New Research Directions in Computer Experiments: ε-Clustered Designs

Selden B. Crary
NewallStreet, Inc., 535 Everett Ave., Apt. 410, Palo Alto, CA 94301
selden@newallstreet.com

Abstract

We explore the existence, properties, and applications of exact optimal designs for computer experiments, under Gaussian-process (GP), fixed-Gaussian-covariance-parameter, zero-nugget assumptions, that prescribe a cluster of two or more design points as closely spaced as practical, without being identically located. We define such designs as *ε-clustered* and define subcases, e.g., *twin-point*, *triplet-point*, etc. designs and review the history of these designs. We also define the *phase* of a design, based on its symmetry properties, and we introduce the concept of *phase transitions* between phases. We prove that the 0^{th}- and 1^{st}-degree terms in the expansion of the determinant of the covariance matrix, in powers of the separation distance from the center of a twin-point cluster to one of the twins, are zero. Using this fact, we outline a proof that, in two or more factors, the IMSE function is a truncated rational function, with leading powers of at least two in the series expansion in powers of the separation of the points, for numerator or denominator. We outline applications of the theory to extrapolation and to inversion of covariance matrices. We demonstrate the use of a *nonuplet-point* (9-point) design as the first stage of a sequential GP fit. We conjecture the form of the power-series expansion of the determinant of the covariance matrix for triplet-point, quadruplet-point, etc. designs. Finally, we conjecture that standard GP fitting does not support triplet-point designs, but a renormalization fixes this problem. We use the renormalization conjecture as a basis for performing GP fits to functions with regions of closely-spaced design points and show that the proposed method avoids the numerical errors observed via standard approaches.

Key Words: Twin-point design, twin points, clustered design, computer experiments, Gaussian processes, IMSE, design of computer experiments, Kriging, essential discontinuity, rational function, covariance function, covariance matrix, ill-conditioning

1. Introduction

We are interested in the existence, properties, and applications of exact optimal designs for computer experiments, under Gaussian-process, fixed-Gaussian-covariance-parameter, zero-nugget assumptions, that prescribe a cluster of two or more points as closely spaced as practical in the design space, without being identically located. We define such designs as ε-clustered. Such designs with exactly two proximal points were called *twin points* by Crary et al., and designs with exactly one pair of twin points were called *twin-point designs* [1]. By extension, such clusters with exactly three, four, etc. points may be named *triplet points*, *quadruplet points*, etc., respectively, and designs with exactly three triplet points, four quadruplet points, etc. may be named *triplet-point designs*, *quadruplet-point designs*, etc., respectively.

The clusters of the present study are distinguished from other types of clusters studied in statistics, e.g., those studied by Cochran [2] or Morris [3], in which design points, while separated by small distances in the design domain, were not specified as being as close as practical.

The clusters of our study are also distinguished from points in the design domain at which a function evaluation and/or one or more directional derivatives are specified to be taken because of the availability or computability of derivative information at those points, via adjoint [4,5] or other [6] methods.

Rather, we report optimal designs, in addition to those reported in [1,7,8], found by treating all design points on an equal footing and using extended-precision arithmetic and algorithms. The frequent proximity of the ε-clustered design points, as well as the complex phase diagram that can be constructed from plotting boundaries of contiguous regions of designs sharing symmetry properties, as a function of the covariance parameters, are emergent properties, made evident via computation.

The closest work to our study is the speculation in Stephenson's dissertation [6] that optimal designs for the construction of emulators may specify some regions (or points) where only function evaluations should be taken and other regions (or points) at which derivative information should be taken.

The foundational paper by Sacks et al. [9] provides the present paper's theoretical background and notation.

1.1 Known ε-clustered designs

d=1, N=2, MMSE-optimal, restricted, twin-point design: The first reference to such clustered designs seems to be Sacks, Welch, Mitchell, and Wynn (SWMW) [10], who reported a pair of twin points in the minimum-MMSE, *N=2* design on a closed interval of length *L*, with one point restricted to be held fixed at the interval's center. They provided the interpretation that the twin points provide both a function evaluation and a directional derivative at the twin-point location. Our close examination of this problem has shown the second design point is optimally located, as follows, depending upon the value of the dimensionless covariance parameter $\xi \equiv \theta L^2$, relative to a critical value ξ_c of approximately *2.78*:

(i) $\xi < \xi_c$: at the interval's center, as one of a pair of twin points
(ii) $\xi > \xi_c$: at either end of the design interval
(iii) $\xi = \xi_c$: indifferently at the interval's center, as one of a pair of twin points, or at either end of the design interval.

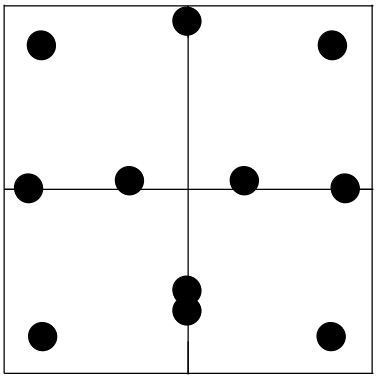
d=1, N=2, IMSE-optimal, restricted, twin-point design and twin-point transition: The day after the oral, SRC 2012 presentation of the present SRC 2012 paper, Hickernell [11] reported observing a continuous transition, upon decreasing ξ below a critical $\xi_c' \cong 3.01$, from a non-twin-point design to a twin-point design, for the problem of the minimum-IMSE, *N=2* design on a closed interval, with one point held fixed at the interval's center. This critical value has been confirmed by the present author.

