Efficient Moment Estimation with Extremely Small Sample Size via Bayesian Inference for Analog/Mixed-Signal Validation

Chenjie Gu
Intel Strategic CAD Labs
chenjie.gu@intel.com

Eli Chiprout
Intel Strategic CAD Labs
eli.chiprout@intel.com

Xin Li
Carnegie Mellon University
xinli@cmu.edu

ABSTRACT

A critical problem in pre-Silicon and post-Silicon validation of analog/mixed-signal circuits is to estimate the distribution of circuit performances, from which the probability of failure and parametric yield can be estimated at all circuit configurations and corners. With extremely small sample size, traditional estimators are only capable of achieving a very low confidence level, leading to either over-validation or under-validation. In this paper, we propose a multi-population moment estimation method that significantly improves estimation accuracy under small sample size. In fact, the proposed estimator is theoretically guaranteed to outperform usual moment estimators. The key idea is to exploit the fact that simulation and measurement data collected under different circuit configurations and corners can be correlated, and are conditionally independent. We exploit such correlation among different populations by employing a Bayesian framework, *i.e.*, by learning a prior distribution and applying maximum a posteriori estimation using the prior. We apply the proposed method to several datasets including post-silicon measurements of a commercial high-speed I/O link, and demonstrate an average error reduction of up to 2×, which can be equivalently translated to significant reduction of validation time and cost.

1. INTRODUCTION

In various product validation disciplines (*e.g.*, pre-Silicon simulation-based validation, post-Silicon measurement-based validation), it is critical to make statistically valid predictions of circuit performances. This requirement boils down to the problem of estimating the probability distribution of circuit performance metrics of interest. From this distribution, we may also compute the probability of failure (PoF) or yield. The common practice is to estimate the moments of a distribution. In particular, if the distribution of performance metrics is Gaussian, the distribution is fully characterized by its first two moments, *i.e.*, mean and variance. With abundant data, sample mean and sample variance converge to the actual mean and variance, as guaranteed by the law of large numbers and central limit theorem [1].

However, in practice, simulation and measurement are both time and cost consuming [2, 3, 4]. For example, post-layout simulation can be slow, especially for circuits such as SRAM/PLL where extremely small time steps are required for high accuracy. As another example, during post-Silicon validation, due to the time-line of product releases, only a limited amount of measurement may be performed within the post-Silicon time-frame. In addition, the measurement of performance metrics, such as Bit-Error-Ratio and Time/Voltage Margins of high-speed I/O links, takes a long time, and requires expensive equipment (such as BER testers) [5, 6, 7].

Furthermore, for the validation of products, there are many corners and configurations to be covered. As an example, in I/O link validation, in addition to common process, voltage and temperature (PVT) corners, we must also validate against different board/add-in card configurations, input patterns, different equalization settings, *etc.*. Therefore, with the time and cost constraints of simulation and measurement, an extremely small number (1 to 5) of data is available at each corner or configuration.

We call the above problem the small-sample-size problem. This problem makes most statistical analysis tools/algorithms not applicable because they are built upon the assumption that "enough" data is available for valid statistical estimation. When the assumption is broken, we obtain low confidence in the estimated quantities. In another word, this means that we may either under-validate or over-validate the circuit. Similar to over-design and under-design, over-validation and under-validation are as harmful, if not more, in terms of cost and time-to-market. Unfortunately, there is few existing satisfying solution to get around this problem. To the best of our knowledge, the usual practice is to increase the sample size as much as possible to reach a certain confidence level, or to set an empirical guard-band on top of the estimation. There is a recent work [8] that considers a similar problem, but for performance modeling. Another recently published technique [9] solves a similar problem for post-layout performance distribution estimation, but with mildly small number of samples (50 or more).

It is also important to point out that in many situations, it is necessary to estimate the distribution at each corner/configuration, for which we only have 1 to 5 samples. For example, to validate I/O interfaces such as PCIE[10] and DDR[11], it is critical to make sure that the interface works properly with different boards and add-in cards/DIMMs.

DAC'13, May 29 – June 07 2013, Austin, TX, USA.

.

Therefore, for each board and add-in card, the distribution of the BER must be estimated separately. It is inappropriate to mix the measurements under different configurations, because even with a low overall PoF, we may obtain a very high PoF at a particular configuration. In this case, combining data from all configurations does not help us to increase the sample size. In fact, estimating the overall distribution can lead to misleading validation results.

In this paper, we propose a technique to efficiently estimate the mean and standard deviation of circuit performance distributions under the small-sample-size constraint. The key idea of the method is to exploit correlation in data collected at multiple populations to improve the accuracy of the proposed estimator. In particular, we emphasize that data collected at different design stages, different configurations and different corners are not independent, but are correlated. Taking advantage of this non-intuitive fact leads to a theoretically guaranteed better estimator. In comparison to sample mean/standard deviation estimators, our method achieves an average error reduction of up to 2×, for examples obtained from measurement of commercial designs.

Mathematically, we employ Bayesian inference[12] to fuse conditionally independent data. The method is composed of two steps. First, the Maximum Likelihood (ML) method is used to learn a prior distribution of mean/standard deviation from data collected at multiple populations. Second, the prior learned in the first step is used to obtain the Maximum A Posteriori (MAP) estimation of mean and standard deviation. The two steps are formulated as two optimization problems. Based on this formulation, we further propose a relaxed algorithm, to alleviate the computational burden.

