IC Spatial Variation Modeling:
Algorithms and Applications

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Wangyang Zhang

B.S., Computer Science, Tsinghua University
M.S., Computer Science, Tsinghua University

Carnegie Mellon University
Pittsburgh, PA

September, 2012

Abstract

Rapidly improving the yield of today's complicated manufacturing process is a key challenge to ensure profitability for the IC industry. In this thesis, we propose accurate and efficient modeling techniques for spatial variation, which is becoming increasing important in the advanced technology nodes. Based on the spatial model, we develop algorithms for two applications that help identify the important yield-limiting factors and prioritize yield improvement efforts. *Variation decomposition* narrows down the sources of variation by decomposing the overall variation into multiple different components, each corresponding to a different subset of variation sources. *Wafer spatial signature clustering* automatically partitions a large number of wafers into groups exhibiting different spatial signatures, which helps process engineers find important factors that prevent the process from stably maintaining a high yield across different lots and wafers.

An important problem in variation decomposition is to accurately model and extract the wafer-level and within-die spatially correlated variation. Towards this goal, we first develop a physical basis function dictionary based on our study of several common physical variation sources. We further propose the DCT dictionary to discover spatially correlated systematic patterns not modeled by the physical dictionary. Moreover, we propose to apply sparse regression to significantly reduce the over-fitting problem posed by a large basis function dictionary. We further extend the sparse regression algorithm to a robust sparse regression algorithm for outlier detection, which provides superior accuracy compared to the traditional IQR method. Finally, we propose several efficient methods to make the computational cost of sparse regression tractable for large-scale problems.

We further develop an algorithm for the wafer spatial signature clustering problem based on three steps. First, we re-use the spatial variation modeling technique developed for variation decomposition to automatically capture the spatial signatures of wafers by a small number of features. Next, we select a complete-link hierarchical clustering algorithm to perform clustering on the features. Finally, we develop a modified L-method to select the number of clusters from the hierarchical clustering result.

Acknowledgement

First of all, I would like to thank my advisors, Prof. Xin Li and Prof. Rob Rutenbar for their continuous support for my Ph. D. study and research. They have provided me thorough academic training in self-motivated learning, problem solving, writing and presentation skills. Their work ethics and technical passion have greatly influenced me.

Each of my Ph. D. committee members deserves my special thanks not only for reviewing my proposal and thesis, but also for their significant contribution to the research presented in this thesis. I would like to thanks Prof. Shawn Blanton for being one of first to collaborate with us when we started to work on spatial variation modeling. I would like to thanks Prof. Andrzej Strojwas for providing a lot of helpful guidance throughout my Ph. D study and especially introducing the wafer spatial signature clustering application to me. I would like to thank Dr. Sharad Saxena for all the continued collaboration and helpful discussions since Fall 2011.

This work cannot be done without the collaboration and support from a lot of companies, including IBM, Intel, TI and PDF Solutions, and my sincere thanks goes to them. I would also like to thank Karthik Balakrishnan and Prof. Duane Boning from MIT for providing their variation characterization measurement data and continued discussion and support. I would like to acknowledge the financial support for this work from the C2S2 Focus Center and the National Science Foundation.

I would like to thank my colleagues during my internships for their collaboration on a number of exciting projects and for greatly broadening my views on the interesting problems in industry. My special thanks goes to Howard Chen and Ming Ting in Mentor Graphics, Amith Singhee, Jinjun Xiong, Peter Habitz, Amol Joshi, Chandu Visweswariah, James Sundquist and Bob Maier in IBM, and Xiaojing Yang, Rakesh Vallishayee and Sharad Saxena in PDF Solutions. I would like to also thank my group-mates and friends in the ECE department for their friendship and making my graduate life more enjoyable.

Last but not the least, I would like to thank my parents and my wife for their continuous encouragement and unconditional support over the past years.

Table of Contents

Chapter 1 Introduction 1

1.1 Process Variations 3

1.2 Variation Characterization 7

1.2.1 Test Structures for Variation Characterization 7

1.2.2 Statistical Analysis of Measurement Data 9

1.3 Spatial Variation Modeling 12

1.4 Thesis Contributions 14

1.5 Thesis Organization 15

Chapter 2 Variation Decomposition via Sparse Regression 18

2.1 Motivation 18

2.2 Dictionaries of Basis Functions 21

2.2.1 Physical Dictionary 22

2.2.2 DCT Dictionary 29

2.3 Basis Selection via Sparse Regression 34

2.4 Numerical Solver for Sparse Regression 38

2.4.1 Orthogonal Matching Pursuit (OMP) 39

2.4.2 Simultaneous Orthogonal Matching Pursuit (S-OMP) 43

2.4.3 Cross-Validation 46

2.5 Numerical Results 47

2.5.1 Quadratic Basis Effects 47

2.5.2 Center/Edge Effects 55

2.5.3 Complex Effects 67

2.6 Summary....73

Chapter 3 Robust Regression for Variation Decomposition....74

3.1 Motivation....74

3.2 Background....75

3.3 Robust Sparse Regression....78

3.4 Numerical Results....83

3.4.1 Comparison of Outlier Detection Methods....83

3.4.2 Results on Silicon Measurement Data....88

3.5 Summary....95

Chapter 4 Fast Implementation for Sparse Regression....97

4.1 Introduction....97

4.2 Inner Product Calculation....98

4.3 Least Squares Fitting....102

4.4 Numerical Results....106

4.5 Summary....108

Chapter 5 Wafer Spatial Signature Clustering....110

5.1 Introduction....110

5.2 Robust Feature Extraction....113

5.3 Clustering Algorithm....115

5.4 Cluster Selection....119

5.5 Numerical Results....124

5.5.1 Results on Synthetic Data....125

5.5.2 Results on Silicon Measurement Data....133

5.6 Summary....138

Chapter 6 Thesis Summary & Future Work....139

6.1 Summary....139

6.2 Future Work....140

Bibliography....144

List of Figures

Figure 1-1. Three key stages of the lithography flow. 4

Figure 1-2. The etching process transfers the image from the resist to the layer under the resist. 5

Figure 1-3. (a) Dishing and (b) erosion effects of the CMP process. 6

Figure 1-4. A 9-stage ring oscillator. 9

Figure 1-5. Overview of the thesis organization. 15

Figure 2-1. (a) Wafer-level CD measurement map and (b) Spatially correlated variation extracted using a quadratic model [6]. 23

Figure 2-2. (a) Depth 1 edge of a wafer and (b) Depth 2 edge of a wafer, where edge dies are marked in red. 24

Figure 2-3. (a) Depth 1 edge and (b) Depth 2 edge of a wafer divided into 4 regions. 25

Figure 2-4. (a) Depth 1 edge and (b) Depth 2 edge of a wafer divided into 4 regions with different partitioning. 26

Figure 2-5. Four different center region definitions of the same wafer. 27

Figure 2-6. (a) Within-die CD measurement map and (b) Spatially correlated variation extracted using a quadratic model [6]. 29

Figure 2-7. (a) Measured ring oscillator (RO) period values (normalized by a randomly selected constant) of 117 ROs from the same wafer. (b) Discrete cosine transform (DCT) coefficients (magnitude) of the normalized RO period measurement show a unique sparse pattern. 31

Figure 2-8. Comparison of DCT and wavelet basis functions on explaining variance of spatial variation based on ring oscillator period measurement data. 33

Figure 2-9. (a) First step of the OMP algorithm on a 2-D example. (b) Second step of the OMP algorithm on a 2-D example. 41

Figure 2-10. A 2-D example of using S-OMP to select the correct basis vector from two choices. 44

Figure 2-11. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation. 47

Figure 2-12. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.48

Figure 2-13. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding large random variation. 49

Figure 2-14. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.50

Figure 2-15. (a) Systematic variation of the synthetic wafer. (b) Synthetic data with checkerboard sampling. 50

Figure 2-16. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.50

Figure 2-17. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation. 52

Figure 2-18. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.52

Figure 2-19. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding large random variation. 53

Figure 2-20. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.54

Figure 2-21. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation. 55

Figure 2-22. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.56

Figure 2-23. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding large random variation. 57

Figure 2-24. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.57

Figure 2-25. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation. 58

Figure 2-26. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.59

Figure 2-27. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding large random variation. 60

Figure 2-28. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.60

Figure 2-29. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation. 61

Figure 2-30. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.61

Figure 2-31. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding large random variation. 62

Figure 2-32. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.62

Figure 2-33. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation. 64

Figure 2-34. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.64

Figure 2-35. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding large random variation. 65

Figure 2-36. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.65

Figure 2-37. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation. 68

Figure 2-38. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries. 69

Figure 2-39. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding large random variation. 69

Figure 2-40. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries. 70

Figure 2-41. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding small random variation. 71

Figure 2-42. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries. 71

Figure 2-43. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation. 72

Figure 2-44. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries. 72

Figure 3-1. (a) Systematic variation of the synthetic wafer. (b) Measurement data created by adding large random variation and three outliers. (c) Estimated spatially correlated variation without outlier detection. 75

Figure 3-2. The IQR method successfully detects the outlier that strongly biases the estimation of mean and standard deviation. 77

Figure 3-3. (a) Measurement data with 3 outliers. (b) Estimated spatially correlated variation. (c) Outliers are more easily distinguished after removing spatially correlated variation. 78

Figure 3-4. (a) the L_2-norm ρ function. (b) the bisquare ρ function. 80

Figure 3-5. The bi-square weight function. 82

Figure 3-6. (a) Systematic variation of the synthetic wafer. (b) Measurement data created by adding random variation. (c) Measurement data after adding 3 outliers. 84

Figure 3-7. (a) Spatially correlated variation extracted by applying sparse regression with the physical dictionary, without outlier detection. (b) Spatially correlated variation extracted by applying sparse regression with the physical dictionary, with traditional IQR outlier detection. (c) Spatially correlated variation extracted by the proposed method with the physical dictionary. 85

Figure 3-8. (a) Systematic variation of the synthetic wafer. (b) Measurement data created by adding random variation. (c) Measurement data after adding 3 outliers. 85

Figure 3-9. (a) Spatially correlated variation extracted by applying sparse regression with the physical dictionary, without outlier detection. (b) Spatially correlated variation extracted by applying sparse regression with the physical dictionary, with traditional IQR outlier detection. (c) Spatially correlated variation extracted by the proposed method with the physical dictionary. 86

Figure 3-10. (a) Spatially correlated variation extracted from 201 wafers of ring oscillator period measurements (normalized). (b) Measurement data from one wafer with outliers. 87

Figure 3-11. (a) Spatially correlated variation extracted by applying sparse regression with the physical and DCT dictionaries, without outlier detection. (b) Spatially correlated variation extracted by applying sparse regression with the physical and DCT dictionaries, with traditional IQR outlier detection. (c) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries. 88

Figure 3-12. I_{dsat} measurement data (normalized) from one of the 15 wafers. 88

Figure 3-13. (a) Spatially correlated variation extracted by quadratic basis functions. (b) Spatially correlated variation extracted by the proposed method with the physical dictionary. (c) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries. 90

Figure 3-14. I_{dsat} measurement data (normalized) from one of the 8 wafers. 90

Figure 3-15. (a) Spatially correlated variation extracted by quadratic basis functions. (b) Spatially correlated variation extracted by the proposed method with the physical dictionary. (c) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries. 90

Figure 3-16. Ring oscillator (RO) period measurement data (normalized) from one of the 201 wafers. 92

Figure 3-17. (a) Spatially correlated variation extracted by quadratic basis functions. (b) Spatially correlated variation extracted by the proposed method with the physical dictionary (c) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries. 93

Figure 3-18. (a) Contact resistance measurement data (normalized) from one of the 24 dies. (b) Spatial distribution of different contact layout patterns in the test chips. 93

Figure 3-19. (a) Spatially correlated variation extracted by the physical dictionary with layout basis functions. (b) Spatially correlated variation extracted by the physical and DCT dictionaries. (c) The spatially correlated variation represented by the quadratic and DCT basis functions. 94

Figure 4-1. Contact resistance measurement data (normalized) from one of the 24 dies that contain significant amount of missing data. 100

Figure 4-2. Contact resistance measurement data (normalized) from one of the 24 dies. 107

Figure 5-1. I_{dsat} measurements (normalized) on two wafers with different spatial signatures. 111

Figure 5-2. Proposed flow to achieve wafer spatial signature clustering. 112

Figure 5-3. (a) Synthetic two-dimensional data with 2 clusters and 3 outliers and (b) k-means clustering result with 5 clusters. 116

Figure 5-4. (a) Synthetic two-dimensional data with 5 points and (b) The dendrogram generated by hierarchical clustering. 117

Figure 5-5. An example of inconsistency coefficient calculation....................120

Figure 5-6. (a) The error curve of complete-link hierarchical clustering on a synthetic data set. (b) The optimal number of clusters can be found by fitting the curve with two lines. 121

Figure 5-7. An synthetic example where the error can be minimized by either (a) c=2. (b) c=3. 123

Figure 5-8. Systematic variation of three different clusters in the synthetic example. 125

Figure 5-9. Three synthetic wafers belonging to different clusters. 126

Figure 5-10. Three synthetic wafers with large random variation belonging to different clusters. 127

Figure 5-11. Five abnormal wafers created by very large random variation. 127

Figure 5-12. Systematic variation of four different clusters in the synthetic example. 129

Figure 5-13. Four synthetic wafers belonging to different clusters. 130

Figure 5-14. Four synthetic wafers with large random variation belonging to different clusters. 131

Figure 5-15. Averaged wafer map (normalized) of four different clusters detected by the proposed method for NMOS I_{dsat} measurement data set 1....................133

Figure 5-16. Averaged wafer map (normalized) of three different clusters detected by k-means for NMOS I_{dsat} measurement data set 1....................134

Figure 5-17. Three different wafers in NMOS I_{dsat} measurement data set 1. 135

Figure 5-18. Averaged wafer map (normalized) of two different clusters detected by the proposed method for PMOS I_{dsat} measurement data set 1. 135

Figure 5-19. Averaged wafer map (normalized) of four different clusters detected by the proposed method for NMOS I_{dsat} measurement data set 2....................136

Figure 5-20. Averaged wafer map (normalized) of two different clusters detected by the proposed method for PMOS I_{dsat} measurement data set 2. 137

List of Tables

Table 1-1. Gate length scaling and 3σ variation predicted by ITRS 2011 [61] 1

Table 2-1. Summary of results with quadratic effects. 54

Table 2-2. Summary of results with edge/center effects. 66

Table 2-3. Summary of results with complex effects. 73

Table 3-1. Variation components of the first I_{dsat} measurement data set. 89

Table 3-2. Variation components of the second I_{dsat} measurement data set. 91

Table 3-3. Variation components of the RO period measurement data set. 92

Table 3-4. Variation components of the contact resistance measurement data set. 94

Table 4-1. Computational time of sparse regression for a single chip...108

Table 4-2. Computational time of sparse regression for 24 chips 108

Chapter 1

Introduction

In 1965, Gordon Moore observed that the number of transistors on integrated circuits doubles approximately every two years, which was soon recognized as Moore's Law [80]. This trend has continued for more than half a century and is still expected to continue for the next few years. A large number of benefits are enabled for integrated circuits (ICs) by transistor scaling: the cost per transistor becomes cheaper, the transistors become faster, and they also consume less power. As a result, integrated circuits (ICs) with more functionality, superior performance and less cost are being produced every year. It has been the key enabler of a large number of technological and social changes in the late 20th and early 21st centuries [81].

Table 1-1.Gate length scaling and 3σ variation predicted by ITRS 2011 [61]

Year	2011	2012	2013	2014	2015
Gate Length (nm)	24	22	20	18	17
3σ variation (nm)	2.9	2.65	2.42	2.21	2.02

As Moore's Law continues to hold, new process technology that achieves deeper scaling is continuously being developed, and more and more new products are being designed and manufactured using newer process. However, one of the key limiting factors to the profitability of these new process/products is the *yield*, which is defined as the proportion of manufactured circuits that are functional and meet their performance requirements [56]. The yield loss of circuits is mainly due to process variations, which can be defined as the deviations of the manufactured circuit compared to the design. Therefore, to ensure profitability, reducing the variability and improving the yield is an important task that is performed throughout the entire lifecycle of any process and product. However, such task becomes increasingly difficult to achieve with scaling [1][61][56]. For example, Table 1-1 shows the predicted trend of gate length and the corresponding 3σ variation from 2011 to 2015 by ITRS 2011 [61], where yellow indicates manufacturable solutions are known and red means manufacturable solutions are unknown. It can be seen that in order to keep process variation under control, the 3σ needs to scale proportionally with the gate length scaling. This poses significant challenges to process engineers, such that no manufacturable solutions are known beyond 2014. To make things more challenging, process engineers are now faced with a much shorter time window for the yield improvement effort. The lifetime of modern electronic products, such as cell phones, may be only several months; moreover, missing important deadlines such as Christmas for consumer electronics will result in significant revenue loss. Based on these observations, rapidly improving the yield for today's complicated manufacturing process is a key enabler for profitability for the IC industry. In order to achieve this goal, process variation must be thoroughly characterized to determine the important yield-limiting factors, and the yield improvement efforts can then be prioritized to focus on these important factors. For example, it is reported by PDF Solutions that facilitated by such an accurate variation characterization methodology, compared to a traditional yield ramp approach, they are able to further improve the yield for an actual product by 5%-28% throughout its lifecycle, which enables cost savings of more than 100 million dollars [103].

To achieve accurate variation characterization, an important observation is that a large number of variation sources cause spatial non-uniformity in process condition across the wafer and/or die surface. These variation sources are becoming increasingly critical in advanced technology nodes, especially with the transition to 450mm wafers. Each variation source often results in a unique spatial variation pattern. Therefore, if we are able to accurately understand the spatial patterns produced by the process, it will provide important insights into the yield-limiting factors. In this thesis, we develop accurate modeling techniques for spatial variation, and further develop two algorithms based on our model that automatically produce relevant results to help identify important yield-limiting factors. For wafers with similar spatial patterns, *variation decomposition* narrows down the sources of variation by decomposing the overall variation into multiple different components. each corresponding to a different subset of variation sources. Especially, the wafer-level and within-die spatial pattern is extracted and their impact in overall variation is estimated. If the spatial pattern can be different for different wafers, *wafer spatial signature clustering* automatically partitions these wafers into groups exhibiting different spatial signatures, which helps process engineers find important factors that prevent the process from stably maintaining a high yield across different lots and wafers. In the rest of this chapter, we will briefly review the background on process variation, variation characterization and spatial variation modeling, and then outline the overall structure of the thesis.

1.1 Process Variations

Process variations are the deviations of the manufactured circuit compared to the design. Process variations can be categorized into *catastrophic* variations and *parametric* variations. Catastrophic variations are mainly due to defects such as metal opens and shorts, while parametric variations are due to variability in process parameters such as gate length and threshold voltage. In this thesis, we focus on parametric variations, which are becoming increasingly significant in new technology nodes [56]. For example, Ref. [2] shows the leakage and frequency variations of Intel microprocessors on a wafer, in which 20× variation in chip leakage and 30% variation in chip frequency can be seen. As a result, both the high leakage and low frequency chips have to be discarded, and the remaining chips still have to go through an expensive and time-consuming frequency binning process. This poses a significant challenge to process engineers and circuit designers in order to ensure yield and profitability.

Modern products typically require hundreds of process steps. First, active devices such as MOS transistors are fabricated on top of the substrate through a series of steps such as deposition, patterning and implantation, which is named the front end of line (FEOL) process. Next, multiple layers of interconnect are created to connect the active devices and power sources, which is named the back end of line (BEOL) process. Each of these process steps is subject to process variation, which ultimately impacts the final product yield. In this section, we will briefly review some of the main variation sources in today's manufacturing process. Note that for different processes/products, the relative importance of these variation sources can be significantly different.

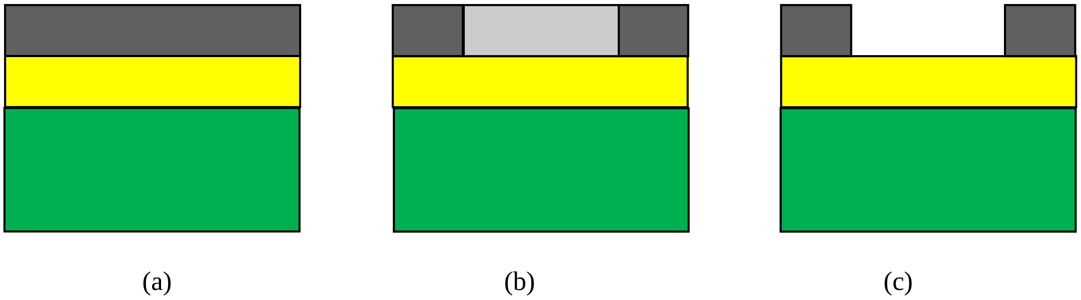


Figure 1-1. Three key stages of the lithography flow.

Multiple main variation sources in the FEOL process can be found within the lithography process. Lithography uses light to transfer a geometric pattern from the mask to the resist, which is a material sensitive to light. In the lithography flow, resist is first applied to the wafer with a spin-coating process. This is shown in Figure 1-1 (a), where the yellow material denotes the thin film on which the pattern needs to be applied, and the black material denotes the resist. The resist coated wafer then goes through a soft baking process to remove excess resist solvent. Next, the resist is exposed to a pattern of intense light where the pattern is defined by the mask. A post-exposure bake (PEB) is then performed after exposure to reduce the standing wave effect. The state of the materials at this stage is shown in Figure 1-1 (b), where the exposed and unexposed regions have different solubility. The soluble resist is finally removed by the development process and the wafer is baked again to solidify the remaining resist. After the aforementioned lithography process, the thin film not protected by the resist will be removed by the etching process. Several possible main variation sources are in the resist spinning step, where various factors such as variability in spin speed, resist viscosity, and adhesive properties between the resist and substrate can lead to significant variation in resist thickness [47]. This will in turn lead to CD variation since resist thickness is strongly related to its sensitivity to exposure dose. A large number of possible variation sources exist in the exposure step, such as optical proximity, exposure dose variation, defocus, misalignment, mask error, lens aberration and line edge roughness (LER). Finally, significant CD variation can be caused by the nonuniformity of the thermal dose across wafer in the PEB step [31][32]. This is caused by the inability to maintain a perfectly uniform spatial PEB temperature distribution in the PEB equipment.

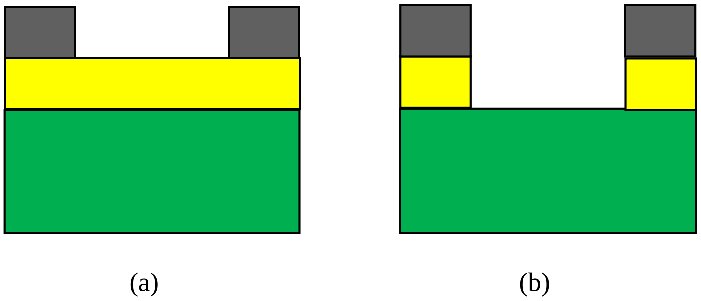


Figure 1-2. The etching process transfers the image from the resist to the layer under the resist.

After the lithography process, the geometric pattern has been formed on the resist, and etching is then applied to transfer the image into the layer under the resist, as shown in Figure 1-2. There are two types of etching processes, wet etching and dry etching. For critical process steps, typically dry etching is applied because of its better controllability. The dry etching process bombards the wafer with an incident flux of ions, radicals, electrons and neutrals and the unwanted material is removed by both physical damage and chemical attack [82]. CD variation caused by the etching process can be due to a number of variation sources in the process conditions, such as temperature, pressure, gas flow and RF power [75]. Moreover, the etch rate can be layout dependent, resulting in the macro and micro loading effects which can be significant sources of variation in modern ICs.

Besides lithography and etching, several other important sources of variation in the FEOL process can be found in the ion implantation and annealing process. In order to define the transistors, different regions such as the n- and p- wells, the transistor source and drain, and the lightly doped drain (LDD) are doped with different ion species and/or concentration. This is achieved by first applying the ion implantation process, where ionized impurity items are accelerated through an electrostatic field to strike the wafer. Next, these impurities are activated by the annealing process to properly distribute them [82]. Variability in the ion implantation is related to the variation of multiple process conditions such as implantation dose, tilt angle, temperature, and uniformity of dopants across the wafer surface [82][75]. For the annealing process, the prevalent rapid thermal annealing (RTA) method is known to be sensitive to pattern non-uniformity across the wafer surface [44]-[46], which makes strongly layout-dependent [41][42]. Finally, in advanced technology nodes, since the device size is extremely small, the number of dopants in the channel area may be only hundreds, so that the actual number of dopants and their placement can cause significant variation to the device threshold voltage, which is known as the random dopant fluctuation (RDF) problem [83].

Besides the aforementioned variation sources, a number of other significant FEOL variation sources exist. For example, many layers of thin films such as the polysilicon gate are deposited using chemical vapor deposition (CVD), and it is difficult to maintain uniform deposition rate across the wafer surface [36] [37]. In the gate oxidation step, temperature non-uniformity across the wafer due to lamp configuration, as well as the gas flow, can cause significant across wafer gate oxide thickness variation [43].

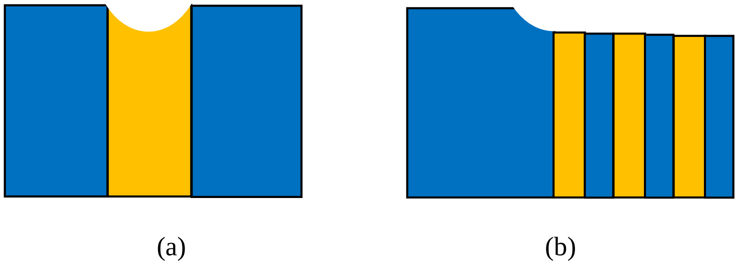


Figure 1-3. (a) Dishing and (b) erosion effects of the CMP process.

In the BEOL process, an important source of metal thickness and inter-layer dielectric (ILD) thickness variations is the chemical mechanical polishing (CMP) process. After depositing the metal and ILD, CMP is applied to achieve a planar surface so that subsequent layers can be fabricated on top of them. However, complete planarity cannot be achieved by CMP and it is subject to a number of variation sources. These variation sources include process condition variation such as pad pressure, pad velocity and temperature [82]. Moreover, metal with large width is subject to the dishing effect, where more metal is removed in the center, as shown in Figure 1-3 (a); different pattern density leads to different removal rate of metal and ILD, which is named the erosion effect shown in Figure 1-3 (b).

In addition to the aforementioned variation sources, significant process variation can be caused by non-ideal matching properties of equipments. Wafers manufactured by different equipments can have significant difference due to the mismatch in the process conditions of equipments; even within the same equipment, process condition mismatch between chambers can cause process variation [55]. The process condition of the same tool can also change over time, resulting in process shifts and drifts [75].

In summary, there exist a large number of variation sources can potentially impact the product yield.

While all these variation sources must be carefully addressed in the manufacturing process, when improving the yield of a particular process/product, due to the stringent requirement of time to market, the process engineers must prioritize their goals and focus their efforts on a smaller subset of the variation sources that has stronger yield impact. Obviously, the dominant variation sources change from process to process or even from product to product. Therefore, to capture these variation sources, process variation must be thoroughly characterized and the measurement data must be carefully analyzed. In the next sub-section, we will review the variation characterization techniques to understand process variation.

1.2 Variation Characterization

In order to understand and ultimately reduce the variation to improve yield, process variation must be thoroughly characterized. Variation characterization is primarily achieved by measuring a set of electrical properties from test structures. These test structures may be placed within test chips, scribe lines, or the product chips. In this sub-section, we will first review some of the most important test structures to characterize parametric variation used in today's manufacturing process. Next, we will discuss several applications that analyze the measurement data obtained from these test structures to derive important information that guides the efforts to yield improvement.

1.2.1 Test Structures for Variation Characterization

A large variety of test structures have been proposed for variation characterization purposes. Some of these test structures focus on characterizing the variation of a particular parameter. For example, electrical linewidth metrology (ELM) measures the gate length by passing a precisely calibrated current through the gate and measuring the voltage across a subsection of the gate [6]. Interconnect resistance can be measured using the Van der Pauw method [95], and a charge based capacitive measurement test structure is proposed in [94] to measure interconnect capacitance. A method to measure the contact resistance of individual contacts is described in [21]. These test structures are related to a small subset of physical variation sources so that the sensitivity to a particular variation source can be more easily determined, but they do not directly provide information on how these variation sources would impact the variation or yield of a finished product.

Other test structures are based on transistors and their performance measurements are therefore more strongly correlated with the performance of the actual product. However, since a large number of process steps must be performed to fabricate a transistor, determining the sensitivity to a particular variation source can be challenging. One important category of the transistor-based test structures measures the properties of a single transistor [84]-[86] [88]. A benefit of this type of test structure is that it is possible to completely characterize a single transistor by gathering its full I-V data. Traditionally, the gate, source and drain of the transistor are required to be directly connected to probing pads. Since it consumes a lot of resources, this type of test structure was primarily used to create the SPICE models for circuit simulation [56], and it was difficult to deploy this test structure in large quantities to gather the statistics required for variation characterization. This problem is addressed by several recent works. For example, Ref. [84] measures the I-V characteristics for a large number of transistors using a scan chain based approach. Ref. [86] presents a large addressable transistor array where the I-V characteristic of each transistor can be measured. To obtain the transistor threshold voltage variation from single-transistor measurements, Ref. [85] presents a large transistor array dedicated to measure threshold voltage variation of each individual transistor efficiently by measuring the gate-to-source voltage variation under the same drain current; Ref. [88] presents another technique which derives the threshold voltage variation from the leakage current measurements of each transistor in a large transistor array. The design dependent variation can be captured by measuring and comparing transistors with different design attributes such as width, length and layout.

SRAM is a key building block in modern chips and hundreds of millions of SRAM cells may be placed on chip as cache memories. Moreover, because of the small device size used in SRAM, it is particularly sensitive to process variation. Therefore, the variability of SRAM cells is typically thoroughly characterized using SRAM arrays. For example, Ref. [87] characterizes the read current and write trip voltage of 1M SRAM cells, and Ref. [93] characterizes the read and write margins of SRAM cells by using several SRAM arrays in a test chip. In practice, other key components of the product chip can be also characterized as test structures to learn their performance and sensitivity to process variations.

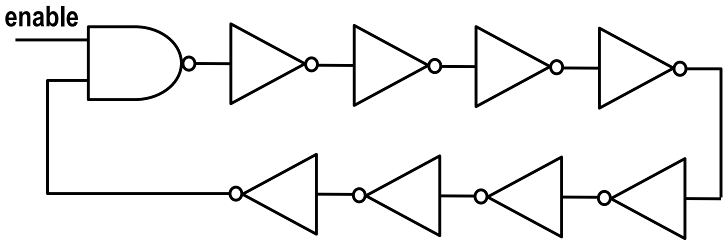


Figure 1-4. A 9-stage ring oscillator.

Another important category of test structures commonly deployed is ring oscillators (ROs). A ring oscillator consists of an odd number of inverting stages. For example, Figure 1-4 shows an RO with 9 stages where one of the stages is a NAND gate connected to a signal for enabling oscillation, and the other 8 stages are inverters. Frequency and leakage measurements can be gathered from RO test structures, in which the frequency measurement can be easily measured with a low-cost frequency counter. Compared to the single transistor test structures, ring oscillators reflect circuits operations under high-speed conditions as in an actual product application, so that it is more strongly related to the performance of actual products [89]. Therefore, ROs are widely applied in variation characterization. For example, Ref. [90] describes a ring oscillator based test chip to characterize the process variation of a 0.25um process under different layout settings. Ref. [91] uses RO frequency and leakage to characterize the delay and leakage variation of a 90nm process. Ref. [92] uses RO frequency and leakage from an array of transistors to characterize the delay and leakage variation of a 45nm process. The problem of identifying sensitivity to process parameters is partly addressed in [89], which proposes to derive a number of parameters such as switching capacitance and threshold voltage variation by comparing ROs with different configurations. However, decomposing the variation and identifying the important variation sources remains a significant challenge.

1.2.2 Statistical Analysis of Measurement Data

After obtaining the measurement results from test structures, the next step is to apply statistical analysis techniques to interpret these measurement data. An important goal of statistical analysis is to derive important information that helps process engineers with the efforts to reduce the variation to improve yield. For any process and product, these yield improvement efforts are made throughout its entire lifecycle. From a product point of view, the lifecycle of its manufacturing process can be partitioned into three stages: process development, product yield ramp and volume manufacturing. In the process development stage, the foundry internally develops, evaluates and optimizes the manufacturing process, and provides the process design kit (PDK) of the process to the customer. In the product yield ramp stage, the process and product are further fine-tuned to optimize the yield. Finally, in the volume production stage, the product is manufactured in large quantities and the goal is to stably maintain a high yield across different lots and wafers.

In order to rapidly improve the yield, we need to inspect the measurements from product representative test structures (e.g. transistor saturation current, transistor leakage, RO frequency) or performance measurements from the product itself (e.g. maximum operating frequency, leakage), and identify the important variation sources that significantly contribute to the variation of these measurements. Once such important variation sources are identified, yield improvement efforts can be made more effective by focusing on these variation sources. However, this goal is extremely difficult to achieve, since modern manufacturing processes typically consist of hundreds of complex process steps. To narrow down the sources of variation, an important first step is to decompose the variation from a geometrical perspective into: *lot-to-lot variation*, *wafer-to-wafer variation*, *wafer-level variation* and *within-die variation*. Different geometrical levels can indicate different physical sources of variation. For example, lot-to-lot variation can be caused by tool-to-tool variations, changes in tool conditions over time, and differences in starting and processing material properties [7]. For single-wafer processing tools, wafer-to-wafer variation may be caused by temporal drift of process conditions as wafers are sequentially processed [7], or chamber condition mismatch of the same tool [55]. For multiple-wafer processing tools, wafer-to-wafer variation can be caused by different process conditions at different spatial locations within the same tool [37]. Wafer-level variation can be caused by process condition non-uniformity across the wafer. For example, ion density in etching, temperature gradients in baking, or process condition variation from reticle to reticle can cause wafer-level variation [7] [56]. Within-die variation can be caused by stepper induced variations, etching variations, mask errors, or random device mismatches such as random dopant fluctuation and line edge roughness [7][56].

For a number of wafers with similar spatial patterns, we would like to further decompose their wafer-level and within-die variations into spatially correlated variation and random variation. In process development and product yield ramp stages, wafers typically have similar patterns, since they are manufactured using a limited set of equipments so that the mismatch between equipments is not a strong concern. Once such decomposition is performed, spatially correlated variation and random variation are related to different physical sources. Wafer-level and within-die spatially correlated variation can uncover systematic variation sources such as temperature gradients due to baking/etching equipment design, lens aberrations in lithography, etc. On the other hand, random variation can be caused by random device mismatches, random fluctuations of equipment condition over time, etc. Since different variation sources can result in completely different spatial patterns, once the spatially correlated component is extracted, it is possible to further search for the important systematic variation source by comparing the extracted spatial pattern with those produced by various process steps/equipments [57], such as the results from the aforementioned test structures dedicated to characterize few process steps. The key question is how to develop a statistical method to automatically achieve variation decomposition from both geometrical and spatial perspectives.

In practice, for a large number of wafers with product chips, different spatial patterns can occur for different wafers, especially during volume production. By detecting such difference, it may reveal a large number of yield-limiting factors, such as process shift/drift, mismatch between equipments, mismatch between different chambers, etc. To monitor the process variations, for each wafer, a number of measurements are collected from test structures deployed on-chip and/or in the scribe line, such as DC characteristics of single transistors and ring oscillator frequency [7][64]. In order to detect yield-limiting factors from these measurement data, an important property that can be utilized is that wafers affected by different major variation sources can exhibit completely different spatial patterns. Therefore, if we can capture the spatial signature of each wafer with an accurate model, and further automatically partition all the wafers into different groups based on such spatial signature, in which each group exhibits a similar spatial signature, it would provide important insight to help process engineers with the yield improvement effort. Especially, process engineers can prioritize the yield improvement goals and focus on the mechanism related to large groups with strong spatial signature, so that reducing the variation sources that correspond to such spatial signature will have a significant impact on the improvement of overall yield.

In summary, we discussed two applications that analyze measurement data to derive important information for yield improvement throughout the product lifecycle. In this thesis, our goal is to derive efficient algorithms for these two applications. It can be seen that for both applications, an accurate model for spatial variation is a key component. Therefore, in the next sub-section, we will first briefly review the previous works on spatial variation modeling and motivate the need for a new modeling technique. We will then summarize the main contributions of this thesis in Section 1.4.

1.3 Spatial Variation Modeling

In order to develop efficient solutions for the two applications discussed in the previous sub-section, a key problem that must be solved is to develop an accurate model for spatial variation. Many modeling techniques have been proposed in the literature based on the spatial correlated property of systematic variation sources, and they can be divided into two categories. In this sub-section, we will first review these two categories of techniques, and then motivate the need for a new spatial variation modeling technique for our applications.

The first category of models represents spatially correlated variation as correlated random variables and the correlation is modeled as a function of distance. The earliest work in this category is the Pelgrom model [3], which states that the variance of multiple process parameters is dependent on the squared distance between transistors, as a result of the spatially correlated systematic variation. For example, it models the threshold voltage mismatch of a transistor as:

$$\sigma^2(V_{T0}) = \frac{A_{V_{T0}}^2}{WL} + S_{V_{T0}}^2 D^2 \tag{1.1}$$

where $A_{V_{T0}}$ and $S_{V_{T0}}$ are technology-dependent constants, W and L are the width and length of a transistor respectively, and D is the distance between instances of devices. Several recent works [4][5][96][97] further explicitly models the spatial variation as a stationary random field, where the correlation between any two points (x_i, y_i) and (x_j, y_j) is a function of their distance:

$$\rho\big(f(x_i, y_i), f(x_j, y_j)\big) = \rho\big(D((x_i, y_i), (x_j, y_j))\big) \tag{1.2}$$

where f denotes the performance of interest, ρ is the correlation coefficient and

$$D((x_i, y_i), (x_j, y_j)) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{1.3}$$

is the Euclidean distance between two points. The difference between these works is mainly the different correlation functions used. Specifically, the correlation function in [96] is a linear function of distance, and a piecewise linear function is used in [97]. Three correlation functions: exponential, Gaussian and linear functions are used in [4] and the actual choice for a particular process/design is determined empirically. Finally, a general family of valid correlation functions was proposed in [5], which allows more flexibility on the shape of correlation function. These models do not fit the need of our applications because of the following two reasons. First, they do not explicitly decompose the spatially correlated variation with the random variation in the measurement data, and therefore cannot be applied for variation decomposition. Second, all these methods follow the assumption of (1.2) which consider the spatial correlation only as a function of the distance between measurements. This assumption is too simplistic to fully capture the spatially correlated systematic variation in manufacturing process.

The second category of models represents spatially correlated variation as an analytical function of the spatial coordinate (x, y). For example, a linear function is used in [7] to model the spatially correlated variation of RO delay:

$$f(x,y) = a_0 + a_1 x + a_2 y \,. \tag{1.4}$$

Several other works further add quadratic terms to model the spatially correlated variation [6][35]:

$$f(x,y) = a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 y^2 \,. \tag{1.5}$$

And the following full quadratic model is applied to model the spatially correlated variation in [34], [98] and [99]:

$$f(x,y) = a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 y^2 + a_5 xy \,. \tag{1.6}$$

These models allow the decomposition of spatially correlated variation and random variation by explicitly extracting the spatially correlated variation with the model. Moreover, the model coefficients such as a_0-a_5 in (1.6) provide efficient representation of the spatial signature of wafers, which could be further utilized for the wafer spatial signature clustering application.

From the above comparison of models, the second category of models is more suitable for our applications. However, the most significant challenge in applying these models is that the simple linear and quadratic functions in (1.4)-(1.6) are only capable of modeling a limited amount of systematic variation sources and may not be sufficient for modern processes. For example, as will be shown in Section 2.1, the difference between edge dies and other parts of a wafer is becoming an increasing difficult problem, which cannot be modeled by the functions in (1.4)-(1.6). More complex models are needed in order to capture more systematic sources such as the edge effect, but overly complicated models will lead to the over-fitting problem [24]. Once over-fitting occurs, it will model random variations as complex spatial patterns, which greatly increases the modeling error. Therefore, we need to re-visit this problem and develop an accurate model for spatial variation that addresses these issues.

1.4 Thesis Contributions

In this thesis, we propose accurate and efficient statistical techniques to solve the aforementioned variation decomposition and wafer spatial signature clustering problems. These techniques facilitate accurate identification of the important variation sources throughout the product lifecycle, which is vital to rapidly improving yield. The major technical contributions of this thesis are:

- We propose to model spatial variation based on sparse regression. We demonstrate that spatially correlated variation can typically be modeled with a small number of pre-determined "templates" (e.g., linear and quadratic functions). However, the most appropriate templates to model the spatially correlated variation may vary for different process or design, and directly applying all possible templates will lead to severe over-fitting problem. To apply the proposed sparse regression technique, only a *dictionary* of templates is needed, which includes all possible patterns of spatially correlated systematic variation. The optimal templates to model the spatially correlated variation of a given wafer/die will be automatically selected by sparse regression to significantly reduce over-fitting.
- We construct two dictionaries that can capture more spatially correlated systematic variation sources than the traditional quadratic modeling approach. We have studied a number of common physical variation sources and constructed a physical dictionary based on them. Furthermore, we construct the Discrete Cosine Transform (DCT) [23] dictionary based on unique sparse structure of spatially correlated variation in frequency domain.
- We develop a robust solver for the sparse regression problem to accurately select the templates and remove measurement outliers.

- We develop a number of efficient numerical algorithms that significantly reduce the computational cost with large problems when the DCT dictionary is applied.
- We propose a method to solve the variation decomposition problem based on the proposed robust sparse regression technique, the physical and DCT dictionaries and the linear mixed model [29]. The method is tested on a large number of synthetic and silicon data sets to demonstrate its efficacy.
- We propose a method for the wafer spatial signature clustering problem by using robust sparse regression to extract the spatial signature of wafers and complete-link hierarchical clustering algorithm to perform clustering, and we develop a modified L-method to accurately determine the number of clusters. The efficacy of the proposed method is demonstrated on a large number of synthetic and silicon data sets.

1.5 Thesis Organization

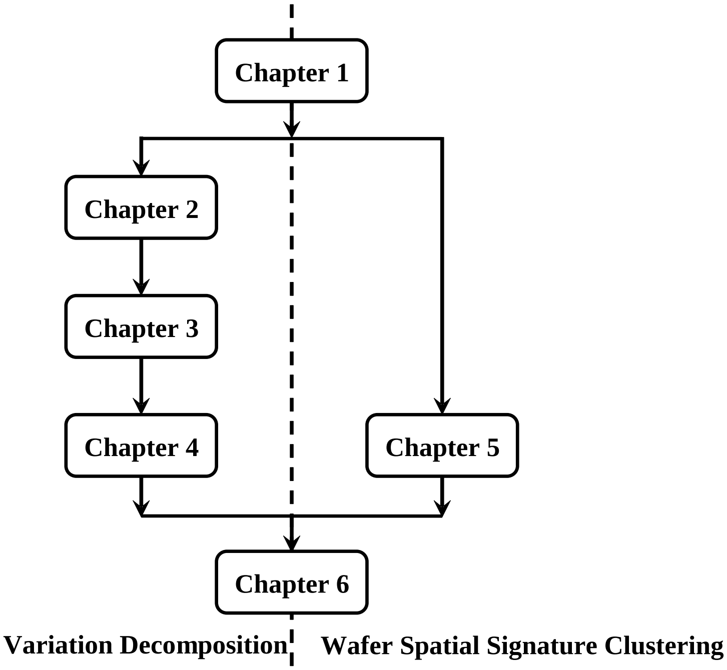


Figure 1-5. Overview of the thesis organization.

The overall structure of the thesis is shown in Figure 1-5. Chapter 2-Chapter 4 focus on variation decomposition, and the spatial variation modeling technique is motivated and developed based on this application. Chapter 5 proposes a method for wafer spatial signature clustering and the spatial variation modeling technique in Chapter 2-Chapter 4 is re-used as an important component. Chapter 6 concludes this thesis. We briefly summarize the contents of Chapter 2-Chapter 6 below:

In Chapter 2, we first present the mathematical formulation for the variation decomposition problem, in which an important goal is to identify the systematic spatially correlated component for wafer-level and within-die variation. Towards this goal, we first develop a physical basis function dictionary based on our study of several common physical variation sources, which captures more spatially correlated systematic variation sources than the traditional quadratic modeling approach, and then propose the DCT dictionary to discover spatially correlated systematic patterns not modeled by the physical dictionary. Moreover, we proposed to apply sparse regression to significantly reduce the over-fitting problem posed by a large basis function dictionary. A large number of synthetic examples are constructed to demonstrate the efficacy of the proposed algorithm and models.