Open research question: SWMW's twin-point design; our observation of a phase transition extending the result of SWMW; Hickernell's twin-point design; and

Hickernell's continuous, twin-point transition all require one of the design points to be fixed. An open research question is whether unrestricted, ε-clustered designs exist for minimum-MMSE, minimum-IMSE, or other useful objective functions.

d=2, N=11, IMSE-optimal, unrestricted, twin-point designs and twin-point transitions: Crary, Woodcock, and Hieke [1] reported the two-factor, $N=11$, $\xi_1=0.512$, $\xi_2=0.276$, IMSE-optimal, twin-point design, over the design domain $[-1,1]^2$, shown in Fig. 1, below. The vertical separation between the twin points was prescribed to be as small as practical, given the available computational resources.

d=4, N=8, IMSE-optimal, unrestricted, twin-point design: Crary [7,8] reported the four-factor, $N=8$, $\xi_1=2.024$, $\xi_2=0.552$, $\xi_3=1.260$, $\xi_4=0.908$, IMSE-optimal, twin-point design, over the design domain $[-1,1]^2$, shown in Table 1, below, where the value of δ was prescribed to be as small as practical.



x1	x2
-0.859071	0.001473
-0.792118	0.779630
-0.782622	-0.803880
-0.314466	0.039516
0.000009	0.906950
0.314470	0.039508
0.782616	-0.803880
0.792121	0.779620
0.859072	0.001461
-0.000017	-0.617140+δ
-0.000017	-0.617140- δ

Figure 1: The dot diagram (left) and listing (right) of the $N=11$, $\xi_1=0.512$, $\xi_2=0.276$, IMSE-optimal twin-point design of [1] are shown the this figure. In this and all subsequent dot diagrams of the present paper, the horizontal and vertical axes are x_1 and x_2, respectively, and the design domains are $[-1,1]^2$.

Table 1: The four-factor, minimum-IMSE design listing from [7,8]. The design domain is $[-1,1]^4$, and the covariance parameters are $\xi_1=2.024$, $\xi_2=0.552$, $\xi_3=1.260$, and $\xi_4=0.908$.

x_1	x_2	x_3	x_4
0.6038	0	-0.6570	0
-0.6038	0	0.6570	0
-0.5355	0	-0.4540	-0.6245
-0.5355	0	-0.4540	0.6245
0.5355	0	0.4540	-0.6245
0.5355	0	0.4540	0.6245
0.0005	0 + δ	-0.0012	0.0001
0.0005	0 – δ	-0.0012	0.0001

The author knows of no other ε-clustered designs in the literature.

1.2 Theory

Crary and Johnson [12], mostly using symbolic algebra, confirmed the earlier result [1], which mostly used high-precision numerical analysis, that the *d=1*, *N=2*, closed-interval, IMSE objective function was not singular, in the limit as the distance between design points approached zero, ($\varepsilon \to 0$) but that there was a jump discontinuity in the function, when the distance between points equaled zero ($\varepsilon=0$). They also reported that the *d=2*, *N=2*, closed-rectangular-design-domain, IMSE objective function was of the following, rational-function form:

$$\frac{IMSE}{\sigma_z^2} = \frac{a\delta_1^2 + b\delta_2^2 + \text{ higher-degree terms}}{c\left(\theta_1\delta_1^2 + \theta_2\delta_2^2\right) + \text{ higher-degree terms}}, \text{ with } c \neq 0,$$

where δ_1 and δ_2 are the Cartesian components of distance from the center of the points to the one of the points; and *a*, *b*, and *c* are real constants. Finally, they showed this function did not possess any singularity, except for an essential discontinuity at $\delta_1=\delta_2=0$ [12].

The following pair of 3D plots shows an example, from [12], of this essential discontinuity from two different viewing angles:

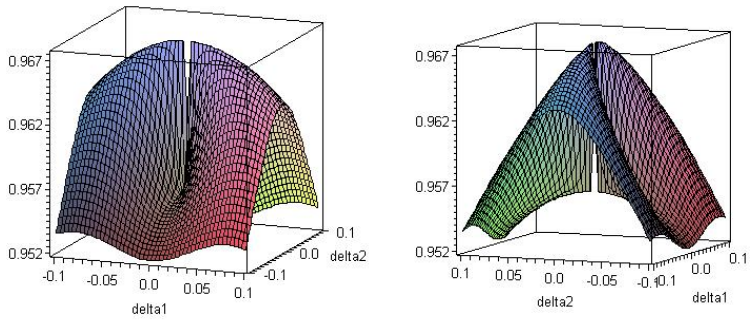


Figure 2: Two views of the essential discontinuity of an IMSE function, from [12].

The purpose of the present paper is to advance the understanding of ε-clustered designs and to encourage others to investigate these interesting designs.

2. Phases and an Example Phase Diagram

In addition to the *d=2*, *N=11*, $\xi_1=0.512$, $\xi_2=0.276$, IMSE-optimal, unrestricted, twin-point design shown in Fig. 1, above, we also identified similarly characterized designs for a variety of pairs of values of ξ_1 and ξ_2. A series of eight of these designs, all with $\xi_1=0.512$, but with ξ_2 ranging from *0.132* to *0.440*, is shown in Fig. 3, below. Two of these designs were twin-point designs. One of the pairs of twin points had a separation direction along the x_2 axis, while the other had a separation direction not along a cardinal direction. (See the figure caption for details.)

