

DESIGNING EXPERIMENTS FOR CAUSAL NETWORKS

William D Heavlin, Advanced Micro Devices, MS 117

One AMD Place, Sunnyvale, CA 94088-3453 <bill.heavlin@amd.com>

Key words: blocking, directed graphs, multidimensional scaling, optimal design, tolerance design.

Abstract: Causal networks are directed graphs that generalize Ishikawa diagrams to encompass multiple responses. Emphasizing tolerance design applications, this work presents an optimal design algorithm when the variables are organized as a causal network. The causal network is first transformed into a causal map, which represents all factors and responses as points in a common D -dimensional metric space. The design approach is algorithmic, optimizing Wynn's entropy criterion. This criterion maximizes dispersion among predicted multivariate responses, using a distance-in-space coefficients (DiSCo) model. A key constraint is block self-containment—the blocks are analyzable without reference to one another; these analyses are to be complemented by a unified all-block analysis. Also explored is the benefit of skewing blocks by setting a few factors off-target.

1. Causal Networks

In 1943, Ishikawa introduced cause-effect (CE) diagrams for quality control. CE diagrams render causal relationships hierarchally, with the single response as the root node, factor groups the primary branches and finer branches from there. Causal networks are directed graphs that generalize CE diagrams in three ways: (1) the number of responses can be more than one, (2) responses can point causally to other responses, and (3) each factor is represented as one and only one node, yet may contribute to more than one response and/or higher-level factor. Regarding (3), causal networks are only fully hierarchal for assembly manufacturing processes, but many are directed acyclic graphs.

Here causal networks are used to design experiments. The application area is roughly that of CE diagrams—manufacturing processes with many (10-50) input factors, all of which contribute to response variation. The number of factors F is large enough to suggest a systematic approach. Factor interactions are also important, but F is too large for a resolution V design.

2. Tolerance Design

A new semiconductor manufacturing process has three phases: (i) Development, where alternatives are evaluated and targets tuned. (ii) Pre-production, where process targets are set, and tolerances tested. (iii) Production, which fixes targets and tolerances, commits manufacturing resources to volume. Pre-production

manages risk, providing time for finding critical steps, checking manufacturability, and testing reliability.

Competitive pressures offer strong incentives for a brief pre-production phase. Better than nominal runs at characterizing processes are designed experiments, the application of which Taguchi terms *tolerance design*.

3. Self-Contained Blocks

Implicit is the need for thorough characterization, involving most or all declared factors. Conventional application of CE diagrams usually result in several experimental designs, not one, each a network subset. Computer scientists term this kind of decomposition *graph partitioning*. Here the concept of partitions aligns to the experimental block structure. The blocks are sub-experiments, each a subset of runs and of factors, yet ultimately taken together as a whole.

To be acceptable to semiconductor manufacturing, experimental designs are constrained: (1) The number of per-block factors F_l is only a fraction (8, say) of F potential factors (50, say). (2) The high cost of experimental lots results in each experiment, each block, being reported separately. In this way, the blocking structure represents the partition into smaller experiments, motivating the following definitions:

Definition 3.1. A factor changed within a given block is called a *split factor*.

Definition 3.2. An experimental block is *self-contained* (a) if its split factors are not confounded with one another, and (b) all other factors are fixed. Definition 3.2 does not require that the other factors be at their nominal (i.e. target or central) level, hence:

Definition 3.3. A factor held constant within a given block is called a *skew factor*. This term is applied especially to a factor (1) that is held constant but not at its nominal level, and (2) that for other blocks, the skew factor is at other than its current level.

Caveat 3.4. If a block has centerpoints, the skew factor may or may not affect them. Skew factors are defined with respect to the factorial part of the block.

Remark 3.5. The labels *split* and *skew* are in the context only of a given experimental block. This classification differs from that of Bingham and Sitter's (1999) optimal split-plot theory, wherein a factor is split (sub-plot) or skew (whole-plot) across *all* blocks.