The existence of outliers is an important problem that widely exists in silicon measurement data. If outliers are not appropriately considered, they will introduce substantial error to variation decomposition. In Chapter 3, we extend the sparse regression algorithm introduced in Chapter 2 to a robust sparse regression algorithm. By solving robust sparse regression, basis functions will be accurately selected in the presence of outliers, and outliers will be automatically detected and removed, before the data is provided to the linear mixed model to perform variation decomposition. Experiments on synthetic and silicon measurement data demonstrate that the proposed robust sparse regression algorithm provides superior accuracy compared to the traditional IQR method for outlier detection. We further performed variation decomposition on several silicon data sets and demonstrated the effectiveness of the proposed variation decomposition flow based on robust sparse regression.

The computational cost for sparse regression with DCT basis functions can become extremely large for problems with large size, which limits the applicability of the variation decomposition methodology introduced in Chapter 2-Chapter 3. Therefore, in Chapter 4, we propose several efficient methods to make the computational cost of sparse regression tractable for large-scale problems. The key idea of these methods is to utilize fast DCT/IDCT computation to speed up the matrix-vector product computation. From the experimental results on a large problem with contact resistance measurement data, we observe nearly

200× speedup compared to the traditional direct implementation.

In Chapter 5, we propose an accurate method to solve the wafer spatial signature clustering problem. The proposed method contains three key components: first, a robust feature extraction method is developed to automatically capture the spatial signatures of wafers by a small number of features based on the robust sparse regression technique developed in Chapter 2-4; second, a complete-link hierarchical clustering algorithm is selected to perform clustering on the features; finally, a modified L-method is developed to select the number of clusters from the hierarchical clustering result. The effectiveness of the proposed method is demonstrated by a number of synthetic and silicon data sets.

Chapter 6 concludes the thesis with a high-level summary of the work, and discusses several future potential directions of research related to this work.

Chapter 2

Variation Decomposition via Sparse Regression

2.1 Motivation

As was discussed in Section 1.2.2, an important goal in variation characterization is to identify important variation sources that contribute significantly to the overall variation. Since modern manufacturing processes typically consist of hundreds of complex process steps, this goal is extremely difficult to achieve. To narrow down the sources of variation, an important first step is to decompose the variation from a geometrical perspective into: *lot-to-lot variation*, *wafer-to-wafer variation*, *wafer-level variation* and *within-die variation*, where different geometrical levels can indicate different physical sources of variation. Therefore, the overall variation can be mathematically represented by the summation of four components:

$$b_{lkji} = \tau_l + \theta_{k(l)} + \gamma_{j(kl)} + \varepsilon_{i(jkl)} \tag{2.1}$$

where b_{lkji} indicates the overall variation, τ_l is the l-th lot variation, $\theta_{k(l)}$ is the k-th wafer variation within the l-th lot, $\gamma_{j(kl)}$ is the j-th die variation within the k-th die and l-th wafer, and finally $\varepsilon_{i(jkl)}$ is the i-th within-die variation within the j-th die, the k-th wafer, and the l-th lot.

To further narrow down the sources of variation for a number of wafers with similar spatial pattern, wafer-level and within-die systematic variation can be modeled by extracting the spatially correlated variation, which is represented by linear combination of a set of pre-defined basis functions:

$$\gamma_{j(kl)} = \sum_{m=1}^{\lambda_1} A_{wafer,m}(x_{die,j}, y_{die,j}) \cdot \alpha_m + \gamma^r_{j(kl)} \tag{2.2}$$

$$\varepsilon_{i(jkl)} = \sum_{m=1}^{\lambda_2} A_{die,m}(x_{site,i}, y_{site,i}) \cdot \beta_m + \varepsilon^r_{i(jkl)} \tag{2.3}$$

where the wafer-level spatially correlated variation is represented by λ_1 basis functions $\{A_{wafer,m}(x_{die,j}, y_{die,j}),$

m = 1, 2, …, λ_1}, where ($x_{die,\ j}$, $y_{die,\ j}$) is the location of the j-th die on the wafer, and the remaining component $\gamma^r_{j(kl)}$ represents the random component of wafer-level variation. Similarly, the within-die spatially correlated variation is represented by λ_2 basis functions {$A_{die,m}$($x_{site,\ i}$, $y_{site,\ i}$), m = 1, 2, …, λ_2}, where ($x_{site,\ i}$, $y_{site,\ i}$) is the location of the i-th measurement site on the die, and the remaining component $\varepsilon^r_{i(jkl)}$ represents the random component of within-die variation. Each basis function can be viewed as a particular "template" to model the spatially correlated variation, and may be related to a small subset of process steps. This is a generalized definition compared to the simple linear and quadratic functions in (1.4)-(1.6). For example, the full quadratic model (1.6) can be expressed by six basis functions: $\{1, x,\ y,\ x^2,\ y^2,\ xy\}$. Once the decomposition in (2.2) and (2.3) is achieved, additional steps can be taken to further analyze the physical sources related to the spatially correlated variation. This can be achieved by two means: first, if the basis functions carry significant physical meaning, we can narrow down the process steps by first locating the important basis functions in (2.2) and (2.3), i.e. the basis functions that explain a significant portion of variance, and then investigate the variation sources that are related to them. Second, the extracted spatially correlated variation will present a unique spatial pattern, and therefore it is possible to search for the important variation source by comparing this spatial pattern with those produced by various process steps/equipments [57].

By combining Equations (2.1), (2.2) and (2.3), we obtain the following representation of the overall variation:

$$b_{lkji} = \tau_l + \theta_{k(l)} + \sum_{m=1}^{\lambda_1} A_m(x_{die,j}, y_{die,j}) \cdot \alpha_m + \gamma^r_{j(kl)} + \sum_{m=1}^{\lambda_2} B_m(x_{site,i}, y_{site,i}) \cdot \beta_m + \varepsilon^r_{i(jkl)} \qquad (2.4)$$

where the overall variation is decomposed into six components: lot-to-lot variation, wafer-to-wafer variation, wafer-level spatially correlated variation, wafer-level random variation, within-die spatially correlated variation and within-die random variation. Eq. (2.4) is referred to as a linear mixed model [29] in statistics. This model can be estimated using the Restricted Maximum Likelihood (REML) method [29], yielding the coefficients {α_m, m = 1, 2, …, λ_1} and {β_m, m = 1, 2, …, λ_2} for wafer-level and within-die spatially correlated variation, and the following variances: variance for lot-to-lot variation σ^2_{lot}, variance for wafer-to-wafer variation σ^2_{wafer}, variance for wafer-level random variation $\sigma^2_{die,r}$, and variance for within-die random variation $\sigma^2_{site,r}$. In order to estimate the contributions of spatially correlated variations in terms of variance, we also estimate the variance for spatially correlated wafer-level and within-die variation by the following sample variance estimation:

$$\sigma^2_{die,s} = \frac{1}{N_{die}-1}\sum_{j=1}^{N_{die}}\left(\sum_{m=1}^{\lambda_1} A_m(x_{die,j}, y_{die,j})\cdot\alpha_m - \mu_{die,s}\right)^2 \tag{2.5}$$

$$\sigma^2_{site,s} = \frac{1}{N_{site}-1}\sum_{i=1}^{N_{site}}\left(\sum_{m=1}^{\lambda_2} B_m(x_{site,i}, y_{site,i})\cdot\beta_m - \mu_{site,s}\right)^2 \tag{2.6}$$

where

$$\mu_{die,s} = \frac{1}{N_{die}}\sum_{j=1}^{N_{die}}\sum_{m=1}^{\lambda_1} A_m(x_{die,j}, y_{die,j})\cdot\alpha_m \tag{2.7}$$

$$\mu_{site,s} = \frac{1}{N_{site}}\sum_{i=1}^{N_{site}}\sum_{m=1}^{\lambda_2} B_m(x_{site,i}, y_{site,i})\cdot\beta_m \tag{2.8}$$

N_{die} is the number of dies on the wafer, and N_{site} is the number of measurement sites in a die. Once these variance values are estimated, the contribution of a particular component is estimated by dividing its variance value with the sum of variance for all components. Note that in practice, due to the limitation of measurements, we may only be able to estimate part of these variance values. For example, in early-stage yield learning, there may be only one wafer and only a single performance value is obtained from each die, we are only able to extract the wafer-level spatially correlated and random components. In this case, the contribution of wafer-level spatially correlated variation will be calculated by $\sigma^2_{die,s}/(\sigma^2_{die,s}+\sigma^2_{die,r})$, and the contribution of wafer-level random variation will be calculated by $\sigma^2_{die,r}/(\sigma^2_{die,s}+\sigma^2_{die,r})$.

An important problem in applying the linear mixed model (2.4) is that the appropriate basis functions must be selected to model the spatially correlated wafer-level and within-die variation. Traditionally, only a small number of simple basis functions are employed, such as linear basis functions [7] and quadratic basis functions [6][34][35][98][99]. These simple basis functions are only capable of modeling a limited amount of variation sources and are not sufficient for modern processes. For example, an important problem for modern processes is that edge dies on a wafer can have significantly lower yield compared to other parts of the wafer. This problem has been identified as an important yield-limiting factor in ITRS 2005 [58] and remains important in all subsequent ITRS editions [59]-[61]. With process scaling, as CD tolerances become tighter, the systematic differences encountered at the wafer's edge are playing a larger role in the yield equation [62]. Moreover, as wafer size has grown, so has the number of dies residing near the edge. For example, the outer 20mm of a 300mm wafer can contain up to 25% of the dies on a wafer [62]. It is expected that the challenges for extreme edge dies will be further increased by the transition to 450mm wafers [63]. As will be shown in Section 2.2, more basis functions are needed in order to capture more systematic sources such as those related to the aforementioned edge effect.

As the number of possible basis functions increases, a large *dictionary* of basis functions can be formed, and the underlying physical sources for a particular process or design may be modeled by only a subset of all basis functions from the dictionary. In this case, as will be shown in the numerical results in Section 2.5, if all basis functions are directly applied, it will result in the over-fitting problem [24]. Once over-fitting occurs, the amount of spatially correlated variation can be overestimated. Also, it may generate overly complicated spatial pattern which is difficult to analyze. Therefore, in this chapter, we further propose to apply a sparse regression technique to accurately select the actual subset of basis functions for a particular process/design from the dictionary, in order to combat such over-fitting problem.

The remainder of the chapter is organized as follows. In Section 2.2 we present dictionaries that can be used to model spatially correlated variation. Then, we formulate basis selection as a sparse regression problem in Section 2.3. The numerical solver for sparse regression will be presented in Section 2.4. The efficacy of sparse regression is demonstrated by several examples in Section 2.5. Finally, we summarize our findings in Section 2.6.

2.2 Dictionaries of Basis Functions

For both wafer-level variation and within-die variation, we need to construct a dictionary of basis functions in order to capture spatial patterns for a large number of physical effects. In this section, we propose two possible dictionaries of basis functions. The first dictionary includes basis functions based on actual physical effects. Different basis functions within this dictionary can correspond to different physical variation sources. Therefore, by observing the actual basis functions selected and the amount of variation explained by basis functions that correspond to different physical sources, additional insights can be provided to further narrow down the major physical sources of variation. The second dictionary is constructed by the basis functions from Discrete Cosine Transform (DCT). This dictionary is based on different signatures of spatially correlated variation and random variation in the frequency domain.

Specifically, spatially correlated variation typically carries a unique sparse structure within the coefficients, and therefore it can be extracted by identifying this structure. The basis functions themselves do not have clear physical meaning, but it can be used to complement the first dictionary, if the physical effect of an actual process/design is not accurately captured within the first dictionary.

2.2.1 Physical Dictionary

For basis functions to model wafer-level variation, each basis function is a function of the die's location on the wafer: $f(x_{die}, y_{die})$. For basis functions to model within-die variation, each basis function is a function of the measurement site's location on the die: $f(x_{site}, y_{site})$. In the remainder of this chapter, for the sake of simplicity in writing the expressions, we will simply use x and y to designate the spatial location. The actual physical meaning of x and y will be explained in the context. In the following, we will describe a number of basis functions based on actual physical effects of wafer-level and within-die variation respectively, which form the physical basis function dictionary. Note that the physical basis function dictionary does not need to be restricted to the functions to be introduced below. In practice, with improved understanding on the manufacturing process, more basis functions can be constructed and added to the physical dictionary.

2.2.1.1 Wafer-level Variation

We first consider basis functions of wafer-level variation. First of all, we include the full quadratic model in the wafer-level physical dictionary, which contains the following six basis functions:

$$f_0(x, y) = 1 \tag{2.9}$$

$$f_1(x, y) = x \tag{2.10}$$

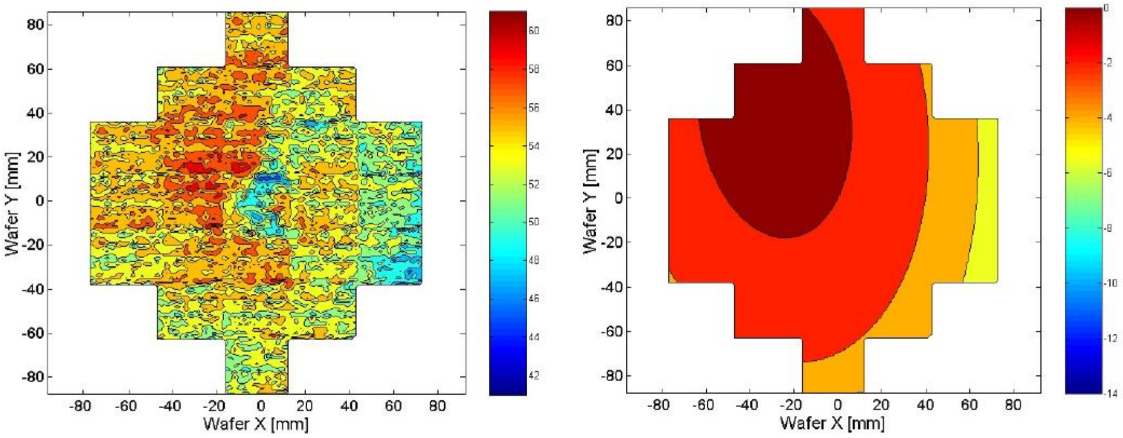
$$f_2(x, y) = y \tag{2.11}$$

$$f_3(x, y) = x^2 \tag{2.12}$$

$$f_4(x, y) = y^2 \tag{2.13}$$

$$f_5(x, y) = xy \tag{2.14}$$

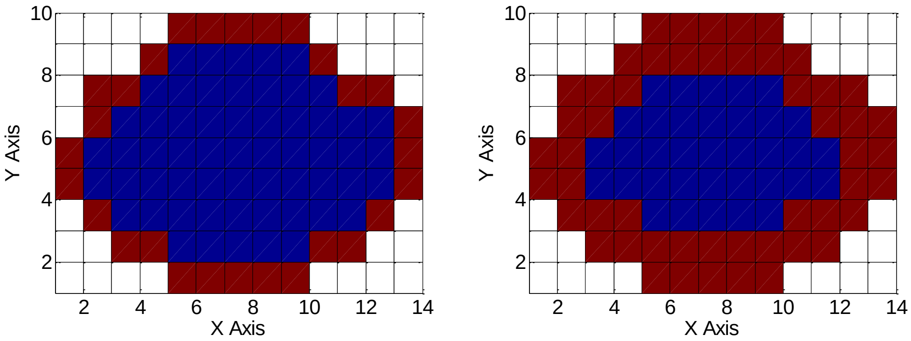
It has been shown that a large number of physical effects can be modeled using a quadratic model of *x* and *y*, such as post-exposure baking (PEB) temperature related CD variation [30]-[32], etching temperature related CD variation [30]-[33], overlay error [35], and deposition rate variation of chemical vapor deposition (CVD) [37]. A quadratic wafer-level pattern on silicon wafers has also been observed by a number of test structure measurements including electrical linewidth metrology (ELM) measurements of gate length [6], ring oscillator frequency [30] and NMOS/PMOS transistor leakage current [30]. An example of the quadratic model is shown in Figure 2-1, where the gate CD measurements on a wafer after removing within-die spatially correlated variation can be modeled using a quadratic function of *x* and *y* [6].



(a) (b)

Figure 2-1. (a) Wafer-level CD measurement map and (b) Spatially correlated variation extracted using a quadratic model [6].

In addition to effects that can be modeled using a quadratic function, it is observed that edge dies of a wafer are often substantially different from other parts of the wafer [35][62][63]. Process condition near wafer edge is usually less well-controlled, which is partly because the density of patterns near wafer edge is substantially different compared to the rest of the wafer. The etching process is known to often generate the edge effect [38]-[40] [44]. Also, rapid thermal annealing is sensitive to pattern density within millimeter-scale interaction distance [41][42], and therefore it can produce significant edge effect [44]-[46].



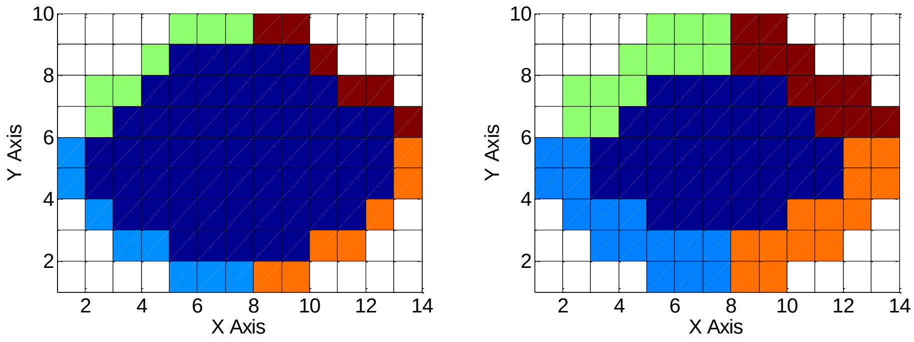
(a) (b)

Figure 2-2. (a) Depth 1 edge of a wafer and (b) Depth 2 edge of a wafer, where edge dies are marked in red.

In order to model the edge effect, indicator functions can be applied. In general, indicator functions have the following form:

$$f(x,y)=\begin{cases}1 & (x,y)\in E\\ 0 & otherwise\end{cases} \tag{2.15}$$

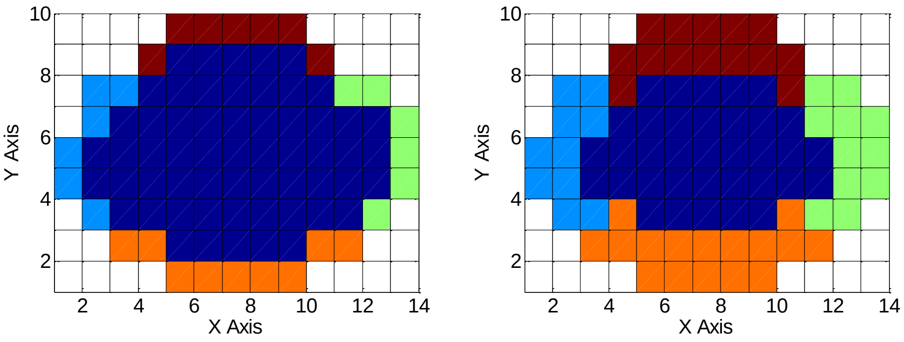
where E is a pre-defined subset of dies that belong to the edge region of the wafer. For example, an edge basis function can be defined according to Figure 2-2 (a), where a die is considered to be an edge die if one or more of its neighbors is not a valid die on the wafer. In reality, the edge effect may not affect only a single layer of dies on the immediate edge of the wafer, but may affect multiple layers of dies. To this end, we define the edge dies corresponding to Figure 2-2 (a) as the depth 1 edge of a wafer, and define edge dies with larger depth recursively: a die belongs to the depth i edge of a wafer, if itself or one of its neighbors belong to depth i-1 edge of a wafer. An example of depth 2 edge of the same wafer is shown in Figure 2-2 (b). In the physical dictionary, edge basis functions with different depth can be included, and the actual basis function that best matches a particular process can be automatically selected by the sparse regression algorithm in Section 2.3-2.4.



(a) (b)

Figure 2-3. (a) Depth 1 edge and (b) Depth 2 edge of a wafer divided into 4 regions.

The basis functions in Figure 2-2 can accurately model the edge effect if it uniformly affects all edge dies on a wafer. However, strong non-uniformity often presents in wafer measurement data: we may only observe edge effect in a portion of edge dies, and edge effect at different regions of a wafer can be different. In order to more accurately capture edge effect under such non-uniformity, we further partition the edge dies of a wafer into multiple regions, and construct a separate basis function for each region of the wafer. Again, in the dictionary, we can allow different configurations for performing the partition, and the actual basis functions that best describes a particular process will be automatically selected by the sparse regression algorithm in Section 2.3-2.4. For example, one method of performing the partition is shown in Figure 2-3, where the wafer edge is divided into 4 regions in the top-left, top-right, bottom-left and bottom-right directions. For each depth, this will generate 4 basis functions. For *K* different depth settings, a total of 4*K* basis functions can be generated. An alternative partitioning of the basis functions is shown in Figure 2-4, where the wafer is partitioned into top, bottom, left and right regions, yield another 4*K* basis functions. In practice, all basis functions in Figure 2-2, Figure 2-3 and Figure 2-4 can be included in the physical dictionary. One or more of these basis functions can be automatically selected by the sparse regression algorithm for a particular process.



(a) (b)

Figure 2-4. (a) Depth 1 edge and (b) Depth 2 edge of a wafer divided into 4 regions with different partitioning.

Other than the quadratic and edge effects, it has been observed that the center region of a wafer can be significantly different from other parts of the wafer. Physically, it can arise due to a number of sources, such as resist spinning and ion implantation [57]. Similarly, we can construct basis functions in the form of indicator functions to capture the center effect:

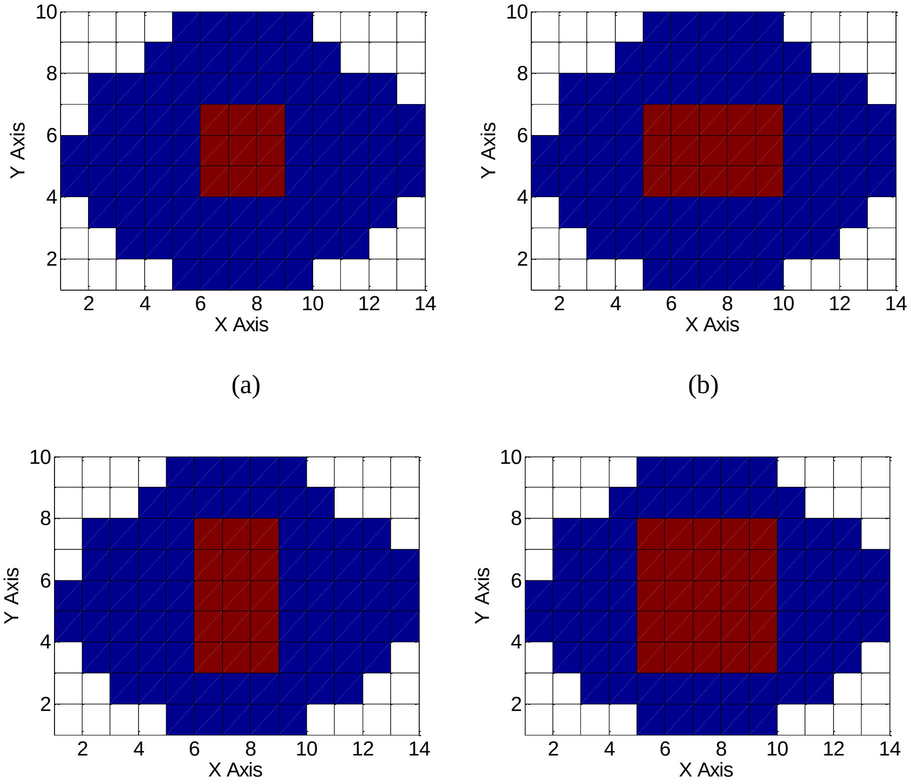
$$f(x,y)=\begin{cases}1 & (x,y)\in C\\ 0 & otherwise\end{cases} \tag{2.16}$$

where *C* is a pre-defined subset of dies that belong to the center region of the wafer. In practice, it is difficult to uniquely define what dies belong to the center region in advance. Therefore, multiple basis functions corresponding to different center region definitions can be included in the physical dictionary. To construct the center basis functions, we start from a small region in the center of the wafer, and then gradually expand the center region in the x- and y- direction. Namely, in the first basis function, we define the center region as the spatial locations (*x*, *y*) that satisfy the following criterion:

$$x\in\begin{cases}\left\{\frac{P-1}{2},\frac{P+1}{2},\frac{P+3}{2}\right\} & odd\ P\\ \left\{\frac{P}{2},\frac{P}{2}+1\right\} & even\ P\end{cases} \tag{2.17}$$

$$y\in\begin{cases}\left\{\frac{Q-1}{2},\frac{Q+1}{2},\frac{Q+3}{2}\right\} & odd\ Q\\ \left\{\frac{Q}{2},\frac{Q}{2}+1\right\} & even\ Q\end{cases} \tag{2.18}$$

where *P* is the number of dies along the x-direction, and *Q* is the number of dies along the y-direction. Figure 2-5 (a) shows an example of the first center basis function on a 13×9 wafer, where the center region is defined as a 3×3 region located in the exact center of the wafer. Next, we construct three more basis functions by symmetrically expanding the first basis function in the x-direction, y-direction, and both directions respectively. Figure 2-5 (b)-(d) show the three resulting basis functions expanded from Figure 2-5 (a). More basis functions can be expanded from Figure 2-5 (d) for larger wafers. For a particular process, if center effect exists, the most suitable basis function will be automatically selected from these candidates by the sparse regression algorithm.



(a)

(b)

(c)

(d)

Figure 2-5. Four different center region definitions of the same wafer.

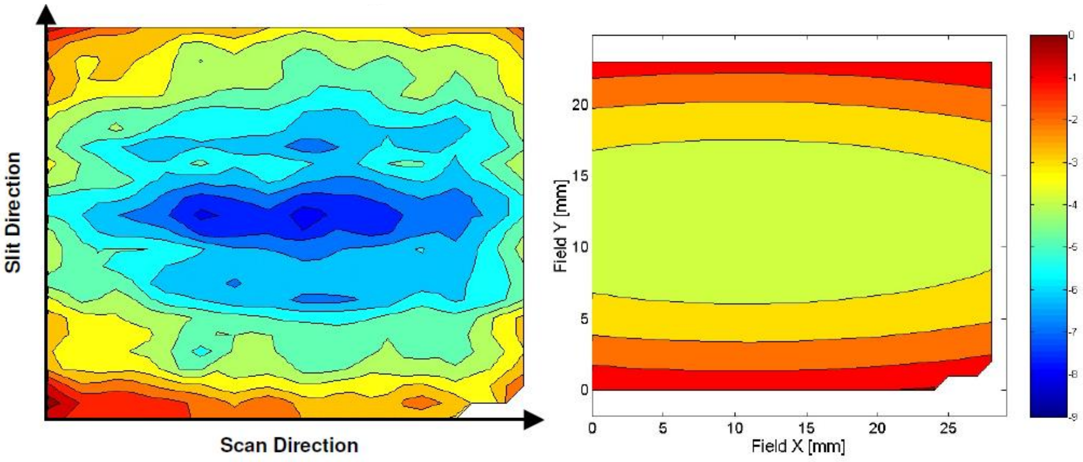
2.2.1.2 Within-die Variation

Process variation within the same die can be layout dependent. Therefore, if the measurements collected within die are not from test structures of the same layout, layout dependent variation can be estimated by constructing basis functions that account for the layout differences. One possible method of modeling layout dependent variation is again by using indicator functions:

$$f_i(x, y) = \begin{cases} 1 & (x, y) \in L_i \\ 0 & otherwise \end{cases} \quad (i = 1,2,..., N) \tag{2.19}$$

where L_i is a set of measurements collected from test structures with layout style *i*. Suppose that there are *N* different layout styles for the within-chip test structures, *N* different basis functions can be used to model the systematic difference in performance caused by different layout styles. In practice, if layout-dependent variation is considered to be more important than other spatial variation sources, the layout-dependent variation can be extracted by these basis functions and removed from the measurement data before applying variation decomposition. Otherwise, these basis functions can be simply added to the within-die physical dictionary and the amount of variation due to layout-dependent effects will be automatically determined within the variation decomposition process.

After the layout dependent variation is considered, devices within the same die across different locations can also present spatially correlated variations due to a number of variation sources, such as exposure dose variation, lens aberrations, etc. These variations can often be modeled using a quadratic function [34]. For example, it is shown in [6] that within-die gate CD variation can be modeled using a quadratic function of *x* and *y*. It is further analyzed in [47] that the variation in the along-slit direction is typically larger than the along-scan direction, and is usually symmetric in nature due to the symmetry of the components in the exposure tool. Therefore, second-order terms are often needed to model within-die variation. In our physical dictionary, the basis functions of a full quadratic model (2.9)-(2.14) are included in the physical dictionary for within-die variation. Note that although the full quadratic model only contains 6 basis functions, in practice it is possible for the number of within-die measurement sites to be very small. In this case, it is still beneficial to apply the sparse regression algorithm to combat the over-fitting problem.



(a) (b)

Figure 2-6. (a) Within-die CD measurement map and (b) Spatially correlated variation extracted using a quadratic model [6].

In Section 2.2.1.1 and 2.2.1.2, we have developed physical basis functions for wafer-level and within-die variation based on several physical variation sources. In practice, more physical basis functions can be defined and included in the physical dictionary, as we gain understanding on more physical variation sources.

2.2.2 DCT Dictionary

We would like to model the spatially correlated variation using physical basis functions as much as possible, because they provide direct insight about the source of variation. However, in practice, because the manufacturing process contains a huge number of process steps and we do not have complete understanding for all of them, it is impossible to model all possible variation sources using physical basis functions. In this sub-section, we introduce another dictionary of discrete cosine transform (DCT) basis functions borrowed from image processing literature to complement the physical dictionary. We will first construct the DCT dictionary based on DCT transform, and then explain why this dictionary can be used to decompose spatially correlated and random variation.

Let $b(x, y)$ be a two-dimensional function representing the spatial variation of interest, where x and y indicate spatial coordinates of either wafer-level or within-die variation. In practice, the sampling points to measure the spatial variation $b(x, y)$ are often collected on a regular grid. In this case, without loss of generality, we can label the x and y in $b(x, y)$ as integer numbers: $x \in \{1, 2, ..., P\}$ and $y \in \{1, 2, ..., Q\}$. The discrete cosine transform (DCT) is a two-dimensional orthogonal linear transform that maps the spatial variation $\{b(x, y); x = 1, 2, ..., P, y = 1, 2, ..., Q\}$ to the frequency domain [23]:

$$D(u,v) = \sum_{x=1}^{P}\sum_{y=1}^{Q}\alpha_u \cdot \beta_v \cdot b(x,y) \cdot \cos\frac{\pi(2x-1)\cdot(u-1)}{2\cdot P} \cdot \cos\frac{\pi(2y-1)\cdot(v-1)}{2\cdot Q} \tag{2.20}$$

where

$$\alpha_u = \begin{cases} \sqrt{1/P} & (u=1) \\ \sqrt{2/P} & (2 \le u \le P) \end{cases} \tag{2.21}$$

$$\beta_v = \begin{cases} \sqrt{1/Q} & (v=1) \\ \sqrt{2/Q} & (2 \le v \le Q) \end{cases} . \tag{2.22}$$

In (2.20), $\{D(u, v); u = 1, 2, ..., P, v = 1, 2, ..., Q\}$ represents the DCT coefficients (i.e., the frequency-domain components) of the spatial variation function $b(x, y)$. Equivalently, the function $\{b(x, y); x = 1, 2, ..., P, y = 1, 2, ..., Q\}$ can be represented as the linear combinations of $\{D(u, v); u = 1, 2, ..., P, v = 1, 2, ..., Q\}$ by inverse discrete cosine transform (IDCT):

$$b(x,y) = \sum_{u=1}^{P}\sum_{v=1}^{Q}\alpha_u \cdot \beta_v \cdot D(u,v) \cdot \cos\frac{\pi(2x-1)(u-1)}{2\cdot P} \cdot \cos\frac{\pi(2y-1)(v-1)}{2\cdot Q} . \tag{2.23}$$

Based on (2.23), we construct the DCT dictionary by including the following PQ basis functions:

$$\begin{gathered} f_{u,v}(x,y) = \alpha_u \cdot \beta_v \cdot \cos\frac{\pi(2x-1)(u-1)}{2\cdot P} \cdot \cos\frac{\pi(2y-1)(v-1)}{2\cdot Q} \\ (u=1,2,...,P; v=1,2,...,Q) \end{gathered} . \tag{2.24}$$

Next, we will explain why decomposition of spatially correlated variation and random variation can be achieved by using the DCT dictionary. We first express this decomposition problem using the following equation:

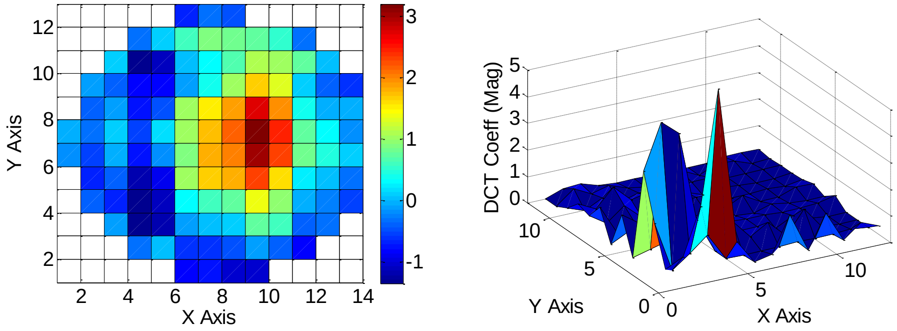
$$b(x,y) = s(x,y) + r(x,y) \tag{2.25}$$

where $s(x, y)$ stands for the spatially correlated variation and $r(x, y)$ stands for the random variation. Since the DCT transform (2.20) is a linear transform [23], the variation decomposition (2.25) can be equivalently performed in the frequency domain:

$$D(u,v) = S(u,v) + R(u,v) \tag{2.26}$$

where $\{S(u, v); u = 1, 2, ..., P, v = 1, 2, ..., Q\}$ and $\{R(u, v); u = 1, 2, ..., P, v = 1, 2, ..., Q\}$ denote the DCT coefficients of the spatially correlated variation $s(x, y)$ and the uncorrelated random variation $r(x, y)$ defined in (2.25). Once $S(u, v)$ and $R(u, v)$ are found, $s(x, y)$ and $r(x, y)$ can be determined by IDCT, similar to the case in (2.23).

An important property of the DCT coefficients is that if the spatial variation exhibits a strong spatial pattern, the DCT coefficients are sparse. For example, Figure 2-7 (a) shows a wafer map of 117 ring oscillators distributed over different spatial locations. Since ring oscillators use a large number of stages to average out the random variation [48], the wafer-level variation should be dominated by spatially correlated variation. A strong spatial pattern can be intuitively seen from Figure 2-7 (a), and as a result, its DCT coefficients in Figure 2-7 (b) only contain a small number of coefficients with large magnitude, while other coefficients are close to zero. This unique property of sparseness has been observed in many image processing tasks and serves as the key component in the compression algorithm in JPEG [67]. Also, it has motivated the compressed sensing research for image recovery using a minimum number of samples [8]-[14].



(a) (b)

Figure 2-7. (a) Measured ring oscillator (RO) period values (normalized by a randomly selected constant) of 117 ROs from the same wafer. (b) Discrete cosine transform (DCT) coefficients (magnitude) of the normalized RO period measurement show a unique sparse pattern.

Based on this sparsity property of spatially correlated variation in DCT domain, there exist a small number of (say, λ where $\lambda << PQ$) dominant DCT coefficients to satisfy:

$$\sum_{(u,v)\in\Omega} S^2(u,v) \approx \sum_{u=1}^{P}\sum_{v=1}^{Q} S^2(u,v) \tag{2.27}$$

where Ω denotes the set of the indices of the dominant DCT coefficients for *S*(*u*, *v*). Eq. (2.27) simply implies that the total energy of all DCT coefficients {*S*(*u*, *v*); *u* = 1, 2, ..., *P*, *v* = 1, 2, ..., *Q*} are almost equal to the energy of the dominant DCT coefficients {*S*(*u*, *v*); (*u*, *v*) ∈ Ω}.

On the other hand, the uncorrelated random variation can be characterized as white noise [23] and evenly distributed among all frequencies. Therefore, given the set of indices Ω, the following equation holds:

$$\sum_{(u,v)\in\Omega} R^2(u,v) \approx \frac{\lambda}{PQ} \cdot \sum_{u=1}^{P}\sum_{v=1}^{Q} R^2(u,v)\,. \tag{2.28}$$

Because of the inequality λ << *PQ*, we have λ/*PQ* << 1 in (2.28). If the value of λ is sufficiently small (i.e., the DCT coefficients of spatially correlated variation are sufficiently sparse), such that

$$\lambda / PQ << \sum_{u=1}^{P}\sum_{v=1}^{Q} S^2(u,v) / \sum_{u=1}^{P}\sum_{v=1}^{Q} R^2(u,v)\,, \tag{2.29}$$

the following inequality holds:

$$\sum_{(u,v)\in\Omega} R^2(u,v) << \sum_{(u,v)\in\Omega} S^2(u,v)\,. \tag{2.30}$$

Based on these assumptions, if the set of dominant DCT coefficients can be identified, an accurate approximation of the DCT coefficients *S*(*u*, *v*) (corresponding to spatially correlated variation) can be expressed as:

$$\tilde{S}(u,v) = \begin{cases} D(u,v) & ((u,v)\in\Omega) \\ 0 & (otherwise) \end{cases}. \tag{2.31}$$

In other words, we simply approximate *S*(*u*, *v*) by the dominant DCT coefficients {*D*(*u*, *v*); (*u*, *v*) ∈ Ω}.Comparing (2.26)and (2.31), it can be easily proven that the approximation error of (2.31) is given by:

$$\sum_{u=1}^{P}\sum_{v=1}^{Q} \left[S(u,v) - \tilde{S}(u,v)\right]^2 = \sum_{(u,v)\in\Omega} R^2(u,v) + \sum_{(u,v)\notin\Omega} S^2(u,v)\,. \tag{2.32}$$

Given the assumptions in (2.27)-(2.30), the error terms in (2.32) are almost negligible. Applying the DCT dictionary requires knowing the set Ω of the indices of the dominant DCT coefficients. This set can be identified using the sparse regression method, which will be introduced in the next sub-section.

It should be noted that other dictionaries of basis functions that offer sparsity for spatially correlated variation have been proposed in the image processing literature, such as the wavelet basis functions. Both DCT and wavelet basis functions have been widely applied in image compression because for almost any real-world image, only a small number of DCT/wavelet coefficients are needed to accurately represent it. Especially, they are employed in the JPEG [67] and JPEG2000 [68] standards respectively. While wavelet basis functions show superior sparsity to DCT for many image examples [69][70], we found that DCT typically outperforms wavelet when representing spatial process variation measurement data. The fundamental reason is because wavelet basis functions are localized in the spatial domain while DCT basis functions are global in spatial domain. In other words, most wavelet basis functions are only constructed from measurements in a relatively small region of the wafer/die, while any DCT basis function is constructed from all measurements on the wafer/die. An important difference of images with spatial process variation is that images typically have very short correlation distance, i.e. the color of a pixel is typically only strongly correlated with other pixels within a small neighborhood. On the other hand, many physical sources of process variation will impact the whole wafer, such as temperature gradients in chemical vapor deposition [37], post exposure baking [30]-[32] and etching [30]-[33]. To model these variations with long correlation distance, the DCT basis functions will be more effective.

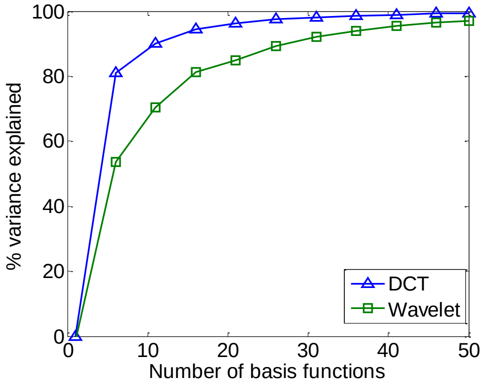


Figure 2-8.Comparison of DCT and wavelet basis functions on explaining variance of spatial variation based on ring oscillator period measurement data.

To demonstrate the superiority of DCT over wavelet basis functions, we compare the sparsity of these two dictionaries by comparing their ability to explain the variance of spatial variation on the silicon measurement data example in Figure 2-7 (a). To obtain the percentage of explained variance in Figure 2-8, we approximate the spatial variation in Figure 2-7 (a) with a list of different λ values (number of dominant basis functions) and calculate the percentage of explained variance according to the following equation based on the definition of coefficient of determination in statistics [71]:

$$R_{\lambda}^{2} = 1 - \frac{Var(e_{\lambda})}{Var(B)} \tag{2.33}$$

where B stands for all the measurement data in Figure 2-7 (a), and

$$e_{\lambda} = B - B_{\lambda} \tag{2.34}$$

stands for the residual after subtracting the spatial variation B_{λ} approximated by λ dominant basis functions. The wavelet dictionary selected in Figure 2-8 is the level 3 Haar wavelets [72]. Using other wavelets does not significantly change the results in Figure 2-8. From Figure 2-8, it can be seen that the percentage of explained variance grows quickly with the first 5-10 basis functions selected in both DCT and wavelet dictionaries, which show a large portion of variation can indeed be represented by a small number of basis functions for both dictionaries. With the same number of basis functions, DCT explains more variance than wavelets, which shows that DCT offers better sparsity than wavelets. Similar results can be observed for other silicon measurement data.

While spatially correlated variation can be modeled by a small number of dominant basis functions from the DCT dictionary, the dominant DCT basis functions themselves do not carry significant physical meaning. Therefore, unlike the physical dictionary, it is not possible to narrow down the potential variation sources based on which DCT basis functions are selected. However, since applying the DCT dictionary relies on little assumptions of the underlying process, as will be shown in Section 2.3, it can be used to complement the physical dictionary and identify important systematic patterns that are missed by the physical dictionary. The resulting spatial pattern generated from the DCT dictionary can still be used to help identify the variation sources by comparing it against the spatial patterns produced by various process steps/equipments.

2.3 Basis Selection via Sparse Regression

In the previous sub-sections, we learned that an important problem in variation decomposition is to apply the best basis functions to model wafer-level and within-die spatially correlated variation. In order to achieve good coverage for a large variety of variation sources, we have developed two dictionaries that contain a large number of possible basis functions which can be used to model the spatially correlated variation. However, as will be shown in the experiments in Section 2.5, directly applying all basis functions in a dictionary can lead to significant over-estimation of spatially correlated variation due to over-fitting. Therefore, for a particular process or design, the actual basis functions need to be selected from the dictionaries to achieve accurate modeling and reduce over-fitting. In this sub-section, we will address this basis selection problem by applying sparse regression. The numerical solver for sparse regression will be discussed in the next sub-section.

It has been shown in (2.25) that for any wafer or die measured, the spatial variation $b(x, y)$ can be represented as the summation of spatially correlated variation $s(x, y)$ and random variation $r(x, y)$. In practice, measurements for the spatial variation can be collected from multiple wafers and/or dies, so that we can represent their spatial variation using a set of two-dimensional functions: $\{b_{(l)}(x, y); l = 1,2,\ldots L\}$, where L denotes the total number of wafers/dies. Each spatial variation function contains two components:

$$b_{(l)}(x,y) = s_{(l)}(x,y) + r_{(l)}(x,y) \tag{2.35}$$

where $s_{(l)}(x, y)$ stands for the spatially correlated variation and $r_{(l)}(x, y)$ stands for the uncorrelated random variation for wafer/die l, respectively.

We first model the spatially correlated variation using basis functions from the physical dictionary. The physical dictionary is prioritized over the DCT dictionary, because its basis functions carry significant physical meaning that can be utilized to further analyze the physical variation sources. To this end, we represent $s_{(l)}(x, y)$ as a linear combination of all basis functions from the physical dictionary:

$$b_{(l)}(x,y) = \sum_{j=1}^{M_{phys}} \eta_{phys(l),j} \cdot A_{phys,j}(x,y) + r_{(l)}(x,y) \tag{2.36}$$

where the spatially correlated variation is represented by all M_{phys} basis functions $\{A_{phys,j}(x, y); j = 1, 2, \ldots, M_{phys}\}$ in the physical dictionary through coefficients $\{\eta_{phys(l),j}; j = 1, 2, \ldots, M_{phys}\}$. Since we wish to identify the subset of basis functions that are relevant to a particular process/product, the coefficients are further required to be sparse. In other words, lots of the coefficients must be 0 in (2.36).