While this paper focuses on derivations for the mean and standard deviation estimation, our formulation is general, and it incorporates estimation of moments of any order. Our formulation is also general to cover many application scenarios, depending on the availability of different data sets. In particular, the following two scenarios are commonly seen in practice:

1. Given early stage data, or empirical results, estimate the mean and standard deviation at a targeted configuration.
2. Given data at multiple configurations, estimate the mean and standard deviation at each configuration.

The rest of paper is organized as follows. Sec. 2 formulates the problem and describes the small-sample-size problem. Sec. 3 discusses and derives the multi-population estimation algorithm based on the Bayesian framework. Sec. 4 presents experimental results on several datasets to demonstrate the advantages of the proposed method.

2. BACKGROUND AND PROBLEM FORMULATION

For simplicity, in this paper, we consider the problem of estimating a single performance metric, denoted by x, which depends on many parameters such as process parameters, voltage, temperature, board, add-in card, *etc.*. The performance metric x also depends (indirectly) on time, because a subset of the parameters, such as process parameters, also change over time.

As an example application, we consider the problem of post-Silicon validation of I/O interfaces. In this application, a configuration is defined by fixing the values of a subset of the parameters. By considering variability of all other parameters, x has a distribution at each configuration. For example, a configuration of an I/O link can be defined by the combination of a specific board and a specific add-in card. The variability of time/voltage margin (of the eye diagram) is caused by parameter variations such as PVT variations. Measurement of margins is repeated at each configuration for each Silicon stepping, and the goal of validation is to ensure that PoF meets the specification at each stepping and at each configuration.

2.1 Problem Formulation

To formalize the above description, we define a *population* to be a specific (corner, configuration, stepping) combination, and suppose that there are P populations. For each population, we define a random variable x_i, $(i = 1, \cdots, P)$ to model the variability of the performance metric at the corresponding (corner, configuration, stepping) combination, and x_i satisfies a Gaussian distribution $x_i \sim N(\mu_i, \sigma_i^2)$ where μ_i is the mean and σ_i^2 is the variance. For notational convenience, we define $\boldsymbol{\mu} = [\mu_1, \cdots, \mu_P]^T$ and $\boldsymbol{\sigma} = [\sigma_1 \cdots, \sigma_P]^T$.

For each population, we obtain a set of independent observations $\mathcal{X}_i = \{x_{i,1}, \cdots, x_{i,N_i}\}$, where N_i is the sample size of the i-th population. (For simplicity, we consider the case where $N_1 = \cdots = N_P = N$ throughout the paper. Extension to the more general case is straightforward. Each element in $\mathcal{X}_i$ corresponds to one independent measurement at the i-th population. The problem we aim to address is to estimate μ_i's and σ_i's given the observations $\{\mathcal{X}_1, \cdots, \mathcal{X}_P\}$, with the special constraint that N_i's are very small.

2.2 Low Confidence under Small Sample Size

For a specific population, the most widely used estimator for mean and variance is the sample mean $\bar{x}_i$ and sample variance S_i, respectively,

$$\bar{x}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{i,j}, \quad S_i = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(x_{i,j} - \bar{x}_i)^2. \qquad (1)$$

Since $\bar{x}_i \sim N(\mu_i, \frac{\sigma_i^2}{N_i})$ and $S_i \sim \frac{\sigma_i^2}{N_i - 1}\chi^2_{N_i - 1}$, we obtain

$$\mathrm{Std}(\bar{x}_i) = \frac{1}{\sqrt{N_i}}\sigma_i, \quad \mathrm{Std}(S_i) = \frac{\sqrt{2}}{\sqrt{N_i - 1}}\sigma_i^2. \qquad (2)$$

If the standard deviation of an unbiased estimator is used as a measure of accuracy and confidence level, Eqn. (2) shows that the accuracy of both sample mean and variance estimators depend on N_i. As N_i approaches infinity, the error converges to 0. However, when N_i is small, both estimators suffer from significant error.

3. MULTI-POPULATION MOMENT ESTIMATION

3.1 Overview

As is evident in Sec. 2.2, if each population is treated independently, there is little room for improvement. In contrast, our method views data at different populations as correlated, and it tries to exploit such correlation to improve the accuracy of the estimator.

Mathematically, out method employs a Bayesian framework, and consists of two steps, as shown in Fig. 1. First, it

learns a prior distribution of $p(\boldsymbol{\mu}, \boldsymbol{\sigma})$ from data at all populations, using maximum likelihood estimation. Second, it applies Maximum A Posteriori estimation to each population using the prior distribution learned from the first step.

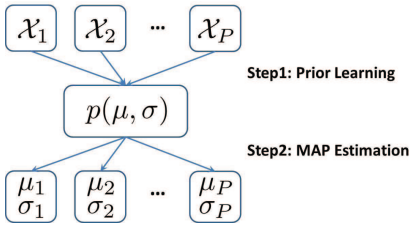


Figure 1: Proposed method consists of two steps.

To give an intuitive idea of why methods other than sample mean/variance can be much better, we consider two special examples. It can be shown that the estimators described in the two examples can be thought of as special cases of our proposed method.

