

5-11-2007

Improving Monitoring and Diagnosis for Process Control using Independent Component Analysis

Thaddeus Taylor Shannon
Portland State University

Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds



Part of the [Systems Science Commons](#)

Let us know how access to this document benefits you.

Recommended Citation

Shannon, Thaddeus Taylor, "Improving Monitoring and Diagnosis for Process Control using Independent Component Analysis" (2007). *Dissertations and Theses*. Paper 6136.

This Dissertation is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.

**IMPROVING MONITORING AND DIAGNOSIS FOR PROCESS CONTROL
USING INDEPENDENT COMPONENT ANALYSIS**

by

THADDEUS TAYLOR SHANNON III

A dissertation submitted in partial fulfillment of the
requirements for the degree of

**DOCTOR OF PHILOSOPHY
in
SYSTEMS SCIENCE**

Portland State University
2007

DISSERTATION APPROVAL

The abstract and dissertation of Thaddeus Taylor Shannon III for the Doctor of Philosophy in Systems Science were presented May 11, 2007, and accepted by the dissertation committee and doctoral program.

COMMITTEE APPROVALS:



Andrew M. Fraser, Chair



George G. Lendaris



James McNames

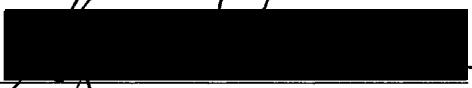


Martin Zwick



Serge Preston,
Representative of the Office of Graduate Studies

DOCTORAL PROGRAM APPROVAL:



George G. Lendaris, Acting Director
Systems Science Ph.D. Program

ABSTRACT

An abstract of the dissertation of Thaddeus Taylor Shannon III for the Doctor of Philosophy in Systems Science presented May 11, 2007.

Title: Improving Monitoring and Diagnosis for Process Control Using Independent Component Analysis

Statistical Process Control (SPC) is the general field concerned with monitoring the operation and performance of systems. SPC consists of a collection of techniques for characterizing the operation of a system using a probability distribution consistent with the system's inputs and outputs. Classical SPC monitors a single variable to characterize the operation of a single machine tool or process step using tools such as Shewart charts. The traditional approach works well for simple small to medium size processes. For more complex processes a number of multivariate SPC techniques have been developed in recent decades. These advanced methods suffer from several disadvantages compared to univariate techniques: they tend to be statistically less powerful, and they tend to complicate process diagnosis when a disturbance is detected.

This research introduces a general method for simplifying multivariate process monitoring in such a manner as to allow the use of traditional SPC tools while facilitating process diagnosis. Latent variable representations of complex processes are developed which directly relate disturbances with process steps or segments. The

method models disturbances in the process rather than the process itself. The basic tool used is Independent Component Analysis (ICA). The methodology is illustrated on the problem of monitoring Electrical Test (E-Test) data from a semiconductor manufacturing process. Development and production data from a working semiconductor plant are used to estimate a factor model that is then used to develop univariate control charts for particular types of process disturbances. Detection and false alarm rates for data with known disturbances are given. The charts correctly detect and classify all the disturbance cases with a very low false alarm rate.

A secondary contribution is the introduction of a method for performing an ICA like analysis using *possibilistic* data instead of *probabilistic* data. This technique extends the general ICA framework to apply to a broader range of uncertainty types. Further development of this technique could lead to the capability to use extremely sparse data to estimate ICA process models.

Dedication

For MaryAnn and Tex

and Daniela

and Robin

and Shersten.

Acknowledgments

I would like to thank the members of my committee for their ideas, guidance, and support. The results on e-test data are joint work with James McNames, David Abercrombie, Manut Rehani, and Bruce Whitefield. My work on possibilistic source separation grew out of many helpful interactions with professor George Klir of SUNY Binghamton. I would also like to acknowledge many helpful interactions with Robert Daasch and his students from the IC Design and Test Laboratory.

This work was generously supported by LSI Logic and the National Science Foundation. Daniela Steiner, Stephen Shervais, Larry Schultz, Kevin Vixie, Shari Matzner, Michael Carroll, Roberto Santiago, Dawn Sharafi, TC Smith, Robin Lane, and Shersten Finley have all been invaluable in this endeavor.

Table of Contents

Acknowledgments.....	ii
List of Tables.....	v
List of Figures.....	vi
Notations.....	viii
1. Introduction.....	1
1.1 Systems Framework.....	3
1.2 Data Driven System Modeling.....	5
1.3 Data Reduction.....	7
1.4 Independent Components Analysis.....	10
1.5 An ICA Based SPC Method.....	11
1.6 Dissertation Overview.....	12
2. Independent Component Analysis and Exploratory Factor Analysis.....	14
2.1 Principal Components.....	14
2.1.1 PCA methods.....	15
2.1.2 Example.....	17
2.2 Factor Analysis.....	20
2.2.1 Extracting Factors.....	21
2.2.2 Exploratory versus Confirmatory Factor Analysis.....	24
2.3 Independent Components.....	25
2.4 The ICA Model and Fundamental Concepts.....	26
2.5 Two-Source BSS Example.....	29
2.6 Standard Approaches to Implementation.....	34
2.6.1 Algebraic Methods.....	35
2.6.2 Estimators of Independence.....	36
2.6.3 Maximum Likelihood Estimation.....	42
2.6.4 Optimization Algorithms.....	43
2.7 BSS of Sources with Time Structure.....	46
2.8 Blind Deconvolution.....	47
2.9 Historical Development of ICA.....	48
3. Statistical Process Control.....	50
3.1 Univariate Control Charts.....	52
3.2 Optimality and Design Issues.....	58
3.3 Multivariate Approaches.....	60
3.4 Advanced Process Control.....	63
4. ICA Based Exploratory Factor Analysis for E-test Data.....	65
4.1 Introduction.....	65
4.2 Methodology.....	66
4.3 Data.....	67
4.4 Results.....	70
4.5 Conclusion.....	74

5. ICA Disturbance Modeling Methodology.....	76
5.1 Methodology	76
5.2 Data.....	78
5.3 Model Estimation	80
5.4 Nominal Distributions	83
5.5 Construction of Control Charts	88
5.6 Validation Results	97
5.7 Discussion.....	99
6. Possibilistic Independent Component Analysis.....	101
6.1 Possibility Theory.....	101
6.2 Possibilistic Measures of Uncertainty	102
6.3 Possibilistic Independence: Example	103
6.4 Proposed Method.....	105
6.5 Crisp Example.....	107
6.6 Discussion.....	112
7. Conclusion	114
7.1 Summary.....	114
7.2 Future Research.....	115
References	118
Appendix A: Information Theory.....	128
Appendix B: Cumulants.....	132

List of Tables

Table 2.1	Eigenvalues and variance explained	18
Table 3.1	Data for control chart example	54
Table 4.1	Percentage of parameter variance explained by each principal component	70
Table 4.2	Percentage of parameter variance explained by each independent component	71
Table 4.3	Percentage of parameter variance explained by each independent component of the reduced model	72
Table 4.4	Percentage of parameter variance explained by each independent component, reduced model, LD and TOX W omitted	73
Table 5.1	Proportion of variance explained by the first seven principal components	82
Table 5.2	Independent component explanation of variance	83
Table 5.3	Detection results for wafers included in ICA	98
Table 5.4	Detection results for wafers not included in ICA	98
Table 5.5	Explanation of variance of five independent components	100

List of Figures

Figure 1.1	Data driven system modeling, standard architecture	6
Figure 1.2	Data driven modeling with unobservable inputs	7
Figure 2.1	Eigenvalues of the Principal Components	17
Figure 2.2	Variance explained by each Principal Component	19
Figure 2.3	Cumulative variance explained by Principal Component	19
Figure 2.4	The joint distribution of the independent sources s_1 and s_2	30
Figure 2.5	The joint distribution of the observed mixtures x_1 and x_2	31
Figure 2.6	The joint distribution of the whitened mixtures y_1 and y_2	32
Figure 2.7	The joint distribution of the estimated sources \hat{s}_1 and \hat{s}_2	34
Figure 3.1	Twenty hours of samples plotted with 2 and 3 σ limits	55
Figure 3.2	Shewart \bar{X} chart with 2 and 3 σ limits	56
Figure 3.3	Upper and lower CUSUMs, $k = 0.1$	58
Figure 4.1	Percent of variance explained by principal component	69
Figure 5.1	Nominal distribution of N implant score wafer mean	84
Figure 5.2	Nominal distribution of P implant score wafer mean	84
Figure 5.3	Nominal distribution of gate oxide score wafer mean	85
Figure 5.4	Nominal distribution of lithography score wafer mean	85
Figure 5.5	Nominal distribution of N implant score wafer variance	86
Figure 5.6	Nominal distribution of P implant score wafer variance	86
Figure 5.7	Nominal distribution of gate oxide score wafer variance	87
Figure 5.8	Nominal distribution of lithography score wafer variance	87
Figure 5.9	Shewart chart for N implant score wafer mean	89
Figure 5.10	Compressed Shewart chart for N implant score wafer mean	89
Figure 5.11	Compressed Shewart chart for P implant score wafer mean	90
Figure 5.12	Compressed Shewart chart for gate oxide score wafer mean	90
Figure 5.13	Compressed Shewart chart for lithography score wafer mean	91
Figure 5.14	Compressed Shewart chart for N implant score wafer variance	91
Figure 5.15	Compressed Shewart chart for P implant score wafer variance	92
Figure 5.16	Compressed Shewart chart for gate oxide score wafer variance	92
Figure 5.17	Compressed Shewart chart for lithography score wafer variance	93
Figure 5.18	Chart for additional lots, N implant score wafer mean	93
Figure 5.19	Chart for additional lots, P implant score wafer mean	94
Figure 5.20	Chart for additional lots, gate oxide score wafer mean	94
Figure 5.21	Chart for additional lots, lithography score wafer mean	95
Figure 5.22	Chart for additional lots, N implant score wafer variance	95

Figure 5.23	Chart for additional lots, P implant score wafer variance	96
Figure 5.24	Chart for additional lots, gate oxide score wafer variance	96
Figure 5.25	Chart for additional lots, lithography score wafer variance	97
Figure 6.1	Observed possibility distribution	104
Figure 6.2	Demixed possibility distribution	105
Figure 6.3	Joint possibility distribution of two noninteractive sources	108
Figure 6.4	Mixed (observed) possibility distribution	109
Figure 6.5	Centered distribution	109
Figure 6.6	Centered distribution with principal axes aligned	110
Figure 6.7	Distribution with principal axes scaled	110
Figure 6.8	Distribution after final separating rotation	111

Notations

s_i	The i th source random variable.
\mathbf{s}	The source random vector.
x_i	The i th observation random variable.
\mathbf{x}	The observation random vector.
$p(s_i)$	The probability density/mass function of variable s_i .
$p(s_i)(\omega)$	The probability density/mass function of variable s_i written as a function with its argument, the variable ω , made explicit.
$h(y)$	The entropy of the random variable y .
$D(p(y) p(z))$	The Kullback-Leibler divergence of the density $p(y)$ from the density $p(z)$.
$g(\omega)$	The zero mean, unit variance Gaussian density.
\mathbf{C}_z	The covariance matrix of the random vector \mathbf{z} .
$E\{\mathbf{z}\}$	The expectation of the random variable \mathbf{z} .
\hat{u}	An estimate of the variable, parameter, vector or function denoted by u .
$ \mathbf{K} $	The determinant of the matrix \mathbf{K} .
$\ \mathbf{v}\ $	The norm of the vector \mathbf{v} (may be taken to be the L_2 norm).

1. Introduction

As humans design and operate ever more complex systems, understanding and monitoring the operation of the systems becomes increasingly difficult. Not only do the systems become more difficult to design, but their operation and maintenance, often by non-specialists, pose additional problems. Statistical Process Control (SPC) is the general field concerned with monitoring the operation and performance of such systems. Originally developed in the manufacturing sector, SPC has been embraced by organizational theorists and management scientists, and its techniques are regularly applied to problems ranging from biomedical engineering to business development and marketing.

SPC consists of a collection of techniques for characterizing the operation of a system using a probability distribution consistent with the system's inputs and outputs. Changes observed in the empirical distribution over time are attributed to changes in the system's operation. The proper design of an SPC scheme for a particular system requires that the disturbances external to the system be distinguished from internal system disturbances.

The usefulness of an SPC system may be judged by its ability to detect disturbances in a timely fashion, to have a low rate of false alarms, and to be able to distinguish between different kinds of disturbances. Classical SPC in manufacturing has focused on the first two criteria, quick detection and minimization of false alarms. As is to be expected with any statistical classification problem, there is a tradeoff

1. Introduction

between these two objectives. Fast detection, signaling an alarm at the first evidence of a disturbance, yields a higher false alarm rate than slow detection. The variance of the monitored variable(s) and the magnitude of the disturbance one is monitoring determine the tradeoff.

Classical control concerns itself with Single Input, Single Output (SISO) problems. More complex plants are analyzed or designed by breaking the complex systems down into subsystems such that the subsystems either do not interact or operate on significantly different timescales. Fast variables are then controlled “inside” and independently of slow variables. In this paradigm, all systems are decomposed into single variable feedback loops. The SPC equivalent of this approach is to monitor a single variable to characterize the operation of a single machine tool or process step using tools such as Shewart charts or Cumulative Sum (CUSUM) charts. This traditional approach works well for simple small to medium size processes, but becomes inefficient when applied to large (many step) processes. When a plant has many steps, even a low false alarm rate at each step produces a high false alarm rate for the overall process.

This approach becomes even less tenable when single process steps are either too expensive to monitor directly or cannot be characterized by a single measurable variable, e.g., a batch fermentation process. In these cases, multiple variables need to be monitored simultaneously as they jointly characterize (as a multivariate probability distribution) the process step(s). A number of multivariate SPC techniques have been

1. Introduction

developed in recent decades to address these situations, including Hotelling T^2 tests, Principal Component Analysis for dimension reduction, and Principal Regression Analysis for Input/Output simplification. These advanced SPC methods work well in many circumstances but suffer from several disadvantages compared to univariate techniques: they tend to be statistically less powerful, and they tend to complicate process diagnosis when a disturbance is detected.

This dissertation presents a general method for simplifying multivariate process monitoring in such a manner as to allow the use of traditional SPC tools while facilitating process diagnosis. The objective is to develop latent variable representations of complex processes which are directly identified with process steps or segments. Our method models disturbances in the process rather than the process itself. The basic tool used is a relatively new method for data analysis known as Independent Component Analysis (ICA). The methodology is illustrated on the problem of monitoring Electrical Test (E-Test) data from a semiconductor manufacturing process. On the sample problem, the methodology successfully distinguishes between the four different types of disturbances it is designed for.

1.1 Systems Framework

In order to present this work in a Systems Science context, one must begin with a suitably abstract framework for discussing systems in general. An appropriate framework that encompasses our issues of interest is Klir's systems epistemology (Klir 1985, Klir 2001). The first type of system considered is a *source system*, a set of

1. Introduction

objects with a relation. Practically, we consider a set of variables defined as a system in some context. The next type of system is a *data system*, a source system together with a set of observed values for the variables. Often one defines a source system of interest only to discover that only some of the variables are observable. It is tempting, in those circumstances, to redefine the source system of interest based on the measurability of the implied data system. In some instances, this can be done with little harm. However, from a methodological point of view, it very much like looking for a lost set of keys under a lamp post because that is where the light is. Many applied disciplines have been forced to develop methods for dealing with this problem. Two prominent examples are the construction of state observers in control theory, and the use of factor models and latent variable analysis in the social sciences and finance. These methods seek to estimate data for specific unobserved variables based on available observations of related variables.

After moving from a source system to a data system description, one normally wants to move on to a *generative system*. A generative system is a time invariant description of the relations between the system variables that is more abstract than an exhaustive enumeration. A generative system description of a data system must be consistent with the observed data values. Generative systems can generally be thought of as models of the source system. The general systems method of moving from source system to data system, and data system to generative system can be described as data driven system modeling. A fourth type of system appears in design and

1. Introduction

analysis of complex systems. *Structure systems* are systems in which the nature of the interaction between variables is recognized. The decomposition of a complex control problem into a collection of SISO feedback loops is a classic example of the use of systems of this type.

To place our work in this framework, we recognize that our objective is to obtain a structured data system that describes the presence of identified disturbances, in which the variables are non-interactive (so that univariate SPC techniques may be used). We start with a data system of interacting variables, estimate a generative model using a data driven modeling method, and then identify the resulting latent variables with particular types of disturbances. Our end product is then a generative model that transforms one data system into another.

1.2 Data Driven System Modeling

Data driven modeling is a popular topic in the statistical learning and soft computing communities. This approach selects a parameterized functional form for the model, then adjusts the parameters to minimize the model error estimated from observations of the original (modeled) system. Figure 1.1 provides a block diagram of the basic modeling setup.

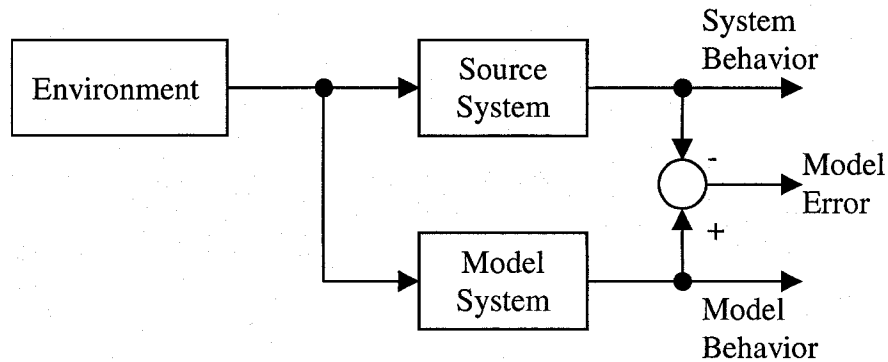


Figure 1.1: Data driven system modeling, standard architecture.

The source system's behavior, its observable outputs, is assumed to be a function of the system's internal states and environmental inputs. The model system's outputs are compared to the source system's outputs for identical input sequences, and the difference (model error) is used to adjust the model function's parameters. This general approach is the basis for a variety of techniques in engineering practice, ranging from adaptive filtering to artificial neural networks and fuzzy systems. It relies on the availability of the environmental inputs for its success. If the inputs are not observable, this approach cannot be used.

If the inputs are unobservable, the modeling architecture used is like that in Figure 1.2. The basic approach here is to choose a parametric form for the model that is invertible, and use it to produce estimates of the environmental inputs that produced the observed system behavior. These estimated inputs are then evaluated in a criterion function that provides a goodness-of-fit or figure-of-merit for the model. The ubiquitous example of this approach is linear time series modeling using ARMA

1. Introduction

models. The classical approach for time series modeling uses the likelihood of the observed data as the model criterion, which is calculated from the residuals. The residuals are the moving average inputs to the model. The key assumption made is that the inputs are independent and identically distributed, zero mean, Gaussian random variables. Without such an assumption, the likelihood cannot be calculated.

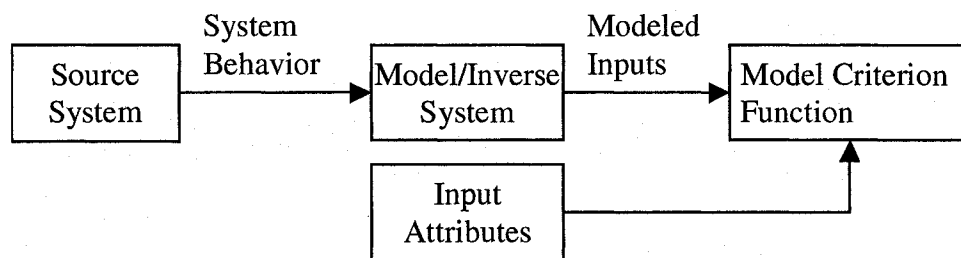


Figure 1.2: Data driven modeling with unobservable inputs.

1.3 Data Reduction

Even when working with the data system of one's choice, one is often faced with data reduction problems, i.e. one is presented with a multivariate set of observations generated by a system, and has to find a transformation of the data that allows the behavior of the observed system to be better characterized. Different senses of better characterizations range from shorter descriptions of the original observations without loss of fidelity (e.g. source compression), to descriptions of the behaviors of individual subsystems of the observed system together with a characterization of how the individual subsystem behaviors generated the observed system behavior (e.g.

1. Introduction

source separation, factor analysis, latent variable analysis). In this context, a variable that describes the underlying behavior is referred to as a *latent* variable. Often, data reduction is merely the first step in an investigation, whose goal is to make the available data easier to work with, without losing any information needed for a solution. From a statistical point of view, one is presented with a set of observations that imply a (often high dimensional) joint probability density/distribution, which one seeks to represent using an alternative coordinate system (often for a much lower dimensional manifold).

The best established and most straightforward approach to this problem is Principal Component Analysis (PCA). In PCA, one finds the eigen decomposition of the covariance matrix of the data. As covariance matrices are positive semi-definite, they are guaranteed to have a full, orthogonal set of eigenvectors, which together specify a linear map that diagonalizes the matrix. This map transforms the original observed variables into a new set of orthogonal, *uncorrelated* variables. The variance of each new variable is specified by the eigenvalue of the associated eigenvector in the transformation. In other words, the eigenvectors of the covariance matrix form a basis for the observation space, for which the observed density is uncorrelated. An alternative means of calculating the principal components is to find the singular value decomposition of the observation vectors stacked as an array.

