Robust Spatial Correlation Extraction with Limited Sample via L1-Norm Penalty

M. Gao, Z. Ye, D. Zeng, Y. Wang and Z. Yu
Institute of Micro-Electronics, Tsinghua Univ.
Beijing, China

Outline

- **Introduction: Spatial Correlation in Random Process Variation**
- Technical Background: Kriging Model and Lasso in Linear Regression
- Our Method: L1-Penalty in Kriging Model
- Experimental Results & Conclusion

Breakdown of Process Variation

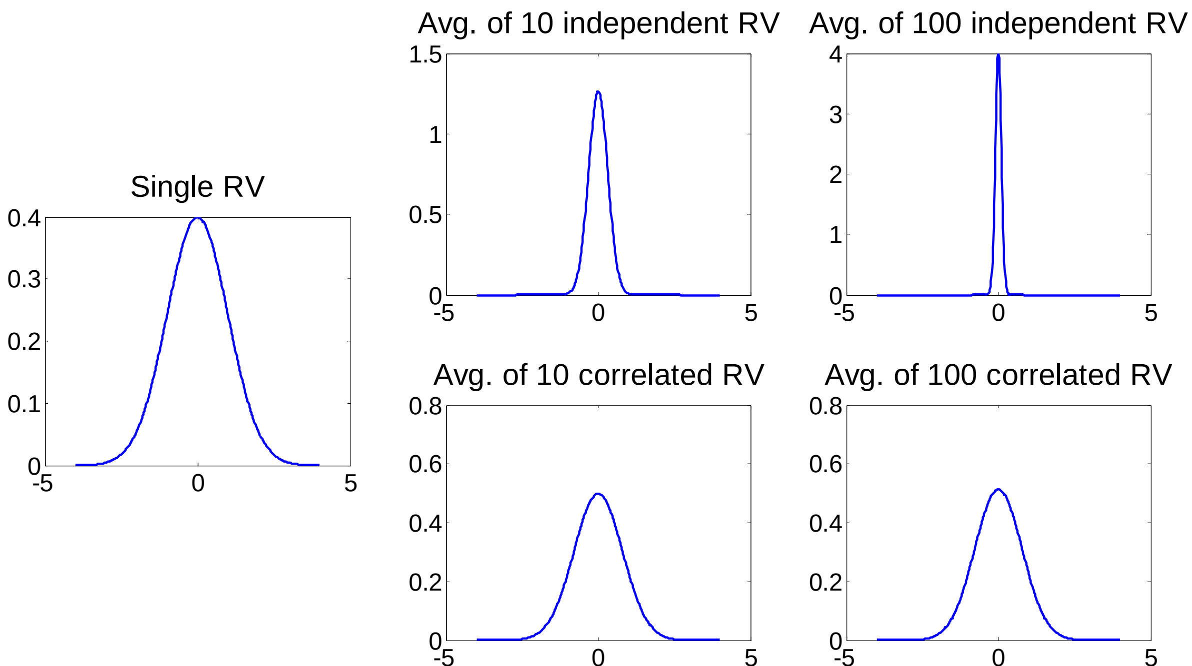
- Systematic:
  - Caused by LPE, MVP or Layout induced stress
  - Essentially predictable

- Purely Random:
  - Caused by some physical or process controlling limitation
  - RDE, LER, Litho, Annealing, …
  - Can occur in different scales

- The two parts are usually assigned to different research regime, since the underneath methodology is different