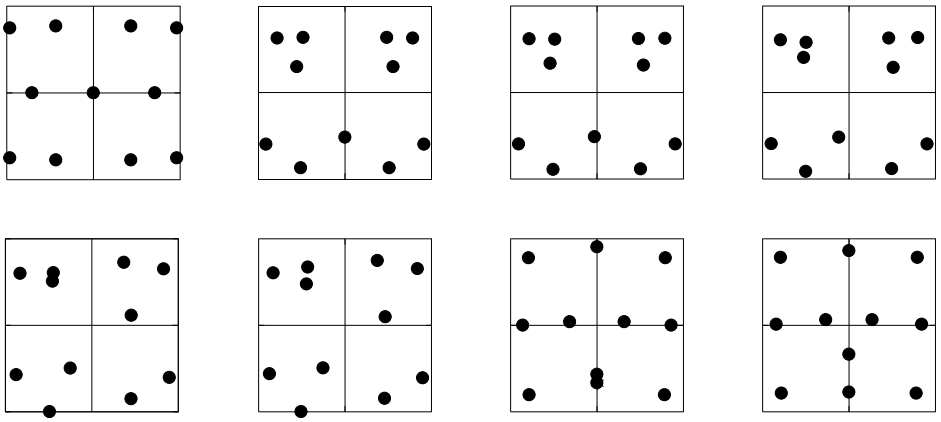


Figure 3: A series of IMSE-optimal designs for $\xi_1=0.512$ is shown here. From top left, reading horizontally, the values of ξ_2 are *0.132*, *0.140*, *0.148*, *0.152*, *0.200*, *0.224*, *0.276*, and *0.440*, respectively and the phases are 1, 2_1, 4, 4, 4T, 4, 2_1T, and 2_1, respectively.

Some of the designs of Fig. 3 had symmetries under mirror reflections about either the x_1 or x_2 axes, i.e., the operations of the Klein four group (also known as the dihedral group, D_2). For example, the design with $\xi_2=0.132$, transformed, under these reflections, only into itself and thus had group membership *1*. By contrast, the design with $\xi_2=0.140$ transformed, under these reflections, into either itself (via reflection about the x_2 axis) or into a different design (via reflection about the x_1 axis) with equal value of the objective function and thus had group membership *2*. In analogy to phases of matter, we identified the *phase* of a design as the proper symmetry group of the design.

For this two-factor case, we chose to label the phases first by their group membership, i.e., by *1*, *2*, or *4*; then to add a subscript *1* or *2* to indicate about which axis a membership *2* phase had its mirror symmetry (*1* for the axis with the smaller θ value); and finally to append the letter "T" for a twin-point design. Because a twin-point design came with a specific direction between the twin points, there could be no Phase 1 twin-point design. Further, we did not consider phases with more than one pair of twin points nor those with triplet points, etc., as such phases were not observed. Thus, the following were the seven possible names of phases, according to our naming convention: 1, 2_1, 2_1T, 2_2, 2_2T, 4, and 4T.

Extending our analogy with phases of matter, we generated the phase diagram in Fig. 4, which, for variable values of the covariance parameters, θ_1 and θ_2, showed a number of well-defined, contiguous regions of constant phase. Phase boundaries were drawn as guides for the eye. Straight-line segment A-B, represented a line of points on the plot with $\xi_1=0.512$, and, thus, this line passed through all the designs of Fig. 3. In order, starting at A, the phases on this line segment, were 1, 2_1, 4, 4T, 4, 2_1T, and 2_1. Thus, along this line segment, five of the seven uniquely named phases were observed.

If line segment A-B were extended down and to the left in Fig. 4, it would pass into another Phase 4 region, which was observed for all designs with $\xi_1=\xi_2$, i.e., along the lower boundary of the plot. Detail: When $\xi_1=\xi_2$, the full symmetry group to be used was the full symmetry group of the square, viz., Z_4, rather than the Klein four group. Group Z_4 allowed for two additional reflection symmetries, viz., those about the two diagonals of

