

High Dimensional Data Mining in
Manufacturing
or
High Dimensional Data Mining in
Complex Manufacturing Processes

—
diss.tex: September 22, 2002

Defense: 11 November 2002, 8:00 in Olsson Hall 111A

A Dissertation Presented to the
Faculty of the School of Engineering and Applied Science
University of Virginia

In Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy (Systems Engineering)

by
David R. Forrest

December 2002

APPROVAL SHEET

This dissertation is submitted in partial fulfillment of the requirements for the degree of:

Doctor of Philosophy (Systems Engineering)

This dissertation has been read and approved by the examining Committee:

James Aylor, Ph.D., Electrical Engineering
Committee Member

Donald Brown, Ph.D., Systems Engineering
Committee Chair

Christina M. Mastrangelo, Ph.D., Systems Engineering
Thesis Advisor

Steven Patek, Ph.D., Systems Engineering
Committee Member

K.P. White, Ph.D., Systems Engineering
Committee Member

Accepted for the School of Engineering and Applied Science:

Dean, School of Engineering and Applied Science

December 2002

Symbol	Meaning
p, p_i	Number of variables, in system i
n, n_i	Number of observations, in system i
σ_{ab}	Covariance of a and b
σ_x^2	Variance of x
$\widehat{\mathbf{Y}}, \widehat{\mathbf{Y}}_i$	Metamodel estimate, submodel i estimate.
\mathbf{X}	Column vector of X variables
$\underline{\mathbf{X}}$	$n \times p$ Data matrix
$\underline{\mathbf{Y}}$	$n \times p_y$ output matrix or vector
β	Vector of model parameters
$\widehat{\beta}$	Vector of estimated model parameters
$\underline{\mathbf{H}}$	the H matrix from linear models

Table 1: Table of Symbols

Contents

Table of Symbols	i
1 Introduction	1
1.1 Characteristics of Complex Systems	4
1.2 Statement of Work	9
1.3 Problem Statement	14
1.3.1 Background of Semiconductor Manufacturing	14
1.3.2 Mathematical Model of the Process	17
1.3.3 Semiconductor Manufacturing Data Collection and Storage	20
1.4 Summary	25
2 Literature Review	26
2.1 Semiconductor modeling	26
2.2 Complexity, Sample Size, and Modeling	29
2.3 High Dimensional Modeling and Visualization	31
2.4 Reduction of dimensionality	32
2.5 Alternate high-dimensional domains and approaches	34
2.5.1 Chemometrics	36
2.5.2 Text Mining	37
3 Application and Methodology	38
3.1 Application	39
3.2 Methodology	39
3.2.1 Hierarchical Modeling	39
Subcomponents	39
Intermediate variables	41
3.2.2 Meta-Analysis applied to high dimensional complex systems	41
3.2.3 The variables of interest	42
Risk, Odds, Probit, and Logit	42
Random Regression Model of log risk defect	43
3.3 Proposed Model	43
3.3.1 Subcomponents as a risk factor	43

	Risk, defect rate, and odds are functions of each other . . .	43
	Baseline risk is the overall, or intercept term.	43
	Benefits of combined submodels	43
3.3.2	Comparison to intermediate surrogate variables	44
3.3.3	(E.g. TEG as an intermediary)	44
3.3.4	Continuity Adjustment ($\pm e$)	44
3.3.5	Combination of models and relative risks	45
3.3.6	Logit and Probit models	45
3.4	Research Plan	45
3.5	Intermediate Variable Model	46
3.6	Initial Approach and Preliminary Results	49
3.7	Results and Summary of Preliminary Approach	49
4	Results	51
4.1	Description of specific problem in Semiconductor Manufacturing	52
4.2	Results of the method	52
4.2.1	Submodels – subprocess estimates of yield	52
4.2.2	Metamodel – yield/defect rate as a function of subprocess estimates	52
4.3	Evaluation	52
4.3.1	Competing models	52
4.3.2	Effectiveness	52
4.4	Recommendations	52
4.5	Summary	52
5	Discussion	53
5.1	Applicability in situations where $p \gg n$	53
5.2	Applicability in systems of subcomponents	54
5.3	Recommendations	54
5.4	Concerns	54
5.4.1	Interactions between variables in submodels	55
5.4.2	Interaction terms between submodels	56
5.5	Correlations between variables	56
	Partitioning of a flat linear model	57
5.6	Other applications	59
6	Conclusion	60
6.1	Overview	60
6.2	Results	60
6.3	Summary	60
	Bibliography	61

Appendix	67
A Worries – the appendix to be cut out	68
A.1 Miscellany	68
A.2 To do List	68
A.3 Research Plan	70

List of Tables

1 Table of Symbols i

List of Figures

1.1	Simple Model	16
1.2	Process Length Variance	21
A.1	Hierarchical Model	72

Listings

Chapter 1

Introduction

High dimensional complex systems present a number of unique challenges in modeling for understanding, prediction, and control. This work focuses on modeling to build understanding of the contributions of the subcomponents of a large complex system to the performance characteristics of the overall system.

As a motivating example, a semiconductor manufacturing plant illustrates a large system composed of many interrelated subcomponents, with several levels of understanding of the system.

Process engineers in the semiconductor industry often focus on the individual processes rather than the overall manufacturing system (Herrmann et al., 2000, p. 1491).

Simulation models of factories focus on measures of system performance such as total lot processing time, throughput, utilization, and work-in-process (Herrmann et al., 2000). However, high level models of *product performance characteristics*, such as yield, conformance to specifications, and quality, require understanding of the influences and contributions of the sub-systems outputs on the end product. For example, it is believed that quality character-

istics in processing at the Gate Contact level can have significant effects on the performance and yield characteristics of the memory chips. An understanding of this effect, and the effects of competing models is important for management decision making.

From a bottom-up perspective, manufacturing systems have embraced statistical process monitoring and control to identify and understand problem areas in manufacturing. Traditionally, these were univariate and tightly bound to the specific manufacturing process. Distinct or interrelated processes require more advanced techniques such as multivariate or multi-stream techniques (Montgomery, 1996, p. 235). As the measurement systems were integrated into the machinery, control methods were used to more accurately monitor and actively control the processes. In complex manufacturing processes, the necessity for control across distinct runs has inspired the run-to-run (R2R) control methodology, in which information external to the process is used to adjust the active controls in the process (Moyne et al., 2001). Optimizing each sub-process to minimize the variation in its outputs may indeed improve the overall product performance, but local optimization of subprocesses does not guarantee global optimization, especially if the local optimizations are not of the same variables. This research builds a framework for combining locally optimized processes into a global model of product performance characteristics.

Hierarchical control of manufacturing systems applied in the semiconductor industry is a new and active area of research Moyne et al. (2001, p. 9, p. 321). The work shown in Leang et al. (1996) is an application of the R2R control scheme to include several machines in a subprocess to form a single layer on a semiconductor wafer. This is an expansion of the single sub-process to several sub-processes and is the first step up from the bottom level of a bottom-up

hierarchical model. On the other hand, simulation models of logistical characteristics of a factory, for example, are the first step down from a top-down hierarchical model. Between these top-down and bottom-up methodologies is a need for understanding the relationship between low level sub-process characteristics and *product performance characteristics*.

Our conceptual approach is to use hierarchical modeling and data fusion as a means of addressing the problem of understanding the interrelationships between low-level process parameters and high level product performance characteristics.

The key characteristics of this data seem to be well studied, however, interactions between the characteristics outlined below combine to make the application of the common solutions difficult.

1. Inconsistent data due to:
 - (a) Frequent process changes, which produce a dynamic process structure.
 - (b) Missing data due to structure changes. E.g., use of a different machine produces a different set of variables
 - (c) Key punch errors due to manual entry erroneously report values against improper lots.
2. High dimensionality with respect to the number of observations (25 lots/day * 30 days/month = 2000 rows by 21000 columns)
3. Data aggregation issues (chip / reticle / wafer/ lot / batch)
4. Interpretation of multiple, competing process/product models

Although these characteristics have been studied separately, these issues combine to produce a problem that is incompletely understood.

1.1 Characteristics of Complex Systems

In this work, a complex system is distinguished from a simpler system by the characteristics of the data and information available. Simple systems have simple data structures with consistent connections between the elements. Examples drawn from the UCI machine learning databases include cardiac data, indexed by patient with results, a cars database with prices and performance for a number of models, housing data. These systems are indexed on individual observations and map one or a number of output variables to one or a number of input variables. Complex systems encompassing a number of simpler systems raise questions of indexing, aggregation, competing models, and the combination of the submodels.

Supposing the cardiac, car, and housing data were all somehow related in a larger system, how can useful models be developed from the data to help provide understanding of the influence of the systems subcomponents on a variable of interest. For example, high blood pressure as a function of housing, medical data, census, and consumer goods data. Medical data is likely indexed on individual patients, while housing and census data may be by the household or by region.

Inconsistent process structure data: Changes in a process fracture a data set and raise questions of whether the change affected, improved, or hindered one or several indices of performance. In the univariate case with one process

change, a simple hypothesis test suffices. With multivariate indices, multiple models and tests must be considered together, necessitating adjustments in hypothesis testing significance levels, (Neter et al., 1996, p. 1024) For example, the effects of a policy change in a school or administration at UVA on a variable that is used to rank colleges might be testable using a two-sample hypothesis test, but one must also consider the many other policy changes that occur at UVA. The many process changes in a semiconductor manufacturing plant also interrelate and confound the analysis process.

Missing data due to sampling plans in the Die Sort testing process produces missing data rates of up to 67%. These data may not be missing at random – wafers and lots with higher failure rates are examined more carefully. Missing data due to process changes are also non-random.

Large number of variables relative to sample size: Many tools exist for dealing with high dimensional data, however, these methods each begin with a well formed data matrix. Combining heterogeneous data sets produces increased problems with missing data and dimensionality. Forming a larger data matrix requires joins of datasets, and the large number of variables exacerbates missing data problems. Considering the Metal 1 layer production data, suppose only 7% of the data in each set was missing at random in each of 10 sub-processes, the resulting table join would have 52% rows with missing data. Established techniques, such as SAS PROC REG, PROC CLUSTER, and PROC PCA, each case-wise delete data with missing elements. Combination of sets with missing data results in larger missing data problems.

Solutions to problems of missing data are generally limited to missing at random (MAR) and missing completely at random (MCAR), (SAS:ProcMI, p.

154). Complex systems that require combination of data from different data sources must treat non-random sources of missing data, as well as the technical problems of indexing and combining many large data sets. Clearly, missing data due to process changes causes systematic changes in the data set. Large blocks of nulls exist after an old process ended or before a new process is started. Missing data due to process changes is systematic, and violates the MAR assumption.

Data aggregation: Differences in data aggregation is normally treated by summarizing the non-aggregated data to match the level of aggregation in the smaller dataset, or by expanding the aggregated data to match the non-aggregated data through the use of duplication, resampling, or simulation. Expansion of the analysis to lower levels of aggregation increases the dimensionality. For example, using chip level information to predict lot level characteristics increases the number of variables by a factor equivalent to the number of chips in a lot.

