Efficient Performance Estimation with Very Small Sample Size via Physical Subspace Projection and Maximum A Posteriori Estimation

Li Yu, Sharad Saxena[1], Christopher Hess[1], Ibrahim (Abe) M. Elfadel[2], Dimitri Antoniadis, Duane Boning
Massachusetts Institute of Technology, [1]PDF Solutions, [2]Masdar Institute of Science and Technology
Email: yul09@mit.edu

***Abstract*—In this paper, we propose a novel integrated circuits performance estimation algorithm through a physical subspace projection and maximum-a-posteriori (MAP) estimation. Our goal is to estimate the distribution of a target circuit performance with very small measurement sample size from on-chip monitor circuits. The key idea in this work is to exploit the fact that simulation and measurement data are physically correlated under different circuit configurations and topologies. First, different groups of measurements are projected to a subspace spanned by a set of physical variables. The projection is achieved by performing a sensitivity analysis of measurement parameters with respect to the subspace variables using a virtual source MOSFET compact model. Then a Bayesian treatment is developed by introducing prior distributions over these subspace variables. Maximum a posteriori estimation is then applied using the prior, and an expectation-maximization (EM) algorithm is used to estimate the circuit performance. The proposed method is validated by post-silicon measurement for a commercial 28-nm process. An average error reduction of 2x is achieved which can be translated to 32x reduction on data size needed for samples on the same die. A 150x and 70x sample size reduction on training dies is also achieved compared to traditional least-square fitting method and least-angle regression method, respectively, without reducing accuracy.**

I. Introduction

Continued scaling of CMOS technology has introduced increased variations of process and design parameters, which affects all aspects of circuit performance [1]. A critical problem in post-Silicon validation is to build statistically valid prediction models of circuit performance based on a small number of measurements. These prediction models could then be used in many circuit applications such as parametric yield prediction and robust circuit design.

A widely adopted performance modeling technique is response surface modeling (RSM) which approximates the circuit performance (e.g., delay, power, etc.) as an analytical (typically linear or quadratic) function of device parameters (e.g., V_{th}, T_{ox}, etc.) [2]. However, a large number of device-level random variables must be used to capture device-level variations, which results in a variation space of very high dimensions. Although principal component analysis (PCA) could be employed as a useful dimensionality reduction method to convert a high-dimensional space into a set of uncorrelated variables, the required training sample size is quite large [3]. When the data set is not large enough to support the variable space, parameter estimates by RSM become very-much data-dependent; this phenomenon is known as over-fitting.

One common strategy for preventing over-fitting in performance modeling is by adding regularization terms to error functions. An example of such a strategy is least-angle regression (LAR) which uses L_1-norm regularization [4]. One major benefit of regularizing with the L_1-norm is that it results in sample complexity logarithmic in the number of features. On the other hand, an L_2 regularization results in sample complexity that is linear in the number of features. Recent work employed Bayesian inference to address the over-fitting problem where sparse model coefficients and correlated performance variability were exploited [5] [6]. Such methods require the reuse of data collected from the same system (either a different mode/corner or a different stage) and 50 or more samples are still required for each measurement.

In post-Silicon performance estimation, a typical situation is to predict the performance of a target system based on a very small number of measurements taken from different configurations of on-chip monitor circuits (e.g., device-array test structure, ring-oscillator structure, etc.). The very small size of the data set is typically due to the following two reasons: (1) a limited number of replicated devices under test (DUTs) per die due to limited area and pads for on-chip monitor circuits; and (2) a limited number of training dies are measured due to test time limitations. To the best of our knowledge, there is no satisfactory solution to the problem of predicting circuit performance based on a very small data set of measurements. This paper proposes one such solution.

The key idea in this work is to exploit two types of physical correlation. The first is the one that exists between simulation and measurements under different circuit configurations and topologies. The second exists between different groups of performance measurements. These correlations can physically be traced back to the hidden intrinsic parameters of the technology such as threshold voltage and source velocity. The core of the proposed method is to first project different groups of measurements onto a subspace spanned by a set of physical variables (e.g., V_{thn} and V_{thp}). The projection is achieved by performing a sensitivity analysis of measurement parameters with respect to subspace variables using the MIT virtual source (MVS) compact model. A prior distribution is defined over these subspace variables and a Bayesian formalism is introduced to estimate the performance parameters. This is achieved using a maximum a posteriori (MAP) estimation defined over all the group measurement distributions and the subspace variable prior. A modification of an expectation-maximization (EM) algorithm is employed to iteratively solve the MAP estimation problem.

II. Background and Problem Definition

Without loss of generality, we consider the problem of estimating a single performance of interest, denoted by g. Assume that g follows a Gaussian distribution $g \sim \mathcal{N}(\mu_{\mathbf{g}}, \sigma_g)$:

$$pdf(g) = \frac{1}{\sqrt{2\pi} \cdot \sigma_g} \cdot exp[-\frac{(g-\mu_g)^2}{2 \cdot \sigma_g^2}] \tag{1}$$

where μ_g and σ_g are, respectively, the mean and standard deviation of the performance distribution. However, due to constraints on testing costs, measurements of g may not be directly available. Instead, groups of measurement data of other performance parameters are provided that we denote by $\mathbf{F} = \{F_1, F_2, ..., F_m\}$.

As an example, consider the problem of post-Silicon validation of a digital system. In this application, the performance metric

978-3-9815370-2-4/DATE14/©2014 EDAA

g might be critical path delays or leakage power across a die and F_i would be measurement results from on-chip monitoring arrays (e.g., threshold voltages, I_{dsat} for transistors, frequencies for ring oscillators (ROs) [7][8]). Here the variability of g is mainly caused by parameter variations such as V_{th} and V_{dd}. Our task therefore is to predict the distribution of g given $\mathbf{F}$ and, consequently, predict the parametric yield.

