# Projection-Based Performance Modeling for Inter/Intra-Die Variations

Xin Li[1], Jiayong Le[2], Lawrence T. Pileggi[1] and Andrzej Strojwas[1]

[1]Dept. of Electrical & Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{xinli, pileggi, ajs}@ece.cmu.edu

[2]Extreme DA
165 University Avenue
Palo Alto, CA 94301, USA
kelvin@extreme-da.com

## Abstract

Large-scale process fluctuations in nano-scale IC technologies suggest applying high-order (e.g., quadratic) response surface models to capture the circuit performance variations. Fitting such models requires significantly more simulation samples and solving much larger linear equations. In this paper, we propose a novel projection-based extraction approach, PROBE, to efficiently create quadratic response surface models and capture both inter-die and intra-die variations with affordable computation cost. PROBE applies a novel projection scheme to reduce the response surface modeling cost (i.e., both the required number of samples and the linear equation size) and make the modeling problem tractable even for large problem sizes. In addition, a new implicit power iteration algorithm is developed to find the optimal projection space and solve for the unknown model coefficients. Several circuit examples from both digital and analog circuit modeling applications demonstrate that PROBE can generate accurate response surface models while achieving up to 12x speedup compared with the traditional methods.

## 1. Introduction

As IC technologies scale to finer feature sizes, it becomes increasingly difficult to control the relative process variations, particularly due to sub-wavelength photolithography [1]-[2]. The increasing fluctuations in manufacturing process have introduced unavoidable and significant uncertainty in circuit performance. Hence, modeling and analyzing these random process variations to ensure manufacturability and improve yield has been identified as a top priority for today's IC design problems.

In order to address this process variation problem, response surface models [3] are utilized to capture the circuit performance variations caused by manufacturing fluctuations. The objective of response surface modeling is to approximate the circuit performance (e.g., delay, gain) as a polynomial (e.g., linear or quadratic) function of variational process parameters (e.g., $V_{TH}$, $T_{OX}$). These models are extensively applied in many applications such as statistical timing analysis [1], analog mismatch analysis [4], yield optimization [5]-[6], etc.

Most of the previous response surface models, e.g., [1], utilize linear approximations, which are efficient and accurate when process variations are sufficiently small. However, two recent changes in advanced IC technologies suggest a need to revisit this assumption. Firstly, process variations are becoming relatively larger. As reported in [1], the gate length variation can reach ±35% in nano-scale technologies. This, in turn, implies the importance of applying high-order (e.g., quadratic) response surface models to guarantee high approximation accuracy [3], [6], [7]. Applying nonlinear response surface models is especially important for analog circuits, since many analog performances (e.g., offset voltage) can be strongly nonlinear in the presence of large-scale variations.

Secondly, but most importantly, intra-die variations (i.e., mismatches) are becoming increasingly important [2], especially for analog circuits [4]. These intra-die variations model the individual, but spatially correlated, local variations within the same die. The intra-die variations must be modeled by using many additional random variables, thereby significantly increasing the number of unknown model coefficients. Therefore, more simulation samples are required in order to determine all these unknown coefficients by solving a larger linear equation. This makes model fitting much more expensive, especially when using high-order response surface models. For example, the number of unknown coefficients (hence the required number of samples and the linear equation size) in a quadratic response surface model will quadratically increase in the number of random process parameters, thereby quickly making the quadratic model fitting infeasible. For this reason, generating accurate high-order (e.g., quadratic) response surface models with affordable computation cost becomes a new challenging problem in nano-scale technologies.

In this paper we propose a novel Projection-Based Extraction (PROBE) for quadratic response surface modeling. The novelty of PROBE lies in our new formulation of the model fitting problem such that quadratic response surface modeling becomes tractable even for large-size problems. Instead of fitting a full-rank quadratic model, PROBE applies *projection* operator and attempts to find an optimal low-rank model by minimizing the approximation error. In PROBE, the modeling accuracy can be easily traded for simplicity by increasing or decreasing the dimension of the projection space. Most importantly, taking advantage of this novel projection scheme, PROBE can dramatically reduce the number of unknown coefficients that need to be solved, thereby significantly reducing the fitting cost and facilitating scaling to much larger problem sizes.

Another important contribution of PROBE is a new *implicit power iteration* algorithm to find the optimal projection space and extract the unknown model coefficients. This iteration solves a sequence of over-determined linear equations and exhibits robust convergence. Using the proposed implicit power iteration algorithm, PROBE can achieve significant speedup for generating low-rank quadratic response surface models. As demonstrated by the numerical examples from both digital and analog circuit modeling applications, PROBE can extract accurate models and reduce the computation cost by up to 12x compared with the traditional full-rank quadratic modeling.

The remainder of the paper is organized as follows. In Section 2 we review the background on response surface modeling. Then, we propose our PROBE approach, including both the theoretical analysis and the implicit power iteration algorithm, in Section 3. The computational efficiency of PROBE is demonstrated by several circuit examples in Section 4, followed by the conclusions in Section 5.