Factor analysis is a method employed by social scientists and financial analysts that attempts to explain a set of observations using a linear model containing

1. Introduction

unobserved factors as independent variables. Frequently an investigator will use such a model to study a particular variable that is immeasurable, yet is highly related to a number of easily measured variables. In such cases, both the value of the factor variable and the parameter values of the factor model may be of interest. From the description of PCA, it should be clear that any orthogonal transformation of the observed data is a valid latent variable or factor representation of the data. Thus, additional constraints must be imposed on the analysis if one hopes to find a unique factor model. Such constraints usually include dimensionality reduction from the observation space to the factor space, the assumption of uncorrelated zero-mean disturbances for the factor model, and often the imposition of priors for the signs of parameters in the factor model.

One interpretation of PCA is that it provides a decomposition of the observation space into a set of uncorrelated subspaces, which may be ordered by the variance of the data each contains. The usual assumption is that the variance of the data in each subspace signifies the relative importance of that subspace. Thus, subspaces with very little variance are not very important for representing the data, and in many circumstances are good candidates for being dropped from a representation. This observation forms the basis for the methods of Principal Component Regression (PCR) and Karhunen-Loeve (KL) transforms. PCR is a method for estimating linear models of a dependent variable based on the most significant principal components of the independent variables. PCR reduces the

1. Introduction

degrees of freedom in a model, thus producing a lower variance (though slightly biased) estimator. KL transforms are projections of the data onto the most significant subspace as determined by PCA. As such, they can provide significant dimension reduction, hence data compression, and are used in a wide range of signal and data processing applications.

1.4 Independent Components Analysis

Independent Component Analysis (ICA) often includes PCA as a first step. The purpose of ICA is to find that orthogonal transformation that maximizes the statistical independence of the transformed variables. Statistical independence provides the additional constraints necessary for source separation. Two random variables are said to be independent when their joint density is factorizable, i.e. their joint density is the product of their separate univariate densities. This implies that knowledge of the value of one variable for a particular trial imparts no information concerning the value of the other variable for that trial. From the point of view of information theory, two variables are independent when their mutual information is zero. Thus, ICA seeks a linear transformation of the PCA basis that minimizes the mutual information between each pair of the output variables. Alternatively, one could say that ICA builds an invertible linear model of the system based on the assumption that the system's inputs are independent. In terms of the architecture of Figure 1.2, ICA uses mutual information as the figure of merit to be minimized.

1. Introduction

The key issues in ICA are 1) how is the mutual information between the sources estimated from the transformed data, 2) how is the mixing transformation parameterized, and 3) how are the parameters adjusted to minimize the mutual information. The first ICA methods used estimates of kurtosis and skewness, fourth and third order cumulants (see Appendix 2), or ad hoc nonlinear correlation measures as their objective functions. More recent methods use estimates of the Kullback-Leibler divergence between the joint and product densities of the sources as the function to be optimized. At the same time, a variety of optimization methods have been refined specifically tailored to ICA problems, e.g. the natural gradient and FastICA algorithms.

1.5 An ICA Based SPC Method

We adopt the position that statistical independence can provide the extra constraint necessary for modeling process disturbances in Multiple Input Multiple Output (MIMO) systems when only observations of the output variables are available. In particular, we use ICA methods for identification and system structure investigation. The general procedure is summarized as:

- Collect process output data that includes both nominal process operation and designed disturbances that qualitatively enumerate the types of disturbances one wishes to detect,
- Develop independence-based factor-analytic models of the process outputs that link specific factors to specific disturbance types,

1. Introduction

- Using the nominal process outputs, develop probability distributions for the factor scores under nominal process,
- Design appropriate univariate SPC charts for the factor scores,
- Validate the factor models by retroactively applying the SPC charts to the designed disturbance data.

1.6 Dissertation Overview

Chapter two is a fairly detailed description of ICA models, methods, and motivation. It begins with a discussion of multivariate methods starting with Principal Component Analysis (PCA) and Factor Analysis (FA) that lead to ICA methods. After a short introduction to the linear ICA model, the basic method is demonstrated with a simple two dimensional blind source separation example. The example is followed by a survey of ICA methods, including a discussion of the relationship between various estimators of independence and maximum likelihood methods. After that, there are sections describing ICA methods for time structured observations and blind deconvolution. Chapter two concludes with a brief sketch of the historical development of ICA.

The third chapter introduces the field of Statistical Process control. It includes illustrative examples of the most common univariate methods: Shewart Charts and CUSUM charts. The optimality of these techniques is briefly discussed along with their implementation issues. Multivariate SPC is then discussed both in relation to the univariate techniques and with respect to process diagnosis. Chapter four introduces

1. Introduction

our example application field, semiconductor manufacturing process quality control. After a description of the problem context, a review of work applying ICA to E-test data in a factor analytic context is provided. The fifth chapter describes our ICA based approach to simplifying multivariate SPC. Development and production data from a working semiconductor plant are used to estimate a factor model that is then used to develop univariate control charts for particular types of process disturbances. Detection and false alarm rates for data with known disturbances are given. The charts correctly detect and classify all the disturbance cases with a very low false alarm rate.

The sixth chapter introduces a method for performing an ICA like analysis using *possibilistic* data instead of *probabilistic* data. This technique extends the general ICA framework to apply to a broader range of uncertainty types. While not as yet directly applicable to our SPC framework, further development of this technique may lead to the capability to use extremely sparse data to estimate ICA process models. Such a capability would allow rare or costly disturbances to be modeled in our framework. The final chapter summarizes this work and reviews the possibilities for its further development.

2. Independent Component Analysis and Exploratory Factor Analysis

Latent variable techniques are numerical methods that delineate the structure of empirical multivariate datasets. The most common technique is Principal Component Analysis, which is often applied as a preprocessing step to achieve dimension reduction. When the data can be divided into input and output subsets, Partial Least Squares Regression can be used to reduce the dimensionality of both subsets while also estimating a linear model of the input-output relationship. Factor Analysis (FA) is a set of methods and heuristics for defining interesting representations of datasets in terms of latent variables that are “meaningful” to the analyst. The beauty in the eyes of the beholder nature of FA arises because the system of equations FA is solving is underspecified. Factor Analysis has two modalities, an exploratory form and a confirmatory form, which differ slightly in their assumptions and procedures.

2.1 Principal Components

The best established latent variable method is Principal Component Analysis (PCA). In PCA, one finds the eigen decomposition of the covariance matrix of the data. The variance of each latent variable is specified by the eigenvalue of the associated eigenvector in the transformation. The eigenvectors of the covariance matrix form a basis for the observation space, for which the components of the observed density are uncorrelated. An alternative means of calculating the principal

2. ICA and Factor Analysis

components is to find the singular value decomposition of the data vectors stacked together to form a matrix.

PCA provides a decomposition of the observation space into a set of uncorrelated subspaces, which may be ordered by the variance of the data each contains. The variance of the data in each subspace signifies the relative importance of that subspace for representing the data. Subspaces with very little variance are not very important for representing the data, and in many circumstances are good candidates for being dropped from a representation. In this manner, PCA finds the set of latent variables providing the most parsimonious representation of the observations.

2.1.1 PCA methods

Methods for performing PCA can be classified as either closed form or on-line learning techniques. Closed form methods start with an estimated covariance matrix formed from normalized observations (zero mean, unit variance). Standard numerical techniques such as QR decomposition are then used to find the eigenvectors and eigenvalues. On-line techniques are used when adaptive estimation is required for tracking nonstationary distributions. Algorithms have been devised using either variance maximization or mean-square error minimization with gradient learning techniques.

There are a variety of criteria for deciding how many components to keep, the simplest of which is the Kaiser or *eigenvalue one* criteria. Under the assumption that the PCA is performed on normalized data (z-scores with unit variance), the total

2. ICA and Factor Analysis

variance of the n -variate will be n . Any component with eigenvalue less than one will capture less variance than any one of the original observations. The criterion keeps only those components with more variance than the observations and drops the others. While easy to understand and apply, the crisp nature of the decision rule can lead to arbitrary results, as it takes no account of the total variance explained by the remaining components. Another simple criterion is to keep enough components to explain some predetermined total variance, such as 90% or 95%. While this method considers the fidelity of the reduced model, it does not consider the complexity of the model. It may also result in arbitrary results in that the least significant component retained may be as explanatory as the most significant component dropped.

The *scree test* attempts to overcome this problem by keeping all the factors more significant than the scree – the relatively flat tail of the eigenvalue distribution. Ideally there is a sharp break where the slope of the eigenvalue distribution changes from significantly negative to nearly zero; this is the threshold the scree criterion selects. Unfortunately, there may be multiple breaks, so there is still arbitrariness to this criterion. Both Akaike's Information Criterion (AIC) and the Minimum Description Length Criterion (MDL) can be used as selection criteria to control the tradeoff between degrees of freedom and description fidelity, though at significant computational cost. Cross-validation is generally used to evaluate the reliability of any particular choice.

2.1.2 Example

This example illustrates the ambiguity in selecting how many components to keep. The data is drawn from 2 dimensional Fourier transforms of satellite images. For each image, there are 197 samples. The full spectrum of eigenvalues has been truncated to the first 31, the number necessary to explain 95% of the total variance of the samples. The Kaiser criterion would suggest keeping just 16 of the components, those with eigenvalue greater than 1. This would also appear to be a good cutoff as it is the minimum number that retains 90% of the total variance. However, the scree test would suggest a break after the twelfth or thirteenth component, and for a dimension reduction purposes, one might select only the first six.

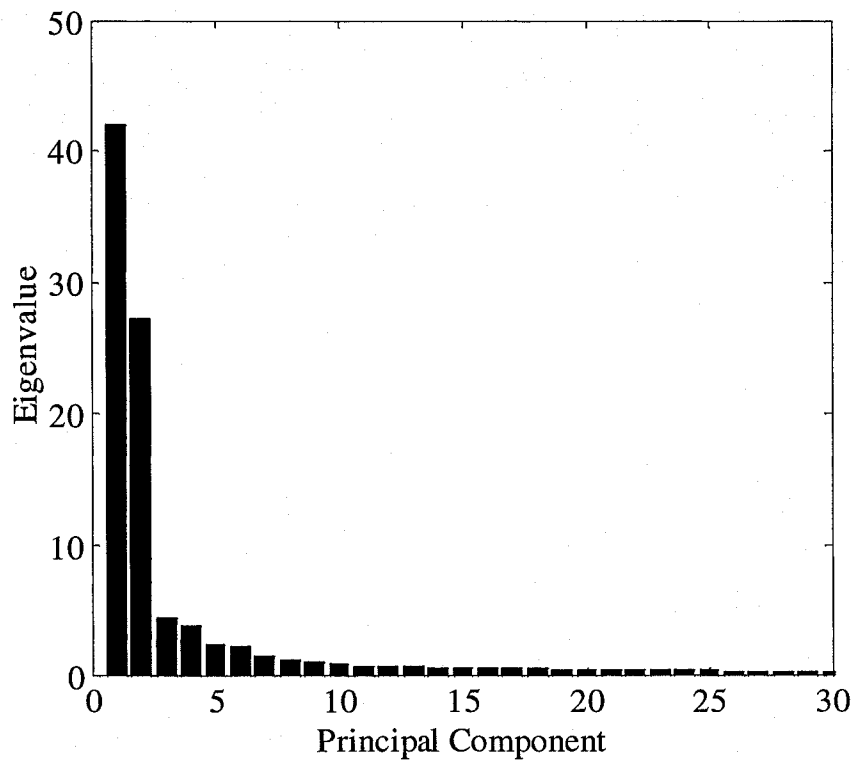


Figure 2.1: Eigenvalues of the Principal Components.

2. ICA and Factor Analysis

Table 2.1: Eigenvalues and Variance Explained

<u>Component</u>	<u>Eigenvalue</u>	<u>Variance Explained</u>	<u>Cumulative Variance</u>
1	81.035	41.99	41.99
2	52.511	27.21	69.19
3	8.551	4.43	73.63
4	7.561	3.92	77.54
5	4.601	2.38	79.93
6	4.236	2.19	82.12
7	2.661	1.38	83.50
8	2.194	1.14	84.64
9	1.828	0.95	85.58
10	1.571	0.81	86.40
11	1.461	0.76	87.15
12	1.276	0.66	87.82
13	1.185	0.61	88.43
14	1.091	0.57	89.00
15	1.054	0.55	89.54
16	1.006	0.52	90.06
17	0.887	0.46	90.52
18	0.883	0.46	90.98
19	0.812	0.42	91.40
20	0.801	0.42	91.82
21	0.746	0.39	92.20
22	0.705	0.37	92.57
23	0.662	0.34	92.91
24	0.623	0.32	93.23
25	0.605	0.31	93.55
26	0.577	0.30	93.85
27	0.524	0.27	94.12
28	0.522	0.27	94.39
29	0.497	0.26	94.64
30	0.484	0.25	94.90
31	0.454	0.24	95.13

2. ICA and Factor Analysis

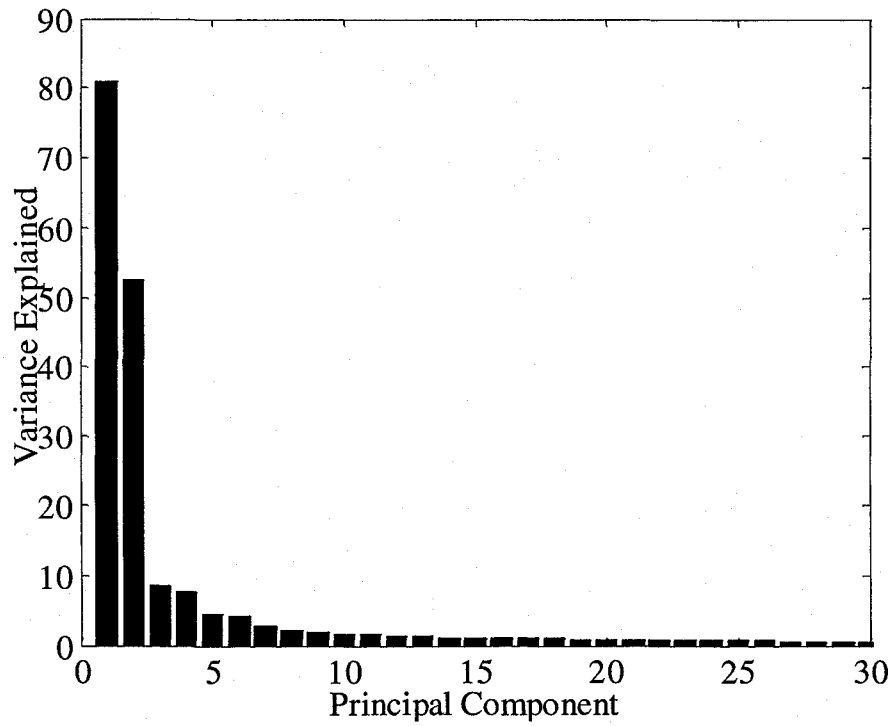


Figure 2.2: Variance explained by each Principal Component.

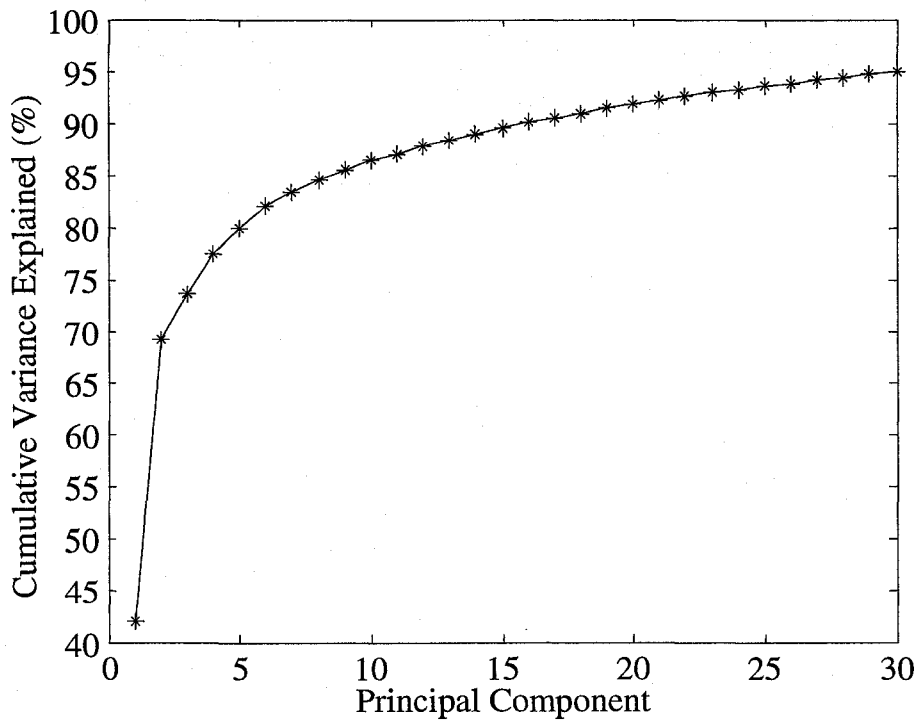


Figure 2.3: Cumulative variance explained by Principal Component.

2.2 Factor Analysis

Factor analysis is a method employed by social scientists and financial analysts that attempts to explain a set of observations using a linear model containing unobserved factors as independent variables. Frequently an investigator will use such a model to study a particular variable that is immeasurable, yet is highly related to a number of easily measured variables. In such cases, both the value of the factor variable and the parameter values of the factor model may be of interest. This quite common form of analysis, construction of factor scales, has the twin benefits of reducing the data requirements for a complex analysis (when the scale is used in place of multiple exogenous variables) and of simplifying the conceptual framework for the analyst.

The basic postulate in Factor Analysis is a generative model

$$x = Ay + n$$

where x is observed, y is the vector of common factor scores (unobserved sources), A is the factor loading matrix (mixing matrix), and n is the vector of specific factors (unobserved sources unique to each observed variable). In general, one postulates a small number of common factors, often only two or three, so the rank of A is small and the system is underspecified. The specific factors are assumed to be uncorrelated and normally distributed, while the common factors are a normal but not necessarily uncorrelated. If one assumes

$$E[yy^t] = I$$

then

$$E[xx^t] = \mathbf{A}\mathbf{A}^t + E[nn^t],$$

where $E[\]$ is expected value. Clearly there are multiple solutions for \mathbf{A} even if the variance of the specific factors is known. If the specific factors are assumed to have zero variance, then the principal components are one solution.

2.2.1 Extracting Factors

To find a factor solution, one first needs an estimate of the *communalities*, i.e. how much of the observed variance in each variable is due to the common factors as opposed to the specific factors. The communality for each variable can be estimated by its multiple correlation with all the other variables, R^2 . Denoting the diagonal matrix containing the communalities as \mathbf{R} and assuming our observations are normalized, we have

$$E[nn^t] = \mathbf{I} - \mathbf{R}.$$

Subtracting the specific factor term from both sides of the previous equation and substituting the above relation produces

$$\mathbf{A}\mathbf{A}^t = E[xx^t] - (\mathbf{I} - \mathbf{R}),$$

which means we are seeking a decomposition of the matrix obtained by substituting the communalities onto the diagonal of the observed covariance matrix (the communality based covariance matrix). The initial factors are found by taking the

2. ICA and Factor Analysis

eigen-decomposition of this matrix. Notice that this is merely principal component analysis of the communality based covariance matrix. All the caveats about how many components (factors) to keep in PCA apply to this stage of FA.

After choosing the number of factors to keep, one is left with

$$\mathbf{A} = \mathbf{E}\mathbf{D}^{1/2}$$

where \mathbf{E} is the eigenvector matrix and \mathbf{D} is the diagonal matrix of eigenvalues. This is a PCA like solution in that its factors are orthogonal and ordered from most to least explanatory. At this stage, the rows of \mathbf{A} can be considered regressions of the observed variables on the latent factors. The sum of the squared coefficients is the R^2 for the regression, which is now the exact communality for the observed variable and the percentage of that variable's variance explained by the factors.

The artistic instincts of the analyst now come into play in choosing a "rotation" to a final model. The primary motivation for the final rotation is to produce an interpretable model by attempting to evenly distribute variance across factors, and having observed variables load strongly on as few factors as possible. The final rotation can either be orthogonal, in which case the factors remain orthogonal, or oblique. Oblique transformations complicate the interpretation significantly, as the factors will be correlated with one another, but are used because they can greatly simplify the factor loadings. Oblique transformations are found by first finding an orthogonal rotation and then tweeking to simplify the complexity of the factors.

2. ICA and Factor Analysis

The standard framework for orthogonal rotations is maximization of the *Orthomax* criterion

$$OM = n \sum_i \sum_j a_{i,j}^4 - \gamma \sum_i \left(\sum_j a_{i,j}^2 \right)^2$$

where n is the number of factors, the rows are usually normalized prior to rotation and the constant γ controls which of the traditional criteria is implemented. With γ set to one, one has the *Varimax* criterion, which maximizes the variance of the variance distribution for each variable. Varimax tends to result in either high or low factor loadings with few moderate loadings, and is the most frequently used rotation criteria. With γ set to zero, this is the *Quartimax* criterion, which maximizes the kurtosis of the loadings. Quartimax has a tendency to produce a common or general factor while minimizing the number of moderate factor loadings. If a common factor is expected, it may be a good choice. A γ of $n/2$, where n is the number of factors retained, produces the *Equimax* criterion, a middle ground between Varimax and Quartimax.

If the results of the orthogonal rotation are unsatisfactory from a factor complexity or interpretability perspective, an oblique transformation can be added. The common procedure is to raise the factor loadings from the orthogonal transformation to a high power, which increases the contrast between high and low factor loadings, to establish a pattern matrix. A linear transformation of the orthogonal factor matrix that produces a best fit to the pattern according to a squared sum of error criterion is then found. This is referred to as a procrustean transformation.

