j_r})] dx.$$
**Remark.** A useful property of $\nu_{u_1,\ldots,u_s}$ is for each $1 \leq k \leq s$,

$$\sum_{0 \leq \tilde{c}_k, u_k \leq b-1} \nu_{u_1,\ldots,u_s[\tilde{c}_{j,1}, \ldots, \tilde{c}_{j,u_j} : 1 \leq j \leq s] = 0.}$$

Loh (2002) showed that

$$W = \frac{1}{b^m \sigma_{0,m,s}} \sum_{i=1}^{b^m} \sum_{u_1,\ldots,u_s \geq 0: u_1 + \ldots + u_s \geq m+1} \times \nu_{u_1,\ldots,u_s[\pi_j(a_{i,j,1}), \ldots, \pi_j(a_{i,j,u_j}) : 1 \leq j \leq s]}. $$

As an illustration of the nature of the function $\nu_{u_1,\ldots,u_s}$, if $b = s = 2$,

$$\nu_{0,0[ , ]} = \mu,$$

$$\nu_{1,0[0, ]} = \int_0^{1/2} \int_0^1 [f(x) - \mu]dx,$$

$$\nu_{1,0[1, ]} = \int_{1/2}^1 \int_0^1 [f(x) - \mu]dx.$$
Step 2. As it is, $W$ appears to be rather difficult to analyze. As such, this step creates a proxy, $\tilde{W}$, for $W$ which is more analytically tractable.

Define $\tilde{m} = \lceil 2s \log_b m \rceil$ and

$$\tilde{W} = \frac{1}{b^n \sigma_{0,m,s}} \sum_{i=1}^{b^n} \sum_{u_1, \ldots, u_s \geq \tilde{m} : u_1 + \cdots + u_s \geq m+1} \times \nu_{u_1, \ldots, u_s} \left[ \pi_j(a_{i,j,1}), \ldots, \pi_j(a_{i,j,1}, \ldots, a_{i,j,1}, u_{j-1}(a_{i,j,u_j}) : 1 \leq j \leq s \right] .$$

Then

Proposition 1 Let $b \geq \max\{s, 2\}$ and $f : [0, 1)^s \to \mathcal{R}$ be smooth such that

$$\int_{[0,1)^s} \frac{\partial^s}{\partial x_1 \cdots \partial x_s} f(x)^2 dx > 0.$$  

Then $W - \tilde{W} \to 0$ in probability as $m \to \infty$.

The above proposition and Slutsky’s theorem show that $W$ and $\tilde{W}$ has a common asymptotic distribution. Hence it remains only to determine the asymptotic distribution of $\tilde{W}$.
**Step 3.**

We shall now use Stein’s method to determine the asymptotic distribution of $\tilde{W}$.

**Theorem 15** Let $b \geq \max\{s, 2\}$ and $f : [0, 1)^s \to \mathcal{R}$ be smooth such that

$$
\int_{[0,1)^s} \left[ \frac{\partial^s}{\partial x_1 \cdots \partial x_s} f(x) \right]^2 dx > 0.
$$

Then as $m \to \infty$,

$$
\sup\{|P(\tilde{W} \leq w) - \Phi(w)| : -\infty < w < \infty\} = O\left( \frac{\log_b m}{m} \right)^{1/2}.
$$

**Proof.** In order to obtain an identity in $\tilde{W}$, we first need to construct an exchangeable pair of random variables $(\tilde{W}, \tilde{W}^*)$. Let $I$ and $J$ be random variables uniformly distributed over $\{1, \cdots, b^m\}$ and $\{1, \cdots, s\}$ respectively.
Furthermore let
\[\{\pi_j^*, \pi_{j;a_1}^*, \pi_{j;a_1,a_2}^*, \ldots : 1 \leq j \leq s, \]
\[0 \leq a_k \leq b - 1, k = 1, 2, \ldots \}\]
be an independent replication of the \(\pi\)'s. We assume that \(I\), \(J\) and the \(\pi^*\)'s are independent and that they are also independent of all previously defined random quantities.

Next define for \(1 \leq i \leq b^m, 1 \leq j \leq s,\)
\[\tilde{\pi}_{j;a_i,j,1,a_{i,j},u_{j-1}} = \begin{cases} 
\pi_{j;a_i,j,1,a_{i,j},u_{j-1}}^* & \text{if } J = j, u_j \geq \tilde{m}, \text{ and} \\
(a_{I,,j,1}, \ldots, a_{I,,j,'m-1}) = \\
(a_{i,,j,1}, \ldots, a_{i,,j,'m-1}), \\
\pi_{j;a_i,j,1,a_{i,j},u_{j-1}} & \text{otherwise,}
\end{cases}\]
where
\[A_{i,j} = \sum_{k=1}^{\infty} a_{i,j,k} b^{-k},\]
for the \((0, m, s)\)-net \(\{A_i : i = 1, \ldots, b^m\}\).
With $\tilde{W}$ as motivation, define

$$
\tilde{W}^* = \frac{1}{b^m \sigma_{0,m,s}} \sum_{i=1}^{b^m} \sum_{u_1, \ldots, u_s \geq \tilde{m}; u_1 + \cdots + u_s \geq m+1} \nu_{u_1, \ldots, u_s} \left[ \tilde{\pi}_j(a_{i,j,1}), \ldots, \tilde{\pi}_{j;a_{i,j,1}, \ldots, a_{i,j,u_j-1}}(a_{i,j,u_j}) : 1 \leq j \leq s \right].
$$

From symmetry, we observe that $(\tilde{W}, \tilde{W}^*)$ is an exchangeable pair of random variables.

Next let $z \in \mathcal{R}$ and $g_z : \mathcal{R} \to \mathcal{R}$ be as Stein’s Lemma. From the exchangeability of $(\tilde{W}, \tilde{W}^*)$, we have

$$
E\{(\tilde{W}^* - \tilde{W})[g_z(\tilde{W}) + g_z(\tilde{W}^*)]\} = 0.
$$

**Remark.** Note that the left hand side of the above equation is an antisymmetric function in $(\tilde{W}, \tilde{W}^*)$. 
Using a conditioning argument and some algebra we obtain the identity

\[
P(\tilde{W} \leq z) - \Phi(z) \\
= E[g_z^{(1)}(\tilde{W}) - \tilde{W} g_z(\tilde{W})] \\
= E \int [g_z^{(1)}(\tilde{W}) - g_z^{(1)}(V + w)] K_{\tilde{W},\tilde{W}^*}(w) dw \\
\quad + E[g_z^{(1)}(\tilde{W})] E[ \int K_{\tilde{W},\tilde{W}^*}(w) dw] \\
\quad - E[g_z^{(1)}(\tilde{W}) \int K_{\tilde{W},\tilde{W}^*}(w) dw] \\
\quad + [E g_z^{(1)}(\tilde{W})] [1 - E \int K_{\tilde{W},\tilde{W}^*}(w) dw].
\]

where \( K_{\tilde{W},\tilde{W}^*} \) is a known kernel function.

It remains now to show that each of the terms on the right hand side of the last equality is of order \( O((m^{-1} \log_b m)^{1/2}) \).

This part of the proof is very tedious and computationally intensive and the details can be obtained from Loh (2002).
2.9 Good Lattice Points

This section will conclude our treatment of the frequentist approach to computer experiments. Again, we shall focus on the estimation of

\[ \mu = \int_{[0,1)^d} f(x)dx \]

using \( n \) evaluations of the integrand \( f : \mathbb{R}^d \rightarrow \mathbb{R} \). An inspection of the error bounds of the numerical integration techniques discussed in previous sections reveals a possibly undesirable feature in that once \( f \) is sufficiently regular, (e.g. of bounded variation in the sense of Hardy and Krause), then any additional regularity (say, smoothness) of \( f \) is not reflected in the order of magnitude of the error bound.

The method of good lattice points was introduced by Korobov (1959). This method enjoys the property that the degree of regularity of \( f \) is reflected in the order of magnitude of the error bound provided \( f \) is periodic (with period 1) in each of its \( d \) covariates, i.e.,

\[ f(t_1, \ldots, t_d) = f(t_1 + m_1, \ldots, t_d + m_d), \]

for all \((t_1, \ldots, t_d) \in \mathbb{R}^d\) and \((m_1, \ldots, m_d) \in \mathbb{Z}^d\).
One construction of good lattice points is as follows [see Chapter 5 of Niederreiter (1992)]. First for each \( t \in \mathbb{R} \), we write \( \{t\} = t \mod 1 \), e.g., if \( t = 3.142 \), then \( \{t\} = 0.142 \).

Let \( g = (g_1, \cdots, g_d)' \in \mathbb{Z}^d \) and that \( 1, g_1, \cdots, g_d \) are linearly independent over the set of rational numbers. Then a set of \( n \) lattice points is given by \( x_1, \cdots, x_n \in [0, 1)^d \) where

\[
x_{k,j} = \left\{ \frac{g_j(k - 1)}{n} \right\}, \\
x_k = (x_{k,1}, \cdots, x_{k,d})',
\]

whenever \( j = 1, \cdots, d \) and \( k = 1, \cdots, n \). The term good lattice points refer to a careful choice of \( n \) and \( g \) usually based on number theory.

Using this design, \( \mu \) is estimated via the sample mean

\[
\hat{\mu}_{GLP} = \frac{1}{n} \sum_{k=1}^{n} f(x_k).
\]
Now with $f$ satisfying (4), we can write

$$f(x_i) = f\left(\frac{g_1(i - 1)}{n}, \ldots, \frac{g_d(i - 1)}{n}\right).$$

If $f$ can be written as a Fourier series, we obtain

$$f(x) = \sum_{h \in \mathbb{Z}^d} \hat{f}(h) \exp(2\pi i \langle h, x \rangle), \quad x \in \mathbb{R}^d,$$

with Fourier coefficients

$$\hat{f}(h) = \int_{[0,1]^d} f(x) \exp(-2\pi i \langle h, x \rangle),$$

and $\langle h, x \rangle$ denotes the usual inner product of 2 vectors $h, x \in \mathbb{R}^d$.

Observing that $\mu = \hat{f}(0)$, we have after some simplification

$$\hat{\mu}_{GLP} - \mu = \frac{1}{n} \sum_{h \neq 0} \hat{f}(h) \sum_{k=1}^{n} \exp(2\pi i \langle x_k, h \rangle).$$

The last inner sum on the r.h.s. is equal to zero if $\langle g, h \rangle \neq 0 \mod n$ and equal to $n$ otherwise.
Hence
\begin{equation}
\hat{\mu}_{GLP} - \mu = \sum_{h} \hat{f}(h),
\end{equation}
where the sum on the r.h.s. is over all \( h \neq 0 \) satisfying \( \langle g, h \rangle = 0 \mod n \).

According to the Riemann-Lebesgue lemma, \( \hat{f}(h) \) tends to zero as \( h \) moves away from the origin 0. The convergence rate of \( \hat{f}(h) \) to zero serves as a regularity condition on \( f \).

For good lattice points, we need to choose \( g \) and \( n \) such that the r.h.s. of (5) is small. Niederreiter (1992), page 145, noted that for \( d \geq 3 \), computer searches are needed to find good choices of \( g \) and \( n \). However when \( d = 2 \), an explicit construction based on Fibonacci numbers giving good lattice points is available.
Appendix A of Fang and Wang (1994) lists several choices of good lattice points but the smallest value of \( n \) there for \( d = 5 \) is 1069. Bates, et. al. (1996) has a practical example on the use of good lattice points in computer experiments. Sloan and Joe (1994) also contains a comprehensive discussion of lattice rules for numerical integration and Fang and Wang (1992) gives numerous applications of good lattice points to statistics.

**Remark.** There are other variations of good lattice points. Koehler and Owen (1996) defined a centered version of good lattice points as \( x_1, \ldots, x_n \) where

\[
x_{i,j} = \left\{ \frac{g_j(i - 1) + 0.5}{n} \right\}.
\]

However the results reported in this section apply to these points as well.
3.1 Bayesian Approach

When a physical process is modeled by a large simulator, experiments can be conducted directly on the computer code. The complexity and non-linearity of the code (plus absence of measurement errors) have given way to two new methodologies

- the use of ‘space filling’ experimental designs which fill up the factor/covariate space in a uniform fashion,

- fitting highly adaptive models that interpolate the design observations.

One possible disadvantage of the use of deterministic models (e.g., multivariate splines, high order polynomials, neural networks, etc.) that interpolate the data points is that the complexity (i.e., the number of parameters) of the model must necessarily increase with the sample size $n$. 
On the other hand, the use of a stochastic model avoids this drawback in that the probabilistic mechanism that generates the model does not need to increase in complexity as sample size increases. A Bayesian approach to modeling simulator output [e.g., Sacks, et. al. (1989), Welch, et. al. (1990)] can be based on a spatial model adapted from geo-statistics [e.g., Cressie (1993), (1996), Koehler and Owen (1996) and Stein (1999)].
3.2 Gaussian Random Fields

This section provides us with some preliminary material on random fields in general and Gaussian random fields in particular.

**Definition.** A random field $Z : \mathbb{R}^d \rightarrow \mathbb{R}$ is a collection of random variables $\{Z(x) : x \in \mathbb{R}^d\}$. It is said to be a Gaussian random field if all its finite dimensional distributions are Gaussian (i.e., multivariate normal). Thus for instance, $(Z(x_1), \ldots, Z(x_m))'$ has a multivariate normal distribution for all $x_1, \ldots, x_m$.

**Definition.** A random field $Z$ is strictly stationary if for all finite $m$, $x, x_1, \ldots, x_m \in \mathbb{R}^d$ and $t_1, \ldots, t_m \in \mathbb{R}$, we have

$$P(Z(x_1 + x) \leq t_1, \ldots, Z(x_m + x) \leq t_m)$$

$$= P(Z(x_1) \leq t_1, \ldots, Z(x_m) \leq t_m).$$
**Definition.** The mean function of a random field $Z$ is defined to be $E[Z(.)]$. The covariance function of a random field is defined to be $\text{Cov}[Z(.), Z(.)]$. A covariance function is said to be an autocovariance function $K(.)$ if

$$K(x - y) = \text{Cov}[Z(x), Z(y)], \quad x, y \in R^d.$$  

A random field is said to be weakly stationary if it has finite second moments, its mean function is a constant and it possesses an autocovariance function.

In the case of a Gaussian random field, the two concepts (strictly stationary and weakly stationary) are equivalent as the law of the random field is uniquely determined by its mean and autocovariance functions. Hence no distinction need be made for Gaussian random fields.
3.3 The Kriging Model

The kriging approach uses a two-component model. The first component consists of a general linear model \( \sum_{j=1}^{k} \beta_j h_j(x) \) while the second (or lack of fit) component is treated as the realization of a stationary Gaussian random field \( Z(x) \).

Letting \([0, 1)^d\) to be the design space, the kriging approach models the response as

\[
(6) \quad f(x) = \sum_{j=1}^{k} \beta_j h_j(x) + Z(x), \quad x \in [0, 1)^d,
\]

where the \( h_j \)'s are known fixed functions, the \( \beta_j \)'s are unknown coefficients to be estimated and \( Z(x) \) is a stationary Gaussian random field with \( E[Z(x)] = 0 \) and autocovariance function

\[
\text{Cov}[Z(x_i), Z(x_j)] = \sigma^2 R(x_i - x_j).
\]

Here \( \sigma^2 = \text{Var}[Z(x)] \) and \( R(.) \) is the autocorrelation function of \( Z \). It is assumed that both these quantities are known.
Remark. The complexity of the kriging model does not increase with the size of the design (i.e., sample size).

For any point \( x \in [0, 1)^d \), the response \( f(x) \) has the Gaussian distribution with mean \( \sum_{j=1}^{k} \beta_j h_j(x) \) and variance \( \sigma^2 \). The linear component models the drift in the response, while the systematic lack of fit is modeled by the second component. The smoothness and other properties of \( Z \) are governed by the autocorrelation function \( R(.) \).