4. Problem Statement

Represent the F factors and R responses in a causal network. Let F_l denote the per-block number of split factors, and n the per-block number of experimental runs; $n = n_0 + n_{\pm}$, n_0 the number of centerpoints per block and n_{\pm} the number of factorial points.

Each block is to be self-contained. Over groups of $b = \text{ceiling}(F/F_1)$ blocks, a higher level block called here a *cycle*, all F factors are split at least once; a cycle of b blocks is the minimum feasible. Upper bounds for F , F_1 , n , and c are roughly 50, 8, 25, and 3.

Goal: Derive a useful experimental design, consistent with (1) the causal network and constrained by both the (2) F , F_1 , n , and c limits and (3) block self-containment requirement.

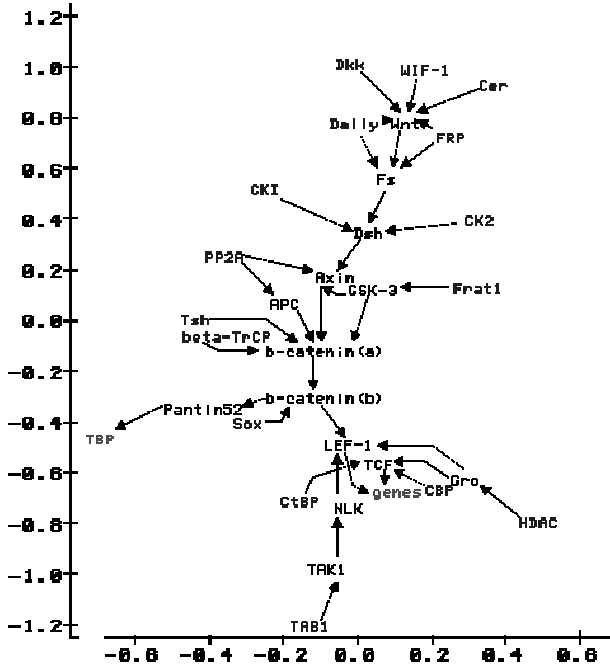


Figure 1. Moon's Wnt/ β -catenin causal network.

5. Causal Maps and the Running Example

The causal networks of semiconductor processes are large, complicated, and proprietary, making them poor examples. Moon's (2000) pathway is more suitable. This causal network has 16 input nodes or factors, and we choose two nodes for responses. Transforming Moon's network into Figure 1 involves these steps: (i) The distance between nodes a and b is the number of links connecting them. (ii) Apply multidimensional scaling (Buja et al, 1998) to this matrix of link-count distances to obtain ($D=$) 2-dimensional coordinates.

This coordinate-based version of a causal network, or *causal map*, has (creates!) extra metric information: (1) A factor close to a response plausibly has a stronger effect. (2) Two factors close together likely share an interaction. (3) Responses sharing many factors cluster, and (4) higher-level responses tend toward the centroid.

6. Data Structures and Algorithm

Base design matrix: Each block has F_1 split factors, and F_2 of skew factors; $0 \leq F_2 \leq F - F_1$. A solution assigns columns of base designs to particular factors.

The algorithm builds a design consisting ultimately of cb blocks, b blocks at a time; the group of b blocks constructed together comprise a *cycle*, and there are c cycles: Let $q=1, \dots, c$; let \mathbf{U} denote the $m \times F$ matrix of previously constructed experiments, $m \geq 0$, \mathbf{X} the generic $bn \times F$ one-cycle base design matrix, and \mathbf{B} the dummy variables labeling blocks. Denote the step q solution as \mathbf{X}_q , and $\mathbf{U}_q = [\mathbf{X}_q^T \mid \mathbf{U}_{q-1}^T]^T$; \mathbf{U}_0 has zero rows; each cycle adds b blocks, and the final solution is \mathbf{U}_c .