To solve the model in (2.36), the performance of interest is measured at a number of spatial locations. In this work, for the sake of simplicity, we directly regard these measurements as samples for the spatial variation $\{b_{(l)}(x_i, y_i); i = 1, 2, \ldots, N_{(l)}\}$. Although the measurements differ from the variation by an average or nominal value, for the purpose of variation decomposition, shifting the measurements by a constant does not alter the results for basis function selection or variation percentages, so that we will not explicitly distinguish measurements with variation. We would like to estimate the sparse coefficients $\{\eta_{phys(l),j}; j = 1, 2, \ldots, M_{phys}\}$ from such measurement data. Therefore, we formulate the following *sparse regression* problem:

$$\begin{array}{ll} \underset{\eta_{phys(l)}}{\text{minimize}} & \left\| A_{phys(l)} \eta_{phys(l)} - B_{(l)} \right\|_2^2 \\ s.t. & \left\| \eta_{phys(l)} \right\|_0 \le \lambda_{phys} \end{array} \quad (l = 1,2,\ldots,L) \qquad (2.37)$$

where $B_{(l)} = [b_{(l)}(x_1, y_1)\ \ b_{(l)}(x_2, y_2)\ \ldots\ b_{(l)}(x_{N_{(l)}}, y_{N_{(l)}})]^T$ is a vector of spatial variation measurements, $\eta_{phys(l)} = [\eta_{phys(l),1}\ \eta_{phys(l),2}\ \ldots\ \eta_{phys(l),M_{phys}}]^T$ is a vector of coefficients for physical basis functions, and $A_{phys(l)}$ is matrix where $A_{phys(l),ij}$ represents the value of the j-th physical basis function at the i-th measurement location. The symbol $||\bullet||_2$ stands for the L_2-norm (i.e., the square root of the summation of the squares of all elements) of a vector, and $||\bullet||_0$ stands for the L_0-norm (i.e., the number of non-zeros) of a vector. The cost function indicates that we would like to fit the measurement data with least-squares error. On the other hand, the constraint controls the *sparsity* of $\eta_{(l)}$, which means out of all possible M_{phys} candidates in the dictionary, there exists a small subset of λ_{phys} basis functions that are applied to model the spatially correlated variation. Therefore, the meaning of (2.37) is to select the best λ_{phys} basis functions to model the spatially correlated variation. The number of basis functions λ_{phys} explores the trade-off between two types of errors: an overly small λ_{phys} will not adequately fit the spatially correlated variation, while an overly large λ_{phys} will fit a significant portion of uncorrelated random variation as spatially correlated. The numerical solver for the problem, as well as the cross-validation method to determine λ_{phys}, will be discussed in Section 2.4.

The optimization (2.37) is solved to select the basis functions for wafer-level and within-die spatially correlated variations respectively, and then the linear mixed model (2.4) is formulated using these selected basis functions and solved using the REML method. The extracted wafer-level spatially correlated variation in (2.4) will be a linear combination of quadratic, edge and center effect basis functions:

$$\gamma_{j(kl)} = \sum_{m=1}^{\lambda_{quad}} A_{quad,m}(x_{die,j}, y_{die,j}) \cdot \alpha_m + \sum_{m=1}^{\lambda_{edge}} A_{edge,m}(x_{die,j}, y_{die,j}) \cdot \alpha_m + \sum_{m=1}^{\lambda_{center}} A_{center,m}(x_{die,j}, y_{die,j}) \cdot \alpha_m + \gamma^r_{j(kl)}. \qquad (2.38)$$

Similar to (2.5)-(2.8), we are able to estimate the sample variance for quadratic, edge and center effects respectively as σ^2_{quad}, σ^2_{edge} and σ^2_{center}. Next, we are able to similarly calculate the contribution of each effect. For example, the contribution of quadratic effect in wafer-level spatially correlated variation is σ^2_{quad} / (σ^2_{quad} + σ^2_{edge}+ σ^2_{center}). Since quadratic, edge and center effects can correspond to different physical variation sources, this further decomposition will allow process engineers to further prioritize the investigation of process steps that cause the wafer-level systematic variation.

As discussed in Section 2.2, it may not be sufficient to model the spatially correlated variation using only the physical basis functions, since not all physical effects can be modeled by the physical dictionary. In this case, the DCT dictionary can be used to complement the physical dictionary to check if there is any significant spatial pattern that has been missed by the physical basis functions. In order to achieve this, we further represent $s_{(l)}(x, y)$ as a linear combination of selected physical basis functions and all DCT basis functions:

$$b_{(l)}(x,y) = \sum_{j\in\Omega_{phys}} \eta_{phys(l),j} \cdot A_{phys,j}(x,y) + \sum_{j=1}^{PQ} \eta_{dct(l),j} \cdot A_{dct,j}(x,y) + r_{(l)}(x,y) \qquad (2.39)$$

where Ω_{phys} represents the subset of physical basis functions selected by solving (2.37). In addition, the spatially correlated variation is further represented by all PQ basis functions $\{A_{dct,j}(x, y); j = 1, 2, \ldots, PQ\}$ in the DCT dictionary through coefficients $\{\eta_{dct(l),j}; j = 1, 2, \ldots, PQ\}$. Again, the DCT coefficients are further required to be sparse, leading to the following sparse regression problem:

$$\begin{array}{lll} \underset{\eta_{phys(l),\Omega_{phys}},\eta_{dct(l)}}{\text{minimize}} & \left\| A_{phys(l),\Omega_{phys}} \eta_{phys(l),\Omega_{phys}} + A_{dct(l)} \eta_{dct(l)} - B_{(l)} \right\|_2^2 & (l = 1,2,\ldots,L) \\ s.t. & \left\| \eta_{dct(l)} \right\|_0 \le \lambda_{dct} & \end{array} \qquad (2.40)$$

where $A_{phys(l),\Omega_{phys}}$ is a sub-matrix of $A_{phys(l)}$ containing the columns that belong to Ω_{phys} and $\eta_{phys(l),\Omega_{phys}}$ is a vector of the corresponding coefficients for physical basis functions; $\eta_{dct(l)} = [\eta_{dct(l),1}\ \eta_{dct(l),2}\ \ldots\ \eta_{dct(l),PQ}]^T$ is a vector of coefficients for all DCT basis functions, and $A_{phys(l)}$ is matrix where $A_{phys(l),ij}$ represents the value of the j-th DCT basis function at the i-th measurement location. Eq. (2.40) selects the best λ_{dct} additional basis functions that can be applied to model the spatially correlated variation. The linear mixed model (2.4) is again formulated using the basis functions selected by (2.40) for wafer-level and within-die variation and solved using the REML method. In the next sub-section, we will present numerical solvers that can be applied to solve the sparse regression problems (2.37) and (2.40), as well as to automatically determine the value of λ_{phys} and λ_{dct}.

2.4 Numerical Solver for Sparse Regression

We notice that these two optimization problems (2.37) and (2.40) are extremely similar in nature. Therefore, we first define a general problem of sparse regression, and show that (2.37) and (2.40) are two specific cases of the general problem. We will then present efficient numerical solvers from the statistics literature for this general problem.

We define the general sparse regression problem as follows:

$$\begin{array}{ll} \underset{\eta_{(l)}}{\text{minimize}} & \left\| A_{(l)}\eta_{(l)} - B_{(l)} \right\|_2^2 \\ s.t. & nnz\left(\left\{\eta_{(l),j} \mid j \notin \Omega_0\right\}\right) \le \lambda \end{array} \quad (l = 1,2,...,L) \tag{2.41}$$

where Ω_0 represents a set of basis functions that are pre-selected, and nnz(•) stands for the number of non-zeros within a set. It can be easily seen that (2.37) is a special case of (2.41) where:

$$A_{(l)} = A_{phys(l)} \tag{2.42}$$

$$\eta_{(l)} = \eta_{phys(l)} \tag{2.43}$$

$$\lambda = \lambda_{phys} \tag{2.44}$$

$$\Omega_0 = \{\ \}. \tag{2.45}$$

Similarly, (2.40) is a special case of (2.41), where

$$A_{(l)} = \left\lfloor A_{phys(l),\Omega_{phys}} \quad A_{dct(l)} \right\rfloor \tag{2.46}$$

$$\eta_{(l)} = \begin{bmatrix} \eta_{phys(l),\Omega_{phys}} \\ \eta_{dct(l)} \end{bmatrix} \tag{2.47}$$

$$\lambda = \lambda_{dct} \tag{2.48}$$

$$\Omega_0 = \{1,2,..., \lambda_{phys}\}. \tag{2.49}$$

Therefore, both sparse regression problems (2.37) and (2.40) can be solved if an efficient solver for the general problem (2.41) can be derived. In the following, we will present the numerical solver for the general problem in three steps. We will first introduce the numerical solver for (2.41) when *L* = 1 (i.e. there is only one wafer/die), then extend the solver to simultaneously solve multiple wafers/dies. We will finally discuss the cross-validation method to select the optimal value of *λ*.

2.4.1 Orthogonal Matching Pursuit (OMP)

We first discuss the numerical solver for a simplified problem of (2.41) when L = 1 (i.e. there is only one wafer/die), Ω_0 = {}, and λ is given in advance:

$$\begin{array}{ll} \underset{\eta}{\text{minimize}} & \|A\eta - B\|_2^2 \\ s.t. & \|\eta\|_0 \le \lambda \end{array} . \tag{2.50}$$

Even with these simplification, (2.50) remains an NP-hard problem and therefore is extremely difficult to solve. Several efficient solvers for the problem (2.50) have been discussed in the compressed sensing literature, including L1-norm relaxation [8]-[10], orthogonal matching pursuit (OMP) [11]-[14], Bayesian method [15], iteratively reweighted L2-norm method [16], etc. The stepwise regression method [104] from statistics can also be applied to solve (2.50). Of all these methods, L1-norm relaxation and OMP are popular choices, because theoretical studies have shown that their accuracy degrades gracefully with increasing amount of random variation [9][14]. We select OMP as the numerical solver for the sparse regression problem. An important reason for choosing OMP is its simplicity. As will be shown in Chapter 3 and Chapter 4, this simplicity of OMP allows us to easily adapt the algorithm for several practical needs, such as outlier detection and fast computation with the DCT dictionary. Moreover, a comparison of OMP and L1-norm relaxation in circuit performance modeling examples have been recently carried out in [17], which shows that the computational cost of OMP and L1-norm relaxation are similar in practice, with the accuracy of OMP slightly superior in most cases. In the following, we will introduce the OMP algorithm to solve (2.50), and then extend the algorithm to handle non-empty Ω_0.

The key idea of OMP is to use the inner product to identify a small number of important basis functions. Namely, we re-write the matrix A by its column vectors:

$$A = \begin{bmatrix} A_1 & A_2 & \cdots & A_M \end{bmatrix} \tag{2.51}$$

where each column corresponds to a different basis function. The inner product between B and a basis function A_j is then defined as:

$$\left\langle B, A_j \right\rangle = \sum_{i=1}^{N} b_i \cdot A_{ij} \tag{2.52}$$

where N is the number of samples, b_i is the i-th element of B and A_{ij} is the i-th element of A_j. Theoretically, the inner product $<B, A_j>$ is equivalent to the actual coefficient η_j when solving the linear equation $A\eta=B$, when the columns of *A* are orthogonal and normalized:

$$\left\langle A_j, A_k \right\rangle = \begin{cases} 1 & \left(j = k\right) \\ 0 & \left(j \neq k\right) \end{cases} \quad . \tag{2.53}$$

For our basis function dictionaries, DCT basis functions are orthogonal and normalized when there is no missing data at any spatial location, and remain approximately so when there exist some missing data. Although the physical basis functions do not guarantee any orthogonality, they can be normalized (i.e. rescale each column vector to unit length) prior to applying the sparse regression algorithm. Overall speaking, even though the ideal conditions may not be fully satisfied in practice, the inner product <*B*, A_j> still remains an important metric to measure the significance of the basis vector A_j. A large inner product between *B* and A_j implies that the basis function *j* is an important component to approximate *B*.

Given the sparse regression problem (2.50), OMP iteratively uses the inner product to select the important basis functions. In the first step, the basis function that results in the largest magnitude of inner product is selected:

$$\max_{s1} \quad \left| \left\langle B, A_{s1} \right\rangle \right| . \tag{2.54}$$

Once this basis function is selected, a least-squares problem is solved for the coefficient that corresponds to the basis function s_1:

$$\min_{\eta_{s1}} \quad \left\| B - A_{s1} \cdot \eta_{s1} \right\|_2^2 . \tag{2.55}$$

The solution of (2.55) is the best representation of the spatial variation *B* using the basis vector A_{s1}. Since (2.55) is an over-determined equation, it will result in a residual *e*, representing the spatial variation that cannot be represented by A_{s1}:

$$e = B - A_{s1} \cdot \eta_{s1} . \tag{2.56}$$

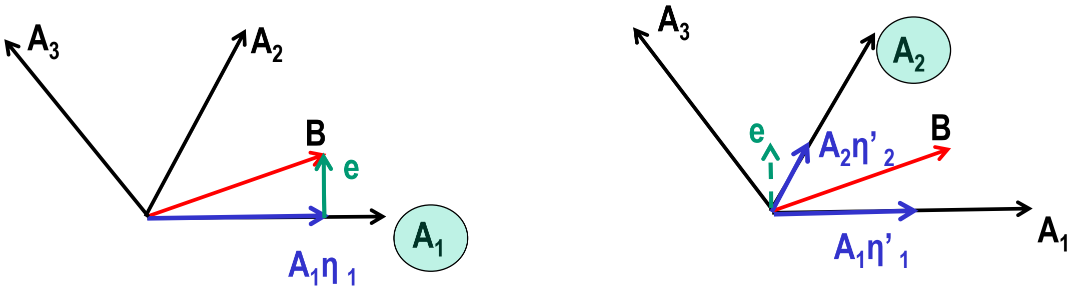
In the next iteration, OMP further identifies the next important basis function by the largest magnitude of inner product with the residual:

$$\max_{s2} \quad \left| \left\langle e, A_{s2} \right\rangle \right| . \tag{2.57}$$

Once the second basis function is selected, another least-squares problem is solved to obtain the best approximation of the spatial variation *B* using the basis function s_1 and s_2:

$$\min_{\eta_{s1},\eta_{s2}} \quad \left\| B - A_{s1} \cdot \eta_{s1} - A_{s2} \cdot \eta_{s2} \right\|_2^2 . \tag{2.58}$$

Note that in (2.58), all the coefficients are re-evaluated to minimize the total sum of squared residual, so that η_{s1} can be changed from the solution in (2.55). This is because the basis functions are not necessarily orthogonal, so that re-evaluation is needed to improve the accuracy. If more basis functions need to be selected, the OMP algorithm will repeatedly select the best basis function according to maximum inner product with the residual similar to (2.57), then re-evaluate all coefficients similar to (2.58), until λ basis functions are selected.



(a) (b)

Figure 2-9. (a) First step of the OMP algorithm on a 2-D example. (b) Second step of the OMP algorithm on a 2-D example.

To intuitively understand the OMP steps (2.54)-(2.58), we use a 2-D example in Figure 2-9 to explain its basic idea. In this example, we would like to approximate the vector B using two out of three basis vectors: A_1, A_2 and A_3. Figure 2-9 (a) shows the first step of the OMP algorithm which corresponds to (2.54)-(2.56). In this example, since A_1 has the largest correlation with B, A_1 is first selected to approximate B. The corresponding coefficient η_1 is determined by least-squares fitting, and the residual e is orthogonal to the basis vector A_1. Figure 2-9 (b) shows the second step of the OMP algorithm where the basis vector A_2 is selected because it has the strongest correlation with e, and the coefficients η_1 and η_2 are re-evaluated using least squares. The OMP algorithm stops here since two basis functions have been selected.

It is straightforward to extend the OMP algorithm to allow a set of basis functions that are pre-selected. In this case, we aim to solve the following generalized problem:

$$\begin{array}{ll}\underset{\eta}{\text{minimize}} & \left\|A\eta - B\right\|_2^2 \\ s.t. & nnz\left(\left\{\eta_j \mid j \notin \Omega_0\right\}\right) \le \lambda\end{array} \tag{2.59}$$

where Ω$_0$ may be non-empty. If the size of Ω$_0$ is a non-zero value λ_0, we can apply the OMP idea by conceptually considering λ_0 steps have been performed to the simplified problem (2.50), and the basis functions selected are in the set Ω$_0$. In this case, the coefficients that correspond to the selected basis functions can be estimated by:

$$\underset{\eta_i, i\in\Omega_0}{\text{minimize}} \quad \left\|\sum_{i\in\Omega_0} A_i \cdot \eta_i - B\right\|_2^2 , \tag{2.60}$$

and the residual can be subsequently estimated by:

$$e = B - \sum_{i\in\Omega_0} A_i \cdot \eta_i \ . \tag{2.61}$$

The OMP iterations can then be “resumed” from the residual in (2.61). The flow of the extended OMP algorithm is summarized in Algorithm 1.

Algorithm 1: Extended OMP

```
1. Start from the optimization problem in (2.59) with a given integer λ.
2. If Ω_0 = {}
       Initialize the residual $e = B$;
   Else
       Initialize the residual $e$ by (2.60)-(2.61).
3. Initialize the set Ω = Ω_0, and the iteration index $p = 1$.
4. Select the new basis vector $A_s$ according to the following criterion:
```

$$\underset{s}{\text{maximize}} \quad \left|\left\langle e, A_s \right\rangle\right| . \tag{2.62}$$

```
5. Update Ω by Ω = Ω∪{$s$}.
6. Solve the least-squares fitting:
```

$$\underset{\eta_i, i\in\Omega}{\text{minimize}} \quad \left\|\sum_{i\in\Omega} A_i \cdot \eta_i - B\right\|_2^2 . \tag{2.63}$$

```
7. Calculate the residual:
```

$$e = B - \sum_{i \in \Omega} A_i \cdot \eta_i \; . \tag{2.64}$$

8. If $p < \lambda$, $p = p + 1$ and go to Step 4.

9. For any $i \notin \Omega$, set $\eta_i = 0$.

2.4.2 Simultaneous Orthogonal Matching Pursuit (S-OMP)

We further derive the numerical solver for the following generalized problem when there exist multiple wafers/dies:

$$\begin{array}{ll} \underset{\eta_{(l)}}{\text{minimize}} & \left\| A_{(l)} \eta_{(l)} - B_{(l)} \right\|_2^2 \\ s.t. & nnz\left(\left\{ \eta_{(l),j} \mid j \notin \Omega_0 \right\}\right) \le \lambda \end{array} \quad (l = 1, 2, ..., L) \; . \tag{2.65}$$

This can be viewed as L sparse regression problems with single wafer/die as in (2.59). Therefore, a naïve method to solve (2.65) is to solve these L problems independently using Algorithm 1. Because of random variation, each problem may result in a different subset of basis functions $\Omega_{(l)}$. Since we would like to identify the systematic variation sources for these wafers/dies, we require these L problems to result in a common subset of basis functions:

$$\Omega_{(1)} = \Omega_{(2)} = \cdots = \Omega_{(L)} = \Omega \; . \tag{2.66}$$

In order to achieve this goal, we further borrow the Simultaneous Orthogonal Matching Pursuit (S-OMP) algorithm [14] from statistics literature to solve a common set of basis functions from (2.65).

With (2.66) in mind, we re-visit the OMP algorithm (i.e., Algorithm 1) where a set of dominant basis functions are selected to approximate the spatially correlated variation. At each iteration of Algorithm 1, a single basis function is chosen according to the inner product in (2.62). For S-OMP, since the index set of dominant coefficients is shared for L different wafers/dies as shown in (2.66), we use the linear combination of multiple inner products as a quantitative criterion for basis vector selection:

$$\underset{s}{\text{maximize}} \quad \sum_{l=1}^{L} \left| \left\langle Res_{(l)}, A_{(l),s} \right\rangle \right| . \tag{2.67}$$

Eq. (2.67) is expected to be more accurate than applying (2.62) to any individual wafer/die, since it is less sensitive to the random noise caused by uncorrelated random variation and/or measurement error. In other words, by adding the inner products over L wafers/dies, the impact of random noise is reduced and the spatial pattern associated with systematic variation can be accurately detected.

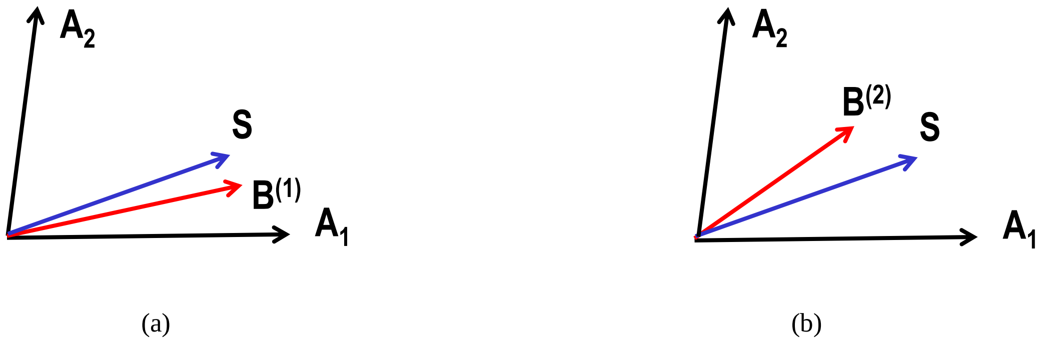


Figure 2-10. A 2-D example of using S-OMP to select the correct basis vector from two choices.

To intuitively explain (2.67), Figure 2-10 shows a 2-D example where a vector S indicating the underlying spatially correlated variation is shared by two data sets in Figure 2-10 (a) and Figure 2-10 (b). Because of random variation, the observed spatial variation is different from S. The observed variation is denoted $B^{(1)}$ and $B^{(2)}$ in the two data sets respectively. Given $B^{(1)}$ and $B^{(2)}$, we would like to select one out of two basis vectors, A_1 and A_2, that best models the spatially correlated variation S. If (2.62) is applied to these two data sets independently, we will get the following result:

$$\left|\left\langle B^{(1)}, A_1 \right\rangle\right| > \left|\left\langle B^{(1)}, A_2 \right\rangle\right| \tag{2.68}$$

$$\left|\left\langle B^{(2)}, A_1 \right\rangle\right| < \left|\left\langle B^{(2)}, A_2 \right\rangle\right| \tag{2.69}$$

where (2.69) leads to the incorrect conclusion that A_2 is preferred over A_1 because of significant random variation. On the other hand, applying (2.67) yields:

$$\left|\left\langle B^{(1)}, A_1 \right\rangle\right| + \left|\left\langle B^{(2)}, A_1 \right\rangle\right| > \left|\left\langle B^{(1)}, A_2 \right\rangle\right| + \left|\left\langle B^{(2)}, A_2 \right\rangle\right| \tag{2.70}$$

which leads to the correct conclusion that A_1 should be selected. This is because the systematic variation is shared across multiple data sets and the random variation can cancel out during the adding process. Therefore, (2.67) will more stably detect the correct basis functions compared to any individual data set. After applying (2.67) to select the basis function in each iteration, least-squares fitting is applied to each wafer/die to solve the coefficients that correspond to the selected basis functions, which is identical to the Algorithm 1. The extension of S-OMP to allow a set of pre-selected basis functions can be performed following the same idea as Algorithm 1.

Algorithm 2 summarizes the major steps of the aforementioned extended S-OMP algorithm. Note that Algorithm 2 can be viewed as a generalized version of Algorithm 1. If there is only one wafer/die (i.e., L = 1), Algorithm 2 is exactly equivalent to Algorithm 1.

Algorithm 2: Extended S-OMP

1. Start from the optimization problem in (2.65) with a given integer λ.
2. If $\Omega_0 = \{\}$

   Initialize the residuals $e_{(l)} = B_{(l)}$;

   Else

   Solve the following L least-squares fitting problems:

$$\underset{\eta_{(l),i},\, i\in\Omega_0}{\text{minimize}} \quad \left\| \sum_{i\in\Omega_0} A_{(l),i} \cdot \eta_{(l),i} - B_{(l)} \right\|_2^2 \quad (l = 1,2,...,L) \tag{2.71}$$

   and initialize the residuals:

$$e_{(l)} = B_{(l)} - \sum_{i\in\Omega_0} A_{(l),i} \cdot \eta_{(l),i} \quad \left(l = 1,2,...,L\right) \quad . \tag{2.72}$$

3. Initialize the set $\Omega = \Omega_0$, and the iteration index p = 1.
4. Select the new basis vector s according to (2.67).
5. Update Ω by $\Omega = \Omega \cup \{s\}$.
6. Solve the following L least-squares fitting problems:

$$\underset{\eta_{(l),i},\, i\in\Omega}{\text{minimize}} \quad \left\| \sum_{i\in\Omega} A_{(l),i} \cdot \eta_{(l),i} - B_{(l)} \right\|_2^2 \quad (l = 1,2,...,L) \quad . \tag{2.73}$$

7. Calculate the following L residuals:

$$e_{(l)} = B_{(l)} - \sum_{i\in\Omega} A_{(l),i} \cdot \eta_{(l),i} \quad \left(l = 1,2,...,L\right) \quad . \tag{2.74}$$

8. If $p < \lambda$, $p = p + 1$ and go to Step 4.
9. For any $i \notin \Omega$, set $\eta_{(l),i} = 0$.

2.4.3 Cross-Validation

The extended S-OMP algorithm (i.e., Algorithm 2) relies on a user defined parameter λ to control the number of basis functions that should be selected. In practice, λ is not known in advance. The appropriate value of λ must be determined by considering the following two important issues. First, if λ is too small, Algorithm 2 cannot select a sufficient number of basis functions to represent the spatially correlated variation, thereby leading to large modeling error. On the other hand, if λ is too large, Algorithm 2 can incorrectly select too many coefficients and some of these coefficients are associated with uncorrelated random variation, instead of spatially correlated variation. It, again, results in large modeling error due to over-fitting. In order to achieve the best accuracy, we must accurately estimate the modeling error for different λ values and then find the optimal λ with minimum error.

In practice, it is extremely difficult to directly estimate the modeling error, since the amount of spatially correlated variation is not known in advance. We cannot simply measure the modeling error from all sampling data, since it will always monotonically decrease with larger number of basis functions, leading to the over-fitting problem. However, comparison of modeling error can be made based on the prediction error: we intentionally leave out a small random portion of measurements as a testing set, and use the other measurements to estimate the spatially correlated variation. The estimated spatially correlated variation for the testing set is compared against its actual measurements. The idea behind this approach is, since spatially correlated variation is the only predictable component within the total variation, the optimal prediction accuracy will be achieved when the spatially correlated variation is best modeled.

In this paper, we adopt the cross-validation method [24] to estimate the modeling error for our variation decomposition application. An F-fold cross-validation partitions the entire data set into F groups. Modeling error is estimated according to the cost function in (2.65) from F independent runs. In each run, one of the F groups is used to estimate the modeling error and all other groups are used to calculate the model coefficients. Note that the training data for coefficient estimation and testing data for error estimation are not overlapped. Hence, over-fitting can be easily detected. In addition, different groups should be selected for error estimation in different runs. As such, each run results in an error value ε_f (f = 1,2,...,F) that is measured from a unique group of data points. The final modeling error is computed as the average of $\{\varepsilon_f; f = 1, 2,...,F\}$, i.e., $\varepsilon = (\varepsilon_1 + \varepsilon_2 + ... + \varepsilon_F)/F$.

2.5 Numerical Results

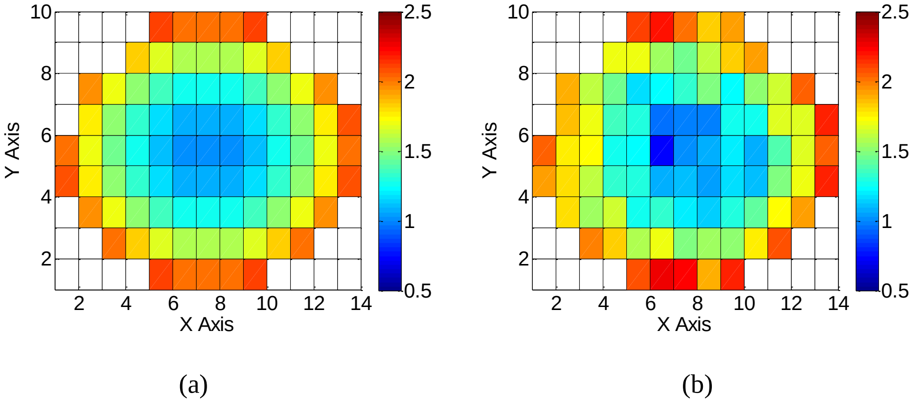
In this section, we demonstrate the efficacy of our proposed variation decomposition algorithm using several synthetic examples. We only show synthetic examples in this chapter because real-world silicon measurement data often contains outliers. If these outliers are not appropriately considered, they will introduce substantial error in the variation decomposition results. We will introduce techniques for detecting and removing these outliers in the next chapter and then show the results on the silicon data.

2.5.1 Quadratic Basis Effects

We first consider several wafer-level examples where the spatially correlated variation can be modeled by using quadratic basis functions. Figure 2-11 shows a synthetic wafer where the systematic variation is created by the following function:

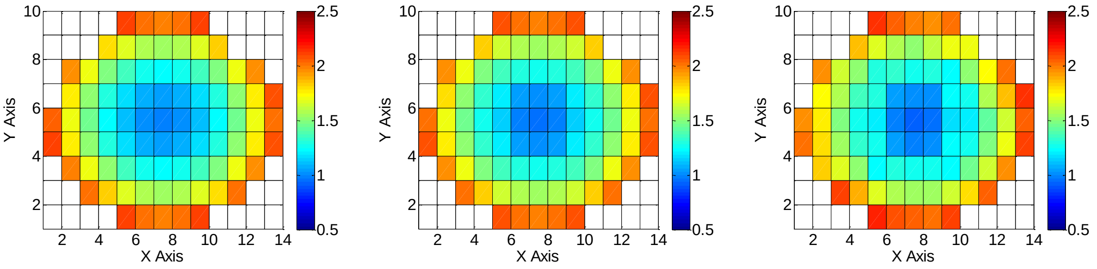
$$s(x,y) = 1 + x^2 + y^2 \tag{2.75}$$

where x and y are coordinates on the wafer with range normalized to [-1 1]. The systematic wafer map is shown in Figure 2-11 (a). The synthetic data is created by adding a small amount of random variation distributed as $N(0, 0.01)$, which is shown in Figure 2-11 (b). After adding the random variation, the systematic variation contributes to 89.6% of the total variance.



(a) (b)

Figure 2-11. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation.

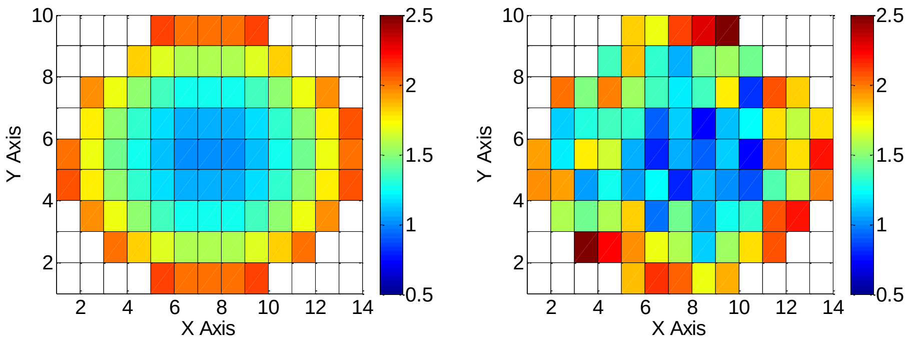


(a) (b) (c)

Figure 2-12. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.

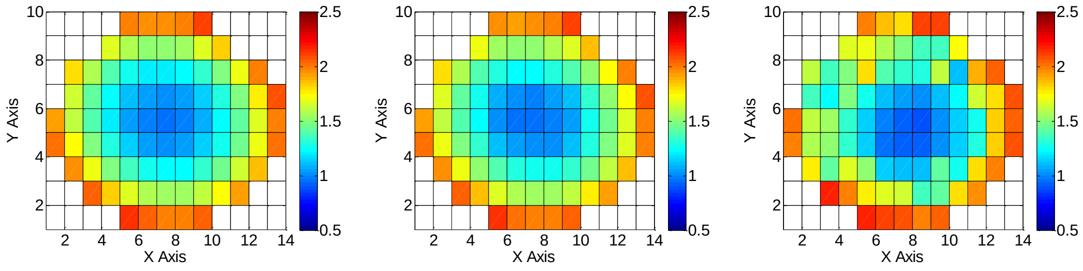
Figure 2-12 compares spatially correlated variation extracted by three methods. Figure 2-12 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary, where the basis functions are determined by sparse regression with the physical dictionary shown in (2.9)-(2.14) and Figure 2-2-Figure 2-5, and then applied in REML. The proposed method identifies 5 basis functions from the dictionary, and the estimated spatially correlated variation is 90.3%. Within the spatially correlated variation, the estimated percentages of quadratic, edge and center effects are 99.4%, 0.2% and 0.5% respectively. Although sparse regression does not completely remove over-fitting, the estimated spatially correlated variation is extremely close to the true amount of systematic variation, and the spatially correlated variation is almost entirely contributed by quadratic basis. Figure 2-12 (b) shows the spatially correlated variation extracted by directly applying REML with the 6 quadratic basis functions in (2.9)-(2.14). Since the systematic variation is created using quadratic basis functions, we expect the results in Figure 2-12 (b) to be accurate. The estimated spatially correlated variation is 90.2%, which shows similar accuracy with the proposed method. Figure 2-12 (c) shows the spatially correlated variation extracted by directly applying REML with the physical dictionary without sparse regression. After removing basis functions that are linearly dependent on other basis functions, there are 24 independent basis functions in total. The estimated spatially correlated variation is 92.4%, and the estimated percentages of quadratic, edge and center effects are 96.3%, 2.6% and 1.1% respectively. The results are less accurate than Figure 2-12 (a) because of over-fitting, but the advantage of sparse regression is not significant. The importance of sparse regression will be much clearer in the following examples where the random variation is large or the number of samples is fewer.

We further consider the same systematic variation but the synthetic data is created by adding a larger amount of random variation distributed as $N(0, 0.09)$, shown in Figure 2-13 (b). After adding the random variation, the systematic variation contributes to 59.5% of the total variance. Figure 2-14 compares spatially correlated variation extracted by three methods. Figure 2-14 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. The proposed method identifies 8 basis functions from the dictionary, and the estimated spatially correlated variation is 60.4%. The estimated percentages of quadratic, edge and center effects are 98.5%, 0.4% and 1.1% respectively. Figure 2-14 (b) shows the spatially correlated variation extracted by the 6 quadratic basis functions. The estimated spatially correlated variation is 60.5%. Therefore, the proposed method shows similar accuracy compared to directly applying the quadratic basis functions. Figure 2-14 (c) shows the spatially correlated variation extracted by the physical dictionary without sparse regression. It can be intuitively seen from Figure 2-14 (c) that the extracted spatial pattern has significant over-fitting. In this case, the estimated spatially correlated variation is 72.3%, and the estimated percentages of quadratic, edge and center effects are 65.2%, 32.5% and 2.3% respectively. Therefore, without sparse regression, the amount of spatially correlated variation is greatly over-estimated, and the results show significant edge effect which does not exist in the actual systematic variation.



(a) (b)

Figure 2-13. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding large random variation.

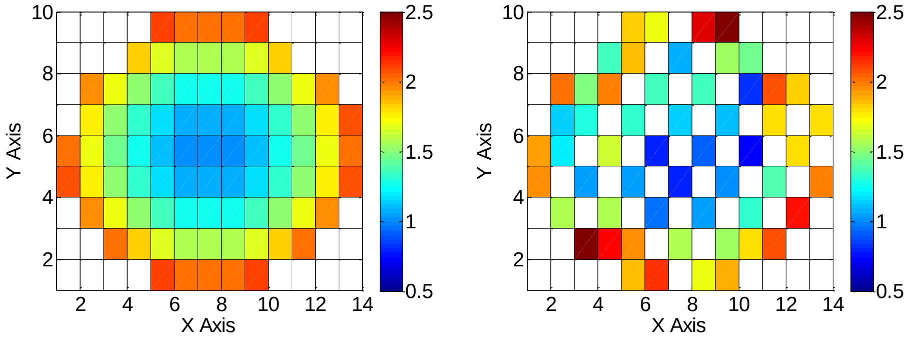


(a)

(b)

(c)

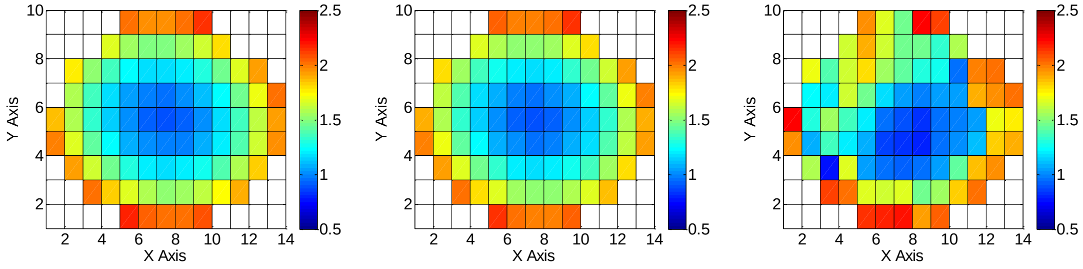
Figure 2-14. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.



(a)

(b)

Figure 2-15. (a) Systematic variation of the synthetic wafer. (b) Synthetic data with checkerboard sampling.



(a)

(b)

(c)

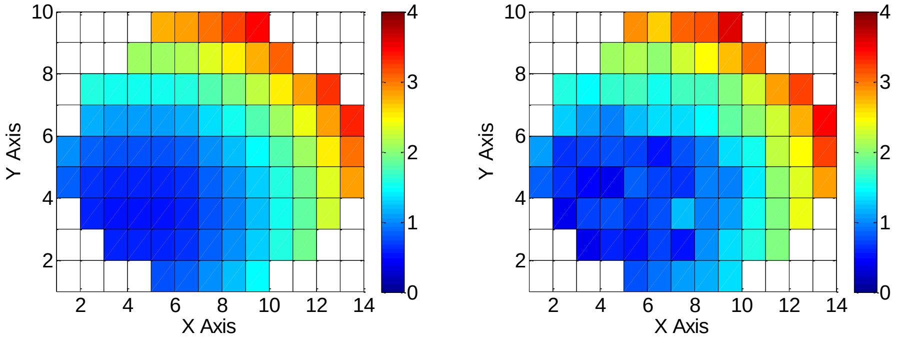
Figure 2-16. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.

In the previous example, we showed that with significant random variation, directly applying the large physical dictionary and cause significant over-fitting, which greatly reduces the accuracy of variation decomposition. Therefore, sparse regression is needed to reduce the over-fitting. In reality, there can exist more significant risk of over-fitting if fewer sampling points are collected. This can happen because of missing data, or intentionally skipping some test sites to reduce test cost. Figure 2-15 (b) shows a scenario where the measurements in Figure 2-13 (b) are further sampled in a "checkerboard" style, resulting in about half the number of sampling points compared to performing full sampling. Figure 2-16 compares spatially correlated variation extracted by three methods. Figure 2-16 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. The proposed method identifies 6 basis functions from the dictionary, and the estimated spatially correlated variation is 60.4%. The estimated percentages of quadratic, edge and center effects are 99.9%, 0.0% and 0.0% respectively. Figure 2-14 (b) shows the spatially correlated variation extracted by the 6 quadratic basis functions. The estimated spatially correlated variation is 60.6%. Therefore, the proposed method shows similar accuracy compared to directly applying the quadratic basis functions. Figure 2-14 (c) shows the spatially correlated variation extracted by directly applying 23 independent basis functions from the physical dictionary without sparse regression. The resulting spatial wafer map shows even more over-fitting compared to Figure 2-14 (c), and the estimated spatially correlated variation is 83.0%. The estimated percentages of quadratic, edge and center effects are 49.1%, 50.6% and 0.3% respectively, which lead to the incorrect conclusion that the spatially correlated variation is dominated by edge effect. Therefore, it is necessary to apply sparse regression to reduce over-fitting.

Figure 2-17 shows another synthetic wafer where the systematic variation is created by the following function:

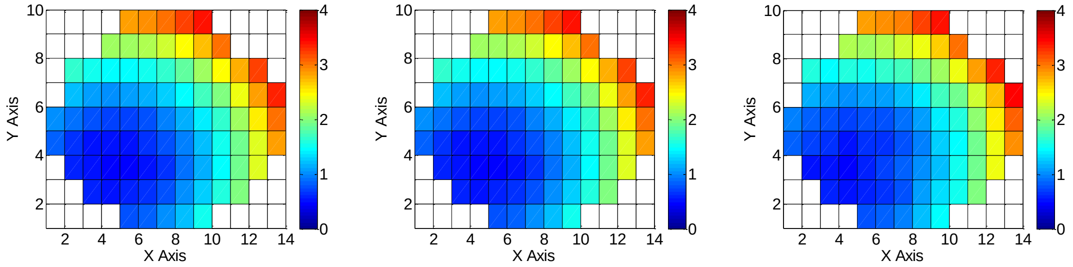
$$s(x, y) = 1 + x + y + x^2 + y^2 \, . \tag{2.76}$$

where x and y are coordinates on the wafer with range normalized to [-1 1]. The systematic wafer map is shown in Figure 2-17 (a). The synthetic data is created by adding a small amount of random variation distributed as $N(0, 0.02)$, which is shown in Figure 2-11 (b). After adding the random variation, the systematic variation contributes to 97.2% of the total variance.



(a) (b)

Figure 2-17. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation.



(a) (b) (c)

Figure 2-18. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.

Figure 2-18 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. The proposed method identifies 8 basis functions from the dictionary, and the estimated spatially correlated variation is 97.3%. The estimated percentages of quadratic, edge and center effects are 100%, 0.0% and 0.0% respectively. Figure 2-18 (b) shows the spatially correlated variation extracted by the 6 quadratic basis functions. The estimated spatially correlated variation is also 97.3%. Figure 2-18 (c) shows the spatially correlated variation extracted by directly applying 24 independent basis functions from the physical dictionary without sparse regression. The estimated spatially correlated variation is 97.7%, and the estimated percentages of quadratic, edge and center effects are 98.6%, 1.0% and 0.5% respectively. In this case, all three methods can capture the systematic variation with good accuracy.

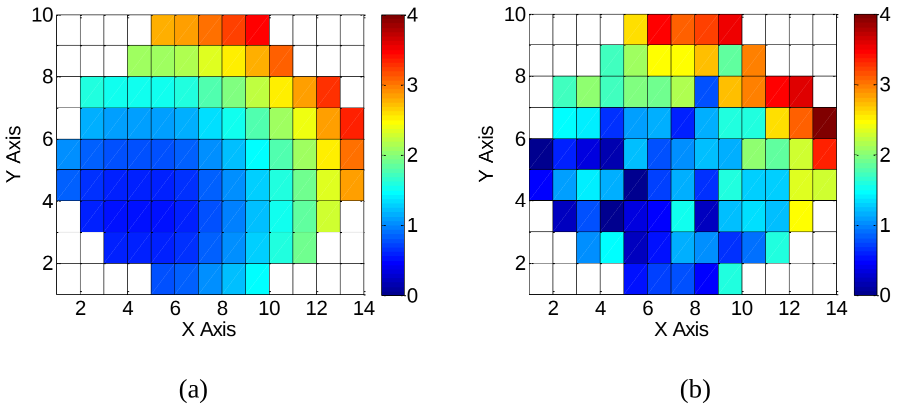
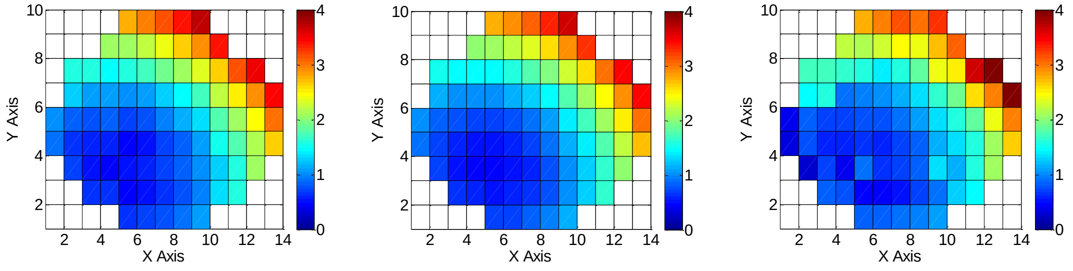


Figure 2-19. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding large random variation.

We further add a larger amount of random variation distributed as $N(0, 0.2)$ to the same systematic variation shown in Figure 2-17 (b). After adding the random variation, the systematic variation contributes to 79.6% of the total variance. Figure 2-20 compares spatially correlated variation extracted by three methods. Figure 2-20 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. The proposed method identifies 8 basis functions from the dictionary, and the estimated spatially correlated variation is 81.6%. The estimated percentages of quadratic, edge and center effects are 98.5%, 0.8% and 0.7% respectively. Figure 2-20 (b) shows the spatially correlated variation extracted by the 6 quadratic basis functions. The estimated spatially correlated variation is 80.9%. Therefore, the proposed method shows similar accuracy compared to directly applying the quadratic basis functions. Figure 2-14 (c) shows the spatially correlated variation extracted by directly applying the 24 independent basis functions from the physical dictionary without sparse regression. Significant over-fitting can be seen from Figure 2-14 (c), and the estimated spatially correlated variation is 85.4%. The estimated percentages of quadratic, edge and center effects are 84.5%, 14.4% and 1.1% respectively.



(a) (b) (c)

Figure 2-20. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.