## 2. Background

Given a circuit topology, the circuit performance (e.g., delay, gain) is a function of the design parameters (e.g., bias current,

0-7803-9254-X/05/$20.00 ©2005 IEEE.

transistor sizes), as well as the process parameters (e.g., $V_{TH}$, $T_{OX}$). The design parameters are optimized and fixed during the design process; however, the process parameters must be modeled as random variables to account for any uncertain variations. Given a set of fixed design parameters, the circuit performance $f$ can be approximated by a linear response surface model [1], [3]:

$$f(X) = B^T X + C \qquad (1)$$

where $X = [x_1, x_2, ..., x_N]^T$ represents the process variations, $B \in R^N$ and $C \in R$ stand for the model coefficients and $N$ is the total number of the variational process parameters.

The linear model in (1) is not sufficiently accurate for modeling the large-scale process variations that are expected for nano-scale technologies. It, in turn, suggests that applying quadratic response surface models might be required to improve the modeling accuracy [3], [6], [7]:

$$f(X) = X^T AX + B^T X + C \qquad (2)$$

where $C \in R$ is the constant term, $B \in R^N$ represents the linear coefficients and $A \in R^{N\times N}$ denotes the quadratic coefficients. The unknown model coefficients $A$, $B$ and $C$ can be determined by solving the over-determined linear equation [3]:

$$X_i^T AX_i + B^T X_i + C = \tilde{f}_i \quad (i = 1,2,\cdots) \qquad (3)$$

where $X_i$ and $\tilde{f}_i$ are the value of $X$ and the exact value of $f$ for the $i$-th sampling point, respectively.

It is straightforward to verify that the number of unknown coefficients in (3) is $O(N^2)$. The overall computation cost for determining all these coefficients consists of two portions:

- *Simulation cost*: i.e., the cost for running a simulator to determine the performance values $\tilde{f}_i$ at the sampling points $X_i$. The number of simulation samples should be greater than the number of unknown coefficients, in order to uniquely solve the linear equation in (3). Therefore, at least $O(N^2)$ sampling points are required for fitting the quadratic model in (2). In practical applications, the number of samples is generally selected to be significantly larger than the unknown coefficient number to avoid over-fitting.

- *Fitting cost*: i.e., the cost for solving the over-determined linear equation in (3). For the quadratic model in (2), the fitting cost is of the order of $O(N^6)$.

The aforementioned high computation cost limits the traditional quadratic response surface modeling approach [3] to small or medium size applications. This observation, therefore, motivates us to propose a novel projection-based response surface modeling algorithm, PROBE, which can significantly reduce the computation cost.

# 3. Projection-Based Extraction

## 3.1 Mathematic Formulation

The key disadvantage of the traditional quadratic response surface modeling is the need to compute all elements of the matrix $A$ in (2). This matrix is often sparse and rank-deficient in many practical problems. Therefore, instead of finding the full-rank matrix $A$, PROBE approximates $A$ by another low-rank matrix $A_L$. Such a low-rank approximation problem can be stated as follows: *given a matrix A, find another matrix $A_L$ with rank p* < rank($A$) *such that their difference* $||A_L–A||_F$ *is minimized.* Here, $||\bullet||_F$ denotes the Frobenius norm, which is the square root of the sum of the squares of all matrix elements. Without loss of generality, we assume that $A$ is symmetric in this paper, since any asymmetric quadratic form $X^TAX$ can be easily converted to an equivalent symmetric form $0.5\cdot X^T(A+A^T)X$ [8].

From matrix theory [8], for any symmetric matrix $A \in R^{N\times N}$, the optimal rank-$p$ approximation with the least Frobenius-norm error is:

$$A_L = \sum_{i=1}^{p} \lambda_i P_i P_i^T \qquad (4)$$

where $\lambda_i$ is the $i$-th dominant eigenvalue, and $P_i \in R^N$ is the $i$-th dominant eigenvector. The eigenvectors in (4) define an orthogonal projector $P_1P_1^T+...+P_pP_p^T$, and every column in $A_L$ is the *projection* of every column in $A$ onto the subspace $span\{P_1,...,P_p\}$. We use this orthogonal projector for response surface modeling in this paper. Fig. 1 intuitively illustrates the low-rank projection for quadratic response surface modeling.

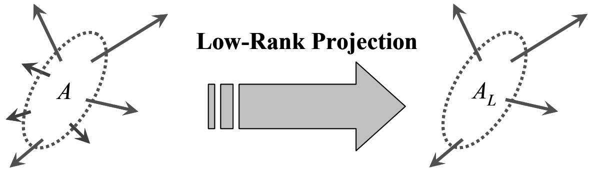


Fig. 1. Illustration of the low-rank projection.

The main advantage of the rank-$p$ projection is that, for approximating the matrix $A \in R^{N\times N}$ in (2), only $\lambda_i \in R$ and $P_i \in R^N$ ($i$ = 1,...,$p$) need to be determined, thus reducing the number of problem unknowns to $O(pN)$. In many practical applications, $p$ is significantly less than $N$ and the number of unknown coefficients that PROBE needs to solve is almost a linear function of $N$. Therefore, compared with the problem size $O(N^2)$ in traditional quadratic modeling, PROBE is much more efficient and can be applied to large-size problems.