To formalize the above description, we define a measurement *group* to be a performance measured under a certain circuit topology and configuration. We assume that there are m such groups. To each group i ($i \in [1, m]$), we associate a random variable F_i to model the variability of the measurement under a certain circuit configuration. Therefore the aforementioned $\mathbf{F}$ could be represented by $\{F_1, F_2, ..., F_m\}$. We also assume that each F_i follows a Gaussian distribution $F_i \sim \mathcal{N}(\mu_{F_i}, \sigma_{F_i})$:

$$pdf(F_i) = \frac{1}{\sqrt{2\pi} \cdot \sigma_{F_i}} \cdot exp[-\frac{(F_i - \mu_{F_i})^2}{2 \cdot \sigma_{F_i}^2}] \tag{2}$$

For each group F_i, we obtain a set of independent observations $\{F_i\} = \{F_i^{(1)}, F_i^{(2)}, ..., F_i^{(N_i)}\}$, where N_i is the sample size of the i-th group. The problem we aim to address is to estimate μ_g and σ_g given the observations $\{F_1, F_2, ..., F_m\}$ with the constraint that N_i are very small. For simplicity, we consider the case where $N_1 = N_2 = ... = N_m = N$.

This problem cannot be addressed by the conventional moment estimation techniques because it is hard to assign a weight to each group and because the relationships between g and the F_i's are unclear [9]. One possible approach is to apply principal component analysis (PCA) to $\mathbf{F}$ and select its top features $\mathbf{X}$. The problem is then converted into a performance modeling problem. If the performance function is approximated as:

$$g(\Delta\mathbf{X}) = \sum_{k=1}^{M} \alpha_{gk} \cdot b_k(\Delta\mathbf{X}) \tag{3}$$

where $\{b_k(\Delta\mathbf{X}); k = 1, 2, ..., M\}$ contains the basis functions (e.g, linear, quadratic, etc), and $\{\alpha_{gk}; (k = 1, 2, ..., M)\}$ are the model coefficients. The unknown model coefficients α_{gi} are usually determined by solving a linear system with N sampling points:

$$\mathbf{G} = \mathbf{B} \cdot \alpha_g \tag{4}$$

where

$$\alpha_g = [\alpha_{g1} \quad \alpha_{g2} \quad ... \quad \alpha_{gM}]^T \tag{5}$$

$$\mathbf{G} = [G^{(1)} \quad G^{(2)} \quad ... \quad G^{(N)}]^T \tag{6}$$

$$G^{(i)} = g(\Delta\mathbf{X}^{(i)})$$

$$\mathbf{B} = \begin{bmatrix} b_1(\Delta\mathbf{X}^{(1)}) & b_2(\Delta\mathbf{X}^{(1)}) & \cdots & b_M(\Delta\mathbf{X}^{(1)}) \\ b_1(\Delta\mathbf{X}^{(2)}) & b_2(\Delta\mathbf{X}^{(2)}) & \cdots & b_M(\Delta\mathbf{X}^{(2)}) \\ \vdots & \vdots & & \vdots \\ b_1(\Delta\mathbf{X}^{(N)}) & b_2(\Delta\mathbf{X}^{(N)}) & \cdots & b_M(\Delta\mathbf{X}^{(N)}) \end{bmatrix} \tag{7}$$

However, the relationship between $\mathbf{X}$ and $\mathbf{G}$ is unknown and we have no prior information on $\{\alpha_{gk}; (k = 1, 2, ..., M)\}$. Under the constraint of very small N, strong over-fitting would appear and the prediction would be inaccurate. Although least-angle regression (LAR) or sparse regression could add a regularization term on $\{\alpha_{gk}; (k = 1, 2, ..., M)\}$, an appreciable number of samples is still required. This is the main motivation for the development of a new performance estimation method via physical subspace projection and Maximum A Posteriori (MAP) estimation. In contrast to PCA, we project $\mathbf{F}$ onto a physical variable subspace $\mathbf{X}$ where a Bayesian inference learns a prior distribution on the $\mathbf{X}$ parameters using measurement data from all the groups. The estimates of μ_g and σ_g are also obtained via a projection onto the $\mathbf{X}$ subspace, and the projection operation itself is facilitated using the virtual source (VS) MOSFET model [10].

III. Physical Subspace Projection

A. Review of MIT Virtual Source (MVS) Model

The VS model is an ultra compact, charge-based MOSFET model that provides a simple, physics-based description of carrier transport in modern short-channel MOSFETs [10][11]. It essentially substitutes the quasi-ballistic carrier transport concept for the concept of drift-diffusion with velocity-saturation. In doing so, it achieves excellent accuracy for the I-V and C-V characteristics of the device throughout the various domains of circuits operation. The number of parameters needed is considerably fewer (11 for DC and 24 in total) than in conventional models[12].

Recently, the statistical extension of MVS was developed with the capability of mapping the variability characterization in device behavior onto a limited number of underlying model parameters, which in turn enables the efficient prediction of variations in circuit performance [13].

B. Definition of Physical Subspace

We define *physical subspace* as a variable space spanned by model parameters in the VS model (e.g., V_{tn}, V_{tp}, etc.). Notice that model parameters are different from measured parameters. For example, V_t is commonly measured through the so-called "constant current method" where threshold voltage is the gate bias corresponding to an arbitrary value of drain current, for instance $0.1\mu A$ [14]. Such measured V_t relates to factors such as transistor geometries and devices under test (DUTs). Hence its absolute value does not have physical meaning. A VS model parameter, in contrast, is a physical parameter with fixed value shared by all transistors with different geometries.

Although parameters measured from different groups have large differences in their absolute values, they are strongly correlated. This assertion is not only valid for the same parameter measured from different configurations (e.g., V_t measurement for transistors with different geometries), but is often also valid for different parameters measured from different configurations (e.g., I_{dsat} for a transistor and frequency for a ring oscillator (RO)). Fig. 1 shows different groups of on-chip monitoring measurements from a real product. All parameters in the red box refer to parameters that are directly measurable. They are governed by a hidden model parameter, namely, V_{tn} (other hidden parameters and their link to measured parameters are not shown in this figure).

