Projection-Based Piecewise-Linear Response Surface Modeling for Strongly Nonlinear VLSI Performance Variations

Xin Li
Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA 15213
xinli@ece.cmu.edu

Yu Cao
Department of Electrical Engineering
Arizona State University, Tempe, AZ 85287
ycao@asu.edu

ABSTRACT

Large-scale process fluctuations (particularly random device mismatches) at nanoscale technologies bring about high-dimensional strongly nonlinear performance variations that cannot be accurately captured by linear or quadratic response surface models. In this paper, we propose a novel projection-based piecewise linear modeling technique, P2M, to address such a modeling challenge with affordable computational cost. P2M borrows the projection pursuit idea from mathematics to convert a high-dimensional modeling problem to a low-dimensional one. In addition, a new piecewise-linear model template is proposed and tuned for strongly nonlinear performance variations. By exploiting the unique piecewise-linear nature of the model template, a robust numerical algorithm is further developed to determine all model coefficients by solving a sequence of over-determined linear equations. Several circuit examples designed in a commercial 65nm CMOS process demonstrate that compared with the traditional quadratic modeling, P2M achieves 2x error reduction with negligible computational overhead.

1. INTRODUCTION

As IC technologies are scaled to nanoscale regime, it becomes increasingly difficult to control the variations in manufacturing process [1]-[2]. Process variations can be classified into two broad categories: inter-die variations and intra-die variations. Inter-die variations model the common/average variations across the die, while intra-die variations model the individual, but spatially correlated, local variations (e.g., random device mismatches) within the same die. Both inter-die and intra-die variations introduce substantial uncertainties in circuit performance and significantly impact parametric yield. Hence, accurately modeling and analyzing process variations to ensure manufacturability has been identified as a top priority for today's IC design.

To address this issue, response surface modeling has been widely applied to solve various statistical circuit analysis problems [1], [3]-[7], [15]. The objective of response surface modeling is to approximate the circuit performance (e.g., delay, gain) as an analytical function of process parameters (e.g., V_{TH}, T_{OX}). Most existing response surface models are either linear or quadratic, assuming that the approximated performance functions are weakly nonlinear. However, two recent changes in advanced IC technologies suggest a need to revisit this assumption.

Firstly, among all sources of variations, random mismatches become dominant at 45nm process and beyond [2]. As a result, any two transistors on the same die can have significantly different electrical performance (e.g., mobility, V_{TH}, etc.). To accurately model this effect, a large number of random variables must be utilized, rendering a *high-dimensional* variation space. Even for a small-size circuit block (e.g., an analog amplifier), the total number of random process parameters can easily reach 50~100 [6], [8].

Secondly, process variations become relatively larger, as IC technologies are scaled to finer feature size. For example, the 3-sigma V_{TH} variation is expected to reach 35% in 2008 and it continuously increases in future technology generations [2]. Such large-scale variations yield *strongly nonlinear* performance variations that cannot be accurately captured by linear or quadratic models. This nonlinearity issue is especially critical for analog and mixed-signal circuits. As will be demonstrated by the numerical examples in Section 5, the error of a quadratic model can reach 13.6% for a commercial 65nm SRAM cell.

To improve accuracy, high-order (e.g., cubic) polynomial models can be used. Directly applying existing response surface modeling techniques to high order, however, results in expensive computational cost. For example, if the total number of random process parameters reaches 100 [6], [8], a cubic polynomial will contain 176,851 unknown model coefficients!

The authors of [8] propose a new projection-based nonlinear modeling technique based on neural network. While it has been successfully applied to various circuit problems, a neural network must be trained by nonlinear optimization where global convergence is difficult to achieve. In other words, the quality of the extracted model heavily depends on the initial guess that is provided to the optimizer. The challenging problem here is how to solve such a high-dimensional strongly nonlinear modeling problem both *robustly* and *efficiently*.

In this paper, we propose a novel projection-based piecewise-linear modeling (P2M) algorithm that is especially tuned for high-dimensional strongly nonlinear fitting problems. P2M borrows the projection pursuit idea that was initially developed by mathematicians in 1980s [9]. It converts a high-dimensional modeling problem to a low-dimensional problem that is easy to solve. In addition to dimension reduction, we propose to capture strongly nonlinear performance variations by piecewise-linear functions. Compared with the traditional quadratic response surface modeling, the proposed P2M approach reduces modeling error by 2x without substantially increasing computational cost, as will be demonstrated by the numerical examples in Section 5.