## 3.2 Coefficient Fitting via Implicit Power Iteration

Since the matrix $A$ in (2) is not known in advance, we cannot use the standard matrix computation algorithm to compute the dominant eigenvalues $\lambda_i$ and eigenvectors $P_i$ that are required for a low-rank approximation. One approach for finding the optimal rank-$p$ model is to solve the following optimization problem for the unknown coefficients $\lambda_i$ and $P_i$ ($i$ = 1,2,...,$p$) and $B$, $C$:

$$\begin{aligned} & \textit{minimize} \quad \psi = \sum_i \left[ X_i^T \left( \sum_{j=1}^{p} \lambda_j P_j P_j^T \right) X_i + B^T X_i + C - \tilde{f}_i \right]^2 \\ & \textit{subject to} \quad \|P_j\|_2 = 1 \quad (j = 1,\cdots,p) \end{aligned} \qquad (5)$$

where $||\bullet||_2$ denotes the 2-norm of a vector.

Compared with (2), equation (5) approximates the matrix $A$ by $\lambda_1P_1P_1^T+...+\lambda_1P_pP_p^T$. Therefore, we can expect that minimizing the cost function $\Psi$ in (5) will converge $\lambda_i$ and $P_i$ to the dominant eigenvalues and eigenvectors of the original matrix $A$, respectively. Unfortunately, $\Psi$ in (5) is a sixth order polynomial and might not be convex. In addition, the constraint set in (5) is specified by a quadratic equation and is not convex either. Therefore, the optimization in (5) is not a convex programming problem and there is no efficient optimization algorithm that can guarantee finding the globally optimal solution for $\Psi$.

Instead of solving the non-convex optimization problem in (5), we propose a novel implicit power iteration method to efficiently extract the unknown coefficients $\lambda_i$ and $P_i$. In what follows, we first develop the implicit power iteration algorithm for rank-one approximation, and then extend it to rank-$p$ approximation.

### *A. Rank-One Approximation*

Fig. 2 outlines the implicit power iteration algorithm for a rank-one approximation. This algorithm repeatedly solves a sequence of over-determined linear equations until the convergence is identified. Next, we explain why the implicit power iteration yields the optimal rank-one approximation $A_L = \lambda_1 P_1 P_1^T$. Note that Step 4 in Fig. 2 approximates the matrix $A$ by $Q_k Q_{k-1}^T$, where $Q_{k-1}$ is determined in the previous iteration step. Finding such an optimal approximation is equivalent to solving the over-determined linear equation:

$$Q_k Q_{k-1}^T = A \tag{6}$$

The least-square-error solution for (6) is given by [8]:

$$Q_k = A Q_{k-1} \cdot \left(Q_{k-1}^T Q_{k-1}\right)^{-1} = A Q_{k-1} \tag{7}$$

In (7), $Q_{k-1} Q_{k-1}^T = \|Q_{k-1}\|_2^2 = 1$, since $Q_{k-1}$ is normalized in Step 3 of Fig. 2. Equation (7) reveals an interesting fact that solving the over-determined linear equation in Step 4 "implicitly" computes the matrix-vector product $AQ_{k-1}$, which is the basic operation required in the traditional power iteration for dominant eigenvector computation [8].

```
1. Start from a set of sampling points $\{X_i, \tilde{f}_i\}$.
2. Randomly select an initial vector $Q_0 \in R^N$ and set $k = 1$.
3. Compute $Q_{k-1} = Q_{k-1}/\|Q_{k-1}\|_2$.
4. Solve the over-determined linear equation for $Q_k$, $B_k$ and $C_k$:
       $X_i^T Q_k Q_{k-1}^T X_i + B_k^T X_i + C_k = \tilde{f}_i \quad (i = 1, 2, \cdots)$
5. If the residue:
       $\psi_k(Q_k, B_k, C_k) = \sum_i \left(X_i^T Q_k Q_{k-1}^T X_i + B_k^T X_i + C_k - \tilde{f}_i\right)^2$
   is unchanged, i.e.:
       $\left|\psi_k(Q_k, B_k, C_k) - \psi_{k-1}(Q_{k-1}, B_{k-1}, C_{k-1})\right| < \varepsilon$
   where $\varepsilon$ is the pre-defined error tolerance, then go to Step 6.
   Otherwise, $k = k+1$ and return Step 3.
6. The rank-one response surface model is:
       $f_1(X) = X^T Q_k Q_{k-1}^T X + B_k^T X + C_k$
```

Fig. 2. Implicit power iteration for a rank-one approximation.

Given an initial vector:

$$Q_0 = \alpha_1 P_1 + \alpha_2 P_2 + \cdots \tag{8}$$

where $Q_0$ is represented as the linear combination of all eigenvectors of $A$, the $k$-th iteration step yields:

$$Q_k = A^k Q_0 = \alpha_1 \lambda_1^k P_1 + \alpha_2 \lambda_2^k P_2 + \cdots \tag{9}$$

In (9), we ignore the normalization $Q_{k-1} = Q_{k-1}/\|Q_{k-1}\|_2$ which is nothing else but a scaling factor. This scaling factor will not change the direction of $Q_k$. As long as $\alpha_1 \neq 0$ in (8), i.e., $P_1$ is not orthogonal to the initial vector $Q_0$, $\alpha_1 \lambda_1^k P_1$ (with $|\lambda_1|>|\lambda_2|>\ldots$) will become more and more dominant over other terms. $Q_k$ will asymptotically approach the direction of $P_1$.