Next let \( x_1, \cdots, x_n \in [0, 1)^d \) be the design points and \( f_D = (f(x_1), \cdots, f(x_n))' \). Consider a linear predictor
\[
\hat{f}(x_0) = \lambda'(x_0) f_D,
\]
of \( f(x_0) \) where \( x_0 \) is an unobserved point. Now treating \( f_D \) as a random variable, the best linear unbiased predictor (BLUP) finds the \( \lambda(x_0) \) that minimizes
\[
MSE[\hat{f}(x_0)] = E[\lambda' f_D - f(x_0)]^2,
\]
subject to the unbiasedness condition
\[
E[\lambda' f_D] = E[f(x_0)].
\]
The BLUP of $f(x_0)$ can be worked out to be

$$
\hat{f}_{BLUP}(x_0) = h'(x_0)\hat{\beta}_{GLS} + v'_x V_D^{-1} (f_D - H_D \hat{\beta}_{GLS}),
$$

where

$$
h'(x_0) = (h_1(x_0), \ldots, h_k(x_0)),
$$

$$(H_D)_{i,j} = h_j(x_i),
$$

$$(V_D)_{i,j} = \text{Cov}[Z(x_i), Z(x_j)],
$$

$$v'_x = (\text{Cov}[z(x_0), Z(x_1)], \ldots, \text{Cov}[Z(x_0), z(x_n)]),$$

and

$$\hat{\beta}_{GLS} = [H'_D V_D^{-1} H_D]^{-1} H'_D V_D^{-1} f_D$$

is the generalized least squares (GLS) estimate of

$$\beta = (\beta_1, \ldots, \beta_k)' .$$

The mean square error of $\hat{f}_{BLUP}(x_0)$ is

$$MSE[\hat{f}_{BLUP}(x_0)]$$

$$= \sigma^2 - (h'(x_0), v'_x) \left( \begin{array}{cc} 0 & H'_D \\ H_D & V_D \end{array} \right)^{-1} \left( \begin{array}{c} h(x_0) \\ v_x \end{array} \right).$$
The first component of the r.h.s. of (7) is the generalized least squares prediction at point \( x_0 \) given the design covariance matrix \( V_D \). The second component ‘pulls’ the GLS response surface through the observed data points. The elasticity of the pull is solely determined by the autocorrelation function \( R(\cdot) \).

As a consequence, the predictions at the design points are exactly the corresponding observations and the MSE is zero. As a prediction point moves away from all the design points, the second component goes to zero, yielding the GLS prediction, while the MSE at that point approaches

\[
\sigma^2 + h'(x_0)[H'_D V_D^{-1} H_D]^{-1} h(x_0).
\]

**Remark.** This section assumes that the covariance function of \( Z \) is known. The kriging approach can be generalized to \( \sigma^2 \) unknown and that the autocorrelation function \( R(\cdot) \) be known up to a finite number of parameters. Putter and Young (2001) consider the effect on covariance function estimation on the accuracy of kriging predictors.
3.4 Fully Bayesian Model

A popular alternative to the kriging model is the fully Bayesian model [see for example Currin, et. al. (1991); O’Hagan (1989)]. The fully Bayesian approach uses the same model (6) as the kriging approach but with a different interpretation of the $\beta_j$’s.

Here the $\beta_j$’s are random variables with prior distributions $\pi_j$’s. The mixed convolution of the $\pi_j$’s with the law of $Z$ yields a prior distribution for $f$. Once the data $f_D$ has been observed, the posterior distribution of $f$ can be calculated. The posterior mean of $f(x_0)$, i.e.,

$$E[f(x_0)|f_D],$$

and the posterior variance

$$\text{Var}[f(x_0)|f_D],$$

are used as the predictor of $f(x_0)$ and as a measure of prediction error respectively.
In general, the kriging and fully Bayesian approaches yield different predictors. However, if the distribution of \( Z(.) \) is Gaussian and the prior distribution of \( \beta \) is diffuse, the two approaches yield identical predictors.

One possible drawback with the fully Bayesian approach is that it is usually much more computationally intensive than the kriging method. Currin, et. al. (1991) has a detailed discussion of the fully Bayesian approach with \( h_j = 1, j = 1, \cdots, k \).
3.5 Correlation Functions

The autocorrelation function $R(.)$ of the random field $Z$ plays a crucial role in the construction of designs and in prediction. In particular the autocorrelation function is closely related to the smoothness of the sample path of $Z(.)$.

3.5.1 Some preliminary results

**Definition.** A function $K : R^d \times R^d \rightarrow R$ is said to be positive definite if

$$\sum_{j,k=1}^{n} c_j c_k K(x_j, x_k) \geq 0$$

for all $x_1, \cdots, x_n \in R^d$, $c_1, \cdots, c_n \in R$ and positive integer $n$.

Given any positive definite function $K$, there exists a Gaussian random field that has covariance function $K$. If in addition, $K(x, x) = 1$ for all $x$, then $K(.,.)$ can be used as a correlation function. It is also known that every covariance function is a positive definite function.
If the random field $Z$ is weakly stationary, then for simplicity it is customary to write

$$K(x_1, x_2) = K(x_1 - x_2),$$

and for correlation functions

$$R(x_1, x_2) = R(x_1 - x_2).$$

We note that the use of such correlation functions does not automatically eliminate the possibility that the response $f(.)$ is nonstationary since the nonstationary behavior of the mean of $f(.)$ can be modeled by the first term of (6).

**Definition.** A random field $Z$ is said to be weakly isotropic if

$$\text{Cov}[Z(x_1), Z(x_2)] = K(||x_1 - x_2||), \quad x_1, x_2 \in \mathbb{R}^d,$$

for some function $K : \mathbb{R} \to \mathbb{R}$ where $\|\cdot\|$ is the usual Euclidean norm in $\mathbb{R}^d$. 

121
Product or multiplicative correlation functions on $R^d$ can be constructed from simplier correlation function on $R$ as follows. For $j = 1, \cdots, d$, let $Z_j$ be a stochastic process on the real line with autocorrelation function

$$\text{Corr}[Z(x_{1,j}), Z(x_{2,j})] = R_j(|x_{1,j} - x_{2,j}|), \quad x_1, x_2 \in R^d.$$ 

Then in $d \geq 2$ dimensions, we have the following multiplicative autocorrelation function

$$R(x_1 - x_2) = \prod_{j=1}^{d} R_j(|x_{1,j} - x_{2,j}|), \quad x_1, x_2 \in R^d.$$ 

It can be shown that $R(.)$ is indeed an autocorrelation function for a (Gaussian) random field on $R^d$. Multiplicative correlation functions are much more analytically tractable than isotropic ones and have been used extensively in computer experiments [for example, Currin, et. al. (1991), Sacks, et. al. (1989), Welch, et. al. (1990)].
On the other hand, Stein (1999), page 55, argues in favor of isotropic correlation functions over multiplicative correlation functions in that the usage of multiplicative correlation functions may lead to physically unrealistic behavior. It is noted that functions that are both isotropic and multiplicative must have the form $c \exp(-a\|x\|^2)$. 
3.5.2 Mean square properties

There is no simple exact relationship between the autocovariance function of a weakly stationary random field and the smoothness of its realizations (or sample paths). However it is possible to rigorously relate the autocovariance function to the mean square properties of a random field.

**Definition.** Suppose $Z$ is a random field on $R^d$. Then $Z$ is mean square continuous at $x$ if

$$
\lim_{y \to x} E\{[Z(y) - Z(x)]^2\} = 0.
$$

It is well known [see for example Stein (1999), page 20] that for $Z$ weakly stationary with autocovariance function $K$, $Z$ is mean square continuous at $x$ if and only if $K$ is continuous at 0. Since the latter statement is independent of $x$, we conclude that $Z$ is either mean square continuous everywhere or nowhere. It is to be noted that mean square continuity of $Z$ does not necessarily imply that the sample path of $Z$ be continuous (with probability 1).
If $K$ is continuous at 0, it is continuous everywhere since

$$|K(x) - K(y)| = |\text{Cov}[Z(x) - Z(y), Z(0)]| \leq \{\text{Var}[Z(x) - Z(y)]\text{Var}[Z(0)]\}^{1/2} = \{2[K(0) - K(x - y)]K(0)\}^{1/2} \to 0,$$

as $y \to x$. On the other hand, if $K$ is not continuous at 0, it may have other discontinuities.

**Definition.** A random field $Z$ on the real line with finite second moments is said to be mean square differentiable at $t$ if there exists a random variable $Z^{(1)}(t)$ such that

$$E\left\{\left[\frac{Z(t + h) - Z(t)}{h} - Z^{(1)}(t)\right]^2\right\} \to 0,$$

as $h \to 0$. 

125
A weakly stationary random field is either mean square differentiable everywhere or nowhere. If $K$ denotes the autocovariance function of $Z$, then $Z$ is mean square differentiable if and only if $K^{(2)}(0)$, the second derivative of $K$ at 0, exists and is finite. If $Z$ is mean square differentiable, then $Z^{(1)}(.)$ as another random field has autocovariance function $-K^{(2)}$.

In order to define higher order mean square derivatives, we say $Z$ is $m$-times mean square differentiable if it is $(m-1)$-times differentiable and $Z^{(m-1)}$ is mean square differentiable. Also it is well known that $Z$ is $m$-times mean square differentiable if and only if $K^{(2m)}(0)$ exists and is finite. In this case, the autocovariance function of $Z^{(m)}$ is $(-1)^mK^{(2m)}$.

Chapter 2 of Stein (1999) contains a very readable survey of the basic properties of random fields. A book length account of correlation theory of stationary random fields can be found in Yaglom (1986).
We shall now present five families of univariate autocorrelation functions that have been used quite widely in computer experiments.

### 3.5.3 Cubic family

The cubic univariate correlation family is parametrized by $\rho \in [0, 1]$ and $\gamma \in [0, 1]$ and is given for $t \in [-1, 1]$ by

$$R(t) = 1 - \frac{3(1 - \rho)t^2}{2 + \gamma} + \frac{(1 - \rho)(1 - \gamma)|t|^3}{2 + \gamma},$$

where

$$\rho \geq \frac{5\gamma^2 + 8\gamma - 1}{\gamma^2 + 4\gamma + 7}.$$

The last inequality is needed to ensure that $R(.)$ is positive definite [see Mitchell, et. al. (1990) and Koehler and Owen (1996)]. Here as can be verified easily, $\rho = R(1)$ is the correlation between the endpoint observations $Z(0)$ and $Z(1)$. Also $\gamma$ is the correlation between the end points of the derivative process, i.e., $Z^{(1)}(0)$ and $Z^{(1)}(1)$. 
Using model (6) and this cubic family of autocorrelations leads to a cubic spline interpolator if \( d = 1 \). If \( d = 2 \) and the multiplicative correlation function is adopted, the interpolation surface is a cubic spline in each covariate.

In general, a weakly stationary stochastic process possessing a cubic autocorrelation function is exactly once mean square differentiable. If however, the process is also Gaussian, then its sample path is (exactly) once differentiable everywhere with probablity 1 [see Cramér and Leadbetter (1967) for details].

A computer experiment application of the use of this cubic family of autocorrelation functions is given in Currin, et. al. (1991).
3.5.4 Exponential family

The univariate exponential autocorrelation family is parametrized by $\theta \in (0, \infty)$ and is given by

$$R(t) = \exp(-\theta |t|), \quad t \in [-1, 1].$$

Stationary processes with exponential autocorrelations are more commonly known as Ornstein-Uhlenbeck processes. These processes are mean square continuous but not mean square differentiable since $R(.)$ is continuous but not differentiable at 0.

If the above stochastic process is also Gaussian, then its sample path is continuous everywhere but nowhere differentiable with probability 1. Furthermore the process is also Markovian [see for example Lamperti (1977)]. Probably because of this, the Ornstein-Uhlenbeck process is much more analytically tractable (compared to random fields with other covariance matrices) and much research has been devoted to studying its theoretical properties [e.g., Ying (1991), (1993), van der Vaart (1996) and Chen, Simpson and Ying (2000)]. More of this later.
Mitchell, et. al. (1990) has found necessary and sufficient conditions on the autocorrelation function such that the derivative process is an Ornstein-Uhlenbeck process.

3.5.5 Gaussian family

Sacks, Welch, Mitchell and Wynn (1989) generalized the exponential family by proposing to use the following Gaussian family of autocorrelations:

\[ R(t) = \exp(-\theta|t|^q), \quad t \in [-1, 1], \]

where \( 0 < q \leq 2 \) and \( \theta \in (0, \infty) \). We observe that \( q = 1 \) retrieves the family of exponential autocorrelations.

An important point to note is that for \( 0 < q < 2 \), the stochastic process is mean square continuous but not mean square differentiable. However if \( q = 2 \), then the process is infinitely mean square differentiable since its autocorrelation function \( R(.) \) is infinitely differentiable.
Moreover if $q = 2$ and the process is Gaussian, then its resulting sample path is infinitely differentiable (not just in the mean square sense) with probability 1.

This family of autocorrelations (especially the case $q = 2$) may be appropriate if it is known apriori that the output of the computer experiment is reasonably smooth.
3.5.6 Matérn family

Stein (1989a) observed that the Gaussian family of autocorrelation functions in the previous subsection leads to processes that are either zero or infinitely mean square differentiable. That may be somewhat restrictive. He recommends a more flexible family of autocorrelation functions, namely the Matérn family, given by

\[ R(t) = \frac{(\theta|t|)^\nu}{\Gamma(\nu)2^{\nu-1}}K_\nu(\theta|t|), \quad t \in [-1, 1], \]

where \( K_\nu(.) \) is a modified Bessel function of order \( \nu \). This family of autocorrelation functions was first proposed by Matérn in 1947 in the context of geostatistics [for more details see the monographs of Matérn (1986) and Stein (1999)].
The associated process is $m$ times mean square differentiable if and only if $\nu > m$. Hence the amount of smoothness can be controlled by the value of $\nu$ and $\theta$ controls the range of the correlations. Again if the process is Gaussian, $m$ times mean square differentiability can be replaced by almost sure $m$ times differentiability of the sample path.

The Matérn family is probably the most flexible family of autocorrelation functions used in computer experiments to date. Practical applications of the use of the Matérn family can be found in Handcock and Stein (1993) and Handcock and Wallis (1994).
3.5.7 Triangular family

The triangular family of autocorrelation functions is defined to be autocorrelation functions that have the form

$$R(t) = c \max\{(a - |t|), 0\}, \quad t \in [-1, 1],$$

where $c$ and $a$ are suitably chosen constants. Stein (1999), page 30, noted that although this family is not commonly used in applications, some best linear unbiased predictors (BLUP) under this model exhibit unusual behavior and hence this family is of interest in order to develop an understanding of the properties of BLUP’s.
3.6 Estimating Covariance Functions

Recall that the kriging approach discussed in Section 3.3 and the fully Bayesian approach discussed in Section 3.4 assume that the covariance function of the random field $Z(.)$ be known. However in applications the variance $\sigma^2$ is usually not available while the autocorrelation function $R(.)$ will only be known up to a finite number of parameters (e.g., the families of autocorrelation functions in Sections 3.5.3 to 3.5.7).