Correlation in Random Process Variation

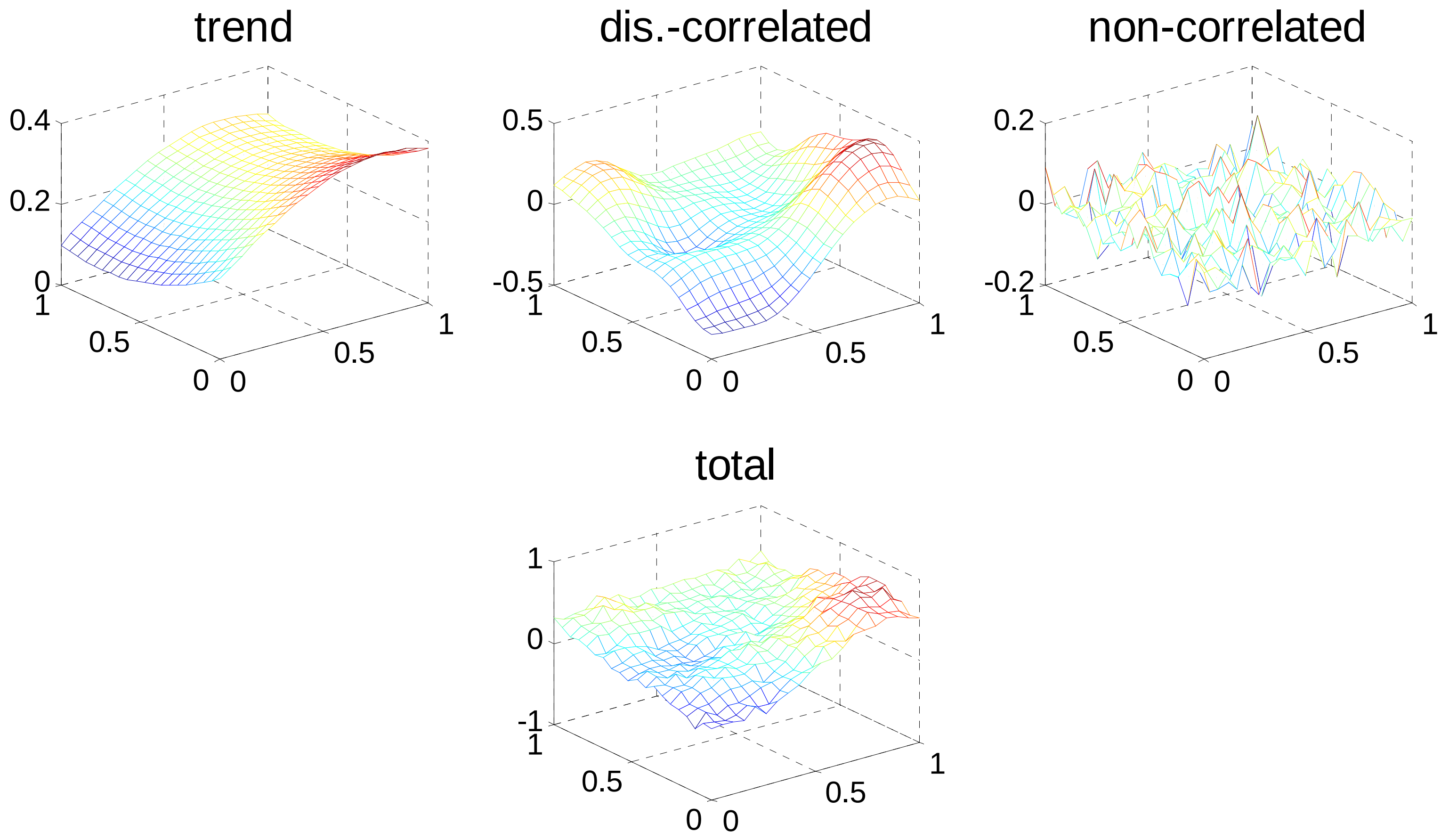
- One of the most nasty characteristics of Process Variation:
  - Usually a result of mixed causes
  - Can not be trivially studied or controlled
  - The killer of a handful of EDA methods



Properties of Correlation in Process Variation

- The correlations are spatial
  - It's up to the location of the device on the chip
  - Location-dependent, aka. trend:
    - due to long range effect (cross-die, cross rectile, cross wafer)
  - Distance-dependent, correlated to neighbors:
    - due to moderate range effect
  - Non-correlated:
    - due to short range effect

Breakdown of the Random Process Variation



The Importance of Correlation Study

1. It is not clear whether each of three types of correlation does play a role
   - To deny the ex. of any, we need solid method as well as solid data
2. The information on correlation help develop or ameliorate the process
3. The form of correlation affect EDA algorithms a lot

An Example: Correlation decides algorithm

- In statistical leakage analysis
  - There’s no correlation: the law of large number works, analytical results available (Rao04’)
  - The correlation are distance dependent: some clever methods exist based on the concepts of `kernal’, provably O(N) for a wide range of situations (Heloue06’, Ye09’)
  - Both location dependent and distance dependent: only some grid based (and PCA based) methods with some annoying limitations (strong correlation, say) can work to some extent (NOT perfectly addressed yet!!)