Interpretation of competing models: Multiple competing models require higher levels of significance than single models, (Neter et al., 1996, p. 1024). Often there is more than one models available for describing a system, and it would be good to use methods which combine the available models?. Decision makers need tools to compare competing models. For example, some process engineers believe the Gate Contact (GC) layer is the most critical layer; others say that lithography is the most important process.

Hierarchical modeling: Hierarchical models are present in a number of disciplines. In manufacturing, a hierarchical nested control methodology establishes clearly defined levels, inputs, outputs, and the relationships between them, and fits process controllers into controllable points in the process (Moyné et al., 2001, p. 8). It is not our intent to deliver control mechanisms, but to develop understanding of the relationships between process variables and product characteristics.

Hierarchical modeling in the form of predicting the intermediate level characteristics and using them in turn to predict high level characteristics is limited to the prediction power of the intermediate data. If the statistics of interest cannot be reliably predicted using the intermediates, then estimates of the intermediates will do no better. Based on this result, the research plan will use a flattened hierarchy to estimate of final variables directly, and will recombine these using regression and more advanced tools.

Hierarchical modeling used in medical research is called meta-analysis, which combines the results of multiple studies to improve the information. Meta-analysis are typically performed on similar sorts of interventions, with the hierarchical levels and elements based on the study, type of model, and an overall effect. These studies typically have univariate outputs, and the studies identify the effects of treatments on an outcome (Sutton et al., 2000). E.g., several drug efficacy studies show that a drug reduces the risk of disease; the studies are grouped together, and pooled estimates of effects by study type and drug are produced. The benefit of meta-analysis is to provide a Bayesian prior distribution of an effect, which tends to moderate and tighten the confidence intervals of each study, increasing the information gained.

Another hierarchical modeling body of literature is structural equation

modeling (SEM), a methodology of specifying causality and confirming a hierarchy of models in a system. SEM studies are primarily low dimensional, and serve to support models of causality in social sciences.

Data fusion is a method for combining sources of data to produce better information about the systems. Meta-analysis is means for combining a number of disparate published studies into an overall model. In the manufacturing environment, states of machine operation (in control /out of control), could be considered treatments, and the yield and failure rate data could be modeled as an log odds ratio. If partitions of the manufacturing system were modeled as separate studies, then the resulting separate models could be combined using meta-analytical techniques. Results from a meta-analysis model would be “out of control conditions in the sub-models under study will produce $\mu_i \pm c\sigma_i\%$ greater failure rates in margin yield for each sub-model i .”

Although I mentioned dynamic programming in my proposal, I use it only to provide a notation for describing manufacturing questions. E.g, the ill-defined “Golden signature” modeling can be used for diagnosis of plant problems, in-line disposition, or future design, all with very different needs for analysis. The observability and controllability of a control system is dependent upon the understanding the system. Building understanding of the overall system must come before prediction and control. This work will help to build understanding of the relationships between process variables and product variables.

1.2 Statement of Work

I will use a top-down approach to establish a hierarchy of sub-systems that could be combined to model the effects of low-level process parameters on product performance level issues. This is neither the top level work-flow simulation model, nor the expansion of low level control systems to higher levels, but the modeling for understanding of the high level product parameters based on the lower level process parameters.

I propose building a two level hierarchical model. The sub-components are the models matching existing data structures and engineering expectations. Samples of sub-models matching the data structures is the electrical testing databases and the logistics database. Samples of sub-models from engineering expectations are the Metal 1 process data, the Gate Contact process data, the Key Quality Control measurement parameters, and the TEG PSL database. If simple models of product outputs can be created for each of these, the models can then be combined using data fusion techniques.

To demonstrate the method, while limiting the scope of the effort, I will use the lot-level aggregated data on three output variables, (margin failure, functional, and DC testing data), process data from two semiconductor layers, (Metal-1 and Gate Contact), the quality control data, and the electrical test data to build models of the margin yield, the functional yield, and the DC testing yield.

Equation 1.1 shows how partitioning the problem based on sub-models can reduce the dimensionality of the problem. Each of the i sub-models is a $\widehat{Y}_i = f(\underline{X}_i)$ where the estimation function could be linear, but may require data cleaning, missing data imputation, variable selection, feature reduction,

or feature extraction. These models may not explain significant portions of the variation in $\underline{\mathbf{Y}}$, but it is hoped that, taken together, the several models can help explain some portion of the variation.

$$\begin{aligned}
 \underset{2000 \times 3}{[\widehat{\mathbf{Y}}]} = f \left(\underset{2000 \times 21000}{\begin{bmatrix} \underline{\mathbf{X}}_1 & \emptyset & \emptyset & \cdots & \underline{\mathbf{X}}_n \\ \emptyset & \underline{\mathbf{X}}_2 & \underline{\mathbf{X}}_3 & \cdots & \underline{\mathbf{X}}_n \\ & & \emptyset & & \end{bmatrix}} \right) = f \left(\underset{2000 \times 3n}{\begin{bmatrix} \widehat{\mathbf{Y}}_1 & \emptyset & \emptyset & \cdots & \widehat{\mathbf{Y}}_n \\ \emptyset & \widehat{\mathbf{Y}}_2 & \widehat{\mathbf{Y}}_3 & \cdots & \widehat{\mathbf{Y}}_n \\ & & \emptyset & & \end{bmatrix}} \right) \quad (1.1)
 \end{aligned}$$

$$\widehat{\mathbf{Y}}_i = f \left(\underline{\mathbf{X}}_i \right) \quad (1.2)$$

Although the partitioning and segmenting of the problem decreases the dimensionality of the sub-problems, it does not solve all the problems in the manufacturing data. Missing data, high dimensionality, aggregation, and structure changes will remain as problems, but will be more manageable in the smaller models.

Techniques for building the submodels (Equation 1.2) include regression, logistic regression, principle components regression, or partial least squares. Since the outputs of the models are to be combined in the upper level model, other techniques could be used, as long as they produce an estimate of the $\widehat{\mathbf{Y}}$ product characteristics. The top level model may use these estimates to build understanding of the effects of the sub-models on the $\widehat{\mathbf{Y}}$, and the interactions between processes.

Theoretical elements – Semiconductor manufacturing has a number of interesting elements, some of which have been treated separately.

- Large number of variables as compared to the number of observations
($p \gg n$)
- Dynamic process structure
- Data stored in a transactional database
- Multiple levels of data (individual/lot/batch)
- Missing data
- Corrupt data
- Multiple output variables
- Process Control
- Process Prediction
- Process Understanding
- Categorization of defects
- Multiple users have differing objectives in analyzing the data
- Combination of models

Automated manufacturing systems, such as those that produce semiconductors, can produce such large quantities of data that understanding of the interrelations between subsystems is difficult. This work produces

In order to control and improve chip production in semiconductor manufacture, a company may seek to use manufacturing and process data already recorded during the process to more fully understand the system and provide avenues for improvements. Semiconductor manufacturers collect a large amount of data, in terms of storage space and number of variables, but small in terms of the number of coherent observations. While any particular operation may have a number of observations, the large number of monitored variables produce effects similar to short run manufacturing processes: insufficient degrees of freedom to reliably model the process. This work will produce a modeling methodology for managing hierarchical manufacturing data and will seek to produce useful models for semiconductor manufacturing, and complex manufacturing systems in general.

Smaller lot sizes and more flexible manufacturing processes, along with the increase process complexity, combine to produce short-run processes. As more automated measuring and recording equipment enters the manufacturing process, a difficulty with the dimensionality of the problem emerges: runs shorter than the dimensionality of the problem. A high dimensional manufacturing process can have fewer distinct observations n than the number of process variables p . In prior work with a semiconductor manufacturer, we established that direct models of $Y_{yield} = \beta X_{process}$ can be ill-defined due to the dimensions of the data matrix $X_{\{n \times p\}}$ where $n \ll p$. These conditions lead to instability in the parameter estimates β for a linear model, and singularity of the $cov(X)$, but they occur in complex manufacturing processes. Realizing that the degrees of freedom in the system is limited by the number of observations, additional constraints must be placed upon the overall model. Assuming a hierarchical structure to the process, i.e. that the outputs of the system are functions of cer-

tain key parameters of the system, which are in turn functions of lower level operations in the production process, may constrain the models into estimable and testable problems.

High dimensional data analysis requires effective visualization methods, since traditional methods such as run plots and scatter plots do not scale well to high dimensional systems (Forrest and Mastrangelo, 2001). Using methods from clustering to sort the variables and observations can aid in high dimensional visualization. A certain manufacturer seeks a “Golden Signature” program which intends to identify production parameters from high quality lots and estimate the effects of deviations from these ideal parameters. This “Golden Signature” program requires a comprehensive model relating the many low level process variables to the high level yield variables.

Although difficult, this project is an ideal application of systems engineering due to the complex system managed by different groups of people. Semiconductor manufacturing is an extreme case of complicated manufacturing systems, with issues of discrete part manufacturing, aggregations and dis-aggregations of data in time, production lots, and production processes. The interactions between production, engineering, information technology, and management indicate a need for an interdisciplinary method integrating the process. The general methodology proposed here is to develop a methodology using hierarchical structures inherent in the manufacturing and engineering processes to manage the complex models and high dimensionality of manufacturing processes.

1.3 Problem Statement

The extraction of meaningful inferences from the large mass of manufacturing data generated by the semiconductor manufacturing process in particular is problematic in two significant ways: first, the data is not often stored in a form amenable to analysis, and second, the large amount of data is often very small when compared to the complexity of a model representing the semiconductor manufacturing plant. This research focuses on these two problems: managing large ($n \ll p$) data for analysis, and using the data to discover interesting relationships.

1.3.1 Background of Semiconductor Manufacturing

Commercial semiconductor devices are manufactured in and on the surface of wafers from large ultra-pure crystals—thin disks, typically 200mm or 300mm in diameter. An area on the wafer containing a single discrete device or integrated circuit (IC) is called a chip or die. Depending on the dimensions of the wafer and the dies, several hundred chips are formed on a single wafer.

During fabrication, wafers are transported and processed in standard lots of twenty-five wafers each. Each lot undergoes hundreds of individual processing steps, in which different parts of the ICs are etched in thin layers of material grown or deposited on the working surface of the wafers. Each process step must be tightly controlled to ensure dimensional tolerances typically measured in nanometers.

Fabrication of a single lot requires approximately three months. Throughout, process settings, engineering parameters, and test data are logged for each fabrication tool at both the wafer and lot level, via a central computer network

called a manufacturing execution system. With as many as 5000 wafer starts a week, process and engineering databases requiring hundreds of gigabytes of memory are normal.

The data details the manufacturing processes involved in the production process. Analysis of the data differs from current data mining techniques developed for business sales information, market-basket analysis, image analysis, or spatial data because of the large number of variables, interactions between sub-processes and relatively small number of observations. For example, a memory device involving 22 layers of semiconductor can involve 524 processing steps over 3 months with 21710 process variables. Figure 1.1 shows a sample of 90 days of lot level production data for one product, the misalignments between separate data tables, and that $n = 221 \ll 21710 = p$. Besides vast amounts of data, another challenge is that the measurements are commonly collected on different aggregations of parts at the chip, wafer, batch and lot levels. Since the measurements for a particular chip are spread out over time, collected at different aggregation levels and are many with respect to the production yield data, current data mining and analysis techniques such as clustering and linear regression modeling are inefficient and difficult to apply to semiconductor manufacturing environments.