A fully Bayesian approach would be to put prior distributions on $\sigma^2$ and the parameters of $R(.)$ and then use the posterior distribution in the statistical inference. The drawback here is that the resulting integration involved is time consuming and complex.
An alternative and perhaps more objective method is to use the data to estimate \( \sigma^2 \) and the unknown parameters of the autocorrelation function. Once that is done, \( \sigma^2 \) and \( R(.) \) are then treated as ‘known’ and prediction is performed as before. Among the methods that can be used to estimate \( \sigma^2 \) and the parameters of \( R(.) \), the maximum likelihood procedure is the most popular and will be considered next.
3.7 Maximum Likelihood Estimation

Consider the kriging model given by (6), namely,

\[ f(x) = h'(x)\beta + Z(x), \]

\( Z(.) \) is a Gaussian random field with autocovariance function \( \sigma^2 R(.) \). With the notation of Section 3.3, e.g. writing \( f_D = (f(x_1), \cdots, f(x_n))' \) to be the responses at the design points \( x_1, \cdots, x_n, \) etc., the \( n \)-variate normal likelihood function is

\[
L(\beta, \sigma^2, R|f_D) \\
= (2\pi)^{-n/2} \sigma^{-n} |R_D|^{-1/2} \\
\times \exp\{-\frac{1}{2\sigma^2} (f_D - H_D\beta)' R_D^{-1} (f_D - H_D\beta)\},
\]

where \( R_D \) is the design correlation matrix.
Consequently, the log-likelihood is

\[
l(\beta, \sigma^2, R|f_D) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log(|R_D|)
\]

\[
(9) \quad -\frac{1}{2\sigma^2} (f_D - H_D\beta)' R_D^{-1} (f_D - H_D\beta),
\]

and hence

\[
\frac{\partial l(\beta, \sigma^2, R|f_D)}{\partial \beta} = -\sigma^{-2} (H_D' R_D^{-1} f_D - H_D' R_D^{-1} H_D \beta).
\]

Equating the r.h.s. of the above equation to zero yields the maximum likelihood estimator \(\hat{\beta}_{ML}\) for \(\beta\) where

\[
\hat{\beta}_{ML} = (H_D' R_D^{-1} H_D)^{-1} H_D' R_D^{-1} f_D.
\]

We note that \(\hat{\beta}_{ML}\) is also the generalized least squares estimator for \(\beta\).
Similarly,

$$\frac{\partial l(\beta, \sigma^2, R| f_D)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(f_D - H_D\beta)'R_D^{-1}(f_D - H\beta),$$

which when set to zero gives the maximum likelihood estimator for $\sigma^2$, namely,

$$\hat{\sigma}^2_{ML} = \frac{1}{n}(f_D - H\hat{\beta}_{ML})'R_D^{-1}(f_D - H_D\hat{\beta}_{ML}).$$

This implies that if the autocorrelation function $R(.)$ is known, then $\hat{\beta}_{ML}$ and $\hat{\sigma}^2_{ML}$ are easily calculated.

The main difficulty arises when $R(.)$ is parameterized by $\theta = (\theta_1, \cdots, \theta_d)'$. Then for $j = 1, \cdots, d$,

$$\frac{\partial l(\beta, \sigma^2, R| f_D)}{\partial \theta_j} = -\frac{1}{2} \text{tr}(R_D^{-1} \frac{\partial R_D}{\partial \theta_j})$$

$$+ \frac{1}{2\sigma^2}(f_D - H_D\beta)'R_D^{-1} \frac{\partial R_D}{\partial \theta_j}R_D^{-1}(f_D - H_D\beta).$$
The r.h.s. of the above equation when set to zero generally leads to an intractable system of equations and the system can only be solved using a numerical computation package.

Another alternative to compute the maximum likelihood estimator for $\theta$ is to use some nonlinear optimization routine to maximize the likelihood function (8) or (equivalently) the log-likelihood function (9).
3.8 Types of Asymptotics

Although the use of maximum likelihood estimators for the (unknown) parameters of the autocorrelation function $R(.)$ is more or less the norm in computer experiments, the properties of the maximum likelihood estimators are not well understood here [Sacks, Welch, Mitchell and Wynn (1989), page 418].

There are typically three types of large sample or asymptotic scenarios that are used to study the behavior of statistical procedures. In the first scenario, one observes a random field at an increasing number of sites such that any two sites are at least a fixed distance apart. In this case, the domain over which the random field is observed must eventually be unbounded as sample size tends to infinity. Such scenarios are commonly known as *increasing domain asymptotics*. An example would be a random field observed on an integer lattice.
For the second type, the domain on which one observes the random field remains fixed (and hence bounded) as the number of observations tends to infinity. For example, the domain can be \([0, 1)^d\). As a result, the minimum distance between the data sites tends to zero as sample size increases. This situation is known as \textit{fixed domain asymptotics}. It is also called \textit{infill asymptotics} [see for example Cressie (1993)].

The third scenario is called \textit{mixed asymptotics} which can be thought to be a compromise between increasing domain and fixed domain asymptotics. In this case as sample size increases to infinity, the domain is unbounded while at the same time the minimum distance between data sites tends to zero.
Most of the literature on spatial asymptotics assumes the increasing domain formulation. Typically, results under increasing domain asymptotics turn out to be spatial or multi-dimensional analogs of well known results on time series.

In the context of computer experiments, it is widely acknowledged that fixed domain asymptotics is the appropriate assumption to make [see for example, Sacks, Welch, Mitchell and Wynn (1989) and Stein (1999)]. Unfortunately, not much is known about the large sample properties of standard statistical inference procedures (e.g., maximum likelihood estimation) under fixed domain asymptotics.

One reason is that repeated sampling from a fixed bounded domain results in a very strong form of dependence among the observations and that the dependence does not decrease to zero between any two observations. This leads to quite different and often surprising (relative to classical results) limiting behaviors of the statistical procedures of interest.
We shall now report on the properties of some usual statistical procedures when the design or sample size is large (i.e., $n$ tends to infinity) under fixed domain asymptotics.

### 3.8.1 Least squares estimator

Consider the following kriging model of Section 3.3.

$$f(x) = \beta' h(x) + Z(x), \quad x \in [0, 1)^d,$$

where $\beta \in \mathbb{R}^k$ is a vector of unknown parameters, $h(.)$ is a vector of known functions and $Z(.)$ is a mean zero, *intrinsically stationary* random field on $[0, 1)^d$, i.e.,

$$E[Z(x)] = 0, \quad x \in [0, 1)^d,$$

$$E[Z(x) - Z(x + h)]^2 = 2\gamma_Z(h), \quad x, x + h \in [0, 1)^d,$$

for some function $\gamma_Z(.)$.

The function $2\gamma_Z(.)$ is known as the variogram and $\gamma_Z(.)$ is the semivariogram of the random field $Z(.)$. Also note that the assumption of intrinsic stationarity is weaker than weak (or second order) stationarity.
Now suppose $f(.)$ is observed on a regular grid, i.e.,

$$f\left(\frac{i_1 - 1}{m}, \ldots, \frac{i_d - 1}{m}\right), \quad 1 \leq i_j \leq m, 1 \leq j \leq d.$$ 

Note that the sample size in this case is $n = m^d$ and that the observation sites become dense in $[0, 1)^d$ as $n \to \infty$. For simplicity, we shall label these sites as $x_1, \ldots, x_n$. Then the least squares estimator for $\beta$ is given as

$$\hat{\beta}_{LS} = \left[\sum_{i=1}^{n} h(x_i)h(x_i)\right]^{-1} \sum_{i=1}^{n} h(x_i)f(x_i).$$

It is assumed that $h$ is chosen such that $\sum_{i=1}^{n} h(x_i)h(x_i)'$ is nonsingular.
The following theorem is taken from Lahiri (1996), page 406.

**Theorem 16** Let $h_1, \ldots, h_k$ denote the components of the function $h$. Assume that $h_1, \ldots, h_k$ are continuous on $[0, 1]^d$ and that the matrix

$$A = \int_{[0,1]^d} h(x)h(x)'dx$$

is nonsingular. If in addition $\gamma_Z(h)$ is continuous at $h = 0$, then $\hat{\beta}_{LS}$ converges in mean square to the random vector

$$\hat{\beta}_\infty = A^{-1} \int_{[0,1]^d} h(x)f(x)dx.$$ 

The random integral $\int_{[0,1]^d} h(x)f(x)dx$ is defined in the mean square sense [see for example Cramer and Leadbetter (1967)].
The limiting random vector $\hat{\beta}_\infty$ has mean $\beta$ and covariance matrix

$$\text{Cov}(\hat{\beta}_\infty, \hat{\beta}_\infty) = A^{-1} \left[ \int_{[0,1)^{2d}} h(x)h(y)' \text{Cov}(f(x), f(y)) dx dy \right] A^{-1}.$$  

This implies that $\hat{\beta}_\infty$ is nondegenerate and hence we conclude from Theorem 16 that the least squares estimator $\hat{\beta}_{LS}$ for $\beta$ is not consistent as $n \to \infty$.

**Remark.** In contrast, it should be noted that the least squares estimator for $\beta$ is consistent and asymptotically normal under (more customary) increasing domain asymptotics [see for example Mardia and Marshall (1984)].
3.8.2 Method of moments variogram estimator

Suppose that \( \{Z(x) : x \in [0, 1)^d \} \) is an intrinsically stationary random field with variogram given by

\[
2\gamma_Z(h) = \text{Var}[Z(x) - Z(x + h)], \quad x, x + h \in [0, 1)^d.
\]

**Remark.** Recall that \( Z(.) \) is intrinsically stationary if \( E[Z(x + h) - Z(x)] = 0 \) and \( \text{Var}[Z(x + h) - Z(x)] \) does not depend on \( x \) [see Cressie (1993) for more details].

Matheron (1962) proposed the following method of moments estimator for the variogram.

\[
2\hat{\gamma}_{Z,n}(h) = \frac{1}{n_h} \sum^* [Z(x_i + h) - Z(x_i)]^2,
\]

where the summation \( \sum^* \) extends over all the design sites such that \( x_i, x_i + h \in [0, 1)^d \) and \( n_h \) is the number of such sites.

Consistency and asymptotic normality of \( 2\hat{\gamma}_{Z,n}(h) \) is well known under increasing domain asymptotics [e.g., Davis and Borgman (1982)].
We shall now consider the behavior of $2\hat{\gamma}_{Z,n}(h)$ under fixed domain asymptotics. As in subsection 3.8.1, suppose $\{Z(x) : x \in [0, 1]^d\}$ is observed on a regular grid. Furthermore, for simplicity suppose that the sites are nested as sample size $n \to \infty$. Then Lahiri (1996), page 408, proved

**Theorem 17** Let $h = x_i - x_j$ for some data sites $x_i, x_j \in (0, 1)^d$. Suppose $\gamma_Z(h)$ is continuous at $h = 0$. Then there exists a random variable $\hat{\gamma}_{Z,\infty}(h)$ such that

$$E\hat{\gamma}_{Z,\infty}(h) = \gamma_Z(h),$$

$$E|\hat{\gamma}_{Z,n}(h) - \hat{\gamma}_{Z,\infty}(h)| \to 0,$$

as $n \to \infty$. If in addition,

$$E[Z(x + h) - Z(x)]^4 \to 0,$$

as $\|h\| \to 0$ for all $x \in (0, 1)^d$, then

$$2\hat{\gamma}_{Z,\infty}(h) = \frac{1}{\text{Vol}.R_1} \int_{R_1} [Z(x + h) - Z(x)]^2 dx,$$

where $R_1 = \{x \in (0, 1)^d : x + h \in (0, 1)^d\}$. 

149
An immediate consequence of Theorem 17 is that the method of moments variogram estimator is not consistent under fixed domain asymptotics.

3.8.3 Empirical c. d. f.

Let \( \{Z(x) : x \in [0, 1]^d \} \) be an intrinsically stationary random field and that \( Z(.) \) is observed at sites \( x_1, \cdots, x_n \in [0, 1]^d \) on a regular grid (as in subsection 3.8.1). The usual empirical cumulative distribution function (c.d.f.) is defined as

\[
\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I\{Z(x_i) \leq t\}, \quad t \in \mathbb{R},
\]

where \( I\{.\} \) denotes the indicator function.

Lahiri (1996) proved that, unlike increasing domain asymptotics, \( \hat{F}_n \) converges to a nondegenerate random cumulative distribution function under fixed domain asymptotics.
More precisely,

**Theorem 18** Let \( \{Z(x) : x \in [0, 1)^d\} \) be an intrinsically stationary random field such that

\[
\gamma_Z(h) \to 0 \text{ as } \|h\| \to 0.
\]

Then \( \hat{F}_n \) converges in probability to \( \hat{F}_\infty \) as \( n \to \infty \) where

\[
\hat{F}_\infty(t) = \int_{[0,1)^d} I\{Z(x) \leq t\} dx, \quad t \in R.
\]

**Remark.** In contrast, recall that in the case of independent and identically distributed observations, the Glivenko-Cantelli theorem [see for example Chung (1974), page 133] states that the empirical c.d.f. converges to the true c.d.f. uniformly pointwise as sample size tends to infinity.

Since many standard statistical procedures are based on statistics of the form \( T(\hat{F}_n) \), Theorem 18 gives an explanation for the inconsistency of many statistical procedures under fixed domain asymptotics. For a detailed discussion of these ideas, see Lahiri (1996).
This also leads us to the following question: are there consistent statistical procedures under fixed domain asymptotics? The answer is yes and the next few sections shall elaborate on this. However to do that, we first need some further definitions.
3.9 Some Further Definitions

Suppose $Z(x)$ and $\tilde{Z}(x)$, $x \in [0, 1]$, are two Gaussian processes. Let $P$ and $\tilde{P}$ denote their respective induced (Gaussian) probability measures.

Definition. We say $P$ and $\tilde{P}$ are mutually absolutely continuous if we have $\tilde{P}(A) = 0$ for any measurable set $A$ satisfying $P(A) = 0$, and vice versa. The probability measures $P$ and $\tilde{P}$ are said to be orthogonal if there is a measurable set $A$ such that $P(A) = \tilde{P}(A^c) = 1$. Here $A^c$ denotes the complement of $A$.

In the case of Gaussian processes, it is well known that $P$ and $\tilde{P}$ are either mutually absolutely continuous or orthogonal [see for example Stein (1999)].

We assume that the processes $Z(.)$ and $\tilde{Z}(.)$ are observed at sites $x_1, \cdots, x_n \in [0, 1)$ on a regular grid. Let $p_n$ and $\tilde{p}_n$ be $n$-variate normal densities corresponding to the distributions of $(Z(x_1), \cdots, Z(x_n))$ and $(\tilde{Z}(x_1), \cdots, \tilde{Z}(x_n))$ respectively.
Then the \textit{divergence} between the two (Gaussian) probability measures $P$ and $\tilde{P}$ is defined as

$$D(P, \tilde{P}) = \lim_{n \to \infty} D(p_n, \tilde{p}_n),$$

where

$$D(p_n, \tilde{p}_n) = E[\log \frac{p_n(Z(x_1), \cdots Z(x_n))}{\tilde{p}_n(Z(x_1), \cdots, Z(x_n))}] + E[\log \frac{\tilde{p}_n(\tilde{Z}(x_1), \cdots \tilde{Z}(x_n))}{p_n(\tilde{Z}(x_1), \cdots, \tilde{Z}(x_n))}].$$

We note from Ibragimov and Rozanov (1978), page 77, that $P$ and $\tilde{P}$ are orthogonal if and only if

$$D(P, \tilde{P}) = \infty,$$

and that $P$ and $\tilde{P}$ are mutually absolutely continuous if and only if

$$D(P, \tilde{P}) < \infty.$$

A point to note is that if $P$ and $\tilde{P}$ are mutually absolutely continuous, then it is not possible with certainty to differentiate the two probability measures even if $n \to \infty.$
Consider two stationary Gaussian processes $Z, \tilde{Z}$ on $[0, 1)$ with variances $\sigma^2, \tilde{\sigma}^2$ and correlation functions $R_{\theta}, R_{\tilde{\theta}}$. Then

$$D(p_n, \tilde{p}_n) = \frac{1}{2} \left\{ \frac{\tilde{\sigma}^2}{\sigma^2} \text{tr}(R^{-1}\tilde{R}) + \frac{\sigma^2}{\tilde{\sigma}^2} \text{tr}(\tilde{R}^{-1}R) \right\} - n,$$

where $R$ is the $n \times n$ correlation matrix with $(k,l)$th element $R_{\theta}(x_l - x_k)$ and $\tilde{R}$ is the $n \times n$ correlation matrix with $(k,l)$th element $R_{\tilde{\theta}}(x_l - x_k)$.