After the iteration in Fig. 2 converges, we have $Q_{k-1} = Q_{k-1}/\|Q_{k-1}\|_2 = P_1$ and $Q_k = AQ_{k-1} = \lambda_1 P_1$. $Q_k Q_{k-1}^T$ is the optimal rank-one approximation $A_L = \lambda_1 P_1 P_1^T$. Thus the proposed implicit power iteration extracts the unknown coefficients $\lambda_1$ and $P_1$ with guaranteed convergence, but in an implicit way (i.e., without knowing the full-rank matrix $A$). This "implicit" property is the key difference between the proposed algorithm and the traditional power iteration in [8].

The above discussion demonstrates that the implicit power iteration is provably convergent if $A$ is symmetric. For an asymmetric $A$, $Q_{k-1}$ and $Q_k$ should iteratively converge to the directions of the dominant left and right singular vectors of $A$ to achieve the optimal rank-one approximation. However, the global convergence of the implicit power iteration is difficult to prove in that case.

### B. *Rank-p Approximation*

Fig. 3 shows the implicit power iteration algorithm for a rank-$p$ approximation. Assuming that the unknown function can be approximated by the full-rank quadratic form in (2), the algorithm in Fig. 3 first extracts its rank-one approximation:

$$g_1(X) = X^T \left(\lambda_1 P_1 P_1^T\right) X + B^T X + C \tag{10}$$

Then, the component of $g_1(X)$ is subtracted from the full-rank quadratic function in Step 3 of Fig. 3, yielding:

$$f(X) - g_1(X) = X^T \left(\sum_{i=2}^{N} \lambda_i P_i P_i^T\right) X \tag{11}$$

Now, $\lambda_2$ and $P_2$ become the respective dominant eigenvalue and eigenvector of the quadratic function in (11), and they are extracted by the rank-one implicit power iteration to generate $g_2(X)$. The rank-one implicit power iteration and the subtraction are repeatedly applied for $p$ times until the rank-$p$ approximation $f_p(X)$ is achieved.

```
1. Start from a set of sampling points $\{X_i, \tilde{f}_i\}$.
   For $k = 1, 2, \ldots, p$
2.     Apply the implicit power iteration algorithm in Fig. 2 to
       compute the rank-one approximation $g_k(X)$.
3.     Update the sampling points:
           $\tilde{f}_i = \tilde{f}_i - g_k(X_i) \quad (i = 1, 2, \cdots)$
   End For
4. The rank-$p$ response surface model is:
       $f_p(X) = g_1(X) + \cdots + g_p(X)$
```

Fig. 3. Implicit power iteration for a rank-$p$ approximation.

The algorithm in Fig. 3 assumes a given approximation rank $p$. In practical applications, the value of $p$ can be iteratively determined based on the approximation error. For example, starting from a low-rank approximation, $p$ should be iteratively increased if the modeling error remains large.

The rank-$p$ implicit power iteration in Fig. 3 requires running the rank-one implicit power iteration for $p$ times. Each rank-one approximation needs to solve $2N$+1 unknown coefficients, for which the required number of samples is of the order of $O(N)$, and solving the over-determined linear equation in Step 4 of Fig. 2 has a complexity of $O(N^3)$. Therefore, a rank-$p$ approximation requires $O(pN)$ simulation samples in total and the overall computation cost for the rank-$p$ implicit power iteration in Fig. 3 is $O(pN^3)$. In many practical applications, $p$ is much less than $N$ and, therefore, PROBE is much more efficient than the traditional full-rank quadratic modeling which requires $O(N^2)$ simulation samplings and has a fitting cost of $O(N^6)$ for solving the over-determined linear equation.

## 3.3 Comparison with Traditional Techniques

There are several traditional techniques, such as principal component analysis [9], variable screening [10] or projection pursuit [11], which aim to reduce the computation cost of response surface modeling. In this subsection, we compare PROBE with these traditional techniques and highlight their differences.

Principal component analysis (PCA) [9] is a statistical method for reducing the number of random variables that are required to represent the process variations. Given $N$ normally distributed process parameters $X = [x_1, x_2, \ldots, x_N]^T$ and their correlation matrix $R$, PCA computes the dominant eigenvalues and eigenvectors of

$R$, and then constructs a set of new random variables $Y = [y_1, y_2, ..., y_M]^T$, where $M < N$, to approximate the original $N$-dimensional random space. The essence of PCA can be interpreted as the coordinate rotation of the original random space $X$ followed by a low-rank projection onto the low-dimensional space $Y$. The new random variables $y_i$ are called the principal components or factors. After PCA, the circuit performances can be approximated as functions of the new random variables $y_i$ using response surface modeling. Since the number of new variables $y_i$ is less than the number of original variables $x_i$, PCA reduces the number of unknown model coefficients.