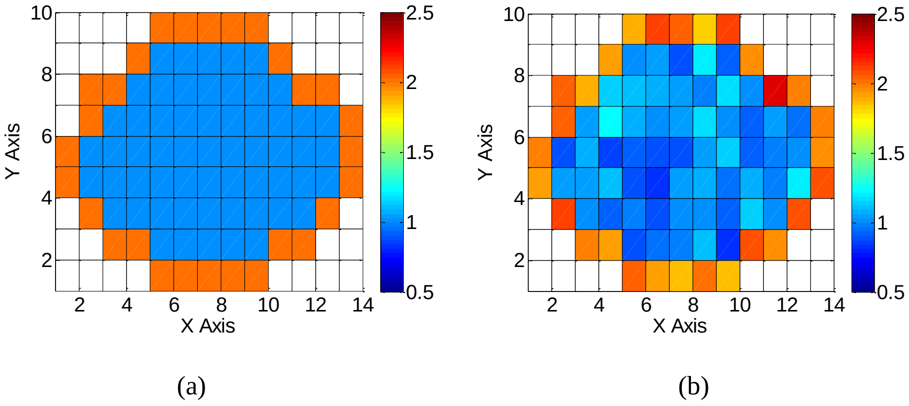
Table 2-1.Summary of results with quadratic effects.

Example	Golden		Proposed physical		Physical w/o SR		Quadratic
Radial, small variation	89.6%	Q: 100% E: 0.0% C: 0.0%	90.3%	Q: 99.4% E: 0.2% C: 0.5%	92.4%	Q: 96.3% E: 2.6% C: 1.1%	90.2%
Radial, large variation	59.5%	Q: 100% E: 0.0% C: 0.0%	60.4%	Q: 98.5% E: 0.4% C: 1.1%	72.3%	Q: 65.2% E: 32.5% C: 2.3%	60.5%
Radial, checkerboard	59.5%	Q: 100% E: 0.0% C: 0.0%	60.4%	Q: 99.9% E: 0.0% C: 0.0%	83.0%	Q: 49.1% E: 50.6% C: 0.3%	60.6%
Asymmetric, small variation	97.2%	Q: 100% E: 0.0% C: 0.0%	97.3%	Q: 100% E: 0.0% C: 0.0%	97.7%	Q: 98.6% E: 1.0% C: 0.5%	97.3%
Asymmetric, large variation	79.6%	Q: 100% E: 0.0% C: 0.0%	81.6%	Q: 98.5% E: 0.8% C: 0.7%	85.4%	Q: 84.5% E: 14.4% C: 1.1%	80.9%

Table 2-1 summarizes the results from examples constructed by quadratic basis functions, where for each example we compare the extracted percentage of spatially correlated variation, as well as the detailed decomposition of quadratic (Q), edge (E) and center (C) effects inside the spatially correlated variation for different methods. The traditional quadratic model can only consider the quadratic effect so that its detailed decomposition will always be 100% quadratic. From these examples, we observe that the proposed method with physical dictionary achieves comparable accuracy with applying REML with the quadratic basis functions, if the systematic variation can be modeled by the quadratic function. The proposed method always achieves superior accuracy compared to directly applying REML with the physical dictionary without sparse regression because over-fitting can be significantly reduced by sparse regression. Sparse regression is especially needed when the amount of random variation is significant, or there exists missing data and/or intentionally skipped dies.

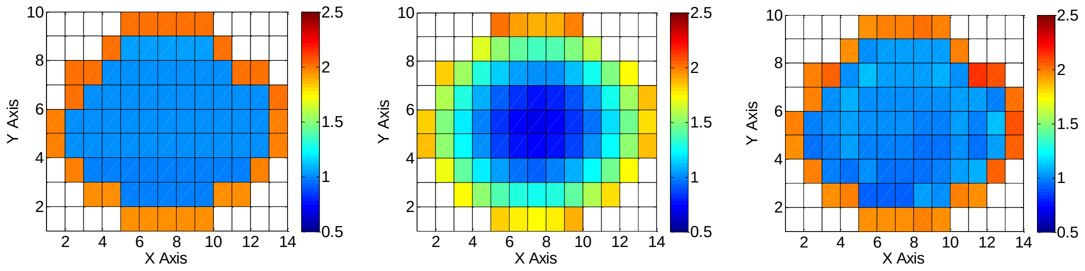
2.5.2 Center/Edge Effects

We further consider several wafer-level examples where the spatially correlated variation contains center and/or edge effects. We first create a systematic wafer map in Figure 2-21 (a), where the dies on the edge have larger performance value than other dies. The synthetic data is created by adding a small amount of random variation distributed as *N*(0, 0.01), which is shown in Figure 2-21 (b). After adding the random variation, the systematic variation contributes to 95.7% of the total variance.



(a) (b)

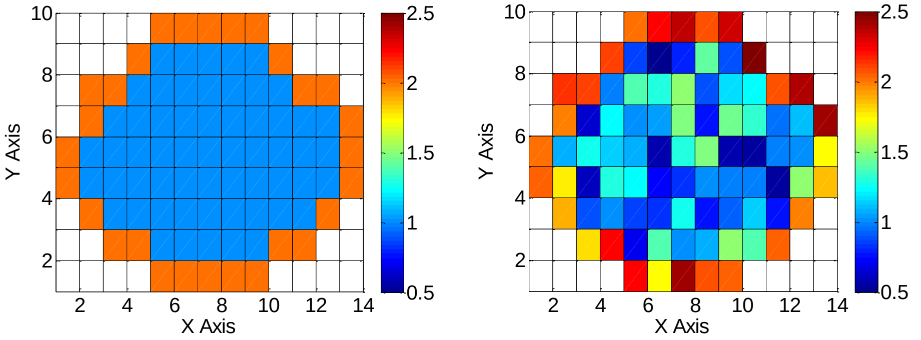
Figure 2-21. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation.



(a) (b) (c)

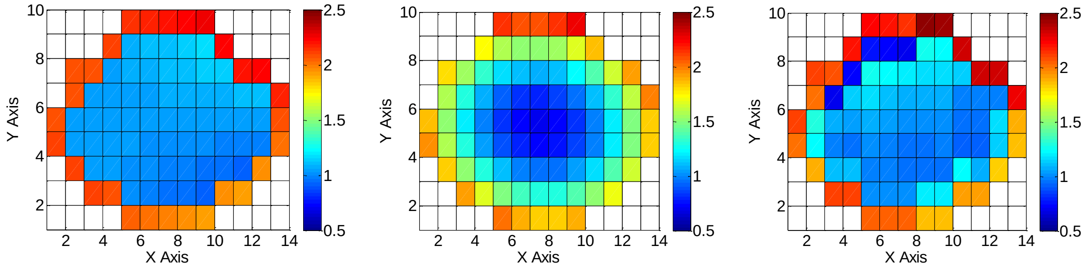
Figure 2-22. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.

Figure 2-22 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. The proposed method identifies 3 basis functions from the dictionary, and the estimated spatially correlated variation is 95.8%. The estimated percentages of quadratic, edge and center effects are 0.2%, 99.8% and 0.0% respectively. Figure 2-18 (b) shows the spatially correlated variation extracted by the 6 quadratic basis functions. The estimated spatially correlated variation is 62.8%. It can be clearly seen that the proposed method achieves excellent accuracy in modeling the edge effect. On the other hand, using quadratic basis functions will not adequately capture the edge effect, and the extracted spatially correlated wafer map will contain significant artifact in the center of the wafer, which may mislead further effort to diagnose the source of variation. Figure 2-22 (c) shows the spatially correlated variation extracted by directly applying 24 independent basis functions from the physical dictionary without sparse regression. The estimated spatially correlated variation is 96.4%, and the estimated percentages of quadratic, edge and center effects are 2.4%, 97.2% and 0.5% respectively. The edge effect is captured but the result is not as accurate as Figure 2-22 (a) because of over-fitting.



(a) (b)

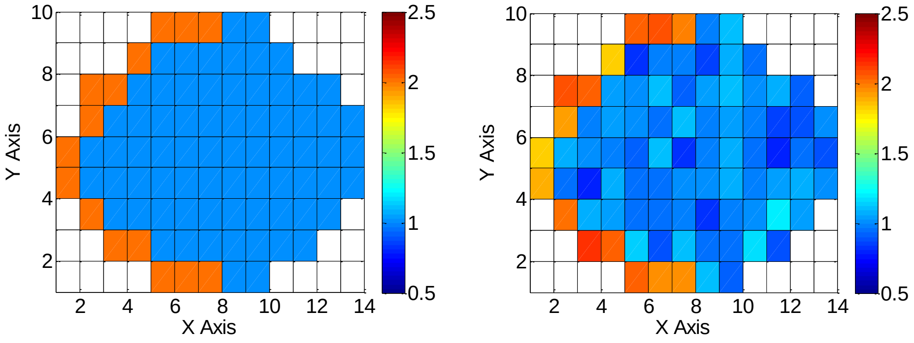
Figure 2-23. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding large random variation.



(a) (b) (c)

Figure 2-24. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.

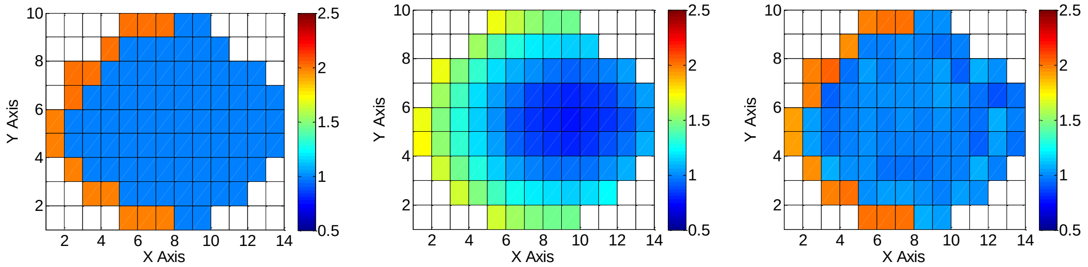
We further add a larger amount of random variation distributed as *N*(0, 0.09) to the same systematic variation shown in Figure 2-23 (b). After adding the random variation, the systematic variation contributes to 72.2% of the total variance. Figure 2-24 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. The proposed method identifies 4 basis functions from the dictionary, and the estimated spatially correlated variation is 75.8%. The estimated percentages of quadratic, edge and center effects are 2.0%, 98.0% and 0.0% respectively. Figure 2-24 (b) shows the spatially correlated variation extracted by the 6 quadratic basis functions. The estimated spatially correlated variation is 51.4%. Figure 2-24 (c) shows the spatially correlated variation extracted by directly applying 24 independent basis functions from the physical dictionary without sparse regression. The estimated spatially correlated variation is 81.4%, and the estimated percentages of quadratic, edge and center effects are 7.1%, 92.5% and 0.4% respectively. It can be clearly seen that the proposed method achieves good accuracy in modeling the edge effect. Using quadratic basis functions does not adequately capture the edge effect and creates artifacts, while not using sparse regression causes inaccuracy due to over-fitting.



(a) (b)

Figure 2-25. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation.

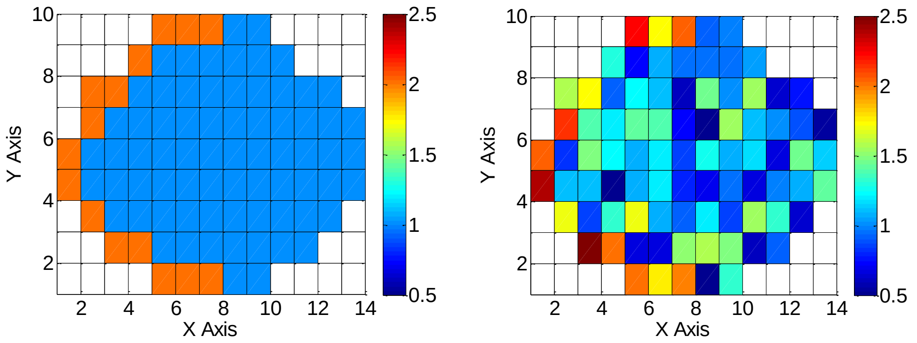
Figure 2-25 shows another synthetic wafer where the systematic variation is created by edge effect on only the left half of the wafer. The systematic wafer map is shown in Figure 2-25 (a). The synthetic data is created by adding a small amount of random variation distributed as $N(0, 0.01)$, which is shown in Figure 2-25 (b). After adding the random variation, the systematic variation contributes to 94.9% of the total variance. Figure 2-26 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. The proposed method identifies 3 basis functions from the dictionary, and the estimated spatially correlated variation is 94.9%. The estimated percentages of quadratic, edge and center effects are 0.0%, 100.0% and 0.0% respectively. Figure 2-26 (b) shows the spatially correlated variation extracted by the 6 quadratic basis functions. The estimated spatially correlated variation is 50.1%. It can be seen that the proposed method achieves excellent accuracy. On the other hand, using quadratic basis functions can only capture extremely weak edge effect, and creates a strong artificial pattern in the right side of the wafer. Figure 2-26 (c) shows the spatially correlated variation extracted by directly applying 24 independent basis functions from the physical dictionary without sparse regression. The estimated spatially correlated variation is 95.9%, and the estimated percentages of quadratic, edge and center effects are 2.8%, 97.0% and 0.3% respectively. The edge effect is captured but the result is not as accurate as Figure 2-24 (a) because of over-fitting.



(a) (b) (c)

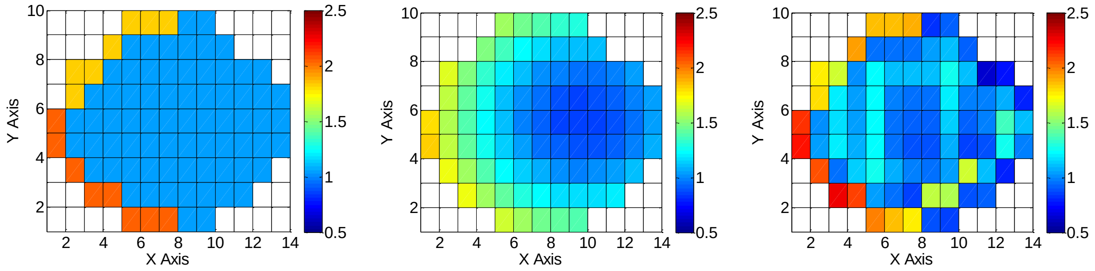
Figure 2-26. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.

We further add a larger amount of random variation distributed as *N*(0, 0.09) to the same systematic variation shown in Figure 2-27 (b). After adding the random variation, the systematic variation contributes to 59.3% of the total variance. Figure 2-28 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. The proposed method identifies 3 basis functions from the dictionary, and the estimated spatially correlated variation is 55.7%. The estimated percentages of quadratic, edge and center effects are 0.0%, 100.0% and 0.0% respectively. Figure 2-28 (b) shows the spatially correlated variation extracted by the 6 quadratic basis functions. The estimated spatially correlated variation is 29.9%. Figure 2-28 (c) shows the spatially correlated variation extracted by directly applying 24 independent basis functions from the physical dictionary without sparse regression. The estimated spatially correlated variation is 69.2%, and the estimated percentages of quadratic, edge and center effects are 16.7%, 78.9% and 4.5% respectively. It can be clearly seen that the proposed method achieves good accuracy in modeling the edge effect. Using quadratic basis functions does not adequately capture the edge effect and creates artifacts, while not using sparse regression causes inaccuracy due to over-fitting.



(a) (b)

Figure 2-27. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding large random variation.



(a) (b) (c)

Figure 2-28. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.

Figure 2-29 shows another synthetic wafer where the systematic variation is created by center effect. The systematic wafer map is shown in Figure 2-29 (a). The synthetic data is created by adding a small amount of random variation distributed as $N(0, 0.01)$, which is shown in Figure 2-29 (b). After adding the random variation, the systematic variation contributes to 93.9% of the total variance. Figure 2-30 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. The proposed method identifies 2 basis functions from the dictionary, and the estimated spatially correlated variation is 93.6%. The estimated percentages of quadratic, edge and center effects are 0.0%, 0.0% and 100.0% respectively. Figure 2-30 (b) shows the spatially correlated variation extracted by the 6 quadratic basis functions. The estimated spatially correlated variation is 38.4%. It can be seen that the proposed method achieves excellent accuracy. On the other hand, using quadratic basis functions can only capture extremely weak center effect, and creates a strong artificial pattern in the edge of the wafer. Figure 2-30 (c) shows the spatially correlated variation extracted by directly applying 24 independent basis functions from the physical dictionary without sparse regression. The estimated spatially correlated variation is 94.9%, and the estimated percentages of quadratic, edge and center effects are 3.2%, 3.3% and 93.6% respectively. The edge effect is captured but the result is not as accurate as Figure 2-30 (a) because of over-fitting.

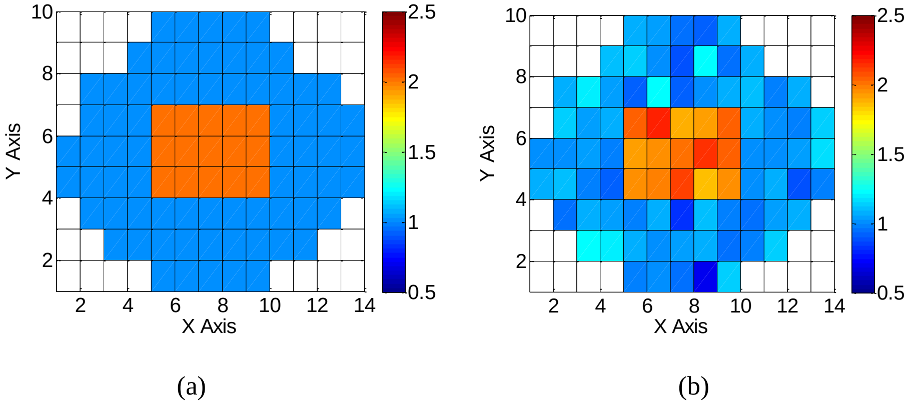


Figure 2-29. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation.

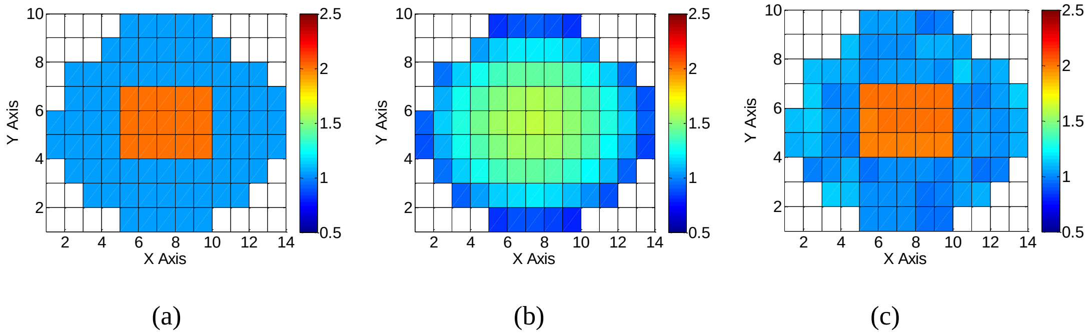


Figure 2-30. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.

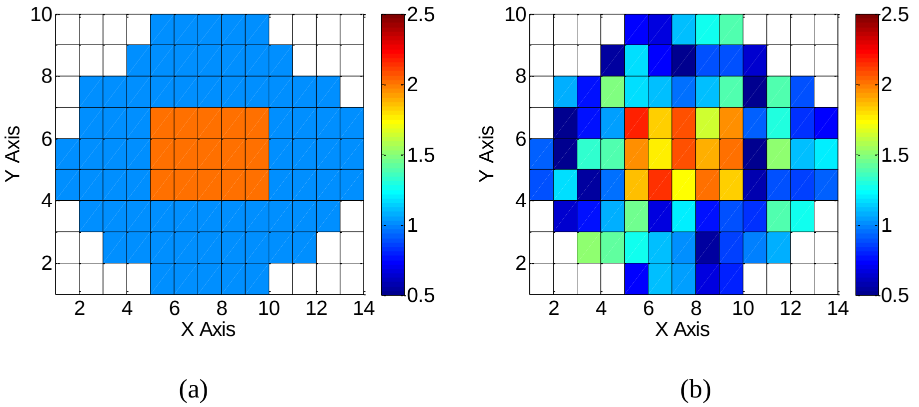


Figure 2-31. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding large random variation.

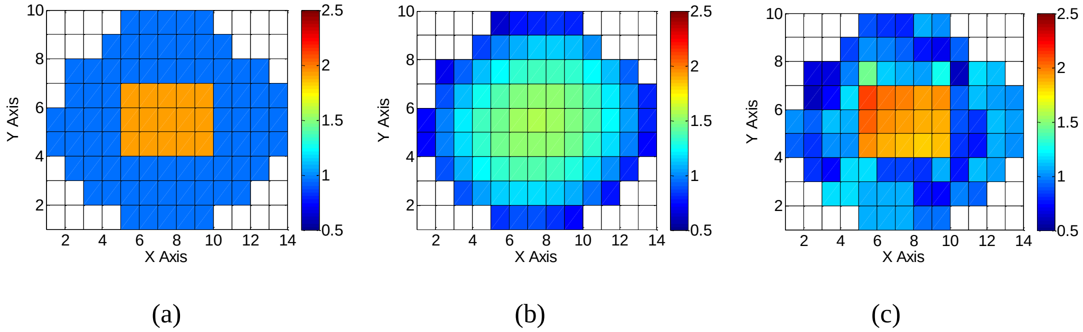


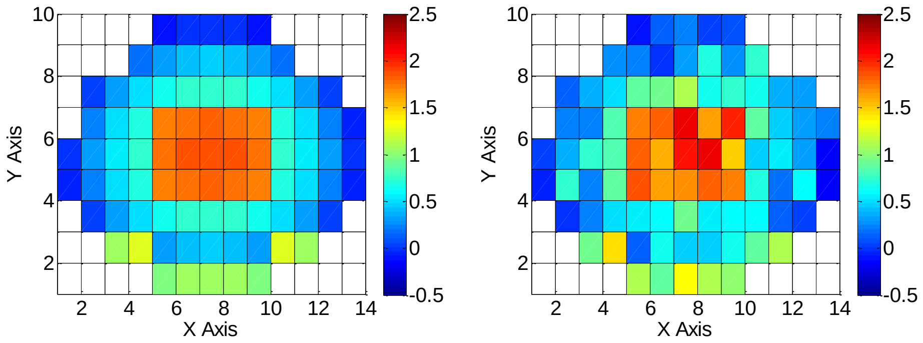
Figure 2-32. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.

We further add a larger amount of random variation distributed as *N*(0, 0.09) to the same systematic variation shown in Figure 2-31 (b). After adding the random variation, the systematic variation contributes to 65.5% of the total variance. Figure 2-32 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. The proposed method identifies 2 basis functions from the dictionary, and the estimated spatially correlated variation is 64.0%. The estimated percentages of quadratic, edge and center effects are 0.0%, 0.0%, and 100.0% respectively. Figure 2-32 (b) shows the spatially correlated variation extracted by the 6 quadratic basis functions. The estimated spatially correlated variation is 30.2%. Figure 2-32 (c) shows the spatially correlated variation extracted by directly applying 24 independent basis functions from the physical dictionary without sparse regression. The estimated spatially correlated variation is 75.0%, and the estimated percentages of quadratic, edge and center effects are 14.5%, 27.1% and 58.3% respectively. It can be clearly seen that the proposed method achieves good accuracy in modeling the center effect. Using quadratic basis functions does not adequately capture the center effect and creates artifacts, while not using sparse regression causes inaccuracy due to over-fitting.

We finally show another synthetic wafer where the systematic variation is created by a combination of quadratic, edge and center effects. The quadratic pattern is created by

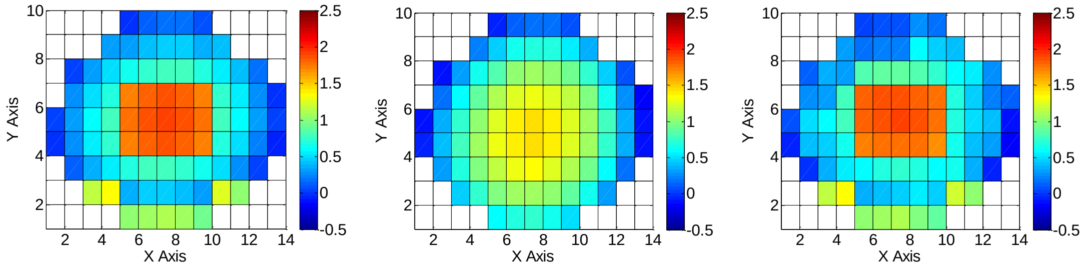
$$s(x, y) = 1 - x^2 - y^2 \tag{2.77}$$

which forms a decreasing radial pattern from center to edge. The edge effect only occurs at the bottom edge of the wafer, and the center effect is similar to Figure 2-29 (a). These three components are rescaled so that each component contributes to one third of the variance in systematic variation. The systematic wafer map is shown in Figure 2-33 (a). The synthetic data is created by adding a small amount of random variation distributed as *N*(0, 0.04), which is shown in Figure 2-33 (b). After adding the random variation, the systematic variation contributes to 90.3% of the total variance. Figure 2-34 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. The proposed method identifies 11 basis functions from the dictionary, and the estimated spatially correlated variation is 90.6%. The estimated percentages of quadratic, edge and center effects are 36.3%, 32.1% and 31.5% respectively. Figure 2-30 (b) shows the spatially correlated variation extracted by the 6 quadratic basis functions. The estimated spatially correlated variation is 54.5%. It can be seen that the proposed method achieves excellent accuracy in estimating the percentage of spatially correlated variation, and the estimated percentages of each component are close to the actual percentage. On the other hand, using quadratic basis functions does not adequately capture the center and edge effects. Figure 2-34 (c) shows the spatially correlated variation extracted by directly applying 24 independent basis functions from the physical dictionary without sparse regression. The estimated spatially correlated variation is 92.6%, and the estimated percentages of quadratic, edge and center effects are 30.1%, 38.8% and 30.3% respectively. The result is not as accurate as Figure 2-34 (a) because of over-fitting.



(a) (b)

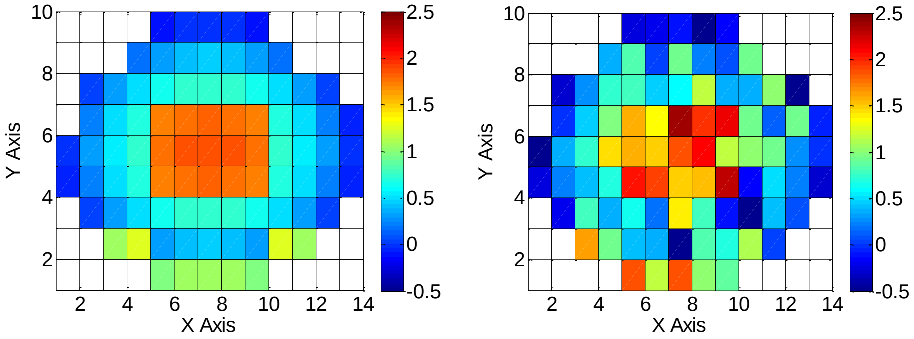
Figure 2-33. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation.



(a) (b) (c)

Figure 2-34. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.

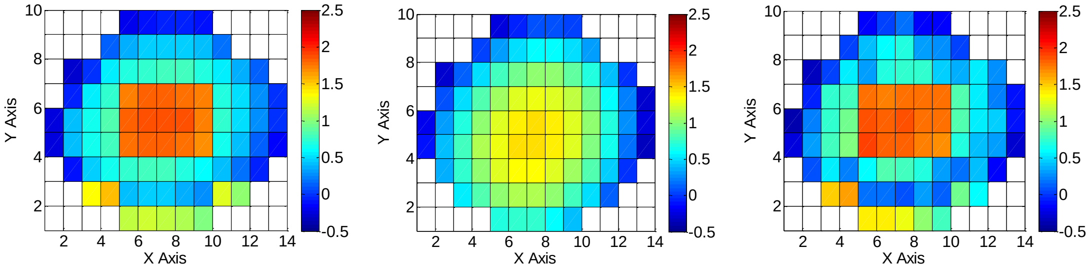
We further add a larger amount of random variation distributed as $N(0, 0.16)$ to the same systematic variation shown in Figure 2-35 (b). After adding the random variation, the systematic variation contributes to 65.9% of the total variance. Figure 2-36 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. The proposed method identifies 11 basis functions from the dictionary, and the estimated spatially correlated variation is 70.1%. The estimated percentages of quadratic, edge and center effects are 32.2%, 39.6%, and 28.2% respectively. Figure 2-36 (b) shows the spatially correlated variation extracted by the 6 quadratic basis functions. The estimated spatially correlated variation is 43.2%. Figure 2-36 (c) shows the spatially correlated variation extracted by directly applying 24 independent basis functions from the physical dictionary without sparse regression. The estimated spatially correlated variation is 74.7%, and the estimated percentages of quadratic, edge and center effects are 40.8%, 34.6% and 24.6% respectively. It can be clearly seen that the proposed method achieves good accuracy. Using quadratic basis functions does not adequately capture the edge/center effects, while not using sparse regression causes inaccuracy due to over-fitting.



(a)

(b)

Figure 2-35. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding large random variation.



(a)

(b)

(c)

Figure 2-36. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by quadratic basis functions. (c) Spatially correlated variation extracted by the physical dictionary without sparse regression.

Table 2-2. Summary of results with edge/center effects.

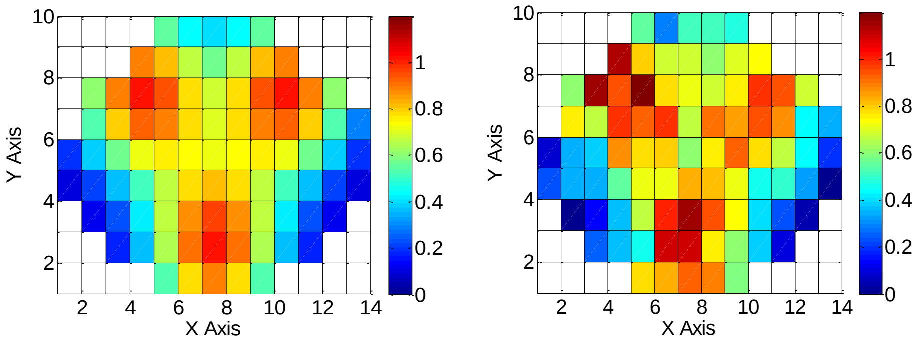
Example	Golden		Proposed physical		Physical w/o SR		Quadratic
Full edge, small variation	95.7%	Q: 0.0%	95.8%	Q: 0.2%	96.4%	Q: 2.4%	62.8%
		E: 100%		E: 99.8%		E: 97.2%	
		C: 0.0%		C: 0.0%		C: 0.5%	
Full edge, large variation	72.2%	Q: 0.0%	75.8%	Q: 2.0%	81.4%	Q: 7.1%	51.4%
		E: 100%		E: 98.0%		E: 92.5%	
		C: 0.0%		C: 0.0%		C: 0.4%	
Partial edge, small variation	94.9%	Q: 0.0%	94.9%	Q: 0.0%	95.9%	Q: 2.8%	50.1%
		E: 100%		E: 100%		E: 97.0%	
		C: 0.0%		C: 0.0%		C: 0.3%	
Partial edge, large variation	59.3%	Q: 0.0%	55.7%	Q: 0.0%	69.2%	Q: 16.7%	29.9%
		E: 100%		E: 100%		E: 78.9%	
		C: 0.0%		C: 0.0%		C: 4.5%	
Center, small variation	93.9%	Q: 0.0%	93.6%	Q: 0.0%	94.9%	Q: 3.2%	38.4%
		E: 0.0%		E: 0.0%		E: 3.3%	
		C: 100%		C: 100%		C: 93.6%	
Center, large variation	65.5%	Q: 0.0%	64.0%	Q: 0.0%	75.0%	Q: 14.5%	30.2%
		E: 0.0%		E: 0.0%		E: 27.1%	
		C: 100%		C: 100%		C: 58.3%	
Mixed, small variation	90.3%	Q: 33.3%	90.6%	Q: 36.3%	92.6%	Q: 30.1%	54.5%
		E: 33.3%		E: 32.1%		E: 38.8%	
		C: 33.3%		C: 31.5%		C: 30.3%	
Mixed, large variation	65.9%	Q: 33.3%	70.1%	Q: 32.2%	74.7%	Q: 40.8%	43.2%
		E: 33.3%		E: 39.6%		E: 34.6%	
		C: 33.3%		C: 28.2%		C: 24.6%	

Table 2-2 summarizes the results from examples containing edge and center basis functions. From these examples, we observe that the proposed method achieves good accuracy when there exist edge and center effects. In this case, the quadratic model will not adequately capture these effects, and may create undesirable artifacts that may mislead the further effort to diagnose the source of variation. The proposed method always achieves superior accuracy compared to directly applying the physical dictionary without sparse regression, which causes significant over-fitting with large random variation.

2.5.3 Complex Effects

As discussed in Section 2.3, not all physical sources of variation can be fully modeled by the physical dictionary. Therefore, after applying (2.37) to select the physical basis functions, we may further apply (2.40) which utilizes the DCT dictionary to discover any significant spatial pattern that has been missed by the physical dictionary. Since the DCT dictionary contains a large number of basis functions, an important concern is over-fitting. If there is only one wafer or die, even if cross-validation is applied, random variation may still accidentally match the pattern of some DCT basis functions. In this case, the proposed algorithm may select more basis functions than that are actually needed. This over-fitting problem can be significantly reduced, if there are multiple wafers/dies. We found that when there are 10 wafers with the same systematic pattern, for all the examples in Section 2.5.1 and 2.5.2, (2.40) will not select any additional basis function compared to (2.37), and therefore the variation decomposition results are exactly the same with applying the physical dictionary only. This result shows that applying the DCT dictionary in addition to the physical dictionary does not cause over-fitting with 10 wafers.

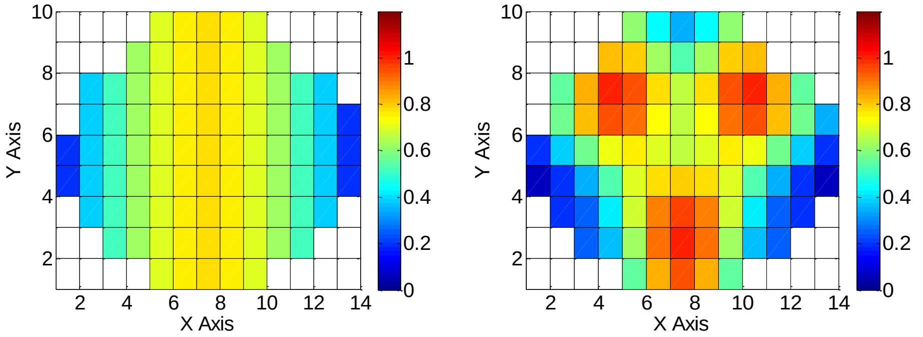
We construct several examples to examine the ability of DCT dictionary to detect complex spatial patterns that are not modeled by the physical dictionary. One possible scenario is wafer-level effects caused non-uniformity of heat sources. In many process steps such as chemical vapor deposition (CVD), thermal oxidation of gate oxide, post exposure baking (PEB) and rapid thermal annealing (RTA), the wafer being processed is heated by heat sources above the wafer. Non-uniformity of heat sources may cause non-uniform temperature distribution across the wafer, which eventually result in wafer-level spatially correlated variation. For example, Ref. [36] shows that in hot-wire CVD process where the heat source is a M-shaped hot wire above the wafer, the film thickness variation shows a spatial pattern which has three peaks at locations where the heat source presents high density. Ref. [43] shows that in a thermal oxidation process of gate oxide where there are three heat sources above the wafer, the gate oxide thickness shows three peaks directly below the heat sources. Based on these observations, we construct a systematic wafer map in Figure 2-37 (a). The systematic wafer map has three peaks similar to the examples in [36] and [43]. We create 10 wafers of synthetic data by adding a small amount of random variation distributed as $N(0, 0.01)$ on each wafer. The synthetic data for one of the wafers is shown in Figure 2-37 (b). After adding the random variation, the systematic variation contributes to 86.1% of the total variance. Figure 2-38 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. Sparse regression identifies 2 basis functions from the dictionary, and the estimated spatially correlated variation is 37.8%. Comparing Figure 2-38 (a) with Figure 2-37 (a), it can be seen that since the physical dictionary is not designed to model the pattern in Figure 2-37 (a), it fails to capture the underlying systematic pattern. Figure 2-38 (b) shows the spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries. It selects 9 DCT basis functions in addition to the physical basis functions, and the estimated spatially correlated variation is 84.9%. The extracted pattern closely matches the actual systematic pattern, which serves as a good basis for further diagnosis of the source of variation.



(a)

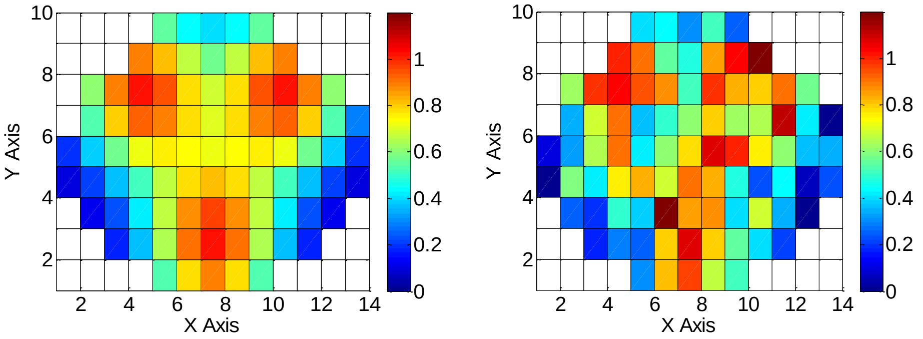
(b)

Figure 2-37. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation.



(a) (b)

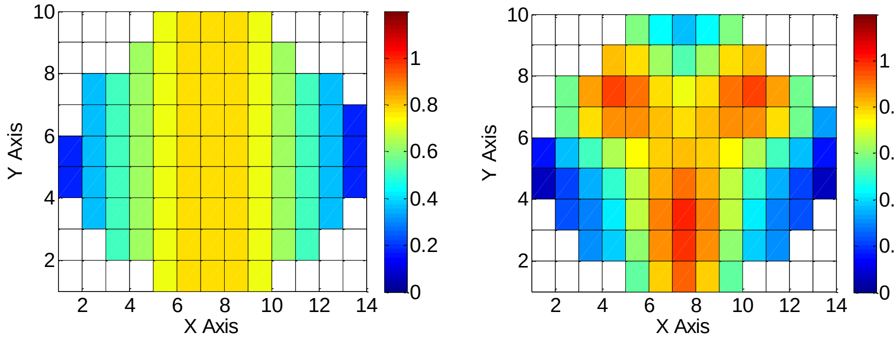
Figure 2-38. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries.



(a) (b)

Figure 2-39. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding large random variation.

We further add a larger amount of random variation distributed as $N(0, 0.04)$ to the same systematic variation shown in Figure 2-39 (b). After adding the random variation, the systematic variation contributes to 60.3% of the total variance. Figure 2-40 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. Sparse regression identifies 2 basis functions from the dictionary, and the estimated spatially correlated variation is 28.6%. Figure 2-40 (b) shows the spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries. It selects 5 DCT basis functions in addition to the physical basis functions, and the estimated spatially correlated variation is 57.0%. Similar to the previous example, the systematic pattern cannot be modeled using the physical dictionary, but an accurate estimate can be found after applying the DCT dictionary.

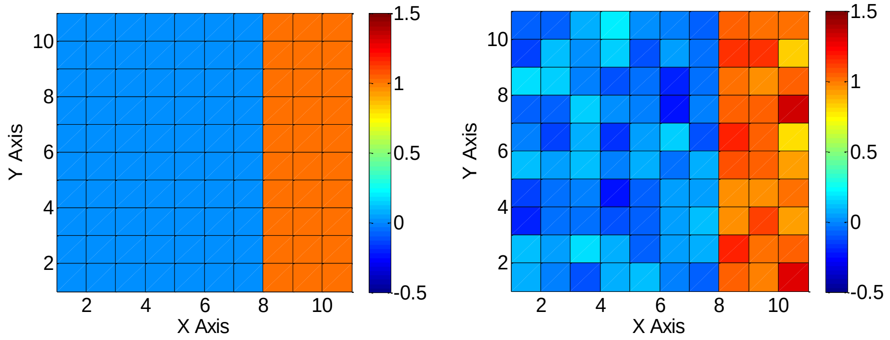


(a)

(b)

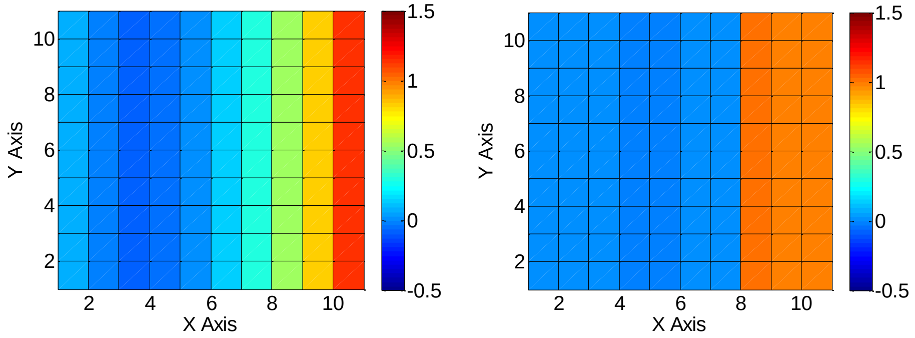
Figure 2-40. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries.

Another example that cannot be easily modeled by the physical dictionary is mask error for within-die variation. Once possible outcome of mask error is that there can exist significant mean shift in two portions of a die. One example of such effect is shown in the contact resistance data in [21], which we will further show in detail in Section 3.4. Based on the data in [21], we construct a systematic within-die variation map in Figure 2-41 (a). The systematic variation map has a significant mean shift at x = 8. We create 10 dies of synthetic data by adding a small amount of random variation distributed as $N(0, 0.01)$ on each die. The synthetic data for one of the dies is shown in Figure 2-41 (b). After adding the random variation, the systematic variation contributes to 95.5% of the total variance. Figure 2-42 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary, which contains 6 basis functions from the quadratic model. Sparse regression identifies 3 basis functions from the dictionary, and the estimated spatially correlated variation is 75.9%. Comparing Figure 2-42 (a) with Figure 2-41 (a), it can be seen that the physical dictionary does not clearly capture the left to right difference. Figure 2-42 (b) shows the spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries. It selects 6 DCT basis functions in addition to the physical basis functions, and the estimated spatially correlated variation is 95.4%. The extracted pattern closely matches the actual systematic pattern, which serves as a good basis for further diagnosis of the source of variation.



(a) (b)

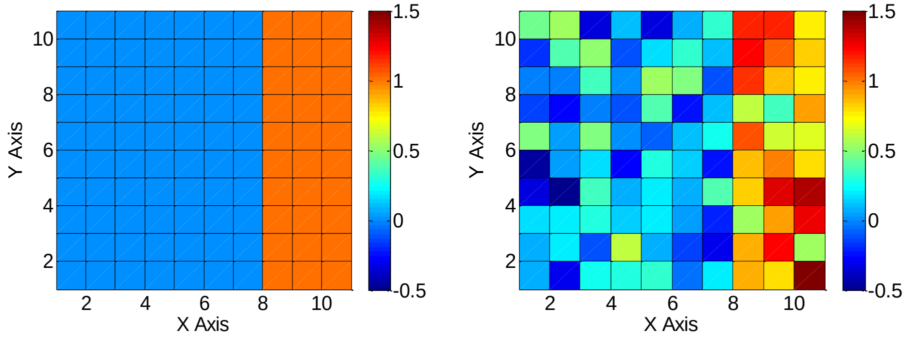
Figure 2-41. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding small random variation.



(a) (b)

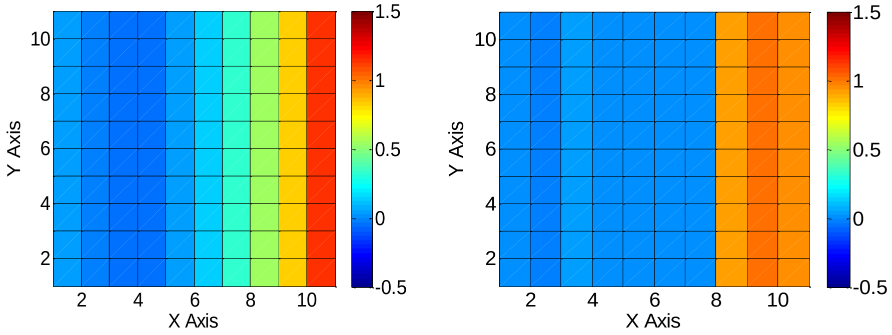
Figure 2-42. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries.

We further add a larger amount of random variation distributed as $N(0, 0.09)$ to the same systematic variation shown in Figure 2-43 (b). After adding the random variation, the systematic variation contributes to 71.2% of the total variance. Figure 2-44 (a) shows the spatially correlated variation extracted by the proposed method with the physical dictionary. Sparse regression identifies 3 basis functions from the dictionary, and the estimated spatially correlated variation is 56.8%. Figure 2-40 (b) shows the spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries. It selects 5 DCT basis functions in addition to the physical basis functions, and the estimated spatially correlated variation is 69.6%. Similar to the previous example, the systematic pattern cannot be modeled using the physical dictionary, but an accurate estimate can be found after applying the DCT dictionary.