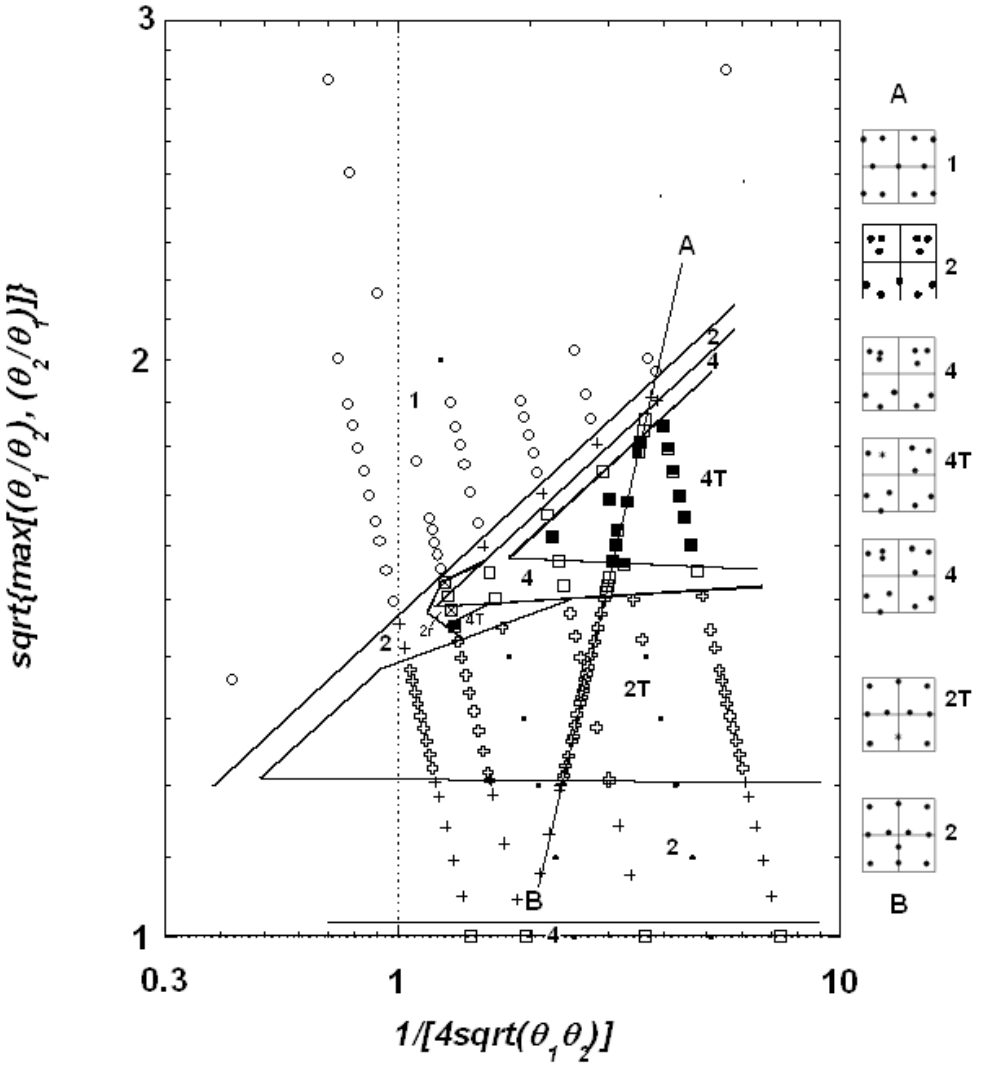


Figure 4: The above dimensionless phase diagram shows contiguous regions of the covariance parameters in which all designs share the same set of invariances under either of the possible mirror symmetries about the x_1 and x_2 axes. The line A-B starts in a Phase 1 region, in which all designs transform to themselves under either or both of the possible mirror symmetries. A prototypical design in this phase is shown in the top-most set of designs to the right of the large, main plot. Line A-B then passes through a phase labeled "2," in which all designs transform either to themselves or to one other design with equal *IMSE*. The line then passes through a phase labeled "4," where all designs transform to themselves or to any of three other designs related to the first by mirror symmetries. Next along the line is a phase labeled "4T," in which all designs include exactly one twin point. The line then appears to reenter phase "4." Then the line enters a phase "2T" in which all designs include a twin point. Finally, the line appears to reenter phase "2." The plot was originally generated using the symbol θ for the covariance parameters, rather than ξ, and the plot has not been updated. The connection between these values is given on the second page of this article.

the square design domain, and this fact led to additional possible phases. As none of the four designs with $\xi_1=\xi_2$ was observed to possess a mirror symmetry about a diagonal, we didn't require, for this study, to add additional phase names, so we simply maintained the original seven names mentioned two paragraphs above. All of the seven phases, except 2_2T were observed.

When both abscissa and ordinate values were approximately *1.5*, we observed a set of especially closely spaced phase transitions. In particular, we observed a single Phase 4T twin-point design on a straight line segment between the only two Phase 2_2 designs observed in this study. Along an extension of this line segment, six of the seven named designs were observed.

3. Theory

In this section, we demonstrate two key results of our theory of clustered designs.

3.1 The 0th- and 1st-degree terms of the power series expansion, in the δ_k's, of the determinant of the covariance matrix of a design with $D \geq 2$ factors and a pair of twin points separated by distance 2δ are zero.

The symmetric covariance matrix of a *D*-factor design with one pair of twin points and Gaussian-covariance function can be written in block form as

$$\frac{V}{\sigma_z^2}=\left[\begin{array}{c|cc} E_{i,j} & F_{i,1} & F_{i,2} \\ \hline \cdot & 1 & J_{1,2} \\ \cdot & \cdot & 1 \end{array}\right],$$

where the sizes of Blocks E, F, and J are $(N-2)^2$, (N-2)x2, and 2x2, respectively.

After changing variables to replace the *2D* coordinate components of the twin points with *D* coordinates of the center of the twins, $\boldsymbol{x_t=(x_1+x_2)/2}$, and *D* components of the vector from the center of the twins to the first of the twin points: $\boldsymbol{\delta = x_1 - x_t}$, the elements of V/σ_z^2 are the following:

$$E_{i,j}=\exp\left[-\sum_{k=1}^{D}\theta_k\left(\mathrm{x_{i,k}}-\mathrm{x_{j,k}}\right)^2\right] \qquad J_{1,2}=\exp\left(-4\sum_{k=1}^{D}\theta_k\delta_{CL1,1,k}^2\right)$$

$$F_{i,1}=\exp\left\{-\sum_{k=1}^{D}\theta_k\left(\mathrm{x_{i,k}}-\mathrm{x_{t,k}}-\delta_k\right)^2\right\}=\exp\left\{-\sum_{k=1}^{D}\theta_k\left[\left(\mathrm{x_{i,k}}-\mathrm{x_{t,k}}\right)^2-2\left(\mathrm{x_{i,k}}-\mathrm{x_{t,k}}\right)\delta_k+\delta_k^2\right]\right\}$$

$$F_{i,2}=\exp\left\{-\sum_{k=1}^{D}\theta_k\left(\mathrm{x_{i,k}}-\mathrm{x_{t,k}}+\delta_k\right)^2\right\}=\exp\left\{-\sum_{k=1}^{D}\theta_k\left[\left(\mathrm{x_{i,k}}-\mathrm{x_{t,k}}\right)^2+2\left(\mathrm{x_{i,k}}-\mathrm{x_{t,k}}\right)\delta_k+\delta_k^2\right]\right\}.$$