Outline

- Introduction: Spatial Correlation in Random Process Variation
- **Technical Background: Kriging Model and Lasso in Linear Regression**
- Our Method: L1-Penalty in Kriging Model
- Experimental Results & Conclusion

Mathematical Form of Correlation

$$Y(x) = F(x)\beta + S(x) + \varepsilon$$

$F(x)\beta$: the location dependent part

$F(x) = (f_1(x),\ f_2(x), \ldots,\ f_p(x))$: base of a function space, which is previously selected

β: the coefficients to be extracted

$\mathrm{S}(x)$: the distance dependent part, with stdev. σ

$\mathrm{Corr}(x_k, x_l) = \rho_\theta(|x_k - x_l|)$, ρ_θ is a family of parameterized functions, dependent on θ, aka correlogram

V : a correlation matrix induced by ρ_θ given the locations

ε : non-correlated part, can be absorbed by S(x) with an extra parameter added to θ

β, σ, θ are parameters to be extracted

Existing Extraction Methods

- Separate extraction:
  - A conventional method to first extract the loc. depend. part as if the residuals are not correlated
  - Mathematically not reasonable

- Kriging method:
  - Maximum likelihood method
  - Two-level optimization process

$$\min_{\beta,\sigma,\theta} g^0(\beta,\sigma,\theta;y,F) = 0.5\left\{\log|\sigma^2 V_\theta| + (y-F\beta)'(\sigma^2 V_\theta)^{-1}(y-F\beta)\right\}$$

y : the measured data at location x

The difficulties with kriging method

- There're not always a full bunch of data
  - Kriging works better only as the data capacity grows
  - Except for specific testing wafer used only for spatial correlation study, the data are much fewer
- The function base F(x) may be a large one
  - Functions should be added to this base for different reasons aiming at process ameiloration or sth. Else
  - Thus make the coefficients possibly sparse
- There's some interaction between the two parts of correlation
  - Physically or mathematically, they aren't strictly distinct

Lasso: add robustness to linear regression when sample is small

When the data set is small, the above method can be less robust.

Lasso proposes to solve the problem with additional L_1 – norm penalty

$$\min_{\beta} 0.5(y - F\beta)'(y - F\beta) + \lambda \sum_{i=1}^{p} |\beta_i|$$

Lasso, aka. compressive sensing, is famous for its non-zero pattern detection.

Actually, it also offers more robust result even the coefficient pattern is dense

Outline

- Introduction: Spatial Correlation in Random Process Variation
- Technical Background: Kriging Model and Lasso in Linear Regression
- **Our Method: L1-Penalty in Kriging Model**
- Experimental Results & Conclusion

Basic idea of the method

- The basic idea is simple. As lasso like L1-norm penalty achieves success in many fields other than linear regression, it may help in our situations.

- Among those numbers to be extracted in our model. We find it is more reasonable to add such penalty on β