2. ICA and Factor Analysis

The full Factor Analysis procedure may be summarized:

- 1) Preprocess data – remove outliers, normalize to unit variance,
- 2) Form the sample covariance matrix,
- 3) Estimate communalities via multiple regression,
- 4) Replace the diagonal terms of the sample covariance matrix with the estimated communalities from (3) and calculate the eigen-decomposition of this communality-covariance matrix,
- 5) Select how many factors to keep based on any of the standard criteria (Kaiser, scree, etc.),
- 6) Calculate a rotation to simplify factor complexity using one of the Orthomax criteria,
- 7) If further factor simplification is desired for interpretability and correlation amongst factors is acceptable, calculate an oblique transformation of the rotated factors.

There are many optional methods, criteria, and variations at each step of the process.

2.2.2 Exploratory versus Confirmatory Factor Analysis

One of the distinctions that is often made in the FA literature is the difference between exploratory and confirmatory factor analysis. The basic difference is that in exploratory factor analysis few or no priors are brought to the analysis. The basic question asked is what relationships exist in the observed data. In confirmatory factor

2. ICA and Factor Analysis

analysis the basic question is whether a specific model is consistent with the observed data. The CFA is for model testing rather than model building. As such it is a form of covariant structural modeling or structural equation modeling.

2.3 Independent Components

Whereas PCA provides a maximally parsimonious representation of the data, and FA seeks to maximize interpretability by minimizing the number of factors contributing to each observed variable (localizing the explanation of variance), ICA tends to spread the explanation of variance out amongst a greater number of components. This is the result of the independence rotation smearing the variance of the major principal components across the minor components. The independent components will generally be less efficient but (hopefully) more interpretable than the principal components.

Our approach to interpreting independent components as representing *physical mechanisms* is based on relating the percentage of variance of each observed variable explained by each component to the mechanism effects expected from first principle knowledge of the modeled system. Just as one considers the explanation of variance offered by a factor analysis model, or a PCA, one can consider the distribution of variance implied by an ICA model. If the data has been centered and the individual variables normalized ahead of time and the model estimated so that the sources have unit variance, the elements of the estimated mixing matrix \mathbf{A} may be considered as standardized regression coefficients of the observations on the independent

2. ICA and Factor Analysis

components. The square of each a_{ij} is the percentage of the variance of the observed variable i explained by the source j . From this perspective, ICA is a principled form of exploratory factor analysis.

2.4 The ICA Model and Fundamental Concepts

The standard form of independent component analysis hypothesizes that the observed signals are linear mixtures of independent sources. One normally assumes there is the same number of sources as there are observations. The resulting mixture model for n observed variables is

$$\begin{aligned}x_1 &= a_{11}s_1 + a_{12}s_2 + \dots + a_{1n}s_n, \\x_2 &= a_{21}s_1 + a_{22}s_2 + \dots + a_{2n}s_n, \\&\vdots \\x_n &= a_{n1}s_1 + a_{n2}s_2 + \dots + a_{nn}s_n,\end{aligned}$$

or in vector form

$$\mathbf{x} = \mathbf{A}\mathbf{s}.$$

The classic example of such a process is the “cocktail party problem” of blind source separation (BSS). The situation is that of many people standing around a large room in small groups having individual conversations. If there are n conversations going on and one has n microphones distributed throughout the room (so that no two microphones pick up exactly the same mix of conversations), the problem of separating out each conversation, $s_i(t)$ from the recorded signals $x_1(t)$ through $x_n(t)$, is that of estimating the inverse of the mixing matrix \mathbf{A} (if it exists). This linear mixing model is of course idealized in that it does not take into account time delays in the

2. ICA and Factor Analysis

propagation of the sound signals, reverberation or echoes that might add into the recordings, or any additional noise sources (external in the room or internal to the recording system).

The key assumption that is made in ICA is that the sources are statistically independent. Two sources, x and y , are independent if they are non-interactive, i.e. their joint distribution factors into the product of their individual densities,

$$p(x, y) = p(x)p(y).$$

This condition holds only if the sources are uncorrelated and all their higher order cross moments are zero. A straightforward way to determine how close to independent two sources are is to calculate the Kullback-Leibler divergence between their joint density and the product of their individual densities, $D(p(x, y) \parallel p(x)p(y))$. For the definition and properties of this divergence and related information theoretic concepts, see Appendix 1.

The standard ICA approach to estimating $\mathbf{W} = \mathbf{A}^{-1}$ is to hypothesize that the sources are all statistically independent, i.e.

$$D(p(\mathbf{s}) \parallel p(s_1)p(s_2)\dots p(s_n)) = 0.$$

The goal of the estimation process is then to find that demixing matrix \mathbf{W} such that

$$D(p(\hat{\mathbf{s}}) \parallel p(\hat{s}_1) p(\hat{s}_2)\dots p(\hat{s}_n)) = 0,$$

where

2. ICA and Factor Analysis

$$\hat{\mathbf{s}} = \mathbf{W}\mathbf{x}.$$

A key issue in implementing ICA is selecting a method for approximating the divergence $D(p(\hat{\mathbf{s}}) \parallel p(\hat{s}_1) p(\hat{s}_2) \dots p(\hat{s}_n))$. Once this has been decided, what remains is application of some nonlinear optimization method (usually gradient based) to find \mathbf{W} . There may be local minima that interfere with an accurate estimation.

Since independent implies uncorrelated and the principal components are uncorrelated, most ICA implementations begin the analysis by centering the data and then applying a whitening step based on PCA. The eigenvectors found in PCA are an orthogonal basis for the observation space. The linear transformation that changes coordinates from the original observations to the orthogonal coordinates uncorrelates the observations. Furthermore, if each coordinate axis is scaled by one over the square root of its respective eigenvalue, then each of the transformed, uncorrelated variables will have unit variance.

The rest of the analysis consists of finding that rotation of the “whitened” coordinate system that minimizes $D(p(\hat{\mathbf{s}}) \parallel p(\hat{s}_1) p(\hat{s}_2) \dots p(\hat{s}_n))$. In other words, if the whitening transformation is denoted by the matrix \mathbf{V} what remains is to find the orthogonal matrix \mathbf{B} such that $\mathbf{W} = \mathbf{B}\mathbf{V}$. Notice that a non-orthogonal \mathbf{B} would have the effect of recorreling the observations, and thus could not produce independent sources. As the space of $n \times n$ orthogonal matrices has dimension $n(n-1)/2$, this optimization problem is somewhat more constrained than the original search in n^2 dimensions. The last architectural detail in ICA is whether all the components of \mathbf{B} are

2. ICA and Factor Analysis

estimated simultaneously, often called *symmetric* estimation, or the components are estimated one at a time, the *deflationary* approach.

2.5 Two-Source BSS Example

We can illustrate the basic ICA procedure with a simple two source example. To emphasize the point that ICA performs source separation based on statistical characteristics, the sources in our example will be characterized only by their density functions. Consider the sources s_1 and s_2 with densities

$$p(s_1) = \frac{\lambda}{2} e^{-\lambda|s_1|},$$
$$p(s_2) = \begin{cases} 1, & \text{if } 0 \leq s_2 \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

The first variable has a Laplacian density, which is an example from the supergaussian or leptokurtic family of distributions. The second variable is uniformly distributed, which is an example from the subgaussian or platykurtic family of distributions. It is important that both source variables have non-gaussian distributions, otherwise we would be no more effective separating the sources using ICA than with PCA. A scatter plot of 1000 samples drawn from the joint distribution $p(s_1)p(s_2)$ is illustrated in Figure 2.4.

Define the mixing matrix

$$\mathbf{A} = \begin{pmatrix} 3 & 10 \\ 5 & 3 \end{pmatrix},$$

2. ICA and Factor Analysis

and consider the set of observations $\mathbf{x} = \mathbf{A}\mathbf{s}$. The observed joint distribution is shown in Figure 2.5.

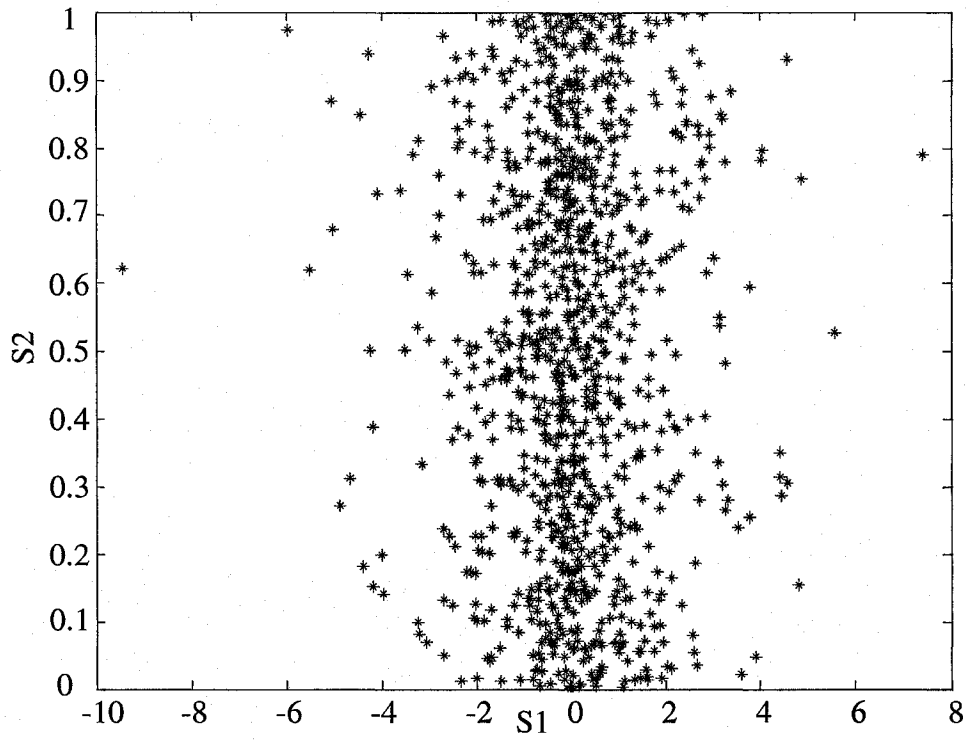


Figure 2.4: The joint distribution of the independent sources s_1 and s_2 .

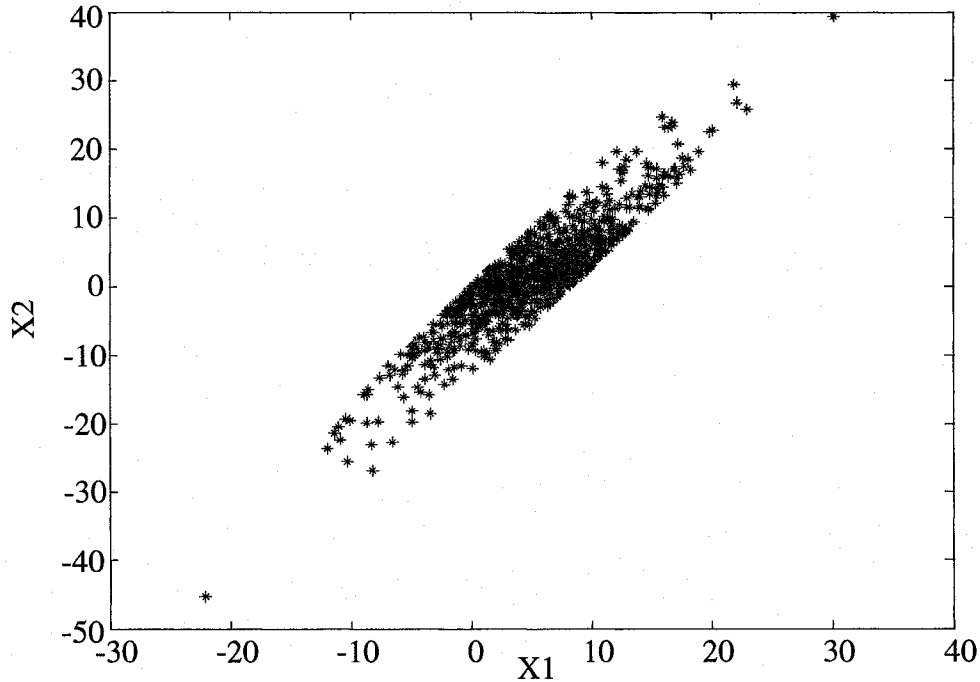


Figure2.5: The joint distribution of the observed mixtures x_1 and x_2 .

It should be clear that the mixed variables are highly correlated. After centering, the observed signals can be whitened using the eigen-decomposition of their covariance matrix. Let \mathbf{C}_X be the covariance matrix of the centered observations,

$$\mathbf{C}_X = E\{\mathbf{xx}^T\}.$$

Let \mathbf{D} be the diagonal matrix containing the eigenvalues of \mathbf{C}_X in descending order, and \mathbf{E} be the matrix whose columns are the unit norm eigenvectors, ordered in a like manner. These matrices are obtained in PCA, and due to the positive definiteness of \mathbf{C}_X , \mathbf{E} is guaranteed to be orthonormal. Define the whitening transformation \mathbf{V} by

$$\mathbf{V} = \mathbf{D}^{-1/2} \mathbf{E}^T.$$

Since

$$\mathbf{C}_X = \mathbf{E}\mathbf{D}\mathbf{E}^T,$$

if we define

$$\mathbf{y} = \mathbf{V}\mathbf{x},$$

we have

$$\mathbf{E}\{\mathbf{y}\mathbf{y}^T\} = \mathbf{V}\mathbf{E}\{\mathbf{x}\mathbf{x}^T\}\mathbf{V}^T = \mathbf{D}^{-1/2}\mathbf{E}^T\mathbf{E}\mathbf{D}\mathbf{E}^T\mathbf{E}\mathbf{D}^{-1/2} = \mathbf{I},$$

thus justifying the name. Note that this transformation is not unique, as it may be composed with any orthogonal transformation (such as the estimated \mathbf{B}) and retain its whitening property. The joint distribution of the whitened observations appears in Figure 2.6.

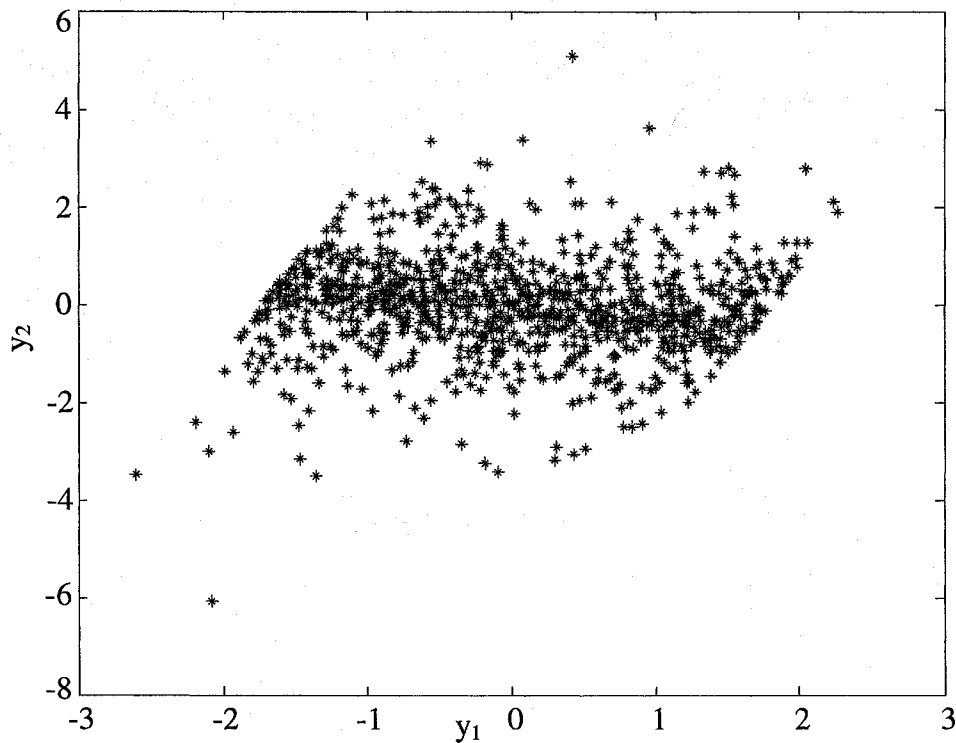


Figure 2.6: The joint distribution of the whitened mixtures y_1 and y_2 .

2. ICA and Factor Analysis

While the whitened observations are no longer correlated, they are clearly not independent. For example, observing a particularly negative value for the first variable (a value below -1.8 for instance) reduces the uncertainty in the value of the second variable (implies the second value will be negative). Since knowing the value of one variable provides information about the value of the other variable, the two variables are not independent. The edges of the uniformly distributed source can be seen as the crisp diagonal edge on the right and left sides of the scatter plot. The orthogonal matrix that we need ICA to find will rotate these edges so that they align with one variable axis or the other. A standard ICA method, FastICA, finds the demixing matrix

$$\mathbf{W} = \begin{pmatrix} -.0535 & .1672 \\ .4181 & -.2499 \end{pmatrix}.$$

Note that this estimated \mathbf{W} is not in fact the inverse of the true mixing matrix \mathbf{A} . In this case, we have

$$\mathbf{WA} = \begin{pmatrix} 0.6759 & -0.0328 \\ 0.0049 & 3.4313 \end{pmatrix}.$$

While this is not the identity matrix, the off diagonal elements are small implying that the estimated sources \hat{s}_1 and \hat{s}_2 are fairly pure representations of s_1 and s_2 , while the non-unity of the diagonal elements implies that we have not recovered the relative scales of the sources with respect to each other. In general, the variance, sign and ordering of the

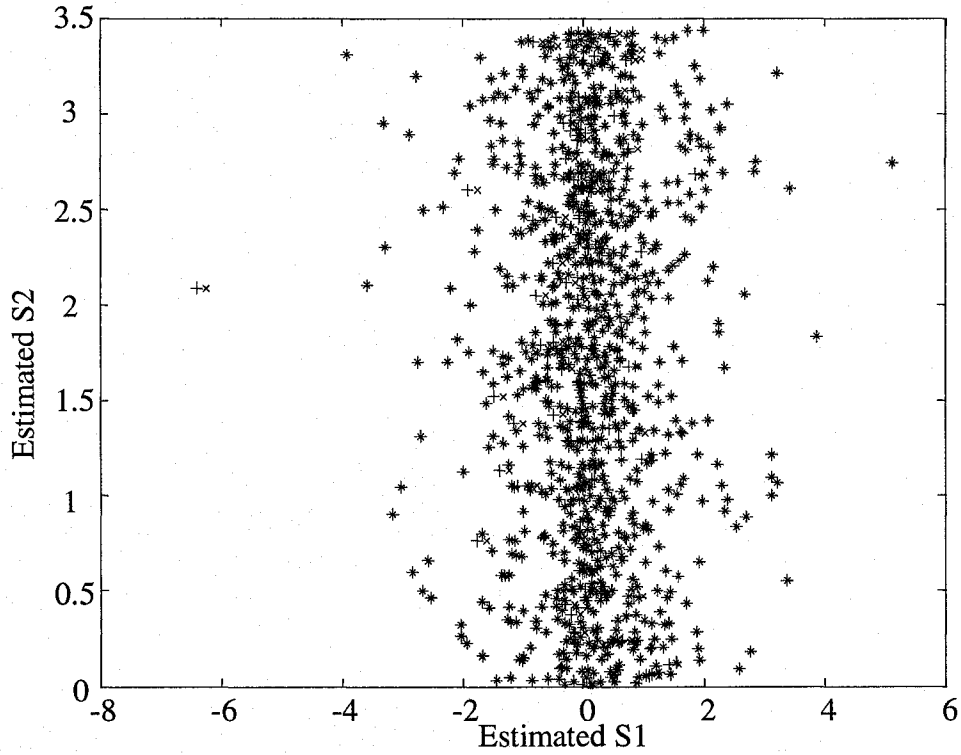


Figure 2.7: The joint distribution of the estimated sources \hat{s}_1 and \hat{s}_2 .

estimated sources is arbitrary, since independence of the sources is invariant to permutations and scalar multiples of the rows of \mathbf{B} . The joint distribution of \hat{s}_1 and \hat{s}_2 is shown in Figure 2.7. Comparing Figure 2.7 to Figure 2.4, we see that they are virtually identical except for scaling: the \hat{s}_1 axis scale has changed by a factor of 0.6759 and the \hat{s}_2 axis scale has changed by a factor of 3.4313.

2.6 Standard Approaches to Implementation

Practical implementations of ICA face three architectural issues: method for approximating the divergence between the joint and factored source distributions,

2. ICA and Factor Analysis

method for adjusting the mixing parameters, and method for constraining the optimization. The choice of optimization method primarily affects computational speed and applicability (online versus offline), and hence is the significant factor determining in what scenarios the algorithm can be effective. The following discussion highlights the major approaches that have been developed over the last decade for implementing ICA.

2.6.1 Algebraic Methods

Since a joint density will factor exactly when all the cross terms in its cumulant tensors are zero, one approach is to estimate specific cross-cumulants for the proposed source variables. For example, the second order cumulant tensor is the covariance matrix and the cross-cumulants are the covariances. Two independent distributions will have zero covariance, and diagonalizing the covariance matrix is a straightforward algebraic problem solved in PCA. Several ICA algorithms have been proposed that apply this principle to the fourth-order cumulant tensor. This is essentially an eigenvalue decomposition problem for an $n^2 \times n^2$ matrix. The method of fourth-order blind identification (FOBI) does just this. The diagonalization can either be done using standard QR factorization methods, which correspond to symmetric or simultaneous estimation of the components, or via the power method for finding individual eigenvectors, which corresponds to deflationary or one at a time estimation of the components. For details of both numerical approaches to finding eigendecompositions, see (Press et al. 1986).