EXAMPLE 3.1 (UNEQUAL MEAN, EQUAL VARIANCE). *Assume that μ_i's are different, and $\sigma_1 = \cdots = \sigma_P = \sigma$, and consider the problem of estimating σ^2. Since $S_i \sim \frac{\sigma^2}{N-1}\chi^2_{N-1}$, we obtain an unbiased estimator for σ^2*

$$\frac{1}{P}[S_1 + \cdots + S_P] \sim \frac{1}{P}\frac{\sigma^2}{N-1}\chi^2_{NP-P}, \tag{3}$$

from which $Std(\frac{1}{P}[S_1 + \cdots + S_P]) = \sigma^2\sqrt{\frac{2}{P(N-1)}}$. Hence, the estimation error decreases as P increases, and is smaller than $Std(S_i)$.

EXAMPLE 3.2 (EQUAL MEAN, UNEQUAL VARIANCE). *Assume that $\mu_1 = \cdots = \mu_P = \mu$, and σ_i's are different, and consider the problem of estimating μ. Since $\bar{x}_i \sim N(\mu, \frac{\sigma_i^2}{N})$, we obtain an unbiased estimator for μ*

$$\frac{1}{P}[\bar{x}_1 + \cdots + \bar{x}_P] \sim N(\mu, \frac{1}{P^2}[\frac{\sigma_1^2}{N} + \cdots + \frac{\sigma_P^2}{N}]). \tag{4}$$

As P increases, the variance of $\frac{1}{P}[\bar{x}_1 + \cdots + \bar{x}_P]$ decreases. This shows that when there are many populations, we may achieve a very accurate estimate of μ.

Note, however, Eqn. (4) is not the "best" estimator. Intuitively, consider an estimator of μ which is a linear combination of $\bar{x}_i$'s. Then, more weight should be given to $\bar{x}_i$ if σ_i is smaller. However, we omit the derivation since the actual expression is rather involved.

3.2 Choice of Prior Distributions

Intuitively, prior distributions for μ_i's and σ_i's, denoted by $p(\mu_i)$ and $p(\sigma_i)$ respectively, describe our *belief* about the correlation among μ_i's and σ_i's. We stress that μ_i's and σ_i's are fixed quantities at each population, and we simply model the variation across populations by imposing a prior distribution. In Example 3.1, $\sigma_1 = \cdots = \sigma_P$ corresponds to a Dirac distribution $p(\sigma_i) = \delta(\sigma_i - \sigma)$. In Example 3.2, $\mu_1 = \cdots = \mu_P$ corresponds to a Dirac distribution $p(\mu_i) = \delta(\mu_i - \mu)$. In a real application, however, it is too strong to claim a priori that μ_i's and σ_i's at all populations are the same.

Rather, it is often seen that μ_i's and σ_i's at different populations are similar, but not equal. This observation makes a lot of sense, especially for circuits designed to account for variability. For example, many circuits have compensation loops and self-reconfigurable features that cancel out the effects due to certain variability, which effectively pushes μ_i's towards each other. On the other hand, the variance in the circuit performance is usually caused by a small set of parameters (such as critical process parameters, temperature, voltage), and the dependency at different configurations tends to be similar, which effectively pushes σ_i towards each other.

Based on the above observation, we choose to use uniform priors for μ_i's and σ_i's, *i.e.*,

$$\mu_i \sim U(a, b), \quad \sigma_i \sim U(c, d), \tag{5}$$

where $a, b, c, d \in \mathbf{R}, c, d \geq 0$. Note that the Dirac prior is an extreme case of the uniform prior when $|b - a|$ and $|d - c|$ become 0.

While the derivations in the rest of the paper will be based on the choice of uniform prior, there is no limitation to incorporate other types of prior distributions in our method in Fig. 1. For example, one may use a Gaussian prior, or even any arbitrary probability distribution. Furthermore, even μ_i and σ_i can be correlated, in which case we may define an arbitrary joint distribution $p(\boldsymbol{\mu}, \boldsymbol{\sigma})$ as the prior distribution. However, in order for the method to work well, the prior distribution must roughly reflect the relationships of μ_i's and σ_i's in reality.

We stress again that although we apply a prior distribution to $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, μ_i's and σ_i's are fixed quantities for each population, rather than random variables. The prior distribution is simply a statistical tool to describe how μ_i's and σ_i's are correlated.

It will be shown later that using a uniform prior distribution effectively applies a bound on the estimated quantities. Therefore, the process of learning a uniform prior can be thought of obtaining a bound on the quantities to be estimated, and the probability distribution can be thought of as a mathematical tool to model correlation.

3.3 Learning a Prior Distribution

The first step in our method is to learn a prior distribution from data collected at all populations. We employ the maximum likelihood approach to learn the prior $p(\mu_i, \sigma_i|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are *hyper-parameters* of the prior distribution. This problem can be formulated as an optimization problem

$$\underset{\boldsymbol{\theta}}{\text{maximize}} \quad p(\mathcal{X}_1, \cdots, \mathcal{X}_P|\boldsymbol{\theta}), \tag{6}$$

where $p(\mathcal{X}_1, \cdots \mathcal{X}_P|\boldsymbol{\theta})$ is the likelihood function. We may either use a nonlinear optimizer to solve for the optimal $\boldsymbol{\theta}$, or we may derive closed-form solutions by solving

$$\frac{d}{d\boldsymbol{\theta}} p(\mathcal{X}_1, \cdots, \mathcal{X}_P|\boldsymbol{\theta}) = 0. \tag{7}$$