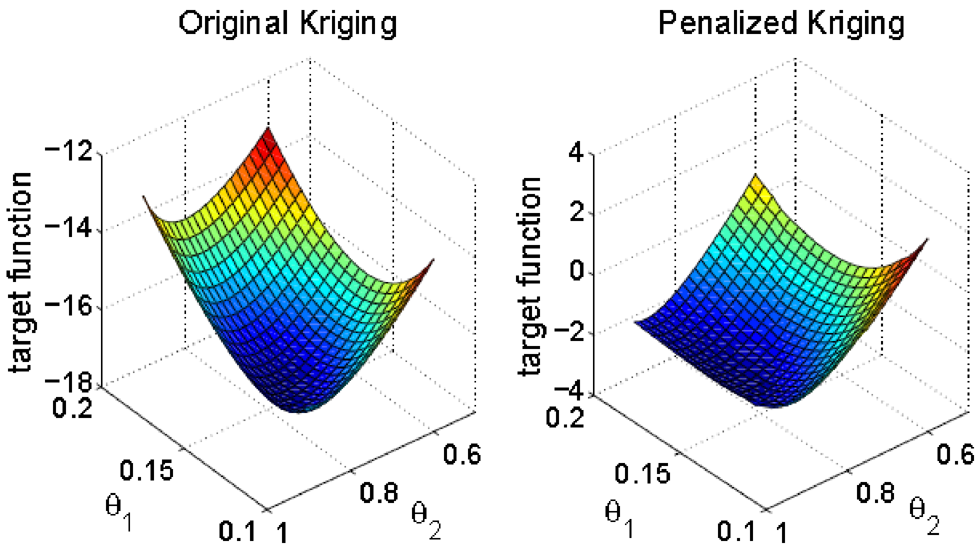
Mathematical Imitation

	Linear Regression	Spatial Correlated Data
Original Problem	$\min_{\beta} 0.5(y - F\beta)'(y - F\beta)$	$\min_{\beta,\sigma,\theta} g^{0}(\beta,\sigma,\theta; y, F) =$ $0.5\left\{\log \lvert \sigma^{2} V_{\theta} \rvert + (y - F\beta)'(\sigma^{2} V_{\theta})^{-1}(y - F\beta)\right\}$
Lasso-like L1-Norm Penalty	$\min_{\beta} 0.5(y - F\beta)'(y - F\beta)$ $+ \lambda \sum_{i=1}^{p} \lvert \beta_i \rvert$	$\min_{\beta,\sigma,\theta} g\ (\beta,\sigma,\theta; y, F, \lambda) = 0.5\{\log \lvert \sigma^{2} V_{\theta} \rvert +$ $(y - F\beta)'(\sigma^{2} V_{\theta})^{-1}(y - F\beta)\} + \lambda \sum_{i=1}^{p} \lvert \beta_i \rvert$

To solve the proposed optimization

- Though it's more complicated than original kriging methods, we still use a similar two-level optimization scheme
  - In the top level, an general purposed optimization is used for a problem with about 3 variables
  - In the bottom level, a problem similar to Lasso is to be solved

The solution space for the top level optimization



Solve Lasso: LAR

- Least Angle Regression (LAR) is a famous technique in statistics society as a power solution to Lasso.

- LAR develops some insights of the special form to be optimized.
  - It keeps going on a 'correct' trajectory, i.e. its temporary solutions are correct for some larger λ , until reaching the required λ
  - The direction for this process is piecewise-linear. Roughly speaking, it keeps going the same direction as long as the set of parameters triggering some 'profitable' condition is not changed.

- Those observations make it of a same level of complexity as OLS, thus enable its use as an inner-loop solver

Pick a right λ

- We choose Akaike Information Criterion (AIC)
  - The complexity of the model will reduce the robustness
  - AIC provides a rigorous way to balance between accuracy and complexity
- Other candidate methods:
  - Cross Validation
  - Another criterions such as BIC etc.

Outline

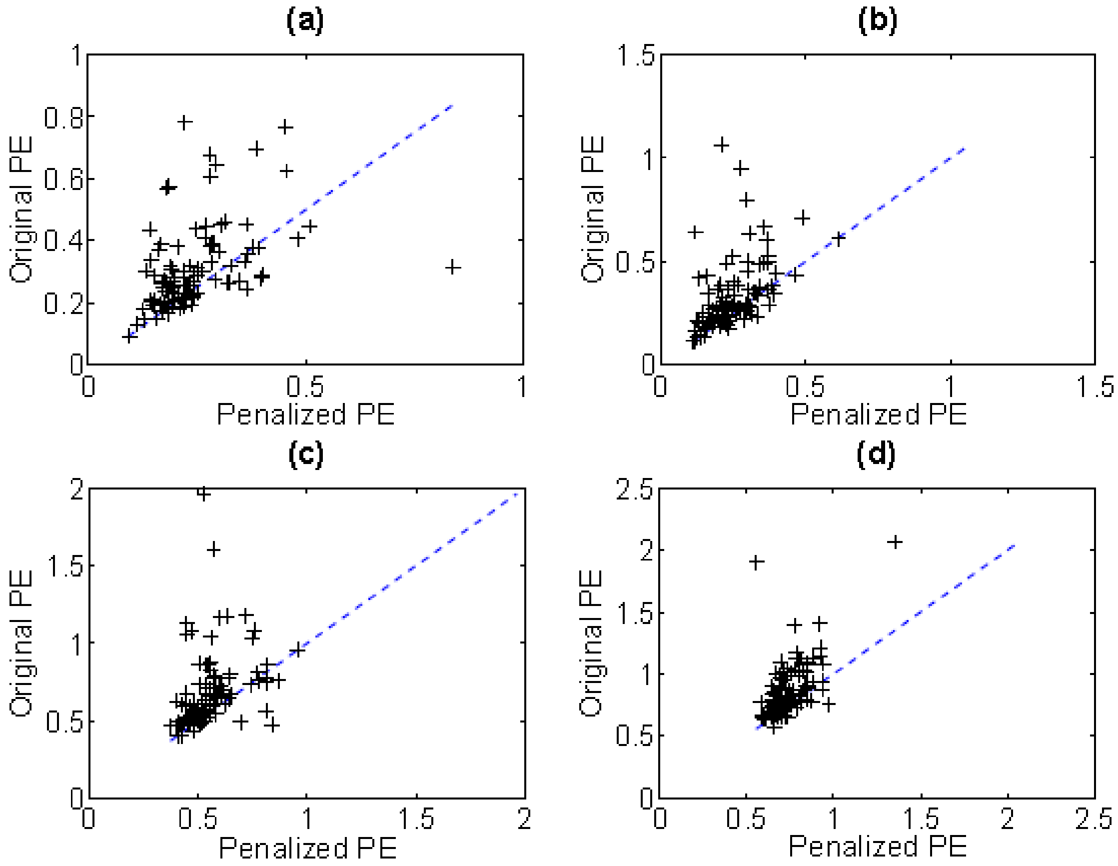
- Introduction: Spatial Correlation in Random Process Variation
- Technical Background: Kriging Model and Lasso in Linear Regression
- Our Method: L1-Penalty in Kriging Model
- **Experimental Results & Conclusion**