An important contribution of this paper is to propose a robust numerical algorithm to determine all unknown model coefficients. Our proposed algorithm recursively approximates a high-dimensional performance function by a number of one-dimensional piecewise-linear functions. Furthermore, by exploiting the unique piecewise-linear nature of the model template, P2M formulates the modeling problem in a special form for which all model coefficients can be solved from a sequence of over-determined linear equations. Therefore, unlike the existing optimization-based techniques that suffer from several numerical issues such as local convergence, our proposed P2M approach offers robust convergence with low computational cost.

The remainder of this paper is organized as follows. In Section 2, we review the background on response surface modeling and projection pursuit. Then, we propose our P2M approach in Section 3 and describe the numerical algorithms in Section 4. The efficacy of P2M is demonstrated by several

0-7695-3117-2/08 $25.00 © 2008 IEEE
DOI 10.1109/ISQED.2008.28

numerical examples in Section 5, followed by the conclusions in Section 6.

2. BACKGROUND

2.1 Response Surface Modeling

Given a circuit design, the circuit performance (e.g., delay, gain) is a function of process parameters (e.g., V_{TH}, T_{OX}). These process parameters must be modeled as random variables to account for uncertain manufacturing fluctuations. A circuit performance f can be approximated as a linear response surface model of process parameters [1], [15]:

$$f(X) = B^T X + C \tag{1}$$

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

The linear approximation in (1) is efficient and accurate when process variations are sufficiently small. As manufacturing variations become relatively large in nanoscale technologies, quadratic response surface models are required to improve modeling accuracy [4], [6], [15]:

$$f(X) = X^T AX + B^T X + C \tag{2}$$

where $C \in R$ is the constant term, $B \in R^N$ contains the linear coefficients, and $A \in R^{N\times N}$ contains the quadratic coefficients.

The unknown model coefficients in (1) and (2) can be determined by solving the over-determined linear equations at a number of sampling points [15]:

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

$$X_i^T AX_i + B^T X_i + C = \tilde{f}_i \quad (i = 1,2,\cdots,S) \tag{4}$$

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, and S is the total number of sampling points.

Even if quadratic response surface models are utilized, however, large modeling error can still be observed in some cases [8]. For instance, as will be demonstrated by the numerical examples in Section 5, the quadratic modeling error can reach 13.6% for an SRAM cell designed in a commercial 65nm CMOS process. It, in turn, motivates us to develop a new piecewise-linear modeling technique to accurately capture strongly nonlinear performance variations.

2.2 Projection Pursuit

One major technical difficulty of fitting high-dimensional nonlinear response surface models stems from the large number of unknown model coefficients. To address this issue, projection pursuit was proposed by mathematicians in 1980s [9] and it has been recently applied to several circuit modeling problems [6]-[8]. The key idea of projection pursuit is to approximate a high-dimensional nonlinear function by the sum of several low-dimensional functions. In particular, a one-dimensional projection has the form of [9]:

$$f(X) = g_1(P_1^T X) + g_2(P_2^T X) + \cdots + g_K(P_K^T X) \tag{5}$$

where $f(X)$ is the approximated high-dimensional nonlinear function, $\{g_i(\bullet);\ i = 1,2,...,K\}$ contains K one-dimensional nonlinear functions, $\{P_i \in R^N;\ i = 1,2,...,K\}$ defines K one-dimensional projection vectors, and K is referred to as the rank of the model.

PROBE was developed in [6] to handle the special case where all nonlinear functions $\{g_i(\bullet);\ i = 1,2,...,K\}$ in (5) are quadratic. A quadratic function defined in (2) can be re-written as [6]:

$$f(X) = \sum_{i=1}^{N} \lambda_i \cdot (Q_i^T X)^2 + B^T X + C \tag{6}$$

where λ_i and $Q_i \in R^N$ are the i-th dominant eigenvalue and eigenvector of the quadratic coefficient matrix A, respectively. In this case, the optimal projection vectors are determined by the eigenvectors and they can be extracted by the implicit power iteration algorithm proposed in [6].

The idea of projection pursuit has been further applied to strongly nonlinear circuit problems in [8]. The SiLVR algorithm developed in [8] determines the optimal projection vectors by nonlinear optimization. Such an optimization-based approach, however, suffers from several numerical issues such as local convergence. In this paper, we propose a new projection-based piecewise linear modeling algorithm, P2M, that aims to robustly solve the high-dimensional nonlinear modeling problem. By exploiting the unique piecewise-linear nature of the model template, P2M determines all unknown model coefficients by solving a sequence of over-determined linear equations, thereby offering low computational complexity and robust convergence.