Such a PCA approach, however, is substantially different from our proposed PROBE method. The PCA operation is completely determined by the statistical characteristics, i.e., the correlation matrix $R$, of random process variations, without depending on a specific circuit performance $f$. In contrast, PROBE reduces the modeling cost by carefully analyzing a specific performance $f$. In other words, PROBE will eliminate (or keep) one eigenvector $P_i$ if $f$ is strongly (or weakly) dependent on $P_i$. Therefore, PCA and PROBE rely on completely different mechanisms to minimize the computation cost. In practical applications, both PCA and PROBE should be simultaneously applied to achieve the minimal modeling cost, as shown in Fig. 4.

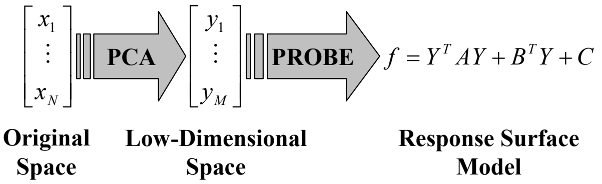

$\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$
PCA
$\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}$
PROBE
$f = Y^T A Y + B^T Y + C$
Original Space
Low-Dimensional Space
Response Surface Model

Fig. 4. Combination of PCA and PROBE to reduce cost.

Variable screening is another traditional approach for reducing the response surface modeling cost [10]. Given a circuit performance $f$, variable screening applies fractional factorial experimental design and tries to identify a subset (hopefully small) of the random process parameters that have much greater influence on $f$ than the others. Compared with variable screening, PROBE also do a similar "variable screening", but with an additional coordinate rotation, as shown in Fig. 5. The additional coordinate rotation offers more flexibility in filtering out insignificant components, thereby achieving better modeling accuracy and/or cheaper modeling cost. From this point of view, the proposed PROBE can be considered as a *generalized variable screening* which is an extension of the traditional variable screening in [10].

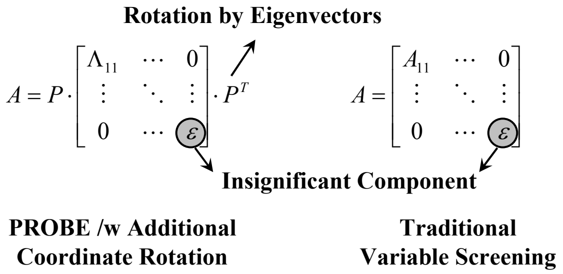

Rotation by Eigenvectors
$A = P \cdot \begin{bmatrix} \Lambda_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varepsilon \end{bmatrix} \cdot P^T$
$A = \begin{bmatrix} A_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varepsilon \end{bmatrix}$
Insignificant Component
PROBE /w Additional Coordinate Rotation
Traditional Variable Screening

Fig. 5. Comparison of PROBE with variable screening.

Projection pursuit [11] tries to approximate the unknown high-dimensional nonlinear function by the sum of several smooth low-dimensional functions. The authors in [11] utilize the one-dimensional projection:

$$f(X) = g_1\left(P_1^T X\right) + g_2\left(P_2^T X\right) + \cdots \quad (12)$$

where $g_i(\bullet)$ is the pre-defined one-dimensional nonlinear function and $P_i \in R^N$ defines the projection space. One of the main difficulties in traditional projection pursuit is to find the optimal projection vectors $P_i$. The authors in [11] apply local optimization with heuristics to search for the optimal $P_i$. Such an optimization can easily get stuck at a local minimum. Our proposed PROBE algorithm is actually a special case of the traditional projection pursuit, where all $g_i(\bullet)$ are quadratic functions. In such cases, the theoretical solution of the optimal projection vectors $P_i$ is known, i.e., they are determined by the dominant eigenvalues and eigenvectors of the original full-rank matrix $A$. These dominant eigenvalues and eigenvectors can be extracted by the proposed implicit power iteration algorithm quickly and robustly. Such a special advantage of using the quadratic $g_i(\bullet)$, however, has not been explored in traditional projection pursuit.

### 3.4 Application of PROBE Models

The low-rank quadratic models extracted by PROBE can be generally applied to any applications that require quadratic response surface modeling, such as [3]-[7]. In addition to these general applications, we emphasize a special link between our PROBE modeling and the APEX algorithm proposed in [7]. In APEX, the most expensive computation is the binomial moment evaluation, which requires diagonalizing the quadratic coefficient matrix $A$ by eigen-decomposition. Using our PROBE modeling, however, the matrix $A$ is approximated by a low-rank one $A_L$. The eigen-decomposition of the low-rank matrix $A_L$ is much cheaper than finding the eigenvalues/eigenvectors of the full-rank matrix $A$. Therefore, the complexity of the APEX algorithm can be significantly reduced if using the PROBE model as its input. The detailed implementation for combining PROBE and APEX is beyond the scope of this paper and, therefore, is not discussed in detail.

## 4. Numerical Examples

In this section we demonstrate the computational efficiency of PROBE using several circuit examples. For each example, two independent sampling sets, called training set and testing set respectively, are generated. The training set is created by Latin hypercube sampling [12], which picks the most important samples based on statistical analysis; this is used for coefficient fitting. For testing and comparison, we collect 500 random samples as the testing set and use them to measure the modeling error. All numerical experiments are performed on a SUN — 1GHz server.

### 4.1 ISCAS'89 S27

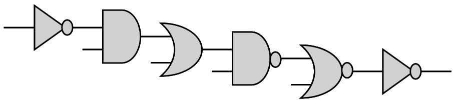


Fig. 6. Longest path in ISCAS'89 S27.