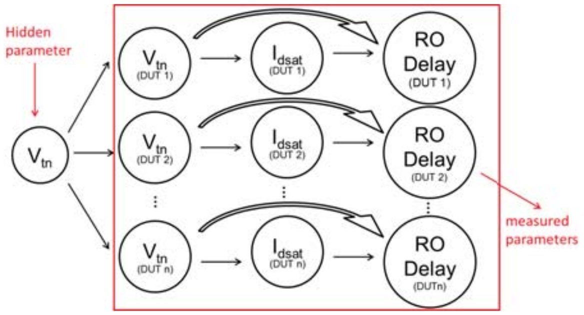

Fig. 1. The graphical model link for parameter correlations.

C. Physical Subspace Selection

The selection of physical subspace $\mathbf{X}$ is a key step in physical subspace projection. Here we propose a least-angle regression (LAR) method to solve this feature selection problem before using any measurement results. A set of MVS model parameters

$\mathbf{Y} = \{Y_1, Y_2, ..., Y_s\}$ are preselected as candidate subspace variables. Then Monte-Carlo simulations are run to compute target performance g by randomly generating samples of each VS model parameter. $\mathbf{X}$ is initially set to be $\{\oslash\}$. Next, LAR finds the vector Y_{si} that is most correlated with g. Once Y_{si} is identified, Y_{si} is removed from $\mathbf{Y}$ and added to $\mathbf{X}$. The model coefficient α is determined by solving the linear equation $\mathbf{G} = \Delta\mathbf{X}\cdot\alpha$ and the residual of the approximation is calculated by:

$$Res = \mathbf{G} - \Delta\mathbf{X}\cdot\alpha \tag{8}$$

Then a new vector Y_{new} which has most correlation with the residual Res is found. The whole process is repeated until Res is smaller than a given threshold. For performance of a typical digital system, we found that $\mathbf{X} = \{V_{tn}, V_{tp}\}$ would be sufficient for a Res criterion of $0.02\mathbf{G}$ and $\mathbf{X} = \{V_{tn}, V_{tp}, v_{xon}, v_{xop}\}$ would be sufficient for a Res criterion of $0.01\mathbf{G}$ where v_{xon} and v_{xop} are the virtual source velocity for NMOS and PMOS, respectively. It is intuitive that V_{tn} and V_{tp} are the top features because random dopant fluctuation and channel length variability are highly important physical sources of performance variation. For simplicity, we select $\mathbf{X} = \{V_{tn}, V_{tp}\}$ for the rest of this paper.

D. Process Shift Calibration

Once the physical subspace $\mathbf{X}$ and its corresponding basis functions $\mathbf{B} = \{b_k(\Delta\mathbf{X}); k = 1, 2, ..., M\}$ are obtained we are able to determine model coefficient α_g and α_{F_i} by solving linear equations $\mathbf{G} = \mathbf{B}\cdot\alpha_g$ and $F_i = \mathbf{B}\cdot\alpha_{F_i}$, respectively. This allow us to build a one-to-many function from $\mathbf{X}$ to target performance g and measurements $\mathbf{F}$. In order to reuse prior information, we assume the coefficients α_g and α_{F_i} of post-layout simulations are identical with the α_g and α_{F_i} of measurement results. While this assumption usually holds in many practical applications, it is sometimes the case that there is mismatch between the nominal performance values of post-layout simulations and the measurements with a typical shift of 15% or less. The shifts in the corresponding performance distributions are due to modeling and extraction inaccuracy. Therefore a very small sample size would be needed to calibrate $\mathbf{F}_{nom}$ and g_{nom}.

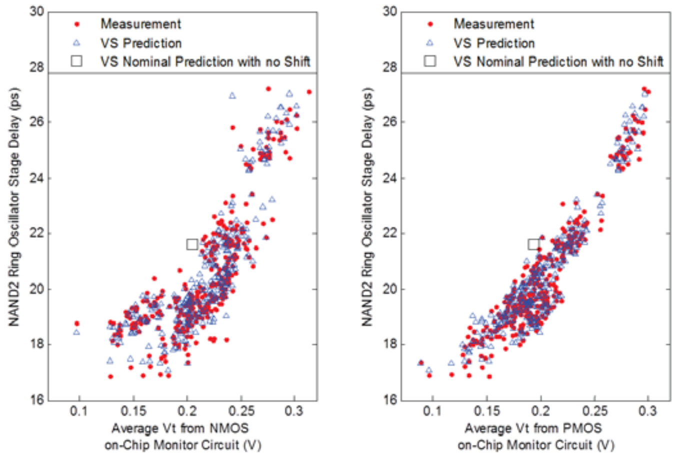

Fig. 2. A comparison of measured and VS model predicted Ring Oscillator (RO) stage delay versus (a) NMOS V_t and; (b) PMOS V_t, respectively. Nominal post-layout simulation without any shift and variation is also marked. V_{tn} on left and V_{tp} on right are hidden but their impact on prediction has been modeled.

To further illustrate the calibration, Fig. 2 shows an example of RO stage delay measurements versus V_{tn} and V_{tp} extracted from same-die test arrays and compared with modeling prediction. A prediction of nominal performance without process shift calibration is also shown in the figure. Note that measurement results are sampled from dies on various wafers and lots, and only a few dies (< 10) are needed to calibrate the nominal shift.

After building the link between physical subspace and performance, we are able to draw the performance map for different systems. Fig. 3 shows the INV and NAND RO stage delay versus V_{tn} and V_{tp}. A high similarity is observed between the two maps; this means that knowing the measurement result for one digital system would give us confidence in predicting other digital system performances.