The above definitions and formulas can be easily generalized to Gaussian random fields $Z$ and $\tilde{Z}$ defined on $[0, 1)^d$ if the multiplicative correlation function is used, say

$$R_{\theta} = \prod_{j=1}^{d} R_{j,\theta}, \quad R_{\tilde{\theta}} = \prod_{j=1}^{d} R_{j,\tilde{\theta}},$$

respectively, where $R_{j,\theta}$ and $R_{j,\tilde{\theta}}$ are univariate correlation functions.
Then the divergence between the associated probability densities $p_n$ and $\tilde{p}_n$ is given by

$$D(p_n, \tilde{p}_n) = \frac{1}{2} \left\{ \frac{\tilde{\sigma}^2}{\sigma^2} \prod_{j=1}^{d} \text{tr}(R_j^{-1} \tilde{R}_j) + \frac{\sigma^2}{\tilde{\sigma}^2} \prod_{j=1}^{d} \text{tr}(\tilde{R}_j^{-1} R_j) \right\} - n^d,$$

where $R_j$ is the $n \times n$ correlation matrix with $(k, l)$th element $R_{j,\theta}(x_l - x_k)$, etc.
3.9.1 Triangular correlation revisited

First, let \( d = 1 \). Let \( Z \) and \( \tilde{Z} \) be two stationary Gaussian processes on \([0, 1]\) with mean 0 and triangular covariance functions

\[
R_\theta(s - t) = \sigma^2(1 - \theta|s - t|),
\]
\[
R_{\tilde{\theta}}(s - t) = \tilde{\sigma}^2(1 - \tilde{\theta}|s - t|), \quad 0 \leq s, t < 1,
\]

\( \theta, \tilde{\theta} \in (0, 2) \) respectively. With the notation of Section 3.9, it can be shown that

\[
\text{tr}(R^{-1}\tilde{R}) = (n - 2)\frac{\tilde{\theta}}{\theta} + 2\frac{\theta + \tilde{\theta} - \theta\tilde{\theta}}{\theta(2 - \theta)}.
\]

Hence \( D(P, \tilde{P}) < \infty \) if and only if \( \theta\sigma^2 = \tilde{\theta}\tilde{\sigma}^2 \) and \( D(P, \tilde{P}) = \infty \) otherwise. This implies that \((\theta, \sigma^2)\) cannot be consistently estimated. However \( \theta\sigma^2 \) is identifiable.

Let \( \hat{\theta} \) and \( \hat{\sigma}^2 \) be the maximum likelihood estimators for \( \theta \) and \( \sigma^2 \). Then the maximum likelihood estimator for \( \theta\sigma^2 \) (based on observations of \( Z(.) \) on a regular grid) is given by

\[
\hat{\theta}\hat{\sigma}^2 = \frac{1}{2} \sum_{i=1}^{n-1} [Z(x_{i+1}) - Z(x_i)]^2.
\]
We observe that
\[ E(\hat{\theta} \hat{\sigma}^2) = \theta \sigma^2, \]
\[ \text{Var}(\hat{\theta} \hat{\sigma}^2) = \frac{2(\theta \sigma^2)^2}{n - 1}. \]
Using Chebyshev’s inequality, this implies that the maximum likelihood estimator \( \hat{\theta} \hat{\sigma}^2 \to \theta \sigma^2 \) in probability as \( n \to \infty \).

Next let \( d \geq 2 \). Let \( Z \) and \( \tilde{Z} \) be two stationary mean zero Gaussian random fields on \([0, 1]^d\) with product triangular covariance functions
\[ R_\theta(s - t) = \sigma^2 \prod_{j=1}^{d} (1 - \theta|s_j - t_j|), \]
\[ R_{\tilde{\theta}}(s - t) = \tilde{\sigma}^2 \prod_{j=1}^{d} (1 - \tilde{\theta}|s_j - t_j|), \quad s, t \in [0, 1]^d, \]
respectively. Then \( D(P, \tilde{P}) = \infty \) unless
\[ \theta_1 = \tilde{\theta}_1, \ldots, \theta_d = \tilde{\theta}_d \text{ and } \sigma^2 = \tilde{\sigma}^2. \]
This implies that the parameters \( \sigma^2, \theta_1, \ldots, \theta_d \) are all identifiable.
Assuming that $Z(.)$ is observed on a regular grid, Abt and Welch (1998), page 132, noted that [using the techniques in Ying (1993)] the maximum likelihood estimators for the above parameters are strongly consistent and asymptotically normal as sample size tends to infinity.

It should be noted the attractiveness of considering the triangular covariance function is that it is analytically tractable. However the triangular correlation function has limited practical applications since the resulting best linear unbiased predictor (BLUP) (see Section 3.3) is a linear interpolating spline. This also implies that BLUP does not depend on the model parameters $\sigma^2, \theta_1, \cdots, \theta_d$. 
3.9.2 Exponential correlation

First, let $d = 1$. Let $Z$ and $\tilde{Z}$ be two stationary Gaussian processes on $[0, 1]$ with mean 0 and exponential covariance functions

$$R_\theta(s - t) = \sigma^2 \exp(-\theta |s - t|),$$
$$R_{\tilde{\theta}}(s - t) = \tilde{\sigma}^2 \exp(-\tilde{\theta} |s - t|), \quad 0 \leq s, t < 1,$$

$\theta, \tilde{\theta} \in (0, \infty)$ respectively. It can be shown that $D(P, \tilde{P}) < \infty$ if and only if $\theta \sigma^2 = \tilde{\theta} \tilde{\sigma}^2$ and $D(P, \tilde{P}) = \infty$ otherwise. This implies that $(\theta, \sigma^2)$ cannot be consistently estimated. However $\theta \sigma^2$ is identifiable.

**Remark.** We observe that the boundedness property of the interval $[0, 1]$ is crucial here because if the domain of $Z(.)$ is $(0, \infty)$, the identifiability problem of the parameters $\theta$ and $\sigma^2$ disappears [see Ibragimov and Rozanov (1978), page 95].
We assume that $Z(.)$ is observed at sites $0 \leq x_1 < \cdots < x_n < 1$. In this case, the $x_i$'s need not be in a regular grid. Using the Markovian property of the Ornstein-Uhlenbeck process $Z(x), \ x \in [0, 1)$, Ying (1991), page 283, showed that the log-likelihood function is given explicitly as

$$l_n(\theta, \sigma^2) = \frac{1}{\sigma^2}Z^2(x_1) + \sum_{i=2}^{n} \frac{[Z(x_i) - e^{-\theta(x_i-x_{i-1})}Z(x_{i-1})]^2}{\sigma^2[1 - e^{-2\theta(x_i-x_{i-1})}]} + \log(\sigma^2) + \sum_{i=2}^{n} \log[\sigma^2[1 - e^{-2\theta(x_i-x_{i-1})}]] + n \log(2\pi).$$

The following theorem is due to Ying (1991), page 284, who proved the strong consistency and asymptotic normality of the maximum likelihood estimator for $\theta \sigma^2$. 

161
Theorem 19 Let $0 < a \leq b < \infty$ be fixed and let $J$ be either $[0, \infty) \times [a, b]$ or $[a, b] \times (0, \infty)$. Define $(\hat{\theta}, \hat{\sigma}^2)$ as a solution of

$$l_n(\hat{\theta}, \hat{\sigma}^2) = \sup_{(\tilde{\theta}, \tilde{\sigma}^2) \in J} l_n(\tilde{\theta}, \tilde{\sigma}^2).$$

Then, with probability 1, $(\hat{\theta}, \hat{\sigma}^2)$ exists for all large $n$ and

$$\hat{\theta}\hat{\sigma}^2 \to \theta\sigma^2, \text{ a.s.}$$

Finally,

$$\sqrt{n}(\hat{\theta}\hat{\sigma}^2 - \theta\sigma^2) \to N(0, 2(\theta\sigma^2)^2)$$

in distribution as $n \to \infty$.

Remark. It is a rather remarkable fact that the results of the above theorem is independent of the choice of design sites $0 \leq x_1 < \cdots < x_n < 1$.

The main difficulty in the proof of Theorem 19 is that as $n$ increases, $Z(x_1), \cdots, Z(x_n)$ becomes highly correlated, thus making its covariance matrix ill-conditioned. The detailed proof can be found in Ying (1991).
Using the reparametrization \((\theta\sigma^2, \sigma^2)\), Abt and Welch (1998) showed that
\[
\lim_{n \to \infty} n(J_{\theta\sigma^2, \sigma^2}^{-1})_{1,1} = 2(\theta\sigma^2)^2,
\]
where \(J_{\theta\sigma^2, \sigma^2}\) is the \(2 \times 2\) Fisher information matrix. Theorem 19 implies that the variance of the asymptotic distribution of \(\sqrt{n}(\hat{\theta}\hat{\sigma}^2 - \theta\sigma^2)\) is equal to the right hand side of the above equation.

For the case \(d \geq 2\), let \(Z\) and \(\tilde{Z}\) be two stationary mean zero Gaussian random fields on \([0, 1)^d\) with product exponential covariance functions
\[
R_\theta(s - t) = \sigma^2 \prod_{j=1}^{d} \exp(-\theta_j|s_j - t_j|),
\]
\[
R_{\tilde{\theta}}(s - t) = \tilde{\sigma}^2 \prod_{j=1}^{d} \exp(-\tilde{\theta}_j|s_j - t_j|), \quad s, t \in [0, 1)^d,
\]
respectively. Then \(D(P, \tilde{P}) = \infty\) unless
\[
\theta_1 = \tilde{\theta}_1, \ldots, \theta_d = \tilde{\theta}_d \text{ and } \sigma^2 = \tilde{\sigma}^2.
\]
This implies that the parameters \(\theta_1, \ldots, \theta_d, \sigma^2\) are all identifiable.
Assume that \( Z(.) \) is observed on a regular grid. Then Ying (1993) proved that the maximum likelihood estimators for \( \theta_1, \cdots, \theta_d, \sigma^2 \) are strongly consistent and asymptotically normal.

More precisely in the case \( d = 2 \), Ying (1993) proved that

**Theorem 20** Let \( \theta_1, \theta_2 \) and \( \sigma^2 \) be the true parameters and \( C \) be a compact region in \( (0, \infty) \times (0, \infty) \) that contains \( (\theta_1, \theta_2) \) as an interior point. Then \( (\hat{\theta}_1, \hat{\theta}_2, \hat{\sigma}^2) \), the maximum likelihood estimator that maximizes the likelihood function over \( C \times (0, \infty) \), is strongly consistent, i.e.,

\[
(\hat{\theta}_1, \hat{\theta}_2, \hat{\sigma}^2) \to (\theta_1, \theta_2, \sigma^2) \quad \text{a.s.}
\]

Furthermore,

\[
n^{1/4}[(\hat{\theta}_1, \hat{\theta}_2, \hat{\sigma}^2) - (\theta_1, \theta_2, \sigma^2)] \to N_3(0, \Sigma)
\]

in distribution as sample size \( n \to \infty \) where \( \Sigma \) is a \( 3 \times 3 \) covariance matrix of rank 2.
Abt and Welch (1998) showed that

$$\Sigma = \lim_{n \to \infty} n^{1/2} J^{-1}_{\theta_1, \theta_2, \sigma^2},$$

where $J_{\theta_1, \theta_2, \sigma^2}$ is the associated Fisher information matrix.

Finally, van der Vaart showed, via certain reparametrization and asymptotic expansions, that the maximum likelihood estimators are asymptotically efficient in that the usual convolution and asymptotic minimax properties hold [see Ibragimov and Hasminskii (1981) and LeCam (1986) for background on asymptotic efficiency of estimators based on dependent observations].
3.9.3 Gaussian correlation

Let $X$ be a stationary Gaussian random field on $[0, 1)^d$ with mean 0 and multiplicative Gaussian covariance function

$$\text{Cov}(X(x), X(y)) = \sigma^2 \prod_{j=1}^{d} \exp[-\theta_j (x_j - y_j)^2], \quad x, y \in [0, 1)^d.$$ 

where $\theta_1, \cdots, \theta_d$ and $\sigma^2$ are unknown positive parameters. It is well-known [see for example Abt and Welch (1998)] that $\theta_1, \cdots, \theta_d$ and $\sigma^2$ are identifiable [or microergodic using the terminology of Stein (1999), page 163].

The main aim of this subsection is to estimate the correlation parameters $\theta_1, \cdots, \theta_d$ using observations of $X(.)$ that are taken on a regular lattice, i.e.,

$$\left\{ X(\frac{i_1}{n}, \cdots, \frac{i_d}{n}) : 1 \leq i_u \leq n, 1 \leq u \leq d \right\},$$

where $n$ is a positive integer.
For simplicity, we order the above observations lexicographically as a $n^d \times 1$ column vector $\tilde{X}_n$. Thus the element $X(i_1/n, \cdots, i_d/n)$ precedes the element $X(j_1/n, \cdots, j_d/n)$ if and only if there exists a $1 \leq k \leq d$ such that $i_u = j_u$ whenever $1 \leq u < k$ and $i_k < j_k$.

Then the covariance matrix of $\tilde{X}_n$ is given by

$$
\Sigma_{\theta_1, \cdots, \theta_d, \sigma^2; n} = \sigma^2 \otimes_{u=1}^{d} R_{\theta_u, n},
$$

where the symbol $\otimes$ denotes Kronecker product [see Anderson (1984), page 599] and for each $1 \leq u \leq d$, $R_{\theta_u, n}$ denotes the $n \times n$ matrix whose $(i, j)$th element is $\exp[-\theta_u (i - j)^2 / n^2]$.

The estimation of $\theta_1, \cdots, \theta_d$ reduces to the estimation of the correlation matrix $\otimes_{u=1}^{d} R_{\theta_u, n}$.

Since $\tilde{X}_n \sim N_{n^d}(0, \Sigma_{\theta_1, \cdots, \theta_d, \sigma^2; n})$, the log-likelihood function $l_n(\theta_1, \cdots, \theta_d, \sigma^2)$ is given by

$$
2l_n(\theta_1, \cdots, \theta_d, \sigma^2) = -n^d \log(2\pi) - n^d \log(\sigma^2) - \log | \otimes_{u=1}^{d} R_{\theta_u, n}| - \frac{1}{\sigma^2} \tilde{X}_n' \left( \otimes_{u=1}^{d} R_{\theta_u, n} \right)^{-1} \tilde{X}_n.
$$
To work with the log-likelihood, we evaluate the determinant $\big| \bigotimes_{u=1}^{d} R_{\theta_u,n} \big|$ and the inverse $(\bigotimes_{u=1}^{d} R_{\theta_u,n})^{-1}$. Let

$$w_u = \exp(-\theta_u/n^2), \quad 1 \leq u \leq d.$$ 

We observe that $R_{\theta_u,n}$ is a patterned matrix given explicitly as

$$R_{\theta_u,n} = \begin{pmatrix}
1 & w_u & w_u^4 & \cdots & w_u^{(n-1)^2} \\
w_u & 1 & w_u & \cdots & w_u^{(n-2)^2} \\
w_u^4 & w_u & 1 & \cdots & w_u^{(n-3)^2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
w_u^{(n-1)^2} & w_u^{(n-2)^2} & w_u^{(n-3)^2} & \cdots & 1
\end{pmatrix}.$$

**Definition.** The Gaussian polynomial $G(m, n; q)$ is defined by

$$G(m, n; q) = \frac{(1 - q^{n-m+1})(1 - q^{n-m+2}) \cdots (1 - q^n)}{(1 - q)(1 - q^2) \cdots (1 - q^m)}$$

if $0 \leq m \leq n$ and $G(m, n; q) = 0$ otherwise.