We create a physical implementation for the ISCAS'89 S27 benchmark circuit using the ST CMOS 90 nm process. Given a set of fixed gate sizes, the longest path delay in the benchmark circuit (shown in Fig. 6) is a function of the process variations (e.g., $\Delta V_{TH}$, $\Delta T_{OX}$, $\Delta L$, etc.). Since the circuit only consists of six gates which can be put close to each other in the layout, inter-die variations will dominate over intra-die variations, and gate delays will dominate over (local) interconnect delays in this example. Therefore, for simplicity, we only consider inter-die variations for CMOS transistors in this example. The probability distributions and the correlation information of the inter-die transistor variations are obtained from the ST Microelectronics design kit. After PCA analysis, 6 principal random factors are identified to represent these process variations. We should note, however, that nothing precludes us from including more detailed intra-die and/or interconnect variation models in PROBE as well.

### *A. Robust Convergence of Implicit Power Iteration*

In order to test the convergence of the proposed implicit power iteration algorithm, we pick 100 random initial vectors $Q_0$ and use them for running power iteration in coefficient fitting. We find that all 100 experiments reliably converge without a single failure.

### *B. Modeling Accuracy*

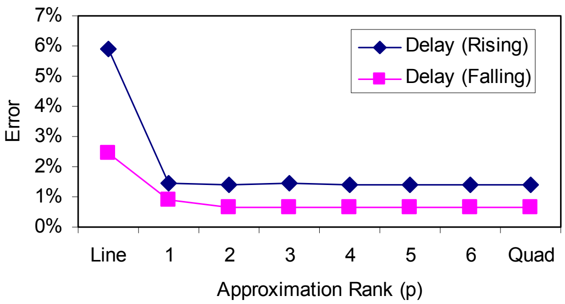


Fig. 7. Response surface modeling error for path delay.

Fig. 7 shows the response surface modeling error when the path delays of both rising and falling transitions for the circuit are approximated by the linear, rank-*p* quadratic (by PROBE) and traditional full-rank quadratic models. All response surface models are fitted using 578 training samples. It is shown in Fig. 7 that as *p* increase, the rank-*p* modeling error asymptotically approaches the full-rank quadratic modeling error. However, after $p > 2$, further increases in *p* do not have a significant impact on reducing error. It, in turn, implies that a rank-2 model, instead of the full-rank quadratic model with rank 6, is sufficiently accurate in this example.

## 4.2 Low Noise Amplifier

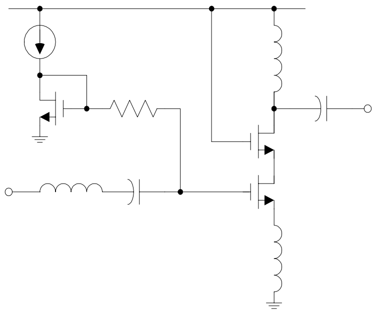
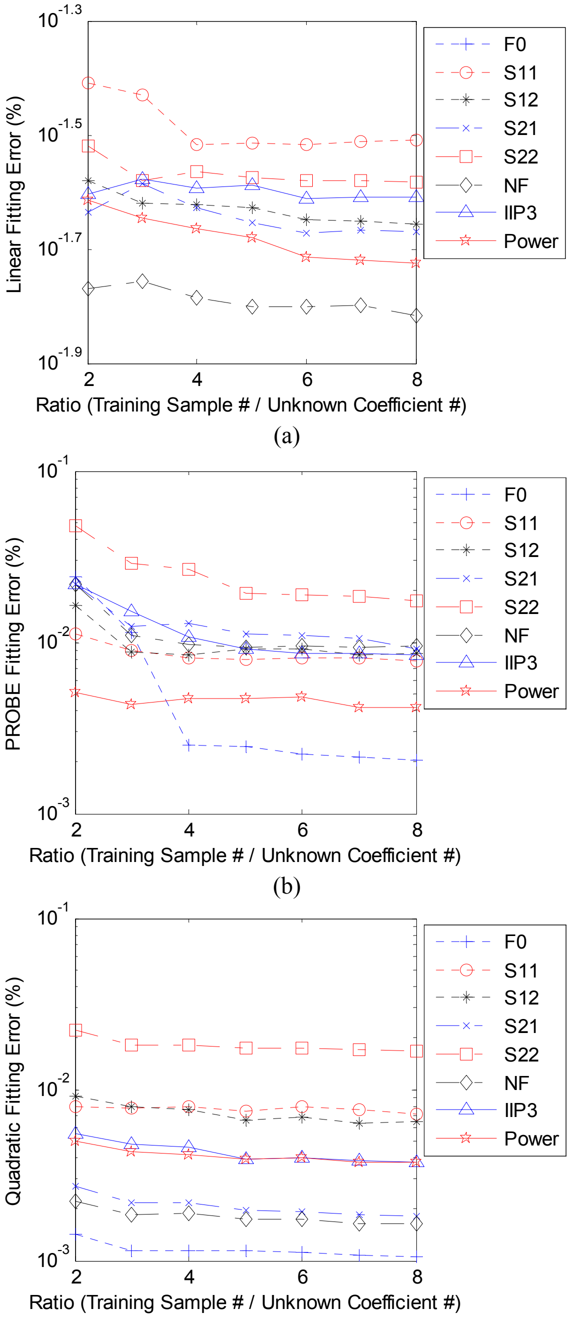


Fig. 8. Circuit schematic for LNA.