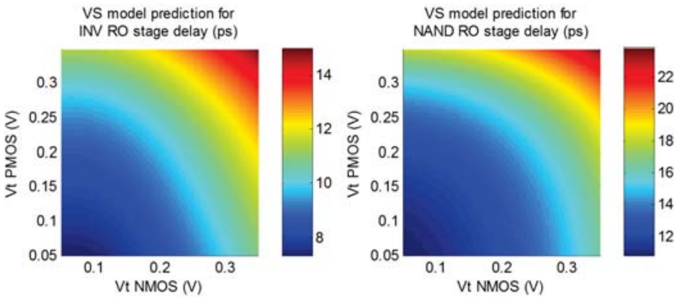

Fig. 3. Sensitivity analysis on (a) INV and (b) NAND2 ring oscillator (RO) stage delay using VS model.

E. Physical Subspace Projection

The purpose of *physical subspace projection* is to transfer measurement data from different groups into a unique physical subspace $\mathbf{X}$. This is a one-to-many function that cannot be resolved using deterministic methods. However, given α and the calibrated process shift, we could calculate the pdf on $\mathbf{X}$, and solve the whole problem using maximum a posteriori (MAP) estimation.

In line with our previous assumptions, the subspace $\mathbf{X}$ satisfy a multivariate Gaussian distribution $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\theta})$:

$$pdf(\mathbf{X}) = \frac{1}{\sqrt{(2\pi)^k\,|\boldsymbol{\theta}|}}\cdot exp[-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_\mathbf{X})^\mathbf{T}\boldsymbol{\theta}^{-1}(\mathbf{X}-\boldsymbol{\mu}_\mathbf{X})] \tag{9}$$

where $\boldsymbol{\mu}_\mathbf{X}$ and $\boldsymbol{\theta}$ are the mean vector and the covariance matrix of $\mathbf{X}$, and k is the dimension of $\mathbf{X}$.

We also assume the "uncertainty" of $\boldsymbol{\mu}_\mathbf{X}$ follows a conjugate Gaussian prior distribution $\boldsymbol{\mu}_\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.

$$pdf(\boldsymbol{\mu}_\mathbf{X}) = \frac{1}{\sqrt{(2\pi)^k\,|\boldsymbol{\Sigma}_\mathbf{0}|}}\cdot exp[-\frac{1}{2}(\boldsymbol{\mu}_\mathbf{X}-\boldsymbol{\mu}_0)^\mathbf{T}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_\mathbf{X}-\boldsymbol{\mu}_0)] \tag{10}$$

μ_{F_i} and σ_{F_i} are calculated by:

$$\mu_{F_i} = \mu(F_i(\Delta\mathbf{X})) = \sum_{k=1}^{M}\alpha_{F_ik}\cdot\mu(b_k(\Delta\mathbf{X})) = \mu_{F_i}(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\theta}) \tag{11}$$

$$\sigma^2_{F_i} = \sigma(F_i(\Delta\mathbf{X}))^2 = \sum_{j=1}^{M}\sum_{k=1}^{M}\alpha_{F_ij}\alpha_{F_ik}\cdot\sigma(b_j(\Delta\mathbf{X}), b_k(\Delta\mathbf{X})) - (\sum_{k=1}^{M}\alpha_{F_ik}\cdot\mu(b_k(\Delta\mathbf{X})))^2 = \sigma^2_{F_i}(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\theta}) \tag{12}$$

where $\mu(b_k(\Delta\mathbf{X}))$ and $\sigma(b_j(\Delta\mathbf{X}), b_k(\Delta\mathbf{X}))$ are the mean and covariance of the basis function, respectively.

Therefore the probability of observing data point $F_i^{(n_i)}$ in ith group associated with subspace distribution $pdf(F_i^{(n_i)}|\boldsymbol{\mu}_\mathbf{X}, \theta)$

$$pdf(F_i^{(n_i)}|\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_{F_i}(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\theta})}exp[-\frac{(F_i^{(n_i)} - \mu_{F_i}(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\theta}))^2}{2\cdot\sigma_{F_i}(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\theta})^2}] \tag{13}$$

which is the complete form of physical subspace projection.

IV. Maximum A Posteriori Estimation

A. Initial setting

Our proposed physical subspace projection method is facilitated by a Bayesian inference which efficiently exploits the correlation between different groups of measurement to improve the accuracy of the estimator. Before we start, a proper physical variable subspace $\mathbf{X}$ is selected and prior knowledge is learned by fitting α_g and α_{F_i}, as described in Section III. Process shifts will be calibrated by a few training dies and initial guess of parameters $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\theta}$ will be selected. $\boldsymbol{\mu}_0$ is the nominal value for the subspace variables and $\boldsymbol{\Sigma}_0$ is the covariance matrix of subspace variables under inter-die variation. The initial value of $\boldsymbol{\theta}$ equals to the covariance of subspace variables under only intra-die variation.

B. Learning a Prior Distribution

The first step is to project very small samples in different measurement groups $\{\{F_i^{n_i}; n_i 1, 2, ..., N_i\}; i = 1, 2..., m\}$ to selected subspace $\mathbf{X}$ and obtain the probability of observing $\mathbf{F_i}$ given $\boldsymbol{\mu}_{\mathbf{X}}$ and θ. Then we further combine it with the prior distribution $pdf(\boldsymbol{\mu}_{\mathbf{X}})$ in (10) to accurately estimate $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\theta}$ and essentially μ_g and σ_g.