(a) (b)

Figure 2-43. (a) Systematic variation of the synthetic wafer. (b) Synthetic data created by adding random variation.



(a) (b)

Figure 2-44. (a) Spatially correlated variation extracted by the proposed method with the physical dictionary. (b) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries.

Table 2-3.Summary of results with complex effects.

Example	Golden	Proposed physical	Proposed physical+DCT
Multiple heat, small variation	86.1%	37.8%	84.9%
Multiple heat, small variation	60.3%	28.6%	57.0%
Mask error, small variation	95.5%	75.9%	95.4%
Mask error, large variation	71.2%	56.8%	69.6%

Table 2-3 summarizes the results from examples containing complex spatial patterns. From these examples, we observe that when the systematic variation pattern differs from those considered in the physical dictionary, applying physical dictionary may not adequately fit the systematic pattern and may create misleading results. In this case, if multiple wafers/dies are present, further applying the DCT dictionary can help discover significant patterns that have been missed by the physical dictionary.

2.6 Summary

Variation decomposition is an important tool to help identify important process steps that cause significant overall performance variations. For wafers with similar spatial pattern, an important goal is to identify the spatially correlated component for wafer-level and within-die variation, which serves as an important basis for further diagnosis of systematic variation sources. In this chapter, we have proposed a wafer-level physical basis function dictionary which models more physical effects than the traditional quadratic modeling approach, and further proposes to use the DCT dictionary to discover systematic patterns not modeled by the physical dictionary. Moreover, we propose a variation decomposition method which uses sparse regression to select basis functions from the physical and DCT dictionaries. By applying sparse regression, the over-fitting problem related to applying a large basis function dictionary can be significantly reduced. A large number of examples are constructed to demonstrate the efficacy of the proposed algorithm and models.

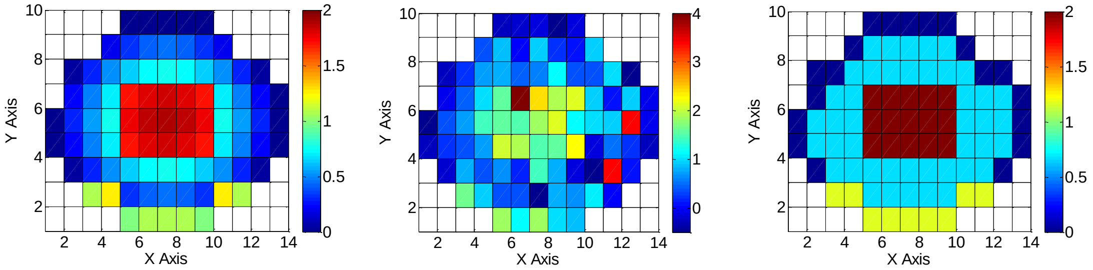
Chapter 3

Robust Regression for Variation Decomposition

3.1 Motivation

In the previous chapter, we have presented a sparse regression based method for performing variation decomposition. The proposed method mainly contains two steps: first, the basis functions are selected from a dictionary by sparse regression; second, a linear mixed model is solved by REML with the selected basis functions to decompose variation. Both of these steps are sensitive to outliers, which typically exist in real measurement data. Outliers are measurement data that significantly deviate from the regular parametric variation range. In practice, outliers can occur because of manufacturing defects or measurement errors. For example, wafer probe test may produce incorrect measurement results due to probe misalignment [20].

If outliers are not appropriately considered, they will introduce substantial error when performing variation decomposition. Two types of errors can be introduced by outliers in the variation decomposition process. First, sparse regression may not select the correct basis functions to model the spatially correlated variation. Specifically, the selected basis functions are typically fewer than the actual basis functions. This is because outliers often strongly deviate from the general spatial pattern, so that the sparse regression algorithm may determine that the spatial pattern does not exist at all. Second, a few outliers may result in an extremely strong random variation component, and will result in underestimation of spatially correlated component. To intuitively demonstrate the errors caused by outliers, Figure 3-1 (a) shows the systematic variation of a synthetic wafer with quadratic, edge and center effects, which is the same as Figure 2-35 (a). Figure 3-1 (b) shows the measurement data, where 3 outliers are randomly added to the measurement data in Figure 2-35 (b). Figure 3-1 (c) shows the extracted spatially correlated variation when the basis functions are selected by Algorithm 2 in the previous chapter with the physical dictionary. Comparing Figure 3-1 (c) with the actual systematic variation in Figure 3-1 (a), the most important difference is that the radial trend in Figure 3-1 (a) is completely lost in Figure 3-1 (c). This is because when performing sparse regression, no quadratic term is selected. The lack of quadratic terms will lead to incorrect interpretation of the spatial pattern. Because of this lack of fit and presence of outliers, the estimated spatially correlated variation is 48.3% of the overall variation, which is significantly less than the actual amount of 65.9%.



(a) (b) (c)

Figure 3-1. (a) Systematic variation of the synthetic wafer. (b) Measurement data created by adding large random variation and three outliers. (c) Estimated spatially correlated variation without outlier detection.

In this chapter, we propose to extend the sparse regression algorithm introduced in Algorithm 2 to a *robust* sparse regression algorithm. By solving robust sparse regression, basis functions will be accurately selected in the presence of outliers, and outliers will be automatically detected and removed, before the data is provided to the linear mixed model to perform variation decomposition. The remainder of the chapter is organized as follows. In Section 3.2 we review the background on the traditional IQR method for outlier detection and its limitation. We present our robust sparse regression algorithm in Section 3.3. In Section 3.4, we will first demonstrate the efficacy of the robust sparse regression algorithm with several examples, and then show the variation decomposition results on several silicon data sets. Finally, we summarize our findings in Section 3.5.

3.2 Background

Traditionally, outlier detection is performed as a pre-processing step to all the subsequent algorithms. Therefore, before applying the sparse regression algorithm described in Algorithm 2, we can estimate the variation range of all the measurement data, and then remove the outliers that sit outside the estimated range.

The classical outlier detection method in statistics is the *Interquartile Range* (IQR) method [26]. *Quartiles* are defined as the three values in ascending order, [Q_1 Q_2 Q_3], which divide the sorted data set into four equal parts. In other words, Q_1 Q_2 and Q_3 correspond to the 25%, 50% and 75% points of the cumulative distribution function (CDF) of the data. Suppose that B represents all the measurement data sorted in ascending order and the total amount of measurement data is N, these three values are computed by the following equations:

$$Q_1 = B_{N/4+0.5} \tag{3.1}$$

$$Q_2 = B_{N/2+0.5} \tag{3.2}$$

$$Q_3 = B_{3N/4+0.5} \,. \tag{3.3}$$

If the index is a non-integer value, the quartile is computed by interpolation from two adjacent data points. For example, if N = 9, the index for Q_1 is 2.75. Therefore, Q_1 is computed by

$$Q_1(N = 9) = \left(B_2 + 3 \cdot B_3\right)/\,4\,. \tag{3.4}$$

Next, we compute the IQR of the data:

$$IQR = Q_3 - Q_1 \tag{3.5}$$

to estimate the variability. If the measurement data satisfy a Normal distribution, IQR is nearly equal to $4/3 \cdot \sigma$ where σ denotes the standard deviation of the distribution.

Finally, we define the variation range of the data based on IQR computed from (3.5). For each measurement data, we consider it as an outlier, if its value is outside the following variation range:

$$R = [Q_1 - 3 \cdot IQR, Q_3 + 3 \cdot IQR] \tag{3.6}$$

where the scaling factor 3 is decided empirically by the statistics community. If the measurement data is normally distributed, the IQR method removes the data outside ±4.7σ range.

The key idea of the IQR method is to use robust metrics in (3.1)-(3.3) to define the variation range. The metrics themselves must not be easily biased by the outliers within the data. To intuitively illustrate this idea of the IQR method, Figure 3-2 shows an outlier detection example containing 9 regular data points and 1 outlier. A naïve method to detect outliers can be constructed by defining the following range:

$$R = [\mu - 3\cdot\sigma, \mu + 3\cdot\sigma] \tag{3.7}$$

where μ and σ are the sample mean and sample standard deviation of all measurement data, respectively. Both μ and σ can be strongly biased by a single outlier, so that outlier remains inside the interval specified by $\mu \pm 3\sigma$. Therefore, outlier detection cannot be performed by using simple statistics (e.g., mean and standard deviation), since they are extremely sensitive to the large measurement error posed by outliers. On the other hand, the boundary defined by is Q_3+3·IQR intuitively a good point to separate regular data from outliers.

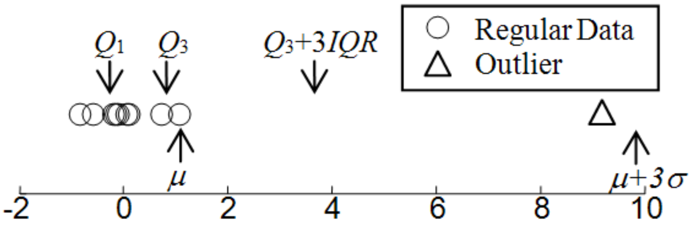
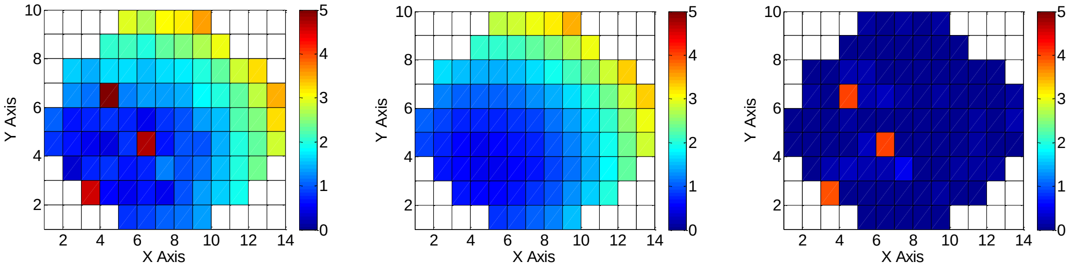


Figure 3-2. The IQR method successfully detects the outlier that strongly biases the estimation of mean and standard deviation.

The aforementioned outlier detection method suffers from strong limitations when processing IC measurements. This is because modern IC processes suffer from a lot of variation sources which present themselves in lot-to-lot, wafer-to-wafer, wafer-level and within-die variations. When directly viewing all measurement data, the accumulation of these variation sources will result in a very large variation range. Therefore, even if the outcome of a particular process step is strongly distorted by defects, such distortion may not be significant compared to the natural variation range of all process steps, making outliers difficult to detect. This has been observed in a number of performance metrics in the testing literature, such as I_{ddq} [49], minimum V_{dd} [50] and maximum frequency [51]. Therefore, an accurate outlier detection method for IC measurement must take into account the fact that the overall variation can be decomposed into multiple components, and explicitly use this information to make the outliers more distinguishable from the regular data, before determining the variation range of regular data. In the next sub-section, we will present a robust sparse regression method that is built based upon this concept.

3.3 Robust Sparse Regression

In this section, we develop a new algorithm to accurately select the basis functions and remove outliers for IC measurements. The key motivation of our algorithm is that when removing the wafer-level/within-die outliers, we would like to define the variation range only based on wafer-level/within-die random variation. By only considering one component of the overall variation, we will be able to significantly reduce the variation range for regular measurements, and therefore making regular data and outliers more easily separable.



(a) (b) (c)

Figure 3-3. (a) Measurement data with 3 outliers. (b) Estimated spatially correlated variation. (c) Outliers are more easily distinguished after removing spatially correlated variation.

We intuitively illustrate the concept of our algorithm by a wafer-level example shown in Figure 3-3. Figure 3-3 (a) is created by adding three outliers to the wafer map in Figure 2-17 (b), which contains only a small amount of random variation. It can be seen that the value of the outlier dies are not very far from regular data. Suppose that we are able to accurately extract the spatially correlated variation of the wafer in Figure 3-3 (b), and subtract this component from Figure 3-3 (a), the resulting residual is shown in Figure 3-3 (c). Figure 3-3 (c) corresponds to the wafer-level random variation component of the wafer in Figure 3-3 (a). It can be obviously seen from Figure 3-3 (c) that the outliers are much more clearly separable from the regular measurement data than Figure 3-3 (a).

While the aforementioned flow improves the accuracy of outlier detection by estimating and removing the spatially correlated variation component, the most important challenge is how to accurately estimate such spatially correlated variation in the presence of outliers. Since Algorithm 2 in the previous chapter is sensitive to outliers, it will only be accurate if the outliers have been removed. This forms a circular dependency with the need to use the sparse regression result to detect outliers. In order to escape from this dependency, we would like to revise the sparse regression algorithm to make it robust to outliers.

We develop our algorithm by first studying the fundamental reason why Algorithm 2 is sensitive to outliers. We re-write the matrix $A_{(l)}$ in the sparse regression problem (2.65) into a row matrix form:

$$A_{(l)} = \begin{bmatrix} A_{1,(l)} \\ A_{2,(l)} \\ \vdots \\ A_{N(l),(l)} \end{bmatrix} \tag{3.8}$$

where each row corresponds to the value of all basis functions for a particular performance measurement. For each measurement, by solving (2.65), we can obtain a residual:

$$e_{(l),i} = b_{(l),i} - A_{i,(l)}\eta_{(l)} \left(i = 1,2,...,N_{(l)}\right) \tag{3.9}$$

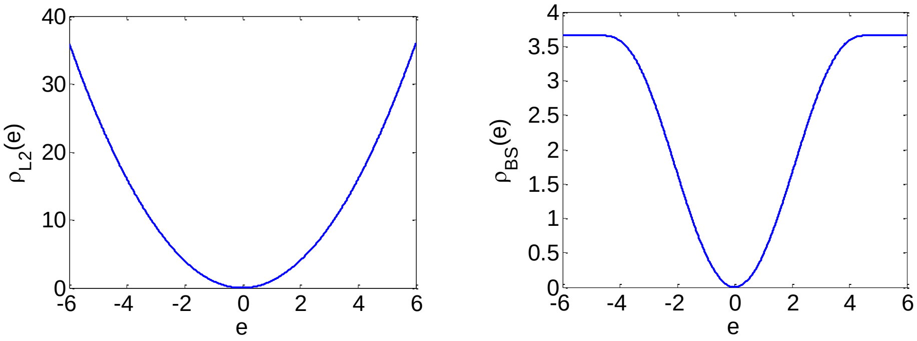
where $b_{(l),i}$ is the i-th element of the vector $B_{(l)}$. Therefore, we can re-write the sparse regression problem (2.65) as

$$\begin{array}{ccc} \underset{\eta_{(l)}}{\text{minimize}} & \sum_{i=1}^{N_{(l)}} \rho_{L2}\left(e_{(l),i}\right) & (l = 1,2,...,L) \\ s.t. & nnz\left(\left\{\eta_{(l),j} \mid j \notin \Omega_0\right\}\right) \le \lambda & \end{array} \tag{3.10}$$

where

$$\rho_{L2}(e) = e^2 . \tag{3.11}$$

In robust statistics, when the cost function of an estimate can be represented as the summation of ρ functions, where ρ is a function of a single sampling point and the estimate, such an estimate is called an M-estimate [52]. For regression problems, ρ is typically a function of the residual, such as the ρ_{L2} function in (3.11). It has been well studied in the robust statistics literature that if the ρ function is not robust, the regression problem will be sensitive to outliers. The ρ function in (3.11) is not a robust function. To understand this concept, we plot the function (3.11) in Figure 3-4 (a). It can be seen that as the residual moves away from zero, the objective function increases rapidly. Therefore, if there exist any strong outliers, even if they are few in number, they can significantly influence the cost function in (3.10) and therefore completely bias the result.



(a) (b)

Figure 3-4. (a) the L$_2$-norm ρ function. (b) the bisquare ρ function.

Based on the M-estimate theory in statistics, we develop a robust regression method by replacing the non-robust ρ function in (3.11) by a robust function. A large number of functions that are robust to outliers have been proposed in the statistical literature [26] [52]. It has been shown that the choice of the function is not critical in many practical situations, and many choices will give similar results that offer great improvements over classical estimates [52]. In this work, we adopt a robust ρ function named bisquare function shown in Figure 3-4 (b). The bisquare function is similar in shape to the L$_2$-norm ρ function, but its value stops to grow after a certain threshold. Therefore, intuitively it would prevent a small number of outliers from significantly changing the result. Mathematically, the bisquare ρ function is defined as:

$$\rho_{BS}(e) = \begin{cases} \dfrac{k^2}{6}\left(1-\left(1-\left(\dfrac{e}{k}\right)^2\right)^3\right) & \left(|e| \le k\right) \\ \dfrac{k^2}{6} & \left(|e| > k\right) \end{cases} \tag{3.12}$$

where k is a tuning constant specifying the cut-off threshold in Figure 3-4 (b). The following tuning constant is often used [53]:

$$k_{(l)} = 4.685 \cdot std_i\left(e_{(l),i}\right) \tag{3.13}$$

where $std_i(e_{(l),i})$ means standard deviation of the residuals $\{e_{(l),i};\ i = 1, 2, \ldots, N_{(l)}\}$. This cut-off threshold is similar to the threshold in (3.6) based on 3·IQR. Traditionally, the following equation is used to estimate the standard deviation:

$$std_i\left(e_{(l),i}\right)=\sqrt{\frac{1}{N_{(l)}-1}\sum_{i=1}^{N_{(l)}}\left(e_{(l),i}-mean_j\left(e_{(l),j}\right)\right)^2} \qquad (3.14)$$

where $mean_j(e_{(l),j})$ means standard deviation of the residuals $\{e_{(l),j};\ j = 1, 2, \ldots, N_{(l)}\}$. However, similar with the cost function in (3.10), Eq. (3.14) can be significantly influenced by large $e_{(l),i}$ values. Therefore, to robustly estimate the standard deviation, we adopt the following equation [26]:

$$std_i\left(e_{(l),i}\right)=median_i\left(\left|e_{(l),i}-median_j\left(e_{(l),j}\right)\right|\right)/\,0.6745\ . \qquad (3.15)$$

Eq. (3.15) is a consistent estimator of the standard deviation when $\{e_{(l),j};\ j = 1, 2, \ldots, N_{(l)}\}$ is normally distributed. Compared with the traditional sample standard deviation (3.14), Eq. (3.15) uses metrics such as median and absolute value to replace the mean and square value to ensure robustness.

Based on the above equations, we formulate the robust sparse regression problem as:

$$\begin{array}{lll}\underset{\eta_{(l)}}{\text{minimize}} & \sum_{i=1}^{N_{(l)}}\rho_{BS}\left(e_{(l),i}\right) & \\ s.t. & nnz\left(\left\{\eta_{(l),j}\ |\ j\notin\Omega_0\right\}\right)\le\lambda & \end{array} \quad (l=1,2,\ldots,L) \qquad (3.16)$$

where the definition of ρ_{BS} is in (3.12), (3.13) and (3.15). Compared with the original sparse regression problem (3.10), Eq. (3.16) replaces the quadratic cost function by a robust cost function. Since the new cost function is a non-convex function, additional approximations are required. A practical method to solve this optimization problem with non-convex cost function is the iteratively reweighted least squares (IRLS) method [53]. To derive the IRLS method, we differentiate the cost function with respect to the coefficients and set the partial derivative to 0:

$$\sum_{i=1}^{N_{(l)}}\psi_{BS}\left(e_{(l),i}\right)\cdot A_{i,(l)}=0 \qquad (3.17)$$

where ψ_{BS} is the derivative of ρ_{BS}. We further define the weight function as

$$w_{BS}\left(e_{(l),i}\right)=\psi_{BS}\left(e_{(l),i}\right)/\,e_{(l),i} \qquad (3.18)$$

and re-write (3.16) as:

$$\sum_{i=1}^{N_{(l)}}w_{BS}\left(e_{(l),i}\right)\cdot e_{(l),i}\cdot A_{i,(l)}=0\ . \qquad (3.19)$$

The IRLS method is an iterative method where in each iteration, $w_{BS}(e_{(l),\, i})$ is estimated from the residual in the previous iteration. The initial estimate can be obtained by solving the traditional sparse regression problem (3.10). Therefore, when solving the optimization, $w_{BS}(e_{(l),\, i})$ is treated as a fixed value $w_{BS(l),\, i}$:

$$\sum_{i=1}^{N_{(l)}} w_{BS(l),i} \cdot e_{(l),i} \cdot A_{i,(l)} = 0 \,. \tag{3.20}$$

It can be easily proved that the solution from (3.20) is equivalent to setting the derivative of the following cost function to 0:

$$c(\eta_{(l)}) = \sum_{i=1}^{N_{(l)}} w_{BS(l),i} \cdot \rho_{L2}\left(e_{(l),i}\right). \tag{3.21}$$

By changing the cost function to (3.21), the following optimization can be formulated:

$$\begin{array}{ll} \underset{\eta_{(l)}}{\text{minimize}} & \sum_{i=1}^{N_{(l)}} w_{BS(l),i} \cdot \rho_{L2}\left(e_{(l),i}\right) \\ s.t. & nnz\left(\left\{\eta_{(l),j} \mid j \notin \Omega_0\right\}\right) \le \lambda \end{array} \quad (l = 1,2,\ldots,L)\,. \tag{3.22}$$

Compared to (3.10), Eq. (3.22) reweights each sampling point according to the following weight function:

$$w_{BS}\left(e_{(l),i}\right) = \begin{cases} \left(1 - \left(\dfrac{e_{(l),i}}{k_{(l)}}\right)^2\right)^2 & \left(\left|e_{(l),i}\right| \le k_{(l)}\right) \\ 0 & \left(\left|e_{(l),i}\right| > k_{(l)}\right) \end{cases} \tag{3.23}$$

where $k_{(l)}$ is defined in (3.13) and (3.15). To intuitively understand such reweighting, we plot the weight function $w_{BS}(e)$ in Figure 3-5, where the unit of the x-axis is the standard deviation of residual. It can be seen from Figure 3-5 that the weight of a sampling point will decrease as its residual moves away from 0. When the residual exceeds the cut-off threshold in (3.13), the corresponding weight will reduce to 0, meaning that the corresponding sampling point will not be used in the next iteration.

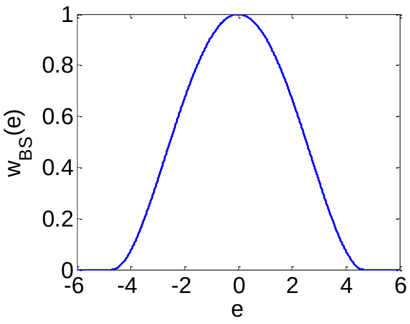


Figure 3-5. The bi-square weight function.

Eq. (3.22) can be re-written into the following form:

$$\begin{array}{ll} \underset{\eta_{(l)}}{\text{minimize}} & \left\| W_{BS(l)}^{0.5} \left(A_{(l)} \eta_{(l)} - B_{(l)} \right) \right\|_2^2 \\ s.t. & nnz\left(\left\{ \eta_{(l),j} \mid j \notin \Omega_0 \right\} \right) \le \lambda \end{array} \quad (l = 1,2,..., L) \tag{3.24}$$

where $W_{BS(l)}$ is an diagonal matrix with $W_{BS(l),ii} = w_{BS(l),i}$. Eq. (3.24) still satisfies the general form of sparse regression in (2.65), and therefore can be solved by Algorithm 2. We summarize the main steps of the robust sparse regression in Algorithm 3:

Algorithm 3: Robust Sparse Regression Based on Spatial Correlation

1. Apply Algorithm 2 to solve the coefficients $\{\eta_{(l)}; l = 1,2,\ldots,L\}$from (3.10).
2. Calculate the weight for each measurement by (3.23).
3. Apply Algorithm 2 to solve the coefficients $\{\eta_{(l)}; l = 1,2,\ldots,L\}$from (3.24).
4. If the change of coefficients $\{\eta_{(l)}; l = 1,2,\ldots,L\}$ is sufficiently small compared to the previous iteration, stop. Otherwise go to step 2.

Algorithm 3 selects the basis functions in the presence of outliers by repeatedly re-weighting the sampling points and solving the sparse regression problem with Algorithm 2. Its computational cost is about the time of performing Algorithm 2 multiplied by the number of iterations. We observe that in most cases Algorithm 3 will converge within 10 iterations. Therefore, it consumes about 2-10× time compared to Algorithm 2. While performing Algorithm 3, measurement data for which the residual exceeds the cut-off threshold in (3.13) will be automatically removed. These data are identified as outliers and removed before solving the linear mixed model.

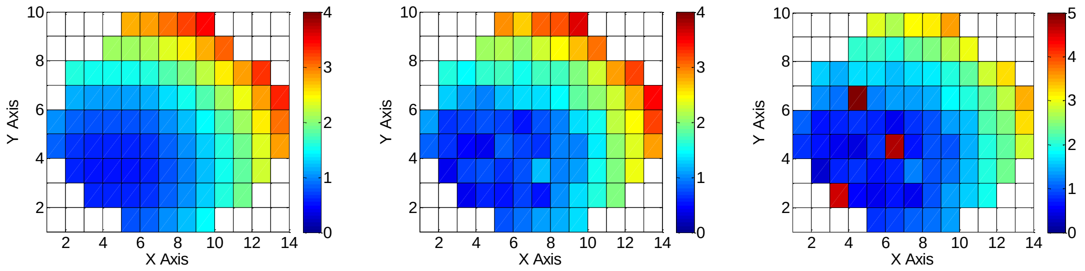
3.4 Numerical Results

In this section, we first use several synthetic and real examples to demonstrate that our proposed robust sparse regression method provides superior accuracy in determining the correct basis functions and remove outliers compared to the traditional outlier detection method. We will then show the variation decomposition results on several silicon data sets.

3.4.1 Comparison of Outlier Detection Methods

We first consider a synthetic wafer constructed based on the measurement data in Figure 2-17.

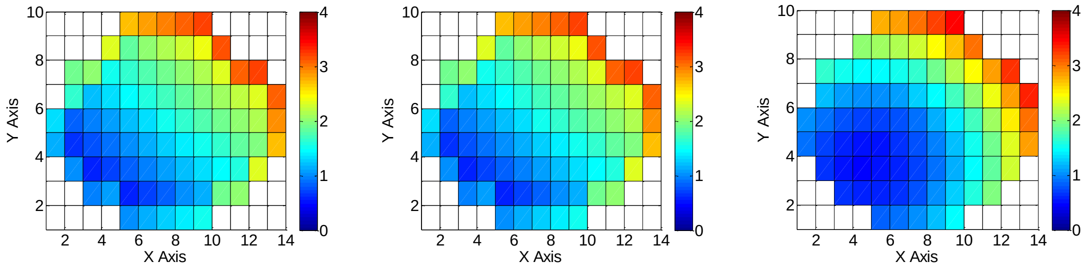
Figure 3-6 (a) shows the systematic variation of the synthetic wafer. Figure 3-6 (b) shows the measurement data after adding small random variation, where the systematic variation contributes to 97.2% of the total variance. Based on Figure 3-6 (b), we further randomly add 3 outliers at 3 random locations in Figure 3-6 (c). For each location, the outlier is created by adding 3·IQR of the wafer to its original value, where IQR is defined in (3.5).



(a) (b) (c)

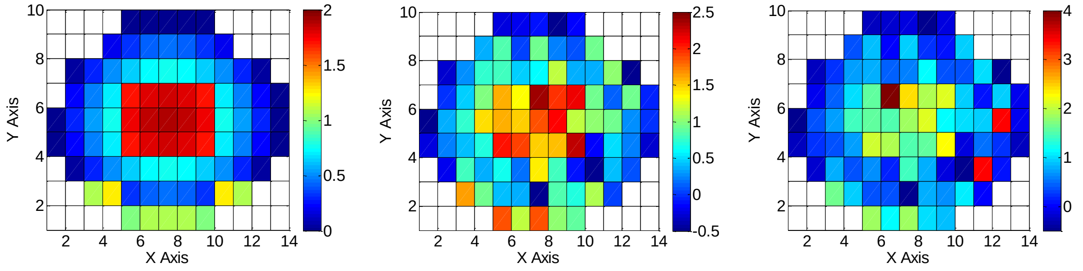
Figure 3-6. (a) Systematic variation of the synthetic wafer. (b) Measurement data created by adding random variation. (c) Measurement data after adding 3 outliers.

We compare the variation decomposition results from three methods. All the three methods use sparse regression to determine the basis functions from the physical dictionary, and then solve the linear mixed model with these basis functions by applying REML. The first method does not apply any outlier detection method, and the extracted spatially correlated variation is shown in Figure 3-7 (a). The second method applies the traditional IQR outlier detection method in Section 3.2 before applying sparse regression, and the extracted spatially correlated variation is shown in Figure 3-7 (b). The third method is our proposed method in this chapter with physical basis functions, which applies the robust sparse regression method in Section 3.3 to determine the basis functions and remove outliers, and the extracted spatially correlated variation is shown in Figure 3-7 (c). The estimated spatially correlated variation for the first method is 45.5%, which significantly underestimates the spatially correlated variation. The estimated percentages of quadratic, edge and center effects are 81.8%, 18.2%, and 0.0% respectively. Examining Figure 3-7 (a), it can be seen that it fails to select the quadratic basis functions that are present in the systematic variation in Figure 3-7 (a). It also determines that there is strong edge effect at the lower-left corner, which does not exist in the systematic variation. In this example, none of the 3 outliers are detected by the IQR method. This is because these locations have relatively small values in the systematic variation map Figure 3-6 (a), which cancels out a significant portion of the outlier effect. Since no outlier is detected, Figure 3-7 (b) is exactly the same as Figure 3-7 (a). The robust sparse regression method correctly detects the 3 outliers, the estimated spatially correlated variation is 97.2%, and the estimated percentages of quadratic, edge and center effects are 100%, 0.0%, and 0.0% respectively. These results exactly match the systematic variation and demonstrate that the robust sparse regression method achieves significantly better accuracy.



(a) (b) (c)

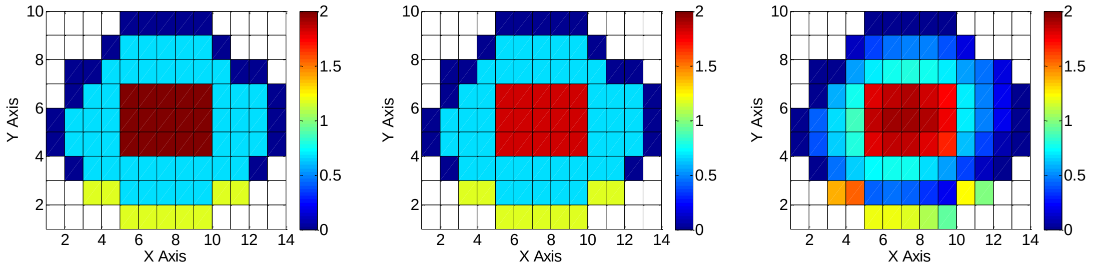
Figure 3-7. (a) Spatially correlated variation extracted by applying sparse regression with the physical dictionary, without outlier detection. (b) Spatially correlated variation extracted by applying sparse regression with the physical dictionary, with traditional IQR outlier detection. (c) Spatially correlated variation extracted by the proposed method with the physical dictionary.



(a) (b) (c)

Figure 3-8. (a) Systematic variation of the synthetic wafer. (b) Measurement data created by adding random variation. (c) Measurement data after adding 3 outliers.

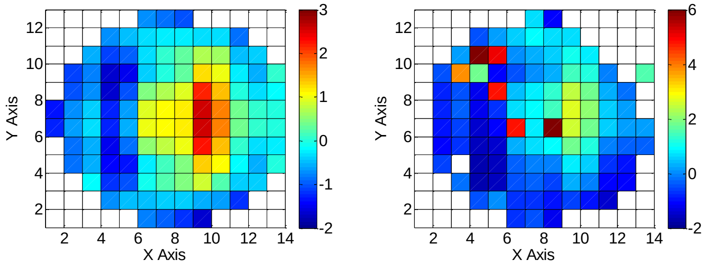
We further consider a synthetic wafer constructed based on the measurement data in Figure 2-35. Figure 3-8 (a) shows the systematic variation of the synthetic wafer. Figure 3-8 (b) shows the measurement data after adding random variation, where the systematic variation contributes to 65.9% of the total variance. Based on Figure 3-8 (b), we further randomly add 3 outliers at 3 random locations in Figure 3-8 (c). For each location, the outlier is created by adding 3IQR of the wafer to its original value, where IQR is defined in (3.5). We again compare the variation decomposition results from three methods in Figure 3-9. Figure 3-9 (a) shows the extracted spatially correlated variation by the first method. The estimated spatially correlated variation method is 48.3%, which significantly underestimates the spatially correlated variation. The estimated percentages of quadratic, edge and center effects are 0.0%, 32.0%, and 68.0% respectively. Examining Figure 3-9 (a), it can be seen that it does not contain the radial pattern produced by quadratic basis functions in Figure 3-8 (a). The traditional IQR method detects only one outlier in this example located in the center of the wafer, and the extracted spatially correlated variation after removing this outlier is shown in Figure 3-9 (b). The estimated spatially correlated variation method is 49.0%, which still significantly underestimates the spatially correlated variation. The estimated percentages of quadratic, edge and center effects are 0.0%, 39.4%, and 60.6% respectively. Examining Figure 3-9 (b), it can be seen that it still does not select the quadratic basis functions and therefore does not fit the radial pattern. This is the main reason why no significant improvement can be seen compared to the first method. The robust sparse regression method correctly detects the 3 outliers, the estimated spatially correlated variation is 71.5%, and the estimated percentages of quadratic, edge and center effects are 32.7%, 39.1%, and 28.1% respectively. These results closely match the systematic variation. Therefore, it again demonstrates that significant accuracy improvement can be achieved by adopting the robust sparse regression method.



(a) (b) (c)

Figure 3-9. (a) Spatially correlated variation extracted by applying sparse regression with the physical dictionary, without outlier detection. (b) Spatially correlated variation extracted by applying sparse regression with the physical dictionary, with traditional IQR outlier detection. (c) Spatially correlated variation extracted by the proposed method with the physical dictionary.

We further consider an example from silicon measurement data in Figure 3-10. The measurement data are collected from 201 wafers from 14 lots at an advanced technology node. Each wafer contains 117 ring oscillators (ROs) distributed over different spatial locations. As will be discussed in more detail in Section 3.4.2, the spatially correlated variation is not adequately fit by physical basis functions only and both physical and DCT basis functions need to be applied. The extracted spatially correlated variation is shown in Figure 3-10 (a), and manual inspection reveals that all wafers closely match the spatial pattern in Figure 3-10 (a). Next, we consider the results by performing variation decomposition on one of the wafers in Figure 3-10 (b), which contains a significant number of outliers.



(a)

(b)

Figure 3-10. (a) Spatially correlated variation extracted from 201 wafers of ring oscillator period measurements (normalized). (b) Measurement data from one wafer with outliers.

We compare the variation decomposition results from three methods in Figure 3-11. Figure 3-11 (a) shows the extracted spatially correlated variation by the first method. The estimated spatially correlated variation method is 29.7%. The traditional IQR outlier detection method detects only 2 outliers, and the extracted spatially correlated variation after removing these outliers is shown in Figure 3-11 (b). The estimated spatially correlated variation method is 33.2%. It can be clearly seen that these two methods greatly underestimates the spatially correlated variation, and the extracted pattern is significantly different from the spatial pattern in Figure 3-10 (a). Robust sparse regression detects 7 outliers in this example, and the extracted spatially correlated variation by the proposed method in shown in Figure 3-11 (c). The 7 outliers intuitively match the 7 dies that are significantly different from overall spatial pattern. The estimated spatially correlated variation is 93.9%. Examining Figure 3-11 (c), it can be seen that it more accurately matches Figure 3-10 (a) comparing to the other two methods.

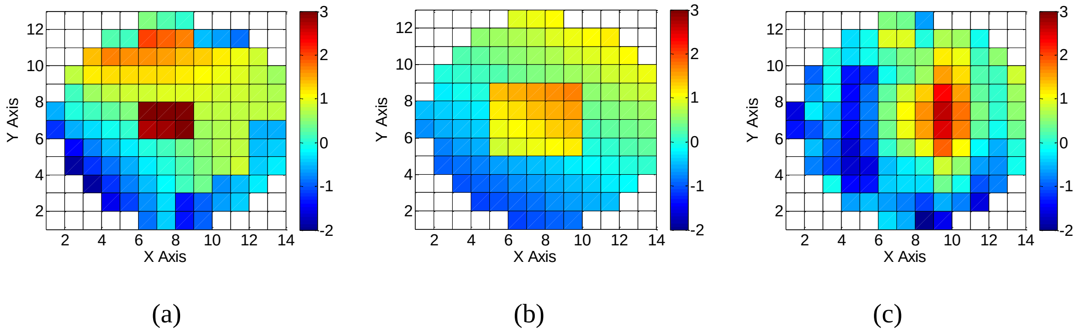


Figure 3-11. (a) Spatially correlated variation extracted by applying sparse regression with the physical and DCT dictionaries, without outlier detection. (b) Spatially correlated variation extracted by applying sparse regression with the physical and DCT dictionaries, with traditional IQR outlier detection. (c) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries.

From the above examples, we find that outliers in measurement data can cause incorrect choice of basis functions in sparse regression, and significant underestimation of spatially correlated variation in variation decomposition. The proposed robust sparse regression method achieves superior accuracy in basis function selection and outlier detection compared to the traditional method in Section 3.2 in several synthetic and silicon examples.

3.4.2 Results on Silicon Measurement Data

From the previous experiments, we observe that by applying robust sparse regression, we are able to accurately find the basis functions and detect the outliers in the measurement data. We will then perform variation decomposition on several sets of silicon measurement data and present the results.

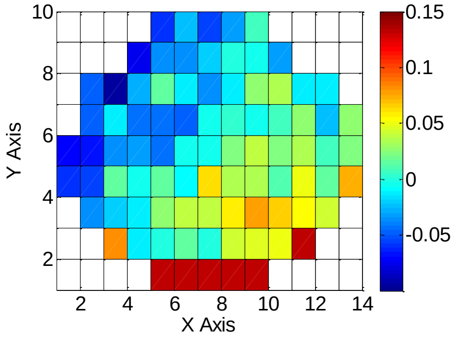
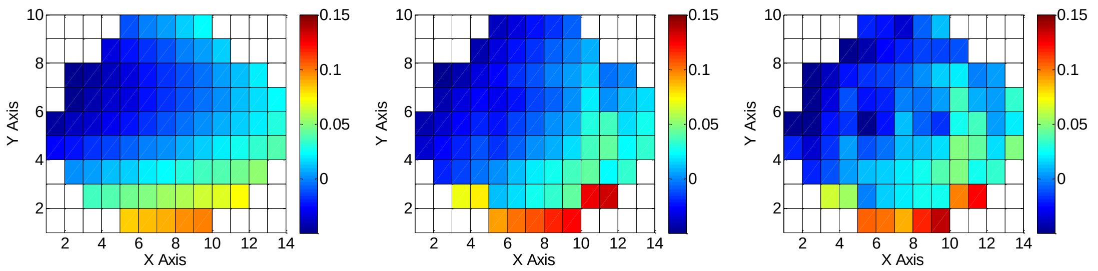


Figure 3-12. I_{dsat} measurement data (normalized) from one of the 15 wafers.

Table 3-1. Variation components of the first I_{dsat} measurement data set.

Method	Wafer-to-wafer	Wafer-level spatially correlated	Wafer-level random
Quadratic	6.7%	55.4%	37.9%
Proposed physical	5.8%	64.9%	29.3%
Proposed physical+DCT	6.3%	67.4%	26.2%

We first consider transistor drain saturation current (I_{dsat}) measurements taken from 15 wafers from a commercial CMOS process. Figure 3-12 shows one of the wafers. Intuitively, the measurement data contains significant random variation. Table 3-1 compares the variation components estimated by three methods. The first method directly performs REML with quadratic basis functions; the second method is our proposed method in this chapter with the physical dictionary, and third method is our proposed method in this chapter with the physical and DCT dictionaries. Figure 3-13 compares the spatially correlated variation extracted by three methods. Figure 3-13 (a) shows the spatially correlated variation extracted by applying the quadratic basis functions. Figure 3-13 (b) shows the spatially correlated variation extracted by robust sparse regression with the physical dictionary, where for wafer-level spatially correlated variation, the estimated percentages of quadratic, edge and center effects are 55.2%, 43.6%, and 1.2% respectively. Compared to using the quadratic basis functions, applying robust sparse regression with physical dictionary explains a significantly larger amount of variation as wafer-level spatially correlated variation. Moreover, it reveals that the spatially correlated variation mainly comes from quadratic and edge effect. The edge effect is non-trivial and contributes to nearly half of the spatially correlated variation. From Figure 3-13 (b), it can be also intuitively seen that a much more obvious edge effect pattern in the bottom is modeled compared to Figure 3-13 (a). These observations tell us that reducing the bottom edge effect is an important task when improving the overall yield. This conclusion would not be easily reached if the quadratic basis functions are simply applied. Figure 3-13 (c) shows the spatially correlated variation extracted by applying robust sparse regression with the physical and dictionaries. After adding the DCT basis functions, we do not observe significant larger amount of wafer-level spatially correlated variation, and Figure 3-13 (c) does not clearly show any meaningful additional pattern compared to Figure 3-13 (b). Therefore, we believe that the physical dictionary in sufficient in modeling the spatially correlated variation in this example.



(a) (b) (c)

Figure 3-13. (a) Spatially correlated variation extracted by quadratic basis functions. (b) Spatially correlated variation extracted by the proposed method with the physical dictionary. (c) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries.

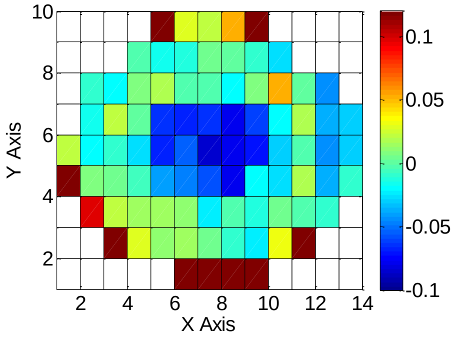
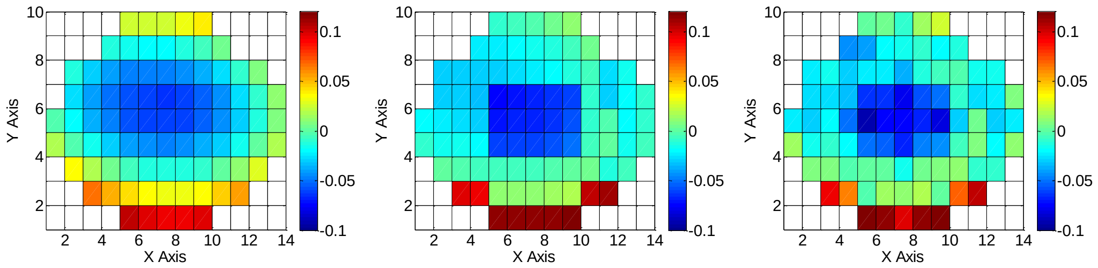


Figure 3-14. I_{dsat} measurement data (normalized) from one of the 8 wafers.



(a) (b) (c)

Figure 3-15. (a) Spatially correlated variation extracted by quadratic basis functions. (b) Spatially correlated variation extracted by the proposed method with the physical dictionary. (c) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries.

Table 3-2.Variation components of the second I_{dsat} measurement data set.

Method	Wafer-to-wafer	Wafer-level spatially correlated	Wafer-level random
Quadratic	30.2%	45.2%	24.6%
Proposed physical	28.2%	53.8%	17.9%
Proposed physical+DCT	30.3%	54.8%	14.9%

We further consider I_{dsat} measurements taken from another 8 wafers with a different spatial signature from the same commercial CMOS process. Figure 3-14 shows one of the wafers. Intuitively, the measurement data also contains significant random variation. Table 3-2 compares the variation components estimated by three methods, and Figure 3-15 compares the spatially correlated variation extracted by three methods. Figure 3-15 (a) shows the spatially correlated variation extracted by applying the quadratic basis functions. Figure 3-15 (b) shows the spatially correlated variation extracted by robust sparse regression with the physical dictionary. For wafer-level spatially correlated variation, the estimated percentages of quadratic, edge and center effects are 25.5%, 56.3%, and 18.1% respectively. Compared to using the quadratic basis functions, applying robust sparse regression with physical dictionary explains a significantly larger amount of variation as wafer-level spatially correlated variation. Moreover, it reveals that quadratic, edge and center effect all significantly contribute to the overall variation. The edge effect is the dominant source of spatially correlated variation for these wafers, and the center effect is also non-trivial. From Figure 3-15 (b), it can be also intuitively seen that much more obvious edge and center effect patterns are modeled compared to Figure 3-15 (a). These observations tell us that reducing the edge and center effects are important tasks when improving the overall yield. This conclusion would not be easily reached if the quadratic basis functions are simply applied. Figure 3-15 (c) shows the spatially correlated variation extracted by applying robust sparse regression with the physical and dictionaries. After adding the DCT basis functions, we do not observe significant larger amount of wafer-level spatially correlated variation, and Figure 3-15 (c) does not clearly show any meaningful additional pattern compared to Figure 3-15 (b). Therefore, we believe that the physical dictionary in sufficient in modeling the spatially correlated variation in this example.