In our formulation, the likelihood function can be com-

puted by

$$\begin{aligned} & p(\mathcal{X}_1, \cdots, \mathcal{X}_P|\boldsymbol{\theta}) \\ =& \int_{\boldsymbol{\mu},\boldsymbol{\sigma}} p(\mathcal{X}_1, \cdots, \mathcal{X}_P|\boldsymbol{\mu},\boldsymbol{\sigma}) p(\boldsymbol{\mu},\boldsymbol{\sigma}|\boldsymbol{\theta}) d\boldsymbol{\mu} d\boldsymbol{\sigma} \\ =& \int_{\boldsymbol{\mu},\boldsymbol{\sigma}} \left(\prod_{i=1}^{P} p(\mathcal{X}_i|\mu_i,\sigma_i) \right) \left(\prod_{i=1}^{P} p(\mu_i,\sigma_i|\boldsymbol{\theta}) \right) d\boldsymbol{\mu} d\boldsymbol{\sigma} \\ =& \prod_{i=1}^{P} \int_{\mu_i,\sigma_i} p(\mathcal{X}_i|\mu_i,\sigma_i) p(\mu_i,\sigma_i|\boldsymbol{\theta}) d\mu_i d\sigma_i, \end{aligned} \tag{8}$$

where the second equality is due to two conditional independences, $(\mathcal{X}_1 \perp \cdots \perp \mathcal{X}_P|\boldsymbol{\mu},\boldsymbol{\sigma})$[1] and $(\{\mu_1,\sigma_1\} \perp \cdots \perp \{\mu_P,\sigma_P\}|\boldsymbol{\theta})$. The integral Eqn. (8) can be computed by numerical integration, or we may derive its closed-form expression for special prior distributions.

In our formulation, we choose a uniform prior on μ_i's, as defined in Eqn. (5). To write it in the form of $p(\mu_i,\sigma_i|\boldsymbol{\theta})$, we define $\boldsymbol{\theta} = [a,b,c,d]^T$. We further assume that μ_i and σ_i are independent given $\boldsymbol{\theta}$.

Therefore, $p(\mu_i,\sigma_i|\boldsymbol{\theta}) = p(\mu_i|a,b)p(\sigma_i|c,d)$, *i.e.*,

$$p(\mu_i,\sigma_i|\boldsymbol{\theta}) = \begin{cases} \frac{1}{b-a}\frac{1}{d-c}, & \text{if } a \le \mu_i \le b, c \le \sigma_i \le d, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

For each population, x_i satisfies the Gaussian distribution $N(\mu_i,\sigma_i^2)$, and therefore

$$p(\mathcal{X}_i|\mu_i,\sigma_i) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{ -\frac{1}{2} \frac{(x_{i,j}-\mu_i)^2}{\sigma_i^2} \right\}. \tag{10}$$

Inserting Eqn. (9) and Eqn. (10) into Eqn. (8), we need to compute for each i,

$$\int_c^d p(\sigma_i|c,d) d\sigma_i \int_a^b p(\mu_i|a,b) d\mu_i p(\mathcal{X}_i|\mu_i,\sigma_i), \tag{11}$$

where the integral with respect to μ_i is

$$\begin{aligned} & \int_a^b p(\mu_i|a,b) d\mu_i p(\mathcal{X}_i|\mu_i,\sigma_i) \\ =& \frac{1}{Z} \left\{ \Phi(\frac{b-\bar{x}_i}{\sigma/\sqrt{N_i}}) - \Phi(\frac{a-\bar{x}_i}{\sigma/\sqrt{N_i}}) \right\}, \end{aligned} \tag{12}$$

where Z is a normalizing constant, and $\Phi(\cdot)$ is the CDF of the standard normal distribution. (Unfortunately, the integral in terms of σ_i is more involved, and we compute it by numerical methods.) Eqn. (11) can then be inserted into Eqn. (8) to compute the likelihood function.

3.4 Maximum A Posteriori Estimation of μ and $\boldsymbol{\sigma}$

Once the prior $p(\mu_i,\sigma_i|\boldsymbol{\theta})$ is learned, MAP estimation can be applied to obtain a point estimate of μ_i's and σ_i's. MAP formulation searches for the values of μ_i's and σ_i's that maximize the posterior distribution, *i.e.*, it solves

$$\underset{\mu_i,\sigma_i}{\text{maximize}} \quad p(\mu_i,\sigma_i|\mathcal{X}_i). \tag{13}$$

According to Bayes' rule,

$$p(\mu_i,\sigma_i|\mathcal{X}_i) \propto p(\mathcal{X}_i|\mu_i,\sigma_i) p(\mu_i,\sigma_i), \tag{14}$$

where $p(\mathcal{X}_i|\mu_i,\sigma_i)$ is derived in Eqn. (10), and $p(\mu_i,\sigma_i)$ is learned as described in Sec. 3.3.

For uniform priors of μ_i and σ_i, the right-hand side of Eqn. (14) is

$$\frac{1}{b-a}\frac{1}{d-c} p(\mathcal{X}_i|\mu_i,\sigma_i), \quad \text{if } \mu_i \in [a,b] \text{ and } \sigma_i \in [c,d]. \tag{15}$$

Therefore, MAP is equivalent to maximum likelihood estimation on the support $\mu_i \in [a,b]$ and $\sigma_i \in [c,d]$. The solution is simply

$$\mu_{i,MAP} = \begin{cases} a & \text{if } \mu_{i,MLE} < a \\ \mu_{i,MLE} & \text{if } a \le \mu_{i,MLE} \le b \\ b & \text{if } \mu_{i,MLE} > b \end{cases}, \tag{16}$$

$$\sigma_{i,MAP} = \begin{cases} c & \text{if } \sigma_{i,MLE} < c \\ \sigma_{i,MLE} & \text{if } c \le \sigma_{i,MLE} \le d \\ d & \text{if } \sigma_{i,MLE} > d \end{cases}, \tag{17}$$

where $\mu_{i,MLE}$ and ${\sigma_{i,MLE}}^2$[2] are equal to the sample mean and standard deviation, respectively[1].