3. PIECEWISE-LINEAR MODELING

Our proposed P2M approach utilizes one-dimensional piecewise-linear functions to approximate $\{g_i(\bullet);\ i = 1,2,...,K\}$ in (5). Such a piecewise-linear model template allows us to approximate strongly nonlinear performance functions both accurately and efficiently, which is one of the major advantages of the P2M method.

A one-dimensional M-segment piecewise-linear function $g_i(P_i^T X)$ is uniquely specified by the projection vector P_i and the M+1 grid points $\{(\alpha_{i,j}, \beta_{i,j}); j = 0,1,...,M\}$:

$$g_i(P_i^T X) = \beta_{i,j-1} + \frac{\beta_{i,j} - \beta_{i,j-1}}{\alpha_{i,j} - \alpha_{i,j-1}} \cdot (P_i^T X - \alpha_{i,j-1}). \tag{7}$$

where

$$\alpha_{i,j-1} \le P_i^T X \le \alpha_{i,j}\ . \tag{8}$$

In other words, the value of $g_i(P_i^T X)$ is determined by the linear interpolation of the M+1 grid points. Fig 1 shows a simple piecewise-linear function example with four segments.

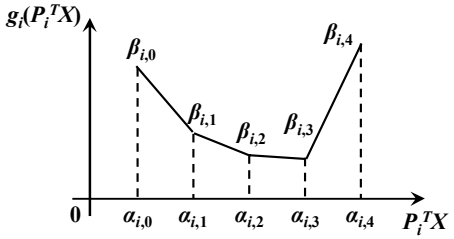

Fig 1. A one-dimensional 4-segment piecewise-linear function.

Given the one-dimensional M-segment piecewise linear functions $\{g_i(\bullet);\ i = 1,2,...,K\}$, the nonlinear model $f(X)$ in (5) is the sum of all $\{g_i(\bullet);\ i = 1,2,...,K\}$. Theoretically, this is equivalent to partitioning the variation space X into M^K polytopes:

$$\begin{array}{ll} \alpha_{1,j_1-1} \le P_1^T X \le \alpha_{1,j_1} & (j_1 = 1,2,\cdots,M) \\ \alpha_{2,j_2-1} \le P_2^T X \le \alpha_{2,j_2} & (j_2 = 1,2,\cdots,M) \\ \vdots & \vdots \\ \alpha_{K,j_K-1} \le P_K^T X \le \alpha_{K,j_K} & (j_K = 1,2,\cdots,M) \end{array} \tag{9}$$

and approximating $f(X)$ as a linear function within each polytope.

Fig 2 shows a two-dimensional space that is partitioned by the projection vectors $P_1 = [1\ 0]^T$ and $P_2 = [1\ 1]^T$. Note that the polytopes in Fig 2 are not rectangular, because P_1 and P_2 are not orthogonal.

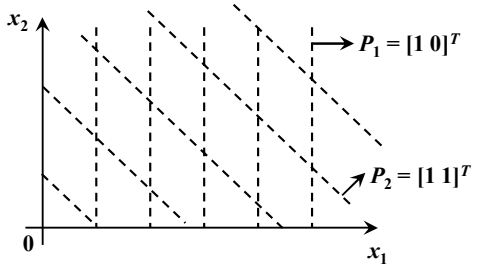

Fig 2. Partitions of a two-dimensional space $[x_1\ x_2]^T$.

The proposed piecewise-linear model $f(X)$ defined in (5), (7)-(8) has two important properties:

- *Continuous*: Since all $\{g_i(\bullet);\ i = 1,2,...,K\}$ are continuous, it is straightforward to verify that $f(X) = g_1(\bullet) + g_2(\bullet) + ... + g_K(\bullet)$ is also continuous.
- *Low-rank*: For many high-dimensional problems, the rank K is substantially smaller than the variation space dimension N. In other words, the performance of interest is only affected by several "dominant" directions of process variations and the variations on other "non-dominant" directions can be ignored [6]-[8]. This rank-deficient property allows us to accurately extract a compact low-rank model with low computational cost.