2. ICA and Factor Analysis

In the case of even moderately large n this algorithm becomes both computationally and memory intensive. FOBI is also limited by its inability to resolve components with identical eigenvalues. This latter problem was resolved with the development of the popular joint approximate diagonalization of eigenmatrices or JADE algorithm (Cardoso 1993). This algorithm attempts to find a transformation that simultaneously diagonalizes all the eigenmatrices of the cumulant tensor. Note that our fourth-order tensor is a linear transformation in a space of $n \times n$ matrices, thus its eigenvectors are $n \times n$ matrices. In practice, these eigenmatrices are not jointly diagonalizable, but it can be shown that the demixing matrix that maximizes the sum of squares of the almost diagonal eigenmatrices also minimizes the cross-cumulants and thus produces sources that are as independent as possible. Aside from the computational limitations of these algebraic methods, the estimation of high order cumulants from finite data samples is problematic. Often times the estimated values depend almost entirely on a few extreme valued observations (outliers) and thus are not robust.

2.6.2 Estimators of Independence

Alternatives to the tensor methods involve actually estimating the Kullback-Leibler divergence, which essentially requires the estimation of differential entropy from data.

2. ICA and Factor Analysis

We have

$$\begin{aligned}
 D\left(p(\mathbf{s}) \parallel \prod_i p(s_i)\right) &= \int p(\mathbf{s}) \log \frac{p(\mathbf{s})}{p(s_1)p(s_2)\dots p(s_n)} ds_1 ds_2 \dots ds_n, \\
 &= \int p(\mathbf{s}) \log p(\mathbf{s}) ds - \sum_i \int p(\mathbf{s}) \log p(s_i) ds_i, \\
 &= \sum_i h(s_i) - h(\mathbf{s}),
 \end{aligned}$$

then given the ICA model

$$\mathbf{x} = \mathbf{A}\mathbf{s},$$

and letting

$$\mathbf{B} = \mathbf{A}^{-1},$$

we get

$$D\left(p(\mathbf{s}) \parallel \prod_i p(s_i)\right) = \sum_i h(s_i) - h(\mathbf{x}) - \log|\mathbf{B}|.$$

Now if we constrain the estimated sources to be uncorrelated and have unit variance, i.e.

$$\mathbf{C}_s = \mathbf{I},$$

we have

$$\begin{aligned}
 1 = |\mathbf{I}| &= |\mathbf{C}_s|, \\
 &= |\mathbf{B}\mathbf{C}_x\mathbf{B}^T|, \\
 &= |\mathbf{B}| |\mathbf{C}_x| |\mathbf{B}^T|, \\
 &= |\mathbf{B}|^2 |\mathbf{C}_x|,
 \end{aligned}$$

2. ICA and Factor Analysis

thus $|\mathbf{B}|$ must be constant. Therefore, the only component of the divergence that varies with the demixing matrix is the sum of the marginal source entropies. The marginal entropies are the terms that must be approximated.

Series Expansion Estimators

The classical approach to this problem uses polynomial density expansions, similar to Taylor expansions of functions, to arrive at a parametric form for the marginal densities, which can then be massaged analytically to approximate the differential entropies or their derivatives as desired. This approach supposes that the desired marginal density $p(s_i)$ is near the standard zero mean, unit variance Gaussian density g . Two alternative expansions can be used, the Edgeworth series

$$p(s_i)(x) = g(x) \left[1 + \frac{k_3(s_i)}{3!} H_3(x) + \frac{k_4(s_i)}{4!} H_4(x) + \frac{10k_3^2(s_i)}{6!} H_6(x) + \dots \right],$$

or the Gram-Charlier series

$$p(s_i)(x) = g(x) \left[1 + \frac{k_3(s_i)}{3!} H_3(x) + \frac{k_4(s_i)}{4!} H_4(x) + \frac{k_6(s_i) + 10k_3^2(s_i)}{6!} H_6(x) + \dots \right],$$

where $H_j(x)$ are the Hermite polynomials. Truncating either of these series provides a closed parameterized form for the marginal densities. A number of researchers have proposed using this approach to implement gradient based ICA algorithms (Comon 1989, Amari, et al. 1996, Haykin 1999). Such gradient based methods can be implemented for both online and offline applications and have relatively modest

2. ICA and Factor Analysis

memory requirements. One improvement in the expansion based methods is the inclusion of both even and odd cumulants in the measure of independence, thus they should theoretically be sensitive to a wider range of statistical features.

However, these methods once again rely on estimates of the cumulants and thus are subject to the same criticisms concerning robustness as the algebraic methods. An additional criticism of cumulant based methods is that the high order polynomials in the cumulants concentrate on the tails of the distribution and are minimally affected by features near the center. Thus even if the cumulants could be estimated perfectly, expansion based estimators of differential entropy would still do a poor job.

Maximum Entropy Estimators

An alternative to the cumulant expansion methods is to estimate the entropy of the maximum entropy distribution consistent with the observed data (or in the ICA case, the hypothesized sources). Note that the smaller the K-L divergence is, the closer the sources are to being independent. As the marginal entropies are the positive terms in the divergence, using an approximation method that bounds each of those terms from above is a conservative approach.

From information theory and statistical mechanics, e.g. (Cover and Thomas, 1991), we know that the density $p(x)$ consistent with m moment constraints

$$\int p(x)F_j(x)dx = c_j,$$

that maximizes differential entropy is of the form

2. ICA and Factor Analysis

$$p(x) = \exp\left(A \sum_j a_j F_j(x)\right).$$

In general, it may not be possible to solve for the constants a_j and A , as the maximum entropy bound may not be achievable.

Several research groups have proposed that by purposefully selecting the moment constraints, it is possible to always solve for the constants to a first order approximation, resulting in the “approximative maximum entropy density”. The nonlinear functions $F_i(x)$ in the moment constraints must form an orthonormal set with respect to the metric defined by $g(x)$, i.e.

$$\int g(x) F_j(x) F_k(x) dx = \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{if } j \neq k, \end{cases}$$

$$\int g(x) F_j(x) x^k dx = 0, \quad \text{for } k = 0, 1, 2.$$

In addition, constraints of zero mean and unit variance are added. Under these assumptions, the approximate density is simply

$$\hat{p}(x) = g(x) \left(1 + \sum_j c_j F_j(x) \right),$$

and the resulting entropy bound is

$$h(\hat{p}(x)) = h(g(x)) - \frac{1}{2} \left(\sum_j E\{F_j(x)\}^2 \right) = h(g(x)) - \frac{1}{2} \left(\sum_j c_j^2 \right)$$

(see Hyvärinen 1998, or Cook et al. 1993 for details). The quantity to be minimized over \mathbf{B} then becomes

2. ICA and Factor Analysis

$$-\sum_{i=1}^n \sum_{j=1}^m E\{F_j(\mathbf{b}_i^T \mathbf{x})\}^2.$$

Implementation of this method requires selection of the nonlinear functions. In addition to the orthogonality constraint, the functions should not increase faster than a quadratic to insure the integrability of the approximate density and to avoid being sensitive to outliers. Furthermore, the functions should capture aspects or features of the modeled distribution. Thus if the form of the density were known a priori, i.e. $p(x)$, the best choice would be the $-\log p(x)$.

A variety of functions have been suggested to measure specific features, e.g. the spike of the Laplace distribution

$$F(x) = |x| \approx \frac{1}{a} \log \cosh(ax),$$

or infinitely heavy tails

$$F(x) = -\exp\left(-x^2/2\right),$$

or asymmetry

$$F(x) = x \exp\left(-x^2/2\right),$$

(Hyvärinen, Karhunen, and Oja 2001).

2.6.3 Maximum Likelihood Estimation

An alternative method for estimating BSS models is maximum likelihood estimation. For the basic linear ICA mixing model, the likelihood considered as a function of the demixing matrix \mathbf{B} is

$$L(\mathbf{B}) = \prod_{t=1}^T \prod_{i=1}^n p_i(\mathbf{b}_i^T \mathbf{x}(t)) |\mathbf{B}|,$$

and the log-likelihood is

$$\log L(\mathbf{B}) = \sum_{t=1}^T \sum_{i=1}^n \log p_i(\mathbf{b}_i^T \mathbf{x}(t)) + T \log |\mathbf{B}|.$$

Maximizing the log-likelihood requires the assumption of some functional form for the marginal densities. These density functions could be estimated simultaneously with the demixing matrix, but density estimation is, in general, a much harder problem. Typical implementations constrain the density estimation problem to choosing between a generic subgaussian density

$$p^-(x) = a - \frac{x^2}{2} - 2 \log \cosh x,$$

and a generic supergaussian density

$$p^+(x) = a - 2 \log \cosh x.$$

It should be clear that maximizing the log-likelihood in this circumstance is exactly the same as minimizing the sum of the marginal entropies using the maximum entropy approximation method for the case when the negative of the hypothesized log density is used as the moment constraint. Note that if the hypothesized or estimated

2. ICA and Factor Analysis

density is subgaussian instead of supergaussian, or vice versa, the maximum likelihood method will give erroneous results. The initial connection of maximum likelihood estimation to the ICA problem, and its connection to minimizing the KL-divergence of the source s was first made by Bell and Sejnowski (Bell and Sejnowski, 1995) in the context of a gradient based algorithm. It was further developed by Amari and Cichocki into the *natural gradient* algorithm (Amari, Cichocki, and Yang 1996).

2.6.4 Optimization Algorithms

The most basic gradient algorithm (Bell and Sejnowski) is of the form

$$\mathbf{W} = \mathbf{W} + \eta \left((\mathbf{W}^T)^{-1} + E \{ F(\mathbf{W}\mathbf{x}) \mathbf{x}^T \} \right),$$

where η controls the size of each adjustment, and $F(\mathbf{z})$ is either the log-density, or nonlinear correlation, or independence estimator adopted (per the above discussion). The observations are not assumed to be decorrelated, unit variance, thus the full demixing matrix \mathbf{W} is estimated rather than just the orthogonal transformation \mathbf{B} . Convergence with this method is rather slow due to the inclusion of the decorrelation into the gradient optimization, and due to the computational load of the matrix inversion for every adjustment. Furthermore, this algorithm is based on the assumption that the matrix space over which the optimization is performed has a Euclidian structure. This assumption is not, in general, true, therefore the adjustments made by the algorithm are not in fact in the direction of steepest descent.

2. ICA and Factor Analysis

The complication arises from the fact that when tangent vectors are translated, they do not maintain their direction unless the space is flat (e.g. Euclidian), so the deflection over an infinitesimal translation must be accounted for in the gradient calculation. This leads to the natural gradient algorithm

$$\begin{aligned}\mathbf{W} &= \mathbf{W} + \eta \left((\mathbf{W}^T)^{-1} + E\{F(\mathbf{W}\mathbf{x})\mathbf{x}^T\} \right) \mathbf{W}^T \mathbf{W}, \\ &= \mathbf{W} + \eta \left(\mathbf{I} + E\{F(\mathbf{W}\mathbf{x})\mathbf{x}^T \mathbf{W}^T\} \right) \mathbf{W}.\end{aligned}$$

Not only does this algorithm follow the direction of steepest descent, it also avoids the matrix inversion of the basic gradient approach.

As we noted earlier, the observations can be made decorrelated with unit variance using PCA as a preprocessing step. In that case, the demixing matrix estimated by the optimization, \mathbf{B} , is orthogonal, thus

$$\mathbf{B}^T \mathbf{B} = \mathbf{I}.$$

Both the gradient and natural gradient algorithms would then reduce to

$$\mathbf{B} = \mathbf{B} + \eta \left(\mathbf{B} + E\{F(\mathbf{B}\mathbf{x})\mathbf{x}^T\} \right),$$

together with the enforcement of the orthogonality constraint

$$\mathbf{B} \leftarrow (\mathbf{B}\mathbf{B}^T)^{-1/2} \mathbf{B},$$

after every update.

Once the observations have been whitened, it turns out there is a computationally more attractive second order method available. Observe that the above methods all seek to find optima of

$$E\{F(\mathbf{b}^T \mathbf{x})\}.$$

The Kuhn-Tucker condition for such optima under the constraint

$$E\{(\mathbf{b}^T \mathbf{x})^2\} = \|\mathbf{b}\|^2 = 1,$$

is

$$E\{\mathbf{x}F'(\mathbf{b}^T \mathbf{x})\} - \lambda \mathbf{b} = 0.$$

The Jacobian matrix of the left hand side considered as a function of \mathbf{b} is

$$J(\mathbf{b}) = E\{\mathbf{x}\mathbf{x}^T F''(\mathbf{b}^T \mathbf{x})\} - \lambda \mathbf{I} \approx (E\{F''(\mathbf{b}^T \mathbf{x})\} - \lambda) \mathbf{I},$$

where the approximation is made using the assumption that the data is sphered, i.e.

$$\mathbf{C}_x = \mathbf{I}.$$

Thus an approximate iteration for a Newton method solution to the Kuhn-Tucker constraint would be

$$\mathbf{b}^+ = \mathbf{b} - \frac{E\{\mathbf{x}F'(\mathbf{b}^T \mathbf{x})\} - \lambda \mathbf{b}}{E\{F''(\mathbf{b}^T \mathbf{x})\} - \lambda},$$

which can be simplified to

$$\mathbf{b}^+ = E\{\mathbf{x}F'(\mathbf{b}^T \mathbf{x})\} - E\{F''(\mathbf{b}^T \mathbf{x})\} \mathbf{b}.$$

This leads to the FastICA algorithm (Hyvärinen and Oja, 1997):

1. Choose an initial component direction \mathbf{b} .
2. Let $\mathbf{b}^+ = E\{\mathbf{x}F'(\mathbf{b}^T \mathbf{x})\} - E\{F''(\mathbf{b}^T \mathbf{x})\} \mathbf{b}$.
3. Let $\mathbf{b} = \mathbf{b}^+ / \|\mathbf{b}^+\|$.
4. Repeat steps 2 and 3 until \mathbf{b} does not change.

This second order method does not require an explicit matrix inversion due to the assumption that the data is sphered and thus the Jacobian is approximately diagonal. It can be applied either to the divergence based objective function or to the likelihood function, and has been shown to have cubic convergence

2.7 BSS of Sources with Time Structure

If the sources in a BSS problem are signals in time rather than just random variables, the correlation structure across time can be used to separate them. In principle, as long as the time dependencies of the signals are not identical, i.e. there exists some lag τ for which the autocorrelation functions differ, the sources can be separated using only second order statistics. If the sources have time structure, the basic ICA model becomes

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t).$$

We form the time lagged covariance matrix

$$\mathbf{C}_X(\tau) = E\{\mathbf{x}(t)\mathbf{x}(t-\tau)^T\},$$

which in principle will be symmetric and whose eigendecompositions would provide the desired demixing transformation if all the eigenvalues are distinct. In practice, the estimated autocovariance matrix may not be symmetric, so instead we work with

$$\bar{\mathbf{C}}_X(\tau) = \frac{1}{2}[\mathbf{C}_X(\tau) + \mathbf{C}_X(\tau)^T],$$

2. ICA and Factor Analysis

which is symmetric by construction. This is the strategy behind the simple algorithm known as AMUSE (Tong, et al. 1991, Molgedey and Schuster, 1994).

The principle drawback to AMUSE is that its success depends on picking a lag value for which the eigenvalues are distinct. A generalization of this idea that is less apt to cause repeated guessing for a good time lag, attempts to jointly diagonalize the averaged, estimated autocovariance matrices for several time lags. This algorithm is known as second order blind identification or SOBI (Belouchrani, et al. 1997).

2.8 Blind Deconvolution

Another related problem context where ICA techniques can be applied is for blind deconvolution (BD). In this context, an observed signal is assumed to be generated by a convolution model

$$x(t) = \sum_{k=-\infty}^{\infty} a_k s(t-k)$$

and the goal is to find a finite impulse response (FIR) deconvolution filter

$$s(t) = \sum_{k=L}^n w_k x(t-k)$$

that recovers the source signal. If the convolving process is minimum phase, then the temporal whitening filter is the deconvolving filter. If the process is not minimum phase (stable and invertible), then a more sophisticated approach is necessary.

Bussgang methods are the earliest approach to BD. Bussgang algorithms look for noncausal filters

2. ICA and Factor Analysis

$$s(t) = \sum_{k=-L}^L w_k x(t-k)$$

using gradient techniques to minimize a heuristically defined cost function (Godard, 1980).

An ICA formulation of the problem defines an observation vector

$$\mathbf{x}(t) = [x(t), x(t-1), \dots, x(t-n+1)],$$

and a source vector

$$\mathbf{s}(t) = [s(t), s(t-1), \dots, s(t-n+1)],$$

with the basic linear mixing model

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t).$$

The ICA solution is then a causal FIR filter (Torkkola 2000, Amari, et al. 2000).

2.9 Historical Development of ICA

One of the first instances of a blind source separation problem being solved by a statistical learning process using independence as an objective was the mapping of observed muscle contraction signals to the angular position and velocity of a joint (Hérault et al. 1982-85). The Hérault-Jutten algorithm adapted the feedback coefficients in a simple two input, two output recurrent network to make the network's outputs independent. This was done by adjusting the feedback coefficients to minimize the nonlinear correlation between the outputs. While this method works for small problems, it does not scale well. A local stability proof exists for convergence (Sorouchyari 1991), but global convergence is not guaranteed.

2. ICA and Factor Analysis

In the late 1980s, interest emerged in high-order spectral analysis. Early papers on the spectral approach to ICA appeared in 1989 (Comon 1989, Cardoso 1989). This approach was based on the algebra of cumulant tensors, which eventually lead to the development of the JADE algorithm (Cardoso 1993). The term independent component analysis appears to originate in a 1994 paper by Comon (Comon 1994).

During the early 1990s, the Héroult-Jutten algorithm was improved and extended by Cichocki, Unbehauen, *et alia*. While still based on the heuristic of nonlinear decorrelation, this approach utilizes a feedforward network, does not require a potentially ill-conditioned matrix inversion, and scales to larger problems.

In 1995, Bell and Sejnowski published their approach to ICA based on the infomax principle (Bell & Sejnowski 1995). The Bell-Sejnowski algorithm is a simple gradient method for maximizing the likelihood of an ICA model. While computationally intensive, this algorithm attracted a lot of attention to ICA and BSS type problems. It was soon refined into the natural gradient algorithm (Amari, Cichocki, and Yang 1996), which was really an outgrowth of Cichocki and Unbehauen's earlier work.

Starting in 1997, Hyvärinen and coworkers introduced their FastICA algorithms (Hyvärinen 1997, Hyvärinen and Oja 1997, Hyvärinen 1999). These algorithms utilize a fixed point optimization method that posses cubic convergence, and thus tend to be computationally fast for off line applications.

3. Statistical Process Control

Statistical Process Control (SPC) is the practice of using statistical inference to monitor and diagnose the performance of production systems so as to reduce the variability of the system's output. In the manufacturing context, this translates to decreasing the variability in the makeup or function of the finished products or components. In the service industry context this amounts to reducing the variability in the system's attributes from the consumer's point of view. In general the goal is to enhance the system's performance by monitoring the distribution of the measured values of variables that describe significant attributes of the system's outputs.

The general strategy is to characterize the distribution of the system's outputs when the system is functioning as desired (is *in control*), for example the radius of a ball bearing could be characterized as a Gaussian random variable with mean equal to the nominal design radius and known variance, then monitor the statistics of the population of finished products based on regular sampling to see if the actual population produced remains consistent with the nominal, in-control distribution. If the measured population is found to be significantly inconsistent with the nominal distribution, the system is assumed to have shifted internal state so as to be *out of control*, and steps to intervene are then taken to return the system to its in-control state.

When the measured variables are directly related to the quality of the system's output, this practice is formally called *statistical quality control* (Montgomery 2005),

3. Statistical Process Control

which also includes acceptance testing and related practices. SPC is the general practice described above of monitoring the internal system state, and usually includes monitoring process variables (attributes of the process rather than the product) as well as quality variables.

The entire endeavor forms a closed loop control mechanism for a productive system. Such an approach was found to be necessary for the control of manufacturing systems in the early twentieth century. The complexity of the systems to be controlled was such that no single closed feedback loop would have been sufficient to maintain the stability of the system in the in-control state, as such systems routinely involved multiple steps, multiple inputs of varying qualities and highly variable human operators.

Historically, SPC grew out of sampling techniques for quality assurance. Basic techniques for acceptance testing can be used for process monitoring. For example, a testing regime designed for testing incoming lots of component assemblies from a supplier to determine if they comply with design specification before incorporating them into finished products could just as well be used by the upstream supplier to see if their assembly process is performing properly.

The field of Quality Engineering considers the various permutations of such testing and their application to increase the quality of finished products. The variables measured in such practice are termed *quality variables* and monitoring such variables is correctly termed Statistical Quality Control. SPC can be thought of as a

3. Statistical Process Control

generalization of these methods that monitors any variable (i.e. not just quality variables) that indicates the internal state of the system to be controlled.

The core techniques of SPC monitor process outputs. There are many cases where input variability needs to be considered to distinguish between internal and external process disturbances. As our proposed methodology is currently developed to apply only to process outputs, the following overview of SPC will only cover process output monitoring techniques. Methods for input-output modeling such as Partial Least Squares (PLS) and model based methods will not be discussed for the sake of brevity. Likewise the myriad time-dependent techniques that correct for autocorrelation will be omitted. Finally the distinction between quality variables and process variables is not made, as it is not germane to the proposed methodology.