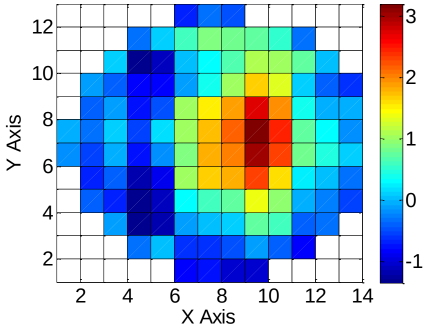
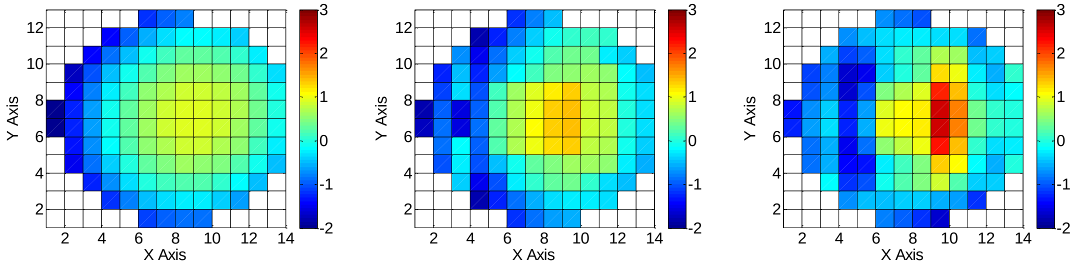


Figure 3-16. Ring oscillator (RO) period measurement data (normalized) from one of the 201 wafers.

Table 3-3. Variation components of the RO period measurement data set.

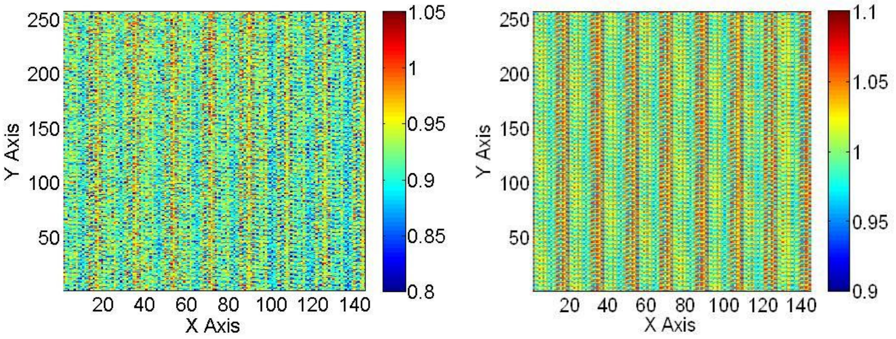
Method	Lot-to-lot	Wafer-to-wafer	Wafer-level spatially correlated	Wafer-level random
Quadratic	17.0%	25.4%	26.7%	30.9%
Proposed physical	17.1%	25.3%	33.5%	24.2%
Proposed physical+DCT	17.9%	26.3%	44.8%	11.0%

We further consider ring oscillator (RO) period measurement data collected from 201 wafers from 14 lots at an advanced technology node. Each wafer contains 117 ROs distributed over different spatial locations. Since ring oscillators use a large number of stages to average out the random variation [48], the wafer-level variation should be dominated by spatially correlated variation. Figure 3-16 shows one of the wafers, and it can be intuitively seen the measurement data already presents a strong spatial pattern. Table 3-3 compares the variation components estimated by three methods, and Figure 3-17 compares the spatially correlated variation extracted by three methods. Figure 3-17 (a) shows the spatially correlated variation extracted by applying the quadratic basis functions, where only a weak pattern of spatially correlated variation is modeled. From Table 3-3, it can be seen that the wafer-level spatially correlated variation is less than the wafer-level random variation, which is against our intuition that random variation is small. Figure 3-17 (b) shows the spatially correlated variation extracted by robust sparse regression with the physical dictionary. For wafer-level spatially correlated variation, the estimated percentages of quadratic, edge and center effects are 72.4%, 23.4%, and 4.2% respectively. Compared to using the quadratic basis functions, robust sparse regression with physical dictionary explains a significantly larger amount of variation as wafer-level spatially correlated variation. Figure 3-17 (c) shows the spatially correlated variation extracted by robust sparse regression with the physical and DCT dictionaries. After adding the DCT basis functions, we see that a significantly larger portion of variation is explained as spatially correlated variation. Manual inspection reveals that all the wafers indeed present a similar pattern to Figure 3-17 (c). Therefore, in this example, the robust sparse regression results with the physical and DCT dictionaries are the most reasonable variation decomposition results. To find the true systematic sources that cause the wafer-level spatial variation, more inspection beyond the common variation sources that cause the quadratic, center and edge patterns is needed.



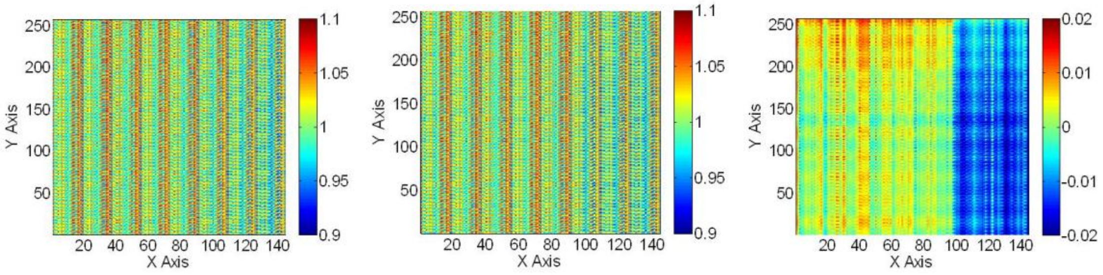
(a) (b) (c)

Figure 3-17. (a) Spatially correlated variation extracted by quadratic basis functions. (b) Spatially correlated variation extracted by the proposed method with the physical dictionary (c) Spatially correlated variation extracted by the proposed method with the physical and DCT dictionaries.



(a) (b)

Figure 3-18. (a) Contact resistance measurement data (normalized) from one of the 24 dies. (b) Spatial distribution of different contact layout patterns in the test chips.



(a) (b) (c)

Figure 3-19. (a) Spatially correlated variation extracted by the physical dictionary with layout basis functions. (b) Spatially correlated variation extracted by the physical and DCT dictionaries. (c) The spatially correlated variation represented by the quadratic and DCT basis functions.

Table 3-4.Variation components of the contact resistance measurement data set.

Method	Wafer-level	Within-die spatially correlated	Within-die random
Proposed physical	51.5%	30.9%	17.6%
Proposed physical+DCT	51.5%	31.5%	17.0%

We finally consider the contact plug resistance measurement data collected from 24 test chips in a 90 nm CMOS process. Each chip contains 36,864 test structures (i.e., contacts) arranged as a 144×256 array, as described in [21]. Among these 24 test chips, three of them contain missing data due to external measurement error. The number of failed measurements are 2936, 864 and 8 for these three chips, respectively. In the test chips, contacts with 55 different layout patterns are regularly distributed over the entire chip. The spatial distribution of different layout patterns is shown in Figure 3-18 (b). Figure 3-18 (a) shows the measured contact plug resistance (normalized) from one of the 24 test chips. Studying Figure 3-18 (a), we would notice that there is a unique spatial pattern due to layout dependency. However, the spatial pattern is not clearly visible because of the large-scale uncorrelated random variation found in this example.

We first extract the spatially correlated variation by performing robust sparse regression with the physical dictionary. In this example, since we know that the difference of layout patterns must be an important component in the spatial variation, we construct the following indicator basis functions that correspond to different layout patterns:

$$f_i(x,y)=\begin{cases}1 & (x,y)\in L_i\\ 0 & otherwise\end{cases}\quad (i=1,2,...,55) \tag{3.25}$$

where L_i is a set of measurements collected from test structures with layout style *i*. These 55 layout basis functions are added to the physical dictionary and pre-selected in the sparse regression process. The extracted spatially correlated variation is shown in Figure 3-19 (a). For within-die spatially correlated variation, the estimated percentages of quadratic and layout effects are 2.7% and 97.3% respectively. Therefore, layout-dependent variation is indeed the dominant variation source that causes the spatially correlated variation. To examine whether there exists any significant variation sources not modeled by quadratic basis functions, we further perform sparse regression after adding the DCT dictionary. Because of the high computational cost, we do not perform robust sparse regression but simply remove the outliers detected with the physical basis functions. The computational cost issue will be discussed in detail in the next chapter. Figure 3-19 (b) shows the spatially correlated variation extracted by sparse regression with the layout, physical and DCT dictionaries. The variation percentages are not significantly different from the previous experiments, which show that there do not exist significantly large additional variation sources. However, comparing Figure 3-19 (b) with Figure 3-19 (a), we notice that there is a subtle left-to-right transition at around *x* = 100. This transition is more obvious if we plot only the quadratic and DCT components in Figure 3-19 (c). This transition can be caused by mask error, which is a common variation source for within-die variation. Although this component is not significant in this example, it demonstrates that this type of variation can be revealed by the DCT dictionary.

3.5 Summary

The existence of outliers is an important problem that widely exists in silicon measurement data. If outliers are not appropriately considered, they will introduce substantial error to variation decomposition. In this chapter, we extend the sparse regression algorithm introduced in Algorithm 2 to a robust sparse regression algorithm. By solving robust sparse regression, basis functions will be accurately selected in the presence of outliers, and outliers will be automatically detected and removed, before the data is provided to the linear mixed model to perform variation decomposition. Experiments on synthetic and silicon measurement data demonstrate that the proposed algorithm provides superior accuracy compared to the traditional IQR method for outlier detection. We further performed variation decomposition on several silicon data sets and demonstrated the effectiveness of the proposed variation decomposition flow based on robust sparse regression.

Chapter 4

Fast Implementation for Sparse Regression

4.1 Introduction

In order to perform variation decomposition, we need to perform the robust sparse regression algorithm described in Algorithm 3 to select the basis functions and detect outliers. Algorithm 3 repeatedly solves the following sparse regression problem:

$$\begin{array}{ll} \underset{\eta_{(l)}}{\text{minimize}} & \left\| W_{BS(l)}^{0.5} \left(A_{(l)} \eta_{(l)} - B_{(l)} \right) \right\|_2^2 \\ s.t. & nnz\left(\left\{ \eta_{(l),j} \mid j \notin \Omega_0 \right\} \right) \le \lambda \end{array} \quad (l = 1,2,...,L) \tag{4.1}$$

where $W_{BS(l)}$ is a diagonal matrix. In order to solve (4.1), it can be re-written into the following equation:

$$\begin{array}{ll} \underset{\eta_{(l)}}{\text{minimize}} & \left\| W_{(l)} A_{(l)} \eta_{(l)} - B_{w(l)} \right\|_2^2 \\ s.t. & nnz\left(\left\{ \eta_{(l),j} \mid j \notin \Omega_0 \right\} \right) \le \lambda \end{array} \quad (l = 1,2,...,L) \tag{4.2}$$

where

$$W_{(l)} = W_{BS(l)}^{0.5} \tag{4.3}$$

$$B_{w(l)} = W_{(l)} \cdot B_{(l)} . \tag{4.4}$$

The sparse regression problem (4.2) is solved by Algorithm 2, and therefore the overall computational cost is proportional to the cost of performing Algorithm 2. As will be shown in Section 4.2-4.3, the computational cost of a straightforward implementation of Algorithm 2 is quadratically dependent on the size of the dictionary. When only the physical dictionary is applied, this computational cost is typically small because the both the wafer-level and within-die physical dictionaries defined in Section 2.2.1.1 and 2.2.1.2 contain only a small number of basis functions. However, Algorithm 2 can be extremely computationally expensive for large-scale problems when the DCT dictionary is used. According to its definition in (2.24), the total number of basis functions in the DCT dictionary is *PQ*, which is equal to the total number of points in the two-dimensional grid of interest. This number can become large for both wafer-level and within-die measurement data, since a wafer with small die size can easily contain thousands of dies and a test chip can contain thousands or even millions of test structures. As a result, as will be shown in Section 4.4, when extracting the spatially correlated variation for contact resistance measurement data in Figure 3-19 (b) using physical and DCT basis functions, a straightforward implementation of Algorithm 2 will cost more than one year, making it impractical to be applied. Therefore, a number of implementation details must be carefully considered in order to make Algorithm 2 computationally efficient for large-scale problems.

Based on the above observation, in this chapter, we will we derive several efficient numerical algorithms to address the computational cost issue when the both physical and DCT basis functions are applied. Namely, according to the definition in Section 2.3, we assume that λ_0 physical basis functions have been pre-selected, and we would like to further select a subset of basis functions from *PQ* DCT basis functions. It can be easily observed from Algorithm 2 that the computational cost is dominated by two steps: the inner product computation in Step 4 and the least-squares fitting in Step 6. In Section 4.2, we first derive an efficient numerical algorithm to calculate the inner product values. We will then discuss the numerical algorithm for least-squares fitting in Section 4.3.

4.2 Inner Product Calculation

In Algorithm 2, in order to appropriately select the basis vectors by (2.67), we will need to compute the following inner product values:

$$\left\langle e_{(l)}, A_{w(l),j} \right\rangle, j = 1,2,..., (\lambda_0 + PQ); l = 1,2,..., L \tag{4.5}$$

where

$$A_{w(l)} = W_{(l)} \cdot A_{(l)} \tag{4.6}$$

λ_0 + *PQ* is the total number of basis functions, and *L* is the number of wafers/dies. A straightforward implementation first computes $A_{w(l)}$ by (4.6), and the computational cost is in the order of O($LPQ(\lambda_0 + PQ)$). Then, if the inner product values are simply calculated by vector-vector multiplications in (4.5), the computational cost is again in the order of O($LPQ(\lambda_0 + PQ)$). Note that the computational cost quadratically increases with the DCT dictionary size *PQ*. Hence, the aforementioned implementation can quickly become computationally intractable, as the problem size increases.

For this reason, an efficient numerical algorithm for inner product computation is needed in order to reduce the computational cost. Towards this goal, we first re-write the inner product <$e_{(l)}$, $A_{w(l),j}$> as:

$$\left\langle e_{(l)}, A_{w(l),j} \right\rangle = A_{w(l),j}^{T} \cdot e_{(l)} . \tag{4.7}$$

For each $l \in \{1, 2, \ldots, L\}$, we need to calculate (4.7) for each basis vector, i.e., $j \in \{1,2,\ldots, \lambda_0 + PQ\}$. The results can be expressed by the following matrix-vector multiplication:

$$\begin{bmatrix} \left\langle e_{(l)}, A_{w(l),1} \right\rangle \\ \left\langle e_{(l)}, A_{w(l),2} \right\rangle \\ \vdots \\ \left\langle e_{(l)}, A_{w(l),M} \right\rangle \end{bmatrix} = A_{w(l)}^{T} \cdot e_{(l)} . \tag{4.8}$$

In other words, by calculating the matrix-vector multiplication in (4.8), we are able to obtain the inner product values for all $\lambda_0 + PQ$ basis vectors.

Since direct computation of $A_{w(l)}$ by (4.6) is expensive, we first re-write (4.8) by substituting (4.6) into (4.8):

$$A_{w(l)}^{T} \cdot e_{(l)} = A_{(l)}^{T} \cdot W_{(l)} \cdot e_{(l)} . \tag{4.9}$$

We notice that the following operation can be easily performed with linear complexity:

$$e_{w(l)} = W_{(l)} \cdot e_{(l)} . \tag{4.10}$$

Therefore, we can further re-write (4.9) into the following equation:

$$A_{w(l)}^{T} \cdot e_{(l)} = A_{(l)}^{T} \cdot e_{w(l)} \tag{4.11}$$

where is $e_{w(l)}$ computed by (4.10). Calculating (4.7) with (4.11) prevents the quadratic computational cost related to the computation of (4.6). Note that this implementation also saves the computational cost, even if $A_{(l)}$ only contains physical basis functions.

Next, we need to efficiently compute the matrix-vector product in (4.11). When both physical and DCT dictionaries are applied, we can re-write $A_{(l)}$ into the following:

$$A_{(l)} = \left[A_{\Omega_0(l)} \quad A_{dct(l)} \right] \tag{4.12}$$

where $A_{\Omega_0(l)}$ are the columns of $A_{(l)}$ that correspond to the λ_0 pre-selected physical basis functions, and $A_{dct(l)}$

are the columns of $A_{(l)}$ that correspond to all PQ DCT basis functions. Therefore, we further re-write (4.11) as:

$$\begin{bmatrix} A_{\Omega_0(l)}^T \cdot e_{w(l)} \\ A_{dct(l)}^T \cdot e_{w(l)} \end{bmatrix}. \tag{4.13}$$

Since the number of all DCT basis functions should be much larger than the number of pre-selected physical basis functions, the key bottleneck for computing (4.13) is the second term $A^T_{dct(l)} \cdot e_{w(l)}$. We observe that if the measurement of the l-th wafer/die does not contain any missing data, the matrix $A_{dct(l)}$ represents the IDCT matrix and it is a full-rank square matrix, as defined in (2.23). In this case, since DCT/IDCT is an orthogonal transform [23], $A^T_{dct(l)} = A^{-1}_{dct(l)}$ is exactly the DCT matrix. Namely, calculating the matrix-vector product $A^T_{dct(l)} \cdot e_{w(l)}$ is equivalent to performing DCT on $e_{w(l)}$. Similar to fast Fourier transform (FFT), there exist a number of fast algorithms for DCT/IDCT. The computational cost of these fast algorithms is in the order of O($PQ \cdot \log(PQ)$) [23]. Therefore, by using a fast DCT algorithm, the computational cost for Step 4 of Algorithm 2 is reduced from O($LPQ(\lambda_0 + PQ)$) to O($LPQ(\lambda_0 + \log(PQ))$). This will bring significant speedup, since λ_0 should be much smaller than PQ.

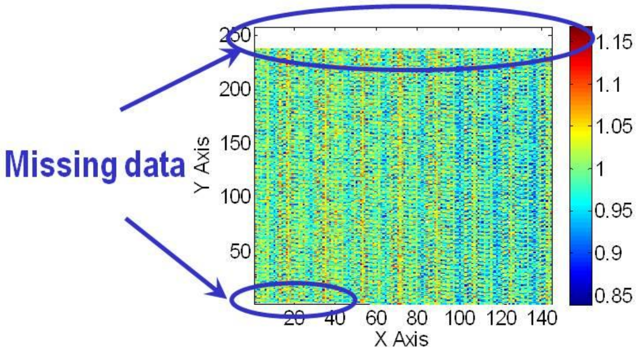


Figure 4-1. Contact resistance measurement data (normalized) from one of the 24 dies that contain significant amount of missing data.

The aforementioned fast DCT algorithm is applicable, if and only if there is no missing data. In this case, the number of available data $N_{(l)}$ is the same as the number of DCT basis functions PQ, and hence, the matrix $A_{(l)}$ is the full-rank square IDCT matrix, which we denote as A^*. However, in practice, a number of missing data often exist in silicon measurement data. For example, Figure 4-1 shows contact resistance measurement data from one of the test chips, in which a significant number of data are missing due to measurement error. A number of missing data can also be artificially introduced by the cross-validation process in Section 2.4.3. If missing data exist, we can construct an augmented vector $e^*_{w(l)} \in R^{PQ}$ where the elements corresponding to missing data are simply filled with zeros. Mathematically, the augmented vector $e^*_{w(l)}$ can be represented as:

$$e^*_{w(l)} = Z_{(l)} \cdot \begin{bmatrix} e_{w(l)} \\ 0 \end{bmatrix} \tag{4.14}$$

where $Z_{(l)}$ is a permutation matrix to map $e_{w(l)}$ and the zero vector to the appropriate elements in $e^*_{w(l)}$. Applying DCT to the augmented vector $e^*_{w(l)}$ yields:

$$A^{*T} \cdot e^*_{w(l)} = A^{*T} \cdot Z_{(l)} \cdot \begin{bmatrix} e_{w(l)} \\ 0 \end{bmatrix} \tag{4.15}$$

where A^* represents the IDCT matrix and, hence, A^{*T} is the DCT matrix. Remember that the matrix $A_{dct(l)}$ contains $N_{(l)}$ rows taken from the IDCT matrix A^*. Hence, the matrix $A^{*T} \cdot Z_{(l)}$ in (4.15) can be re-written as:

$$A^{*T} \cdot Z_{(l)} = \begin{bmatrix} A^T_{dct(l)} & A^T_{dct(\tilde{l})} \end{bmatrix} \tag{4.16}$$

where the matrix $A_{dct(\tilde{l})}$ contains the $PQ - N_{(l)}$ rows of A^* that are not included in $A_{dct(l)}$ due to missing data. Substituting (4.16) into (4.15), we have:

$$A^{*T} \cdot e^*_{w(l)} = \begin{bmatrix} A^T_{dct(l)} & A^T_{dct(\tilde{l})} \end{bmatrix} \cdot \begin{bmatrix} e_{w(l)} \\ 0 \end{bmatrix} = A^T_{dct(l)} \cdot e_{w(l)} . \tag{4.17}$$

Note that the DCT results in (4.17) are exactly equal to the second matrix-vector product in (4.13). It, in turn, demonstrates that by filling the missing data with zeros, we can efficiently calculate the inner product values by using a fast DCT algorithm. In this case, the computational cost for Step 4 of Algorithm 2 is again reduced from O($LPQ(\lambda_0 + PQ)$) to O($LPQ(\lambda_0 + \log(PQ))$).

In addition to the reduction in computational cost, the aforementioned fast algorithm based on DCT can also efficiently reduce the memory consumption. Note that the direct matrix-vector multiplication in (4.11) requires to explicitly form a dense matrix $A_{(l)}$ with about $PQ(\lambda_0 + PQ)$ entries. While it is possible to calculate each inner product in (4.11) one by one without forming the matrix $A_{(l)}$, such an approach leads to large computational time since each column of $A_{(l)}$ must be repeatedly formed during the iterations of Algorithm 2. For these reasons, the direct approach based on matrix-vector multiplication or vector-vector multiplication is expensive in either memory consumption or computational time. On the other hand, our proposed method only needs to form the sub-matrix $A_{\Omega 0(l)}$ with $PQ\cdot\lambda_0$ entries. A fast DCT algorithm can be applied to $e^*_{w(l)}$ without explicitly building the DCT matrix $A_{dct(l)}$ in memory, thereby significantly reducing the memory consumption for large-scale problems.

4.3 Least Squares Fitting

In addition to inner product computation, least-squares fitting is another computationally expensive operation that is required by Step 6 of Algorithm 2. The goal is to solve the following optimization problem:

$$\underset{\eta_{(l),i}, i\in\Omega}{\text{minimize}} \quad \left\| \sum_{i\in\Omega} W_{(l)} \cdot A_{(l),i} \cdot \eta_{(l),i} - B_{w(l)} \right\|_2^2 \quad (l = 1,2,..., L) \tag{4.18}$$

where Ω contains the indices of selected basis vectors. In this sub-section, we will develop an efficient numerical algorithm to reduce the computational cost of (4.18).

We first re-write (4.18) for the *l*-th wafer/die at the *p*-th iteration step:

$$\underset{\eta_{(l),(p)}}{\text{minimize}} \quad \left\| W_{(l)} \cdot A_{(l),(p)} \cdot \eta_{(l),(p)} - B_{(l)} \right\|_2^2 \tag{4.19}$$

where the matrix $A_{(l),(p)}$ contains λ_0+p column vectors selected from $A_{(l)}$ and the vector $\eta_{(l),(p)}$ contains the coefficients corresponding to these selected basis vectors. Similarly, we can re-write $A_{(l),(p)}$ into the following:

$$A_{(l),(p)} = \begin{bmatrix} A_{\Omega_0(l)} & A_{dct(l),(p)} \end{bmatrix} \tag{4.20}$$

where the matrix $A_{dct(l),(p)}$ contains p column vectors selected from $A_{dct(l)}$. The relation between $A_{dct(l),(p)}$ and $A_{dct(l)}$ can be further expressed as:

$$A_{dct(l)} \cdot Z_{(p)} = \begin{bmatrix} A_{dct(l),(p)} & A_{dct(l),(\tilde{p})} \end{bmatrix} \tag{4.21}$$

where $Z_{(p)}$ is a permutation matrix, and the matrix $A_{dct(l),(\tilde{p})}$ contains the DCT basis vectors that are not included in $A_{dct\ (l),(p)}$.

The least-squares solution $\eta_{(l),(p)}$ of (4.19) satisfies the following normal equation [27]:

$$\left(W_{(l)} \cdot A_{(l),(p)}\right)^T \cdot \left(W_{(l)} \cdot A_{(l),(p)}\right) \cdot \eta_{(l),(p)} = \left(W_{(l)} \cdot A_{(l),(p)}\right)^T \cdot B_{w(l)} \ . \tag{4.22}$$

Traditionally, the solution $\eta_{(l),(p)}$ of (4.22) is solved by QR decomposition [27]. In order to perform QR decomposition, we first need to explicitly perform the multiplication of $W_{(l)}$ and $A_{(l),(p)}$:

$$A_{w(l),(p)} = W_{(l)} \cdot A_{(l),(p)} \, . \tag{4.23}$$

Next, QR decomposition is performed on the matrix $A_{w(l),(p)}$:

$$A_{w(l),(p)} = Q_{w(l),(p)} \cdot R_{w(l),(p)} \tag{4.24}$$

where $Q_{w(l),(p)}$ is an $N_{(l)}$-by-(λ_0+p) matrix with orthonormal columns and $R_{(l),(p)}$ is a (λ_0+p)-by-(λ_0+p) upper triangular matrix. Substituting (4.23) and (4.24) into (4.22) yields:

$$R_{w(l),(p)} \cdot \eta_{(l),(p)} = Q^T_{w(l),(p)} \cdot B_{w(l)} \, . \tag{4.25}$$

In (4.25), since $R_{w(l),(p)}$ is upper triangular, $\eta_{(l),(p)}$ can be solved by back substitution. The computational cost of the aforementioned least-squares fitting is dominated by the QR decomposition step and it is in the order of O($N_{(l)}$·(λ_0+p)^2).

The traditional least-squares solver based on QR decomposition is not computationally efficient for large-scale problems. Similar to the previous sub-section, we would like to utilize the fact that matrix-vector products can be efficiently computed by fast DCT/IDCT algorithms. Therefore, we need an iterative solver for the least-squares problem (4.19). A naïve method to solve (4.19) with an iterative solver is to apply the conjugate gradient method to the normal equation (4.22), but it is known to be numerically unstable [54]. An improved iterative solver for (4.19) is referred to as the LSQR method [22]. Unlike the conjugate gradient method based on (4.22), LSQR aims to directly solve (4.19) in order to improve numerical stability. LSQR relies on bi-diagonalization of the matrix $A_{w(l),(p)}$. During its iterations, LSQR generates a sequence of solutions to approximate $\eta_{(l),(p)}$. These solutions are exactly identical to the results calculated by the conjugate gradient method for the normal equation in (4.22). The details of LSQR can be found in [22].

When applying LSQR, it is not necessary to explicitly form the matrix $A_{w(l),(p)}$. Instead, in each iteration, only two matrix-vector multiplications need to be performed, $A_{w(l),(p)}$·α and $A^T{}_{w(l),(p)}$·β, where α is a (λ_0+p)-by-1 vector and β is an $N_{(l)}$-by-1 vector. These matrix-vector multiplications can be efficiently calculated by applying a fast numerical algorithm based on fast DCT/IDCT transform. In what follows, we will show the mathematical formulation of our proposed fast algorithm.

First, we notice that to efficiently compute $A_{w(l),(p)}$·α, it is not necessary to compute (4.23), since

$$A_{w(l),(p)} \cdot \alpha = W_{(l)} \cdot \left(A_{(l),(p)} \cdot \alpha\right). \tag{4.26}$$

Therefore, we only need to efficiently compute $A_{(l),(p)}\cdot\alpha$, and $A_{w(l),(p)}\cdot\alpha$ will then be calculated by simply performing a vector-vector product. Next, we re-write $A_{(l),(p)}\cdot\alpha$ into the following equation:

$$A_{(l),(p)} \cdot \alpha = \begin{bmatrix} A_{\Omega_0(l)} & A_{dct(l),(p)} \end{bmatrix} \cdot \begin{bmatrix} \alpha_{\Omega_0} \\ \alpha_{dct} \end{bmatrix} = A_{\Omega_0(l)} \cdot \alpha_{\Omega_0} + A_{dct(l),(p)} \cdot \alpha_{dct} \tag{4.27}$$

where α_{Ω_0} is an λ_0-by-1 vector and α_{dct} is a p-by-1 vector. Of these two matrix-vector multiplications, $A_{\Omega_0(l)}\cdot\alpha_{\Omega_0}$ has to be computed using the traditional matrix vector product, but we are able to efficiently compute $A_{dct(l),(p)}\cdot\alpha_{dct}$. To efficiently compute $A_{dct(l),(p)}\cdot\alpha_{dct}$, we construct an augmented vector $\alpha^*_{dct} \in R^{PQ}$:

$$\alpha^*_{dct} = Z_{(p)} \cdot \begin{bmatrix} \alpha_{dct} \\ 0 \end{bmatrix} \tag{4.28}$$

where $Z_{(p)}$ is the permutation matrix defined in (4.21). If we conceptually consider α_{dct} as a vector of selected DCT coefficients, then α^*_{dct} represents all DCT coefficients with the unselected DCT coefficients filled by 0. We then apply IDCT to the augmented vector α^*_{dct}:

$$A^* \cdot \alpha^*_{dct} = A^* \cdot Z_{(p)} \cdot \begin{bmatrix} \alpha_{dct} \\ 0 \end{bmatrix} \tag{4.29}$$

where A^* denotes the IDCT matrix as defined in (4.15). On the other hand, we can derive the following equation from (4.16):

$$A^* = Z_{(l)} \cdot \begin{bmatrix} A_{dct(l)} \\ A_{dct(\tilde{l})} \end{bmatrix}. \tag{4.30}$$

Substituting (4.30) into (4.29) yields:

$$A^* \cdot \alpha^*_{dct} = Z_{(l)} \cdot \begin{bmatrix} A_{dct(l)} \cdot Z_{(p)} \\ A_{dct(\tilde{l})} \cdot Z_{(p)} \end{bmatrix} \cdot \begin{bmatrix} \alpha_{dct} \\ 0 \end{bmatrix}. \tag{4.31}$$

In (4.31), $A_{dct(l)}\cdot Z_{(p)}$ can be represented as two sub-matrices as shown in (4.21). If we similarly re-write $A_{dct(\tilde{l})}\cdot Z_{(p)}$ as two sub-matrices:

$$A_{dct(\tilde{l})} \cdot Z_{(p)} = \begin{bmatrix} A_{dct(\tilde{l}),(p)} & A_{dct(\tilde{l}),(\tilde{p})} \end{bmatrix} \quad . \tag{4.32}$$

Eq. (4.31) becomes:

$$A^{*}\cdot\alpha_{dct}^{*}=Z_{(l)}\cdot\begin{bmatrix}A_{dct(l),(p)} & A_{dct(l),(\tilde{p})}\\ A_{dct(\tilde{l}),(p)} & A_{dct(\tilde{l}),(\tilde{p})}\end{bmatrix}\cdot\begin{bmatrix}\alpha_{dct}\\ 0\end{bmatrix}=Z_{(l)}\cdot\begin{bmatrix}A_{dct(l),(p)}\cdot\alpha_{dct}\\ A_{dct(\tilde{l}),(p)}\cdot\alpha_{dct}\end{bmatrix}. \tag{4.33}$$

Since $Z_{(l)}$ is a permutation matrix, Eq. (4.33) is equivalent to:

$$\begin{bmatrix}A_{dct(l),(p)}\cdot\alpha_{dct}\\ A_{dct(\tilde{l}),(p)}\cdot\alpha_{dct}\end{bmatrix}=Z_{(l)}^{T}\cdot A^{*}\cdot\alpha_{dct}^{*}\ . \tag{4.34}$$

Eq. (4.34) reveals an important fact that the matrix-vector multiplication $A_{dct(l),(p)}\cdot\alpha_{dct}$ can be efficiently computed by applying IDCT to the augmented vector α^{*}_{dct}. The value of $A_{dct(l),(p)}\cdot\alpha_{dct}$ is determined by selecting the appropriate elements from the IDCT result $A^{*}\cdot\ \alpha^{*}_{dct}$. If a fast IDCT algorithm is applied [23], the computational cost of this matrix-vector calculation is in the order of O($PQ\cdot\log(PQ)$). Therefore, the computational cost of the matrix-vector product $A_{w(l),(p)}\cdot\alpha$ is O($PQ\cdot(\lambda_0+\log(PQ))$).

Next, we consider the other matrix-vector multiplication $A^{T}_{w(l),(p)}\cdot\beta$ that is required by the LSQR algorithm. Similar to the computation in (4.9)-(4.11), we first re-write $A^{T}_{w(l),(p)}\cdot\beta$ by substituting (4.23) into it:

$$A_{w(l),(p)}^{T}\cdot\beta=A_{(l),(p)}^{T}\cdot W_{(l)}\cdot\beta\ . \tag{4.35}$$

We first simply perform the following operation with linear complexity:

$$\beta_{w}=W_{(l)}\cdot\beta\ . \tag{4.36}$$

Then, we can further re-write $A^{T}_{w(l),(p)}\cdot\beta$ into the following equation:

$$A_{w(l),(p)}^{T}\cdot\beta=A_{(l),(p)}^{T}\cdot\beta_{w}\ . \tag{4.37}$$

Next, we re-write $A^{T}_{(l),(p)}\cdot\ \beta_w$ into the following equation based on (4.20):

$$A_{(l),(p)}^{T}\cdot\beta_{w}=\begin{bmatrix}A_{\Omega_0(l)}^{T}\cdot\beta_{w}\\ A_{dct(l),(p)}^{T}\cdot\beta_{w}\end{bmatrix}. \tag{4.38}$$

Of these two matrix-vector multiplications, $A^{T}_{\Omega_0(l)}\cdot\beta_w$ has to be computed using the traditional matrix vector product, but we are able to efficiently compute $A^{T}_{dct(l),(p)}\cdot\beta_w$. Similarly, we first construct an augmented vector $\beta^{*}_{w}\in R^{PQ}$:

$$\beta_{w}^{*}=Z_{(l)}\cdot\begin{bmatrix}\beta_{w}\\ 0\end{bmatrix} \tag{4.39}$$

where $Z_{(l)}$ is the permutation matrix defined in (4.14). Similar to (4.14), if we conceptually consider β_w as a vector of available measurement, then β^*_w represents all measurements with the missing measurements filled by 0. We then apply DCT to the augmented vector β^*_w:

$$A^{*T}\cdot\beta_w^* = A^{*T}\cdot Z_{(l)}\cdot\begin{bmatrix}\beta_w\\ 0\end{bmatrix} \tag{4.40}$$

where A^{*T} is the DCT matrix as defined in (4.15). Substituting (4.30) into (4.40) yields:

$$A^{*T}\cdot\beta_w^* = \begin{bmatrix}A^T_{dct(l)} & A^T_{dct(\tilde{l})}\end{bmatrix}\cdot Z^T_{(l)}\cdot Z_{(l)}\cdot\begin{bmatrix}\beta_w\\ 0\end{bmatrix} = A^T_{dct(l)}\cdot\beta_w\,. \tag{4.41}$$

Based on (4.21), Eq. (4.41) can be further re-written as:

$$\begin{bmatrix}A^T_{dct(l),(p)}\cdot\beta_w\\ A^T_{dct(l),(\tilde{p})}\cdot\beta_w\end{bmatrix} = Z^T_{(p)}\cdot A^{*T}\cdot\beta_w^*\,. \tag{4.42}$$

Hence, the matrix-vector multiplication $A^T_{dct(l),(p)}\cdot\beta_w$ can be calculated by applying DCT to the augmented vector β^*_w. The value of $A^T_{dct(l),(p)}\cdot\beta_w$ is determined by selecting the appropriate elements from the DCT result $A^{*T}\cdot\beta^*_w$. The computational cost is in the order of O($PQ\cdot\log(PQ)$). Therefore, the computational cost of the matrix-vector product $A^T_{w(l),(p)}\cdot\beta$ is also O($PQ\cdot(\lambda_0+\log(PQ))$).

Finally, it is worth mentioning that similar to other iterative solvers, a good initial guess should be provided to LSQR to achieve fast convergence. If the initial guess is close to the actual solution, LSQR can reach convergence in a few iterations [22]. In this paper, LSQR is required at each iteration step of the Algorithm 2. When Algorithm 2 is applied, the solution from the previous iteration step can serve as a good initial guess for the current iteration step. By adopting such a heuristic, LSQR typically converges in only 2-3 iterations in our tested examples.

4.4 Numerical Results

The aforementioned fast implementation of sparse regression solver will be extremely useful if the wafer/die collects data from a very dense grid. For example, when the die size is small, a modern wafer can easily contain several thousand dies. Also, a test chip can contain a very large array of test structures. In this sub-section, we will use the contact resistance measurement data to demonstrate the efficiency improvements of the proposed fast implementation on a large problem.

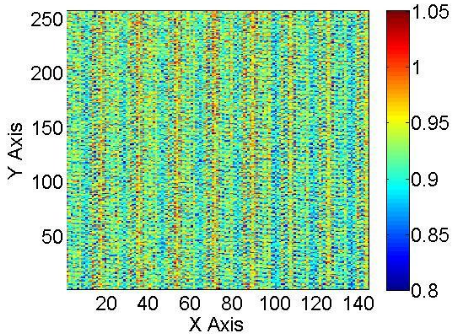


Figure 4-2. Contact resistance measurement data (normalized) from one of the 24 dies.

The contact plug resistance measurement data are collected from 24 test chips in a 90 nm CMOS process. Figure 4-2 shows contact resistance measurement data collected from one of the dies, in which data is collected from a 144×256 array. Therefore, the total number of test structures per chip is 36,864. Based on the definition of DCT basis functions in (2.24), the total number of DCT basis functions in the DCT dictionary is also 36,864. After performing sparse regression with the physical dictionary, 60 basis functions are selected, and the problem we examine in this experiment is to select additional basis functions from 36,864 candidates in the DCT dictionary. The DCT dictionary size is much larger than the number of pre-selected physical basis functions, which agrees with our previous discussion.

To demonstrate the efficiency of the fast numerical algorithms proposed, we implement three different versions of Algorithm 2 where the inner product and the least-squares fitting are calculated by different methods. In the first implementation, the inner product is directly computed by (4.5) and the least-squares fitting is directly computed by the QR decomposition in (4.23)-(4.25). In the second implementation, the traditional inner product calculation is replaced by the fast algorithm proposed in Section 4.2. Finally, in the third implementation, both the inner product and the least-squares fitting are calculated by the fast algorithms proposed in Section 4.2-4.3.

For testing and comparison purposes, we first run Algorithm 2 with the aforementioned three implementations on only the die in Figure 4-2. Table 4-1 shows the computational time for the proposed variation decomposition of a single test chip. Note that the fast algorithm for inner product computation achieves 91× speed-up and the fast least-squares fitting further brings 2.2× speed-up. The overall speed-up achieved by our proposed fast algorithms is 199×, compared to the traditional direct implementation.

Table 4-1. Computational time of sparse regression for a single chip

Inner product	Least-squares fitting	CPU time (Sec.)
Direct	Direct	1.85×10^{6}
Fast	Direct	2.03×10^{4}
Fast	Fast	9.31×10^{3}

Next, we run Algorithm 2 for all 24 test chips and Table 4-2 compares the computational time for two different implementations. Once Algorithm 2 is applied to all test chips, the computational time increases significantly. The simple implementation with direct inner product calculation and least-squares fitting is not computationally feasible. Hence, its result is not shown in Table 4-2. In this example, the proposed fast algorithm for least-squares fitting achieves 2.1× speed-up over the direct implementation. This is consistent with the speedup in Table 4-1. We infer that if the direct implementation is adopted in this example, it would take more than one year to obtain the results, which makes Algorithm 2 inapplicable. Therefore, by applying the fast implementation of Algorithm 2, we are able to extend its applicability to problems with large size.

Table 4-2. Computational time of sparse regression for 24 chips

Inner product	Least-squares fitting	CPU time (Sec.)
Fast	Direct	4.89×10^{5}
Fast	Fast	2.35×10^{5}

4.5 Summary

The computational cost for sparse regression with DCT basis functions can become extremely large for problems with large size, which limits the applicability of the variation decomposition methodology introduced in Chapter 2-Chapter 3. In this chapter, we proposed several efficient methods to make the computational cost of sparse regression tractable for large-scale problems. The key idea of these methods is to utilize fast DCT/IDCT computation to speed up the matrix-vector product computation. From the experimental results on the contact resistance data, we observe nearly 200× speedup compared to the traditional direct implementation.

Chapter 5

Wafer Spatial Signature Clustering

5.1 Introduction

In Chapter 2-Chapter 4, we introduced a statistical framework for variation decomposition which assists the identification of major physical variation sources for wafers with similar spatial signature, so that the process engineers can focus on the variation sources that have a major contribution on overall yield. While wafers during process development and product yield ramp stages typically have similar spatial signature, in volume production different wafers may exhibit completely different spatial signatures. For example, Figure 5-1 shows normalized I_{dsat} measurements on two different wafers, where significant difference in the spatial signatures can be seen. Such difference in spatial signature may suggest a number of underlying yield-limiting factors, such as process shift/drift, mismatch between equipments, mismatch between different chambers, etc. Therefore, if we can automatically partition all the wafers into different groups, in which each group exhibits a similar spatial signature, it would provide important insight to help process engineers with the yield improvement effort. Especially, process engineers can prioritize the yield improvement goals and focus on the mechanism related to large groups with strong spatial signature, so that reducing the variation sources that correspond to such spatial signature will have a significant impact on the improvement of overall yield.

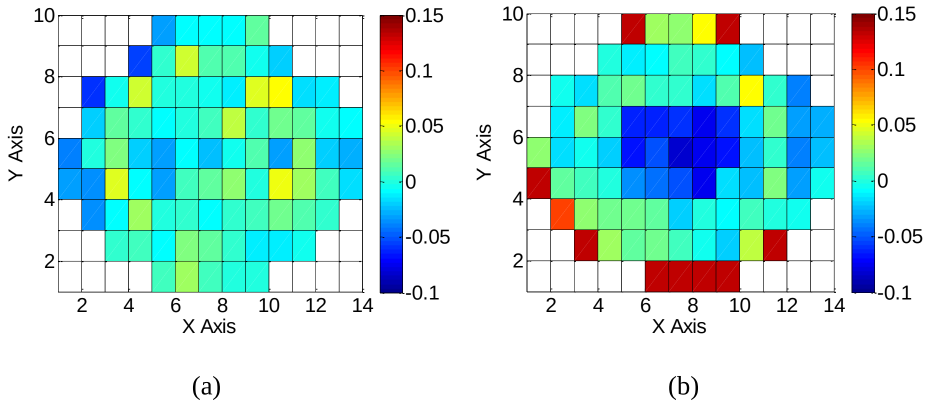


Figure 5-1. I_{dsat} measurements (normalized) on two wafers with different spatial signatures.

The problem of automatically grouping wafers with similar spatial signatures can be defined as a clustering analysis problem in statistics [65]. While clustering analysis has been extensively studied in the statistics literature, a number of unique characteristics of the wafer spatial signature clustering problem must be carefully considered in order to obtain accurate clustering results:

- *Large random variation*: the performance measurements collected from test structures may be subject to large random variation. The random variation will be more significant with technology scaling, where fundamental device variability is playing an increasingly important part in overall variation. Large random variation will obscure the spatial signature, making it difficult to identify the difference between different spatial signatures.
- *Missing and outlier measurements*: Defects in the manufacturing process, as well as the measurement error may generate missing measurements, where no data can be collected from some test structures, or outlier measurements, where the collected data significantly deviate from the regular variation range. Meaningful clustering results cannot be obtained if these problems are not properly addressed.
- *Abnormal wafers*: Because of equipment malfunction, there can be a small number of abnormal wafers whose spatial signature is significantly different from the vast majority of wafers [66]. We would like to automatically detect these abnormal wafers, rather than merging them into the main clusters. This poses a significant challenge to the clustering algorithm, which will be explained in detail in Section 5.3.

- *Unknown number of clusters*: Most clustering algorithms require knowing the number of clusters, or have user-defined parameters related to the number of clusters. In our wafer spatial signature clustering application, the number of clusters cannot be known in advance. Therefore, additional efforts must be made to determine the number of clusters from data.

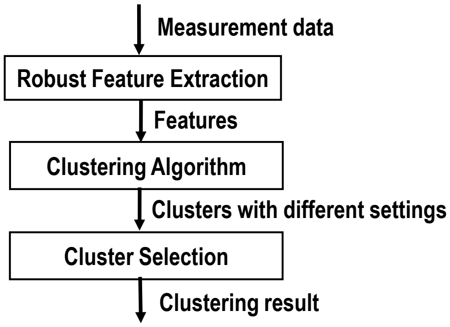


Figure 5-2. Proposed flow to achieve wafer spatial signature clustering.