To determine a rank-K piecewise-linear model $f(X)$, we must determine the projection vectors $\{P_i;\ i = 1,2,...,K\}$ in (5) and the grid points $\{(\alpha_{i,j}, \beta_{i,j});\ i = 0,1,...,K, j = 0,1,...,M\}$ in (7)-(8). In what follows, we propose an efficient numerical algorithm to solve these unknown model coefficients.

4. IMPLEMENTATION OF P2M

The proposed P2M algorithm is facilitated by two key techniques, including: (1) a nonlinear sensitivity analysis to determine the projection vectors; and (2) an iterative algorithm to decompose a rank-K modeling problem into multiple rank-one problems. In this section, we first develop the numerical algorithm for rank-one P2M approximation, and then extend it to rank-K approximation.

4.1 Rank-One Approximation

Given a rank-one M-segment P2M model, the unknown model coefficients include the projection vector P_1 and the M+1 grid points $\{(\alpha_{1,j}, \beta_{1,j});\ j = 0,1,...,M\}$. Ideally, P_1 and $\{(\alpha_{1,j}, \beta_{1,j});\ j = 0,1,...,M\}$ should be optimized concurrently to extract the optimal model. However, such a co-optimization can be computationally expensive and does not guarantee global convergence. For this reason, we propose a heuristic algorithm to decompose the modeling process into two separate steps. We first apply a nonlinear sensitivity analysis to find the projection vector P_1 and then perform a least-square fitting to determine the grid points $\{(\alpha_{1,j}, \beta_{1,j});\ j = 0,1,...,M\}$. Our numerical examples in Section 5 demonstrate that the proposed heuristic algorithm yields excellent results for most circuit problems.

To determine the projection vector P_1, we need to find out the dominant direction along which the performance function $f(X)$ varies significantly. If $f(X)$ is linear, the projection vector P_1 is determined by linear sensitivities and it can be easily found by fitting a linear response surface model. However, such a linear-sensitivity-based approach does not work well if $f(X)$ is strongly nonlinear. Motivated by this observation, we propose to borrow the PROBE algorithm [6] to fit a rank-one quadratic model from which the projection directions for both linear and quadratic terms are determined. Such an approach is referred to as *nonlinear sensitivity analysis* in this paper.

A rank-one PROBE model contains the constant term, the linear term, and the first dominant quadratic term [6]:

$$f(X) = \lambda_1 \cdot \left(Q_1^T X\right)^2 + B^T X + C\ . \tag{10}$$

From (10), we obtain two projection directions: B for the linear term and Q_1 for the quadratic term. For rank-one P2M approximation, we need to select one of them as the dominant projection vector P_1.

Toward this goal, we define the following expected "energies" and use them as a criterion to compare the significance of the linear and quadratic terms:

$$Energy_{Linear} = E\left[\left(B^T X\right)^2\right] = E\left[f_L^2\right] \tag{11}$$

$$Energy_{Quadratic} = E\left[\lambda_1^2 \cdot \left(Q_1^T X\right)^4\right] = E\left[\lambda_1^2 \cdot f_Q^4\right] \tag{12}$$

where $E(\bullet)$ denotes the expected value [14] and

$$f_L = B^T X \tag{13}$$

$$f_Q = Q_1^T X\ . \tag{14}$$

To compute the statistical measures in (11)-(12), we assume that all random variables $X = [x_1\ x_2\ ...\ x_N]^T$ are mutually independent and standard Normal (i.e., zero mean and unit variance). If $\{x_i;\ i = 1,2,...,N\}$ are correlated Normal distributions, they can be converted into independent Normal distributions by principal component analysis (PCA) [12].

Given the definitions in (13)-(14), both f_L and f_Q are Normal, since they are the linear combinations of multiple Normal distributions. In addition, the first-order and second-order moments of f_L and f_Q can be determined by [12]:

$$E\left[f_L\right] = 0 \quad and \quad E\left[f_L^2\right] = \left\|B\right\|_2^2 \tag{15}$$

$$E\left[f_Q\right] = 0 \quad and \quad E\left[f_Q^2\right] = \left\|Q_1\right\|_2^2 \tag{16}$$

where $\|\bullet\|_2$ denotes the 2-norm of a vector. Substituting (15) into (11) yields the expected energy for the linear term B^TX:

$$Energy_{Linear} = E\left[f_L^2\right] = \left\|B\right\|_2^2\ . \tag{17}$$

The expected energy for the quadratic term $\lambda_1 \cdot (Q_1^T X)^2$ is determined by the eigenvalue λ_1 and the fourth order moment of f_Q [14]:

$$Energy_{Quadratic} = E\left[\lambda_1^2 \cdot f_Q^4\right] = 3\lambda_1^2 \cdot \left\|Q_1\right\|_2^4\ . \tag{18}$$