The target of the proposed work is at the system operational level and is to extract knowledge from data from sophisticated processes in order to improve operations - that is to improve productivity, decrease ramp-up time, identify and validate quality control parameters, these will ultimately increase yield. The anticipated research will focus on two areas: operational modeling of manufacturing data and data representation and manipulation. I will develop a methodological approach to solving the complex modeling problems that arise

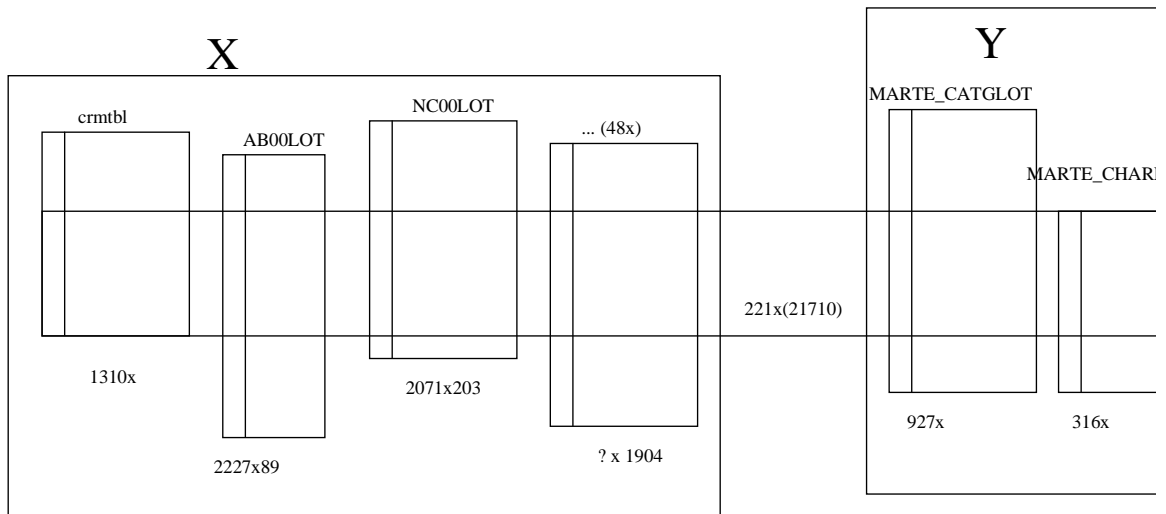


Figure 1.1: Schematic of Simple Model of DS and QC data. Note that the number of rows (lots) differ in each data table, and that the intersection of all data rows for this 90 day sample of data is $n = 221, p = 21710$.

in semiconductor manufacturing. I will also show how subsystems of the manufacturing process could be combined to produce an overall model suitable for process monitoring and improvement.

1.3.2 Mathematical Model of the Process

One way to think of a manufacturing process is as a dynamic program with a stage for each processing step. In semiconductor manufacturing, the various testing results, such as wafer yield, die sort yield, bit-fail maps, margin yield, bit retention time, refresh rates, and others are the outputs of the process at the last stage. The data collected at each step adds to the total information available at that stage. Mathematically, the path of a wafer through the production process could be represented as:

$$\begin{aligned}
 n &= 0, \dots, N \\
 I_0 &= \text{What is known before starting production} \\
 X_0 &= \text{The initial state of the wafer} \\
 A_0 &= \text{The initial Action} \\
 I_n &= \{I_{n-1}, a_{n-1}, Y_{n-1}\}; \text{ Information vector} \\
 A_n &= S_n(I_n); \text{ Action vector} \\
 X_n &= f_n(X_{n-1}, a_{n-1}, d_n); \text{ State vector} \\
 Y_n &= M_n(X_n, e_n); \text{ Results vector}
 \end{aligned}$$

where: n is the stage or step of the process, I_n is the what we know at step n , A_n is the action we choose, (which is based on some strategy S_n that uses the information we know), X_n is the “true” state or condition of the wafer after the

step (which is unknowable, and may be dependent on some disturbance d_n), and Y_n is the new data we learn from the process step through a measurement process M_n including errors e_n . Y_N is the last information we take from a part, and would be the final test data. If you know Y_N for a wafer, you can calculate the die sort yield, the margin yield and the sales price, however, profit also depends on costs, which in turn depend on the production history.

If we think of each uppercase I_n, A_n, X_n, Y_n as a row vector of the various parameters in each processing step, the current database system records the different A_n vector of machine settings and Y_n vector of measurements at each processing step. For example, if step $n = 5$ is a visual inspection that is always done the same way, A_5 is the constant procedure that is used for inspection, Y_5 is the results of the inspection, d_5 is any change in the wafer due to the inspection (e.g. a mote of dust became stuck to the part), and the part changed from an uninspected wafer without dust, (X_4), to an inspected part with dust (X_5), and for the next step, we know everything we did before, plus the facts that it was inspected, and the results of the inspection. For a more complex operation, masking for example, the vectors would be much more complicated: the action would have many more options and machine settings, and the step would produce much more data.

This model captures several important facts. We may not know everything important about the wafer at every stage, we measure some features that may or may not be representative of the state of the part, we can do different things at different stages, and we know more and more about the part as the part steps through the process. The I_N vector holds every item of recorded data about a particular wafer, and may be thousands or millions of attributes wide. This model is very general, but has enough elements to represent several

problems of interest to a manufacturer: yield improvement, design of the best recipe, in-line classification and disposition, and identification of new defects.

Each of these alternative manufacturing problems depend on estimating Y_N from different $I_{0,\dots,N}$. This mathematical model is general enough to capture all of the elements of these problem, but it might also be intractable due to a number of problems:

- The information matrix is high dimensional, representing the many variables in the semiconductor matrix.
- The relatively low number of observations resulting from the frequent process changes results in insufficient degrees of freedom for parameter estimation.
- The aggregations and dis-aggregations of lots also result in complications in the structure of the information matrix.
- Reworked production results in repeated measures.
- Key punch errors result in missing and erroneous data.

Although the mathematics of a dynamic program provides a good structure for describing a complex semiconductor manufacturing process, estimation of the parameters and solution of a single dynamic program is not feasible due to these problems. Each of these problems might be soluble through the use of smaller, more manageable models focusing on segments of the process.

1.3.3 Semiconductor Manufacturing Data Collection and Storage

In a typical manufacturing plant, the semiconductor manufacturing data is recorded automatically from a number of machines in several distinct manufacturing areas. Assuming the process to operate on uniform discrete lots, typical production consists of about 25 lot starts/day flowing through about 270 operations over about 90 days. Using these estimates, one can see that 2250 lots of work in process accrue 6750 operations per day. However, this simplifying assumption of discrete lots is invalid due to batching and detailed processing; for example several lots are batched together for annealing in a furnace, while each wafer in a lot may have multiple die-shot reticle exposures. These aggregations and dis-aggregations complicate the understanding of the manufacturing process, and produce challenges in data collection. Short runs, rework, and process changes provide another source of complexity in semiconductor manufacturing as shown in Figure 1.2. These histograms of the process length assigned to different production lots show that there is not a consistent manufacturing process; the lower of the two histograms shows a spread of the number of check-in/check-out operations from 450 to 550, none of them more numerous than 12 lots. In terms of the mathematical model presented in 1.3.2, the length of the dynamic program is constantly changing, and identification of a consistent information matrix is impossible. Frequent process changes and manufacturing dispositions result in an amorphous process.

To expedite the collection of data, each machine operation is recorded in a transactional database whose structure mirrors the physical production and testing machinery. This optimizes data collection, but hinders data analy-

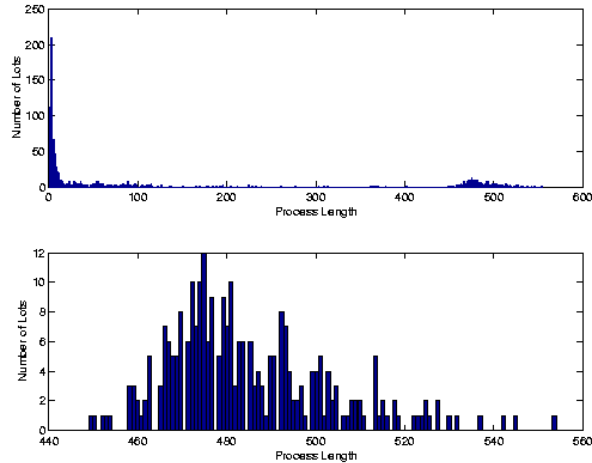


Figure 1.2: Process Length Variance – The count of the number of steps used to produce a lot of wafers shows significant variability in production process length. In the lower figure, notice the mode at approximately 475 process operations. In the upper histogram, note that there are large numbers of very short processes.

sis. Each machine is capable of emitting a number of different measurement records, and tables corresponding to each machine and each record type are automatically updated as production flows through the machines. While this data recording was initially driven by contractual agreements with the parent companies of the manufacturing plant, current efforts seek to use this data to improve the production process. The transactional database holds all the required data, but since the results are not aligned with the batch, lot, wafer, reticle shot, or chip, using the data for analysis is not possible without intricate database queries.

High dimensional data often requires reduction of the number of dimensions in order to build knowledge. Examples of data domains similar to semiconductor manufacturing data include high dimensional data from image analysis, radar and spectral data, text recognition and mining, speech recognition,

genetic code sequences, and chemometrics. Interpretation of high dimensional data is difficult, as is understanding of a high dimensional model. Several of these fields have underlying spatial or theoretical models on which to base further analysis. If the theory is lacking, however, then the use of data mining techniques to build models may help to develop theory about these complex domains. Semiconductor manufacturing differs from these domains in that the structure of complex manufacturing data is not well organized for analysis.

Examination of the semiconductor manufacturing process leads one to the question: What are appropriate methods to use large dimension, small sample manufacturing data for prediction and understanding complex manufacturing processes?

XXX

High dimensional complex systems present a number of unique challenges in the data analysis for understanding, prediction, and control. This work studies the top level system of a semiconductor manufacturing factory in order to build understanding of the effects of lower level processes on the upper level goals and objectives. Process engineers in the semiconductor industry often focus on the individual processes rather than the overall manufacturing system (Herrmann et al., 2000, p. 1491).

Simulation models of factories focus on measures of system performance such as total lot processing time, throughput, utilization, and work-in-process (Herrmann et al., 2000). However, high level models of *product performance characteristics*, such as yield, conformance to specifications, and quality, require understanding of the influences and contributions of the sub-systems outputs on the end product. For example, it is believed that quality characteristics in processing at the Gate Contact level can have significant effects on the

performance and yield characteristics of the memory chips. An understanding of this effect, and the effects of competing models is important for management decision making.