$G(m, n; q)$’s are also known as $q$-binomial coefficients in the combinatorics literature [see Chapter 3 of Andrews (1976)].
By making use of the Cholesky decomposition of $R_{\theta,u,n}$, Loh and Lam (2000) proved that

**Proposition 2** The determinant of the $n^d \times n^d$ matrix $\bigotimes_{u=1}^{d} R_{\theta,u,n}$ is given by

$$| \bigotimes_{u=1}^{d} R_{\theta,u,n} | = \prod_{u=1}^{d} \left\{ \prod_{k=1}^{n-1} [(1 - w_u^{2k})^{n-k}] \right\}^{n^d-1},$$

where $w_u = \exp(-\theta_u/n^2)$, $1 \leq u \leq d$.

and

**Proposition 3** The inverse of $\bigotimes_{u=1}^{d} R_{\theta,u,n}$ is given by

$$(\bigotimes_{u=1}^{n} R_{\theta,u,n})^{-1} = \bigotimes_{u=1}^{d} (R_{\theta,u,n}^{-1}),$$

where for $1 \leq i, j \leq n$,

$$(R_{\theta,u,n}^{-1})_{i,j} = (-1)^{i-j} \sum_{m=1}^{i\wedge j} \frac{w_u^{i+j-2m}}{\prod_{k=1}^{n-m}(1 - w_u^{2k})} G(i - m, n - m; w_u^2) \times G(j - m, n - m; w_u^2).$$
The use of symbolic computation software (e.g., Mathematica or Maple) features significantly in the motivation and proof of the above two propositions.

Propositions 2 and 3 enable us to obtain an explicit closed-form expression for the log-likelihood function $l_n(\theta_1, \cdots, \theta_d, \sigma^2)$.

An interesting and important point to note is that $R_{\theta_u,n}$ tends to a singular matrix extremely quickly as $n \to \infty$. For example using Proposition 2, a very crude bound for $| \bigotimes_{u=1}^d R_{\theta_u,n} |$ is $O(n^{-d/2})$.

An immediate consequence is that any fixed order (e.g., a first-order) Taylor expansion of the log-likelihood function $l_n(\theta_1, \cdots, \theta_d, \sigma^2)$ will not be able to give a reasonable approximation of the behavior of $l_n$ for large $n$. The noise swamps the signal here. Thus obtaining an exact analytically tractable formula for $l_n$ may be the only way to proceed.
Now we assume that there exist known constants \( \nu, \alpha_u, \beta_u, \ 0 \leq u \leq d \) such that \( 0 < \nu < 1, \ 0 < \alpha_0 < \sigma^2 < \beta_0 < \infty \) and \( 0 < \alpha_u < \theta_u < \beta_u < \infty, \ 1 \leq u \leq d \). In other words, we are assuming that the unknown parameters \( \theta_1, \cdots, \theta_d, \sigma^2 \) lie in a (known) compact subset (bounded away from 0 and infinity) of the parameter space.

**Sieve maximum likelihood estimation**

We define a sieve \( \Theta_n \) on the parameter space of \( (\theta_1, \cdots, \theta_d) \), namely,

\[
\Theta_n = \left\{ \left( \frac{i_1}{n^{\nu}}, \cdots, \frac{i_d}{n^{\nu}} \right) : \alpha_u \leq \frac{i_u}{n^{\nu}} \leq \beta_u, 1 \leq u \leq d \right\}.
\]

The sieve maximum likelihood estimator (MLE) for \( (\theta_1, \cdots, \theta_d) \) is that element \( (\hat{\theta}_{1,n}, \cdots, \hat{\theta}_{d,n}) \in \Theta_n \) such that

\[
\sup \{ l_n(\hat{\theta}_{1,n}, \cdots, \hat{\theta}_{d,n}, \tilde{\sigma}^2) : \tilde{\sigma}^2 \in [\alpha_0, \beta_0] \} = \sup \{ l_n(\tilde{\theta}_1, \cdots, \tilde{\theta}_d, \tilde{\sigma}^2) : \tilde{\sigma}^2 \in [\alpha_0, \beta_0], \ (\tilde{\theta}_1, \cdots, \tilde{\theta}_d) \in \Theta_n \}.
\]
The following strong consistency property of the sieve MLE was established by Loh and Lam (2000).

**Theorem 21** With the above notation and assumptions, let $4/5 \leq \nu < 1$ and

$$(\sigma^2, \theta_1, \cdots, \theta_d) \in \prod_{u=0}^{d}(\alpha_u, \beta_u).$$

Then for all $0 < \rho < 1 - \nu$, with probability 1

$$(\hat{\theta}_{1,n}, \cdots, \hat{\theta}_{d,n}) \in \prod_{u=1}^{d}(\theta_u - d^{-1}\theta_un^{-\rho}, \theta_u + \theta_un^{-\rho})$$

for sufficiently large $n$. 
The proof of Theorem 21 involves essentially two steps [see Loh and Lam (2000) for details]. The first step is to establish that

\[
\frac{1}{n^d} \inf \left\{ l_n(\theta_1, \ldots, \theta_d, \sigma^2) - l_n(\tilde{\theta}_1, \ldots, \tilde{\theta}_d, \tilde{\sigma}^2) : \tilde{\sigma}^2 \in [\alpha_0, \beta_0], \right. \\
\left. (\tilde{\theta}_1, \ldots, \tilde{\theta}_d) \in \Theta_n \setminus \prod_{u=1}^d \left( \theta_u - \frac{\theta_u}{dn^\rho}, \theta_u + \frac{\theta_u}{n^\rho} \right) \right\}
\to \infty,
\]
as \( n \to \infty \).

The second step involves showing that

\[
\frac{1}{n^d} [l_n(\theta_1, \ldots, \theta_d, \sigma^2) - l_n(\tilde{\theta}_1, \ldots, \tilde{\theta}_d, \sigma^2)] \to 0,
\]
as \( n \to \infty \) uniformly over

\[
(\tilde{\theta}_1, \ldots, \tilde{\theta}_d) \in \prod_{u=1}^d \left( \theta_u - \frac{1}{n^\nu}, \theta_u + \frac{1}{n^\nu} \right).
\]
Some related open problems

Theorem 21 shows that the sieve MLE of \((\theta_1, \cdots, \theta_d)\) is strongly consistent. Is the same result true for “honest” maximum likelihood estimators of the \(\theta_j\)’s? We conjecture that it is also true.

The consistency of the MLE for \(\sigma^2\) is still an open question in the case of Gaussian correlation functions. The analysis of the log-likelihood function indicates that in terms of accuracy, maximum likelihood estimation of \(\sigma^2\) is an order more difficult than sieve likelihood estimation of the \(\theta_j\)’s.

Abt and Welch (1998) reported that numerical simulations indicate (but not prove) that the maximum likelihood estimators for \(\sigma^2\) as well as the \(\theta_j\)’s are consistent.

Finally, since the log-likelihood function has a relatively simple closed-form expression, we believe the above questions have a good chance of being resolved through more research work.
3.9.4 Matérn correlation

This section reports on my current (not complete) on-going work on fixed-domain asymptotics. Recall from Section 3.5.6 that the Matérn family of multiplicative covariance functions on $[0, 1)^d$ can be written as

$$
\prod_{t=1}^{d} \frac{\pi^{1/2} \phi}{\Gamma(\alpha + 1/2)2^{\alpha-1}\theta_t^{2\alpha}} (\theta_t | x_t - y_t |)^\alpha K_\alpha(\theta_t | x_t - y_t |),
$$

$\forall x, y \in [0, 1)^d$, where $K_\alpha(.)$ is a modified Bessel function of order $\nu$ [see also Stein (1999), page 31]. In this section, we take $\alpha = 3/2$. It is well known that $K_{3/2}(s) = (\pi s^{-3/2})^{1/2}(1 + s)e^{-s}$.

More precisely, let $X$ be a stationary Gaussian random field on $[0, 1)^d$ with mean 0 and multiplicative Matérn-type covariance function

$$
\text{Cov}(X(x), X(y)) = \frac{\pi^d \phi^d}{2^d \theta_1^3 \cdots \theta_d^3 \prod_{t=1}^{d} (1 + \theta_t | x_t - y_t |) e^{-\theta_t | x_t - y_t |}},
$$

for all $x, y \in [0, 1)^d$. We observe from Section 3.5.6 that the above Gaussian random field is exactly once mean square differentiable and that its sample path is also exactly once differentiable almost surely.
Stein (1999), page 164, that if $d = 1$, $\phi$ is microergodic (or identifiable) while $\theta_1$ is not. Hence $\theta_1$ cannot be estimated consistently.

As usual, we assume that observations of $X(.)$ are made on a regular grid, i.e.,

$$\left\{ X\left( \frac{i_1}{n}, \ldots, \frac{i_d}{n} \right) : 1 \leq i_1, \ldots, i_d \leq n \right\},$$

where $n$ is a positive integer. The observations are ordered lexicographically as a $n^d \times 1$ vector $\tilde{X}_n$ [see Section 3.9.3]. The covariance matrix $\Sigma_{\phi, \theta_1, \ldots, \theta_d; n}$ of $\tilde{X}_n$ is given by

$$\Sigma_{\phi, \theta_1, \ldots, \theta_d; n} = \frac{\pi^d \phi^d}{2^d \theta_1^3 \cdots \theta_d^3} \otimes_{t=1}^d R_{\theta_t, n},$$

where $\otimes$ denotes the Kronecker product and for each $1 \leq t \leq d$, $R_{\theta_t, n}$ is the $n \times n$ matrix whose $(i, j)$th element is $(1 + \theta_t|i - j|/n) \exp(-\theta_t|i - j|/n)$.
In particular, writing \( w = \theta / n \) and \( u = e^{-w} \), we have

\[
R_{\theta, n} = \begin{pmatrix}
1 & (1 + w)u & \cdots \\
(1 + w)u & 1 & \cdots \\
(1 + 2w)u^2 & (1 + w)u & \cdots \\
\vdots & \vdots & \ddots \\
(1 + (n - 1)w)u^{n-1} & (1 + (n - 2)w)u^{n-2} & \cdots & 1
\end{pmatrix}.
\]

We are concerned with the estimation of the parameters \( \phi, \theta_1, \cdots, \theta_d \). Since \( \tilde{X}_n \sim N_{n^d}(0, \Sigma_{\phi, \theta_1, \cdots, \theta_d ; n}) \), the likelihood function \( L_n(\phi, \theta_1, \cdots, \theta_d) \) satisfies

\[
2 \log L_n(\phi, \theta_1, \cdots, \theta_d)
= -n^d \log(2\pi) - n^d \log\left(\frac{\pi^d \phi^d}{2^d \theta_1^3 \cdots \theta_d^3}\right)
- \log \left| \bigotimes_{t=1}^d R_{\theta_t, n} \right|
- \frac{2^d \theta_1^3 \cdots \theta_d^3}{\pi^d \phi^d} \tilde{X}'_n (\bigotimes_{t=1}^d R_{\theta_t, n})^{-1} \tilde{X}_n.
\] (10)
As in the Gaussian correlation case, it is not enough to approximate the log-likelihood function by its one or two-term Taylor series expansion. Fortunately using symbolic computation software, one can obtain exact closed-form formulas for the determinant $|R_{\theta_t,n}|$ and the inverse $R_{\theta_t,n}^{-1}$. These expressions unfortunately are not simple enough for manual theoretical analysis. However they are amenable to analysis via Mathematica or Maple.

Fisher information

Section 6.6 of Stein (1999) reports numerical calculations of the Fisher information matrix for a mean 0 Gaussian process with a Matérn-type covariance function on the real line. This motivates the following result.
Theorem 22 Let the likelihood $L_n$ be as in (10). Then the elements of the Fisher information matrix for the parameters are given by

$$-E[\frac{\partial^2}{\partial \phi^2} \log L_n(\phi, \theta_1, \cdots, \theta_d)]$$

$$= \frac{d^2n^d}{2\phi^2},$$

$$-E[\frac{\partial^2}{\partial \theta_t^2} \log L_n(\phi, \theta_1, \cdots, \theta_d)]$$

$$= \frac{n^{d-1}(2\theta_t + 5)}{\theta_t^2} + O(n^{d-2}),$$

$$-E[\frac{\partial^2}{\partial \phi \partial \theta_t} \log L_n(\phi, \theta_1, \cdots, \theta_d)]$$

$$= -\frac{n^{d-1}(\theta_t + 2)}{\phi \theta_t} + O(n^{d-2}),$$

and

$$-E[\frac{\partial^2}{\partial \theta_s \partial \theta_t} \log L_n(\phi, \theta_1, \cdots, \theta_d)] = O(n^{d-2}),$$

as $n \to \infty$ whenever $1 \leq s \neq t \leq d$. 
We observe from Theorem 22 that in the case $d = 1$, the Fisher information for $\theta_1$ is bounded whereas the Fisher information for $\phi$ tends to infinity as $n \to \infty$. This is consistent with the fact that $\theta_1$ cannot be consistently estimated under fixed-domain asymptotics.

In contrast when $d \geq 2$, the Fisher information for $\phi$ and $\theta_t$, $1 \leq t \leq d$, all tend to infinity as $n \to \infty$. This indicates the possibility that the parameters can be estimated consistently here.

**Estimation of $\phi$**

Let $\hat{X}_n \sim N_{n^d}(0, \Sigma_{\phi, \theta_1, \ldots, \theta_d}; n)$ be as before. Also let $0 < \beta_{0,t} < \beta_{1,t} < \infty$, $t = 0, \ldots, d$, be known constants such that $\beta_{0,0} < \phi < \beta_{1,0}$ and $\beta_{0,t} < \theta_t < \beta_{1,t}$, $t = 1, \ldots, d$. 
We shall now construct a consistent maximum likelihood type estimator for the scale parameter \( \phi \). First set the estimator for \((\theta_1, \cdots, \theta_d)\) to be a known arbitrary but fixed vector, say

\[
(\tilde{\theta}_1, \cdots, \tilde{\theta}_d) \in \prod_{t=1}^{d} (\beta_{0,t}, \beta_{1,t})
\]

and define the estimator \( \hat{\phi}_{\tilde{\theta}_1, \cdots, \tilde{\theta}_d} \) for \( \phi \) to be that value of \( \tilde{\phi} \) that maximizes the likelihood function \( L_n(\tilde{\phi}, \tilde{\theta}_1, \cdots, \tilde{\theta}_d) \). In particular,

\[
\hat{\phi}_{\tilde{\theta}_1, \cdots, \tilde{\theta}_d} = \left\{ \frac{2^d \tilde{\theta}_1^3 \cdots \tilde{\theta}_d^3}{\pi^d n^d} \tilde{X}_n' \left( \otimes_{t=1}^{d} R_{\tilde{\theta}_t,n} \right)^{-1} \tilde{X}_n \right\}^{1/d}.
\]
We have the following result on $\hat{\phi}_{\tilde{\theta}_1, \ldots, \tilde{\theta}_d}$.

**Theorem 23** With the above notation and conditions,

$$
\frac{2^d \tilde{\theta}_1^3 \cdots \tilde{\theta}_d^3}{\pi^d n^d} E[\tilde{X}'_n (\otimes_{t=1}^d R_{\tilde{\theta}_t,n})^{-1} \tilde{X}_n] = \phi^d + O(n^{-1}),
$$

and

$$
\text{Var} \left\{ \frac{2^d \tilde{\theta}_1^3 \cdots \tilde{\theta}_d^3}{\pi^d n^d} \tilde{X}'_n (\otimes_{t=1}^d R_{\tilde{\theta}_t,n})^{-1} \tilde{X}_n \right\} = O(n^{-d}),
$$

as $n \to \infty$ uniformly over

$$(\phi, \theta_1, \cdots, \theta_d) \in \prod_{t=0}^d (\beta_0, t, \beta_1, t).$$

Consequently, $\hat{\phi}_{\tilde{\theta}_1, \ldots, \tilde{\theta}_d} \to \phi$ in probability as $n \to \infty$. 