Based on the above characteristics, we propose a wafer spatial signature clustering method based on the flow shown in Figure 5-2, which consists of three components. Robust feature extraction is first performed on the measurement data, which represents the spatial signature of each wafer by a small number of features. The impact of random variation will be significantly reduced in the feature space, and the feature extraction process must be insensitive to the missing and outlier measurements. Next, a clustering algorithm will be performed on the extracted features. Since the number of clusters is not known in advance, the clustering algorithm will not directly generate the final clustering result. Instead, a set of possible clustering results will be generated based on different settings of the clustering algorithm. In the final step, a cluster selection algorithm will be applied to adaptively choose the clustering result that best explains the data.

The remainder of the chapter is organized as follows. In Section 5.2 we present the robust feature extraction algorithm. Then we discuss the choice of the clustering algorithm in Section 5.3. The algorithm for selecting the final clustering result will be presented in Section 5.4. The efficacy of the proposed method is demonstrated by several examples in Section 5.5. Finally, we summarize our findings in Section 5.6.

5.2 Robust Feature Extraction

As stated in the previous sub-section, the goal of robust feature extraction is to represent the spatial signature of each wafer by a small number of features that reduce the impact of random variation, missing data and outlier measurements. Similar to the basis function selection problem in Section 2.3, suppose that we collected measurements from L wafers, the spatial variation of these L wafers can be represented by L two-dimensional functions: $\{b_{(l)}(x, y);\ l = 1, 2, \ldots, L\}$. Each spatial variation function contains two components:

$$b_{(l)}(x, y) = s_{(l)}(x, y) + r_{(l)}(x, y) \tag{5.1}$$

where $s_{(l)}(x, y)$ stands for the spatially correlated variation and $r_{(l)}(x, y)$ stands for the uncorrelated random variation for wafer l, respectively. To reduce the impact of random variation, we will represent the spatial signature of each wafer by only using its spatially correlated component. Specifically, if the spatially correlated variation is modeled by the linear combination of λ basis functions:

$$b_{(l)}(x, y) = \sum_{j=1}^{\lambda} \eta_{(l),j} \cdot A_j(x, y) + r_{(l)}(x, y) \quad , \tag{5.2}$$

we define the features of wafer l as the following vector:

$$\eta_{(l)} = \begin{bmatrix} \eta_{(l),1} & \eta_{(l),2} & \cdots & \eta_{(l),\lambda} \end{bmatrix}^T . \tag{5.3}$$

By using the λ features in (5.3) to represent the spatial signature of wafer l, the uncorrelated variation $r_{(l)}(x, y)$ will not be considered in the subsequent clustering process, and the clustering result will be less sensitive to random variation.

In practice, we do not know in advance what spatial signatures are present in the wafers. If a pre-determined set of basis functions are used, too few basis functions may not be enough to cover all possible signatures, while too many basis functions will limit the ability to remove random variation. Therefore, following the same idea as Chapter 2-Chapter 3, a large dictionary of possible basis functions can be employed, and the relevant basis functions can be automatically selected by cross-validation. Namely, we solve the following robust sparse regression problem to generate the features:

$$\begin{array}{lll} \underset{\eta_{(l)}}{\text{minimize}} & \sum_{i=1}^{N_{(l)}} \rho_{BS}\left(b_{(l),i} - A_{i,(l)}\eta_{(l)}\right) & \\ s.t. & \left\|\eta_{(l)}\right\|_0 \le \lambda & \end{array} \quad (l = 1,2,..., L) \tag{5.4}$$

where $b_{(l),i}$ represents the i-th performance measurement on wafer l, and

$$A_{(l)} = \begin{bmatrix} A_{1,(l)} \\ A_{2,(l)} \\ \vdots \\ A_{N(l),(l)} \end{bmatrix} \tag{5.5}$$

contains M columns that correspond to M basis functions in the dictionary, and each row $A_{i,(l)}$ correspond to the basis function values for the i-th performance measurement in wafer l, ρ_{BS} is the bi-square cost function defined in (3.12), (3.13) and (3.15). The index set of selected basis functions are required to be the same for each wafer, similar to the basis function selection problem in Section 2.3. The optimization problem (5.4) can be solved by the iteratively reweighting algorithm in Algorithm 3, where the optimal λ value is automatically determined by cross-validation described in Section 2.4.3. Note that the optimization (5.4) explicitly detects and removes outlier measurements by employing a robust cost function, so that the features can still be accurately determined in the presence of outliers. Also, the sparse regression is extremely insensitive to missing measurements, which is demonstrated in both compressed sensing literature [8]-[14] and our experiments [19]. Therefore, by applying robust sparse regression to extract the features, the subsequent clustering process will be shielded from the missing and outlier measurements problems in the raw measurement data.

An important problem when applying sparse regression is how to select the dictionary of basis functions. For the wafer spatial signature clustering application, two important factors must be taken into account when selecting the dictionary: *sparsity* and *orthogonality*. Sparsity means that only a small number of basis functions are needed to accurately represent any spatial signature. Obviously, a dictionary with better sparsity offers better protection against random variation. Orthogonality requires that the basis functions in the dictionary must not be correlated. If such correlation exists, it is possible for a small difference in spatial signature of measurement data to be translated into a significant difference in the features. This will in turn yield counter-intuitive clustering results. The physical dictionary introduced in Section 2.2.1 does not guarantee orthogonality, and therefore is not used in our wafer spatial signature clustering application. Several dictionaries of basis functions that offer both sparsity and orthogonality have been proposed in the image processing literature, including discrete cosine transform (DCT) and wavelet basis functions. As was explained in Section 2.2.2, since we found that DCT typically provides better sparsity than wavelet when representing spatial process variation measurement data, we employ DCT basis functions described in Section 2.2.2 as the dictionary in the optimization problem (5.4).

5.3 Clustering Algorithm

After the features describing the wafer signatures are extracted, the next step is to apply a clustering algorithm to partition the wafers into clusters. Many clustering algorithms have been proposed in the statistics literature, such as k-means clustering [73], density-based clustering [74] and hierarchical clustering [65]. Each clustering algorithm is based on a different assumption on the data and not all algorithms are suitable for clustering wafer spatial variation data. We first briefly review the most traditional clustering method – k-means clustering [73] and explain why it is not an appropriate method for our wafer spatial signature clustering application.

The k-means algorithm partitions the data into K clusters that minimizes the following cost function:

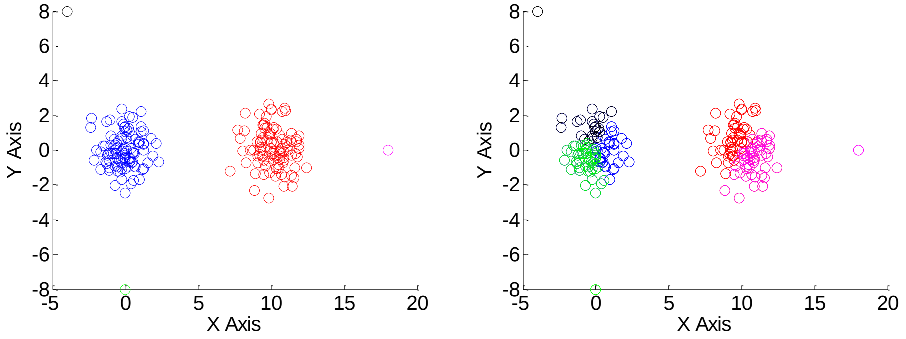
$$\sum_{i=1}^{K}\sum_{l\in c_i}\left\|\eta_{(l)}-\mu_i\right\|_2^2 \tag{5.6}$$

where c_i is the index of wafers that belong to cluster i, $\eta_{(l)}$ is the feature vector of wafer l defined in (5.3) and μ_i is the centroid of cluster i, defined as:

$$\mu_i=\frac{1}{|c_i|}\sum_{l\in c_i}\eta_{(l)} \tag{5.7}$$

where $|\ c_i\ |$ stands for the size of c_i. The number of cluster K is a parameter that has to be specified by the user. When K is not known in advance, another algorithm must be developed to determine K from the data, which will be discussed in more detail in the next sub-section.

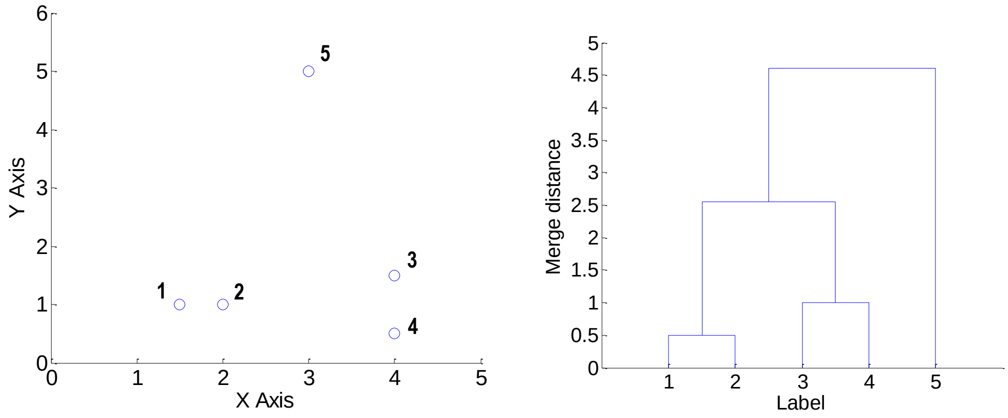
In order to solve the k-means problem defined in (5.6)-(5.7), the first step is to select initial cluster centers from K randomly selected wafers as the seeds. The algorithm then moves the cluster centers around in space in order to minimize (5.6). This is done iteratively by repeating two steps until a stopping criterion is met: reassigning wafers to the cluster with the closest centroid and recomputing each centroid based on the current members of its cluster by (5.7). The cost function (5.6) will monotonically decrease with the iterations and the k-means algorithm will stop once the value of (5.6) has converged. The k-means algorithm does not guarantee a global minimum for (5.6). Therefore, in practice, k-means algorithm is often repeatedly performed many times with different initial seeds and the clustering result with the minimum cost function (5.6) will be selected as the final result. More details of k-means can be found in [73] and [65].



(a) (b)

Figure 5-3. (a) Synthetic two-dimensional data with 2 clusters and 3 outliers and (b) k-means clustering result with 5 clusters.

For our wafer spatial signature clustering application, an important problem that prevents k-means from achieving accurate clustering results is the existence of abnormal wafers. Unlike the outlier measurements discussed in the previous sub-section, abnormal wafers are a small number of wafers whose spatial signatures are significantly different from any of the main clusters. In the feature space, the abnormal wafers are typically far away from any normal wafer. An accurate clustering result will produce separate clusters with very small size that reflect the abnormal wafers, rather than merging them into the large clusters. However, this goal is often not achievable with the k-means clustering algorithm. In order to demonstrate this, we constructed a synthetic 2-D data set in Figure 5-3 (a). Figure 5-3 (a) contains two main clusters and three abnormal data points that are far away from the main clusters. Intuitively, this data set should be partitioned into 5 clusters represented by different colors in Figure 5-3 (a), where each abnormal point forms a single cluster since it is not close to any other data. The k-means clustering result is shown in Figure 5-3 (b), where the number of clusters is pre-specified as 5. From Figure 5-3 (b), it can be seen that the k-means algorithm splits the large clusters, rather than assign separate clusters for the abnormal data. Comparing the k-means cost function (5.6) of Figure 5-3 (b) with Figure 5-3 (a), we find the cost function of Figure 5-3 (b) is indeed lower. This is because each point has equal weight in (5.6), so that small clusters would naturally have lower total weight. In other words, k-means favors clusters with similar size. Therefore, the k-means clustering algorithm is not suitable for situations where the size of clusters can be significantly different.



(a)

(b)

Figure 5-4. (a) Synthetic two-dimensional data with 5 points and (b) The dendrogram generated by hierarchical clustering.

An alternative algorithm that does not suffer from the aforementioned problem with k-means is hierarchical clustering [65]. Unlike the k-means algorithm which explicitly minimizes a cost function, hierarchical clustering builds clusters in a greedy manner. Suppose that there are N data points in total, hierarchical clustering starts by assigning each an individual cluster for each point. Then, N-1 merging steps are performed iteratively, in which each step merges the two clusters that are closest in distance. Therefore, the data points that are close will be merged first, and those that are far away will not be merged until the end of the algorithm. To intuitively explain the idea of hierarchical clustering, we construct a synthetic data set with 5 points in Figure 5-4 (a). When hierarchical clustering is applied on this data set, the first two steps will merge data point 1 with 2, and data point 3 with 4. These two clusters will be merged in the third step, and data point 5 will be merged into it in the last step. The clustering result can be represented as a tree called the dendrogram in Figure 5-4 (b), which reflects how the data points are merged. In Figure 5-4 (b), the height of each node reflects its merge distance, which means the distance of two clusters that are merged to form this node. Note that points that are close will be merged early, while distant points will not be merged until the last steps. However, hierarchical clustering does not directly generate the final clustering result (i.e. cluster labels for each data point). We will discuss algorithms to determine the cluster labels from hierarchical clustering result in the next sub-section.

An important component that must be assigned when performing the hierarchical clustering algorithm is how to define the distance between clusters. While the distance between two individual data points can be simply defined by their Euclidean distance:

$$dist\left(\eta_{(l)}, \eta_{(k)}\right) = \left\|\eta_{(l)} - \eta_{(k)}\right\|_2 , \tag{5.8}$$

the definition of distance between clusters that may contain multiple points is not unique. Different definitions of cluster distance have been proposed, each corresponds to a different assumption about the cluster structure. We need to select the distance definition whose assumption best matches our goal in wafer spatial signature clustering. In this work, the following definition is selected, which determines the distance between two clusters as the maximum of the distance between any two points in these two clusters:

$$dist(c_l, c_k) = \underset{i \in c_l, j \in c_k}{\text{maximum}} \quad \left\|\eta_{(i)} - \eta_{(j)}\right\|_2 . \tag{5.9}$$

The hierarchical clustering algorithm based on the distance metric in (5.9) is named complete-link hierarchical clustering [65]. The physical meaning of (5.9) is that a cluster will be formed only if *all* members in the clusters are completely connected, i.e. within a small distance to each other. This definition matches our goal in wafer spatial signature clustering: since all members in a cluster should correspond to the same spatial signature, we would like any two wafers to be similar. To further explain why (5.9) is the appropriate choice for wafer spatial signature clustering, we compare it with another most commonly used definition which uses the minimum distance to define the distance of two clusters:

$$dist(c_l, c_k) = \underset{i \in c_l, j \in c_k}{\text{minimum}} \quad \left\|\eta_{(i)} - \eta_{(j)}\right\|_2 . \tag{5.10}$$

The hierarchical clustering algorithm based on the distance metric in (5.10) is named single-link hierarchical clustering [65]. The assumption behind single-link hierarchical clustering is that two points should belong to the same cluster as long as there exists a path connecting these two points, such that any adjacent pair of points along this path is close in distance. As a result, single-link hierarchical clustering is known to often generate elongated clusters, where extremely distant points are connected by a long path in between. While this type of cluster is suitable for many applications, it is undesirable in our wafer spatial signature clustering. For example, in the manufacturing process, the change in process condition may not occur abruptly, but instead gradually drift from one state to another. This can happen because of, for example, equipment aging [75]. By employing single-link hierarchical clustering, we are unable to split the wafers into clusters to reflect the change of process condition in this scenario. Note that several other clustering techniques, for example density-based clustering [74], are also based on the idea of a connecting path when forming clusters, and therefore are not suitable for the wafer spatial signature clustering application. On the other hand, complete-link hierarchical clustering will naturally break down a long string of data points into smaller clusters, and is therefore more suitable for this application. In Section 5.5, we will show several examples where the natural clusters detected by complete-link hierarchical clustering cannot be found by either single-link hierarchical clustering or k-means.

5.4 Cluster Selection

In the previous sub-section, we propose to apply complete-link hierarchical clustering for wafer spatial signature clustering. However, as was discussed previously, hierarchical clustering does not directly generate the clustering labels. In practice, the clustering labels may be obtained by asking the user to visit the dendrogram in a top-down manner and decide whether to keep each merge action. To achieve automatic clustering and minimize human efforts, a separate algorithm needs to be applied to automatically select the most intuitive cluster labels from the hierarchical clustering result.

The traditional approach to select the clusters from the hierarchical clustering result is the inconsistency coefficient method [76]. The inconsistency coefficient method visits each node in the dendrogram and compares its merge distance with the average merge distance of nodes below it. Such difference is quantitatively defined by the following inconsistency coefficient:

$$I_k = \frac{d_k - \mu_k}{\sigma_k} \tag{5.11}$$

where I_k represents the inconsistency coefficient of a node k, d_k is the merge distance of k, μ_k is the average merge distance of k and all nodes below k, and σ_k is the standard deviation of the merge distances of k and all nodes below k. Figure 5-5 shows an example where the inconsistency coefficient of the node connecting the cluster {1,2} with the cluster {3,4} is calculated by:

$$I_0 = \frac{d_0 - mean(d_0, d_1, d_2)}{std(d_0, d_1, d_2)}. \tag{5.12}$$

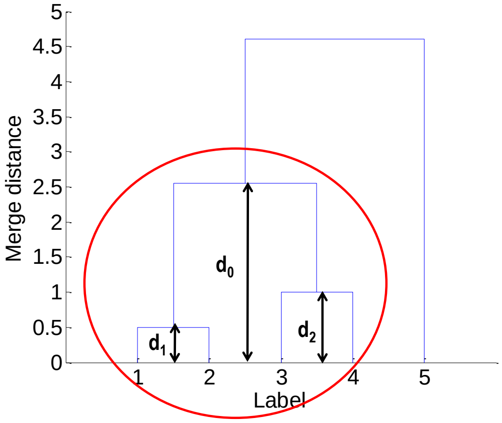
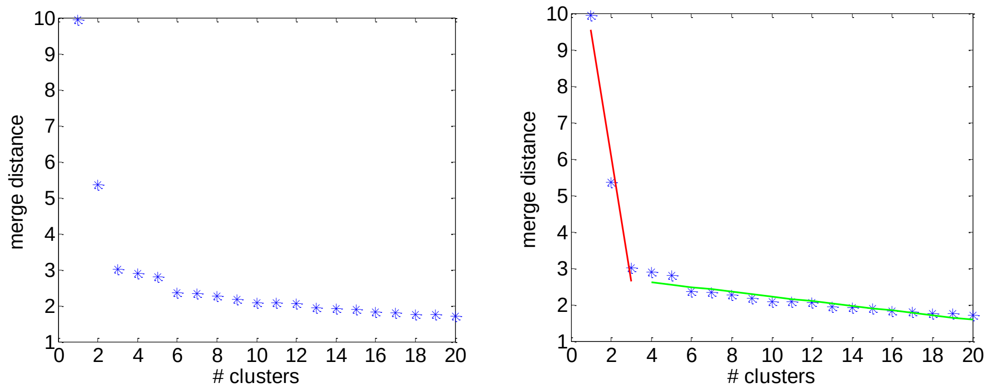


Figure 5-5. An example of inconsistency coefficient calculation.

The fundamental idea behind the inconsistency coefficient method is that nodes that join distinct clusters should have a high inconsistency coefficient, while nodes that join indistinct clusters should have a low inconsistency coefficient. Therefore, the clusters can be generated by breaking the nodes with inconsistency coefficient higher than a certain threshold. However, the threshold itself still has to be specified by the user. This threshold value is typically empirically assigned [77][78], but its optimal value can vary significantly for different applications or even different data sets. Moreover, the clustering result is extremely sensitive to this threshold value. Therefore, it is extremely difficult to develop a fully automatic clustering process based on the inconsistency coefficient method.

An alternative method to select the number of clusters that has gained popularity in recent years is the L-method [79]. The L-method is based on the fact that for many clustering algorithms, it is possible to plot an error curve where the x-axis is the number of clusters, and the y-axis is the evaluation function internally used by the clustering algorithm. For example, the evaluation function for k-means can be defined as the cost function (5.6). For hierarchical clustering, the evaluation function for having i clusters can be defined as the merge distance of the (N-i)-th merge, which corresponds to the clustering result by performing the first *N-i* merges in hierarchical clustering. While the error curve generally presents a decreasing trend with the number of clusters, it typically has a sharp transition at the most intuitive clustering of the data. For example, Figure 5-6 (a) plots the error curve of complete-link hierarchical clustering on the first synthetic data set described in Section 5.5, where synthetic wafers are created based on three distinct underlying signatures with random variation. It can be easily noticed by inspection that the transition point is at x=3. The L-method attempts to match human intuition by defining the following criterion: If we find the two consecutive lines that optimally fit the error curve, the transition point of the two lines is determined as the transition point of the error curve. For example, Figure 5-6 (b) accurately fits the error curve by two lines where one line fits the data with x=1-3, and another line fits the data with x=4-20. We can then determine that the data can be partitioned into 3 clusters.



(a) (b)

Figure 5-6. (a) The error curve of complete-link hierarchical clustering on a synthetic data set. (b) The optimal number of clusters can be found by fitting the curve with two lines.

In what follows, we will first describe the L-method in detail, and then discuss its limitation. Consider a number of clusters vs evaluation metric graph such as Figure 5-6 (a), with values on the x-axis up to x=B. We partition the data points at x=c into left and right sequences. The left sequence has points with x_{lc}=1...c, and the right sequence has points with x_{rc}=c+1… B, where c=2…B-1. Next, we find the optimal two lines that minimize the least squares error for fitting the left and right parts of the graph respectively:

$$\underset{a_{lc},b_{lc}}{\text{minimize}} \quad \left\| y_{lc} - a_{lc} - b_{lc} \cdot x_{lc} \right\|_2^2 \tag{5.13}$$

$$\underset{a_{rc},b_{rc}}{\text{minimize}} \quad \left\| y_{rc} - a_{rc} - b_{rc} \cdot x_{rc} \right\|_2^2 \tag{5.14}$$

where

$$x_{lc} = \begin{bmatrix} 1 & 2 & \cdots & c \end{bmatrix}^T \tag{5.15}$$

$$x_{rc} = \begin{bmatrix} c+1 & c+2 & \cdots & B-1 \end{bmatrix}^T , \tag{5.16}$$

y_{lc} and y_{rc} are the value of the evaluation metric at x_{lc} and x_{rc} respectively. Eq. (5.13) and (5.14) can be solved by least-squares fitting, yielding the following root mean squared error:

$$RMSE_{lc} = \frac{1}{\sqrt{c}} \left\| y_{lc} - a_{lc} - b_{lc} \cdot x_{lc} \right\|_2 \tag{5.17}$$

$$RMSE_{rc} = \frac{1}{\sqrt{B-c}} \left\| y_{rc} - a_{rc} - b_{rc} \cdot x_{rc} \right\|_2 \tag{5.18}$$

where a_{lc} and b_{lc} are from the solution of (5.13), and a_{rc} and b_{rc} are from the solution of (5.14). The total root mean squared error at $x=c$ is then defined as the weighted sum of the left and right errors:

$$RMSE_c = \frac{c}{B} RMSE_{lc} + \frac{B-c}{B} RMSE_{rc} \tag{5.19}$$

The optimal number of clusters is then defined by selecting the c value that minimizes the total error defined by (5.19):

$$\underset{c}{\arg\min} \quad RMSE_c \ . \tag{5.20}$$

In practice, the number of clusters is typically much smaller than the number of data points. Therefore, when directly applying the criterion (5.20) to the entire data set, a large number of values representing merges at extremely fine-grain clusterings (large values of x) are irrelevant and may inaccurate result due to highly imbalanced left and right sides. Therefore, the L-method is applied iteratively to the data points. Starting from solving (5.20) on the entire data set, each iteration reduces the number of data points included in the next iteration to:

$$B_{next} = \max(2 \cdot c, 20). \tag{5.21}$$

where c is the optimal number of clusters determined in the current iteration, and $2 \cdot c$ is a number to keep the left and right side balanced. The total number of data points is not permitted to drop below 20, which is an empirical number proposed in [79] to keep a reasonable number of points to fit the lines. The L-method stops when the number of data points included has converged.

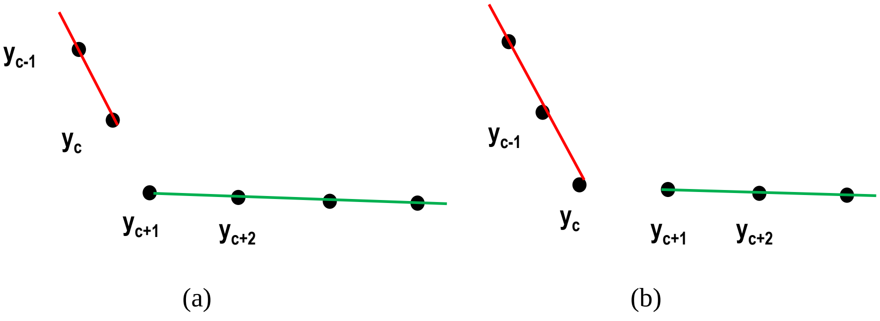


Figure 5-7. An synthetic example where the error can be minimized by either (a) c=2. (b) c=3.

While the L-method attempts to match human intuition in finding the transition point of the error curve, we noticed that its definition of the transition point is counter-intuitive. Specifically, the transition point can often be reasonably fit by either the left or the right curve. To explain this idea, we constructed a synthetic example in Figure 5-7 where 6 points can be exactly fit with two lines. It is obvious that the optimal solution with human inspection is $c = 3$. However, according to the L-method, two possible values of c would result in a minimum error of 0. Figure 5-7 (a) and Figure 5-7 (b) show lines fitted by the L-method when setting $c = 2$ and $c = 3$ respectively, from which it can be clearly seen that both of these solutions are valid solutions of the L-method. In practice, the points cannot be perfectly fit by two lines and the choice of $c = 2$ or 3 by the L-method is arbitrary: small changes in the error curve may cause the number of clusters to shift from one solution to another.

Based on the above observation, we propose to add a post-processing step to the L-method to more accurately determine the number of clusters. This is done by detecting if a sharper transition occurs at the currently selected point c or the next point c+1. The number of clusters is added by one, if the next point causes a sharper transition in the error curve. Specifically, we propose to use the following quantity to represent the transition rate at $x = c$:

$$s_c = \left(\log\left(y_{c+1}\right) - \log\left(y_c\right)\right) - \left(\log\left(y_c\right) - \log\left(y_{c-1}\right)\right) = \log\left(\frac{y_{c+1} \cdot y_{c-1}}{y_c^2}\right), \qquad (5.22)$$

where $y_{c\text{-}1}$, y_c and y_{c+1} are the value of the evaluation metric at c-1, c and c+1 respectively. Similarly, we define the transition rate at $x = c$+1 by:

$$s_{c+1} = \left(\log\left(y_{c+2}\right) - \log\left(y_{c+1}\right)\right) - \left(\log\left(y_{c+1}\right) - \log\left(y_{c}\right)\right) = \log\left(\frac{y_{c+2} \cdot y_{c}}{y_{c+1}^{2}}\right) \tag{5.23}$$

where y_{c+2} is the value of the evaluation metric at c+2. Eq. (5.22) and (5.23) are essentially the second-order difference of the series log(y). The second-order difference is applied, because we would like to compare the difference from the previous point to the current point with the difference from current point to the next point. A large second-order difference means a significant difference between these two values, which indicates a strong transition in the trend of the error curve. We take the logarithm for the evaluation metric y, because comparing the ratio between two consecutive points is more intuitive than comparing the absolute difference. For example, in Figure 5-6 (b), y_2-y_1 is significantly different from y_3-y_2, yet c = 3 remains the intuitive solution for the transition point in Figure 5-6 (b) by human inspection. We summarize the main steps of the modified L-method for cluster number determination in Algorithm 4:

Algorithm 4: Modified L-method for cluster number selection

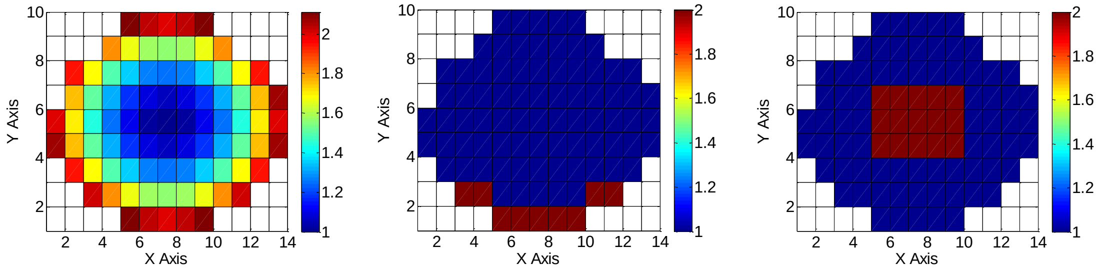
1. Start from a vector $y^{B\times 1}$ representing the value of the evaluation metric when the number of clusters is 1, 2, …, B.
2. Find the optimal number of clusters c according to the criterion (5.20).
3. Compare s_c and s_{c+1} defined by (5.22) and (5.23) respectively. If $s_{c+1} > s_c$, $c = c + 1$.
4. Calculate the number of data points included in the next step B_{next} by (5.21).
5. If $B_{next} = B$, stop. Otherwise, $B = B_{next}$ and go to Step 1.

Note that the applicability of Algorithm 4 is not restricted to hierarchical clustering. Instead, it can be applied to select the number of clusters for any clustering algorithm for which a number of clusters vs evaluation metric graph can be generated. For example, it can be applied to the k-means clustering algorithm, where the evaluation metric is the cost function (5.6). We will show in Section 5.5 that accurate clustering results can also be generated by k-means with Algorithm 4 when the data does not contain any abnormal wafers.

5.5 Numerical Results

In the previous sub-sections, we have proposed a wafer spatial signature clustering method which mainly consists of three components: robust feature extraction by (5.4), complete-link hierarchical clustering described in Section 5.3, and the modified L-method for cluster number selection described by Algorithm 4. In this sub-section, we will use several synthetic examples and silicon measurement data sets to demonstrate the effectiveness of the proposed method.

5.5.1 Results on Synthetic Data



(a) (b) (c)

Figure 5-8. Systematic variation of three different clusters in the synthetic example.

We first consider several examples where the data set contains three clusters with distinct spatial signatures shown in Figure 5-8. Figure 5-8 (a) is created by the following quadratic function:

$$s(x, y) = 1 + x^2 + y^2 \tag{5.24}$$

where x and y are coordinates on the wafer with range normalized to [-1 1]. Figure 5-8 (b) contains edge effect at the bottom of the wafer and it is created by:

$$s(x, y) = \begin{cases} 2 & (x, y) \in E \\ 1 & otherwise \end{cases} \tag{5.25}$$

where E is the bottom edge region of the wafer. Figure 5-8 (c) contains center effect and it is created by:

$$s(x, y) = \begin{cases} 2 & (x, y) \in C \\ 1 & otherwise \end{cases} \tag{5.26}$$

where C is the center region of the wafer. For each spatial signature in Figure 5-8, we generate 20 synthetic wafers by randomly adding 10% random variation and Figure 5-9 (a)-(c) show randomly selected three wafers from each signature. This forms a synthetic data set with 60 wafers in total.

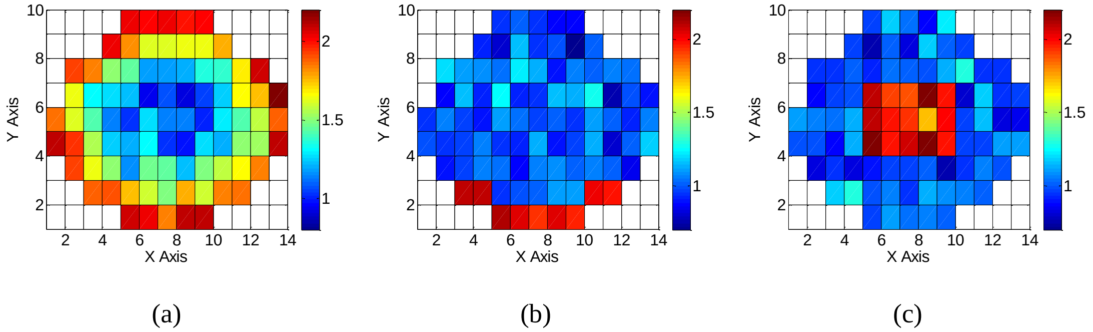
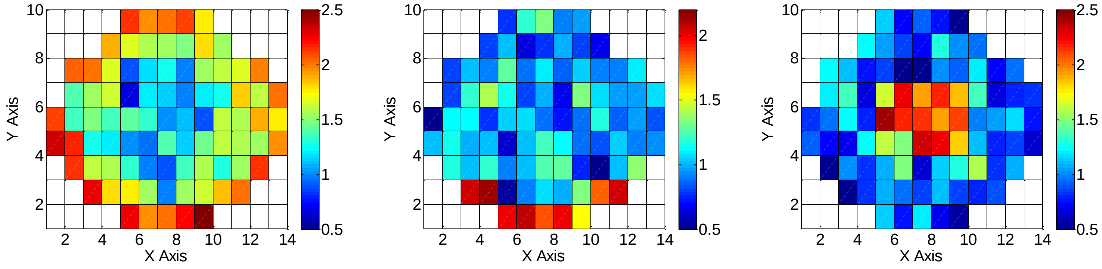


Figure 5-9. Three synthetic wafers belonging to different clusters.

We apply the proposed method to cluster these 60 wafers and it produces the correct clustering result: the 60 wafers are partitioned into 3 clusters, each containing 20 wafers that correspond to the systematic variation in Figure 5-8 (a)-(c) respectively. For comparison purposes, we implement two alternative methods for the clustering algorithm in the wafer spatial signature clustering flow shown in Figure 5-2, the k-means and single-link hierarchical clustering algorithms. In this example, both k-means and single-link hierarchical clustering also produce the correct clustering result. However, accurate clustering results cannot be easily achieved if the number of clusters is selected by the traditional inconsistency coefficient method, which is extremely sensitive to the user-defined inconsistency coefficient. We find that setting inconsistency coefficient to 1.16 will produce only one cluster, while setting it to 1.15 will produce 5 clusters, which unnecessarily divides the 3 natural clusters. Setting the coefficient below 1.15 will result in even larger number of clusters, and no value between 1.15 and 1.16 generates the correct clustering result.

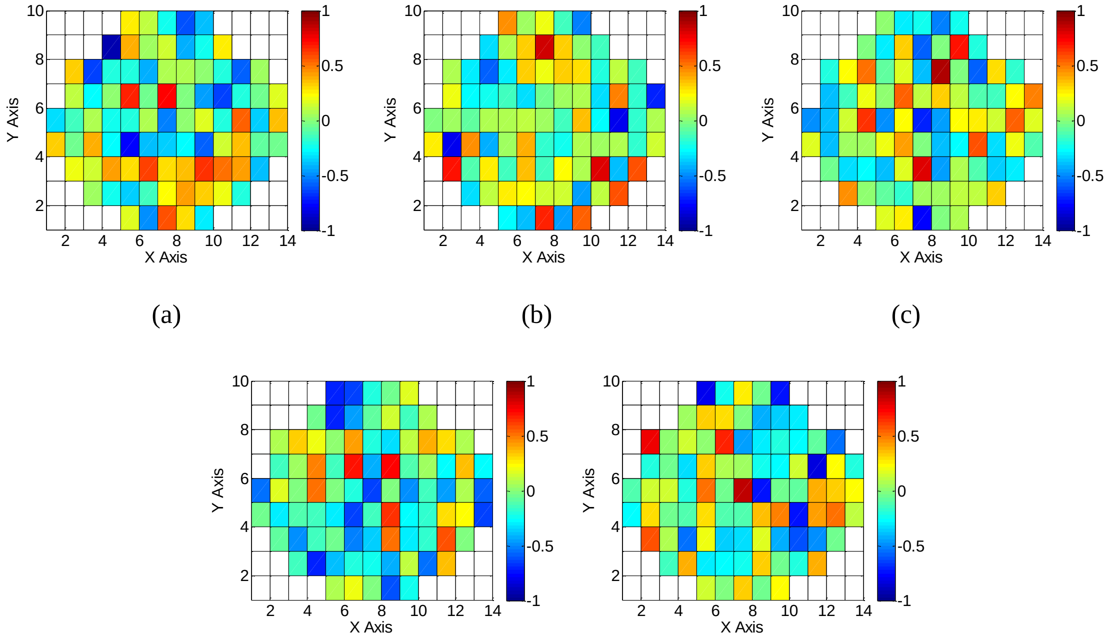
To further examine the effectiveness of the proposed method against random variation, we further increase the percentage of random variation in the data set to 30%. Figure 5-10 (a)-(c) show three randomly selected wafers from each signature, from which it can be seen that the spatial signature is much less clear compared to Figure 5-9. In this example, the proposed method, as well as applying k-means and single-link hierarchical clustering still correctly find the three clusters. This result shows the robustness of the proposed flow against random variation. In this example, accurate result can be obtained by the inconsistency coefficient method when the inconsistency coefficient is set to 1.15, but setting it to 1.16 and 1.14 will result in 1 cluster and 7 clusters respectively, which shows that this method is extremely sensitive to the inconsistency coefficient.



(a) (b) (c)

Figure 5-10. Three synthetic wafers with large random variation belonging to different clusters.

To further examine the effectiveness of the proposed methodology against outliers, we create 5 abnormal wafers with no spatially correlated variation and very large random variation in Figure 5-11. The variance of the random variation is selected to match the variance of the wafer-level variation of the synthetic wafers with 10% random variation, which was shown in Figure 5-9. These 5 abnormal wafers are then added to the data set, resulting in 65 wafers in total.



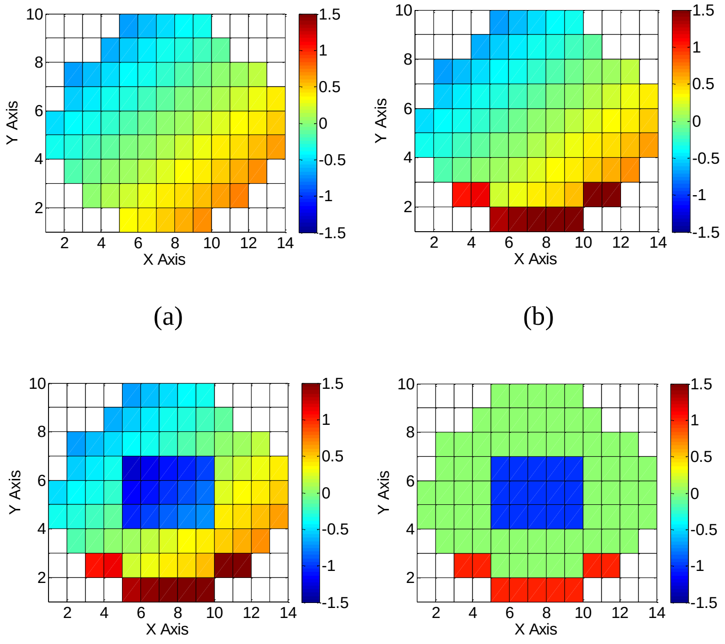
(a) (b) (c)

(d) (e)

Figure 5-11. Five abnormal wafers created by very large random variation.

The proposed method correctly clusters these 65 wafers: 4 clusters are created, where 20 wafers from each spatial signature form a cluster respectively, and 5 abnormal wafers are combined into one cluster. In this example, applying k-means will only detect 2 clusters: wafers with quadratic and edge signatures in Figure 5-8 (a)-(b) are merged into one cluster, and wafers with the center pattern are merged with outlier wafers to form another cluster. K-means fails to produce accurate results because of its lack of ability to handle abnormal wafers. Single-link hierarchical clustering generates 8 clusters, where 20 wafers from each spatial signature form a cluster respectively, and 5 abnormal wafers each form an individual cluster. The three main clusters are correctly identified, and the outliers are detected and separated from the three main clusters. This result also correctly identifies the clusters and abnormal wafers, which shows that the hierarchical clustering approach is robust to abnormal wafers. The complete-link hierarchical clustering result is more desirable than the single-link hierarchical clustering, because it is more concise and accurately detects that the abnormal wafers are in fact from the same distribution. The abnormal wafers can be clustered by complete-link hierarchical clustering for two reasons: First, although the 5 abnormal wafers in Figure 5-11 (a)-(e) look quite different, their random variation is significantly reduced by the robust feature extraction process in Section 5.2, so that they are much more similar in the feature space. Second, all of them are not far away from the spatial signature with no variation, so that they are connected from a complete-link point of view. In this example, the correct clusters again cannot be identified if the inconsistency coefficient method is applied. We find that setting inconsistency coefficient to 1.15 or above will produce only one cluster, while setting it to 1.14 will produce 12 clusters. No value between 1.14 and 1.15 generates the correct clustering result.

We further evaluate the effectiveness of the algorithms by adding 5 abnormal wafers to the data set with 30% random variation. The variance of the abnormal wafers is adjusted to match the variance of the wafer-level variation of other synthetic wafers. The results are very similar to the case with small random variation. K-means still only detects 2 clusters: wafers with quadratic and edge patterns are merged with outliers into one cluster, and wafers with the center pattern form another cluster. The clusters formed by single-link hierarchical clustering and the proposed method are exactly the same as the previous example. These results show that hierarchical clustering is in general robust to abnormal wafers, with complete-link hierarchical clustering produces slightly better results because it agrees better with our intuition for this application. The correct clusters cannot be identified if the inconsistency coefficient method is applied. We find that setting inconsistency coefficient to 1.15 will produce 3 clusters where the outliers are not detected, while setting it to 1.14 will produce 7 clusters where the wafers with the center pattern form 4 clusters. No value between 1.14 and 1.15 generates the correct clustering result.



(a) (b)

(c) (d)

Figure 5-12.Systematic variation of four different clusters in the synthetic example.

We further consider several other examples where the data set contains four clusters where the spatial signature is not as clearly distinct as the previous examples. The systematic variation signatures of these four clusters are shown in Figure 5-12. Figure 5-12 (a) is created by the following linear function:

$$s_l(x,y) = -0.5x - 0.5y \tag{5.27}$$

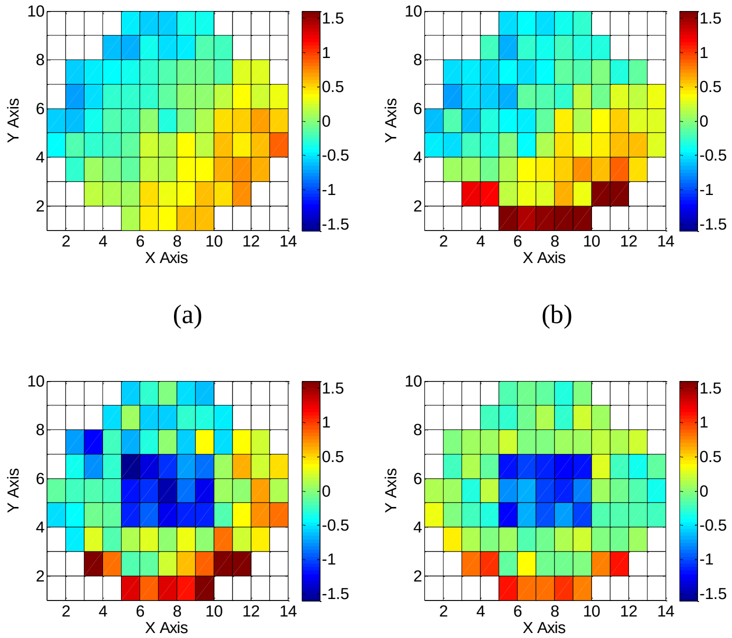
where *x* and *y* are coordinates on the wafer with range normalized to [-1 1]. Figure 5-12 (b) further adds the following edge effect function at the bottom of the wafer to the spatial signature in Figure 5-12 (a):

$$s_e(x,y) = \begin{cases} 1 & (x,y) \in E \\ 0 & otherwise \end{cases} \tag{5.28}$$

where *E* is the bottom edge region of the wafer. Figure 5-12 (c) further adds the following center effect function to Figure 5-12 (b):

$$s_c(x,y) = \begin{cases} -1 & (x,y) \in C \\ 0 & otherwise \end{cases} \tag{5.29}$$

where *C* is the center region of the wafer. Figure 5-12 (d) subtracts the linear function from Figure 5-12 (c), so that the wafer only contains center and edge effects. For each spatial signature in Figure 5-12, we generate 20 synthetic wafers by randomly adding 10% random variation and Figure 5-13 (a)-(d) show randomly selected four wafers from each of the four different signatures. This forms a synthetic data set with 80 wafers in total.



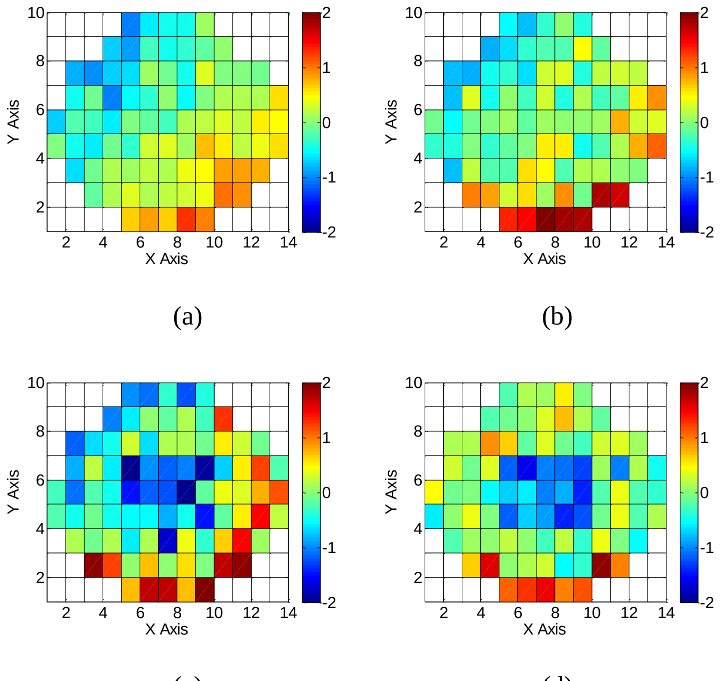
(a) (b)

(c) (d)

Figure 5-13. Four synthetic wafers belonging to different clusters.