From a bottom-up perspective, manufacturing systems have embraced statistical process monitoring and control to identify and understand problem areas in manufacturing. Traditionally, these were univariate and tightly bound to the specific manufacturing process. Distinct or interrelated processes require more advanced techniques such as multivariate or multistream techniques (Montgomery, 1996, p. 235). As the measurement systems were integrated into the machinery, control methods were used to more accurately monitor and actively control the processes. In complex manufacturing processes, the necessity for control across distinct runs has inspired the run-to-run (R2R) control methodology, in which information external to the process is used to adjust the active controls in the process (Moyne et al., 2001). Optimizing each sub-process to minimize the variation in its outputs may indeed improve the overall product performance, but local optimization of subprocesses does not guarantee global optimization, especially if the local optimizations are not of the same variables. This research will build a framework for combining locally optimized processes into a global model of product performance characteristics.

Hierarchical control of manufacturing systems applied in the semiconductor industry is a new and active area of research Moyne et al. (2001, p. 9, p. 321). The work shown in Leang et al. (1996) is an application of the R2R control scheme to include several machines in a subprocess to form a single layer on a semiconductor wafer. This is an expansion of the single sub-process to several sub-processes and is the first step up from the bottom level of a bottom-up hierarchical model. On the other hand, simulation models of logistical char-

acteristics of a factory, for example, are the first step down from a top-down hierarchical model. Between these top-down and bottom-up methodologies is a need for understanding the relationship between low level sub-process characteristics and *product performance characteristics*.

Our conceptual approach is to use hierarchical modeling and data fusion as a means of addressing the problem of understanding the interrelationships between low-level process parameters and high level product performance characteristics.

The key characteristics of this data seem to be well studied, however, interactions between the characteristics outlined below combine to make the application of the common solutions difficult.

1. Inconsistent data due to:
 - (a) Frequent process changes, which produce a dynamic process structure.
 - (b) Missing data due to structure changes. E.g., use of a different machine produces a different set of variables
 - (c) Key punch errors due to manual entry erroneously report values against improper lots.
2. High dimensionality with respect to the number of observations (25 lots/day * 30 days/month = 2000 rows by 21000 columns)
3. Data aggregation issues (chip / reticle / wafer/ lot / batch)
4. Interpretation of multiple, competing process/product models

Although these characteristics have been studied separately, these issues combine to produce a problem that is incompletely understood.

1.4 Summary

In manufacturing systems such as factories, the high level performance characteristics have typically been work-flow and factory efficiency characteristics. The control of low-level processes has expanded to include information from prior and post-processes. An understanding of *product performance characteristics* based on sub-processes would be an advance in manufacturing and management of large systems.

In order to examine the relationships between low-level processes and high level system performance characteristics, a hierarchical model of process outputs based on sub-processes can provide support for decision making in allocating resources between sub-processes.

Although the problem domain is broad, the focus on a methodology for understanding the *product performance characteristics* as related to the process subcomponents is more tightly defined. This methodology is applicable beyond the factory supplying my data; large scale systems with heterogeneous sub-components have a need for understanding the effects of the sub-components on the entire system. Fowler et al. (2000)

Chapter 2

Literature Review

The literature supporting this research consists of several areas: modeling semiconductor manufacturing, sample size and model complexity, domains of similar complexity, and hierarchical modeling. Semiconductor manufacturing literature consists of high level yield modes and low level device performance models. Similar complexity domains exist in image analysis, bioinformatics, and chemometrics.

2.1 Semiconductor modeling

Yield modeling is a key goal of semiconductor manufacturers. Improvements in yield prediction and yield provide direct financial gain to manufacturers. Yield models depend primarily on the combination of defect rates from the various sub-processes to predict yield rates. Unfortunately, the defect rate information attributed to specific sub-process is expensive and slow to collect, so manufacturers are interested in methods of using the process and test data to predict physical characteristics, and further, defect rates.

Van Zant (1997) provides a good overview of the entire semiconductor manufacturing process, while Horton (1998) further explains a number of yield modeling techniques and formulas. Nurani et al. (1998) predicts yield based on defect density information drawn from multiple-layer inspection information using in-line pattern defect density information. Shindo et al. (1998) model the effect of defects on lower layers in a semiconductor sandwich on upper layers.

Hess and Weiland (1999) use a sampling plan across a wafer and lot to produce defect density distributions, which can help to model yield. The paper seems most applicable to predicting yield of one program based on the defect maps of another program, (e.g. 128MiB DRAM based on 64MiB DRAM).

Cunningham and MacKinnon (1998) show a number of defect characterization statistics in order to more fully understand low yield. The methods of quadrat statistics (defect per die), spatial point pattern statistics for spatial randomness, and spatial clustering, and collinearity identification are discussed. A fuller explanation of spatial clustering monitoring by Hansen et al. (1997) provides guidance in creating test statistics from a pick-map and using them to monitor wafers for spatial randomness. Friedman et al. (1997) explain a method for monitoring large area defects by separating a smoothed cluster from the underlying spatially random component.

Chaudry et al. (1998), working with SEMATECH, propose an object-oriented database to provide responsive control during the manufacturing process. This method recognizes the hierarchy of the semiconductor manufacturing process, but requires good models of inter-process interactions (i.e. detailed models of downstream features based on upstream parameters). Given these detailed models, an “active-database” could generate novel recipes as the process proceeds.

Richards and Shen (2000) develop a model of the physical characteristics of a semiconductor device based on some electrical test data. This is the inverse of the problem of predicting in-line electrical test results based on process parameters.

Fowler et al. (2000) surveys a number of modeling methods applied to semiconductor manufacturing data from the probe testing machine in a micromechanical accelerometer production process. The data is in several families, (i.e. min, max, ave, std, quartiles, and quartile range) of the several monitored variables. Although this work uses wafer-level semiconductor data, it uses data from only one step in the manufacturing process. They use some ad-hoc clustering to stratify the yields before applying some tools: On the set of low yield wafers, the models did not work well. Fowler et al. (2000) used PCA based regression to reduce the dimensionality, but the resulting model was poor, and interpretation is difficult. The model complexity is $128 = 23 \times 6$ parameters based on 1123 observations, which is about a 1:10 ratio. The micromechanical accelerometer in Fowler et al. (2000) may be well described by 23 features, but a 64MiB memory chip is a much more complicated device with many more elements.

Shin and Park (2000) discuss data volume in semiconductor manufacturing as 1,000,000 wafer transactions per day, this is consistent with what we see in other fabrication plants, 25 lots of 25 wafers starting and progressing through a 500 step process in 90 days, assuming three transactions per wafer per operation. They use a hybrid Neural Network with memory (implemented as a k-nearest neighbor vector). The k-NN approach helps the interpretation of the reasoning done by the Neural Net, and improves the straight neural net performance by adding the outputs of a k-Nearest Neighbor model to the inputs

of a neural net.

These papers have shown that yield modeling in semiconductor manufacture is an area of significant current interest, that attempts are being made to estimate physical characteristics based on easily measured electrical characteristics, and that yield modeling is still not perfectly understood. Opportunities exist for linking the various levels of semiconductor modeling to produce useful models of yield.

2.2 Complexity, Sample Size, and Modeling

Models generally require a sample size n of about 6-10 times the number of parameters p in the model.(Neter et al., 1996; Montgomery, 1996; Box et al., 1978). From this, one might assume a model of with dimensionality p of $\frac{1}{10}n$ to $\frac{1}{6}n$ might be a reasonable model. This relation between model data and model complexity reflects a limitation on the degrees of freedom and the number of estimable parameters in a model. Some domains, such as semiconductor manufacturing, image analysis, text mining, genomic data, and speech analysis, have small numbers of observations compared to the variables present. Managing the complexity of these complex domains requires methods which exploit the structure, theory, and knowledge of these fields. A method which exploits the hierarchy inherent in a manufacturing process can provide a tool to manage the complexity of the manufacturing data.

An examination of different types of models and the dimensionality of their data provides insight into the use of degrees of freedom in model building. Large data can be large a number of different ways depending on the shape, size, and storage requirements of data. Data with large storage requirements

can cause slow processing. Data with large numbers of observations can also impact processing speed. Data with large numbers of dimensions or variables with respect to the number of observations can cause problems with modeling through the estimation of the covariance of a data set.

A simple linear model $Y = \beta X$ which does not require a covariance matrix estimate, but only estimates of point values of the coefficients requires a β for each x in the model. Even if the β terms are zero, it requires a degree of freedom to estimate them as zero. This simple model requires at least p observations in order to make estimates of the process. A more rigorous linear model additionally estimates standard errors of the model coefficients in order to determine if the parameters are significant or not. These two sets of parameters imply that $2 \times p$ degrees of freedom are consumed by a simple modeling process.

Using the simple Hotelling multivariate process monitor, T^2 for example, requires estimating p^2 covariance and p mean coefficients. The number of degrees of freedom consumed by these estimates exceed those of linear models by including covariance terms between each pair of variables.

The small sample sizes available over a time span of interest provide for only small models relative to the potential dimensionality of the problem. In order to provide useful models with only a limited number of observations, the models should be limited to a complexity smaller than the number of observations. Complexity in this sense is the number of parameters in the model.

A very small model of the manufacturing process might estimate two terms representing an index of a process step and its effect on yield. For example: Process yield is 0.75 plus some factor times the anneal temperature in step 255. Estimating confidence intervals of the intercept and factor would consume four

observations, leaving the rest of the observations to estimate the uncertainty in the predicted yield. The problem with models like these is that there are a great number of competing models, and the uncertainty in model parameters nearly guarantees acceptance of invalid models.

More data would be consumed to validate models, and to choose between competing models Kennedy and et al. (1998). As an extension to the general wisdom of a sample size of 6-10 times the complexity of the dataset, He and Shau (2000) establish bounds on the increase of complexity of a model as the sample size increases. Their limits are based on the types of functions being estimated (i.e. linear and logistic regressions and a spatial median), and the continuity of the functions. Discontinuous functions can support less complexity on the same data, while increasingly large samples can support more complex models, but not at the rate of increase of the sample size (e.g, a sample of 100 points supporting a 10 term model would better satisfy asymptotic assumptions than a 1000 samples of a 100 term model). Under one reference they cite, a linear model without discontinuities would support only about 3 times as many terms with 10 times as many samples: 31 terms on 1000 samples is similar to 10 terms on 100 samples.

2.3 High Dimensional Modeling and Visualization

Understanding hyperdimensional ($p > 3$) datasets is difficult; although many automated systems exist for fitting multivariate models, comparing and understanding multidimensional models is challenging. Although model fitness

can often be reduced to a single score, the relevance and meaning of the model inputs is often uninterpretable. Visualization tools for high dimensional data are essential for understanding the data and models (Tufté, 1983). McLeod and Provost (2001) survey multivariate visualization software and suggest that quantitative programming environments such as MatLab, Mathematica, and S/S-Plus/R provide powerful, flexible, and reproducible graphics and explorations. Examples of visualization used to confirm and explore multivariate modeling are shown in Wilhelm et al. (2001). Essentially, the problem of visualization of a hyperdimensional dataset is accomplished in three ways: selection of small subsets of the original variables, transformation to a small dimensional space, or decomposition of the problem into several less complex sub-problems.