3.5 Algorithm and Relaxation

Summarizing Sec. 3.3 and Sec. 3.4, our proposed algorithm consists of two steps, as shown in Algorithm 1.

Algorithm 1 Multi-Population Moment Estimation

Given: $\mathcal{X}_1, \cdots, \mathcal{X}_P$.
Outputs: (μ_i,σ_i), $i = 1, \cdots, P$.
1: Solve $\underset{\boldsymbol{\theta}}{\text{maximize}}\ p(\mathcal{X}_1, \cdots, \mathcal{X}_P|\boldsymbol{\theta})$ (Eqn. (6)) for $\boldsymbol{\theta}$
2: **for** $i = 1 \to P$ **do**
3: Solve $\underset{\mu_i,\sigma_i}{\text{maximize}}\ p(\mu_i,\sigma_i|\mathcal{X}_i)$ (Eqn. (13)) for (μ_i,σ_i)
4: **end for**

As mentioned, the computation of the integral for σ_i in Eqn. (11) is quite involved. To migrate this problem, we may relax the optimization to an easier one

$$\underset{a,b}{\text{maximize}} \quad p(\mathcal{X}_1, \cdots, \mathcal{X}_P|a,b,\boldsymbol{\sigma}), \tag{18}$$

where $\boldsymbol{\sigma}$ is known, and only hyper-parameters (a,b) of the mean prior are searched for. However, in this formulation σ_i obtained in Eqn. (13) is dependent on μ_i (and hence (a,b)). The algorithm has to be modified to ensure that all the estimated quantities converge.

The modified algorithm with the above relaxation is shown in Algorithm 2. In step 1, we use sample mean/variance computed at each population as an initial guess for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, and the guess for (a,b) is computed by $a = \min(\mu_1, \cdots, \mu_P)$ and $b = \max(\mu_1, \cdots, \mu_P)$. Then we iteratively solve for (a,b), μ_i's and σ_i's until the convergence criteria in Algorithm 2 is satisfied.

3.6 Connections to Empirical Bayes Estimators

The ideas presented in this paper follow the philosophy of a class of Bayesian estimators, called *Empirical Bayes estimators* (EB)[13]. EB applies Bayes' rule to obtain either a point estimation or a posterior distribution of the parameters to be estimated. Unlike standard Bayesian methods that specify an arbitrary prior, EB learns the prior distribution from data. In particular, if a Gaussian prior is used for

[1]The notation $(A \perp B|C)$ means that A and B are conditionally independent given C.

[2]$\sigma_{i,MLE}$ is a biased estimator. To eliminate the bias, we may replace $\sigma_{i,MLE}$ in Eqn. (17) by its unbiased estimator.

Algorithm 2 Multi-Population Moment Estimation with Relaxation

Given: $\mathcal{X}_1, \cdots, \mathcal{X}_P$; ϵ (tolerance for convergence).
Outputs: (μ_i, σ_i), $i = 1, \cdots, P$.

1: Compute the initial guess for $a, b, \boldsymbol{\mu}, \boldsymbol{\sigma}$.
2: **repeat**
3: $a_{\text{old}} = a, b_{\text{old}} = b, \boldsymbol{\mu}_{\text{old}} = \boldsymbol{\mu}, \boldsymbol{\sigma}_{\text{old}} = \boldsymbol{\sigma}$.
4: Solve $\underset{a,b}{\text{maximize}}\ p(\mathcal{X}_1, \cdots, \mathcal{X}_P | a, b, \boldsymbol{\sigma})$ (Eqn. (18)) for (a, b)
5: **for** $i = 1 \to P$ **do**
6: Solve $\underset{\mu_i, \sigma_i}{\text{maximize}}\ p(\mu_i, \sigma_i | \mathcal{X}_i, a, b)$ (Eqn. (13)) for (μ_i, σ_i)
7: **end for**
8: **until** $|a - a_{\text{old}}|^2 + |b - b_{\text{old}}|^2 + ||\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{old}}||_2^2 + ||\boldsymbol{\sigma} - \boldsymbol{\sigma}_{\text{old}}||_2^2 < \epsilon$

the mean, EB gives the so-called *James-Stein estimator*[14] for the mean.

Particularly, a nice feature of the James-Stein estimator is that it is "superior" to the sample mean estimate, in the sense that the expected sum of mean square error of μ_i's at all populations is smaller than that of the sample mean estimator, *i.e.*

$$E\{\sum_{i=1}^{P}(\mu_i - \mu_i^{JS})^2\} < E\{\sum_{i=1}^{P}(\mu_i - \bar{x}_i)^2\}, \qquad (19)$$

where μ_i is the actual mean, μ_i^{JS} is the James-Stein estimator and $\bar{x}_i$ is the sample mean. One can show that if the Gaussian prior on μ_i's is used in our method, we obtain an estimator very similar to the James-Stein estimator, and Eqn. (19) still holds.

Unlike the James-Stein estimator, our method allows for more general prior distributions. Specifically, we have derived the case for uniform priors. We will show in Sec. 4 that our method can significantly out-perform sample mean/variance estimators.