182
It is interesting to note that incorrect specification of the values of \( \theta_1, \ldots, \theta_d \) can still lead to consistent estimation of \( \phi \). This may not be surprising as Theorem 22 shows that the amount of Fisher information contained in the sample on \( \phi \) is an order of magnitude greater than those on \( \theta_1, \ldots, \theta_d \).

This phenomenon has been observed quite broadly in other fixed-domain asymptotic situations [see for example, Stein (1999), page 175]. In short, parameters that cannot be consistently estimated can be assigned arbitrary values without affecting the consistent estimation of the remaining parameters.
Sieve maximum likelihood estimation

Let $d \geq 3$. This section deals with sieve maximum likelihood estimation of the parameters $\phi, \theta_1, \cdots, \theta_d$. Let $0 < \nu < (d-2)/(d+1)$ and $0 < \beta_{0,t} < \beta_{1,t} < \infty$, $t = 0, \cdots, d$, be known constants such that $\beta_{0,0} < \phi < \beta_{0,1}$ and $\beta_{0,t} < \theta_t < \beta_{1,t}$, $t = 1, \cdots, d$.

We define a sieve $\Omega_n$ on the parameter space of $(\phi, \theta_1, \cdots, \theta_d)$, namely,

$$\Omega_n = \left\{ \left( \frac{i_0}{n^{\nu}}, \cdots, \frac{i_d}{n^{\nu}} \right) : \beta_{0,t} \leq \frac{i_t}{n^{\nu}} \leq \beta_{1,t}, 0 \leq t \leq d \right\}.$$

The sieve maximum likelihood estimator for $(\phi, \theta_1, \cdots, \theta_d)$ is that element $(\hat{\phi}, \hat{\theta}_1, \cdots, \hat{\theta}_d) \in \Omega_n$ such that

$$L_n(\hat{\phi}, \hat{\theta}_1, \cdots, \hat{\theta}_d) = \sup \{ L_n(\tilde{\phi}, \tilde{\theta}_1, \cdots, \tilde{\theta}_d) : (\tilde{\phi}, \tilde{\theta}_1, \cdots, \tilde{\theta}_d) \in \Omega_n \}.$$

For $\varepsilon > 0$ sufficiently small, it is convenient to define

$$\Omega_{n, \varepsilon} = \Omega_n \setminus (\phi - \varepsilon, \phi + \varepsilon) \times (\theta_1 - \varepsilon, \theta_1 + \varepsilon) \times \cdots \times (\theta_d - \varepsilon, \theta_d + \varepsilon).$$
We have the following result that shows that the sieve MLE defined above is weakly consistent.

**Theorem 24** Let $d \geq 3$ and let $(\tilde{\phi}, \tilde{\theta}_1, \cdots, \tilde{\theta}_d) \in \Omega_n$ be a constant vector (depending only on $n$) such that $|\phi - \tilde{\phi}| < n^{-\nu}$ and $|\theta_t - \tilde{\theta}_t| < n^{-\nu}$, $t = 1, \cdots, d$. Then for any $\varepsilon > 0,$

$$P \left\{ \frac{\sup_{(\tilde{\phi}, \tilde{\theta}_1, \cdots, \tilde{\theta}_d) \in \Omega_{n, \varepsilon}} \left\{ L_n(\tilde{\phi}, \tilde{\theta}_1, \cdots, \tilde{\theta}_d) \right\}}{L_n(\tilde{\phi}, \tilde{\theta}_1, \cdots, \tilde{\theta}_d)} \leq \varepsilon \right\} \to 1,$$

as $n \to \infty.$ Consequently the sieve maximum likelihood estimator $(\hat{\phi}, \hat{\theta}_1, \cdots, \hat{\theta}_d)$ tends to $(\phi, \theta_1, \cdots, \theta_d)$ in probability as $n \to \infty.$

**Related open problem**

The results in this section pertain to Matérn-type Gaussian random fields that are exactly once mean square differentiable. We conjecture that these results generalize to Matérn-type Gaussian random fields that are exactly $m$-times, $m \geq 2$, mean square differentiable.
3.9.5 Minimum norm quadratic estimation

The idea of minimum norm quadratic estimation can be traced back to Rao (1971). In the context of kriging, Stein (1989b) considered the following model.

\[ f(x) = \beta' h(x) + Z(x), \quad x \in [0, 1]^d, \]

where \( \beta = (\beta_1, \cdots, \beta_q)' \) is a vector of unknown constants, \( h(.) = (h_1(.), \cdots, h_q(.))' \) is a vector of known functions and \( Z(.) \) is a stationary Gaussian random field with mean 0 and covariance function \( V(., .) \). It is postulated that

\[
\mathrm{Cov}(Z(x), Z(y)) = \sum_{i=1}^{k} \theta_i K_i(x, y)
\]

(11)

\[ = K_\theta(x, y), \quad x, y \in [0, 1]^d. \]

Here \( \theta = (\theta_1, \cdots, \theta_k)' \) is a vector of parameters to be estimated and \( K_i(., .), i = 1, \cdots, k \) are known covariance functions. We observe that (11) defines a parametric class of covariance functions. The true covariance function \( V \) need not be a member of this parametric class.
**Definition.** Recall that corresponding to every mean function \( m(.) \) and covariance function \( K(., .) \), there is a unique probability measure [denoted by \((m, K)\)] of a Gaussian random field with those mean and covariance functions. We say that the covariance functions \( K_1(., .) \) and \( K_2(., .) \) are *compatible* if \((0, K_1) \) and \((0, K_2) \) are equivalent (or mutually absolutely continuous). Further recall that continuous covariance functions are either compatible or orthogonal.

**Remark.** Note that the definition of compatibility of covariance functions does not assume that the random field is Gaussian (even though it is the case in this section).

Let \( e \) denote the \( k \times 1 \) vector of 1’s. Stein (1987) observed that if \( e' \theta \neq e' \theta^* \), then \((0, K_\theta) \) and \((0, K_{\theta^*}) \) are orthogonal. However if \( p \) is not proportional to \( e \), then there exist \( \theta, \theta^* \) such that \( p' \theta \neq p' \theta^* \) and \((0, K_\theta) \) and \((0, K_{\theta^*}) \) are equivalent.

This implies that \( e' \theta \) is the only linear combination of \( \theta \) that we can hope to estimate consistently via fixed-domain asymptotics.
We shall now construct the minimum norm quadratic estimator MINQE\((U, I)\) for \(e'\theta\). Here \(U\) stands for\textit{ unbiasedness} and \(I\) stands for\textit{ invariance} with respect to changes in \(\beta\).

We write

\[
 f_D = (f(x_1), \cdots, f(x_n))', \\
 h_D = (h(x_1), \cdots, h(x_n)).
\]

If \(\text{rank}(h_D) = r\), choose a \((n - r) \times n\) matrix \(F\) of full rank whose row space is the orthogonal complement of the row space of \(h_D\). Define \(Y = F f_D\) so that \(Y\) is a set of contrasts, i.e., \(E(Y) = 0\) for all \(\beta\).

Next let \([K_l(x_i, x_j)]\) denote the \(n \times n\) matrix whose \((i, j)\)th element is \(K_l(x_i, x_j)\), and let

\[
 \Psi_l = F[K_l(x_i, x_j)]F', \\
 \Psi(\theta) = \sum_{l=1}^{k} \theta_l \Psi_l.
\]
For a given vector \( p = (p_1, \cdots, p_k)' \), a quadratic unbiased invariant estimator of \( p'\theta \) will be of the form \( Y'HY \), where
\[
\text{tr}(H\Psi_l) = p_l, \quad l = 1, \cdots, k,
\]
and \( H \) is a \( (n - r) \times (n - r) \) matrix.

If there exists such a matrix \( H \), \( p'\theta \) is said to be estimable. Choose \( \alpha = (\alpha_1, \cdots, \alpha_k) \) such that
\[
[K_\alpha(x_i, x_j)] = \sum_{l=1}^{k} \alpha_l[K_l(x_i, x_j)]
\]
is a \( n \times n \) strictly positive definite matrix. Since
\[
\text{Var}(Y'HY) = 2\text{tr}(H\Psi(\theta))^2,
\]
the MINQE(\( U, I \)) chooses \( H \) to minimize the norm \( \text{tr}(H\Psi(\alpha))^2 \). For a given \( \alpha \), we note that \( F \) can be chosen such that \( \Psi(\alpha) = I \). Hence if \( p'\theta \) is estimable, the MINQE(\( U, I \)) for \( p'\theta \) is given by
\[
p'[\text{tr}\Psi_i\Psi_j]^{-}[Y'\Psi_iY],
\]
where \( [\text{tr}\Psi_i\Psi_j]^{-} \) is any generalized inverse of \( [\text{tr}\Psi_i\Psi_j] \) [see Rao (1973), page 305]. Here \( [Y'\Psi_iY] \) denotes the \( k \times 1 \) vector whose \( i \)th component is \( y'\Psi_iY \).
Rao (1979) showed that $p'\theta$ is estimable for all $p$ if and only if $[\text{tr}\Psi_i\Psi_j]$ is invertible, in which case it is said that \( \text{MINQE}(U, I) \) of $\theta$ exists and is given by

$$\hat{\theta}_n(\alpha) = [\text{tr}\Psi_i\Psi_j]^{-1}[Y'\Psi_iY].$$

Finally we write

$$G_n = [\text{tr}\{(\Psi_i(n) - I)(\Psi_j(n) - I)\}].$$

The following theorem is adapted from Stein (1989b), page 992.

**Theorem 25** With the above notation and assumptions, suppose that $K_0, K_1, \ldots, K_k$ are compatible, $e'\alpha = 1$, and there exists $c > 0$ such that $V$ and $cK_0$ are also compatible. Then there exists a finite-valued $G$ such that $\lim_{n \to \infty} G_n = G$. Under the additional assumption that

$$\begin{pmatrix} G & e' \\ e & 0 \end{pmatrix}$$

is invertible, we also have:

(i) $\hat{\theta}_n(\alpha)$, the \( \text{MINQE}(U, I) \) for $\theta$ exists for sufficiently large $n$.

(ii) $(n/2)^{1/2}(e'\hat{\theta}_n(\alpha) - c)/c$ tends in distribution to $N(0, 1)$. 

190
(iii) $E[\hat{\theta}_n(\alpha) - \hat{\theta}(\alpha)]^2 \rightarrow 0$ as $n \rightarrow \infty$, where $\hat{\theta}(\alpha)$ is a finite-valued well-defined random vector.

As a final remark, it is noted that MINQE’s are only defined when the covariance structure is linear in the unknown parameters. Hence this technique may not be as widely applicable as maximum likelihood estimation. An advantage of MINQE($U, I$) is that it appears to be more analytically tractable under fixed-domain asymptotics.
3.10 Prediction

Sections 3.3 and 3.4 describe two approaches to prediction using the kriging model and fully Bayesian model respectively. One important underlying assumption there is that the covariance function is completely known.

On the other hand, Section 3.9 reports that in most applications, the covariance function is specified up to a finite number of parameters. To compound the difficulty, it is known that it is possible that some of these parameters cannot be estimated consistently with respect to fixed-domain asymptotics.

This section addresses the problem of prediction with a misspecified covariance function. Again, we shall use the kriging model, name,

\[ f(x) = \beta' h(x) + Z(x), \quad x \in [0, 1)^d, \]

where \( \beta \) is a \( k \times 1 \) vector of unknown coefficients, \( h(.) \) is a known vector-valued function and \( Z(.) \) is a stationary random field with mean 0 and covariance function

\[ \text{Cov}(Z(x), Z(y)) = K(x, y). \]
We observe \( f_D = (f(x_1), \cdots, f(x_n))' \) and wish to predict the value of \( f(x_0) \) for some \( x_0 \in [0, 1]^d \). If \( K(.) \) is known, the BLUP of \( f(x_0) \) is known to be

\[
\hat{f}_n(x_0) = c' C^{-1} f'_D + (h(x_0) - h_D C^{-1} c)' (h_D C^{-1} h'_D)^{-1} h_D C^{-1} f'_D,
\]

where

\[
h_D = (h(x_1), \cdots, h(x_n)),
\]

\[
c = (K(x_0, x_1), \cdots, K(x_0, x_n))'
\]

and \( C \) is the \( n \times n \) matrix with the \((i, j)\)th element \( K(x_i, x_j) \).

Yakowitz and Szidarovszky (1985) showed that if the design sites \( x_1, x_2, \cdots \) have a limit point \( x_0 \), \( h(x) = 1 \) for all \( x \in [0, 1]^d \) and \( k = 1 \), then \( \hat{f}_n(x_0) \) is a consistent predictor of \( f(x_0) \) even if the covariance function \( K(.) \) is incorrect, as long as \( f(.) \) is continuous almost surely, and there exist positive numbers \( q \) and \( M \) such that

\[
\lim \inf |w|^q S(w) \geq M, \quad \text{as } |w| \to \infty,
\]

where \( S(.) \) is the spectral density for \( K(.) \).
Stein (1988) observed that for isotropic covariance functions \( K_0(.) \) and \( K_1(.) \) on \([-1, 1]\), if they behave similarly at the origin, they will commonly be compatible, although exceptions exist. For example, if \( K_1(.) \) has \( 2m \) derivatives on some interval \((-T, T)\) and \( K_1^{(2m+1)}(0^+) \neq 0 \) [so that the Gaussian process with covariance function \( K_1(.) \) has \( m \) derivatives, but not \( m + 1 \) derivatives almost surely], then for \( K_0(.) \) and \( K_1(.) \) to be compatible on \([0, T]\), \( K_0(.) \) must also have \( 2m \) derivatives on \((-T, T)\) and \( K_0^{(2m+1)}(0^+) = K_1^{(2m+1)}(0^+) \).

There are exceptions to the converse though. Let

\[
K_0(x) = e^{-|x|},
\]

\[
K_1(x) = \max(0, 1 - |x|).
\]

Then it follows from Ibragimov and Rozanov (1978), page 100, that \( K_0(.) \) and \( K_1(.) \) are compatible on \([0, T]\) if and only if \( T \leq 1 \), despite the fact that \( K_0^{(1)}(0^+) = K_1^{(1)}(0^+) \).
3.10.1 Asymptotic efficiency of BLUP

For simplicity, let \( \hat{f}_{i,n}(x) \) denote the best linear unbiased predictor (BLUP) of \( f(x) \) based on the observations \( f(x_1), \ldots, f(x_n) \), assuming that \( K_i(.) \) is the actual covariance function of the process. Also define

\[
e_{i,n}(x) = \hat{f}_{i,n}(x) - f(x),
\]

the error of this predictor at \( x \). Denote \( E_i \), \( \text{Var}_i \) and \( \text{Cov}_i \) to be the expectation, variance and covariance under \( P_i \), the probability measure of a Gaussian random field with mean \( \beta' h(.) \) and covariance function \( K_i(.,.) \).