Huffer and Park (2000) show a test for structure, basically by removing the first and second moments in the data, then studying the multivariate distribution with a chi-squared test. This leads into other methods of high dimensional visualization, such as the Sliced Inverse Regression (SIR) (Li, 2001). SIR, which bins the output variable, calculates the corresponding means and covariance of the input variables, reduces the dimension of the predictors, and then examines the output variables in the reduced space (Basilevsky, 1994).

2.4 Reduction of dimensionality

Liu and Motoda (1998) consider algorithms for reducing the dimensionality of an input data space, distinguishing between different transformations of features used for data mining. Feature subset selection chooses a reduced set of the original variables (e.g. stepwise forward selection). Feature extraction pro-

duces a new set of variables as a simple function of the original variables (e.g.: principal components analysis). Feature construction creates new variables, (e.g. $power = voltage \times amperage$).

Several feature extraction methods, such as Principal Component Analysis (PCA), Singular Value Decomposition, (SVD), Factor Analysis, and Partial Least Squares (PLS), can be used to code the original variables in a smaller dimensional space. PCA produces a set of uncorrelated linear combinations of the initial variables ranked by their contributions to the overall variance (Johnson and Wichern, 1992). Each PCA component includes each of the original variables, encoded in the associated eigenvector. Singular value analysis, is a method of characterizing a data matrix of less than full rank with eigenvalues, eigenvectors, and an orthonormal basis matrix Basilevsky (1994). Alter (2000) uses SVD to reduce the dimensionality of a high dimensional gene data ($\{n \times p\} = \{14 \times 5981\}$) to a smaller space of ‘eigengenes’. These feature extraction techniques map the high dimensional data into a different space, then truncate the dimension of the new space into a smaller dimension.

In contrast to feature extraction methods are feature selection methods that attempt to choose a subset of the initial variables while maintaining the information required to reliably model the process. Feature selection methods choose and exclude variables from an analysis based on some measure of relevance. Methods of the variable subset selection include nested model methods such as backwards elimination or forward selection in regression using changes in model R^2 ; decision trees such as C4.5 or CART that use an information measure to rank variables, and manual methods using expert advice from domain experts.

Bocchieri and Wilpon (1993) discuss the addition and elimination of fea-

tures in a speech recognition problem. As equipment becomes faster, new features and higher order transformations of the original features become available. The new speech recognition variables can improve the accuracy of speech recognition algorithms, but the computational complexity of the algorithms becomes an issue. Bocchieri and Wilpon (1993) suggest a method for limiting the number of features based on a misclassification distance in each of the dimensions. John et al. (1994) suggest a elimination of irrelevant variables using a “wrapper” technique based on stepwise selection or elimination of features and applying the data mining technique to each of the subsets. John et al. (1994) develop a definition of weak relevance based on conditional dependence on a subset of variables. Hall and Holmes (2000) compare several methods of attribute selection and suggest information gain and a correlation based method for high dimensional data. Hall (2000) develops a correlation based feature selection method as a heuristic search of all subsets of features. Wu and Urpani (1999) suggest eliminating the least relevant features rather than selecting the most relevant in order to handle messy data. Liu and Setiono (2001) propose several random search methods for selecting subsets from high dimensional data.

2.5 Alternate high-dimensional domains and approaches

Image analysis, text mining, speech recognition, spectral analysis, and bioinformatics databases are commonly stored in formats tailored to the specifics of the data and the processing algorithms. Large databases of high dimensional

data such as images, text, or speech recognition require dimension reduction to produce summaries and indexes suitable for using this data. Much work has been done on the transformation algorithms in these specialized fields, but the software is often single use or proprietary to the specific application. Bioinformatics data consists of high dimensional genetic information on a limited number of samples, with relatively few genes in the genome responsible for a particular phenomenon, possibly similar to a small number of semiconductor processing parameters being responsible for a particular defect. Spectral analysis of chemical mixtures uses high dimensional spectrogram data to estimate low dimensional mixtures of compounds. Although work has been done in many high-dimensional systems, the single-purpose and tailored systems are not directly applicable to semiconductor manufacturing because of the complexity of the data structure.

Feature extraction methods are often used to summarize and index high dimensional databases for similarity searches. Aslandogan and Yu (1999) survey several systems for image storage and retrieval. Dimensional reduction of color or spectral histograms, and texture signatures derived from fourier transforms of the images are also discussed.

Each semiconductor chip, wafer, lot, or batch carries a large number of independent process variables and characteristic measurements which may differ with each chip/wafer/lot/batch. Image analysis contains a large numbers of pixels and their associated characteristics which may differ for each image. For example, digital cameras routinely produce 1.3 megapixel images, reducing these to a simple greyscale image of 100x200 pixel by 8 bit depth for internet web presentation produces an array of pixels containing 20000 variables with 256 levels. Dimensional reduction is a strategy of creating summary or sig-

nature features (or variables) that may give an analyst a better perspective than a pixel-by-pixel representation. For example, an analyst could query the database for images that are ‘green’, or in manufacturing, creating indices by lot number, by process and by yield. This facilitates similarity signature modeling in that an analyst could request all of the lots similar in yield to lot X, Y, and Z for example. In addition, a practical consideration of indexing strategies is that the number of attributes or fields in relational tables is limited. For example, the commercial database program, Oracle 7, limits the number of attributes to 256 in a table. Ng and Tam (1999) use a multiple level filter to manage high dimensional data. They find that color, texture and ‘eigen-face’ representations of image data may generate 256, 240, or 400 dimensions, respectively. Compared to the original dimensionality of the image data, the reduction is dramatic, but search through an index with > 20 dimensions essentially degenerates into a sequential search. Ng and Tam (1999) solve this problem with a system for storing a multidimensional index in a hierarchy based on transformed features.

2.5.1 Chemometrics

Frank and Friedman (1993) explain some chemometric tools for prediction, such as ridge regression, partial least squares, and principal components regression, in domains where the number of variables far exceeds the number of observations. For example, spectrum analyses use high dimensional data ($p > 1000$) to estimate fractions of chemicals in mixtures.

Gene expression data is a domain with small sample sizes and large domain in which a large number of the variables are irrelevant to the problem

of interest. Eisen et al. (1998) describes data of $n \approx 10^2, p \approx 10^5$ and a method for clustering variables based on a correlation coefficient using average linkage, then displaying them for human interpretation. The general model using gene arrays is to take samples of the biological manifestation, and then compare gene arrays in order to identify genes that are related to the question of interest. Kamimura et al. (2000) use mean hypothesis testing, checking for statistical significance of difference between variables given classes of outputs to determine relevance of genetic information to a problem of interest.

2.5.2 Text Mining

Mining of text databases for relevance and indexing is a high dimensional system, with authors, sources, lengths, ages, names, and words forming a complex parameter set that causes difficulty in extracting reliable information. Yang and Pedersen (2001) compare a number of feature subset selection methods applied to a text mining and propose a method similar in performance to χ^2 and information gain procedures. Ahonen et al. (2001) consider text mining as a time series through multiple sensors and find it tractable using rule discovery techniques. In text mining, Dumais et al. (1998) consider information gain to be an effective feature selection method for the high dimensional problems. Li et al. (1999) use clustering to reduce the dimensionality of a text and image documents for more effective searches. Liu and Setiono (1998) propose a scalable Las Vegas algorithm for selecting subsets of features for data sets with large numbers of features and large number of observations.

Chapter 3

Application and Methodology

The problem of using high-dimensional and low observation complex manufacturing data for prediction and understanding requires a method for managing the complexity. Since complex systems do not often generate data as a simple data matrix, but instead as a set of subsystem-specific distinct datasets, common methods such as standard regression, feature selection, and feature extraction are not easily applied. Changes in the structure of the holistic system produce missing data and make construction of a full rank data matrix impossible. Modeling a complex system using a hierarchy congruent to the component subsystems process can provide a method for managing the complexity of the entire system system. By using the “natural” divisions in the process and data collection and storage system, the entire system can be broken down into sub-components, examined and modeled separately, and recombined in a hierarchy of interlocking models. The smaller sub-models, that still may have problems of dimensionality, irrelevant data and missing data, but will be more tractable than the overall model.

3.1 Application

3.2 Methodology

3.2.1 Hierarchical Modeling

Hierarchical modeling, as used here, builds submodels and meta-models to help develop an understanding of the relationships between components in a complex system. The three key elements to hierarchical modeling is the identification of subcomponents, the identification of the connections or communications between subcomponents, and the unit of interest. In manufacturing systems, the separate manufacturing processes are clear candidates for subcomponents of the system. For observing connections between subsystems, either high level system outputs or low level subprocess outputs may be useful. The unit of interest is the common granularity of the process.

Subcomponents

A system is a set of interrelated subsystems working together to provide for a common goal. ? A multiple step manufacturing process provides clear subsystems. Each of which consists of a inputs and outputs and some sort of functional relationship between the two. Subcomponents may overlap and interrelate: for example, the heating, ventilation, and air conditioning system may affect and be affected by many other systems operating in a manufacturing plant. Identification of a set of subcomponents helps to understand the effects of the subcomponents on the high level system. Without grouping the elements into subsystems or subcomponents, the output process can only be understood

as a function of all of the possible input variables. Subdividing the system into subcomponents can help manage the complexity.

Identification of subcomponents can come from examining the form of the data, the division of the people studying the problem, the models applied to the data, and the reports used to examine the system.

In the semiconductor manufacturing plant under study, data is stored along functional or departmental lines. Manufacturing data is collected automatically for each machine and stored in a large database with a table for each machine. Testing measurements are recorded by a different department in a different database. Process engineers use a summary report of historically critical data to monitor the process. Quality control engineers use a different set of databases to monitor the processes. Production managers use routing and output data to monitor the efficiency of the plant. Each of these functional groups produces, stores, accesses and analyzes data drawn from portions of the entire process, draws conclusions, and makes decisions based on their mental models of the process. Considering each of these systems as a subcomponent of the larger production system raises a question of how to combine and understand these models in relation to one another.

If a large system can be decomposed into meaningful subunits, understanding may be gained by examining the relationships of subunits to the high level system outputs and between subunits. If users of the system conceive of the system as a agglomeration of subsystems, a modeling methodology that explains the interdependencies and effects of the subcomponents may help build understanding of the system.

Intermediate variables

Two obvious selections of intermediate variables are first, the measurable outputs of the subsystems, and second, an estimate of the meta-system output. Using measurable process outputs is a bottom up approach where the process outputs are assumed to be a significant contributor to the meta-process outputs. Estimating the meta-level output directly from the subprocess data is an alternate method, that provides for finding novel relationships between the low level process variables and the high level system outputs.

Measurable process outputs If the intermediate monitoring variables are used as the sole predictor of the higher level system output, the resulting predictions can only be as good as the best prediction based on the intermediate variables. Several studies have identified an increase in US citizen's heights with an increase in nutrition; using height as a sole surrogate variable to summarize the nutritional variables in prediction of health will limit the results to the information present in the height variable. The intermediate variables can be used, but they should be augmented with other information to produce an improved estimate.