As a second example, we consider a low noise amplifier designed in the IBM BiCMOS 0.25 µm process, as shown in Fig. 8. In this example, the variations on both MOS transistors and passive components (resistors, capacitors and inductors) are considered. The probability distributions and the correlation information of these variations are provided in the IBM design kit. After PCA analysis, 8 principal factors are identified to represent the process variations.

### *A. Effect of Training Set Size*

Fig. 9 shows the relation between the modeling error and the training set size for three modeling approaches. From Fig. 9 we observe that the number of training samples should be around 3~4 times greater than the number of unknown coefficients to avoid over-fitting. Further increasing the size of training set does not have a significant impact on reducing fitting error. This observation implies that the required number of training samples depends on the number of unknown coefficients. As the unknown coefficient number is reduced in PROBE, we not only decrease the computation time for coefficient fitting, but also save a large portion of circuit simulation cost because of the smaller training set.



(a)

(b)

(c)

Fig. 9. Effect of the training set size for LNA. (a) Linear fitting error. (b) Rank-one PROBE fitting error. (c) Full-rank quadratic fitting error.

### B. *Modeling Accuracy and Cost*

Table 1 compares the fitting errors for the linear, rank-one PROBE and full-rank quadratic models. As we would expect, the rank-one PROBE modeling error is smaller than the linear modeling error, but larger than the full-rank quadratic modeling error.

Table 2 compares the response surface modeling cost for these three modeling approaches. The training set size in Table 2 is selected to be sufficiently large to avoid over-fitting. Since the rank-one PROBE model contains substantially fewer unknown coefficients and, therefore, requires much less training samples than the full-rank quadratic model, PROBE achieves 2.6x speedup in simulation cost due to the smaller training set. Compared with the simulation cost, the fitting cost is almost neglectable in this example, since the problem size is small and solving the over-determined linear equations only takes a few seconds for all performance metrics.

Table 1. Response surface modeling error for LNA

| Performance | Linear | PROBE (Rank-1) | Quad (Rank-8) |
|---|---|---|---|
| F0 | 1.04% | 0.25% | 0.11% |
| S11 | 3.04% | 0.81% | 0.79% |
| S12 | 2.39% | 0.84% | 0.77% |
| S21 | 2.35% | 1.28% | 0.22% |
| S22 | 2.72% | 2.68% | 1.80% |
| NF | 1.64% | 0.97% | 0.19% |
| IIP3 | 2.55% | 1.07% | 0.46% |
| Power | 2.16% | 0.47% | 0.41% |

Table 2. Response surface modeling cost for LNA

| Performance | Linear | PROBE (Rank-1) | Quad (Rank-8) |
|---|---|---|---|
| Unknown Coeff # | 9 | 17 | 45 |
| Training Sample # | 36 | 68 | 180 |
| Simulation Cost (Sec.) | 2620 | 4949 | 13100 |

Table 1 and Table 2 reveal an important fact that PROBE can easily facilitate the tradeoff between accuracy and cost during response surface modeling. Traditionally, if the linear model cannot provide sufficient accuracy, the full-rank quadratic model is immediately utilized which might provide over-accurate results and require expensive modeling cost. PROBE, however, offers an intermediate step between linear modeling and full-rank quadratic modeling. Depending on the modeling accuracy requirement, PROBE can iteratively select a correct $p$ value and create a rank-$p$ model. In this example, the rank-one PROBE model already provides sufficient accuracy, as shown in Table 1.

## 4.3 Scaling with Problem Size

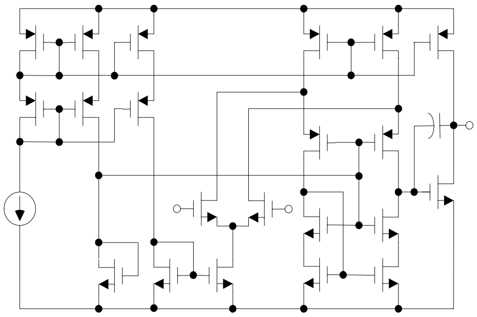


Fig. 10. Circuit schematic of a two-stage Op Amp.

Next, we consider a two-stage folded-cascode operational amplifier designed in the IBM BiCMOS 0.25 μm process, as shown in Fig. 10. In this example, 49 principal random factors are extracted by PCA analysis to represent the process variations, including both inter-die variations and device mismatches. The probability distributions and the correlation information of these random variations are obtained from the IBM design kit.

Due to the inclusion of mismatches, the problem size becomes significantly larger in this example. However, modeling mismatches is extremely important for the Op Amp in Fig. 10, since the device mismatches can significantly impact the performance of the input differential pair.