Setting of the Experiments

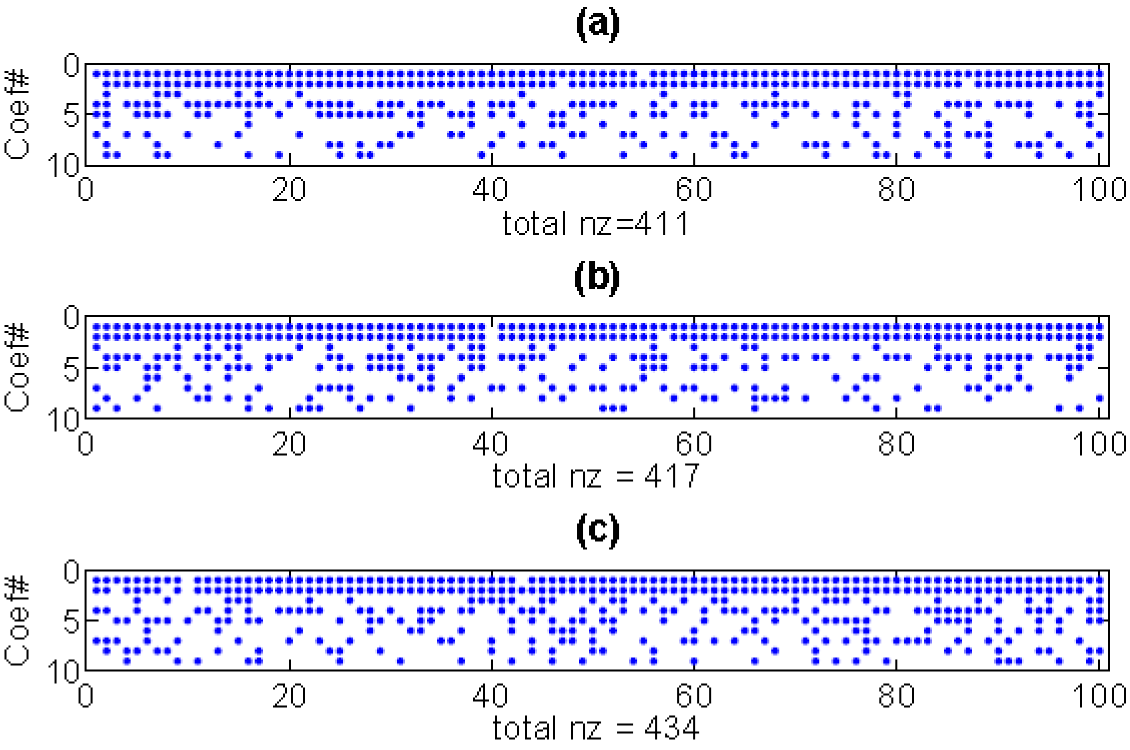
- Experiments are carried out with atificial data, we will publish the result with silicon data in the future
- We use 9 functions in the location dependent part and Gaussion family as the correlogram family, totally 9+3=12 parameters to extract
- Noisy data are generated with various settings, both sparse and dense loc. depend. coef.
- A sample of 50 data are assumed
- Results of proposed method, original kriging, separate extraction are compared together with the oracle model.
- The major concerned criteria are prediction error and the correlation length

Comparison with original kriging

The pairwise results on predicting error
with different settings



To detect the non-zero pattern



The 1st, 2nd , 4th coefficients are set as non-zero.

In most case the three coeficients are detected with about 1 other more in average

Conclusions

- The original kriging is modified to be most robust with Lasso-like L1-norm penalty.
- A solution flow, comprising of least angle regression together with criteria to pick proper weight factor of the L1-norm penalty, has been discussed in details.
- From numerical experiments, the L1-norm penalized kriging model shows improved accuracy and robustness in prediction. The results form a rigid base for applying the method to actual data.