End result variables

3.2.2 Meta-Analysis applied to high dimensional complex systems

To apply meta-analysis to a high dimensional problem requires identification of the variable of interest and the subcomponents of a system. Basically, the

meta-analytical model is a two level system comprised of several submodels each estimating the variable of interest, and the high-level metamodel which combines the estimates from the subsystems to produce an overall estimate of the variable of interest.

3.2.3 The variables of interest

In a complex system composed of many subcomponents working together to produce a common goal, the indices of performance towards that goal are clear variables of interest or outputs from the system. ? For ease in combining and understanding the interrelationships between the various estimates of the variable of interest, clear understanding of the variable is important.

In the specific case of manufacturing, process yield is an important measure of the performance of a manufacturing system. Alternate important measures might be time expended or resources used. From the standpoint of the process engineer attempting to understand yield, a focus on defects and their root causes points towards defect rate as a meaningful measure of process performance. If, for a particular defect, the defect rate can be understood and attributed to root causes, then the path toward improving the process or reducing the defect rate can more clearly be found.

Risk, Odds, Probit, and Logit

Yields, success rates, and failure rates are all examples of probabilities. Risks are a ratio of ($\frac{\text{failures}}{\text{successes}+\text{failures}}$) and have the domain $[0, 1]$. Two methods of characterizing process losses are an odds and as risk. Odds of failure ($\frac{\text{failures}}{\text{successes}}$) are a simple ratio of failures to successes and lie within the domain $(0, +\infty)$.

In a manufacturing process, risks or failure rate might be more commonly used, but models with $(-\infty, +\infty)$ provide more tractable models.

Random Regression Model of log risk defect

$$Y_{risk\ i} = \hat{Y}_{i1} + \hat{Y}_{i2} + \dots + \hat{Y}_{ip} + e_i$$

3.3 Proposed Model

3.3.1 Subcomponents as a risk factor

Risk, defect rate, and odds are functions of each other

Baseline risk is the overall, or intercept term.

Benefits of combined submodels

Equally contributing components will individually have low predictive power, but taken together will have stronger predictive power.

Supposing that the estimate of the defect rate due to some submodel M_i $\hat{y}_i = f_i(\vec{\beta}_i, \vec{X}_i)$ where \vec{X}_i is a vector of variables in submodel i and $\vec{\beta}_i$ is a vector of parameters, and $f_i(\cdot)$ is an estimation model, the model \hat{y}_i will be able to predict the system output with some level of accuracy. Supposing further that $i = 10$ subsystems are independent and that each is responsible for $1/20$ of the variance observed in the output, with the remaining $1/2$ of the variance due to other sources. Each of the 10 models can explain at most 5% of the variation, and may seem insignificant, but a combination of the subsystems has the potential to explain $1/2$ of the variance in the entire system.

If the entire system is small enough to be modeled in one system, that would assuredly produce better prediction results, however, segmentation into sub-components can perhaps aid in the interpretation of the resulting model.

Interdependencies between the subsystems can be reflected in the correlations between the subsystem outputs.

3.3.2 Comparison to intermediate surrogate variables

3.3.3 (E.g. TEG as an intermediary)

3.3.4 Continuity Adjustment ($\pm e$)

The existence of observations with 100% failure or 100% success poses problems for logistic or probit modeling. Since in transforming these values into Logit space, $\ln(-\ln(p))$, the extreme values $\ln(0)$, $\ln(-\ln(1))$ are undefined, or in probit space, $z = \Phi p$, the extreme values $\Phi^{-1}(0)$, or $\Phi^{-1}(1)$ are also undefined, these discrete probabilities are adjusted inward to the

Continuity of x/n at $x = n$ and $x = 0$: for the discrete case:

$$\begin{aligned}
 \text{if } x = 0 & : p = (0.5)/n \\
 0 < x < n & : p = x/n \\
 x = n & : p = (n - 0.5)/n
 \end{aligned} \tag{3.1}$$

This is centered, and invertable. The assumption is testable by checking for equal proportions:

$$H_0 : p_{sample}(x = 0) = p_{theo}(x < 0.5/n) \tag{3.2}$$

$$H_a : p_{sample}(x = 0) <> p_{theo}(x < 0.5/n) \tag{3.3}$$

If we cannot reject this null hypothesis, then the adjustment is reasonable. Alternately, if the adjustment is unreasonable, the underlying problem is a mixture of two processes: One with the base distribution, and another process which generates many of the $x = 0$ cases.

3.3.5 Combination of models and relative risks

3.3.6 Logit and Probit models

3.4 Research Plan

The problem of using high-dimensional and low observation complex manufacturing data for prediction and understanding requires a method for managing the complexity. Since the data is not stored in a simple data matrix, but instead in a set of process-related distinct databases, standard regression, feature selection, and feature extraction methods are not easily applied. Process changes produce missing data and make construction of a full rank data matrix impossible. Modeling the semiconductor manufacturing system using a hierarchy congruent to the manufacturing and engineering process may provide a method for managing the complexity of the manufacturing system. By using the “natural” divisions in the process and data collection and storage system, the entire system can be broken down into subcomponents, examined and modeled separately, and recombined in a hierarchy of interlocking models. The smaller sub-models, that still may have problems of dimensionality, irrelevant data and missing data, will be more tractable than the overall model.

3.5 Intermediate Variable Model

The intermediate variable model uses system variables identified as important by process experts as intermediaries between the subprocesses and the high-level process. In contrast to looking for the impact of each of the approximately 21000 variables on the outputs of interest, a hierarchical model using the engineering variables of interest could provide a way to manage the problem of dimensionality and lack of degrees of freedom.

The submodels XXX

Two internal reports produced at the manufacturer summarize production and testing data: the TEG electrical characterization data through WIPNavigator, and the SIView quality control and process data reports. These two reports represent variables of special interest to the process and design engineers. The internal reports presents the variables in these datasets with scatter plots against yield and other output variables, encouraging univariate models. Discussions with process and design engineers help to build the relationships in Figure A.1. Combining these variables with analysis is a first step in creating a model of yield or other process outputs. Each of these variables could be used as response variables for subsets of the process data. The resulting hierarchical model could use the process data to predict the engineering variables, and then in turn predict the yield response. Although the global model uses the same data, the structure imposed by the constraints reduces the number of parameters estimates required by a multivariate model, and helps manage the short run and small sample sizes.

Specifically, DS contains 27 lot level TEG contains 71 measurements, each a vector of lot summary statistics, monitoring Prime Specification Limits (PSLs)

thought important by process engineers. QC contains 21710 variables separated into measurement and process data. Quality control and process engineers have selected 31 of the QC measurement variables as key Engineering Specifications (ES), each a vector of several lot summary statistics and wafer samples.

Modeling the PSL variables as outputs of lower level processes, such as the ES, and the ES parameters as outputs from QC measurement and test data divides the manufacturing system into “natural” subcomponents. Also, modeling the yield performances in the die sort database as functions of the TEG PSLs may aid in assessing the relative benefits of measuring the PSLs.

Figure A.1 shows a number of modeling efforts that may succeed in relating the low level production data to the high level yield and efficiency data. Each connecting line in figure A.1 represents a smaller, more manageable model, that may provide insight into the wafer production process. For instance, at the bottom of the figure, each of the several dotted lines link some subset of production data to an engineering specification. It may be possible to discover and control the process parameters contributing to the TV nitride measure using partial least squares or other regression techniques to develop lower and intermediate models. By monitoring and controlling the engineering specifications, it should be possible to understand, predict, and control of some characteristics in electrical test and die sort.

Extraction of some die sort and engineering test data from a production system has already been accomplished. What remains is extracting production and measurement information, clarifying the hierarchy between elements of these data, building some of the many possible sub-models, and showing that the sub-models can be combined in a hierarchical model.

Referring to Figure A.1, I will build a limited number of models at different levels of the process, relating the outputs of lower level processes to higher level variables of interest. Specifically, data will be cleaned and feature-selected from a database table, aligned with higher level output data, and models will be built using methods such as multiple regression, principle components regression and partial least squares. The smaller models may be amenable to linear regression, general linear models or partial least squares, depending on dimensionality and the output variables. The organizational hierarchy of the existing data, shown in figure A.1, allows for testing of the sub-models. Relating the elements of the production hierarchy this way should demonstrate that although a simple multiple regression on the process parameters to the yield output is not possible, a detailed overall model of the production process may be successful.

At this writing, domains of similar complexity have been explored in the literature, data has been extracted, preliminary models of die sort variables as a function using in-line testing data have been attempted. I have used SAS summary data, visualization tools, and elicited expert knowledge to manually select relevant feature subsets, and built models of in-line-testing data.

Separating the production margin yields into the contributors to margin yield failures allows modeling of the different defect rates separately, and can provide a more detailed explanation of the correlations between electrical test and die sort data.

Preliminary models on limited data indicate that ordinary least squares models of these separate yield failures are stronger than models of the overall yield. XXX

3.6 Initial Approach and Preliminary Results

3.7 Results and Summary of Preliminary Approach

The organization of the data at the semiconductor manufacturer has not supported easy data analysis. The internal tools, SIView and WIPNavigator, produce a large number of graphs and scatter plots which results in a number of competing ad hoc univariate models of production yield problems based on single variables and processes. By building a few multivariate models and linking them together in a hierarchical model that mirrors the production organization, I hope to provide a mechanism for combining models from the entire process. Our work in this project has inspired the manufacturer to build new data extraction tools with improved query performance, and larger, more relevant data sets are now available. The production and quality control data (QC) are now available in a form which can be aligned with the in-line testing (TEG) and die sort (DS) data.

Multivariate models including the entire process are impractical due to data structure problems. By developing a framework for relating smaller models in a hierarchical manner, a complex manufacturing process can be broken down into more manageable smaller models which can be combined to produce an overall model.

Opportunities exist to build a number of sub-models relating process parameters to key engineering parameters, key engineering parameters to electrical test data, and electrical test data to die sort and yield data. Feature

selection and feature selection methods can help to reduce the dimensionality of the sub-models in cases where the sub-models have rank problems. Combining these models to build a hierarchical model could increase understanding of the semiconductor manufacturing process. Although the combined model could represent the entire process, it is hoped that the sub-models and overall model will enhance understanding of the process as compared to an under-determined overall model and the many parallel univariate models. Some of the sub-models may prove useful and provide avenues for improvements of the process.

The extraction of useful information from the large storage space and high dimensional databases of semiconductor manufacturing may provide substantial benefits in yield modeling. Using a methodology which captures the hierarchical nature of the semiconductor manufacturing process can provide a method for managing short-run, data-poor complex processes. By segmenting the model into discrete units, multiple less-complex models can be created from the available data. Ultimately, the development of a hierarchical model can help manage the complexity of the semiconductor process and relate high-level yield information and electrical parameters to specific manufacturing process parameters. The methodology used to produce and combine these models into a comprehensive model of the semiconductor manufacturing process will be applicable to other complex manufacturing processes, and also to other complex systems with $n \ll p$.

Chapter 4

Results

This chapter presents the results of the methodology. The results are good, showing that there exist significant models of the subsystems and that the metamodel can combine these submodels into a model better than any of the component submodels.