Assuming samples from different measurement groups are independent, we can write the likelihood function $pdf(\mathbf{F}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta})$ as:

$$pdf(\mathbf{F}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) = \prod_{i=1}^{m} pdf(F_i|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) \tag{14}$$

Assuming samples from the same measurement groups are independent, the likelihood function $pdf(F_i|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta})$ is written as:

$$pdf(F_i|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) = \prod_{n_i=1}^{N_i} pdf(F_i^{(n_i)}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) \tag{15}$$

According to Bayes' theory, the joint distribution $p(\mathbf{F}, \boldsymbol{\mu}_{\mathbf{X}}|\boldsymbol{\theta})$ is given by the product of the prior $pdf(\boldsymbol{\mu}_{\mathbf{X}})$ and the likelihood function $pdf(\mathbf{F}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta})$, described by *posterior distribution*:

$$pdf(\mathbf{F}, \boldsymbol{\mu}_{\mathbf{X}}|\boldsymbol{\theta}) = pdf(\boldsymbol{\mu}_{\mathbf{X}}|\boldsymbol{\theta}) \cdot pdf(\mathbf{F}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) \tag{16}$$

Substituting (14) and (15) into (16) and noticing that $pdf(\boldsymbol{\mu}_{\mathbf{X}}|\boldsymbol{\theta}) = pdf(\boldsymbol{\mu}_{\mathbf{X}})$ yield:

$$\begin{aligned} pdf(\mathbf{F}, \boldsymbol{\mu}_{\mathbf{X}}|\boldsymbol{\theta}) &= pdf(\boldsymbol{\mu}_{\mathbf{X}}) \cdot \prod_{i=1}^{m} \prod_{n_i=1}^{N_i} pdf(F_i^{(n_i)}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) \\ &= pdf(\boldsymbol{\mu}_{\mathbf{X}}) \cdot pdf(F_1^{(n_1)}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta})) \cdot ... \cdot pdf(F_m^{(N_m)}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) \end{aligned} \tag{17}$$

This demonstrates the sequential nature of Bayesian learning in which the current posterior distribution forms the prior when a new data point is observed. Fig. 4 shows the results of Bayesian learning on $\boldsymbol{\mu}_{\mathbf{X}}$ as the portfolio of the measurement groups expanded. The first column of this figure corresponds to the situation before any data points are observed and shows a plot of the prior distribution $\boldsymbol{\mu}_{\mathbf{X}} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. The first row shows the likelihood function $pdf(F_i|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta})$ for different measurement alone. The second row shows posterior distribution $pdf(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{F}|\boldsymbol{\theta})$ by multiplying its likelihood function from the top row by the prior. As this process continues, the posterior distribution is much sharper and in the limit of an infinite number of data points, the posterior distribution would become a delta function centered on the true parameter values.

C. Maximum A Posteriori Estimation

Our final goal is to find an optimal estimation of $\boldsymbol{\mu}_{\mathbf{X}}$ which maximize the log likelihood of posterior distribution $\ln pdf(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{F}|\boldsymbol{\theta})$. However, a key step is still missing, which is to determine the hidden variable $\boldsymbol{\theta}$ which maximizes the log likelihood function

$$\mathbf{ln} pdf(\mathbf{F}|\boldsymbol{\theta}) = \mathbf{ln} \int_{\mathbf{X}} pdf(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{F}|\boldsymbol{\theta}) d\boldsymbol{\mu}_{\mathbf{X}} \tag{18}$$

The difficulty arises from the presence of integration that appears inside the logarithm in (18), so that the logarithm function no longer acts directly on the Gaussian. If we set the derivatives of the log likelihood to zero, we will no longer obtain a closed form solution. The idea presented in this paper follows the expectation maximization (EM) algorithm [15].

For any normalized distribution $q(\boldsymbol{\mu}_{\mathbf{X}})$, we have

$$\begin{aligned} &\mathbf{ln} pdf(\mathbf{F}|\boldsymbol{\theta}) = 1 \cdot \mathbf{ln} pdf(\mathbf{F}|\boldsymbol{\theta}) = \textstyle\int_{\mathbf{X}} q(\boldsymbol{\mu}_{\mathbf{X}}) d\boldsymbol{\mu}_{\mathbf{X}} \cdot \mathbf{ln} pdf(\mathbf{F}|\boldsymbol{\theta}) \\ &= \textstyle\int_{\mathbf{X}} q(\boldsymbol{\mu}_{\mathbf{X}}) \mathbf{ln} pdf(\mathbf{F}|\boldsymbol{\theta}) d\boldsymbol{\mu}_{\mathbf{X}} = \int_{\mathbf{X}} q(\boldsymbol{\mu}_{\mathbf{X}}) \mathbf{ln} \frac{pdf(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{F}|\boldsymbol{\theta})}{pdf(\boldsymbol{\mu}_{\mathbf{X}}|\mathbf{F}, \boldsymbol{\theta})} d\boldsymbol{\mu}_{\mathbf{X}} \\ &= \textstyle\int_{\mathbf{X}} q(\boldsymbol{\mu}_{\mathbf{X}}) \left(\mathbf{ln} pdf(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{F}|\boldsymbol{\theta}) - \mathbf{ln} q(\boldsymbol{\mu}_{\mathbf{X}}) - \mathbf{ln} \frac{pdf(\boldsymbol{\mu}_{\mathbf{X}}|\mathbf{F}, \boldsymbol{\theta})}{q(\boldsymbol{\mu}_{\mathbf{X}})} \right) d\boldsymbol{\mu}_{\mathbf{X}} \end{aligned} \tag{19}$$

Here the second item $-\int_{\mathbf{X}} q(\boldsymbol{\mu}_{\mathbf{X}}) \mathbf{ln} q(\boldsymbol{\mu}_{\mathbf{X}}) d\boldsymbol{\mu}_{\mathbf{X}}$ is always a constant. The third item $\int_{\mathbf{X}} q(\boldsymbol{\mu}_{\mathbf{X}}) \mathbf{ln} \frac{pdf(\boldsymbol{\mu}_{\mathbf{X}}|\mathbf{F}, \boldsymbol{\theta})}{q(\boldsymbol{\mu}_{\mathbf{X}})} d\boldsymbol{\mu}_{\mathbf{X}}$ is the Kullback-Leibler divergence between $pdf(\boldsymbol{\mu}_{\mathbf{X}}|\mathbf{F}, \boldsymbol{\theta})$ and $q(\boldsymbol{\mu}_{\mathbf{X}})$ which is ≥ 0, with equality if and only if $q(\boldsymbol{\mu}_{\mathbf{X}}) = pdf(\boldsymbol{\mu}_{\mathbf{X}}|\mathbf{F}, \boldsymbol{\theta})$.