Based on (17) and (18), we make P_1 equal the linear projection direction B (or the quadratic projection direction Q_1) if the expected linear energy $Energy_{Linear}$ is greater (or smaller) than the expected quadratic energy $Energy_{Quadratic}$. This heuristic rule is summarized by the following equation:

$$P_1 = \begin{cases} B & \left(if \quad \left\|B\right\|_2^2 \geq 3\lambda_1^2 \cdot \left\|Q_1\right\|_2^4\right) \\ Q_1 & \left(if \quad \left\|B\right\|_2^2 < 3\lambda_1^2 \cdot \left\|Q_1\right\|_2^4\right) \end{cases}. \tag{19}$$

Next, given the projection vector P_1, we need to determine the grid points $\{(\alpha_{1,j}, \beta_{1,j});\ j = 0,1,...,M\}$. We evenly partition the axis $P_1^T X$ into M segments, resulting in M+1 equally-spaced grid

points $\{\alpha_{1,j}; j = 0,1,...,M\}$, as shown in Fig 1. Then, using a set of sampling points, we list the following linear equations for $\{\beta_{1,j}; j = 0,1,...,M\}$:

$$\beta_{1,j-1} + \frac{\beta_{1,j} - \beta_{1,j-1}}{\alpha_{1,j} - \alpha_{1,j-1}} \cdot \left(P_1^T X_i - \alpha_{1,j-1}\right) = \tilde{f}_i \quad \left(\alpha_{1,j-1} \le P_1^T X_i < \alpha_{1,j} \quad i = 1,2,\cdots,S\right) \tag{20}$$

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, and S is the total number of sampling points. Solving the over-determined linear equations in (20) yields the optimal values of $\{\beta_{1,j}; j = 0,1,...,M\}$. It, in turn, determines the approximated one-dimensional piecewise-linear model.

Algorithm 1: Rank-one piecewise-linear approximation

1. Start from a set of sampling points $\{(X_i, \tilde{f}_i); i = 0,1,...,S\}$.
2. Fit the rank-one quadratic model in (10) using PROBE [6].
3. Calculate $Energy_{Linear}$ and $Energy_{Quadratic}$ using (17)-(18).
4. Determine the projection vector P_1 using (19).
5. Evenly partition the axis $P_1^T X$ into M segments, resulting in M+1 equally-spaced grid points $\{\alpha_{1,j}; j = 0,1,...,M\}$.
6. Solve the linear equations (20) for $\{\beta_{1,j}; j = 0,1,...,M\}$.
7. The rank-one piecewise-linear model is determined by substituting the solved model coefficients P_1 and $\{(\alpha_{1,j}, \beta_{1,j}); j = 0,1,...,M\}$ into (7)-(8).

Algorithm 1 summarizes the major steps of the proposed rank-one piecewise-linear modeling. Note that both the PROBE algorithm in Step 2 and the piecewise-linear fitting in Step 6 only require solving a sequence of over-determined linear equations. No further nonlinear optimization is involved in Algorithm 1. Therefore, the proposed piecewise-linear modeling algorithm completely eliminates the local convergence issue incurred by nonlinear optimization.

4.2 Rank-*K* Approximation

Algorithm 2: Rank-*K* piecewise-linear approximation

1. Start from a set of sampling points $\{(X_i, \tilde{f}_i); i = 0,1,...,S\}$.
2. For $k = 1,2,...,K$
3. Apply Algorithm 1 to the sampling points $\{(X_i, \tilde{f}_i); i = 0,1,...,S\}$ and extract the rank-one model $g_k(X)$.
4. Update the sampling points:
$$\tilde{f}_i = \tilde{f}_i - g_k(X_i) \quad (i = 1,2,\cdots,S). \tag{21}$$
5. End For
6. The rank-K piecewise-linear model is:
$$f_K(X) = g_1(X) + g_2(X) + \cdots + g_K(X). \tag{22}$$

Algorithm 2 shows the proposed P2M algorithm for rank-K piecewise-linear approximation. Starting from a set of sampling points, P2M first extracts a rank-one piecewise-linear model $g_1(X)$. Then, the sampling points are updated in (21) to calculate the residue which is further approximated as a new rank-one piecewise-linear model in the next iteration. The rank-one piecewise-linear fitting and the residue update are repeatedly applied for K times until the rank-K model $f_K(X)$ in (22) is achieved.