4.1 Description of specific problem in Semiconductor Manufacturing

4.2 Results of the method

4.2.1 Submodels – subprocess estimates of yield

4.2.2 Metamodel – yield/defect rate as a function of subprocess estimates

4.3 Evaluation

4.3.1 Competing models

(comparison to a $p < n$ full regression model)

4.3.2 Effectiveness

4.4 Recommendations

4.5 Summary

Chapter 5

Discussion

This chapter generalizes the results of the methodology to to apply to systems beyond the example developed previously. Although the example developed here was applied to a specific semiconductor manufacturing system, the framework can be applied to other manufacturing systems and to other complex systems.

Two reasons for using the decomposition and metamodeling approach are for systems with more variables than observations, and for understanding of complex systems and the relationships between subcomponents.

5.1 Applicability in situations where $p \gg n$

Many realms of large data where the number of variables can $p \gg n$ data exist, as noted in the literature review. Also a relatively small number of variables can quickly balloon to a large number. An examples of this is a single nominal variable of n options in the initial dataset being represented by a set of $n - 1$ indicator variables. A submodel containing the new indicator variables

as a group can help to build understanding of the nominal variable distinct from the other variables in the analysis.

5.2 Applicability in systems of subcomponents

In systems with clearly distinct subsystems, a meta modeling approach can help to understand the subsystems as components. Metamodeling, in the form of Structural Equation Modeling (SEM) has long been used for this purpose, but on small systems of variables with few components to confirm theories of explanatory latent variables. The generalization here, is that for large systems without strong theories for latent variables, estimates of the system outputs can serve as the intermediate variables. By this means, some system level understanding of

5.3 Recommendations

5.4 Concerns

Concerns with the methodology arise two areas: loss of information through the simplified structure of the metamodel, and a contamination of the metamodel inputs through the submodel estimation process.

The information loss is a real concern, and is apparent from a comparison of a partitioned flat model and a hierarchical model. In the flat model, the sub-model covariance matrices along the diagonal are represented in the sub-models, while the between sub-model covariances are represented as the off-diagonal elements in the metamodel covariance matrix. Since the information

is used in a flat model, but is not used in this meta-modeling methodology, loss of information indicates that the methodology should not be used if a flat model is adequate. However, if a flat model is infeasible for some reason, e.g.: $p \gg n$, it is not a fair comparison. An argument for using a meta-model even if an alternative flat model exists, is that it may help in interpretation: If there is a clear hierarchy to the entire system, intermediate variables may help explain the interaction between and the contributions of the subsystems to the process output.

The contamination of the meta-model inputs by the submodel estimation process points to a clear methodological solution: use a separate training set for the sub-model fitting and for the meta-model fitting. If the sub-models and meta-model are trained on the same data set, then from (5.13) we see that \widehat{Y}_1 is a function of Y .

5.4.1 Interactions between variables in submodels

One concern with the disaggregation of data into subcomponents is the possible loss of interaction between the variables in the subcomponents. The loss of significant interaction terms is possible if the model does not specifically include them. However, since the number of interaction terms is of order $O(n^2)$, it may be infeasible to estimate the interaction terms from the un-segmented data. The segmentation and recombination allows more possibility for interactions both within and between submodels than the unsegmented case.

5.4.2 Interaction terms between submodels

Suppose the true model is $Y = \beta_{x_1}x_1 + \beta_{x_2}x_2 + \beta_{x_{12}} + \epsilon$ and x_1 and x_2 are separated into two subsystems with the addition of noise variables d . If the submodels are adequate, they will each be functions of the respective x variable, and the interactions can be modeled in the top-level model:

$$\hat{B} = (X^T X)^{-1}(X^T Y) \quad (5.1)$$

$$\hat{Y} = \hat{B}X = (X^T X)^{-1}(X^T Y)X \quad (5.2)$$

$$\hat{Y}_1 = \hat{B}_{m1}x_1 \quad (5.3)$$

$$\hat{Y}_2 = \hat{B}_{m2}x_2 \quad (5.4)$$

$$\hat{Y} = \hat{B}_1\hat{Y}_1 + \hat{B}_2\hat{Y}_2 + \hat{B}_{12}\hat{Y}_1\hat{Y}_2 \quad (5.5)$$

$$= \quad (5.6)$$

The dimensionality of the interactions are reduced however, by channeling the

5.5 Correlations between variables

One concern with model segmentation is that correlations between variables will be lost. Within submodel correlations will be maintained in each submodel, while between submodel correlations will be summarized into a one term per pair of submodels.

Partitioning of a flat linear model

A flat linear model can be examined through the covariance matrix. Partitioning the component variables into subsets, $\mathbf{X}_a, \dots, \mathbf{X}_n$ (5.7), and examining the structure of the covariance matrix (5.8) helps to understand the structure of the interrelationships between the submodels.

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_a \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \quad (5.7)$$

$$\text{cov}[\mathbf{X}] = \begin{bmatrix} \text{var}[X_a] & \text{cov}[X_a, X_b] & \cdots & \text{cov}[X_a, X_n] \\ \text{cov}[X_a, X_b]^t & \text{var}[X_b] & \cdots & \text{cov}[X_b, X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}[X_a, X_n]^t & \text{cov}[X_b, X_n]^t & \cdots & \text{var}[X_n] \end{bmatrix} \quad (5.8)$$

$$\widehat{\mathbf{Y}}_i = f_i(\underline{\mathbf{X}}_i) \quad (5.9)$$

If a submodel (5.9) is generated for each partition i (5.10), the covariances of the $\widehat{\mathbf{Y}}$ estimates from the submodels shows the structure of the meta model.

$$\mathbf{Y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} \quad (5.10)$$

$$var[\widehat{\mathbf{Y}}] = \begin{bmatrix} \sigma_{\widehat{\mathbf{Y}}_1}^2 & \sigma_{\widehat{\mathbf{Y}}_1\widehat{\mathbf{Y}}_2} & \cdots & \sigma_{\widehat{\mathbf{Y}}_1\widehat{\mathbf{Y}}_n} \\ \sigma_{\widehat{\mathbf{Y}}_2\widehat{\mathbf{Y}}_1} & \sigma_{\widehat{\mathbf{Y}}_2}^2 & \cdots & \sigma_{\widehat{\mathbf{Y}}_2\widehat{\mathbf{Y}}_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\widehat{\mathbf{Y}}_n\widehat{\mathbf{Y}}_1} & \sigma_{\widehat{\mathbf{Y}}_n\widehat{\mathbf{Y}}_2} & \cdots & \sigma_{\widehat{\mathbf{Y}}_n}^2 \end{bmatrix} \quad (5.11)$$

$$\widehat{\mathbf{Y}}_1 = \widehat{\beta}_1 \underline{\mathbf{X}}_1 \quad (5.12)$$

$$= (\underline{\mathbf{X}}_1^t \underline{\mathbf{X}}_1)^{-1} \underline{\mathbf{X}}_1^t \mathbf{Y} \quad (5.13)$$

Assuming the form of the submodel (5.9) is a linear least squares fit, $\widehat{\mathbf{Y}}$ on $\underline{\mathbf{X}}$, where each is centered with mean zero, the least squares model gives $\widehat{\mathbf{Y}} = \underline{\mathbf{X}}\widehat{\beta}$ where $\widehat{\beta} = (\underline{\mathbf{X}}^t \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^t \mathbf{Y}$ then the covariance of $\widehat{\mathbf{Y}}$ is $\sigma_{\widehat{\mathbf{Y}}}^2 = (\widehat{\mathbf{Y}}^t \widehat{\mathbf{Y}})/(n-1)$ which is:

$$\begin{aligned} \sigma_{\widehat{\mathbf{Y}}}^2 &= (\widehat{\mathbf{Y}}^t \widehat{\mathbf{Y}})/(n-1) \\ &= (\underline{\mathbf{X}}\widehat{\beta})^t (\underline{\mathbf{X}}\widehat{\beta})/(n-1) \\ &= \widehat{\beta}^t \underline{\mathbf{X}}^t \underline{\mathbf{X}} \widehat{\beta}/(n-1) \\ &= ((\underline{\mathbf{X}}^t \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^t \mathbf{Y})^t (\underline{\mathbf{X}}^t \underline{\mathbf{X}}) (\underline{\mathbf{X}}^t \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^t \mathbf{Y}/(n-1) \\ &= \mathbf{Y}^t ((\underline{\mathbf{X}}^t \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^t)^t \underline{\mathbf{X}}^t \mathbf{Y}/(n-1) \\ &= \mathbf{Y}^t \underline{\mathbf{X}} ((\underline{\mathbf{X}}^t \underline{\mathbf{X}})^{-1})^t \underline{\mathbf{X}}^t \mathbf{Y}/(n-1) \\ &= \mathbf{Y}^t \underline{\mathbf{X}} (\underline{\mathbf{X}}^t \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^t \mathbf{Y}/(n-1) \\ &= \mathbf{Y}^t \underline{\mathbf{H}} \mathbf{Y}/(n-1) \end{aligned}$$

where $\underline{\mathbf{X}}(\underline{\mathbf{X}}^t \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^t = \underline{\mathbf{H}}$, the $\underline{\mathbf{H}}$ matrix

When examining the covariance of two models, the development is similar,

but the simplification is not as clear:

$$\begin{aligned}
 \sigma_{\widehat{\mathbf{Y}}_1 \widehat{\mathbf{Y}}_2} &= (\widehat{\mathbf{Y}}_1^t \widehat{\mathbf{Y}}_2) / (n - 1) \\
 &= (\underline{\mathbf{X}}_1 \widehat{\beta}_1)^t (\underline{\mathbf{X}}_2 \widehat{\beta}_2) / (n - 1) \\
 &= \widehat{\beta}_1^t \underline{\mathbf{X}}_1^t \underline{\mathbf{X}}_2 \widehat{\beta}_2 / (n - 1)
 \end{aligned}$$

the simplification of the model is to replace the large covariance matrix of the flat model, 5.8 with the summary structure of the covariance matrix of the metamodel 5.11.

Assuming linear models, the each element of the meta model covariance matrix (5.11) is generated as:

$$\text{if } i = j \quad : \quad \sigma_i^2 = \mathbf{Y}^t \underline{\mathbf{H}}_i \mathbf{Y} / (n - 1) \tag{5.14}$$

$$\text{if } i \neq j \quad : \quad \sigma_{ij} = \widehat{\beta}_i^t \underline{\mathbf{X}}_i^t \underline{\mathbf{X}}_j \widehat{\beta}_j / (n - 1) \tag{5.15}$$

5.6 Other applications

Chapter 6

Conclusion

A conclusion and summary of the work performed.