Figure 2 gives examples of base design matrices \mathbf{X} consisting of $F_1=4$ split factors, Figure 2a with no skew factors ($F_2=0$), Figure 2b with $F_2=3$. In Figures 2a and 2b, each factor is split exactly once per cycle.

Let \mathbf{P} denote an $F \times F$ permutation matrix. Observe that $\mathbf{X}\mathbf{P}$ constitutes a rearrangement of columns of \mathbf{X} . Construct a design matrix $\mathbf{W}(\mathbf{P}) = [(\mathbf{X}\mathbf{P})^T \mid \mathbf{U}^T]^T$; \mathbf{P} rearranges the columns of base design \mathbf{X} , and $\mathbf{W}(\mathbf{P})$ then appends rows \mathbf{U} . Denote by $\mathbf{Y}(\mathbf{P})$ the $(bn+m) \times R$ matrix of predicted responses given design $\mathbf{W}(\mathbf{P})$.

DiSCo Model: Let \mathbf{F} and \mathbf{R} denote the $F \times D$ and $R \times D$ matrices whose rows are the coordinates mapping the F causal factors and the R responses into a given D -dimensional causal map. $D=2$ or 3 are most visual.

Let \mathbf{A} denote an $F \times R$ coefficient matrix, elaborated below. Observe that the $n \times R$ matrix $\mathbf{Y}(\mathbf{P}) = \mathbf{W}(\mathbf{P})\mathbf{A}$ gives linear-model response predictions. To describe models with interactions, let \mathbf{Z} denote some $q \times F$ matrix, and let $\mathbf{g}(\mathbf{Z})$ denote the $q \times (F(F+1)/2)$ matrix formed by joining to \mathbf{Z} the additional $F(F-1)/2$ columns formed as products of all pairs of columns. Applied to $\mathbf{W}(\mathbf{P})$, the result is $\mathbf{g}(\mathbf{W}(\mathbf{P})) = [\mathbf{g}(\mathbf{X}\mathbf{P})^T \mid \mathbf{g}(\mathbf{U})^T]^T$. Let \mathbf{A}_g denote some $(F(F+1)/2) \times R$ coefficients matrix. The $n \times R$ matrix $\mathbf{Y}_g(\mathbf{P}) = \mathbf{g}(\mathbf{W}(\mathbf{P}))\mathbf{A}_g$ holds the response predictions with respect to an interaction model. For typical element a_{fr} of matrix \mathbf{A} , corresponding to factor f and response r , take $a_{fr} = \exp\{-\alpha\|\mathbf{f}_f - \mathbf{r}_r\|^2\}$, \mathbf{f}_f a row of \mathbf{F} , \mathbf{r}_r a row of \mathbf{R} , $\alpha \sim 1$. For the element of \mathbf{A}_g corresponding to response r , and the interaction of factors f_1 and f_2 , take it equal to $a_{f_1 r} \times a_{f_2 r} \times \exp\{-\gamma\|\mathbf{f}_1 - \mathbf{f}_2\|^2\}$, where $\gamma \sim 2$. This \mathbf{A}_g resembles the non-additive model proposed by Tukey (1949). The linear model form $\mathbf{Y}_g(\mathbf{P}) = \mathbf{g}(\mathbf{W}(\mathbf{P}))\mathbf{A}_g$, the function $\mathbf{g}(\cdot)$, and the choice of coefficients \mathbf{A}_g together form the *distance-in-space coefficients* ("DiSCo") model.

A connection between DiSCo and resolution IV: The latter has $(b-1)+2F$ terms. When $D=2$ and $R=1$, the term count of the DiSCo model equals this. As a function of D , the number of DiSCo terms grows linearly, so for $D \geq 2$, resolution is between IV and V.