Algorithm 2 assumes a given approximation rank K. In practical applications, the value of K can be iteratively determined based on the approximation error. For example, starting from a low-rank approximation, K should be iteratively increased if the modeling error remains large.

5. NUMERICAL EXAMPLES

In this section we demonstrate the efficacy of P2M using two circuit examples. For each example, two independent sampling sets, called training set and testing set respectively, are generated. The training set contains 1000 sampling points that are created by Latin hypercube sampling [11]; it 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 2.8GHz Linux server.

5.1 SRAM Cell

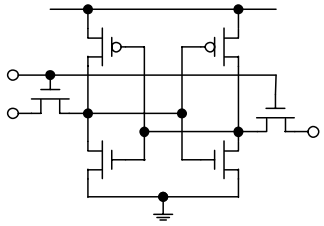

Fig 3. Circuit schematic of a 6T SRAM cell.

Fig 3 shows the circuit schematic of a 6T SRAM cell designed in a commercial 65nm CMOS process. We consider three important performance metrics for this SRAM cell: static noise margin (SNM), read margin (RM) and write margin (WM). These performance metrics are functions of both inter-die variations and device mismatches. The probability distribution and the correlation information of all variations are specified in the process design kit provided by the foundry.

We first apply fractional factorial experiment [16] to identify a subset of important process parameters that have significant influence on the performances of interest. After such a variable screening, 32 random variables are left to model process variations in this example.

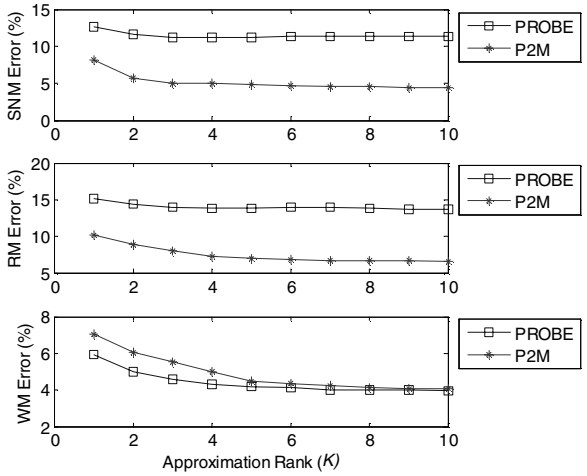

Fig 4. PROBE and P2M modeling error of SRAM cell.

We create performance models over the 5-sigma variation range using two different techniques: the traditional quadratic modeling (PROBE [6]) and the proposed piecewise-linear modeling (P2M). Fig 4 compares the modeling error for PROBE and P2M. As shown in Fig 4, the error of both PROBE and P2M decreases, as the approximation rank K increases. However, after

$K \geq 6$, further increases in K do not have a significant impact on reducing the error. It, in turn, implies that selecting the rank $K = 6$, instead of the full rank $K = 32$, is sufficient in this example.

Studying Fig 4, one would notice that the proposed P2M is substantially more accurate than PROBE when modeling the static noise margin and the read margin in this example. For instance, the modeling error of read margin is reduced by 2.1x, from 13.6% (PROBE) to 6.6% (P2M). It should be noted that as IC technologies are scaled to finer feature sizes, process variations are expected to become increasingly larger. It, in turn, would make the performance nonlinearities even more pronounced.

To intuitively understand the strongly nonlinear performance variations, we plot all training samples of read margin over the first dominant projection direction $P_1^T X$, as shown in Fig 5. Note that the strongly nonlinear relation between read margin and process variations cannot be predicted by a simple linear model. Namely, the first dominant projection vector P_1 of read margin cannot be extracted by a simple linear sensitivity analysis. This observation demonstrates the importance of the nonlinear sensitivity analysis proposed in Section 4.1. On the other hand, although applying quadratic sensitivity analysis yields the correct projection vector P_1, a simple quadratic model fails to accurately approximate the performance function. In this example, the proposed piecewise-linear model is required to achieve sufficiently small modeling error.