3.1 Univariate Control Charts

A simple SPC example is the classic Shewart \bar{X} chart (Shewart 1931). This simplest of all monitoring charts plots the average value of a fixed number of samples drawn from the process output at regular intervals. Alarm and warning limits are calculated for the process based on the variance of the in-control process together with a sensitivity level for detecting process change. From a statistical point of view, the chart implements a straight forward hypothesis test about the mean of the observed process. The null hypothesis is that the sample was drawn from a population with a mean equal to the nominal process mean, and the α level or probability of type I error determines the alarm threshold. A type I error represents a false alarm for the

3. Statistical Process Control

monitoring system, while a type II error represents a missed detection. The α level chosen determines the false alarm rate for the monitoring system, which in practice determines the expected run time for the process, i.e. the expected length of time the process will operate while in control before a false alarm is generated. This is a tunable parameter, and is selected based on the relative cost of false alarms versus missed detections.

Now, if we postulate a process with a normal distribution, a nominal mean of 50.0 and a process variance of 0.01, and sample the process output 5 times an hour, we could obtain a dataset as in Table 3.1. The mean values of the samples within each hour are tallied in the right hand column. The distribution of the sample means has variance

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{.01}{5} = 0.002.$$

The control limits for a Shewart chart based on the hourly sample means are

$$UCL = \mu_{\bar{x}} + L\sigma_{\bar{x}} = 50 + (3)\sqrt{0.002} = 50.1342,$$

$$LCL = \mu_{\bar{x}} - L\sigma_{\bar{x}} = 50 - (3)\sqrt{0.002} = 49.8658,$$

where L is the distance of the control limit from the mean in standard deviations.

Based on the 3 sigma control limit ($p = 0.0027$), the average run length will be

$$ARL = \frac{1}{p} \cong 370$$

3. Statistical Process Control

so we can expect the process to run in control for an average of 370 hours before generating a false alarm.

Table 3.1: Data for control chart example

Hour	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Mean
1	49.9369	50.0485	49.8971	49.9104	50.0811	49.9748
2	49.7675	49.9995	50.0243	50.0135	50.0636	49.9737
3	49.8768	49.9724	49.8743	49.9861	50.1310	49.9681
4	50.1056	50.1276	49.9653	49.8837	50.0327	50.0230
5	49.9887	50.1863	49.9059	50.1184	49.9327	50.0264
6	50.0379	49.9477	49.8825	49.9985	49.9851	49.9703
7	50.0944	50.0103	49.8979	50.0536	49.7551	49.9623
8	49.7880	49.9192	49.9598	49.9284	50.0473	49.9285
9	49.9355	50.0680	50.0174	49.9344	50.0117	49.9934
10	49.9296	49.7635	49.9884	50.1314	50.0409	49.9708
11	49.9982	50.1990	50.2064	50.1107	50.0345	50.1098
12	50.0818	50.1219	50.0755	50.2848	49.9919	50.1112
13	50.2521	50.1262	49.9482	50.0725	50.0952	50.0988
14	50.0962	50.2213	50.1010	50.3213	50.1379	50.1755
15	50.2227	50.0725	50.1071	50.2509	50.0670	50.1440
16	50.0304	50.0867	50.1317	49.9055	50.0500	50.0408
17	50.1008	49.9729	50.1500	49.9319	50.0964	50.0504
18	50.0217	49.9336	50.2278	50.0426	50.0825	50.0617
19	50.1587	50.0296	50.0452	50.0814	50.0043	50.0638
20	50.0749	50.1281	50.1261	50.1009	50.2293	50.1318

In this example, a process change occurs between the third and fourth sample during the tenth hour that causes the process mean to shift up by 0.1. Two versions of a Shewart chart for the process mean are given in Figures 3.1 and 3.2. Figure 3.1 is a simple chart of the individual process measurements. Figure 3.2 charts the hourly mean of the process measurements. It is standard practice to organize process samples into logical groups and chart the means, ranges, variances, etc. of the group. This has

3. Statistical Process Control

the advantage of increasing the time between false alarms as far fewer tests are made, though no alarm can be generated until all the samples for the group have been gathered. Both charts alarm in the fourteenth hour (sample 69).

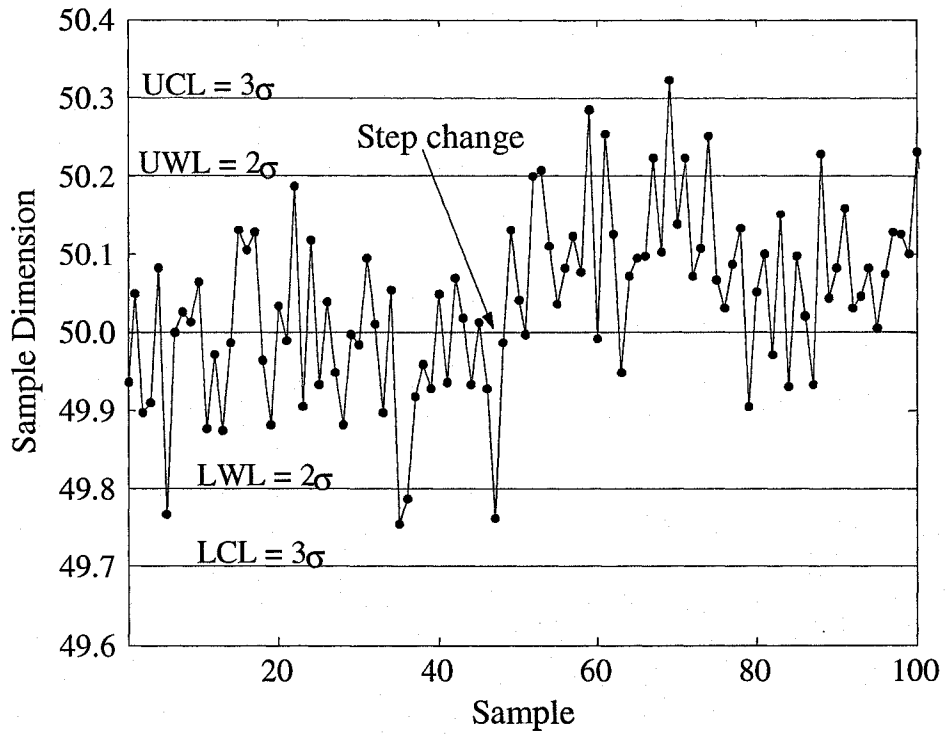


Figure 3.1: Twenty hours of samples plotted with 2 and 3 σ limits.

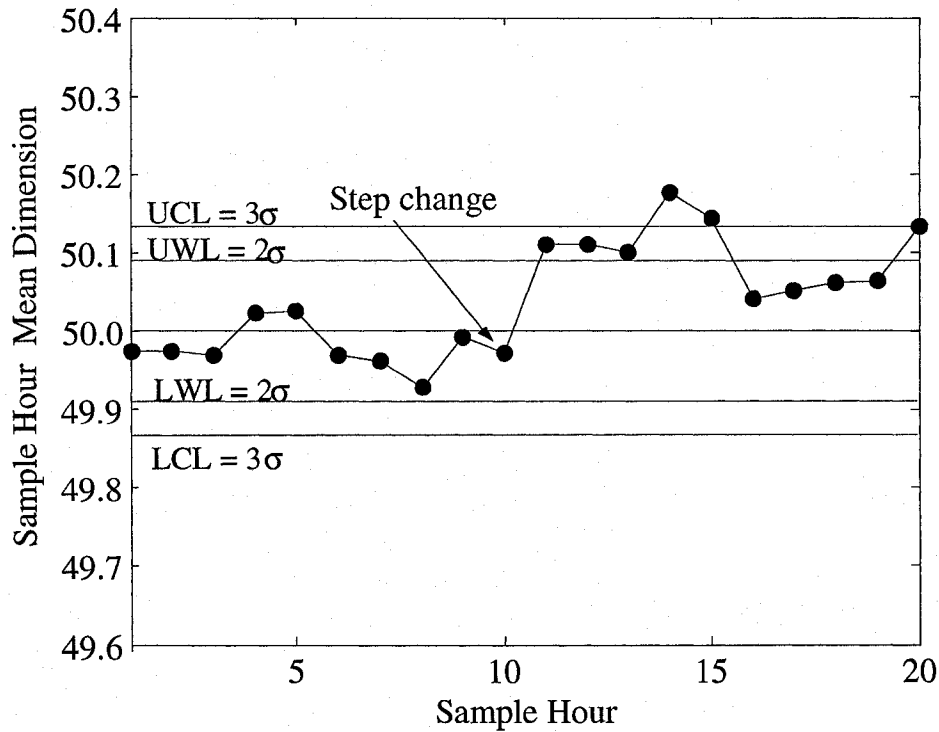


Figure 3.2: Shewart \bar{X} chart with 2 and 3 σ limits.

An alternative to the Shewart chart is the Cumulative Sum (CUSUM) chart, (Page 1954, Hawkins and Olwell 1998). The cumulative sum for a process with target mean μ_* is

$$C_i = \sum_{j=1}^i (\bar{x}_j - \mu_*),$$

$$= C_{i-1} + (\bar{x}_i - \mu_*).$$

If $\bar{x} \sim N(\mu_*, \sigma^2)$ then $C_i \sim N(0, n\sigma^2)$, the cumulative sum is a zero mean random variable. If at $i = 0$ a step change occurs in the process mean of size d , then we have instead $C_i \sim N(nd, n\sigma^2)$. A steadily increasing or decreasing CUSUM indicates a shift

3. Statistical Process Control

in the process mean. This is operationalized by forming two one-sided Tabular CUSUMs

$$C_i^+ = \max(0, x_i - (\mu_* + k) + C_{i-1}^+),$$

$$C_i^- = \min(0, x_i - (\mu_* - k) + C_{i-1}^-),$$

where k is one-half the magnitude of the expected shift in the mean. If either one-sided CUSUM exceeds an alarm limit h , a shift in process mean is indicated. As with the Shewart charts, the false alarm rate is controlled by the selection of the alarm limit. Unlike the Shewart charts, CUSUMs are designed to be sensitive to a particular magnitude of shift in parameter value. This magnitude is usually expressed in terms of standard deviations for the monitored variable, e.g. a CUSUM chart is developed to monitor for a 1.5σ change in the process mean. A CUSUM chart for our example process is in Figure 3.3. This chart detects the step change in the twelfth hour (sample 58), 2 full hours before the Shewart chart. The CUSUM was designed with $k = 0.1$ and an alarm limit based on 3σ . The ARL for this chart is comparable to that of the Shewart chart.

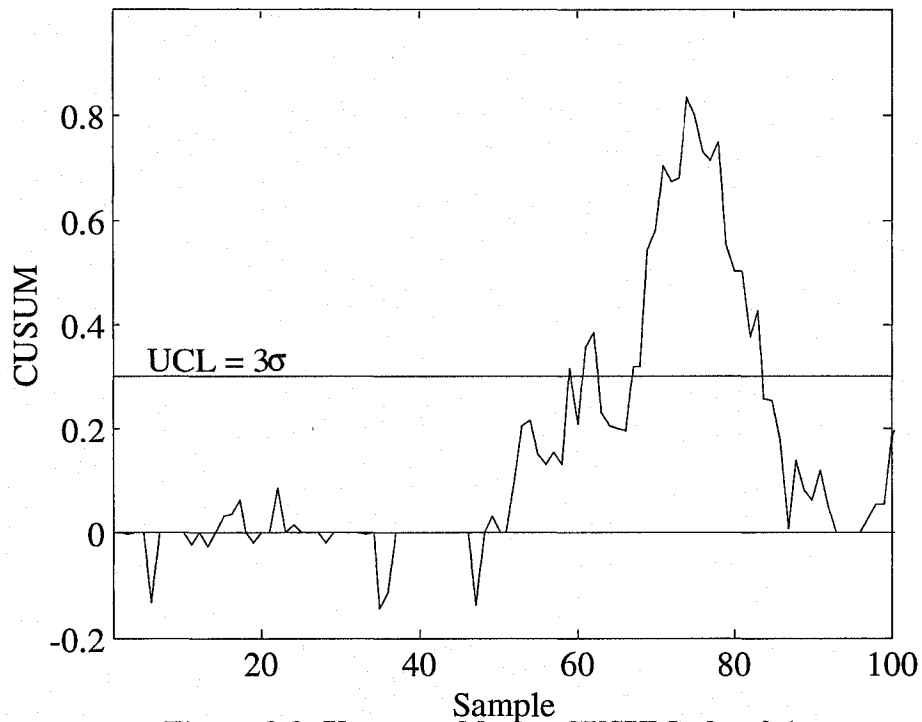


Figure 3.3: Upper and lower CUSUMs, $k = 0.1$.

3.2 Optimality and Design Issues

Which is better a Shewart cart or a CUSUM? Are either of them optimal or is there some other superior method? The proper form of the question is: given all procedures for determining if a process has shifted from one known distribution to a second known distribution and identical false alarm rate, which procedure has the smallest expected time to signal a change? Moustakides proved that CUSUM are the optimal detector for this problem (Moustakides 1986). Put another way, among all techniques with a given false alarm rate, the optimal detector *for a known step change*

3. Statistical Process Control

in a process parameter is a CUSUM chart. While any particular CUSUM is only optimal for a specific magnitude step change, their performance is robust as long as the change is not too far from the designed magnitude.

However, CUSUMs do have several weaknesses. They can be slow to respond to large shifts. They can fail to respond to transient shifts. Both of these cases are better handled by Shewart charts. For example, the optimal test for a one time step transient of unknown size will be a Shewart chart. Therefore CUSUMs and Shewart charts are often used together.

Another consideration when designing a process monitoring scheme is the complexity of implementation. No matter what type of chart is to be used, one must first characterize the nominal, "in control" distribution of the monitored variable. The general observation applicable in all cases is that large samples are always better. Mischaracterization of the parameters variance can radically affect the false alarm rate and sensitivity of any monitoring scheme. A variety of procedures, rules and methods have been developed for both Shewart charts and CUSUMs in many process contexts, e.g. (Montgomery 2005). Most of these apply to variables with normal distributions. Additional work is needed when implementing charts for non-normal distributions. The computational cost of implementing a monitoring scheme is also a consideration. For example, Shewart R charts monitor the range of values occurring within a logical group and are very cheap to implement compared to variance monitoring.

3. Statistical Process Control

When implementing a new production process, it has become common to use an experimental design methodology to characterize the major factors affecting process outputs. Candidate factors are identified and experiments are run that systematically vary the factors so as to capture the effects of both individual factors and factor interaction. The standard framework uses analysis of variance (ANOVA) to analyze the data from the experiments.

3.3 Multivariate Approaches

Let \mathbf{X} be an n dimensional normal random variable $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with known mean and covariance. We may test the null hypothesis

$$H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$$

against the alternative that the mean has shifted

$$H_1: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0.$$

We reject the null hypothesis if

$$T^2 = (\mathbf{X} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}_0) \sim \chi_n^2$$

is sufficiently large. This is the optimal affine invariant test for an unknown shift in the process mean. However, if the alternative hypothesis specifies a direction for the shift in process mean $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \boldsymbol{\varepsilon}_\lambda$,

$$H_1: \boldsymbol{\mu} = \boldsymbol{\mu}_1,$$

then the optimal test statistic is

3. Statistical Process Control

$$Z = (\mathbf{X} - \mu_0)' \Sigma^{-1} (\mu_1 - \mu_0) \sim \chi_1^2.$$

It is important to note that only the direction of the disturbance of the mean is significant, the magnitude of the shift makes no difference to the reduction in degrees of freedom.

In practice T^2 , Hotelling's statistic, is used in a Shewart chart manner (Hotelling 1947). An estimate S based on the sample covariance matrix of the in-control process gives the test statistic

$$T^2 = (\mathbf{X} - \mu_0)' S^{-1} (\mathbf{X} - \mu_0) \sim \chi_n^2,$$

when a large enough sample size is available for estimating S . This is the fundamental tool used for multivariate process monitoring.

One major practical difficulty with this approach is that T^2 alarms are hard to interpret. A variety of approaches have been tried, starting with running univariate control charts on all the individual variables in the hope that only a small subset will alarm with the T^2 . As most multivariate control problems have a significant degree of correlation between measured variables, this approach tends to be swamped by uninformative false positives. A number of authors have suggested using different control limits for the univariate charts to limit the false positives (Alt 1985, Hayter and Tsui 1994), but these approaches still avoid directly linking alarms to causes. A more direct approach is to calculate the relative contribution of each variable to the alarm via

3. Statistical Process Control

$$d_i = T^2 - T_{(i)}^2,$$

where $T_{(i)}^2$ is the test statistic calculated by omitting the i th variable (Runger, Alt and Montgomery 1996). A large value of d_i is an indication that variable i is responsible for the alarm.

Another issue with T^2 charts is that they are inherently Shewart charts. No clean CUSUM implementation exists for Hotelling statistics. One issue is that T^2 is a quadratic variable and so the linear drift interpretation of the CUSUM does not carry over. It is possible to define a related statistic that accumulates on a linear scale, (Crosier 1988), but the optimality arguments that apply to the univariate CUSUM no longer hold.

Another line of development in multivariate process monitoring is the application of PCA to reducing the dimensionality of the problem (Jackson 1980). So long as the disturbances of interest are in the space spanned by the retained components, reducing the dimensionality of the problem reduces the degrees of freedom in the test statistic and increases the power of the test (Scranton et al. 1996). If the data are considered to be multivariate normal, then the principal component scores are independent of each other and can be treated individually with univariate charts. Based on these observations, one common approach is to chart each of the major principal components individually and combine the minor components into a single T^2 chart. Interpretation of alarms is still a problem with this approach. The percentage of variance a component explains of each variable can be used as a

3. Statistical Process Control

contribution plot indicating which measured variables are involved when a particular component generates an alarm. This is about as informative as the relative contribution statistics for T^2 alarms discussed earlier. In both cases a process engineer familiar with the details of the system is required for accurate trouble shooting.

3.4 Advanced Process Control

One of the active areas of research in Advanced Process Control (APC) is the automation of alarm interpretation for multivariate SPC. Several themes are of note. The first approach is based on template matching of alarms to previously identified faults stored in a data base. This approach tends to use PCA for dimension reduction and charting (Kourti and McGregor 1996). Clustering of expert interpreted contribution plots is used to create a library of known faults against which new alarms can be checked (Kourti 2002). A library of diagnosed faults is required and the problem of when to expand the library and what kind of matching criteria to use are key issues. A critical issue in this work is checking the observability of known faults by the monitoring scheme. Kourti *et al.* suggest constructing special Hotelling charts to detect some important types of disturbances.

Another approach for batch processes where data over a time course is available is to attempt to match the eigen decomposition of an alarm to an established library (Singhal and Seborg 2002). Both this and the previous approach fit into a standard machine learning framework, being in essence expert systems for process diagnosis.

3. Statistical Process Control

An alternative APC approach is to divide the variables into subsets associated with different process segments (Ündey and Çinar 2002). PCA is applied to each subset and one or two principal components from each subset are kept. Alarms are then attributed to the process segment that is implicated by the contribution plots. This approach has the advantage of not depending on a data base of historical alarms. However process disturbances in one segment may cause alarms in components associated with other segments due to process structure.

The design issue of needing large samples of nominal process data in order to characterize the process and set appropriate alarm limits, has lead to work that merges data from related processes or products to improve overall monitoring. Data paucity is particularly a problem in industries with short production runs or plants in which multiple products are fabricated simultaneously. Under the assumption that the eigen structures of the multivariate distributions from different products run through a common plant are identical, nominal and disturbance data from the products can be merged via normalizing the observations based on product means and variances (Lane Martin, Kooijmans and Morris 2001, Martin, Morris and Lane 2002). A weighted average of the product sample covariance matrices is used for pooling.

4. ICA Based Exploratory Factor Analysis for E-test Data

4.1 Introduction

This chapter reviews preliminary work using ICA as a method for isolating sources of variance in E-test data. The sources of variation we seek to isolate are fundamental physical phenomena such as lateral diffusion or effective gate length, gate oxide thickness, and channel implant. Such phenomena can be tied directly to individual process steps, and thus process control can be affected. A natural result of our method is that the isolated sources are as statistically independent as possible. We seek to develop a general model of how the manufacturing process generates variability in the observed E-test parameters. Observations of a normally functioning manufacturing system (including well calibrated and poorly calibrated process steps) were assembled, and ICA models were estimated that explained the observations as a function of independent meta-parameters. Meaningful physical interpretations are given to the meta-parameters so that the ICA model provides an explanation of the E-test variability.

The next section provides a brief overview of our methodology, and the following section a description of our data, including variance profile based on PCA. The fourth section includes our results along with a short discussion interpreting the physical significance of the independent components.

4.2 Methodology

In semiconductor manufacturing, the build up of individual circuits is the result of a multi-step process where each step contributes variability to the effectiveness and robustness of the circuits produced. Microelectronic parametric test (E-test) structures are incorporated on production wafers in an attempt to monitor specific physical phenomena in the semiconductor manufacturing process. Although, test structures are designed to isolate specific phenomena, in practice there is usually a high degree of correlation between measured test parameters, due to the influence of multiple variance mechanisms in any structure. This correlation confounds our ability to diagnose production process variation using conventional ANOVA methods on E-test and lot-equipment-history (LEH) data. Traditional statistical reduction methods, such as principal component analysis (PCA), do not help in this context.