Algorithm 1 Algorithm to solve Maximum A Posteriori estimation

```
Require: a joint distribution $pdf(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{F}|\boldsymbol{\theta})$ over observed variables $\mathbf{F}$ and latent variables $\boldsymbol{\mu}_{\mathbf{X}}$, governed by parameters $\boldsymbol{\theta}$, convergence requirement $\epsilon$.
Ensure: $\boldsymbol{\theta}$ which maximize the likelihood function $\mathbf{ln} pdf(\mathbf{F}|\boldsymbol{\theta})$ and $\boldsymbol{\mu}_{\mathbf{X}}$ which maximize the likelihood function $pdf(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{F}|\boldsymbol{\theta})$
Choose an initial setting for the parameters $\boldsymbol{\theta}^{new}$;
repeat
    $\boldsymbol{\theta}^{old} = \boldsymbol{\theta}^{new}$;
    Evaluate $pdf(\boldsymbol{\mu}_{\mathbf{X}}|\mathbf{F}, \boldsymbol{\theta}^{old}) = \frac{pdf(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{F}|\boldsymbol{\theta}^{old})}{pdf(\mathbf{F}|\boldsymbol{\theta}^{old})}$;
    Evaluate $\boldsymbol{\theta}^{new}$given by
            $\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$;
    $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \int_{\mathbf{X}} pdf(\boldsymbol{\mu}_{\mathbf{X}}|\mathbf{F}, \boldsymbol{\theta}^{old}) \ln pdf(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{F}|\boldsymbol{\theta}) d\mathbf{X}$;
until $|\boldsymbol{\theta}^{old} - \boldsymbol{\theta}^{new}| < \epsilon$
$\boldsymbol{\mu}_{\mathbf{X}} = \arg\max_{\boldsymbol{\mu}_{\mathbf{X}}} \ln pdf(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{F}|\boldsymbol{\theta}^{new})$;
```

This suggests an iterative algorithm, as summarized in Algorithm 1. Given an initial value of $\boldsymbol{\theta}^{old}$, the first step is to maximize likelihood $\mathbf{ln} pdf(\mathbf{F}|\boldsymbol{\theta}^{old})$ with respect to $q(\boldsymbol{\mu}_{\mathbf{X}})$, which yields to $q(\boldsymbol{\mu}_{\mathbf{X}}) = pdf(\boldsymbol{\mu}_{\mathbf{X}}|\mathbf{F}, \boldsymbol{\theta}^{old})$. The second step is to fix the distribution $q(\boldsymbol{\mu}_{\mathbf{X}})$ and maximize $\mathbf{ln} pdf(\mathbf{F}|\boldsymbol{\theta}^{old})$ with respect to $\boldsymbol{\theta}^{old}$. The whole process is repeated until convergence and estimations of $\boldsymbol{\theta}$ and $\boldsymbol{\mu}_{\mathbf{X}}$ are obtained.

Once $\boldsymbol{\theta}$ and $\boldsymbol{\mu}_{\mathbf{X}}$ are obtained, we could further estimate the mean and standard deviation of target performance μ_g and σ_g:

$$\mu_g = \mu(g(\Delta\mathbf{X})) = \sum_{k=1}^{M} \alpha_{gk} \cdot \mu(b_k(\Delta\mathbf{X})) \tag{20}$$

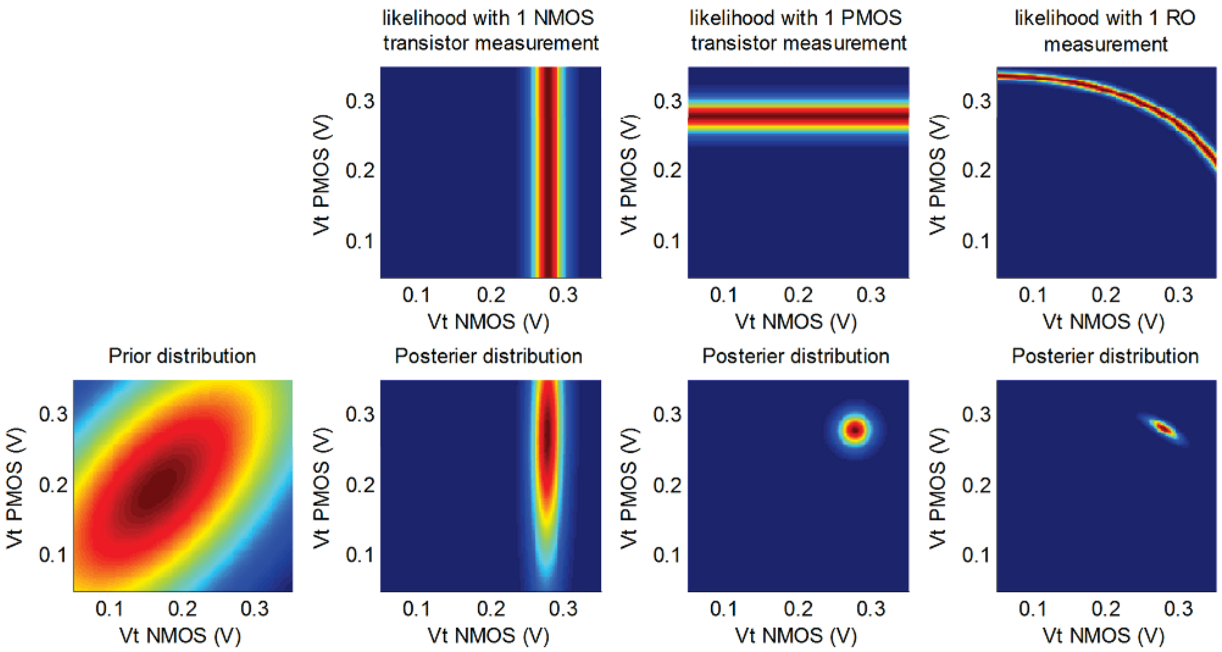

Fig. 4. Illustration of sequential Bayesian learning of $\boldsymbol{\mu}_{\mathbf{X}}$ from prior and on-chip monitor circuit.