Optimality Criteria: The optimality criterion is Wynn's (Sacks et al, 1989). Define $\mathbf{Y}_g(\mathbf{P}) = \mathbf{g}(\mathbf{W}(\mathbf{P})) \times \mathbf{A}_g$, and $\mathbf{Y}_{(\mathbf{B})}(\mathbf{P}) = (\mathbf{I} - \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T) \times \mathbf{g}(\mathbf{W}(\mathbf{P}))\mathbf{A}_g = (\mathbf{I} - \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T) \times \mathbf{Y}_g(\mathbf{P})$. $\mathbf{Y}_{(\mathbf{B})}(\mathbf{P})$ is $\mathbf{Y}_g(\mathbf{P})$ with the effects attributable to blocks \mathbf{B} removed.

With $p=bn+m$, a $p \times R$ matrix \mathbf{Y} , let $\mathbf{C}(\mathbf{Y})$ be the $p \times p$ matrix with typical element $\exp\{-\|\mathbf{y}_a - \mathbf{y}_b\|^2\}$, where

y_a is the a th row of \mathbf{Y} . The optimum design (larger-is-better) criterion is

$$\det(\mathbf{C}(\mathbf{Y}_g(\mathbf{P}))) \times \det(\mathbf{C}(\mathbf{Y}_B(\mathbf{P}))). \quad (6.1)$$

The function $\det(\mathbf{C}(\cdot))$ measures dispersion among $\{y_a\}$. For network monitoring, Federov and Flanagan (1998) propose criteria that generalize Wynn's.

Algorithm: The algorithm maximizes (6.1) with respect to the permutation matrix \mathbf{P} , isomorphic to a traveling salesman problem, with solution $\mathbf{X}_q = \mathbf{X}\mathbf{P}_q$.

7. Choice of matrix \mathbf{X}

In choosing the base design \mathbf{X} , for concreteness take the split factor pattern to be 2_{IV}^{4-1} . The dimensions are (a) F_2 , the number of skew parameters, and (b) b , the number of blocks per cycle. In the following, $F_2=0, 1, 2$, and 3 are considered, hence eight base designs: $\{F_2=0, 1, 2, \text{ and } 3\} \times \{(b=4, c=4), (b=1, c=16)\}$, denoted as $\{\text{IB}(F_2, c \times b): F_2=0 \dots 3, c \times b=4 \times 4 \text{ or } 16 \times 1\}$; 16 blocks.

In each block, include three centerpoints; the total number of runs is 176. The centerpoints separate block from skewed factor effects. Set the centerpoints to nominal for all $F=16$ factors, even for skew factors.

8. Benchmark Designs and Performance Criteria

Two-level Benchmark Designs: With $cb=16$ blocks and $F=16$, a linear model has 32 terms; second-order interactions add $F(F-1)/2=120$, and a resolution V model therefore has $32+120=152$ terms. Call this term set V120. Of the 120 second-order interactions, the 60 with causal map distances less than the median distance are more plausible; call this term set V60.

The blocked 128-run two-level designs have theoretically best efficiency. In constructing these, adapt the algorithm above: (a) a particular criterion is designated, such as the rank with respect to V120, and (b) then select 16 columns maximizing the criterion. Four benchmark designs result: (1) L128 V120; (2) L128 V60; (3) L128 V120+60, using the sum of the two criteria, (4) L128 Wynn, using 6.1.

Performance Criteria are five: (i) criterion 6.1; column ranks of the (ii) V120 and (iii) V60 term sets; the factor standard errors (iv) that is largest, and (v) smallest. Criteria (ii-v) are in Table 1.

9. Results

Each L128 design does well on its own criterion, L128 Wynn poorly on the column rank measures, L128 V120+V60 is equivocally best. For the IB designs, (a) $\text{IB}(F_2, c \times b=16 \times 1)$ offers greater response dispersion than $\text{IB}(F_2, c \times b=4 \times 4)$, yet (b) consistently worse performance on V120 and V60 column ranks and (c) worse efficiency; (d) indeed, only $\text{IB}(3, 16 \times 1)$ is able to estimate coefficients for all 16 factors. From (b-d), one would limit $\text{IB}(F_2, c \times b=16 \times 1)$ to special circumstances, e.g. factor pre-screening, where it can be quite useful.