The following theorem is taken from Stein (1988), page 57, which gives the performance of BLUP when the covariance function is misspecified.
Theorem 26 Suppose $K_0(\cdot, \cdot)$ and $K_1(\cdot, \cdot)$ are two compatible covariance functions on $[0,1)^d$. Let $x \in [0,1)^d$ and $x_1, x_2, \cdots$, be a sequence of points in $[0,1)^d$ not containing $x$ but having $x$ as a limit point. Assume

$$
\lim_{n \to \infty} \text{Var}_0(e_{0,n}(x)) = 0.
$$

Then

$$
\lim_{n \to \infty} \frac{\text{Var}_0(e_{0,n}(x))}{\text{Var}_0(e_{1,n}(x))} = 1.
$$

Remark. $\hat{f}_{i,n}(\cdot)$ is said to be asymptotically efficient if (12) holds [see Stein (1988)]. However it should be noted that the class of predictors is only confined to those predictors that are linear in the observations.
Proof of Theorem 26

We observe that Theorem 26 involves only the first two moments of the random field. Thus without lost of generality, assume that the random field \( f(.) \) is Gaussian with mean \( \beta' h(.) \) and covariance function \( K_i(.,.) \). By the definition of BLUP, we clearly have

\[
\frac{\text{Var}_0[e_{0,n}(x)]}{\text{Var}_0[e_{1,n}(x)]} \leq 1.
\]

Also,

\[
\frac{\text{Var}_0[e_{0,n}(x)]}{\text{Var}_0[e_{1,n}(x)]} = \frac{\text{Var}_0[e_{0,n}(x)] \text{Var}_1[e_{0,n}(x)] \text{Var}_1[e_{1,n}(x)]}{\text{Var}_1[e_{0,n}(x)] \text{Var}_1[e_{1,n}(x)] \text{Var}_0[e_{1,n}(x)]}.
\]

Now,

\[
\frac{\text{Var}_1[e_{0,n}(x)]}{\text{Var}_1[e_{1,n}(x)]} \geq 1.
\]
Hence to prove the theorem it suffices only to show

\[
\liminf_{n \to \infty} \frac{\text{Var}_0[e_{0,n}(x)]}{\text{Var}_1[e_{0,n}(x)]} \geq 1,
\]

since this also implies (via symmetry) that

\[
\liminf_{n \to \infty} \frac{\text{Var}_1[e_{1,n}(x)]}{\text{Var}_0[e_{1,n}(x)]} \geq 1.
\]

Define

\[
Y_n = \frac{e_{0,n}(x)}{\{\text{Var}_0[e_{0,n}(x)]\}^{1/2}},
\]

so that \(E_0(Y_n^2) = 1\).

Now suppose (13) is false. Then there exists a subsequence \(n_1, n_2, \ldots\) satisfying

\[
\lim_{l \to \infty} E_1(Y_{n_l}^2) = \lim_{l \to \infty} \frac{\text{Var}_1[e_{0,n_l}(x)]}{\text{Var}_0[e_{0,n_l}(x)]} = c^{-1} > 1.
\]

(14)

Note that \(c > 0\) since \(K_0(., .)\) and \(K_1(., .)\) are compatible.
Since $E_0(Y_{n_l}^2) = 1$, there exists a further subsequence such that

$$\frac{1}{M} \sum_{l=1}^{M} Y_{n_{kl}}^2 \to 1, \quad \text{as } M \to \infty$$

in probability under $P_0$.

As the probability measures $P_0$ and $P_1$ are mutually absolutely continuous, we must also have

$$\frac{1}{M} \sum_{l=1}^{M} Y_{n_{kl}}^2 \to 1, \quad \text{as } M \to \infty$$

in probability under $P_1$.

Using the assumption that the process is Gaussian, we conclude that

$$\lim_{M \to \infty} E_1\{\frac{1}{M} \sum_{l=1}^{M} Y_{n_{kl}}^2 \} = 1.$$  

This contradicts (14). \hfill \Box

**Remark.** A detailed proof of Theorem 26 can be found in Stein (1988), page 58.
3.10.2 Mean and covariance misspecified

Stein (1990a) investigates the effect of misspecifying the mean and covariance function of a random field on optimal (minimum mean-squared error) linear predictions of the random field.

Let the statistical model be

\[ f(x) = m(x) + Z(x), \quad x \in [0, 1)^d, \]

where \( m(\cdot) \) is a deterministic function and \( Z(\cdot) \) is a mean 0 random field with covariance function

\[ \text{Cov}(Z(x), Z(y)) = K(x, y). \]

The pair \((m, K)\) defines the second-order structure of \( f(\cdot) \).

Recall that if \((m, K)\) is known and \( f \) is observed at sites \( x_1, \cdots, x_n \in [0, 1)^d \), then the optimal linear predictor of \( f(x_0) \) is

\[
\hat{f}_n(x_0) = Ef(x_0) + \text{Cov}(f(x_0), f_D')[\text{Cov}(f_D, f_D')]^{-1}(f_D - Ef_D),
\]

where \( f_D = (f(x_1), \cdots, f(x_n))' \).
Now let \( x_1, x_2, \ldots \), be a dense sequence of points in \([0, 1]^d\). This section investigates the large sample behavior of \( \hat{f}_n(x_0) \) when \((m, K)\) is misspecified.

**Definition.** Corresponding to every second-order structure \((m_i, K_i)\) on \([0, 1]^d\), there is a unique Gaussian measure \(P_i\) with mean function \(m_i\) and covariance function \(K_i\). We say that \((m_0, K_0)\) and \((m_1, K_1)\) are *compatible* if \(P_0\) and \(P_1\) are mutually absolutely continuous. This definition is an extension of compatibility of covariance functions in Section 3.9.5.

Next, if \(f(x), x \in [0, 1]^d\), is a random field with finite second moments and second-order structure \((m, K)\) with \(m\) and \(K\) continuous, then \(H(m, K)\) denotes the closed linear hull of \(f(.)\) with respect to the inner product

\[
\langle f(x), f(y) \rangle_{m, K} = K(x, y) + m(x)m(y).
\]

\(H(m, K)\) is a separable Hilbert space of real random variables. For simplicity, the subscript \(i\) will be used to denote expectations under \((m_i, K_i)\). For example, \(E_i h_1 = m_i(h_1)\) and \(K_i(h_1, h_2) = \text{Cov}_i(h_1, h_2)\) for \(h_1, h_2 \in H(m_i, K_i)\).
Suppose that $(0, K_0)$, $(m_0, K_0)$ and $(m_1, K_1)$ are all compatible on $[0,1]^d$ with continuous mean and covariance functions. Then $H(0, K_0) = H(m_0, K_0) = H(m_1, K_1)$ [see Ibragimov and Rozanov (1978), page 71]. Let $\eta_1, \eta_2, \cdots$, be a linearly independent basis for $H(0, K_0)$ and $\psi_1, \psi_2, \cdots$, be its Gram-Schmidt orthogonalization. For $\psi \in H(0, K_0)$, it follows that the best linear predictor for $\psi$ using the size $n$ “sample”

$$\Psi_n = (\psi_1, \cdots, \psi_n)'$$

under $(m_0, K_0)$ is

$$\hat{\psi}_n = E_0(\psi) + k'_n(\Psi_n - E_0\Psi_n),$$

where $k_n = (K_0(\psi, \psi_1), \cdots, K_0(\psi, \psi_n))'$. Finally, let

$$e_{0,n}(\psi) = \hat{\psi}_n - \psi$$

be the error of this prediction. Analogously, $e_{1,n}(\psi)$ is the prediction error of the best linear predictor for $\psi$ under $(m_1, K_1)$.
Then

**Theorem 27** With the above notation and assumptions,

\[
\sup_{\psi \in H(0, K_0)} \frac{E_0\{[e_{1,n}(\psi) - e_{0,n}(\psi)]^2\}}{E_0[e_{0,n}(\psi)^2]} \to 0,
\]

as \( n \to \infty \).

The proof of the above theorem (plus a lot more related details) is given in Stein (1990a), page 856.
3.10.3 Back to BLUP

Recall the kriging model:

\[ f(x) = \beta' h(x) + Z(x), \quad x \in [0, 1]^d. \]

where \( \beta \) is a \( p \times 1 \) vector of unknown parameters, \( h(.) = (h_1(.), \cdots, h_p(.))' \) is a known \( p \times 1 \) vector function that is continuous at \( x_0 \) and \( Z(.) \) is a mean 0 random field with a covariance function \( K(., .) \) that is continuous at \( (x_0, x_0) \).

Suppose \( f(.) \) is observed at points \( x_1, x_2, \cdots \), where \( \lim_{n \to \infty} x_n = x_0 \). Then Theorem 5.1 of Stein (1990a), page 869, proves that the BLUP \( \hat{f}_n(x_0) \) for \( f(x_0) \) is mean-square consistent. Next let

\[ e_n(f(x_0)) = \hat{f}_n(x_0) - f(x_0), \]

the prediction error of BLUP and \( e_{\beta,n}(f(x_0)) \) denote the analogous prediction error if the optimal linear predictor for \( f(x_0) \) is used (with \( \beta \) known).
Then Stein (1990a), page 870, showed that

**Theorem 28** Suppose that \((h_1, K), \cdots, (h_p, K)\) and \((0, K)\) are all compatible. Then with the above notation and assumptions,

\[
\frac{E[e_{\beta,n}(f(x_0))^2] - E[e_n(f(x_0))^2]}{E[e_{\beta,n}(f(x_0))^2]} = o(1),
\]

as \(n \to \infty\) uniformly over \(x_0 \in [0, 1]^d\).

**Remark.** Theorem 28 gives conditions where the BLUP is asymptotically optimal relative to the best linear predictor with \(\beta\) known.

The proof of Theorem 28 uses techniques from Hilbert space theory and is quite involved.

For the sake of completeness, we note that Stein (1990b) gives explicit bounds on the convergence rate of Theorem 28 and further applications. Stein (1999b) investigates the errors of optimal linear predictors using spectral techniques.
3.10.4 Estimated covariance function

Sections 3.10.1 to 3.10.3 consider the effect of mis-specifying the mean and covariance functions on the performance of the best linear predictor. In practice of course, the covariance function is estimated from the data. Thus given a large number of observations in \([0, 1)^d\), it is important to obtain an estimated covariance function that is “nearly” compatible to the true covariance function. Also it is unimportant to distinguish between covariance functions that are compatible.

The above paragraph complements the fact that under fixed-domain asymptotics, some of the parameters in the statistical model cannot be estimated consistently. However this should not matter (at least heuristically) since they should lead essentially to compatible covariance functions. It would be useful to clarify or rigorize this statement though.
3.10.5 *Open related problem*

The prediction problem treated in Section 3.10 deals with interpolation and it is obvious that consistent prediction for random fields (with continuous sample paths almost surely) is feasible here under fixed-domain asymptotics. However much less (if anything) is known regarding extrapolation. In fact, I have not come across any published work on extrapolation with respect to fixed-domain asymptotics.
4.1 Sequential Design

This section describes some of the sequential strategies used in computer experiments that are found in the literature. In particular, we shall discuss the works of

- Bernardo, Buck, Liu, Nazaret, Sacks and Welch (1992) on integrated circuit design optimization,
- Jones, Schonlau and Welch (1998) on efficient global optimization,

In contrast to the previous two sections, the theory is not as developed in the setting of sequential design (possibly because this is a much more difficult problem). The sequential solutions and approaches proposed in the literature are more ad hoc but nonetheless appear to work well in many situations.
4.1.1 IC design optimization

A major problem for integrated circuit designers is how to design circuits so that the performances, as predicted by a circuit simulator, are insensitive to uncontrollable variations in the manufacturing processes and in the operating conditions. Many of these variations can be modeled statistically and hence statistical modeling of circuit and process simulators to achieve consistently good performance has become increasingly prominent.

Bernardo, et. al. (1992) propose a sequential strategy for designing IC circuits using CAD methods. They noted that their approach allows designers to obtain performances insensitive to unwanted variations in device and process model parameters, environmental fluctuations, material impurities, and other errors which are avoidable in design and manufacturing processes.
In particular, they suggest the use of multistage experimentation and, during each stage, a statistical predictor of the simulator is constructed. This predictor in turn reveals subranges of the parameters in which the optimal solution lies. More observations are taken in this smaller region, a more accurate predictor is obtained and so on.

Let $f(.)$ denote the output of the circuit simulator and $x = (x_1, \cdots, x_d)'$ denote the $d$-dimensional vector of varying input parameters to $f(.)$ (all other inputs are assumed to remain fixed). Furthermore, we assume that $x = c + u$ where $c$ and $u$ denote the controllable and uncontrollable components of $x$ respectively. The detailed approach consists of 6 steps (given below).
Step 1: Postulate a tentative approximating model for the performance.

If \( x \) denotes the vector of input parameters, let

\[
    f(x) = \beta + Z(x).
\]

Here \( Z(.) \) is a Gaussian random field having a correlation function

\[
    \text{Corr}(Z(x), Z(y)) = R(x, y),
\]

and having a constant variance

\[
    \text{Var} Z(x) = \sigma^2.
\]

Bernardo, et. al. (1992) observed that \( f(.) \) is often continuous and thus \( Z(.) \) should be continuous too (almost surely). The specific \( R(., .) \) that they use is of the form

\[
    R(x, y) = \prod_{j=1}^{d} \exp(-\theta_j |x_j - y_j|^{p_j}).
\]

The model parameters \( \beta, \sigma^2, \theta_1, \ldots, \theta_d, p_1, \ldots, p_d \) are unknown and need to be estimated from the data.
Step 2: Plan an experiment and run the simulator to collect the data.

Use a Latin hypercube sampling scheme to select the inputs. This scheme is easy to generate and they cover the experimental region fairly uniformly.

The rule of thumb for initial sample size $n$ is that at least 3 observations are needed per parameter. On top of that, chances are some outlying data will be discarded if the goal is optimization. As an example if there are 17 parameters to be estimated, they suggest using a sample size of $n = 75$ or so.

Step 3: Use the data to fit the model.

Estimate the correlation parameters in $R(.,.)$ via maximum likelihood and obtain the BLUP $\hat{f}_{BLUP}(.)$ for $f(.)$. 

212
Step 4: Check the accuracy of prediction and plot the parameter effects.

Compute the root mean square error of $\hat{f}_{BLUP}(.)$ at a number of randomly selected sites, say 20 sites. The range of these errors is a good indicator of the prediction accuracy of $\hat{f}_{BLUP}(.)$. Note we can do this *without* explicitly running the simulator at these sites.

To get a good understanding of the behavior of $\hat{f}_{BLUP}(.)$, decompose $\hat{f}_{BLUP}(.)$ into a mean value, main effects arising from individual parameters, and the residual (or second-order interactions) interactions among the parameters. This is the usual ANOVA decomposition. The main effects can be plotted and these plots can provide an excellent, yet simple, indication of how a parameter affects the performance. Contour plots of second-order effects may also be useful.

Here we reach a fork in the procedure. If $\hat{f}_{BLUP}(.)$ is sufficiently accurate, go to Step 6; otherwise proceed to Step 5.
Step 5: Choose a subregion for the next experiment.

An optimization routine (as in Step 6) can be used to find the center of the updated subregion. The main effect plots are useful in choosing updated limits for the parameters. It should be noted that the uncontrollable variations cannot be restricted. Then repeat Steps 1 to 4, using data drawn from the updated subregion.
Step 6: Optimize the loss function.

The loss depends of course on the performance function. The yield of a process is a natural and common choice for the performance function. However one concern with yield is the fact that the performance constraints defining yield are often arbitrary and create impractical differences between circuits that just pass and those that just fail.

On the other hand, Taguchi (1986) proposed the objective of minimizing a “loss” as a measure of variability around a target performance. The motivation is that the resulting choice of designable parameters is intended to be robust, i.e., to make the circuit insensitive to manufacturing and other uncontrollable variabilities.

An example of a Taguchi loss function is

\[ L_{MAX}(c) = \max_u |f(c + u) - 1|, \]

where the target value of \( f(.) \) is 1 and the input vector is \( x = c + u \). Here recall that \( c \) denote the controllable component and \( u \) the uncontrollable component of \( x \).
In Step 6, replace \( f(.) \) by \( \hat{f}_{BLUP}(.) \) in the loss function. This gives a "predicted" loss function. Using the example in the previous paragraph, we have

\[
\hat{L}_{MAX}(c) = \max_u |\hat{f}_{BLUP}(c + u) - 1|.
\]

A (random) search algorithm can be used for the optimization (i.e., minimization over \( c \)). Inspection of the main effect plots may be helpful in choosing a starting point.