3.7 Possible Limitations

Although our method may obtain a theoretically better overall estimate according to conclusions such as Eqn. (19), it can be the case, theoretically, that for a specific population, our method introduces a large bias.

As an extreme example, consider 100 populations, each with 1 observation, and $\mu_1 = \cdots = \mu_{99} = 0, \mu_{100} = 1$, $\sigma_1 = \cdots = \sigma_{100} = 1$.Effectively, our method will shrink the estimated mean towards 0. Therefore, for the 100-th population, the bias can be large.

However, due to the reasons mentioned in Sec. 3.2, such extremely pathological cases are unlikely to happen. Even if it happens, the outliers can be easily identified in a preprocessing step, and therefore accuracy will not be compromised by outliers.

3.8 Practical Implementation

It should be noted that the optimization problems in Algorithm 1 and Algorithm 2 may not be convex, and may have multiple local optimal points. There is no guarantee that our method will find the global optima. However, since initial guesses are estimated from the same data, the optimizer has a good guess start with, and is less affected by local optimal points.

To alleviate the computational cost associated with solving the optimization problems, we may impose an empirical prior distribution, instead of learning one from data. For example, experienced designers may have a good idea of the range of σ_i at each population – in this case, a uniform prior for σ_i's can be applied. However, empirical priors should be used with great caution, since it may incur unexpected bias. To be less biased, one may apply cross-validation [15] to check the validity of the empirical prior.

4. EXPERIMENTAL RESULTS

In this section, we apply the proposed method to two examples to show its accuracy and efficiency compared to sample mean/variance estimators. The first example is a set of artificial datasets to illustrate the advantage of our proposed method. The second example is a data set composed of time margin (eye width) measurements of a commercial high-speed I/O link. For notational convenience, we denote $\hat{\mu}_i$ and $\hat{\sigma}_i$ to be the estimation computed by our method, and $\mu_{i,\text{sample}}$ and $\sigma_{i,\text{sample}}$ to be the sample mean and sample standard deviation.

4.1 Illustrative Examples

In this example, we generate two sets of artificial data to illustrate the advantages of the proposed method in comparison with sample mean and variance estimators.

The data is generated as follows:

1. Choose a, b, c, d, P, N,
2. Randomly sample $\mu_i \sim U(a, b)$ and $\sigma_i \sim U(c, d)$ for $i = 1, \cdots, P$,
3. For each population, sample $x_{i,j} \sim N(\mu_i, \sigma_i)$ for $j = 1, \cdots, N$.

To compare $(\hat{\mu}_i, \hat{\sigma}_i)$ and $(\mu_{i,\text{sample}}, \sigma_{i,\text{sample}})$, we independently sample $\mathcal{X}_i$ (with μ_i's and σ_i's fixed) 500 times, and compute both estimators. We compare histograms of both mean/standard deviation estimators, as well as histograms of overall error for mean ϵ_μ and for standard deviation ϵ_σ, defined by

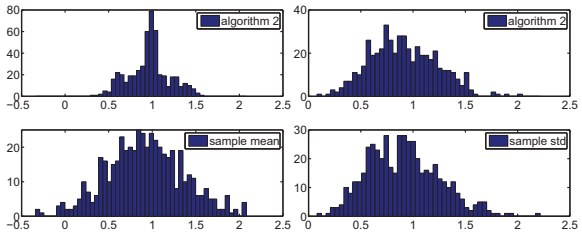
$$\epsilon_\mu = \sqrt{\frac{1}{P}\sum_{i=1}^{P}|\mu_i - \mu_{i,\text{est}}|^2}, \quad \epsilon_\sigma = \sqrt{\frac{1}{P}\sum_{i=1}^{P}|\sigma_i - \sigma_{i,\text{est}}|^2}. \qquad (20)$$

4.1.1 Dataset 1

In the first dataset, we set $a = 0.9$, $b = 1.1$, $c = 0.9$, $d = 1.1$, $N = 5$, $P = 20$ which corresponds to the scenario where μ_i's are similar, σ_i's are similar, and the standard deviation of μ_i's is comparable to the value of σ_i's. Fig. 2 show the histograms of the mean and standard deviation estimations for 500 repeats at one population where $\mu_i = 0.9067$, $\sigma_i = 0.9786$. It is seen that the variance of $\hat{\mu}$ is much smaller than that of μ_{sample}. By using a prior learned from multiple populations, we successfully reduce the variance of the estimator. On the other hand, $\hat{\sigma}$ is only slightly better than σ_{sample}. This is partly due to the fact that in Algorithm 2, the prior for σ_i's is removed, and therefore σ_i's are estimated separately. However, due to the accuracy improvement of μ_i, the accuracy of σ_i is also improved.

It should also be mentioned that we choose to show the histogram at this configuration because its μ_i is the smallest among 20 populations, and therefore is likely to be biased by

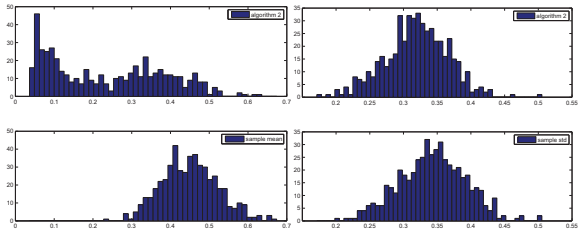
the prior distribution. However, the bias is almost negligible as can be seen from Fig. 2a.



(a) Mean. (b) Standard deviation.