Among the $\text{IB}(F_2, c \times b=4 \times 4)$ designs, (e) $\text{IB}(F_2, c \times b=4 \times 4)$ shows greater response dispersion for larger values of F_2 , but (f) less than $\text{IB}(F_2, c \times b=16 \times 1)$ and the L128 designs. (g) $\text{IB}(0, 4 \times 4)$ achieves its largest feasible V120 column rank. (h) Both V120 and V60 column ranks increase in F_2 . (i) $\text{IB}(3, 4 \times 4)$ has the largest feasible V120 column rank, 144, while (j) its V60 column rank is comparable to L128 designs. (k) The worst-case linear efficiencies of $\text{IB}(F_2, 4 \times 4)$ are about 25 percent, (l) offering 33 percent on the best-case scale. Skew factors improve estimation efficiency inconsistently, but per (e), increase response dispersion.

10. Comments and Conclusions

The arbitrary and general nature of the causal network data structures implicitly favors an optimal design approach. Practical experiments tend toward large F and multiple blocks, each block emphasizing a different sub-networks and response subsets. Unlike proposals developed here, graph partitioning blocks such that factors close together tend to be split within a common block. Rather, the algorithm presented here emphasizes dispersion among all responses, and avoids preferring any single set of responses.

Interblock designs offer (i) increased information on interactions and (ii) greater response dispersion; the trade-off involves (iii) increased complexity (F_2). (iv) Centerpoints, separating block from skew factor effects, are key to their benefit. The choice of \mathbf{X} dominates (i), (iii), and (iv), while the algorithm affects (ii). Because \mathbf{X} does not depend on \mathbf{A} , the experimental designs are likely robust to causal map changes.

11. References

- Bingham, D and Sitter, R (1999), "Minimum-aberration two-level fractional factorial split-plot designs," *Technometrics*, vol 41, pp 62-70.
- Buja, A, Swayne, DF, Littman, ML, and Dean, N (1998), "XGvis: interactive data visualization with multidimensional scaling," *Journal of Computational and Graphical Statistics*, Preprint.
- Fedorov, V and Flanagan, D (1998), "Optimal monitoring of computer networks," in *New Developments and Applications in Experimental Design*, IMS Lecture Notes, vol 34, pp 1-10.
- Moon, Randall T (2000), "Pathway: Wnt/ β -catenin," http://www.stke.org/cgi/cm/CMP_5533, accessed May 18, 2000.
- Sacks, J, Welch, WJ, Mitchell, TJ, and Wynn, HP (1989), "Design and analysis of computer experiments," with discussion, *Statistical Science*, vol 4, pp 409-435.
- Tukey, JW (1949), "One degree of freedom for non-additivity," *Biometrics*, vol 5, pp 232-242.