$$\sigma_g^2 = \sigma^2(g(\Delta\mathbf{X})) = \sum_{i=1}^{M}\sum_{j=1}^{M}\alpha_{gi}\alpha_{gj}\cdot\sigma(b_i(\Delta\mathbf{X}), b_j(\Delta\mathbf{X})) - (\sum_{k=1}^{M}\alpha_{gk}\cdot\mu(b_k(\Delta\mathbf{X})))^2 \tag{21}$$

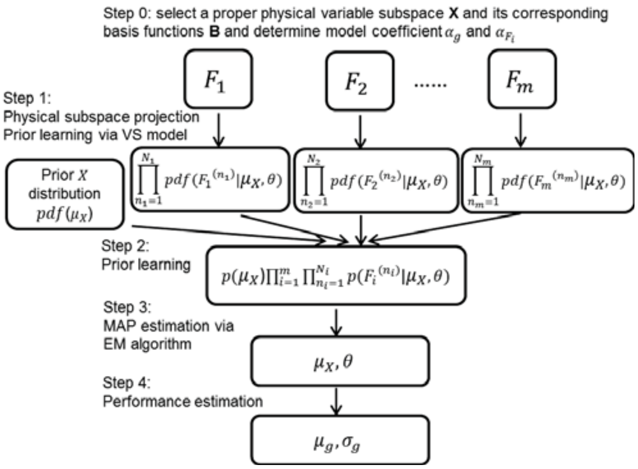
Step 0: select a proper physical variable subspace X and its corresponding basis functions B and determine model coefficient α_g and α_{F_i}
F_1
F_2
......
F_m
Step 1:
Physical subspace projection
Prior learning via VS model
Prior X distribution $pdf(\mu_X)$
$\prod_{n_1=1}^{N_1} pdf(F_1{}^{(n_1)}|\mu_X,\theta)$
$\prod_{n_2=1}^{N_2} pdf(F_2{}^{(n_2)}|\mu_X,\theta)$
$\prod_{n_m=1}^{N_m} pdf(F_m{}^{(n_m)}|\mu_X,\theta)$
Step 2:
Prior learning
$p(\mu_X)\prod_{i=1}^{m}\prod_{n_i=1}^{N_i} p(F_i{}^{(n_i)}|\mu_X,\theta)$
Step 3:
MAP estimation via
EM algorithm
μ_X,θ
Step 4:
Performance estimation
μ_g,σ_g

Fig. 5. Proposed method employing a Bayesian interface and Maximum A Posteriori estimation.

A summary of our physical subspace projection and maximum a posteriori estimation is shown in Fig. 5.

V. Validation

In this section, we demonstrate the accuracy and efficacy of our proposed physical subspace projection and MAP algorithm using measurement results. We consider on-chip measurement results collected from 3186 dies in 27 wafers in a 28-nm bulk CMOS process. Each chip contains different test arrays which include device-array and ring oscillator (RO)-array, which are often used as monitor circuits due to their simplicity and small area overhead. Information about each measurement group is summarized in Table I.

TABLE I
A summary of measurement groups

# of meas group	1	2	3	4	5	6
DUT	INV	NAND	NOR	INV	NAND	NOR
Circuit topology	Device	Device	Device	RO	RO	RO
# of replicas	4	4	4	4	4	4

Table II shows cross-group validation errors in RO frequency predictions using the proposed physical subspace projection method, with different mixtures of device- and RO-array measurements. We use RO frequency as the performance of interest to mimic the operation of a digital system. As we compare cross-group prediction errors, we observe smaller prediction errors as the number of different measurement groups grows. Fig. 6 shows relative error on group #6 frequency predictions as a function of N_i replicas on the same die. All 3186 dies are used for training and the number of samples per die varies. Our proposed "VS+MAP" method is able to achieve higher accuracy using multiple-group measurements compared with response surface method (RSM) using single group measurements. We observe consistently 2x sample accuracy improvement over sample mean, which represents a 32.5x sample size reduction.

Fig. 7 shows how the average performance prediction varies with the number of training dies. All samples on training dies are used but the number of training dies varies in this case. Compared with standard "PCA+LSR" and "PCA+LAR" method, our proposed "VS+MAP" method is able to achieve substantially higher accuracy, with 70x and 150x sample reduction, respectively. Here we split the data set into 27-folds while dies on the same wafer remain in the same fold. Each time we select one or several folds to train the model and use the rest to do validation. The whole process is repeated and the prediction error is averaged.

VI. Conclusion

In this paper, we have propose a novel performance estimation algorithm through a physical subspace projection and

TABLE II
RELATIVE PREDICTION ERROR FOR CROSS-GROUP VALIDATION.

#	Measurement group	Prediction group	%error
1	1,2,3	4	3.22
2	1,2,3	5	2.99
3	1,2,3	6	2.7
4	5,6	4	2.4
5	4,6	5	2.19
6	5,6	4	3.54
7	1,2,3,4	5	2.26
8	1,2,3,4	6	2.17
9	1,2,3,5	5	2.26
10	1,2,3,5	6	2.06
11	1,2,3,6	4	2.32
12	1,2,3,6	5	2.15
13	1,2,3,4,5	6	2.1
14	1,2,3,4,6	5	1.98
15	1,2,3,5,6	4	2.01

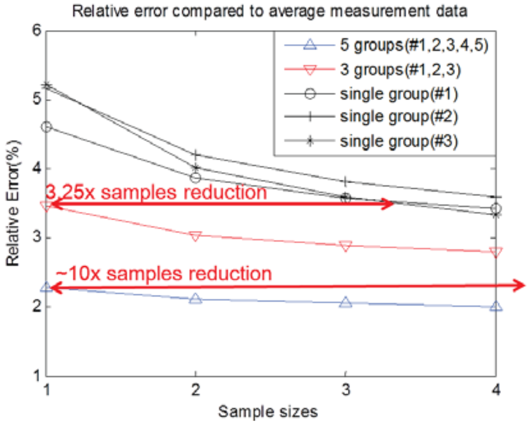

Fig. 6. Relative prediction error for group #6 versus replicate samples per die. A mixture of measurement groups is compared.