After the estimate of the optimum is obtained, a confirmatory run on the circuit simulator at that point is done. If the confirmatory run is unsatisfactory, steps should be taken to improve the statistical model or a further stage with more data may be needed.

**Remark.** Bernardo, et. al. (1992) illustrated the above ideas and approach on sequential design with two concrete examples on integrated circuits. See the original paper for more details regarding these examples.
4.1.2 Efficient global optimization

Jones, Schonlau and Welch (1998) proposed a response surface methodology that is especially good at modeling nonlinear, multimodal functions that often occur in engineering. The paper emphasized the need to first validate the model and the selection of additional sites (beyond the initial sample) is rigorously based on a statistical model.

The authors claimed that their method has the following three advantages.

• First they observed that their technique often requires the fewest function evaluations of all competing methods. This is probably due to the flexibility of the model and that typical functions in engineering are multimodal and nonlinear.

• Second, the response surface approach provides a credible stopping rule based on expected improvement from further searching.

• Third, the response surface approach provides a fast approximation to the computer model that can be used to identify important variables, visualize the nature of input-output relationships, and quantify tradeoffs among multiple objectives.
DACE stochastic process model

Let \( f : [0, 1)^d \rightarrow \mathbb{R} \) be the deterministic function of interest. In the stochastic processes approach of Jones, et. al. (1998), it is assumed that

\[
f(x) = \mu + Z(x),
\]

where \( \mu \) is a real-valued parameter and \( Z(.) \) is a stationary mean 0 Gaussian random field with

\[
\text{Var}[Z(x)] = \sigma^2
\]

and correlation function

\[
\text{Corr}[Z(x), Z(y)] = \exp[- \sum_{j=1}^{d} \theta_j |x_j - y_j|^{p_j}],
\]

where \( \theta_j > 0, p_j \in [1, 2], j = 1, \cdots, d. \)

The above model is sometimes called the DACE stochastic process model (for Design and Analysis of Computer Experiments). The DACE model has \( 2d + 2 \) parameters \( \mu, \sigma^2, \theta_1, \cdots, \theta_d, p_1, \cdots, p_d. \) These parameters are estimated by maximizing the likelihood of the sample.
In particular, let the observed sample be
\[ f_D = (f(x_1), \ldots, f(x_n))' \]
and \( R \) denotes the \( n \times n \) matrix whose \((i, j)\)th element is \( \text{Corr}(Z(x_i), Z(x_j)) \). Letting \( e \) be the \( n \times 1 \) vector of 1’s, the likelihood function is given by
\[
\frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}|R|^{1/2}} \exp\left[ -\frac{(f_D - e\mu)'R^{-1}(f_D - e\mu)}{2\sigma^2} \right].
\]

Given the correlation parameters \( \theta_1, \ldots, \theta_d, p_1, \ldots, p_d \), the values of \( \mu \) and \( \sigma^2 \) that maximize the likelihood function are given in closed-form as
\[
\hat{\mu} = \frac{e' R^{-1} f_D}{e' R^{-1} e},
\]
\[
\hat{\sigma}^2 = \frac{1}{n} (f_D - e\hat{\mu})' R^{-1} (f_D - e\hat{\mu}).
\]

The expressions for \( \hat{\mu} \) and \( \hat{\sigma}^2 \) are substituted back into the original likelihood function. The resulting “concentrated likelihood function” is then maximized with respect to \( \theta_j, p_j, j = 1, \ldots, d \). The associated \( \theta_j \)'s and \( p_j \)'s are the maximum likelihood estimates \( \hat{\theta}_j \)'s and \( \hat{p}_j \)'s respectively.
Now treat the MLE’s $\hat{\theta}_1, \cdots, \hat{\theta}_d, \hat{p}_1, \cdots, \hat{p}_d$ as if these are the true values of $\theta_1, \cdots, \theta_d, p_1, \cdots, p_d$. The best linear unbiased predictor (BLUP) for $f(x)$ is given by

$$\hat{f}(x) = \hat{\mu} + r' R^{-1} (f_D - e\hat{\mu}),$$

where $r$ is a $n \times 1$ vector whose $i$th element is given by $\text{Corr}(Z(x), Z(x_i))$. $\hat{f}(.)$ is our desired response surface.

Also it is useful to note that the mean square prediction error at $x \in [0,1]^d$ is

$$s^2(x) = \sigma^2 [1 - r' R^{-1} r + \frac{(1 - e' R^{-1} r)^2}{e' R^{-1} e}].$$

Note that $s^2(x_i) = 0$, $i = 1, \cdots, n$. That is the mean square prediction error of $f$ at each design site is 0.
**Remark.** One possible drawback of this approach is that in the computation of the mean square prediction error, it is assumed that $\hat{\theta}_j$’s and $\hat{p}_j$’s are the true parameter values. This probably leads to an underestimation of prediction errors. It is also interesting to note that for the DACE model, the consistency of the MLE’s (under fixed-domain asymptotics) is still an open problem.

**Model validation**

Jones, Schonlau and Welch (1998) suggested the use of leave-one-out cross-validation to assess the accuracy of the statistical model without sampling any more points beyond those used to fit the model.

More precisely, given the sample $f(x_1), \ldots, f(x_n)$, leave one point, say $f(x_1)$ out. Use the remaining samples points to estimate the model parameters and to obtain the response surface (i.e. BLUP) $\hat{f}_{-1}(.)$. The corresponding mean square prediction error $s^2_{-1}(.)$ is also calculated. Since $f(x_1)$ is known, the prediction error $f(x_1) - \hat{f}_{-1}(x_1)$ can then be computed at the site $x_1$. 

221
We standardize the prediction error by using the statistic
\[
\frac{f(x_1) - \hat{f}_{-1}(x_1)}{s_{-1}(x_1)}.
\]
If the model is valid, the value of the standardized statistic should be between $-3$ and $3$ (from the asymptotic normality of the statistic).

This is repeated by leaving $f(x_2)$ out and the above procedure repeated and so on until $f(x_n)$ is left out.

It is also advantageous to look at the usual residual plots as well as the quantile-quantile plot to try to better understand the behavior of $f$.

If these diagnostics fail to show a good fit, it is sometimes possible to improve the fit of the DACE model by transforming the function, e.g., the log transform $\log(f)$ or the inverse transform $-1/f$. 

222
**Global optimization**

One intuitively obvious way to optimize (say, minimize) a function $f$ is to fit an approximating surface to $f$ and then find the global minimum of the surface. However if the surface is nonlinear or multimodal, it is very likely that a local minimum is found and not the global minimum.

Jones, Schonlau and Welch (1998) noted that in the estimation of the global minimum of a function $f$, not only is the approximating response surface important but also the accuracy of the surface as a predictor of $f$. 
The following is an example that illuminates the ideas of Jones, et. al. (1998). Suppose the unknown function is shown as the solid line in Figure 8. The function is observed at 5 points (shown as dots). The dotted line is the DACE predictor. The DACE predictor is almost exactly minimized at the local minimum \( x = 2.8 \). If we were only to sample around the local minimum of DACE, it should be clear that we shall completely miss the global minimum.
To ensure that this does not happen, some emphasis must be put on the prediction uncertainty of the DACE predictor. Figure 9 shows the DACE predictor as well as its standard error. The standard error is maximized at around $x = 8.3$. So at least at this point, evaluating $f$ at $x = 8.3$ is a plausible alternative to evaluating $f$ at, say $x = 2.8$. 
To decide which is the more promising option to take, we need further notation. Let

\[ f_{MIN} = \min\{f(x_1), \ldots, f(x_n)\}, \]

the current minimum value of the function at the design sites. The improvement at \( x \) is defined to be

\[ I(x) = \max\{f_{MIN} - f(x), 0\}, \]

where \( f(x) \) is taken to be a normal random variable with mean and standard deviation given by the DACE predictor \( \hat{f}(x) \) and its standard error.

The expected improvement at \( x \) is then \( E[I(x)] \). Jones, et. al. (1998) suggest choosing the next site \( x \) (for the evaluation of \( f \)) to be the site that maximizes \( E[I(.)] \).
This of course would mean going from one optimization problem to another. The latter maximization problem has the objective function $E[I(.)]$ which has a relatively simple closed-form expression, i.e.,

\[
E[I(x)] = [f_{MIN} - \hat{f}(x)]\Phi\left(\frac{f_{MIN} - \hat{f}(x)}{s(x)}\right)
\]

\[
+s(x)\phi\left(\frac{f_{MIN} - \hat{f}(x)}{s(x)}\right),
\]

where $\Phi$, $\phi$ denote the c.d.f., p.d.f. of the standard normal distribution and $s(x)$ is the standard deviation of the DACE predictor $\hat{f}(x)$. On the other hand, $E[I(.)]$ has a lot of local maxima that increases with sample size.
In the context of the previous DACE example, Figure 11(a) plots the function $E[I(.)]$ when only 5 points have been sampled and Figure 11(b) gives the plot of $E[I(.)]$ after adding a point at $x = 2.8$.

We remark that Figures 8, 9 and 11 are taken from Jones, et. al. (1998).
Jones, et. al. (1998) has an algorithm, called EGO for Efficient Global Optimization, for finding the maximum of $E[I(.)]$. EGO makes use of the fact that

$$\frac{\partial E[I(x)]}{\partial \hat{f}(x)} = -\Phi \left( \frac{f_{MIN} - \hat{f}(x)}{s(x)} \right) > 0,$$

$$\frac{\partial E[I(x)]}{\partial s(x)} = \phi \left( \frac{f_{MIN} - \hat{f}(x)}{s(x)} \right) > 0.$$

Jones, et. al. (1998) applied their technique to a number of examples with objective functions with up to 6 arguments (i.e. $d = 6$). See their paper for details.
4.1.3 Integrated response function minimization

Williams, Santner and Notz (2000) considered the commonly occurring situation in computer experiments in which there are two types of input variables: suppose \( x = (x_c, x_e) \) where \( x_c \) is a set of “manufacturing” (control) variables and \( x_e \) is a set of “environmental” (noise) variables. Manufacturing variables can be controlled while environmental variables are not controllable but have values governed by some (known) distribution.

Williams, et. al. (2000) introduced a sequential experimental design for finding the optimum of

\[
L(x_c) = E[f(x_c, X_e)],
\]

where \( f \) is the objective function and the expectation is taken over the distribution of the environmental vector \( X_e \).
Their method can be briefly described as:

1. Calculate $f$ on an initial space filling design.

2. Use the sample information to select the next site according to a modified expected improvement criterion.

3. Continue selecting sites using all available sample information until a stopping criterion is met.

**Statistical model**

Williams, et. al. (2000) assume that the objective function can be represented as

$$f(x) = \beta + Z(x), \quad x \in [0, 1)^d,$$

where $\beta$ is an unknown constant and $Z(.)$ is a stationary mean 0 Gaussian random field with

$$\text{Var}[Z(x)] = \sigma^2,$$

and a multiplicative Matérn-type correlation function $R_\zeta(.)$. 
In particular,

\[ R_\zeta(x - y) = \prod_{j=1}^{d} \left\{ \frac{1}{\Gamma(\nu)2^{\nu-1}} \left( \frac{2\sqrt{\nu}|x_i - y_i|}{\theta_i} \right)^\nu \times K_\nu \left( \frac{2\sqrt{\nu}|x_i - y_i|}{\theta_i} \right) \right\}, \]

where \( x, y \in [0,1]^d \), \( \zeta = (\nu, \theta_1, \ldots, \theta_d) \) is an unknown vector parameter and \( K_\nu(.) \) is the modified Bessel function of order \( \nu \).

Next, Williams, et. al. (2000) placed the noninformative distribution

\[ (\beta, \sigma^2, \zeta)' \propto \sigma^{-2}, \]

on the parameter space of \( (\beta, \sigma^2, \zeta)' \).

Finally, let \( x_c \) and \( X_e \) represent the control and environmental variable vectors. It is assumed that \( X_e \) has a probability distribution with finite support \( \{x_{e,i}\}_{i=1}^{n_e} \) and associated probabilities \( \{w_i\}_{i=1}^{n_e} \).
The loss function $L(.)$ is given by

$$L(x_c) = \sum_{i=1}^{n_c} w_i f(x_c, x_{e,i}).$$

The goal is to identify the control variable setting $x^*_c$ that minimize $L(.)$.

**Remark.** Prior uncertainty in $L(.)$ is induced by the uncertainty in $f(.)$.

**Minimization algorithm**

The minimization algorithm of Williams, et. al. (2000) can be summarized in 5 steps:

**Step 0.** Choose the initial set of design sites

$$S_n = \{x_1, \cdots, x_n\}$$

according to a maximin distance design in the set of Latin hypercube sampling designs.
Denote the control variable portion of $S_n$ by

$$S_n^c = \{x_{c,1}, \cdots, x_{c,n}\}.$$

Let

$$f_{S_n} = (f(x_1), \cdots, f(x_n))',
L_{S_n^c} = (L(x_{c,1}), \cdots, L(x_{c,n}))',
L_{1:n} = \min\{L(x_{c,1}), \cdots, L(x_{c,n})\},$$

and the \textit{improvement} at the control variable site $x_c$
is defined to be

$$I_n(x_c) = \max\{0, L_{1:n} - L(x_c)\}.$$

Observe that apriori, $f(x)$ is a random variable and hence $L_{1:n}$ is also random.

\textbf{Step 1.} Estimate the correlation parameter vector $\zeta$ by the maximizer $\hat{\zeta}$ of the posterior density of $\zeta$ given $f_{S_n}$.

For Steps 2 to 4, assume that $\hat{\zeta}$ is the true correlation parameter vector.
**Step 2.** Choose the \((n+1)\)st control variable site \(x_{c,n+1}\) to maximize the posterior expected improvement given the current sample, i.e.,

\[
x_{c,n+1} = \arg\max_{t_c} E[I_n(t_c)|f_{S_n}, \zeta].
\]

Step 2 makes use of the fact that

\[
E[I_n(x_c)|f_{S_n}, \zeta] = E_{L_{S_n}|f_{S_n}, \zeta}\{E[I_n(x_c)|f_{S_n}, L_{S_{n_i}}, \zeta]\}.
\]

The inner expectation on the r.h.s. has a closed-form expression and this reduces the time needed for numerical integration.

**Step 3.** Choose the environmental variable site corresponding to the control site \(x_{c,n+1}\) to minimize the posterior mean square prediction error given the current sample, i.e.,

\[
x_{e,n+1} = \arg\min_{t_e} E\{[\bar{L}_{n+1}(x_{c,n+1}, t_e) - L(x_{c,n+1})]^2|f_{S_n}, \zeta\},
\]

where \(\bar{L}_{n+1}(x_{c,n+1}, t_e)\) is the posterior mean of \(L(x_{c,n+1})\) given \(f_{S_n}, f(x_{c,n+1}, t_e)\) and \(\zeta\).
**Step 4.** Determine if the algorithm should be terminated. If the stopping criterion is not met, set

\[ S_{n+1} = S_n \cup (x_{c,n+1}, x_{e,n+1}) \]

and compute \( f(x_{c,n+1}, x_{e,n+1}) \). Set \( n \) to \((n + 1)\) and go to Step 1.

If the stopping criterion is met, the global minimizer is set to the minimizer of BLUP based on the current sampling design. A typical stopping criterion would be that the expected improvements of the last 3 sites added are all small.

Williams, et. al. (2000) has two examples. In the first example, the objective function has 6 variables (4 of them are controllable variables and 2 environmental variables).

The other example is the Branin function with 4 variables (2 controllable and 2 environmental).