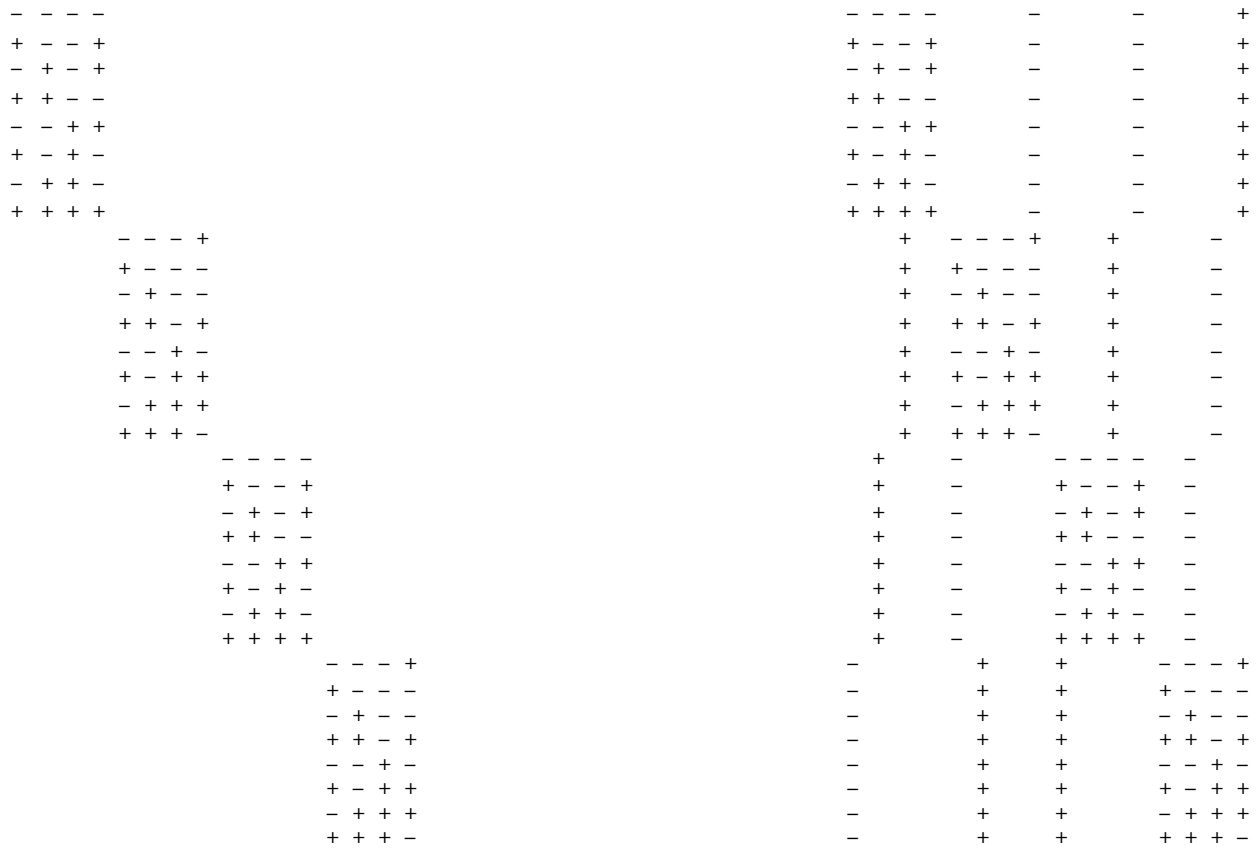


Figure 2. (a) Left-hand side, a split base design in four 8-run blocks, but no skews. (b) Right-hand side, a split base design in four 8-run blocks with three skew factors per block.

design	V120 column rank	V60 column rank	worst factor's efficiency	best factor's efficiency	notes
IB(0, 4x4)	80	69	0.250	0.252	
IB(0,16x1)	72	55	0.055	0.786	3 factors lost
IB(1, 4x4)	117	76	0.245	0.369	
IB(1,16x1)	84	59	0.068	0.831	2 factors lost
IB(2, 4x4)	130	83	0.248	0.359	
IB(2,16x1)	96	59	0.030	0.778	1 factors lost
IB(3, 4x4)	144	90	0.270	0.336	
IB(3,16x1)	120	75	0.089	0.691	
L128 Wynn	87	61	0.973	1.000	
L128 V120	128	85	1.000	1.000	
L128 V60	117	91	0.976	1.000	
L128 V120+V60	128	89	0.974	1.000	

Table 1. Numerical performance criteria for the 8 interblock designs and 4 benchmark designs. All designs consist of 16 factors in 16 blocks; each block has 8 factorial runs and 3 centerpoints. IB(0,16x1), IB(1,16x1), and IB(2,16x1) estimate only 13, 14, and 15 of the 16 factors, respectively.