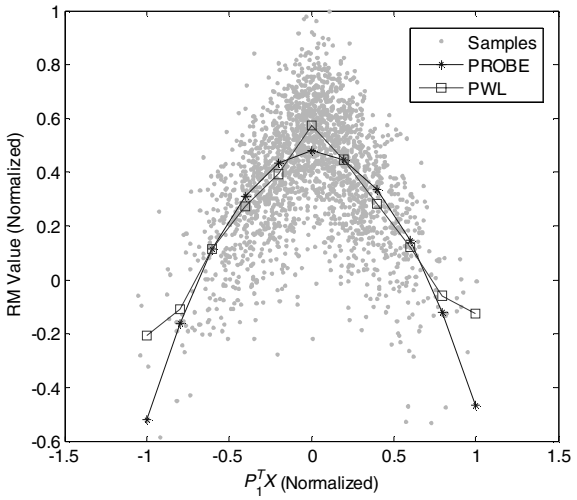

Fig 5. Sampling points of read margin over the first dominant projection direction $P_1^T X$.

Table 1. Computational cost of PROBE and P2M for SRAM cell performance modeling

Rank	Simulation Cost (Sec.)	Fitting Cost (Sec.)		Total Cost (Sec.)	
		PROBE	P2M	PROBE	P2M
1	980	5.35	10.45	985.35	990.45
2		7.63	13.05	987.63	993.05
3		11.74	16.79	991.74	996.79
4		15.33	21.69	995.33	1001.69
5		19.33	28.80	999.33	1008.80
6		23.20	37.46	1003.20	1017.46
7		26.93	47.65	1006.93	1027.65
8		29.45	57.48	1009.45	1037.48
9		32.59	67.97	1012.59	1047.97
10		34.89	80.24	1014.89	1060.24

Table 1 compares the computational cost for PROBE and P2M. The overall computational cost consists of two portions: simulation cost and fitting cost. The simulation cost is the computational time to run a numerical simulator (e.g., SPICE) to generate a number of sampling points. The fitting cost is the computational time to solve all unknown model coefficients from a sequence of over-determined linear equations. Studying Table 1, one would find that P2M has higher fitting cost than PROBE. However, since the simulation cost is dominant, the overall computational overhead of P2M is negligible (within 5%) in this example.

5.2 Operational Amplifier

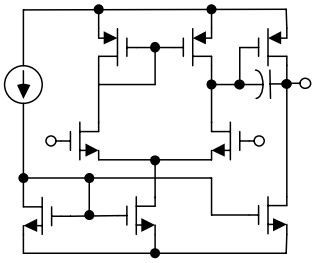

Fig 6. Circuit schematic of a two-stage operational amplifier.

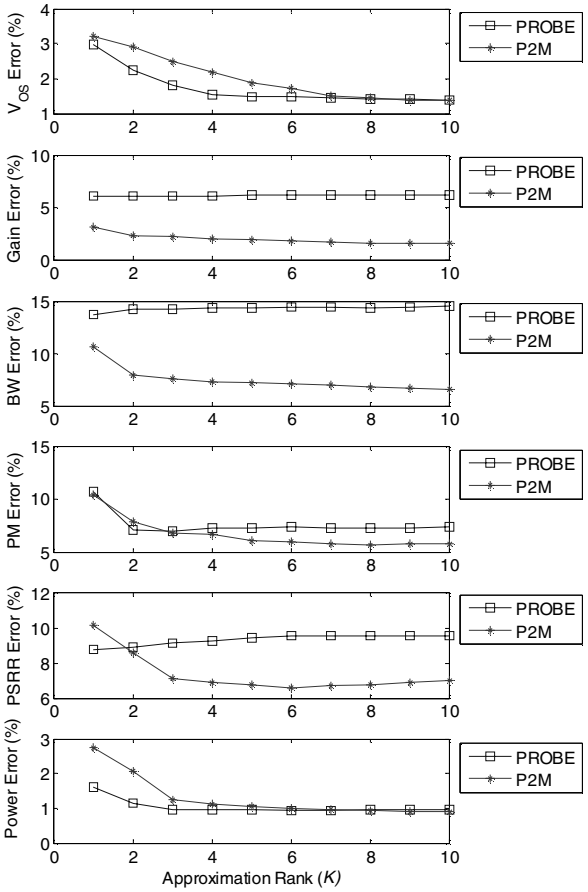

Fig 7. PROBE and P2M modeling error of operational amplifier.