Our approach to interpreting independent components as representing physical mechanisms is based on relating the percentage of variance of each observed variable explained by each component to the mechanism effects expected from first principle engineering knowledge. This can be done both after the fact and by including E-test parameters calculated by first principle models that should represent isolated mechanisms, and then seeing how they are represented by the independent components.

The issue of how many independent components (mechanisms) to look for was handled by calculating confidence intervals for the components using a bootstrap

4. ICA Factor Analysis for E-test

technique (Meinecke, et al. 2001, Davison and Hinkley 1997). The confidence intervals allow one to determine which components are significantly different from all the other components, as opposed to those which cannot be reliably distinguished from other components. Insignificant components appear in groups of at least two, and jointly describe “independent subspaces” within which no separation or preferred basis can be calculated. Such subspaces could represent a mixture of minor variance sources, whose sum is too gaussian to separate. Our approach was to collapse these subspaces and force the variance they contain to be represented by either a single component or by the existing significant components. This is accomplished by dropping the most minor components from the PCA basis, and re-estimating the ICA model. Following the re-estimation of the reduced dimension mixing model, confidence intervals were again calculated and the process continued until all the components that remained were significant.

4.3 Data

This work focused on estimating models of 0.25 μ m CMOS transistor test structures embedded in the scribe lines of 8” CMOS wafers. Our raw data set contained over 10,000 test sites from 230 regular production lots of a single ASIC product made by LSI Logic. The data represented all lots, both high and low yield, manufactured during a two month period at a single fabrication facility. In general, data were sampled from 20% of the wafers in a lot, though some lots were under or

4. ICA Factor Analysis for E-test

over represented. No effort was made to censor or oversample lots as part of this study.

At each test site 107 test parameters were available, of which we chose to include eight n -channel measurements:

- saturation current, long-wide channel (ID LW),
- saturation current, short-wide channel (ID SW),
- saturation current, short-narrow channel (ID SN),
- threshold voltage, long-wide channel (VT LW),
- threshold voltage, short-wide channel (VT SW),
- threshold voltage, short-narrow channel (VT SN),
- lateral diffusion (LD),
- gate oxide thickness (TOX W).

Of these variables, the saturation currents and voltage thresholds are directly measured values, while lateral diffusion and gate oxide are values calculated using analytic first principles models. Lateral diffusion was calculated using Whitfield's method (Whitfield 1985, Ng and Brews 1990) from current-voltage relationships measured across a variety of gate lengths. Gate oxide was calculated using a quantum correction to a capacitance measurement (Lo et al. 1997).

The data were cleaned by censoring sites containing one or more out-of-range test measurements (attributable to tester malfunction) or extreme outliers. Less than

4. ICA Factor Analysis for E-test

1% of the sites were removed. Each parameter was then normalized to zero mean and unit variance.

Applying PCA to our data we found that over 95% of the variance for either channel can be explained by four components, Figure 4.1. However, when we consider which components explain which parameters, we find that the principal components are as uninformative as they are efficient, Table 4.1. As the principal components offer little possibility for source isolation, we are forced to look for an additional constraint. Statistical independence is a reasonable constraint to add, in that fabrication process variability arises from variability in the individual steps of the process, and the variability introduced at each step should (hopefully) be independent of the other steps. ICA enables us to implement this constraint.

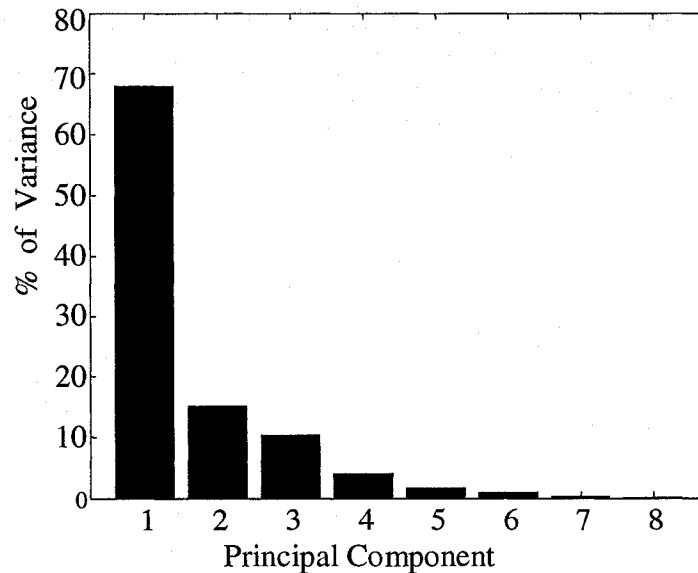


Figure 4.1: Percent of variance explained by principal component.

4. ICA Factor Analysis for E-test

Table 4.1: Percentage of parameter variance explained by each principal component.

	<i>PC1</i>	<i>PC2</i>	<i>PC3</i>	<i>PC4</i>	<i>PC5</i>	<i>PC6</i>	<i>PC7</i>	<i>PC8</i>
ID LW	75	22	1	1	0	1	0	0
ID SW	95	1	0	2	0	1	1	0
ID SN	80	3	0	11	6	0	0	0
VT LW	72	24	0	2	0	1	0	0
VT SW	90	0	5	2	0	3	0	0
VT SN	78	1	7	8	6	0	0	0
LD	24	70	0	5	0	1	0	0
TOX	30	1	68	1	0	0	0	0

4.4 Results

Visual inspection of the parameter distributions suggested that their distributions were both tailed and skewed, so Fast ICA estimation was performed using a hyperbolic tangent moment estimator followed by refinement using a one sided skewness estimator. A symmetric estimation process was used to preserve the ordering of the components during confidence interval estimation. Confidence intervals were developed for every component using a bootstrap approach. This was done by resampling the original data set with replacement to produce new datasets of equal size. The resampling was done at the lot level rather than the wafer or site level, as there should be some dependence between sites on a wafer, and between wafers in a lot. The ICA estimation was repeated for each of the resample datasets using the original model as the starting point for the nonlinear optimization.

4. ICA Factor Analysis for E-test

It is informative to look at the percentage of variance of each test parameter explained by each estimated independent component. Note that this is possible even though we cannot estimate the variance of the sources, because the percentage of variance explained is the product of the source and the model together and is unambiguous. Percentages of parameter explanation by source are in Table 4.2.

Table 4.2: Percentage of parameter variance explained by each independent component.

	<i>IC1</i>	<i>IC2</i>	<i>IC3</i>	<i>IC4</i>	<i>IC5</i>	<i>IC6</i>	<i>IC7</i>	<i>IC8</i>
ID LW	0	0	85	0	0	15	0	0
ID SW	2	28	45	6	1	17	0	1
ID SN	2	31	30	3	0	12	4	17
VT LW	0	1	94	0	1	4	0	0
VT SW	0	38	55	4	0	2	0	0
VT SN	0	31	38	4	0	2	25	0
LD	5	79	1	0	7	6	0	2
TOX W	1	1	2	1	1	94	0	0

In this example, component 1 is not significantly different from components 4 and 5. The three components specify a three dimensional subspace for which there is no preferred ICA basis, i.e. the basis for this subspace does not have a persistent set of basis vectors under bootstrap re-estimation. To reduce this subspace, the most minor component was dropped from the PCA basis and the model and confidence intervals re-estimated. This process was repeated until five significant components were found. Table 4.3 contains the percentages of variance explained by these final n channel model components. Note that the end result of the collapsing or regularization process

4. ICA Factor Analysis for E-test

is that the original components 1, 4, and 5 have been eliminated, while almost all the variance they accounted for is still explained by the five significant components.

Table 4.3: Percentage of parameter variance explained by each independent component, of the reduced model.

	<i>IC1</i>	<i>IC2</i>	<i>IC3</i>	<i>IC4</i>	<i>IC5</i>
	<i>CDW</i>	<i>CDL</i>	<i>Implant</i>	<i>CDW</i>	<i>Gate</i>
	<i>ID</i>			<i>VT</i>	<i>Oxide</i>
ID LW	0	0	85	0	14
ID SW	0	34	47	0	17
ID SN	24	33	31	3	10
VT LW	0	1	94	0	4
VT SW	0	39	56	0	2
VT SN	1	36	39	23	1
LD	1	91	1	1	5
TOX W	0	2	2	1	95

To evaluate whether the isolation accomplished with ICA is a result of having already included the answer in the problem by having included LD and TOX W in the data, the model was also estimated without LD and TOX W. The minor component was dropped from the PCA basis, and five significant independent components were estimated, Table 4.4. The resulting components are very similar to those illustrated above with identical physical interpretations. The mechanisms are clearly present in the drive currents and voltage thresholds alone.

It is clear that a few significant, statistically independent sources of variance account for almost all the variance in the observed test parameters. Based on our knowledge of the device physics and the manufacturing process, the sources can be labeled and related to specific manufacturing steps. Examining the percent of variance

4. ICA Factor Analysis for E-test

of the original parameters explained in light of common understanding of basic semiconductor device physics leads us to suggest that five significant physical mechanisms have been isolated in the eight-parameter transistor model.

For the *n*-channel device, IC5 explains almost all of the gate oxide thickness (TOX W). This component has no effect on the threshold voltage parameters, which makes sense as these are surface channel devices. IC2 explains most of the lateral diffusion parameter and has a strong influence on the short channel drive currents and threshold voltages. It appears to have isolated the poly gate critical dimension influence on the device.

Table 4.4: Percentage of parameter variance explained by each independent component, reduced model, LD and TOX W omitted.

	<i>IC1</i> <i>CDW</i> <i>ID</i>	<i>IC2</i> <i>CDL</i>	<i>IC3</i> <i>Implant</i>	<i>IC4</i> <i>CDW</i> <i>VT</i>	<i>IC5</i> <i>Gate</i> <i>Oxide</i>
ID LW	0	0	85	0	14
ID SW	1	32	46	0	20
ID SN	15	38	30	3	14
VT LW	0	1	96	0	3
VT SW	1	41	56	0	2
VT SN	0	37	39	22	2

IC3 explains significant portions of the threshold and drive current variation without explaining the LD or TOX parameter. This suggests that this component has isolated the channel implant impact. Finally, IC1 and IC4 appear to explain two

4. ICA Factor Analysis for E-test

separate narrow active channel mechanisms. One of these has an effect on narrow channel drive current, while the other affects narrow channel threshold voltage. Neither affects the wide active devices. It is expected that there would be an active critical dimension influence on the narrow active devices. That two different mechanisms were isolated suggests that an additional phenomenon such as field edge profile or gate oxide thinning at the field edge may be influential. This was unexpected, not only has ICA isolated the expected mechanisms, but it has also isolated a separate mechanism previously unrecognized in the test structures.

4.5 Conclusion

A very important point about this modeling technique is that no engineering knowledge or assumptions are placed into the model up front. The ICA technique blindly isolates the phenomena. We have simply interpreted the mechanisms that are isolated by looking at what each ICA component explains about the original parameters. We believe that this method is successful because we know from theory that the underlying physical phenomenon has distinct isolated impacts on device performance. Consider the lateral diffusion parameter, LD, which is not directly measurable from the E-Test structures. LD is a calculated parameter estimated by fitting the relationship of threshold voltage and drive current across devices of several different sizes. It is by first principles of device physics that we know that this method isolates the lateral diffusion with fair accuracy. ICA appears to isolate this mechanism based only on the statistics of the test parameters.

4. ICA Factor Analysis for E-test

E-test parameters are used for process monitoring and diagnosis. Control charts are set for each parameter, both those directly measured and those calculated through analytic models, and alarms set to detect process drift and abrupt change. Our estimated independent components can be utilized in the same manner.

Our approach successfully improves the isolation of specific physical mechanisms in semiconductor E-test parametric data when compared to the isolation provided by specific test structures. The isolated sources have very good agreement with expected physical mechanisms at the device level. Furthermore, the sources can be tied back to specific sets of fabrication steps, enabling improvements in statistical process control and diagnosis.

5. ICA Disturbance Modeling Methodology

5.1 Methodology

Motivated by the results summarized in chapter 4, we formulated a process monitoring methodology using independent components in place of principal components. The advantage of the independent components is that they can be linked to process factors, which can simplify process diagnosis. Kourti's suggestion (Kourti 2002) to check SPC methods to see if they are capable of detecting known faults implies that the appropriate way to validate an independent-component-based scheme is to test the scheme's ability to detect and classify disturbances in its linked factors. The availability of data from factorial experiments provides such an opportunity. Designed factorial experiments are essentially process runs with specific disturbances in one or more factors, so a set of factor identified process control charts should generate alarms specific to each run.

As the factors in designed experiments are purposefully varied to uniformly cover all factor effects and interactions, the designed disturbances are independent sources of variance. If we use the data from the experimental runs to estimate an inverse mixing model, we should obtain a model that isolates the disturbances in separate components. Since we know how many major sources of variance have been designed into the experimental data, we even know how many principal components to keep in our estimation.

5. ICA Disturbance Modeling

One difficulty in using independent component scores for control charts is that by assumption, the scores are not normally distributed. Determining appropriate control limits depends on knowing the distribution of the charted variable. Given that a “large” amount of data is always desired for characterizing nominal distributions, the simplest solution to this problem is to form nonparametric empirical distributions of nominal component scores. Control limits can then be set using the scores given by the percentile corresponding to the desired false alarm rate.

ICA model estimation is an unsupervised learning process. Our method estimates a linear transformation that minimizes mutual information between estimated sources and generates confidence intervals for nominal values for each source. We must then check to see if the resulting control charts correctly identify known disturbances. This validation step was performed both for process data used for ICA and for additional data not included in estimating the inverse mixing process.

The overall procedure for identifying disturbances using ICA is:

1. Collect nominal data and experimental data,
2. Center and normalize each variable using sample means and variances from the experimental data,
3. Apply PCA to the normalized experimental data,
4. Keep only as many principal components from (3) as there are designed disturbance types in the experimental data,

5. ICA Disturbance Modeling

5. Perform ICA to separate principal components into disturbance identified components,
6. Transform nominal data into disturbance identified component loadings using the normalization factors from the experimental data together with the demixing model from (5),
7. Form nominal empirical probability density functions for each disturbance identified component using the scores obtained from (6),
8. Select test values for each disturbance identified component based on a desired probability of committing a type 1 error,
9. Use the ICA transform from (5) and test limits from (8), form \bar{X} and σ^2 control charts for the scores of each disturbance identified component,
10. Evaluate chart performance on the experimental data,
11. Evaluate chart performance on out-of-sample data with known disturbances (if available).

5.2 Data

Continuing in the same application context described in chapter four, our data comes from 8" CMOS wafers with 0.185 μ m CMOS transistor test structures embedded in the scribe lines. The amount of data available from any one experimental lot of wafers is limited. Multiple experiments for any single product are rare so we must pool data in the manner of (Martin, Morris and Lane 2002). The pooling of data

5. ICA Disturbance Modeling

between products manufactured using identical processes has many potential advantages, including proper calibration for short product runs and new products.

Twenty three experimental lots for 18 different products are used. Data was obtained from anywhere from 2 to 25 wafers in each lot. Yield from these lots ranged from 0% to over 90%. In all, 225 wafers are included for model estimation containing a total of 2470 data points for each variable. Two additional experimental lots for two more distinct products, comprising 16 wafers containing 141 data points are used for evaluation purposes. The number of sites sampled on each wafer varies by product. While not affecting the estimation of the ICA model, this variability in sampling impacts the wafer level statistics we generate for the component score control charts. While negligible with respect to our central results, this variability is undesirable from a statistical sampling perspective. In addition to the experimental lots, data from 1022 nominal production wafers containing 9685 data points is used to generate the nominal component score distributions. These nominal wafers are all from a single product included in the experimental lots. The designed disturbances available to us in the e-test context are a two-way factorial analysis of n and p channel implant, each with high, low, and nominal values, a cross-wafer skewing of the lithography process, and a single factor analysis of gate oxide thickness with hi, low, and nominal values.

Based on the exploratory work described in the last chapter and the type of experimental disturbances available, we chose to include eleven measurements from both n -channel and p -channel test structures:

5. ICA Disturbance Modeling

- saturation current, long-wide channel (ID LW),
- saturation current, medium-wide channel (ID MW),
- saturation current, short-wide channel (ID SW),
- saturation current, long-narrow channel (ID LN),
- saturation current, medium-narrow channel (ID MN),
- threshold voltage, long-wide channel (VT LW),
- threshold voltage, medium-wide channel (VT MW),
- threshold voltage, short-wide channel (VT SW),
- threshold voltage, long-narrow channel (VT LN),
- threshold voltage, medium-narrow channel (VT MN),
- gate oxide thickness (TOX).

These 22 parameters should, from an expert's perspective, capture the signatures of the 4 kinds of disturbances included in the experimental lots. Unfortunately, data drawn for p -channel long-narrow saturation current was badly corrupted by truncation, so our analysis is based on the 21 remaining parameters.

5.3 Model Estimation

Each variable was normalized by product to be zero mean by subtracting the product mean from each observation

$$\tilde{z}_i = x_i - \bar{x}_{product}(i).$$

5. ICA Disturbance Modeling

The variance of the combined products was then normalized to be one

$$z_i = \frac{\bar{z}_i}{\sqrt{\text{Var}(\bar{z})}}$$

While (Lane, *et al.* 2001) normalized both parameter mean and variance by product, our normalization is justified by the relatively small sample size of some of our products. The sample mean is relatively easy to estimate and we have the ability to check our estimate against the known nominal parameter value for each product. While estimating variance is in general harder, it is further confounded by some products being represented by lots which contain disturbances designed to increase variance while others do not. In this case, normalizing variance by product would not remove product-to-product variation, but rather experiment-to-experiment variation.

The covariance matrix of the normalized data was then formed and PCA was performed. The proportion of variance explained by the first seven principal components is given in Table 5.1. As there are four designed disturbance factors in our experimental lots, we need at least the first four principal components. The first four components capture 92.4% of the variance in the 21 parameters while the next three components together represent just over 5% of the variance, and the least significant 14 comprise the remaining variance. Keeping only the first four components is therefore not only justified by our hypothesis that our designed disturbances are the significant sources of variance in our data, but is also consistent with the scree and eigenvalue-one criteria commonly used for data reduction.

5. ICA Disturbance Modeling

Table 5.1: Proportion of variance explained by the first seven principal components

Principal Component	1	2	3	4	5	6	7	Total
TOX N	16.3%	1.3%	37.3%	43.3%	1.6%	0.0%	0.0%	99.7%
VT NLW	31.1%	46.1%	7.9%	12.1%	1.4%	0.0%	0.3%	98.8%
ID NLW	36.2%	43.4%	9.8%	7.5%	1.4%	0.0%	0.2%	98.4%
VT NMW	68.2%	23.6%	4.3%	0.7%	0.7%	0.0%	0.1%	97.6%
ID NMW	68.9%	11.7%	10.6%	1.8%	0.8%	1.2%	0.0%	95.0%
VT NSW	69.8%	14.6%	11.9%	0.0%	0.5%	0.0%	0.1%	97.0%
ID NSW	67.7%	7.9%	16.3%	2.7%	0.7%	1.0%	0.0%	96.3%
VT NMN	53.0%	25.7%	0.0%	0.8%	0.9%	13.0%	0.0%	93.3%
ID NMN	53.3%	12.0%	0.3%	2.7%	23.1%	2.8%	1.2%	95.5%
VT NLN	32.6%	44.5%	9.4%	8.7%	1.6%	0.0%	0.4%	97.2%
ID NLN	26.0%	30.3%	16.1%	3.0%	7.4%	11.6%	2.2%	96.5%
TOX P	15.4%	1.3%	37.8%	43.3%	2.0%	0.1%	0.0%	99.7%
VT PLW	13.6%	59.3%	16.9%	8.1%	0.2%	0.1%	0.5%	98.8%
ID PLW	15.3%	57.7%	18.7%	5.9%	0.3%	0.1%	0.6%	98.6%
VT PMW	48.5%	47.5%	1.1%	0.6%	0.4%	0.1%	0.5%	98.6%
ID PMW	60.1%	28.5%	8.1%	0.8%	0.5%	0.1%	0.7%	98.8%
VT PSW	54.2%	35.4%	7.4%	0.0%	0.4%	0.1%	0.4%	97.8%
ID PSW	60.2%	22.2%	13.8%	1.5%	0.4%	0.1%	0.5%	98.6%
VT PMN	39.2%	45.6%	0.7%	0.8%	1.1%	0.0%	10.0%	97.3%
ID PMN	50.2%	30.8%	0.5%	1.9%	8.4%	1.0%	3.8%	96.6%
VT PLN	15.4%	54.6%	20.4%	5.0%	0.2%	0.1%	0.4%	96.0%
Total	42.6%	30.7%	11.9%	7.2%	2.6%	1.5%	1.0%	97.4%

ICA was then performed using the FAST ICA algorithm with a hyperbolic tangent moment function using a scaling parameter of 1.3 and fine tuned using a skewness criterion. The resulting inverse mixing model produced factors which were easily identified with the four types of designed disturbances. The labeled factors explanation of variance is in Table 5.2.