Figure 2: Histograms of both estimators for population ($\mu_i = 0.9067, \sigma_i = 0.9786$).

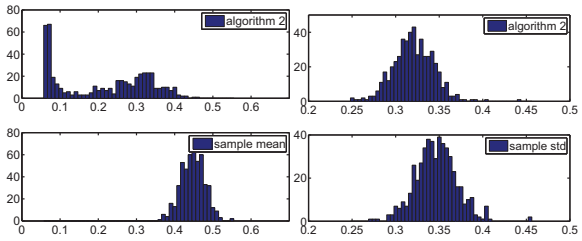
Fig. 3 shows histograms of ϵ_μ and ϵ_σ of both methods. From the histograms, we can compute $E(\epsilon_{\hat{\mu}}) = 0.2375$, $E(\epsilon_{\mu_{\text{sample}}}) = 0.4499$, $E(\epsilon_{\hat{\sigma}}) = 0.3191$ and $E(\epsilon_{\sigma_{\text{sample}}}) = 0.3431$. Hence, on average, our method achieves 2× accuracy improvement on $\boldsymbol{\mu}$. Moreover, the peak of ϵ_μ appears around 0.05, which implies that most likely, our method may achieve much more than 2× accuracy improvement.



(a) ϵ_μ. (b) ϵ_σ.

Figure 3: Histograms of ϵ_μ and ϵ_σ, $P = 20$.

We also study the effect of the number of population on the estimation error. Fig. 4 shows histograms for $P = 100$ (a, b, c, d, N are the same as in dataset 1). Similar to the previous case, we can compute $E(\epsilon_{\hat{\mu}}) = 0.2130$, $E(\epsilon_{\mu_{\text{sample}}}) = 0.4478$, $E(\epsilon_{\hat{\sigma}}) = 0.3206$ and $E(\epsilon_{\sigma_{\text{sample}}}) = 0.3467$. The average accuracy for $P = 100$ is slightly better than that of $P = 20$. However, if we compare Fig. 4 with Fig. 3, we observe that with P larger, the peak at $\epsilon_\mu \simeq 0.05$ is higher, and there is a much clearer separation between the histograms of two estimators. In particular, for mean estimation, our method almost dominates the sample mean estimator, *i.e.*, with high probability, our method obtains less error than the sample mean estimator.

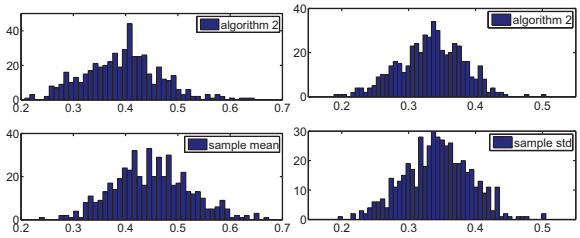


(a) ϵ_μ. (b) ϵ_σ.

Figure 4: Histograms of ϵ_μ and ϵ_σ, $P = 100$.

4.1.2 Dataset 2

In the second dataset, we set $a = 0$, $b = 2$, $c = 0.9$, $d = 1.1$, $N = 5$, $P = 20$. The major difference from the first dataset is that the standard deviation of the mean at all populations is much larger, and is comparable to the standard deviation for the Gaussian distributions. Intuitively, if $(b-a)$ is large, the data at different population is less correlated. However, for this dataset, our method still out-performs sample mean/standard deviation estimators. Fig. 5 shows the histogram of ϵ_μ and ϵ_σ of both methods, from which we can compute $E(\epsilon_{\hat{\mu}}) = 0.3936$, $E(\epsilon_{\mu_{\text{sample}}}) = 0.4499$, $E(\epsilon_{\hat{\sigma}}) = 0.3329$ and $E(\epsilon_{\sigma_{\text{sample}}}) = 0.3431$. Although the improvement is not as evident as dataset1, we still achieve a relative error improvement of 5% and 1% for mean and standard deviation, respectively. For $P = 100$, we observe similar trends as in dataset1.



(a) ϵ_μ. (b) ϵ_σ.

Figure 5: Histograms of ϵ_μ and ϵ_σ, $P = 20$.

4.2 Validation of High-Speed I/O Links

In I/O link validation, one critical performance metric is Bit-Error-Ratio (BER). For the state-of-the-art high-speed links, the BER is extremely small. For example, in the latest PCIE specification [10], $\text{BER}_{\text{spec}} = 10^{-12}$ with 8Gb/sec data rate. This makes BER measurement a very time-consuming process. An alternative is to measure the eye width and eye height (*a.k.a.*, *time margin* (TM) and *voltage margin* (VM), respectively) of the eye diagram at the receiver, which can be converted to BER under reasonable assumptions. Margin measurement, although much faster than direct BER measurement, is still expensive in terms of time and cost. For a limited time period, only a small number of data can be measured for each configuration.

In this example, we have measured the time margin of 50 dies (randomly sampled) for 8 different configurations. (Note that we measured 50 dies simply for the purpose of validating our algorithm.) The mean and standard deviation at different configurations are shown in Fig. 6. We have also observed from the histogram that the distribution of time margin can be well approximated by Gaussian distributions.