We note that all the elements of V/σ_z^2 can be expanded in Laurent series in the δ_k, with leading terms unity. The determinant of V can also be expanded as a Laurent series as

$$\det(V) = \det^{(0)}(V) + \sum_{k=1}^{D}\left[\det^{(1,k)}(V)\cdot\delta_k\right] + \sum_{k=1}^{D}\left[\det^{(2,k)}(V)\right]\cdot\delta_k^2$$

$$+\sum_{k=1}^{D-1}\sum_{k'=k+1}^{D}\left[\det^{(2,k)}(V)\right]\cdot\delta_k\cdot\delta_{k'}+\cdots,$$

which, from the linear independence of the δ_k, may be expressed alternatively as

$$\det(V) = \det(V^{(0)}) + \sum_{k=1}^{D}\left[\det(V^{(1,k)})\cdot\delta_k\right] + \sum_{k=1}^{D}\left[\det(V^{(2,k)})\right]\cdot\delta_k^2$$

$$+\sum_{k=1}^{D-1}\sum_{k'=2}^{D}\left[\det(V^{(2,k,k')})\right]\cdot\delta_k\cdot\delta_{k'}+\cdots.$$

From the fact that the constant term of the power series of each element is unity and that the determinant of an arbitrary matrix ***A*** can be expressed, via a Leibniz formula, as a sum of products of elements of ***A***, it follows that the constant term in the power series of *det(**V**)* is an *NxN* matrix of *1*'s, and therefore that $\det^{(0)}(V) = \det\left(V^{(0)}\right)$=0. Further, *det(**V**)* must be invariant under interchange of the twin points, and this leads to the result that all odd powers of *det(**V**)* are zero. Thus, we see that the 0th- and 1st-degree terms in the power series of *det(**V**)* are identically zero, which was the result sought. It is evident that this result holds, *mutatis mutandis*, for almost any useful covariance function.

3.2 The IMSE objective function, in the vicinity of a pair of twin points separated by distance 2δ, is almost always a low-degree-truncated rational function with leading terms δ^2 in both numerator and denominator.

We start with two matrix identities.

Identity 1: The trace of the matrix product of two, equal-sized, square matrices, ***A*** and ***B***, with ***B*** symmetric, equals the sum of the element-by-element products.

Proof: $\mathrm{tr}\left(\mathrm{A}\otimes\mathrm{B}\right) = \sum_{i=1}^{N}\left(\mathrm{A}\otimes\mathrm{B}\right)_{i,i} = \sum_{i=1}^{N}\sum_{j=1}^{N}A_{i,j}B_{j,i} = \sum_{i,j}A_{i,j}B_{i,j}$.

Identity 2: The elements of the inverse of symmetric, invertible matrix ***L*** are $L^{-1}_{i,j} = (-1)^{i+j}M_{i,j}/\det(L)$.

Proof: From $\left(L^{-1}\right)_{i,j} = C_{i,j}^{T} / \det\left(L\right) = C_{j,i} / \det\left(L\right)$, where $C_{i,j}$ is the cofactor of $L_{i,j}$, and from $C_{j,i} = \left(-1\right)^{i+j} M_{j,i}$, where $M_{i,j}$ is the i,j minor of $\boldsymbol{L}$, then $\left(L^{-1}\right)_{i,j} = \left(-1\right)^{i+j} M_{j,i} / \det\left(L\right) = \left(-1\right)^{i+j} M_{i,j} / \det\left(L\right)$.

We now outline the proof of the statement in the heading of this subsection.

Using the definition of *IMSE* from [9]**,** as well as the two identities, above,

$$IMSE = \sigma_z^2 - tr\left(L^{-1} \otimes R\right) = \sigma_z^2 - \sum_{i,j}\left[\left(L^{-1}\right)_{i,j} R_{i,j}\right] = \left\{\sigma_z^2 \det(L) - \sum_{i,j}\left[\left(-1\right)^{i+j} M_{i,j} R_{i,j}\right]\right\} / \det(L).$$

For a twin-point design, the numerator and denominator of *IMSE* can be expanded as a Laurent series in the δ_k's, with the numerator having the following, possibly zero-valued, 0th- and 1st-degree terms, for *D>1*, independent of the $\boldsymbol{\delta}$'s:

$$numerator^{(0)}\left(IMSE\right) = -\sum_{i,j}\left[\left(-1\right)^{i+j} M_{i,j}^{(0)} R_{i,j}^{(0)}\right] \text{ and}$$

$$numerator^{(1)}\left(IMSE\right) = -\sum_{i,j}\left[\left(-1\right)^{i+j}\left(M_{i,j}^{(0)} R_{i,j}^{(1)} + M_{i,j}^{(1)} R_{i,j}^{(0)}\right)\right],$$

where we have used the fact $\det^{(0)}\left(L\right) = \det^{(1)}\left(L\right) = 0$, which follows simply from arguments similar to the one used in Sec. 3.1, above, that showed $\det^{(0)}\left(V\right) = \det^{(1)}\left(V\right) = 0$. The denominator of *IMSE*, for *D>1*, is *det(L)*, for which the 0th- and 1st-degree powers of δ_k's were shown, in Subsection 3.1, above, to vanish. When *D=1*, *IMSE* can be written as a polynomial in δ [12].