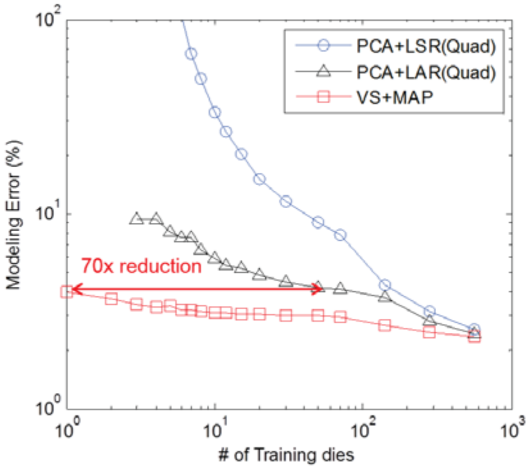

Fig. 7. Relative prediction error for group #6 versus number of training dies. Various algorithms are compared.

maximum a posteriori estimation with very small sample size. The key idea in this work is to exploit the fact that simulation and measurement data are physically correlated under different circuit configurations and topologies. First, different population of measurements are projected to a subspace expanded by a set of physical variables. The projection is achieved by performing a sensitivity analysis of measurement parameters on subspace variables using virtual source compact model. Then we develop a Bayesian treatment by introducing prior distributions over these projected variables. Maximum a posteriori estimation is also applied using the prior. The proposed method is validated by post-silicon measurement of a commercial 28-nm process. A superior sample size reduction is shown in two aspects. First, an average prediction error is reduced by a factor of 2 which can be translated to a 32x reduction on data needed for samples on the same die. Second, a 150x and 70x sample size reduction on training dies is achieved compared to traditional least-square fitting method and least-angle regression method, respectively, without surrendering any accuracy.

ACKNOWLEDGMENT

This work was funded by the Cooperative Agreement between the Masdar Institute of Science and Technology (Masdar Institute), Abu Dhabi, UAE and the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, Reference 196F/002/707/102f/70/9374.

REFERENCES

[1] D. Boning, K. Balakrishnan, H. Cai, N. Drego, A. Farahanchi, K. Gettings, D. Lim, A. Somani, H. Taylor, D. Truque, and X. Xie. Variation. In *International Symposium on Quality Electronic Design (ISQED)*, pages 15 – 20, Mar. 2007.

[2] X. Li, J. Le, L.T. Pileggi, and A. Strojwas. Projection-based performance modeling for inter/intra-die variations. In *International Conference on Computer-Aided Design (ICCAD)*, pages 721–727, 2005.

[3] I.T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

[4] X. Li. Finding deterministic solution from underdetermined equation: Large-scale performance modeling by least angle regression. In *Design Automation Conference*, pages 364–369, 2009.

[5] W. Zhang, T. Chen, M. Ting, and X. Li. Toward efficient large-scale performance modeling of integrated circuits via multi-mode/multi-corner sparse regression. In *Design Automation Conference (DAC)*, pages 897–902, 2010.

[6] F. Wang, W. Zhang, S. Sun, X. Li, and C. Gu. Bayesian model fusion: large-scale performance modeling of analog and mixed-signal circuits by reusing early-stage data. In *Design Automation Conference (DAC)*, pages 64:1–64:6, 2013.

[7] S. Saxena, C. Hess, H. Karbasi, A. Rossoni, S. Tonello, P. McNamara, S. Lucherini, S. Minehane, C. Dolainsky, and M. Quarantelli. Variation in transistor performance and leakage in nanometer-scale technologies. *IEEE Transactions on Electron Devices*, 55(1):131–144, 2008.

[8] D. Boning, J. Panganiban, K. Gonzalez-Valentin, S. Nassif, C. McDowell, A. Gattiker, and F. Liu. Test structures for delay variability. In *Timing Issues in the Specification and Synthesis of Digital Systems*, page 109, 2002.

[9] C. Gu, E. Chiprout, and X. Li. Efficient moment estimation with extremely small sample size via bayesian inference for analog/mixed-signal validation. In *Design Automation Conference (DAC)*, pages 65:1–65:7, 2013.

[10] A. Khakifirooz, O.M. Nayfeh, and D. Antoniadis. A simple semiempirical short-channel MOSFET current-voltage model continuous across all regions of operation and employing only physical parameters. *IEEE Transactions on Electron Devices*, 56(8):1674 – 1680, Aug. 2009.

[11] L. Wei, O. Mysore, and D. Antoniadis. Virtual-source-based self-consistent current and charge FET models: From ballistic to drift-diffusion velocity-saturation operation. *IEEE Transactions on Electron Devices*, (99):1 – 9, 2012.

[12] L. Yu, O. Mysore, Lan Wei, L. Daniel, D. Antoniadis, I. Elfadel, and D. Boning. An ultra-compact virtual source FET model for deeply-scaled devices: Parameter extraction and validation for standard cell libraries and digital circuits. In *Asia and South Pacific Design Automation Conference (ASPDAC)*, pages 521–526, 2013.

[13] L. Yu, L. Wei, D. Antoniadis, I. Elfadel, and D. Boning. Statistical modeling with the virtual source MOSFET model. In *Design, Automation Test in Europe Conference Exhibition (DATE)*, pages 1454–1457, 2013.

[14] A. Ortiz-Conde, F.J. Garca Snchez, J.J. Liou, A. Cerdeira, M. Estrada, and Y. Yue. A review of recent MOSFET threshold voltage extraction methods. *Microelectronics Reliability*, 42(45):583 – 596, 2002.

[15] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society, Series B*, 39(1):1–38, 1977.