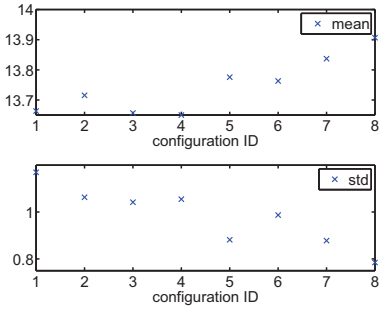


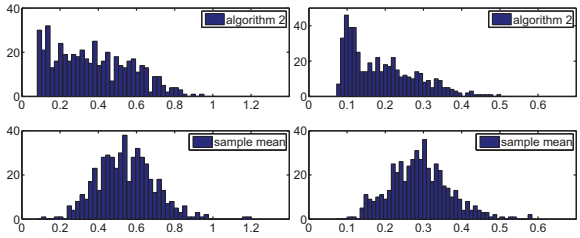
Figure 6: Mean and standard deviation at 8 configurations.

To compare the results of our method and sample mean/standard deviation estimators, we sample N_i data for each configuration from a distribution fitted from the 50 measurements,

and apply both methods. We repeat this experiment for 500 times, and compare the statistics of ϵ_μ and ϵ_σ.

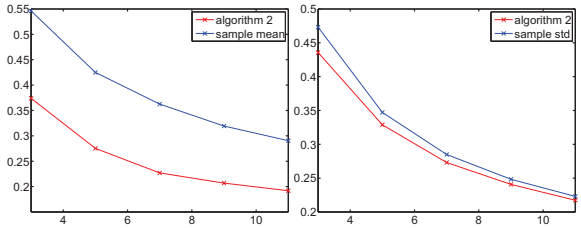
The histogram of ϵ_μ for $N_i = 3$ is shown in Fig. 7a. Similar to dataset1, our method out-performs the sample mean estimator. We also study how ϵ_μ is affected by the sample size at each configuration. Fig. 7b shows the histogram of ϵ_μ for both methods for $N_i = 11$. With N_i larger, the accuracy of both $\hat{\mu}$ and μ_{sample} is improved, and ϵ_μ of $\hat{\mu}$ is peaked around 0.1.

The trend with respect to N_i can be better illustrated by Fig. 8 which shows ϵ_μ and ϵ_σ as a function of N_i. Particularly for ϵ_μ, we observe a consistently 1.5× accuracy improvement over sample mean estimator. It is worth mentioning that as N_i becomes very large, ϵ_μ of both $\hat{\mu}$ and μ_{sample} converge towards 0, and there is little advantage of applying our method. However, if N_i is small, our method is much more accurate.



(a) $N_i = 3$. (b) $N_i = 11$.

Figure 7: Histogram of ϵ_μ.



(a) ϵ_μ vs N_i. (b) ϵ_σ vs N_i.

Figure 8: ϵ_μ and ϵ_σ decreases as N_i increases.

5. CONCLUSION

In this paper, we have proposed an efficient method for mean/standard deviation estimation under extremely small sample size. This problem is commonly seen in practice, and directly affects the time and cost associated with both pre-silicon and post-silicon validation, especially for complex analog/mixed-signal circuits. The validation of our method on several datasets, including measurement of commercial I/O links, shows that our method is consistently better than the sample mean/standard deviation estimators, and can achieve up to 2× average accuracy improvement. Furthermore, the accuracy improvement can also be equivalently translated to a potentially large test/validation time reduction.

6. REFERENCES

[1] A. Papoulis and S. Pillai, *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Science/Engineering/Math, 2001.

[2] G. Balamurugan, B. Casper, J. Jaussi, M. Mansuri, F. O'Mahony, and J. Kennedy, "Modeling and Analysis of High-Speed I/O Links," *Advanced Packaging, IEEE Transactions on*, vol. 32, no. 2, pp. 237 –247, May 2009.

[3] J. Keshava, N. Hakim, and C. Prudvi, "Post-Silicon Validation Challenges: How EDA and Academia can Help," in *Design Automation Conference (DAC), 2010 47th ACM/IEEE*, June 2010, pp. 3 –7.

[4] C. Gu, "Challenges in Post-Silicon Validation of High-Speed I/O Links," in *Computer-Aided Design (ICCAD), 2012 IEEE/ACM International Conference on*. IEEE, 2012.

[5] Intel Corp., "Intel Platform and Component Validation." [Online]. Available: http://download.intel.com/design/chipsets/labtour/PVPT_WhitePaper.pdf

[6] E. E. Lior Shkolnitsky, "Electrical System-Validation Methodology for Embedded DisplayPort," June 2010. [Online]. Available: http://download.intel.com/design/intarch/PAPERS/323931.pdf

[7] K. Gambill, "System Margin Validation," December 2008. [Online]. Available: http://download.intel.com/design/intarch/papers/321078.pdf

[8] 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), 2010 47th ACM/IEEE*. IEEE, 2010, pp. 897–902.

[9] X. Li, W. Zhang, F. Wang, S. Sun, and C. Gu, "Efficient Parametric Yield Estimation of Analog/Mixed-Signal Circuits via Bayesian Model Fusion," in *Computer-Aided Design (ICCAD), 2012 IEEE/ACM International Conference on*. IEEE, 2012.

[10] [Online]. Available: http://www.pcisig.com

[11] [Online]. Available: http://www.jedec.org

[12] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, New York, 2006, vol. 4.

[13] G. Casella, "An Introduction to Empirical Bayes Data Analysis," *The American Statistician*, vol. 39, no. 2, pp. 83–87, 1985.

[14] B. Efron and C. Morris, "Data Analysis using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association*, vol. 70, no. 350, pp. 311–319, 1975.

[15] S. Arlot and A. Celisse, "A Survey of Cross-Validation Procedures for Model Selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.