The statement of the heading now requires only for it to be shown that both $numerator^{(0)}(IMSE)=0$ and $numerator^{(1)}(IMSE)=0$. This can be accomplished for all possible cases. Contact the author for details.

4. Application: Interpolation

Because the IMSE function, for small δ_k, almost always takes the form

$$\frac{IMSE}{\sigma_z^2} \equiv 1 - tr\left(L^{-1}R\right) = \frac{a\delta_1^2 + b\delta_2^2 + \cdots + \text{higher-degree terms}}{c\delta_1^2 + d\delta_2^2 + \cdots + \text{higher-degree terms}},$$

computation via evaluation of the inverse of $\boldsymbol{L}$ leads to ill-conditioning, as is well-known. An alternative is to perform the computation using symbolic-manipulation software, but this method also seems intractable, except for very small problems. Here, we outline an approach to computing IMSE, with small δ_k, that avoids the ill-conditioning altogether, thus allowing for accurate computation of IMSE.

The problem statement is as follows: For *D=2* factor and *N=2* points, which are a pair of twin points centered on the origin of the square prediction region $[-1,1]^2$, and $\theta_1=\theta_2=1$, find $\lim_{\substack{\delta_1\to 0 \\ \delta_2=0}} IMSE$.

From the discussion in this paper, *IMSE* can be expanded as a power series in δ_1 and δ_2, about the center of the twin points, as

$$\frac{IMSE}{\sigma_z^2} = \frac{a\delta_1^2 + b\delta_1^3 + c\delta_1^4 + \mathrm{O}\left(\delta_1^5\right)}{e\delta_1^2 + f\delta_1^3 + g\delta_1^4 + \mathrm{O}\left(\delta_1^5\right)},$$

and, along any straight linethrough $\delta_1=\delta_2=0$, this can be expressed as

$$\frac{IMSE}{\sigma_z^2} \cong h + i\delta_1 + j\delta_1^2 + \mathrm{O}\left(\delta_1^3\right).$$

By symmetry under interchange of the twin points, along a discussion similar to one in Subsection 3.1, above, we know that the odd terms of the power-series expansion of *IMSE* are zero, so we are left with

$$\frac{IMSE}{\sigma_z^2} \cong h + j\delta_1^2 + \mathrm{O}\left(\delta_1^4\right). \tag{1}$$

We seek the numerical value of *h*.

We now make reference to Fig. 5, below. For small $\boldsymbol{\delta}$, there is an "excluded region" centered on $\boldsymbol{x_t}$, in which the traditional, all-numeric, matrix computations fail, or are inaccurate, due to ill-conditioning of the *L* matrix. We made all-numeric computations exterior to the excluded region, specifically at points A and B, where $\delta_1=0.0012$ and *0.0014*, respectively, as given in Table 2, below. Extrapolating from these values, it was possible to approximate the values of $IMSE/\sigma_z^2$ in the excluded region, using Eq. 1, above. In particular, it was possible to determine *IMSE=0.7460942972*, at the center of the twin points. This agreed with the value obtained, in this simple case, via symbolic analysis.

5. Application: Inversion of Ill-Conditioned Covariance Matrices

The method of the last section was also applied to inversion of highly ill-conditioned covariance matrices and performed well. Outstanding research questions include determination of the radius of convergence of the power-series expansions, as well as demonstration on a variety of problems. Interested parties may contact the author.

6. Application: Borehole Model

We sought an example where a design with a cluster containing a few to several points might be useful. We decided to explore the well-known, *D=8*, deterministic-error-model, borehole model [13] and to use *N=27* design points, so the method could be compared to

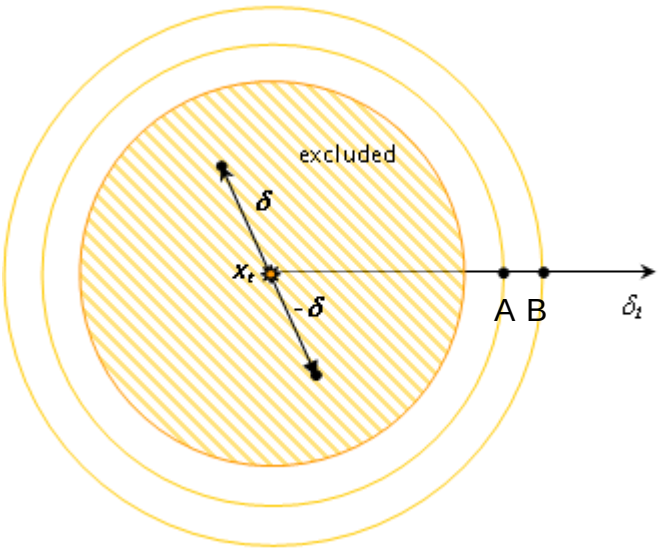


Figure 5: Centered on the center of the twin points, is an exclusion region in which the traditional, all-numeric, matrix computations fail. All-numeric computations made exterior to this region, e.g., at points A and B, are allowed.