Shown in Fig 6 is the circuit schematic of a two-stage operational amplifier designed in a commercial 65nm CMOS process. We consider six performance metrics in this example:

offset voltage (VOS), gain, bandwidth (BW), phase margin (PM), power supply rejection ratio (PSRR), and power. These performance metrics depend on both inter-die variations and device mismatches. The probability distribution and the correlation information of all variations are specified in the process design kit provided by the foundry.

Similar to the SRAM cell example, we first apply fractional factorial experiment [16] to identify a subset of important process variations. After such a variable screening, 47 random variables are left to model process variations in this example.

We create performance models over the 4-sigma variation range using two different techniques: the traditional quadratic modeling (PROBE [6]) and the proposed piecewise-linear modeling (P2M). Fig 7 compares the modeling error for PROBE and P2M. Two important observations can be made from Fig 7. Firstly, for both PROBE and P2M, selecting the rank $K = 6$, instead of the full rank $K = 47$, is sufficient in this example. Secondly, compared with PROBE, P2M achieves significant error reduction for most performance metrics. Taking bandwidth (BW) as an example, the PROBE error is 14.5%, while the P2M error is 6.6% (2.2x difference).

Table 2. Computational cost of PROBE and P2M for operational amplifier performance modeling

Rank	Simulation Cost (Sec.)	Fitting Cost (Sec.)		Total Cost (Sec.)	
		PROBE	P2M	PROBE	P2M
1	7760	15.42	23.13	7775.42	7783.13
2		33.48	31.23	7793.48	7791.23
3		54.04	46.86	7814.04	7806.86
4		68.43	62.53	7828.43	7822.53
5		86.96	82.84	7846.96	7842.84
6		104.97	104.14	7864.97	7864.14
7		122.85	124.82	7882.85	7884.82
8		141.60	142.24	7901.60	7902.24
9		159.48	184.80	7919.48	7944.80
10		186.25	237.65	7946.25	7997.65

Table 2 shows the computational cost for PROBE and P2M. In this example, P2M has slightly higher fitting cost than PROBE. However, since the simulation cost is dominant, the overall computational overhead of P2M is negligible (within 1%) in this example.

6. CONCLUSIONS

In this paper, we propose a novel projection-based piecewise-linear modeling (P2M) algorithm to capture high-dimensional strongly nonlinear performance variations that are observed in nanoscale technologies. P2M borrows the projection pursuit idea from mathematics to achieve dimension reduction. As such, a high-dimensional modeling problem can be converted to a low-dimensional problem that is tractable. In addition, piecewise-linear functions are utilized by P2M to model strongly nonlinear performance variations. By exploiting the unique piecewise-linear nature of the model template, a robust numerical algorithm is proposed to determine all model coefficients by solving a sequence of over-determined linear equations. Our numerical examples demonstrate that compared with the traditional quadratic modeling, P2M achieves 2x error reduction without substantially increasing the computational complexity. The response surface models created by P2M can be further incorporated into a statistical analysis/optimization environment for accurate and efficient parametric yield analysis/optimization.

7. ACKNOWLEDGEMENT

This work has been supported in part by the Semiconductor Research Corporation (SRC) and the National Science Foundation (NSF).

8. REFERENCES

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

[2] Semiconductor Industry Associate, *International Technology Roadmap for Semiconductors*, 2005.

[3] 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.

[4] 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.

[5] F. Schenkel, M. Pronath, S. Zizala, R. Schwencker, H. Graeb and K. Antreich, "Mismatch analysis and direct yield optimization by spec-wise linearization and feasibility-guided search," *IEEE DAC*, pp. 858-863, 2001.

[6] X. Li, J. Le, L. Pileggi and A. Strojwas, "Projection-based performance modeling for inter/intra-die variations," *IEEE ICCAD*, pp. 721-727, 2005.

[7] Z. Feng and P. Li, "Performance-oriented statistical parameter reduction of parameterized systems via reduced rank regression," *IEEE ICCAD*, pp. 868-875, 2006.

[8] A. Singhee and R. Rutenbar, "Beyond low-order statistical response surfaces: latent variable regression for efficient, highly nonlinear fitting," *IEEE DAC*, pp. 256-261, 2007.

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

[10] X. Li, J. Le and L. Pileggi, "Projection-based statistical analysis of full-chip leakage power with non-log-Normal distributions," *IEEE DAC*, pp. 103-108, 2006.

[11] 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.

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

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

[14] A. Papoulis and S. Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 2001.

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

[16] D. Montgomery, *Design and Analysis of Experiments*, John Wiley & Sons, 2005.