6.1 Overview

6.2 Results

6.3 Summary

Bibliography

- Ahonen, H., Heinonen, O., Klemettinen, M., and Verkamo, A. I. (2001). Applying data mining techniques for descriptive phrase extraction in digital document collections. *Not Published*.
- Alter, O. (2000). Singular value decomposition for genome-wide expression data processing and modeling. Technical report, Uxxxx. Mathematical and Computational Challenges in Molecular and Cell Biology, Berkley, California.
- Aslandogan, Y. A. and Yu, C. T. (1999). Techniques and systems for image and video retrieval. *Knowledge and Data Engineering*, 11(1):56–63.
- Basilevsky, A. (1994). *Statistical Factor Analysis and related methods: theory and applications*. John Wiley and Sons.
- Bocchieri, E. L. and Wilpon, J. G. (1993). Discriminative feature selection for speech recognition. *Computer Speech and Language*, 7:229–246.
- Box, G., Hunter, W., and Hunter, J. (1978). *Statistics for Experimenters*. New York: John Wiley & Sons, Inc.
- Chaudry, N., Moyne, J., and Rundenstiener, A. (1998). Active controller: Utilizing active databases for implementing multistep control of semiconductor

- manufacturing. *IEEE Transactions on Components, Packaging and Manufacturing Technology–Part C*, 21(3).
- Cunningham, S. P. and MacKinnon, S. (1998). Statistical methods for visual defect methodology. *IEEE Transactions on Semiconductor Manufacturing*, 11(1):48–53.
- Dumais, S. T., Platt, J., Heckerman, D., and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of ACM-CIKM98*, pages 148–155.
- Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. In *Proceedings of the National Academy of Sciences*, volume 95, pages 14863–14868.
- Forrest, D. and Mastrangelo, C. (2001). Gemini Die Sort and TEG data analysis for Dominion Semiconductor. Technical report, University of Virginia.
- Fowler, J. W., McCarville, D. R., Montgomery, D. C., Rhoads, T. R., Runger, G. C., Skinner, K. R., and Stanley, J. D. (2000). Multivariate statistical methods for modeling and analysis of wafer probe test data. *Private Communication*.
- Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.
- Friedman, D. J., Hansen, M. H., Nair, V. N., and James, D. A. (1997). Model-free estimation of defect clustering in integrated circuit fabrication. *IEEE Transactions in Integrated Circuit Manufacturing*, 10(3):344–359.

- Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann Publishers.
- Hall, M. A. and Holmes, G. (2000). Benchmarking attribute selection data for data mining. Technical report, Department of Computer Science, University of Waikato.
- Hansen, M. H., Nair, V. N., and Friedman, D. J. (1997). Monitoring wafer map data from integrated circuit fabrication processes for spatially clustered defects. *Technometrics*, 39(3).
- He, X. and Shau, Q.-M. (2000). On parameters of increasing dimension. *Journal of Multivariate Analysis*, 73:120–135.
- Herrmann, J. W., Coughlin, B. F., Henn-Lecordier, L., Mellacheruvu, P., Nguyen, M.-Q., Rubloff, G. W., and Shi, R. Z. (2000). Understanding the impact of equipment and process changes with a heterogeneous semiconductor manufacturing simulation environment. In Joines, J. A., Barton, R. R., Kang, K., and Fishwick, P. A., editors, *Proceedings of the 2000 Winter Simulation Conference*, pages 1491–1498.
- Hess, C. and Weiland, L. H. (1999). Extraction of wafer-level defect density distributions to improve yield prediction. *IEEE Transactions on Semiconductor Manufacturing*, 12(2):175–183.
- Horton, D. (1998). Modeling the yield of mixed-technology die. *Solid State Technology*, pages 109–119.

- Huffer, F. W. and Park, C. (2000). A test for multivariate structure. *Journal of Applied Statistics*, 27(5):633–650.
- John, G. H., Kohavi, R., and Pflieger, K. (1994). Irrelevant features and the subset selection problem. In Cohen, W. W. and Hirsh, H., editors, *Machine Learning: Proceedings of the Eleventh International Conference*, pages 121–129. San Francisco, CA: Morgan Kaufman Publishers.
- Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*. Englewood Cliffs, NJ:Prentice Hall.
- Kamimura, R. T., Bicciato, S., Shimizu, H., Alford, J., and Stephanopoulos, G. (2000). Mining of biological data I: Identifying discriminating features via mean hypothesis testing. *Metabolic Engineering*, 2:218–227.
- Kennedy, R. L. and et al. (1998). *Solving Data Mining Problems through Pattern Recognition*. Prentice Hall.
- Leang, S., Ma, S.-Y., Thomson, J., Bombay, B., and Spanos, C. (1996). A control system for photolithographic sequences. *IEEE Transactions on Semiconductor Manufacturing*, 9(2):191–206.
- Li, C., Chang, E. Y., Garcia-Molina, H., and Widerhold., G. (1999). Clindex: Clustering for similarity queries in high-dimensional spaces.
- Li, K. C. (2001). Dimension reduction and visualization. <http://www.stat.ucla.edu/~kcli/sir-PHD.pdf>.
- Liu, H. and Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Boston.

- Liu, H. and Setiono, R. (1998). Some issues on scalable feature selection. *Expert Systems with Applications*, 15:333–339.
- Liu, H. and Setiono, R. (2001). Some issues on scalable feature selection. *Applied Intelligence, Kluwer preprint*.
- McLeod, A. I. and Provost, S. B. (2001). Multivariate data visualization. *Preprint*.
- Montgomery, D. C. (1996). *Introduction to Statistical Process Control*. New York: John Wiley & Sons, Inc.
- Moyne, J., del Castillo, E., and Hurwitz, A. M., editors (2001). *Run-to-Run Control in Semiconductor Manufacturing*. CRC Press.
- Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Statistics*. Chicago,IL:Irwin, 4th edition.
- Ng, R. T. and Tam, D. (1999). Multilevel filtering for high dimensional image data: Why and how. *IEEE Transactions on Knowledge and Data Engineering*, 11(6):916–928.
- Nurani, R. K., Strojwas, A. J., Maly, W. P., Ouyang, C., Shindo, W., Akella, R., McIntyre, M. G., and Derret, J. (1998). In-line yield prediction using patterned wafer inspection information. *IEEE Transactions on Semiconductor Manufacturing*, 11(1).
- Richards, W. R. and Shen, M. (2000). Extraction of two-dimensional metal-metal-oxide-semiconductor field effect transistor structural information from electrical characteristics. *Journal of Vacuum Science Technology*, 18(1):533–539.

- Shin, C. K. and Park, S. C. (2000). A machine learning approach to yield management in semiconductor manufacturing. *International Journal of Production Research*, 38(17):4261–4271.
- Shindo, W., Nurani, R. K., and Strojwas, A. J. (1998). Effects of defect propagation/growth on in-line defect based yield prediction. *IEEE Transactions on Semiconductor Manufacturing*, 11(4):546–551.
- Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., and Song, F. (2000). *Methods for Meta-analysis in Medical Research*. New York: John Wiley & Sons, Inc.
- Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire CT:Graphics Press.
- Van Zant, P. (1997). *Microchip fabrication: A practical guide to semiconductor processing*. New York:McGraw Hill, 3rd edition.
- Wilhelm, A. F. X., Weman, E. J., and Symanzik, J. (2001). Visual clustering and classification: The Oronsay particle size data revisited. *preprint*.
- Wu, X. and Urpani, D. (1999). Induction by attribute elimination. *IEEE Transactions on Knowledge and Data Engineering*, 11(5):805–812.
- Yang, Y. and Pedersen, J. O. (2001). A comparative study on feature selection in text categorization. *Not Published*.

Appendix

Appendix A

Worries – the appendix to be cut out

A.1 Miscellany

Intermediate variables as the links in a hierarchy: Models of intermediate variables, TEG to DS have correlation coefficients of $r=0.007$ in some cases. This concept turned out to be flawed because the meta-model will at best predict only as well as the model of the top level parameters as a function of the sub-model outputs. In the specific, using TEG variable intermediaries shows limits the performance of the metamodel to the model of the process outputs based on actual TEG data. In this case, a correlation coefficient of $r = 0.007$ indicates that the metamodel will not be very useful.

A.2 To do List

- Set a date 11 November 2002

- Talk to James Groves
- Write a schedule
- Finish modeling by Aug 30
 - Other submodels GC: lq1hgc, lx1hgc, vd1hxa, ho1hgc, M1: re2hm1, m1cd, m1mask, m1over, TEG
 - Metamodel, which combines the estimates
 - * Logistic, a second level logisitc model of the default rates
 - * Bayesian updating
 - Models of other outputs
 - * Current: Marginal Defect Rate at S1P1 lot-wise data
 - * DC and Functional Yields of same
- Draft chapters 1,2,3 by September 2nd
 - Rework previous writings
 - Draft chapter 4 (Results/Application) by Sept. 17
 - Draft chapter 5 (Discussion) by Sept 17
 - Rewrite chapter 3 per outline – 17
- Final draft to Sept 20
- Final to Committee Sept 30/Oct 2
- Organize talk
- Defend week of Oct 13

- Publication
 - JQT paper
 - Dissertation paper

A.3 Research Plan

My overall plan for developing a methodology for managing complex manufacturing data as found in semiconductor manufacturing includes several stages and tasks:

- Literature review:
 - Semiconductor manufacturing
 - Model complexity, hierarchical and multi-level modeling
 - High dimensional visualization, variable subset selection, knowledge discovery, and dimension reduction using PCA, SVD, and PLS
 - High dimension data mining and data warehousing in image analysis, text mining, spectral analysis, and bioinformatics
- Data characterization of real semiconductor manufacturing data
 - Exploration of a production data set
 - Identification of database elements as variables of interest
 - Extraction of die sort, in-line testing, quality control and production data
 - Alignment of lots across the datasets

- Hierarchical modeling of the data
 - Development of a “natural” hierarchy
 - Visualization of datasets: scatter plots and data images
 - Eliminate nulls and constants, other feature selection
 - Model in line testing data to defect rates and yield. (Fig A.1 TEG PSLs to DS Yields)
 - Multiple models of quality control information to in line testing data (QC Engineering Specifications to TEG PSLs)
 - Multiple models of production equipment to quality control information (Figure A.1 QC production Data to QC Engineering Specifications)

- Validation
 - In-line testing data to yield and failure rate models
 - Sub-model prediction performance (e.g.: TEG actuals versus predicted based on QC data)
 - Relative performance of alternate QC data models

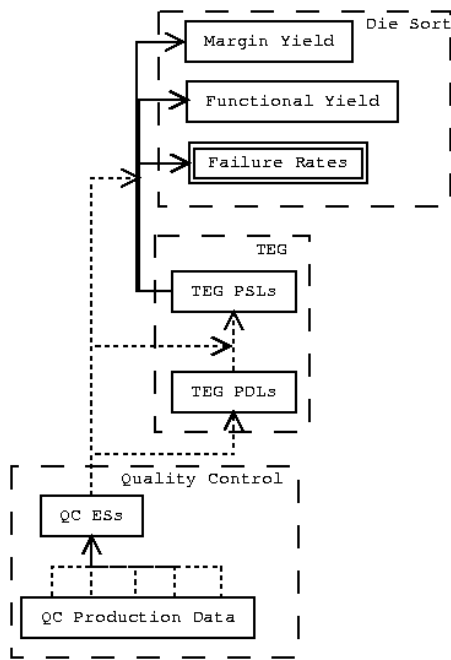


Figure A.1: Hierarchical Model of DS, TEG and QC data. Solid lines represent current linkages and models, dashed lines represent databases, and dotted lines represent future linkages and models.