Table 5.2: Independent component explanation of variance

	Gate	N	P	Diffusion	Total
TOX N	97.7%	0.4%	0.1%	0.0%	98.2%
VT NLW	3.7%	93.0%	0.5%	0.1%	97.2%
ID NLW	7.7%	88.9%	0.3%	0.0%	96.9%
VT NMW	2.7%	59.6%	0.1%	34.4%	96.8%
ID NMW	5.5%	30.3%	0.0%	57.1%	92.9%
VT NSW	1.7%	42.0%	0.0%	52.7%	96.4%
ID NSW	4.1%	23.0%	0.0%	67.5%	94.6%
VT NMN	5.8%	58.6%	0.1%	15.0%	79.5%
ID NMN	20.2%	30.0%	0.0%	18.2%	68.3%
VT NLN	6.2%	88.5%	0.5%	0.0%	95.2%
ID NLN	12.8%	61.9%	0.1%	0.6%	75.4%
TOX P	97.3%	0.3%	0.1%	0.0%	97.7%
VT PLW	1.1%	0.3%	94.9%	1.8%	98.0%
ID PLW	2.4%	0.3%	93.4%	1.5%	97.6%
VT PMW	0.5%	0.1%	72.0%	25.1%	97.6%
ID PMW	1.4%	0.0%	43.3%	52.7%	97.5%
VT PSW	0.2%	0.0%	51.9%	44.8%	97.0%
ID PSW	1.0%	0.0%	33.0%	63.6%	97.6%
VT PMN	2.2%	0.0%	73.8%	10.2%	86.2%
ID PMN	11.8%	0.0%	52.1%	19.5%	83.4%
VT PLN	3.3%	0.3%	90.0%	1.7%	95.3%
Total	13.8%	27.5%	28.9%	22.2%	92.4%

5.4 Nominal Distributions

Nominal distributions for each of the component scores were then generated by normalizing parameter measurements from healthy production wafers using the same factors used above and transforming them using the inverse mixing matrix found by ICA. The resulting population of scores is used as an empirical probability distribution for determining critical values for one and two tailed tests. Histograms of the nominal scores appear in Figures 5.1 – 5.8.

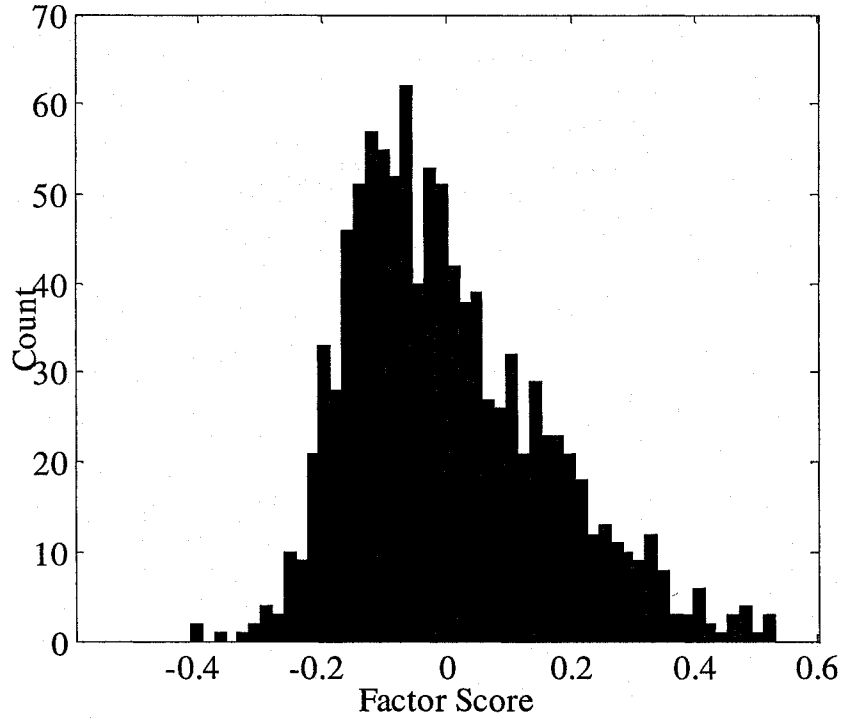


Figure 5.1: Nominal distribution of N implant score wafer mean.

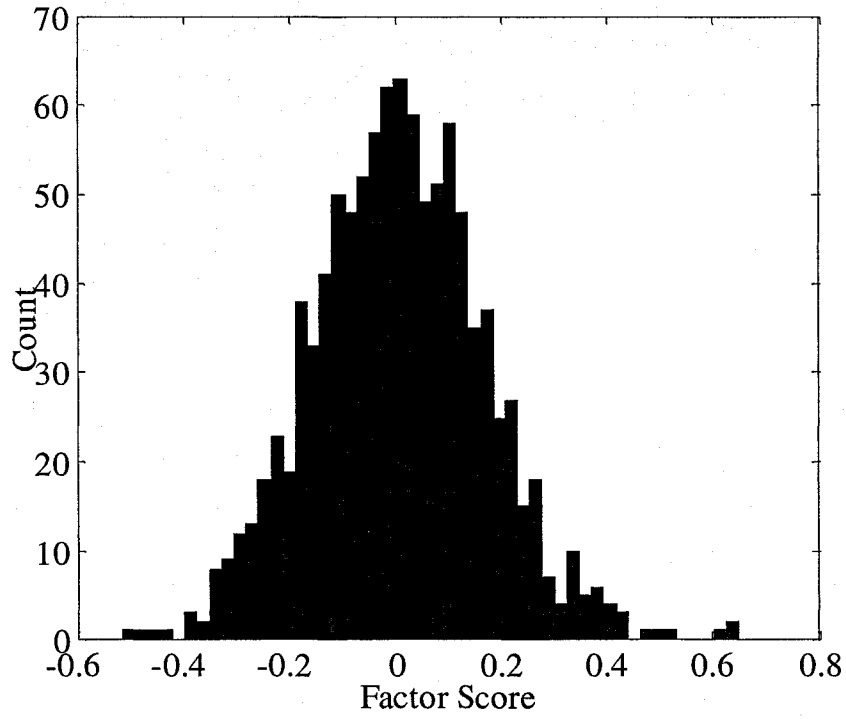


Figure 5.2: Nominal distribution of P implant score wafer mean.

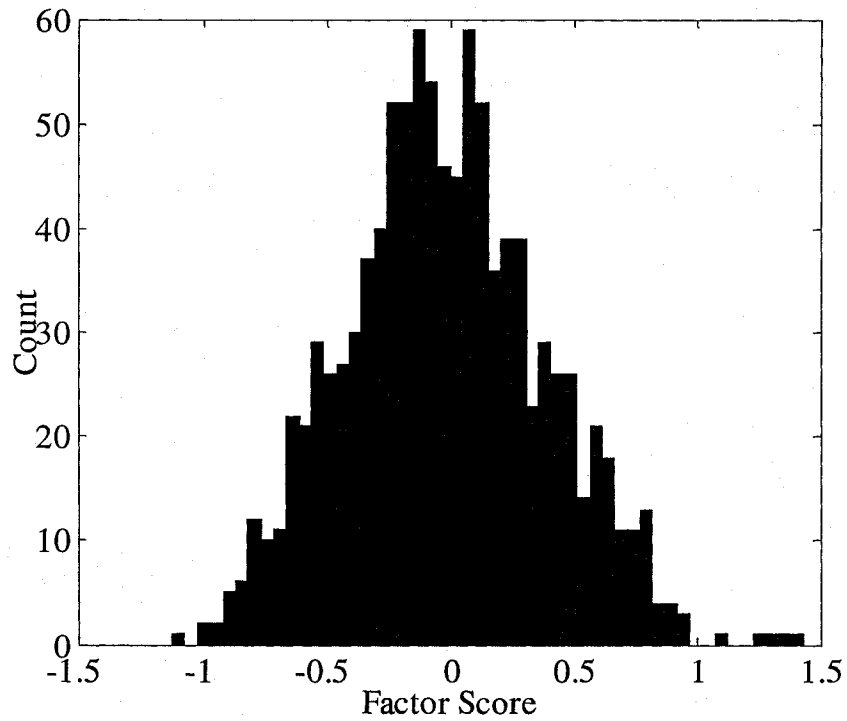


Figure 5.3: Nominal distribution of Gate Oxide score wafer mean.

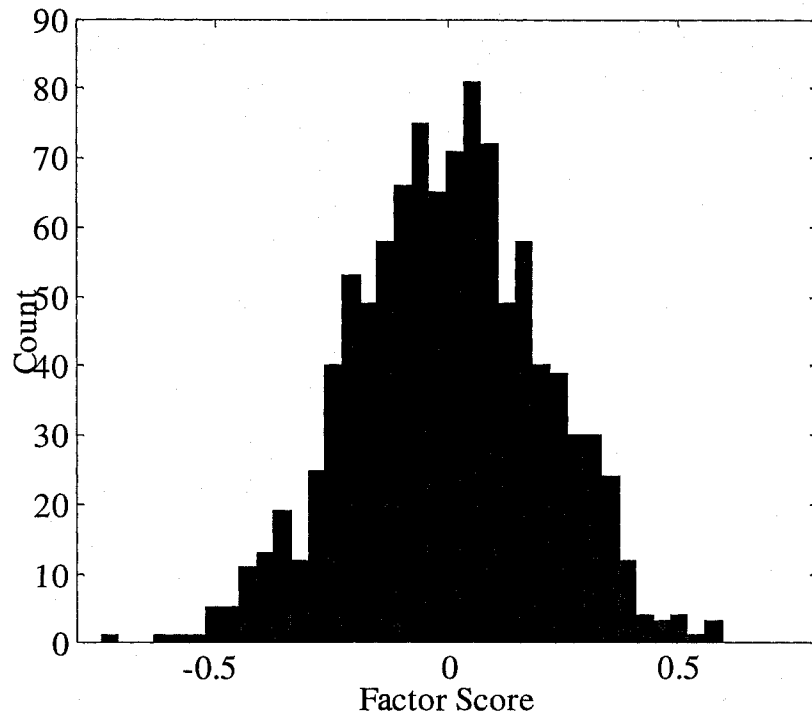


Figure 5.4: Nominal distribution of Lithography score wafer mean.

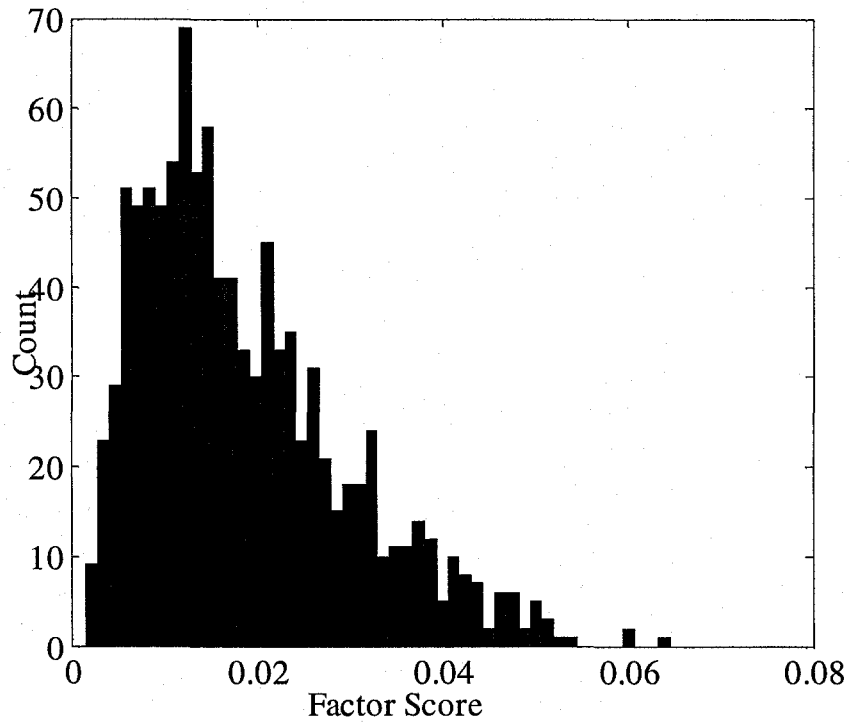


Figure 5.5: Nominal distribution of N implant score wafer variance.

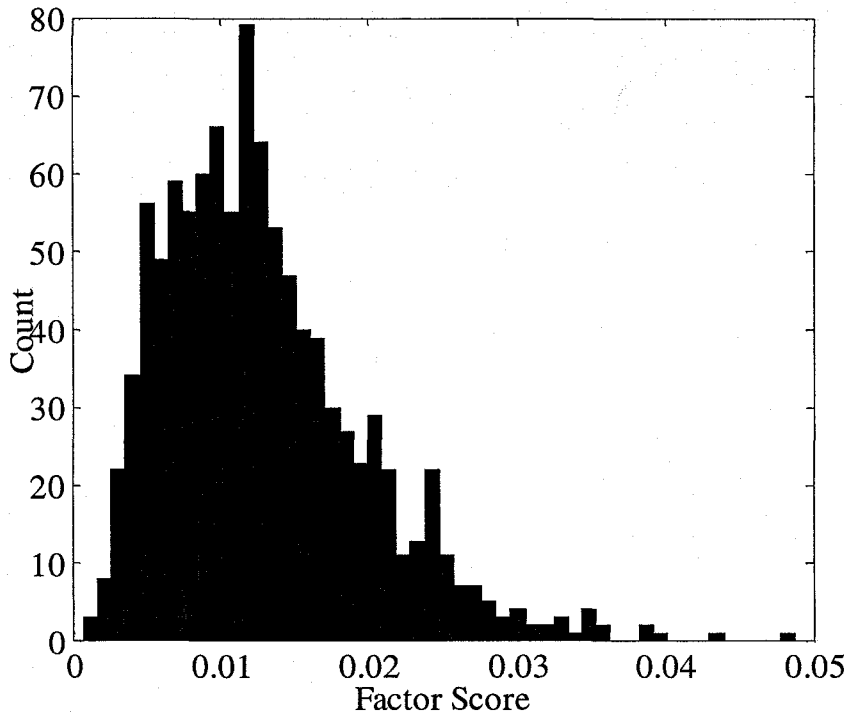


Figure 5.6: Nominal distribution of P implant score wafer variance.

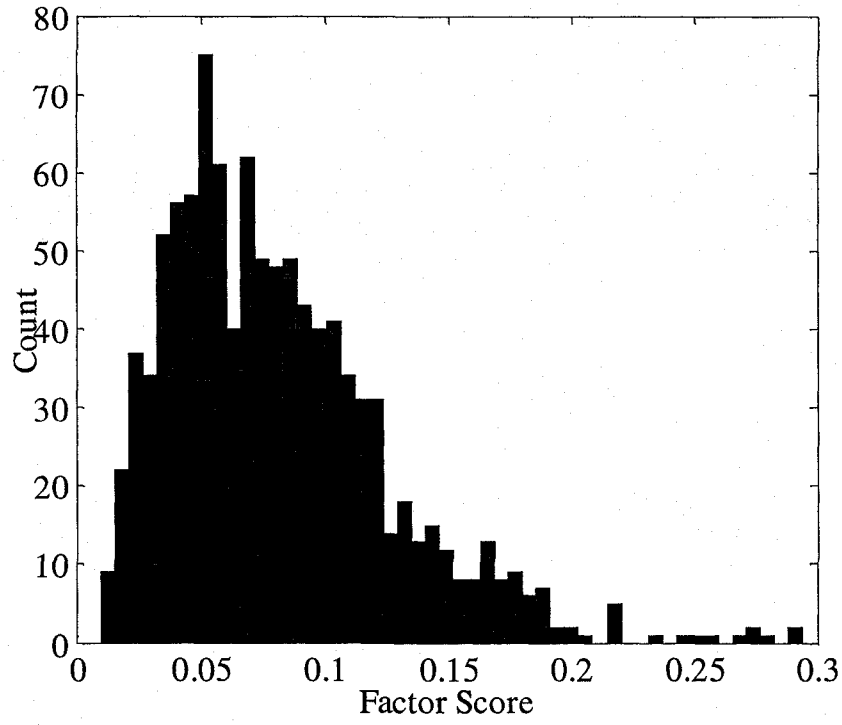


Figure 5.7: Nominal distribution of Gate Oxide score wafer variance.

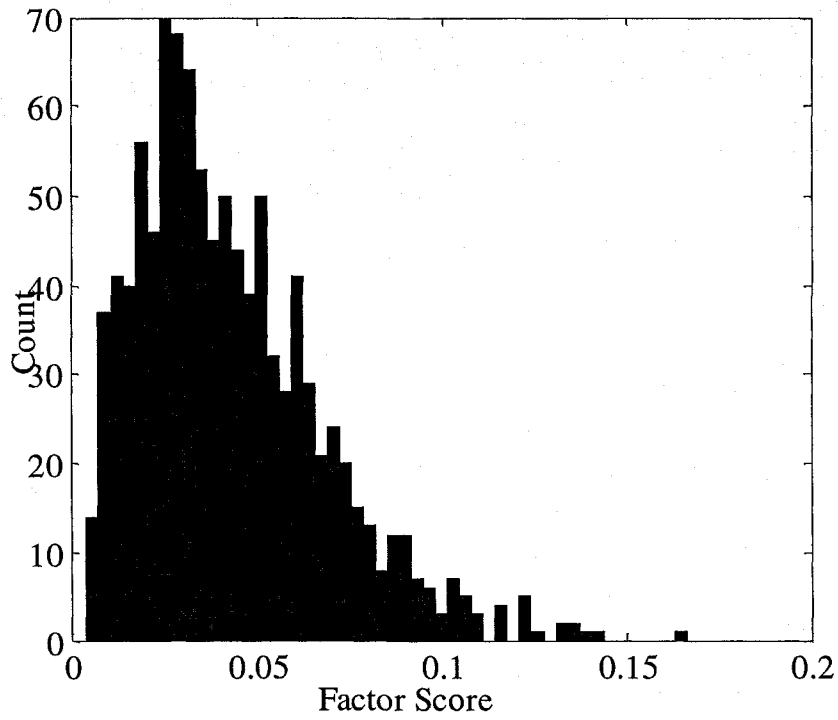


Figure 5.8: Nominal distribution of Lithography score wafer variance.

5.5 Construction of Control Charts

Based on the above empirical distributions, control limits for Shewart charts can be selected and control charts implemented. The basic logical group for our process is the wafer, so the natural control charts to use are for wafer mean and variance. We choose a false alarm probability of 0.01 and implement a two-sided control chart for wafer mean and a one-sided control chart for wafer variance. An example chart for N implant wafer means appears in Figure 5.9. While this is the normal form of a Shewart chart, our data is not time ordered and so the sequence in which the wafer scores are plotted is arbitrary. Figures 5.10 – 5.17 are scatterplots of wafer mean or variance versus lot. Each horizontal line on the plot represents a lot, and the control limits for the Shewart chart appear as vertical lines. This display makes it easy to see which lots contain which types of disturbances. These plots are essentially compressed Shewart charts organized by lot. Two additional experimental lots not used for model estimation are plotted in Figures 5.18 – 5.25.

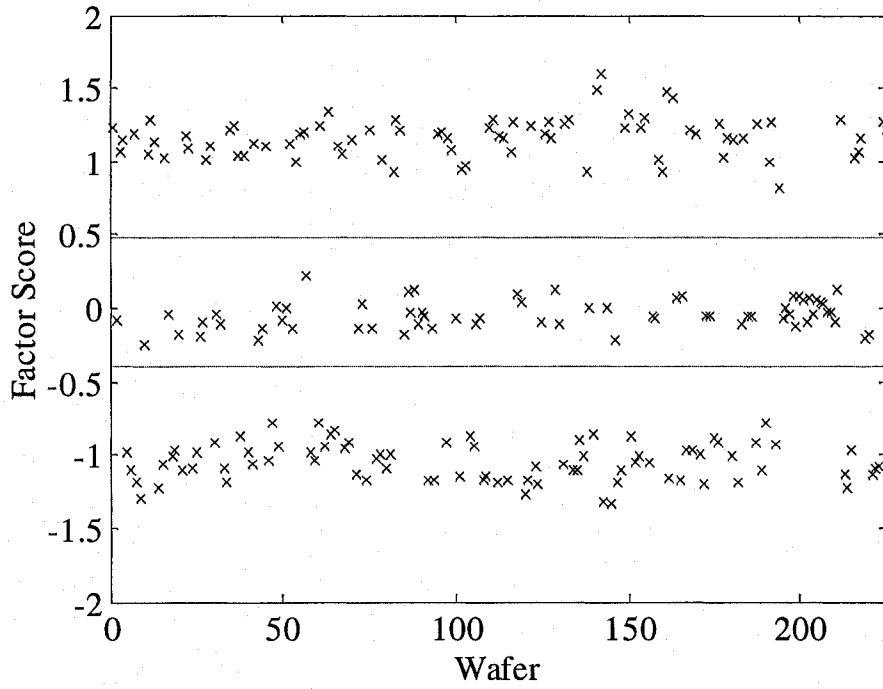


Figure 5.9: Shewart chart for N implant score wafer mean.