Table 2: Design-point labels and values used in the extrapolation

Design-point label	δ_1	δ_1^2	$IMSE/\sigma_z^2$
A	0.0012	0.0000144	0.7460920868…
B	0.0014	0.0000196	0.7460912887…
x_t, via extrapolation	0	0	0.7460942972…
x_t, via symbolic analysis	0	0	0.7460942972…

the blind kriging method of [14]. We took the radical approach of considering, as a first stage in a sequential approach to this problem, a nonuptal-point (9-point) design located near the origin of the scaled variables. Our thinking was that such a design, while leading to extreme ill-conditioning of the ***L*** matrix of the traditional approaches, might be a good means of screening for irrelevant factors and for identifying possible symmetric influence on the response function. (Of course, we knew, a priori, from the problem statement, that variables H_u and H_l appeared anti-symmetrically, but, we adopted, as much as possible, the attitude that we were blind to this fact.) For the first attempt at this problem, we compromised our approach and actually used only a coarsely spaced cluster of nine points, viz., the origin and one step of length *0.1* from the origin in each of the eight Cartesian directions. The results were evaluated using traditional algebraic methods.

The analysis showed there were likely only four highly relevant factors for the response, in the vicinity of the origin, and that two of these had *exactly* anti-symmetric influence on the response. Using these results, we updated our beliefs about the importance of each factor on the response, over the entire domain of the problem. It was possible (but certainly not proved) that there were four irrelevant variables and that two of the

remaining four, viz, H_u and H_l, appeared anti-symmetrically in the response. By noting that the units of H_u and H_l, were identical, viz., lengths, this increased our belief, in a Bayesian sense, that these variables appeared anti-symmetrically over the entire domain of the problem and that these two variables could be replaced by the single, new variable H_u-H_l. If the dimensions of these two variables had not been identical, then it would have been impossible for them to appear in the response function in this way.

For the next design stage we used a four-factor, IMSE-optimal design from JMP Version 7, using thetas of unity, which choice reflected our assumed blindness to the exact nature of the problem, and our knowledge of the well-known fact that the exact specification of thetas might not be critical. As the simulator allowed for the other variables to be inputed, we designed a separate IMSE-optimal design for these variables.

We continued with a final 9-point design and, after analysis, generated the histogram of errors shown in Fig. 6, below, based on 10,000 random evaluations of the final metamodel. This histogram had roughly twice the height (and half the width) of the comparable histogram for blind kriging [14].

Lessons learned: Nothing dramatic was learned from this example, but it was an interesting example of how clustered designs, once they are allowed as design candidates, might play an important role in metamodeling. We were pleasantly surprised to see that the first-stage, quasi-nonuptal-point design led to a final histogram of errors comparable to a design and analysis in which IMSE-optimal designs were used throughout. In addition, clustered designs could potentially identify variables that appear symmetrically or anti-symmetrically in the response. They also allow for highly accurate exploration of the response local to the cluster, a feature not possible with designs that spread out all points. Finally, the highly accurate derivative information provided by clustered designs may prove useful in equation discovery. Further research is clearly indicated.

7. Extensions

Conjecture on the form of det(V) for M-uplet-point designs: We expect that the 0^{th} through *M(M-1)-1*'st-degree terms of the power series expansion of *det(V)* in the δ_k's are all zero. A proof along these lines would be helpful in extending the theory and practice of clustered designs to triplet-point designs, quadruplet-point designs, etc.

Predictand: Twin-point designs: Responding to a question posed by Prof. Thomas J. Santner at SRC 2012, we found the following sensible formula for the Gaussian-process predictand for the *D=1*, *N=2* case, with a pair of twin point centered at x_t:

$$\widehat{Y} = y\left(x_t\right) + \left[y'\left(x_t\right)\cdot\left(x - x_t\right)\right]\cdot\exp\left[-\theta\left(x - x_t\right)\right],$$

where $y'(x_t)$ is the first derivative of $y(x)$, evaluated at the center of the twin points.

Predictand: Triplet-point designs: However, for three equally spaced, triplet points in *D=1*, we found, and here report as an intriguing, tentative finding, that the usual GP fitting includes some nonsensical terms. Discarding the offending terms led to the following sensible formula:

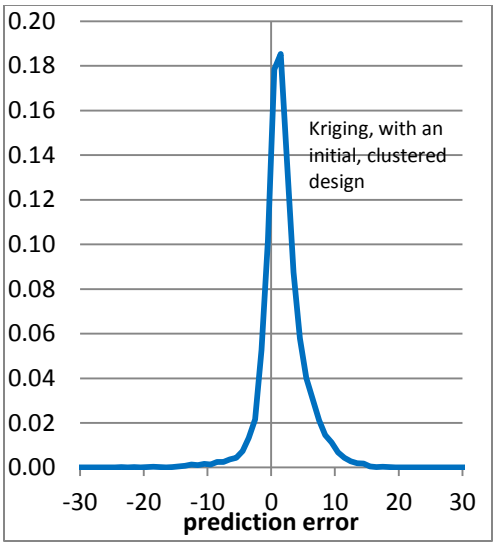


Figure 6. This plot shows the histogram of errors for the borehole example, achieved using an initial quasi-nonuptal-point design near the origin of the design domain, as described in this section.

$$\widehat{Y} = y(x_t) + \left[y'(x_t)\cdot(x - x_t) + \frac{1}{2} y''(x_t)\cdot(x - x_t)^2 \right] \exp\left[-\theta(x - x_t)\right].$$

Predictand Conjecture I: Tight, M-uplet-point clusters: The above results on predictands leads us to conjecture that the correct way to extrapolate from any *M*-uplet-point cluster centered as $\boldsymbol{x_t}$ is via a formula in which the response and all the appropriate directional derivatives are evaluated at the center of the cluster and with the derivative parts decaying with the appropriate multivariate equivalent of the form $\exp\left[-\theta\left(x - x_t\right)\right]$.