The proposed method accurately partitions these 80 wafers into 4 clusters, each containing 20 wafers that correspond to the systematic variation in Figure 5-12 (a)-(d) respectively. In this example, the same correct result can also be obtained if k-means or single-link hierarchical clustering are applied as the clustering algorithm. On the other hand, no correct result can be obtained if the inconsistency coefficient method is applied to select the number of clusters. We find that setting inconsistency coefficient to 1.16 or above will produce only one cluster, while setting it to 1.15 will produce 7 clusters. No value between 1.15 and 1.16 generates the correct clustering result.

We further increase the percentage of random variation in the data set to 30%. Figure 5-14 (a)-(d) show randomly selected four wafers from each of the four different signatures, from which it can be seen that the difference of the spatial signature is much less clear compared to Figure 5-13. In this example, the proposed method again correctly finds the four clusters, and the same result can be obtained if k-means is applied as the clustering algorithm. However, single-link hierarchical clustering only detects 2 clusters: wafers with signatures in Figure 5-12 (a)-(c) are merged into one cluster, and wafers with the signature in Figure 5-12 (d) form another cluster. This is because the signatures in Figure 5-12 are not significantly different; with large random variation, it is possible for a small number wafers from two different spatial signatures to become similar. Therefore, single-link hierarchical clustering may merge these wafers into the same cluster because they are "connected". Complete-link hierarchical clustering requires all wafers within the same cluster to be connected, and is therefore less sensitive to the aforementioned problem due to large random variation. No correct result can be obtained if the inconsistency coefficient method is applied to select the number of clusters. We find that setting inconsistency coefficient to 1.16 or above will produce only one cluster, while setting it to 1.15 will produce 10 clusters. No value between 1.15 and 1.16 generates the correct clustering result.



(a)

(b)

(c)

(d)

Figure 5-14. Four synthetic wafers with large random variation belonging to different clusters.

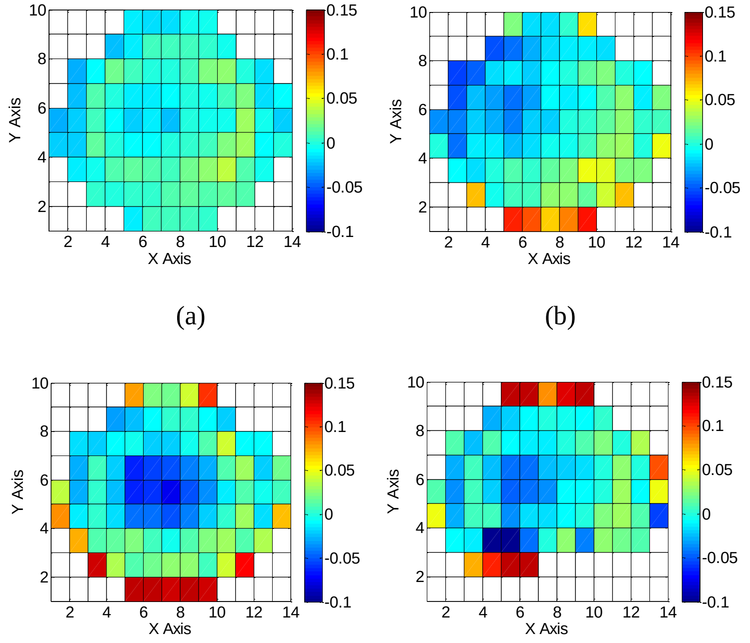
To further examine the effectiveness of the proposed method against outliers, similar to the previous experiments, we add 5 abnormal wafers to the data sets with 10% and 30% random variation respectively, resulting in 85 wafers in total. The variance of the abnormal wafers is selected to match the variance of the wafer-level variation of other synthetic wafers. We observe similar results to the previous experiments with abnormal wafers: reasonably accurate results can be provided by the proposed method, or using single-link hierarchical clustering as the clustering algorithm, but k-means fails to provide accurate clustering results. Namely, in the 10% random variation example, applying k-means will detect 4 clusters, where the abnormal wafers are merged into the cluster with the systematic spatial signature in Figure 5-12 (a), and the other 3 clusters are detected correctly. Both single-link and the proposed method detect 8 clusters where the 4 main clusters are accurately detected and the abnormal wafers are further divided into 4 clusters. In the 30% random variation example, applying k-means will only detect 3 clusters, where the abnormal wafers are merged into the cluster with the systematic spatial signature in Figure 5-12 (a), and the wafers with the systematic spatial signature in Figure 5-12 (b)-(c) are further merged into one cluster. Single-link hierarchical clustering detects 8 clusters where the 4 main clusters are accurately detected and the abnormal wafers are further divided into 4 clusters. The proposed method detects 5 clusters where the 4 main clusters are accurately detected and the abnormal wafers are merged into one cluster, which is more accurate than the single-link hierarchical clustering. No correct result can be obtained if the inconsistency coefficient method is applied to select the number of clusters. We find that setting inconsistency coefficient to 1.16 or above will produce only one cluster in both cases, while setting it to 1.15 will produce 7 and 10 clusters for the 10% and 30% random variation example respectively. No value between 1.15 and 1.16 generates the correct clustering result.

In summary, from the synthetic examples, we observed that the proposed wafer spatial signature clustering flow is capable of producing accurate clusters with either k-means or hierarchical clustering as the clustering algorithm, when the random variation is small or the different spatial signatures are relatively distinct, and the data does not contain abnormal wafers. When abnormal wafers exist, applying k-means would fail to produce accurate clusters because of its inherent lack of ability to handle abnormal wafers. Applying single-link hierarchical clustering may fail when there exists large random variation. In all examples, the proposed method with complete-link hierarchical clustering accurately detects the clusters in the presence of large random variation and abnormal wafers. Such accurate result cannot be obtained if the traditional inconsistency coefficient method is applied to select the number of clusters.

5.5.2 Results on Silicon Measurement Data

From the previous experiments, we observe that the proposed wafer spatial signature clustering method accurately detects the clusters in various synthetic data examples. We will further demonstrate the effectiveness of our method and perform comparison on several sets of silicon measurement data.

We first consider I_{dsat} measurements obtained by single NMOS test structures from 69 wafers. We first apply the proposed method to cluster these wafers, and then compare the clustering result with various other options. In this example, complete-link hierarchical clustering generates 4 clusters, and the number of wafers for these 4 clusters is 34, 23, 9 and 3 respectively. Figure 5-15 shows the averaged wafer map for these 4 clusters, where it can be seen that these clusters contain distinct spatial signatures. Wafers in cluster 1 do not have significant spatially correlated variation. Since this cluster also contains the largest number of wafers, it can be considered as the baseline cluster. Wafers in cluster 2 have strong edge effect and an increasing trend from top-left to bottom-right corner; wafers in cluster 3 have strong edge and center effect; wafers in cluster 4 have a large number of missing measurements in the bottom of the wafer.



(a)

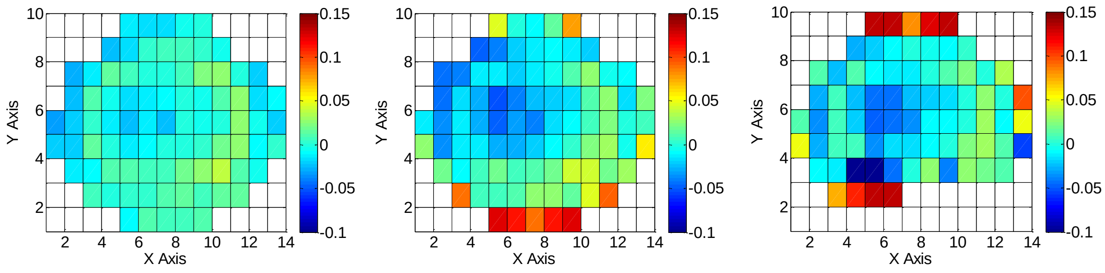
(b)

(c)

(d)

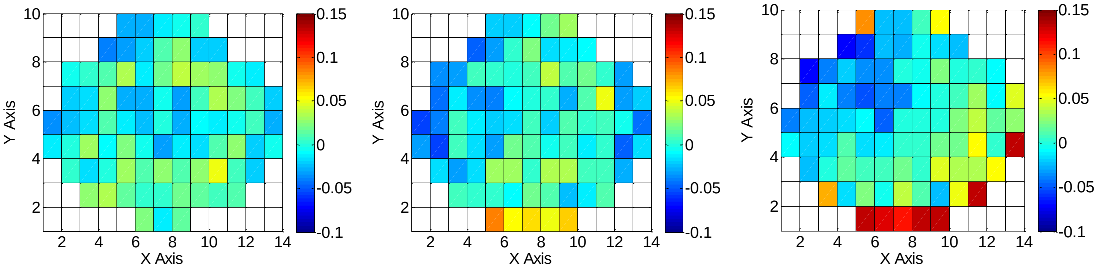
Figure 5-15. Averaged wafer map (normalized) of four different clusters detected by the proposed method for NMOS I_{dsat} measurement data set 1.

In this example, these different signatures cannot be completely detected by applying k-means or single-link hierarchical clustering as the clustering algorithm. K-means only detects 3 clusters and the number of wafers for these 3 clusters is 38, 28 and 3 respectively. Figure 5-16 shows the averaged wafer map for these clusters. Roughly speaking, these three clusters match cluster 1, cluster 2-3 and cluster 4 in the complete-link hierarchical clustering result. Therefore, k-means fails to detect the different spatial signatures presented by cluster 2 and 3 in Figure 5-15. If single-link hierarchical clustering is applied, cluster 1 and 2 in Figure 5-16 are further merged into one cluster, so that it fails to distinguish the spatial signature between cluster 1, 2 and 3 in Figure 5-16. The fundamental reason for the failure of single-link hierarchical clustering is that there does not exist a clear boundary between clusters. To intuitively explain this, we plot the wafer maps for three different wafers in this data set in Figure 5-17. It can be seen that while there is a clear difference in the spatial signature between the wafers in Figure 5-17 (a) and Figure 5-17 (c), there exist wafers such as the wafer in Figure 5-17 (b) whose spatial signature is similar to both. This may happen because of, for example, process drift. In this case, single-link hierarchical clustering will merge Figure 5-17 (a) and Figure 5-17 (c) into the same cluster because they are connected by Figure 5-17 (b). On the other hand, complete-link hierarchical clustering requires all wafers in the same cluster to be similar and therefore does not suffer from this problem. If the inconsistency coefficient method is applied to select the number of clusters, setting inconsistency coefficient to 1.16 or above will produce only one cluster, while setting it to 1.15 will produce 5 clusters, where cluster 2 in Figure 5-15 will be split into two clusters with no significant difference in spatial signature. More clusters will be unnecessarily created if the inconsistency coefficient is set to 1.14 or below.



(a) (b) (c)

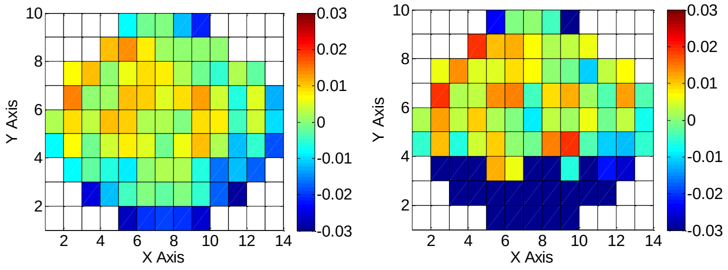
Figure 5-16. Averaged wafer map (normalized) of three different clusters detected by k-means for NMOS I_{dsat} measurement data set 1.



(a) (b) (c)

Figure 5-17. Three different wafers in NMOS I_{dsat} measurement data set 1.

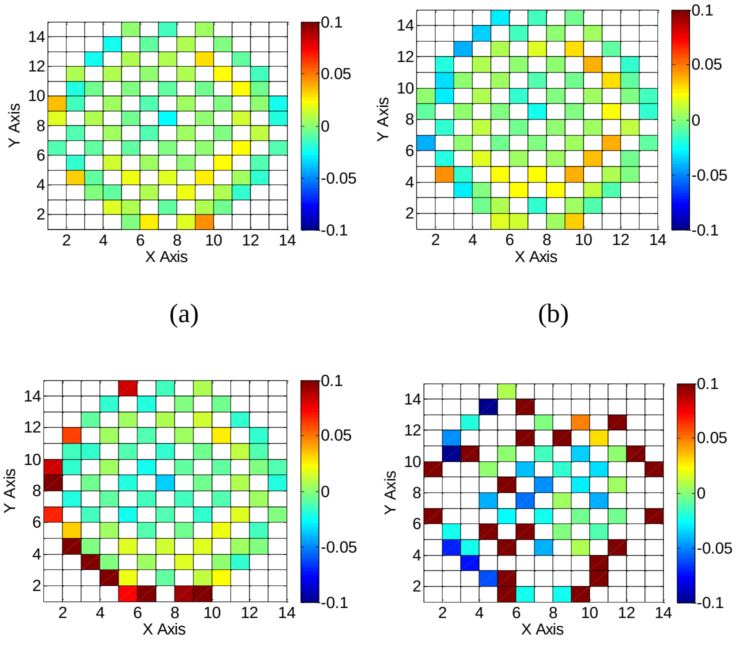
Next, we consider I_{dsat} measurements obtained by single PMOS test structures from the same 69 wafers. In this example, the proposed method generates 2 clusters, and the number of wafers for these 2 clusters is 64 and 5 respectively. Figure 5-18 shows the averaged wafer map for these 2 clusters, where it can be seen that these clusters indeed contain distinct spatial signatures. Cluster 1 can be considered as a baseline cluster and cluster 2 contains several abnormal wafers. In this example, applying k-means or single-link hierarchical clustering produces consistent results with complete-link hierarchical clustering and no significant difference in spatial signature can be further found by manually inspecting cluster 1. Therefore, the proposed flow produces accurate clustering result regardless of the clustering algorithm employed in this example. If the inconsistency coefficient method is applied to select the number of clusters, setting inconsistency coefficient to 1.16 or above will produce only one cluster, while setting it to 1.15 will produce 5 clusters, where cluster 1 in Figure 5-18 will be split into four clusters with no significant difference in spatial signature. More clusters will be unnecessarily created if the inconsistency coefficient is set to 1.14 or below.



(a) (b)

Figure 5-18. Averaged wafer map (normalized) of two different clusters detected by the proposed method for PMOS I_{dsat} measurement data set 1.

We further consider I$_{dsat}$ measurements obtained by single NMOS test structures from another 82 wafers. In this data set, not all test structures are measured; instead, they are sampled in a “checkerboard” style to reduce the test cost. In this example, the proposed method generates 4 clusters, and the number of wafers for these 4 clusters is 43, 18, 20 and 1 respectively. Figure 5-19 shows the averaged wafer map for these 4 clusters. Inspecting Figure 5-19, it can be seen that although cluster 2 presents larger spatially correlated variation compared to cluster 1, overall speaking the difference in spatial signature between these two clusters is not significant. Therefore, they can be simply merged into one cluster and considered as the baseline cluster after simple manual inspection on Figure 5-19 (a) and Figure 5-19 (b). Note that although the clustering result does not exactly match human intuition in this example, such inspection and merging process requires very little human effort. Cluster 3 and cluster 4 detected by the proposed method indeed contain completely different spatial signature compared to cluster 1-2: wafers in cluster 3 have significant edge effect at the bottom-left portion of the wafer, and cluster 4 contains an abnormal wafer with completely different signature compared to any other wafer.



(a)

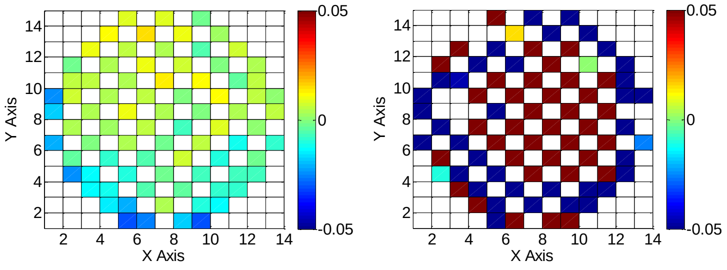
(b)

(c)

(d)

Figure 5-19. Averaged wafer map (normalized) of four different clusters detected by the proposed method for NMOS I$_{dsat}$ measurement data set 2.

In this example, applying k-means or single-link hierarchical clustering as the clustering algorithm would fail to detect all the distinct signatures in Figure 5-19. Namely, k-means generates 2 clusters where clusters 1, 2 and 4 in Figure 5-19 are merged into one cluster and cluster 3 in Figure 5-19 forms another cluster. While the k-means method correctly merges clusters 1 and 2 in Figure 5-19, the abnormal wafer is also merged into the baseline cluster and cannot be detected by simple inspection. Single-link hierarchical clustering generates 2 clusters where clusters 1, 2 and 3 in Figure 5-19 are merged into one cluster and cluster 4 in Figure 5-19 forms another cluster. Therefore, it fails to detect the wafers with edge effect. Therefore, the proposed method with complete-link hierarchical clustering provides the best accuracy. If the inconsistency coefficient method is applied to select the number of clusters, setting inconsistency coefficient to 1.16 or above will produce only one cluster, while setting it to 1.15 will produce 9 clusters, where clusters 1-3 in Figure 5-19 will all be further split into multiple clusters with no significant difference in spatial signature. More clusters will be unnecessarily created if the inconsistency coefficient is set to 1.14 or below.



(a) (b)

Figure 5-20. Averaged wafer map (normalized) of two different clusters detected by the proposed method for PMOS I_{dsat} measurement data set 2.

We finally consider I_{dsat} measurements obtained by single PMOS test structures from the same 82 wafers. In this example, the proposed method generates 2 clusters, and the number of wafers for these 2 clusters is 81 and 1 respectively. Figure 5-20 shows the averaged wafer map for these 2 clusters, where cluster 1 can be considered as the baseline cluster and cluster 2 contains an abnormal wafer. No significant difference in spatial signature can be further found by manually inspecting cluster 1. In this example, applying single-link hierarchical clustering produces the same results as the proposed method. K-means further splits cluster 1 into two clusters with no significant difference in spatial signature, but these two clusters can be merged after simple manual inspection, similar to the previous example. If the inconsistency coefficient method is applied to select the number of clusters, setting inconsistency coefficient to 1.16 or above will produce only one cluster, while setting it to 1.15 will produce 7 clusters, where cluster 1 in Figure 5-20 will be split into six clusters with no significant difference in spatial signature. More clusters will be unnecessarily created if the inconsistency coefficient is set to 1.14 or below.

5.6 Summary

Wafer spatial signature clustering provides important insight to help process engineers prioritize the yield improvement goals. In this chapter, we propose to solve the wafer spatial signature clustering problem based on a three-step process: first, the spatial signatures of wafers are automatically captured by a small number of features based on robust sparse regression with the DCT dictionary; second, complete-link hierarchical clustering is performed on the features; finally, a modified L-method is performed on the hierarchical clustering result to select the clusters. The effectiveness of the proposed method is demonstrated by a number of synthetic and silicon data sets. One of the key decisions in the proposed method is to adopt a complete-link hierarchical clustering algorithm and its superiority over k-means and single-link hierarchical clustering is demonstrated by several synthetic and silicon examples. Moreover, numerical results demonstrate that the accurate clustering result cannot be obtained if the traditional inconsistency coefficient method is applied to select the number of clusters.

Chapter 6

Thesis Summary & Future Work

6.1 Summary

With the continued scaling of CMOS technology, it becomes increasingly difficult to keep process variations under control. At the same time, process engineers are facing increasingly stringent time-to-market requirements for modern products. Therefore, rapidly improving the yield of today's complicated manufacturing process is a key challenge to ensure profitability for the IC industry.

In this thesis, we propose accurate and efficient modeling techniques for spatial variation, which is becoming increasing important in the advanced technology nodes. Based on our spatial model, we propose accurate and efficient techniques for two applications that help process engineers identify the important yield-limiting factors in the manufacturing process, so that process engineers can prioritize their yield improvement efforts. One of these applications is *variation decomposition*, where the overall variation is decomposed into multiple different components, each corresponding to a different subset of variation sources. This allows process engineers to narrow down the main sources of variation for wafers with similar patterns, especially at the process development and product yield ramp stages.

An important problem in variation decomposition is to accurately model and extract the wafer-level and within-die spatially correlated variation, and separate them from random variations. Towards this goal, we first develop a physical basis function dictionary based on our study of several common physical variation sources, which captures more spatially correlated systematic variation sources than the traditional quadratic modeling approach, and then further proposes the DCT dictionary to discover spatially correlated systematic patterns not modeled by the physical dictionary. Moreover, we propose to apply sparse regression to significantly reduce the over-fitting problem related to applying a large basis function dictionary.

Substantial error can be introduced to variation decomposition if outliers are not appropriately detected and removed. We further extend the sparse regression algorithm to a robust sparse regression algorithm, which provides superior accuracy compared to the traditional IQR method for outlier detection. By solving robust sparse regression, basis functions will be accurately selected in the presence of outliers, and outliers will be automatically detected and removed. Experiments on synthetic and silicon measurement data demonstrate the effectiveness of the proposed variation decomposition methodology based on robust sparse regression.

The computational cost for sparse regression with DCT basis functions can become extremely large for problems with large size, which limits the applicability of the variation decomposition methodology based on sparse regression. We propose several efficient methods to make the computational cost of sparse regression tractable for large-scale problems. The key idea of these methods is to utilize fast DCT/IDCT computation to speed up the matrix-vector product computation. From the experimental results on a large problem with contact resistance measurement data, we observe nearly 200× speedup compared to the traditional direct implementation.

The second application we target at is the wafer spatial signature clustering problem. The goal is to automatically partition a large number of wafers into different groups, in which different groups exhibit different spatial signatures. The results would help process engineers find important factors that prevent the process from stably maintaining a high yield across different lots and wafers. Our proposed method contains three key components: first, a robust feature extraction method is developed to automatically capture the spatial signatures of wafers by a small number of features by re-using the robust sparse regression technique developed for variation decomposition; second, a complete-link hierarchical clustering algorithm is selected to perform clustering on the features; finally, a modified L-method is developed to select the number of clusters from the hierarchical clustering result. The effectiveness of the proposed method is demonstrated by a number of synthetic and silicon data sets.

6.2 Future Work

There are multiple directions that can be explored to extend this work to further benefit the manufacturing, design and testing community.

First, more efforts can be made to further automate the process of identifying important variation sources in the spatially correlated systematic variation. This can be done via two avenues: first, we can encode more of our knowledge in the form of physical basis functions and add them to the physical basis functions dictionary. By using the physical dictionary instead of having to rely on the DCT dictionary, we can gain more insight into the physical variation sources. Second, as was mentioned in Section 1.2.1, in variation characterization there can exist a number of test structures dedicated to monitoring a single parameter (e.g. gate length, transistor threshold voltage, etc.). If we also accurately extract the spatially correlated variations related to the single-parameter test structures, and then develop an accurate method to automatically find those that are most strongly related to the spatially correlated variations of product representative test structures, it would provide important information for identifying the physical sources of variation.

Second, the current spatial variation modeling technique based on sparse regression focuses on capturing the spatially correlated systematic variation at the within-wafer and within-die level. It may be possible to extend the sparse regression idea to capture some systematic variation sources at the wafer-to-wafer level, such as process drift, chamber mismatch, etc. Similar to the idea of spatial variation modeling, it may be possible to represent these possible systematic variation sources as a dictionary containing a large number of basis functions, and then apply sparse regression to automatically select the variation sources for a particular process. This allows process engineers to gain more insight into the wafer-to-wafer variation, similar to the wafer-level and within-die variation in the current method.

Third, the over-fitting issue in spatial variation modeling is an extremely important problem that deserves further investigation. In this thesis, we have used sparse regression with cross-validation to significantly reduce over-fitting, but from the experiments, it can be seen that over-fitting is not completely removed. A possible research direction is to investigate alternatives to cross-validation. For example, various information criteria can be applied to replace cross-validation, such as AIC and BIC [24]. Furthermore, the modified L-method may be applied to the trade-off curve between fitting error and number of basis functions to detect the number of basis functions.

Fourth, the wafer spatial signature clustering application we study in this thesis can only be applied to off-line analysis, because the amount of time it takes to build the spatial model and perform clustering makes it impractical to be performed on-line. In practice, if we can detect abnormal wafer spatial signatures on-line during statistical process control (SPC), corrective actions can be taken more promptly to prevent yield loss. A possible way to achieve this is to model the common spatial signatures in advance, and check the spatial signature of each wafer against the common spatial models.

Fifth, in this thesis we have stated variation decomposition and wafer spatial signature clustering as two independent applications mainly applicable to different stages of the process lifecycle, but it may be possible for these two tools to work together to enable more efficient detection of variation sources. For example, in volume production data, after performing wafer spatial signature clustering, it is possible to apply variation decomposition to a particular cluster, in order to find out the primary variation sources related to this cluster. On the other hand, having a large lot-to-lot component in variation decomposition may indicate the existence of clusters. There may also exist other scenarios where these two tools can work together.

Sixth, an important direction to study is how the spatial variation modeling techniques in this thesis can benefit the design. It may be possible to integrate the systematic spatial model extracted by the variation decomposition process into the device model, so that as long as the spatial location of a transistor is know, a large number of its variation can be determined by the spatial model so that they no longer need to be treated as random variables. They can greatly reduce the margin the designer has to leave for random variation. It would be an interesting direction to further investigate how various CAD tools for design such as statistical library characterization, statistical static timing analysis and statistical circuit optimization can be adapted to consider such spatial model.

Finally, another direction that can be explored is to apply the wafer-level spatial variation modeling technique to reduce the testing cost. Our preliminary research has indicated that for many performance measurements such as flush delay (the time for a transition to traverse the entire scan chain) and leakage current, it is possible for the wafer-level variation to be dominated by spatially correlated variation. In this case, the model for spatially correlated variation can be accurately extracted by sampling a small number of dies on the wafers, and the performance for other dies can be predicted based on the model with little error [18][100][101]. A test cost reduction methodology was recently proposed in [102] based on this idea, where it has been applied to 175 wafers with more than one million chips, and each chip was tested for 51 performance metrics. The experimental results demonstrate that the test cost of 39 out of these 51 performance metrics using this idea, resulting in 2.36× reduction in test time with negligible increase in escape rate and yield loss. However, additional data needs to be analyzed to fully understand the trade-off between reduced testing time and increased escape rate/yield loss by this method. Furthermore, how this method can be combined with other test cost reduction approaches (e.g. exploring the correlation between test items) is an interesting topic for future research.

Bibliography

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

[2] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, “Parameter variations and impact on circuits and microarchitecture,” *IEEE Design Automation Conference*, pp. 338-342, 2003.

[3] M. Pelgrom, A. Duinmaijer, and A. Welbers, “Matching properties of MOS transistors,” *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433-1439, Oct. 1989.

[4] F. Liu, “A general framework for spatial correlation modeling in VLSI design,” *IEEE Design Automation Conference*, pp. 817-822, 2007.

[5] J. Xiong, V. Zolotov, and L. He, “Robust extraction of spatial correlation,” *IEEE Trans. Computer-Aided Design*, Vol. 26, No. 4, pp. 619-631, Apr. 2007.

[6] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos, “Modeling within-field gate length spatial variation for process-design co-optimization,” *Proceedings of SPIE*, vol. 5756, pp. 178-188, May. 2005.

[7] A. Gattiker, “Unraveling variability for process/product improvement,” *IEEE International Test Conference*, pp. 1-9, 2008.

[8] D. Donoho, “Compressed sensing,” *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289-1306, Apr. 2006.

[9] E. Candes, “Compressive sampling,” *International Congress of Mathematicians*, 2006.

[10] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” *Journal of Royal Statistical Society*, vol. 58, no. 1, pp. 267-288, 1996.

[11] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” *27th Asilomar Conference on Signals, Systems and Computers*, vol. 1, pp. 40-44, Nov. 1993.

[12] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximation," *Constructive Approximation*, vol. 13, no. 1, pp. 57-98, Mar. 1997.

[13] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Information Theory*, vol. 53, no. 12, pp. 4655-4666, Dec. 2007

[14] J. Tropp, A. Gilbert, and M. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Processing*, vol. 86, pp. 572-588, Mar. 2006.

[15] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 56, no. 6, pp. 2346-2356, Jun. 2008.

[16] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted norm minimization algorithm," *IEEE Trans. Signal Processing*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[17] X. Li, "Finding deterministic solution from underdetermined equation: large-scale performance modeling of analog/RF circuits," *IEEE Trans. Computer-Aided Design*, vol. 29, no. 11, pp. 1661-1668, Nov. 2010.

[18] W. Zhang, X. Li, E. Acar, F. Liu and R. Rutenbar, "Multi-wafer virtual probe: minimum-cost variation characterization by exploring wafer-to-wafer correlation," *IEEE International Conference on Computer Aided Design*, pp. 47-54, 2010.

[19] W. Zhang, K. Balakrishnan, X. Li, D. Boning and R. Rutenbar, "Toward efficient spatial variation decomposition via sparse regression," *IEEE International Conference on Computer Aided Design*, pp. 162-169, 2011.

[20] W. Mann, F. Taber, P. Seitzer and J. Broz, "The leading edge of production wafer probe test technology," *IEEE International Test Conference*, pp. 1168-1195, 2004.

[21] K. Balakrishnan and D. Boning, "Measurement and analysis of contact plug resistance variability," *IEEE Custom Integrated Circuits Conference*, pp. 416-422, 2009.

[22] C. Paige and M. Saunders, "LSQR: An algorithm for sparse linear equations and sparse least squares," *ACM Trans. on Mathematical Software*, vol. 8, no. 1, pp. 43-71, Mar. 1982.

[23] R. Gonzalez and R. Woods, *Digital Image Processing*, Prentice Hall, 2007.

[24] C. Bishop, *Pattern Recognition and Machine Learning*, Prentice Hall, 2007.

[25] T, Hastie, R, Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.

[26] R. Maronna, R. Martin, and V. Yohai, *Robust Statistics: Theory and Methods*, John Wiley and Sons, 2006.

[27] W. Press, S. Teukolsky, W. Vetterling and B. Flannery, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, 2007.

[28] G. Casella and R. Berger, *Statistical Inference*, Duxbury Press, 2001.

[29] S. Searle, G. Casella and C. McCulloch, *Variance Components*, John Wiley and Sons, 1992.

[30] L. Pang, K. Qian, C. Spanos, and B. Nikolic, "Measurement and analysis of variability in 45 nm strained-Si CMOS technology," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 8, pp. 2233-2243, Aug. 2009.

[31] Q. Zhang, K. Poola, and C. Spanos, "One step forward from run-to-run critical dimension control: Across-wafer level critical dimension control through lithography and etch process," *Journal of Process Control*, vol. 18, no. 10, pp. 937-945, Dec. 2008.

[32] Q. Zhang, K. Poola, and C. Spanos, "Across wafer critical dimension uniformity enhancement through lithography and etch process sequence: concept, approach, modeling, and experiment," *IEEE Trans. Semiconductor Manufacturing*, vol. 20, pp. 488-505, 2007.

[33] S. Kanno, G. Miya, J. Tanaka, T. Masuda, K. Kuwahara, M. Sakaguchi, A. Makino, T. Tsubone, and T. Fujii, "Controlling gate-CD uniformity by means of a CD prediction model and wafer-temperature distribution control," *Thin Solid Films*, vol. 515, pp. 4941-4944, Apr. 2007.

[34] K. Qian and C. J. Spanos, "A comprehensive model of process variability for statistical timing optimization," *Proceedings of SPIE*, vol. 6925, pp. 178-182, 2008.

[35] L. Cheng, P. Gupta, C. Spanos, K. Qian and L. He, "Physically justifiable die-level modeling of spatial variation in view of systematic across wafer variability," *IEEE Trans. Computer-Aided Design*, vol. 30, no. 3, pp. 388-401, Mar. 2011.

[36] J. Sali, S. Patil, S. Jadkar, and M. Takwale, "Hot-wire CVD growth simulation for thickness uniformity," *Thin Solid Films*, vol. 395, no. 1-2, pp. 66-70, Sep. 2001.

[37] S. Sakai, M. Ogino, R. Shimizu, and Y. Shimogaki, "Deposition uniformity control in a commercial

scale HTO-CVD reactor," *Materials Research Society Symposium Proceedings*, 2007.

[38] J. Brcka and R. Robison, "Wafer redeposition impact on etch rate uniformity in IPVD system," *IEEE Trans. Plasma Science*, vol. 35, no. 1, pp. 74-82, Feb. 2007.

[39] T. Kim and E. Aydil, "Investigation of etch rate uniformity of 60 MHz plasma etching equipment," *Japanese Journal of Applied Physics*, vol. 40, pp. 6613-6618, Nov. 2001.

[40] T. Kim and E. Aydil, "Effects of chamber wall conditions on Cl concentration and Si etch rate uniformity in plasma etching reactors," *Journal of the Electrochemical Society*, vol. 150, no. 7, pp. G418-G427, Jun. 2003.

[41] I. Ahsan et al., "RTA-driven intra-die variations in stage delay, and parametric sensitivities for 65 nm technology," *2006 Symposium on VLSI Technology*, pp. 170-171, 2006.

[42] S. Springer, S. Lee, N. Lu, E. Nowak, J. Plouchart, J. Watts, R. Williams and N. Zamdmer, "Modeling of variation in submicrometer CMOS ULSI technologies," *IEEE Trans. Electron Devices*, vol. 53, no. 9, pp. 2168-2178, Sep. 2006.

[43] R. Deaton and H. Massoud, "Manufacturability of rapid thermal oxidation of silicon: Oxide thickness, oxide thickness variation, and system dependency," *IEEE Trans. Semiconductor Manufacturing*, vol. 5, no. 4, pp. 347-358, Nov. 1992.

[44] T. Futase, T. Kamino, Y. Inaba, and H. Tanimoto, "Uniform, low-resistive Ni-Pt silicide fabricated by partial conversion with low metal-consumption ratio," *IEEE Trans. Semiconductor Manufacturing*, vol. 24, no. 4, pp. 545-551, Nov. 2011.

[45] C. Chao, S. Hung, and C. Yu, "Thermal stress analysis for rapid thermal processor," *IEEE Trans. Semiconductor Manufacturing*, vol. 16, no. 2, pp. 335-341, May. 2003.

[46] J. Hebb and K. Jensen, "The effect of patterns on thermal stress during rapid thermal processing of silicon wafers", *IEEE Trans. Semiconductor Manufacturing*, vol. 11, no. 1, pp. 99-107, Feb. 1998.

[47] P. Friedberg, *Spatial modeling of gate length variation for process-design co-optimization*, PhD dissertation, Univ. of California, Berkeley, Dept. of Electrical Engineering and Computer Science, 2007.

[48] M. Bhushan, A. Gattiker, M. Ketchen and K. Das, "Ring oscillators for CMOS process tuning and variability control," *IEEE Trans. Semiconductor Manufacturing*, vol. 19, no. 1, pp. 10-18, Feb.

2006.

[49] W. Daascb, J. McNames, D. Blockelman, K. Cota, and B. Madge, “Variance reduction using wafer patterns in Iddq data,” *IEEE International Test Conference***,** pp. 189-198, 2000.

[50] R. Madge, B. Goh, V. Rajagopalan, C. Maccbietto, R. Daasch, C Schuennyer, C. Taylor, and D. Tumer, “Screening MinVdd outliers using feed-forward voltage testing,” *IEEE International Test Conference*, pp. 673-682, 2002.

[51] B. Benware, R. Madge, C. Lu, and C. Daasch, “Effective comparison of outlier screening methods for frequency dependent defects on complex ASICs,” *IEEE VLSI Test Symposium*, pp. 39-46, 2003.

[52] P. Huber, *Robust Statistics*, John Wiley and Sons, 1981.

[53] J. Fox, “Robust Regression”, (Online). http://cran.r-roject.org/doc/contrib/FoxCompanion/appendix-robust-regression.pdf, 2002.

[54] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, 1996.

[55] S. Ikeda, K. Nemoto and M. Funabashi, “Single-wafer technology in a 300-mm wafer fab,” *Ulsi Process Integration III: Proceedings of the International Symposium*, pp. 163-167, 2003.

[56] M. Orshansky, S. Nassif, and D. Boning, *Design for Manufacturability and Statistical Design: A Constructive Approach*, Springer, 2007.

[57] S. Saxena, private communications, 2012.

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

[59] Semiconductor Industry Associate, *International Technology Roadmap for Semiconductors*, 2007.

[60] Semiconductor Industry Associate, *International Technology Roadmap for Semiconductors*, 2009.

[61] Semiconductor Industry Associate, *International Technology Roadmap for Semiconductors*, 2011.

[62] V. Vahedi, M. Srinivasan and A. Bailey, “Raising the bar on wafer edge yield—an etch perspective,” *Solid State Technology*, vol. 55, no. 11, Nov. 2008.

[63] O. Yavas, E. Richter, C. Kluthe and M. Sickmoeller, “Wafer-edge yield engineering in leading-edge DRAM manufacturing,” *Semiconductor Fabtech,* no. 39, Mar. 2009.

[64] M. Bhushan, A. Gattiker, M. Ketchen and K. Das, “Ring oscillators for CMOS process tuning and variability control,” *IEEE Trans. Semiconductor Manufacturing*, vol. 19, no. 1, pp. 10-18, Feb. 2006.

[65] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Minin*g. Addison-Wesley, 2006.

[66] A. Strojwas, “Conquering process variability: A key enabler for profitable manufacturing in advanced technology nodes,” *IEEE International Symposium on Semiconductor Manufacturing*, pp. xxv–xxxii, 2006.

[67] W. Pennebaker and J. Mitchell, *JPEG still image data compression standard*, Springer, 1993.

[68] D. Taubman and M. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*, Springer, 2001.

[69] Z. Xiang, K. Ramchandran, M. Orchard, and Y. Zhang, “A comparative study of DCT- and wavelet-based image coding,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 9, no. 5, pp. 692-695, Apr. 1999.

[70] S. Grgic, M. Grgic and B. Zovko-Cihlar, “Performance analysis of image compression using wavelets,” *IEEE Trans. Industrial Electronics*, vol. 48, no. 3, pp. 682-695, Jun. 2001.

[71] G. Casella and R. Berger, *Statistical Inference*, Duxbury Press, 2001.

[72] C. Chui, *An Introduction to Wavelets*, Academic Press, 1992.

[73] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” *5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297, 1967.

[74] M. Ester, H. Kriegel, J. Sander and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” *International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.

[75] G. May and C. Spanos, *Fundamentals of Semiconductor Manufacturing and Process Control*, Wiley-IEEE Press, 2006.

[76] A. Jain and R. Dubes, *Algorithms for Clustering Data*, Prentice Hall, 1988.

[77] D. Cordes, V. Haughtonb, J. Carewc, K. Arfanakisd and K. Maravilla, “Hierarchical clustering to measure connectivity in fMRI resting-state data,” Magnetic Resonance Imaging, vol. 20, no. 4, pp. 305-317, May 2002.

[78] P. Jiang and M. Singh, “SPICi: a fast clustering algorithm for large biological networks,” *Bioinformatics*, vol. 26, no. 8, pp. 1105-1111, Apr. 2010.

[79] S. Salvador and P. Chan, “Determining the number of clusters/segments in hierarchical

clustering/segmentation algorithms," *International Conference on Tools with AI*, pp. 576-584, 2004.

[80] G. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, April 1965.

[81] R. Keyes, "The impact of Moore's Law," *IEEE Solid-State Circuits Newsletter*, vol. 11, no. 5, pp. 25-27, Sep 2006.

[82] S. Campbell, *The Science and Engineering of Microelectronic Fabrication*, Oxford university press, 2001.

[83] R. Keyes, "The effect of randomness in the distribution of impurity atoms on FET thresholds," *Applied Physics*, vol. 8, pp. 251-259, 1975.

[84] K. Gettings and D. Boning, "Study of CMOS process variation by multiplexing analog characteristics," *IEEE Trans. Semiconductor Manufacturing*, vol. 21, no. 4, pp. 513-525, Nov. 2008.

[85] K. Agarwal, J. Hayes, and S. Nassif, "Fast characterization of threshold voltage fluctuation in MOS devices," *IEEE Trans. Semiconductor Manufacturing*, vol. 21, no. 4, pp. 526-533, Nov. 2008.

[86] K. Agarwal, F. Liu, C. McDowell, S. Nassif, K. Nowka, M. Palmer, D. Acharyya, and J. Plusquellic, "A test structure for characterizing local device mismatches," *Symposium on VLSI Circuits*, pp. 67-68, 2006.

[87] T. Fischer, E. Amirante, P. Huber, T. Nirschl, A. Olbrich, M. Ostermayr, and D. Schmitt-Landsiedel, "Analysis of read current and write trip voltage variability from a 1-MB SRAM test structure," *IEEE Trans. Semiconductor Manufacturing*, vol. 21, no. 4, pp. 534-541, Nov. 2008.

[88] N. Drego, A. Chandrakasan, and D. Boning, "A test-structure to efficiently study threshold-voltage variation in large MOSFET arrays," *IEEE International Symposium on Quality Electronic Design*, pp. 281-286, 2007.

[89] M. Bhushan, A. Gattiker, M. Ketchen, and K. Koushik, "Ring oscillators for CMOS process tuning and variability control," *IEEE Trans. Semiconductor Manufacturing*, vol. 19, no. 1, pp. 10-18, Feb. 2006.

[90] J. Panganiban, *A Ring Oscillator Based Variation Test Chip*, MEng thesis, Massachusetts Institute of Technology, Dept. of Elect. Engineering and Comp. Science, June 2002.

[91] L. Pang and B. Nikolic, "Impact of layout on 90nm CMOS process parameter fluctuations," pp. 69-70, *Symposium on VLSI Circuits*, 2006.

[92] L. Pang and B. Nikolic, “Measurement and analysis of variability in 45nm strained-Si CMOS technology,” *IEEE Custom Integrated Circuits Conference*, pp. 129-132, 2008.

[93] Z. Guo, A. Carlson, L. Pang, K. Duong, T. Liu, and B. Nikolic, “Large-Scale Read/Write Margin Measurement in 45nm CMOS SRAM Arrays,” *IEEE Symposium on VLSI Circuits*, pp. 42-43, 2008.

[94] J. Chen, D. Sylvester, C. Hu, H. Aoki, and S. Oh, “An on-chip, interconnect capacitance characterization method with sub-femto-farad resolution,” *IEEE International Conference on Microelectronic Test Structures*, pp. 77-80, 1997.

[95] L. J. van der Pauw, “A method of measuring specific resistivity and Hall effect of discs of arbitrary shape,” *Philips Research Reports*, vol.13, no.1, pp. 1-9, 1958.

[96] J. Doh, D. Kim, S. Lee, J. Lee, Y. Park, M. Yoo, and J. Kong, “A unified statistical model for inter-die and intra-die process variation,” *International Conference on Simulation of Semiconductor Processes and Devices*, pp. 131-134, 2005.

[97] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos, “Modeling within-die spatial correlation effects for process-design co-optimization,” *International Symposium on Quality Electronic Design*, 2005, pp. 516-521, 2005.

[98] S. Reda and S. Nassif, “Accurate spatial estimation and decomposition techniques for variability characterization,” *IEEE Trans. Semiconductor Manufacturing*, vol. 23, no. 3, pp. 345-357, Aug. 2010.

[99] K. Qian, C. Spanos, and B. Nikolic, “Hierarchical modeling of spatial variability with a 45nm example,” *Proceedings of SPIE*, vol. 7275, pp. 727505-1-727505-12, 2009.

[100] W. Zhang, X. Li, F. Liu, E. Acar, R. Rutenbar and R. Blanton, "Virtual probe: a statistical framework for low-cost silicon characterization of nanoscale integrated circuits," *IEEE Trans. Computer-Aided Design*, vol. 30, no. 12, pp. 1814-1827, Dec. 2011.

[101] W. Zhang, X. Li and R. Rutenbar, "Bayesian virtual probe: minimizing variation characterization cost for nanoscale IC technologies via Bayesian inference," *IEEE Design Automation Conference*, pp. 262-267, 2010.

[102] H. Chang, K. Cheng, W. Zhang, X. Li and K. Butler, "Test cost reduction through performance prediction using virtual probe," *IEEE International Test Conference*, pp. 1-9, 2011.

[103] PDF Solutions, “Integrated Yield Ramps”, (Online). http://www.pdf.com/integrated-yield-ramp, 2012.

[104] N. Draper and H. Smith, *Applied Regression Analysis*, Wiley-Interscience, 1998.