Table 3. Response surface modeling error for Op Amp

| Performance | Linear | PROBE (Rank-1) | Quad (Rank-49) |
|---|---|---|---|
| Gain | 4.20% | 2.00% | 1.74% |
| Offset | 24.83% | 10.28% | 9.09% |
| UGF | 1.23% | 0.48% | 0.48% |
| Gain Margin | 1.03% | 0.55% | 0.55% |
| Phase Margin | 1.20% | 0.44% | 0.44% |
| Slew Rate (+) | 0.92% | 0.93% | 0.70% |
| Slew Rate (–) | 1.38% | 0.53% | 0.48% |
| Power | 1.05% | 0.77% | 0.68% |

Table 4. Response surface modeling cost for Op Amp

| Performance | Linear | PROBE (Rank-1) | Quad (Rank-49) |
|---|---|---|---|
| Unknown Coeff # | 50 | 99 | 1275 |
| Training Sample # | 200 | 396 | 5100 |
| Simulation Cost (Sec.) | $7.88 \times 10^3$ | $1.56 \times 10^4$ | $2.01 \times 10^5$ |
| Fitting Cost (Sec.) | 12.68 | 54.13 | 5192.06 |

Table 3 compares the response surface modeling errors for three different approaches: linear approximation, rank-one approximation by PROBE and traditional full-rank approximation. As we would expect, the Op Amp offset is strongly nonlinear in device mismatches. Therefore, the simple linear approximation yields an extremely large error (i.e., 24.83%) as shown in Table 3. Compared with the linear modeling, both the rank-one PROBE modeling and the full-rank quadratic modeling achieve more than 2x error reduction. Although higher-order (e.g., cubic) response surface models can be applied to further improve the accuracy, these higher-order models are rarely utilized in practical applications as they will inevitably lead to an unaffordable computation cost.

Table 4 shows the response surface modeling cost for these three approaches. The training set size in Table 4 is selected to be sufficiently large to avoid over-fitting. As shown in Table 4, while the full-rank quadratic modeling takes more than *2 days* to generate all training samples, PROBE reduces the simulation cost to 4.3 hours (12x smaller). In addition, 96x additional speedup is achieved by PROBE for coefficient fitting (i.e., solving the unknown model coefficients) compared with the full-rank quadratic modeling, although the fitting cost is not the dominant one in this example.

# 5. Conclusions

We propose a novel projection-based extraction approach, PROBE, for quadratic response surface modeling of circuit performances with consideration of both inter-die and intra-die process variations. PROBE utilizes a new projection scheme to facilitate the tradeoff between modeling accuracy and cost. In addition, a novel implicit power iteration algorithm is developed to find the optimal projection space and solve the unknown model coefficients. By using the proposed implicit power iteration algorithm, PROBE significantly reduces the modeling cost (i.e., both the required number of samples and the linear equation size), thereby facilitating scaling to much larger problem sizes. As demonstrated by numerical examples in this paper, PROBE can generate accurate response surface models and achieve up to 12x speedup compared with the traditional quadratic modeling approach. The response surface models generated by PROBE can be incorporated into a statistical analysis/optimization environment for accurate and efficient yield analysis/optimization.

## 6. Acknowledgements

This work was funded in part by the MARCO Focus Center for Circuit & System Solutions (C2S2, www.c2s2.org) under contract 2003-CT-888.

## 7. References

[1] S. Nassif, “Modeling and analysis of manufacturing variations,” *IEEE CICC*, pp. 223-228, 2001.

[2] M. Orshansky; L. Milor and C. Hu, “Characterization of spatial intrafield gate CD variability, its impact on circuit performance, and spatial mask-level correction,” *IEEE Trans. Semiconductor Manufacturing*, vol. 17, no. 1, pp. 2-11, Feb. 2004.

[3] R. Myers and D. Montgomery, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, Wiley-Interscience, 2002.

[4] C. Michael and M. Ismail, “Statistical modeling of device mismatch for analog MOS integrated circuits,” *IEEE Journal of Solid-State Circuits*, vol. 27, no. 2, pp. 154-166, Feb. 1992.

[5] Z. Wang and S. Director, “An efficient yield optimization method using a two step linear approximation of circuit performance,” *IEEE EDAC*, pp. 567-571, 1994.

[6] A. Dharchoudhury and S. Kang, “Worse-case analysis and optimization of VLSI circuit performance,” *IEEE Trans. CAD*, vol. 14, no. 4, pp. 481-492, Apr. 1995.

[7] X. Li, J. Le, P. Gopalakrishnan and L. Pileggi, “Asymptotic probability extraction for non-Normal distributions of circuit performance,” *IEEE ICCAD*, pp. 2-9, 2004.

[8] G. Golub and C. Loan, *Matrix Computations*, The Johns Hopkins Univ. Press, 1996.

[9] G. Seber, *Multivariate Observations*, Wiley Series, 1984.

[10] K. Low and S. Director, “An efficient methodology for building macromodels of IC fabrication processes,” *IEEE Trans. CAD*, vol. 8, no. 12, pp. 1299-1313, Dec. 1989.

[11] J. Friedman and W. Stuetzle, “Projection pursuit regression,” *Journal of the American Statistical Association*, vol. 76, no. 376, pp. 817-823, 1981.

[12] M. Mckay, R. Beckman and W. Conover, “A comparison of three methods for selecting values of input variables in the analysis of output from a computer code,” *Technometrics*, vol. 21, no. 2, pp. 239-245, May. 1979.