Predictand Conjecture II: Loose, M-point clusters: The literature contains examples of GP fits that get worse with the addition of points. One example is in Haaland and Qian's recent paper [17], which showed, via a series of panels in its Fig. 1, the deterioration of the numerical accuracy of the predictand, with increasing *N*, for the function

$$f(x)=\exp\{(x+1/2)^2 \sin(\exp\{(x+1/2)^2\}),$$

over the interval [0,1] with a Gaussian-covariance fits with $\theta=1$. Based on the last three paragraphs, above, we conjecture that a sensible fit for this problem would be to make any sensible local fit to the loose cluster of data points, e.g., via a cubic spline, or even a more radical *M-1'*st-order spline, and then connect the loose cluster with other regions of the domain of interest, via the appropriate multivariate equivalent of the form $\exp\left[-\theta\left(x - x_t\right)\right]$. In this way, the predictand, local to the cluster improves with increasing *N*, while maintaining the presumably correct mid-range and long-range behavior. Details of this approach will be discussed by the present author at ACAS 2012.

The nugget: The present author thinks the community should maintain an open view on whether or not to include a nugget in GP fits. We point out that the last conjecture, above, leads to an alternative approach that does not require the invocation of a nugget.

Acknowledgments

This paper is dedicated to the memory of David M. Woodcock, the author's co-researcher at the University of Michigan, Ann Arbor, when the first free-ranging, twin-point designs were discovered and explored. David will always be remembered for being extraordinarily sound of mind and spirit.

We thank Prof. Art Owen of Stanford University for his interest, encouragement, and suggestions, especially during the development of twin-point theory in Y2011; Rachel Silvestrini of the Naval Postgraduate School and co-author of [12], for pointing out that SWMW had observed a twin-point design and for independently suggesting the method for inverting covariance matrices, given in of Section 5 of this report; Dr. Bradley Jones of SAS Institute, JMP Division, for informing the author of SRC 2012; Erin Leatherman of Ohio State University (OSU) for identifying a discrepancy between [1] and the first example in the author's SRC-2012 presentation; Prof. Tom J. Santner of OSU for asking, at SRC 2012, a key question about the GP predictand, in the presence of ε-clustered designs; Fred J. Hickernell of the Illinois Institute of Technology for his immediate, spirited interest in twin-point designs, and David Deacon of Los Altos, California for a variety of thoughtful suggestions.

References

1. Selden B. Crary, David M. Woodcock, and Andreas Hieke, "Designing efficient computer experiments for metamodel generation," Published in the *Proceedings of the Fourth International Conference on Modeling of Microsystems, MSM 2001*, Hilton Head, SC, March 19-21, 2001, pp. 132-135.
2. William G. Cochran, *Sampling Techniques* (3rd ed.), New York: John Wiley (1977).
3. Max D. Morris, "Factorial Sampling Plans for Preliminary Computational Experiments," *Technometrics* **33**, pp. 161-714 (1991).
4. Max D. Morris, Toby J. Mitchell, and Donald Ylvisaker, "Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction," *Technometrics* **33**, pp. 243-255 (1993).
5. K. V. Mardia, J. T. Kent, C. R. Goodall, and J. A. Little, "Kriging and splines with derivative information," *Biometrika* **83**, pp. 207-221 (1996).
6. Gemma Stephenson (2010) "Using derivative information in the statistical analysis of computer models," University of Southampton, School of Ocean and Earth Science, Ph.D. Dissertation, pp. i-197.
7. Selden B. Crary, "Statistical design and analysis of computer experiments for the generation of parsimonious metamodels," Published in *Design, Test, Integration, and Packaging of MEMS/MOEMS 2001*, B. Courtois, J.M. Karam, S.P. Levitan, K.W. Markus, A.A.O. Tay, and J.A. Walker, Editors, Proceedings of SPIE Vol. 4408, pp. 29-39 (2001).
8. Selden B. Crary, "Design of computer experiments for metamodel generation," *Analog Integrated Circuits and Signal Processing* **32**, pp. 7-16 (2002).

9. Jerome Sacks, Susannah B. Schiller, and William J. Welch, “Design of computer experiments,” *Technometrics* **31**, pp. 41-47 (1989).
10. Jerome Sacks, William J. Welch, Toby J. Mitchell, and Henry P. Wynn, “Design and analysis of computer experiments,” *Statistical Science* **4**, 409-423 (1989).
11. Fred J. Hickernell, private communication at and during the two weeks, immediately following SRC 2012. The design and transition are reported, here, with permission.
12. Selden B.. Crary and Rachel Johnson, “Validation of the Twin-Point-Design Concept in the Design of Computer Experiments,” Section on Statistical Computing – JSM 2011, pp. 5495-5505.
13. Handbook of statistics, 22: Statistics in industry, R. Khattre and C. R. Rao (eds.), Amsterdam: Elsevier Science B. V. (2003). An on-line description of the problem is available at URL: http://www.jmp.com/support/help/Borehole_Model_3a_A_Sphere-Packing_Example.shtml.
14. V. Roshan Joseph, Ying Hung, and Agus Sudjianto, “Blind kriging: A new method for developing metamodels,” *ASME Journal of Mechanical Design* **130**, 031102-1-8 (2008).
15. Ben Haaland and Peter Z.G. Qian, Accurate emulators for large-scale computer experiments, *Annals of Statistics* **39**, pp. 2974-3002 (2011).