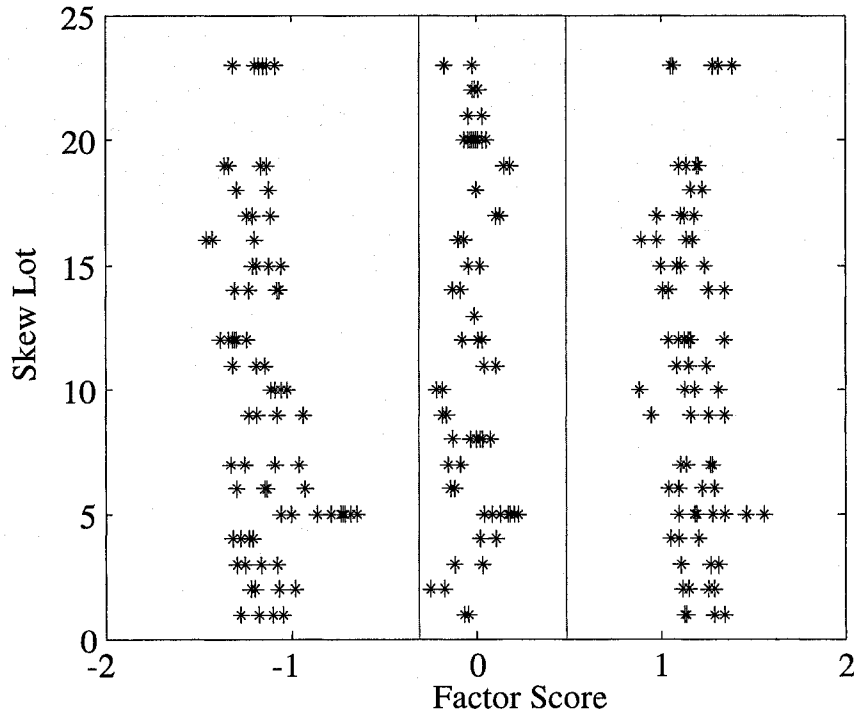


Figure 5.10: Compressed Shewart chart for N implant score wafer mean. Each row represents an experimental lot of wafers. The vertical lines are the test limits.

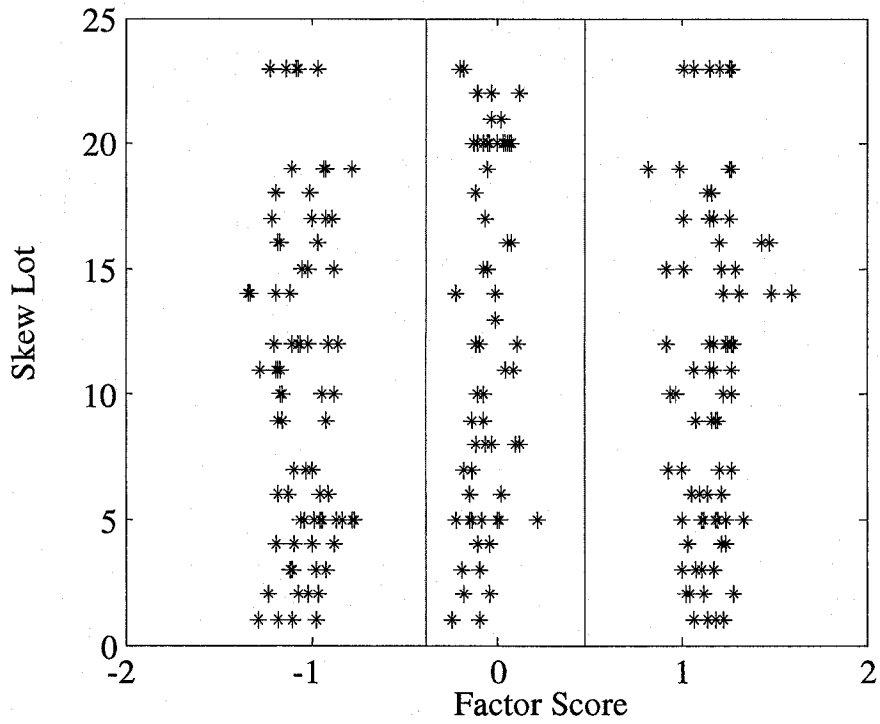


Figure 5.11: Compressed Shewart chart for P implant score wafer mean.

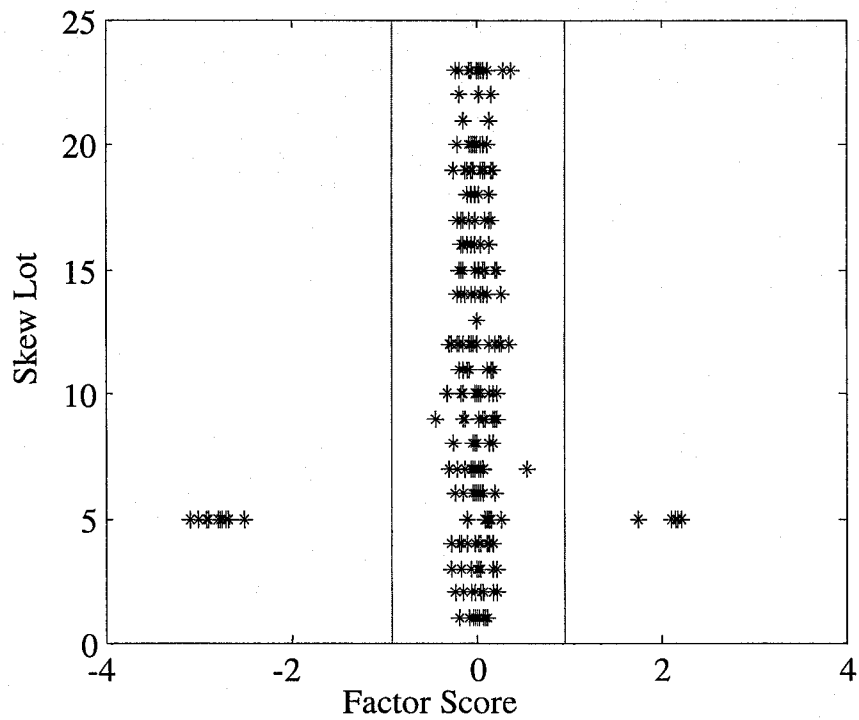


Figure 5.12: Compressed Shewart chart for Gate Oxide score wafer mean.

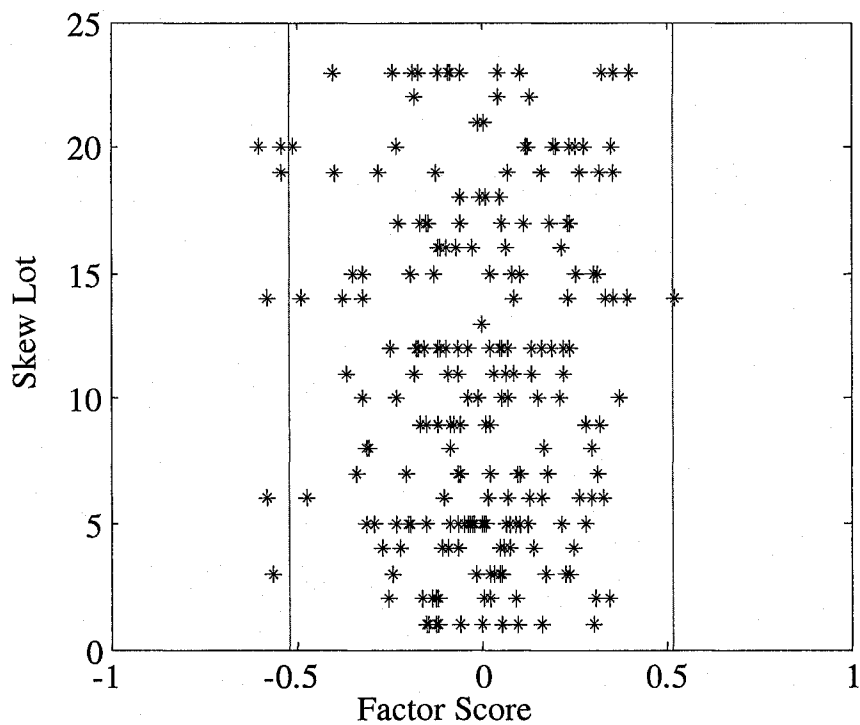


Figure 5.13: Compressed Shewart chart for Lithography score wafer mean.

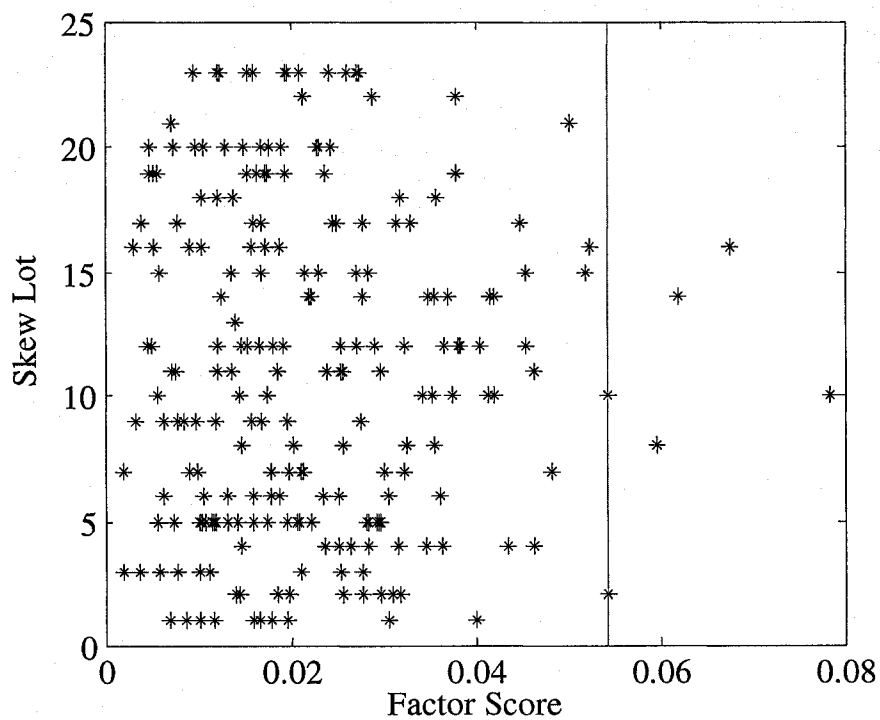


Figure 5.14: Compressed Shewart chart for N implant score wafer variance.

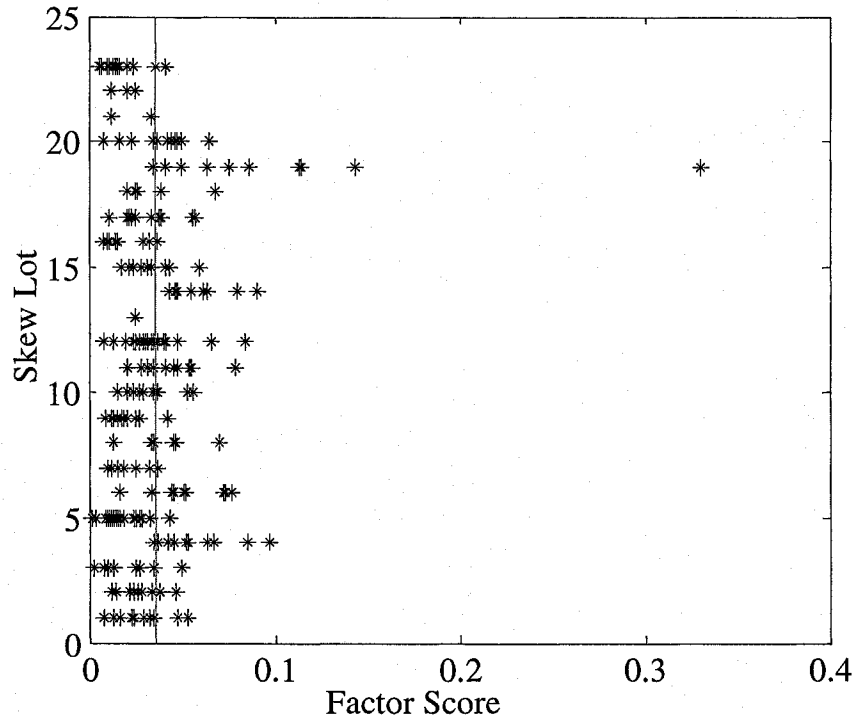


Figure 5.15: Compressed Shewart chart for P implant score wafer variance.

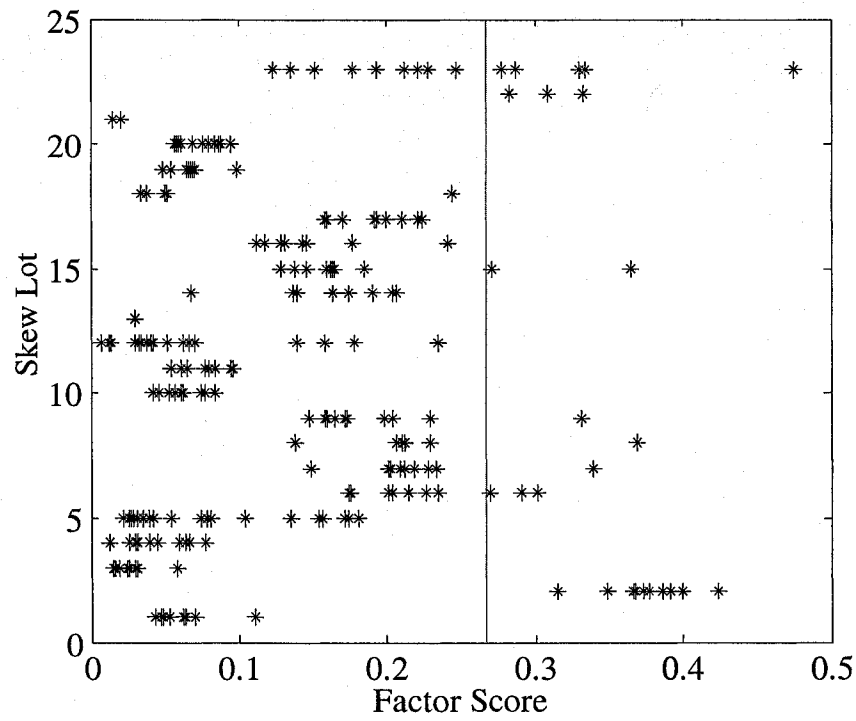


Figure 5.16: Compressed Shewart chart for Gate Oxide score wafer variance.

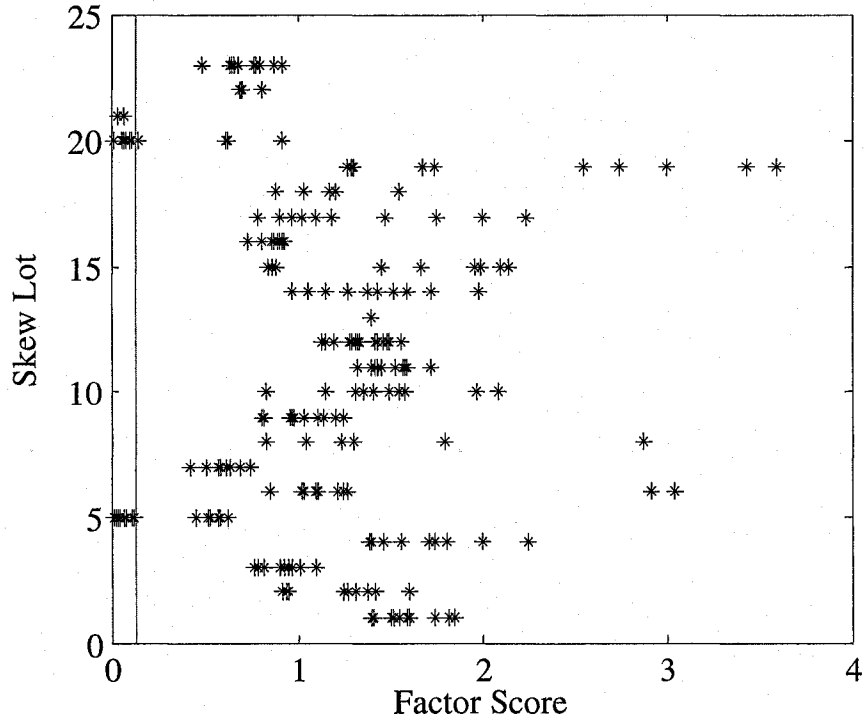


Figure 5.17: Compressed Shewart chart for Lithography score wafer variance.

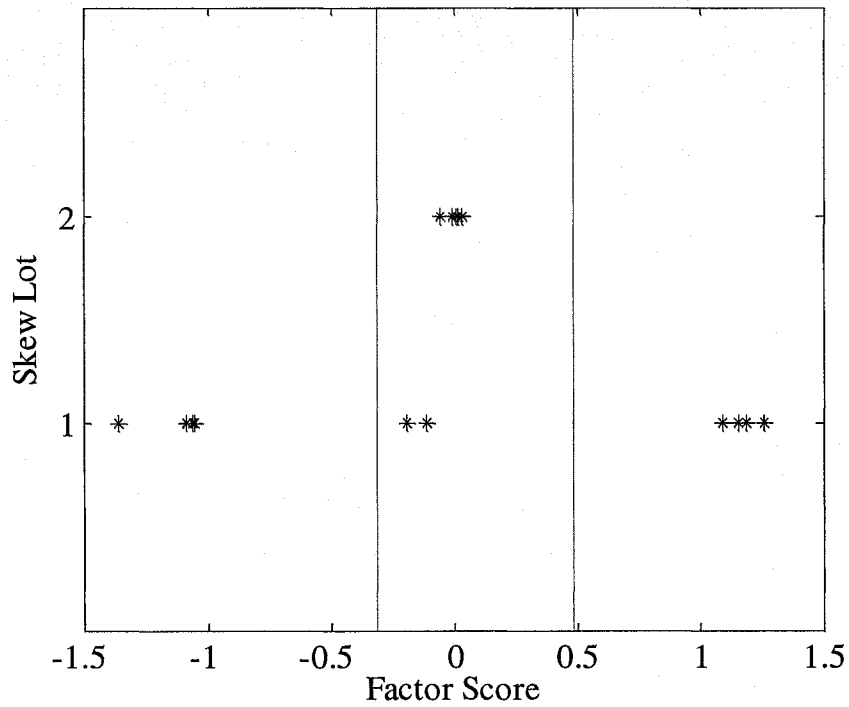


Figure 5.18: Chart for additional lots, N implant score wafer mean.

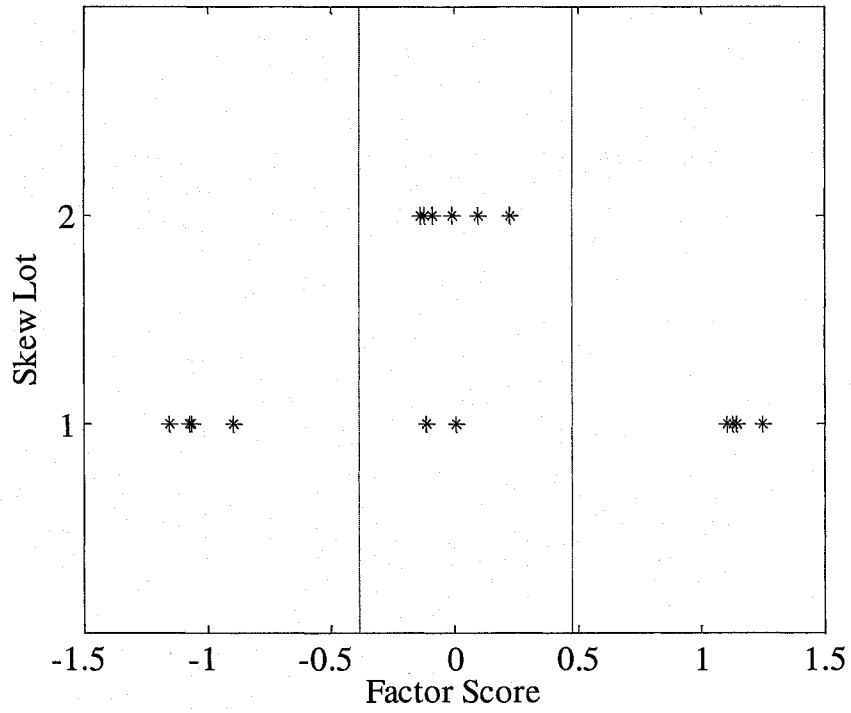


Figure 5.19: Chart for additional lots, P implant score wafer mean.

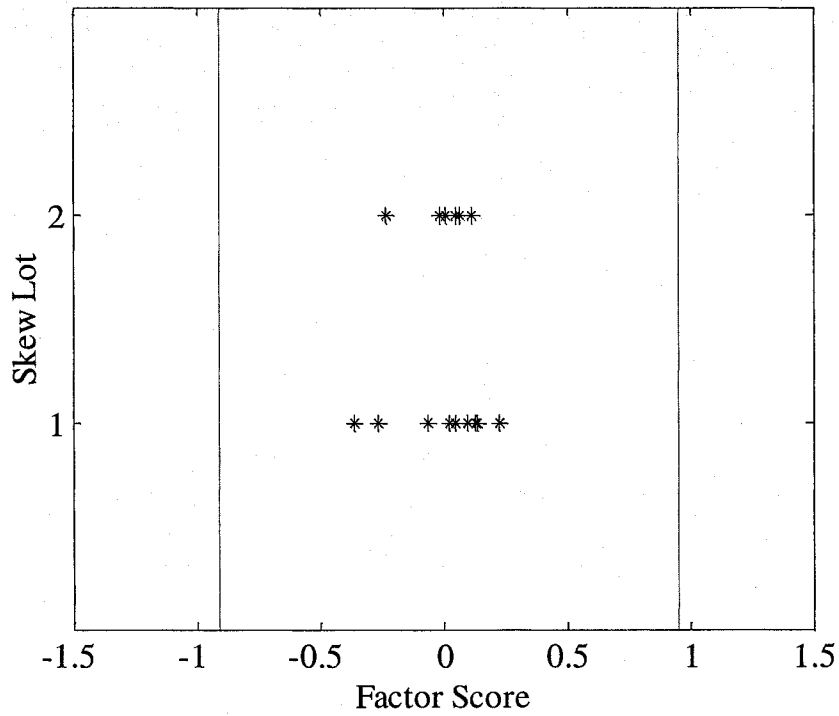


Figure 5.20: Chart for additional lots, Gate Oxide score wafer mean.

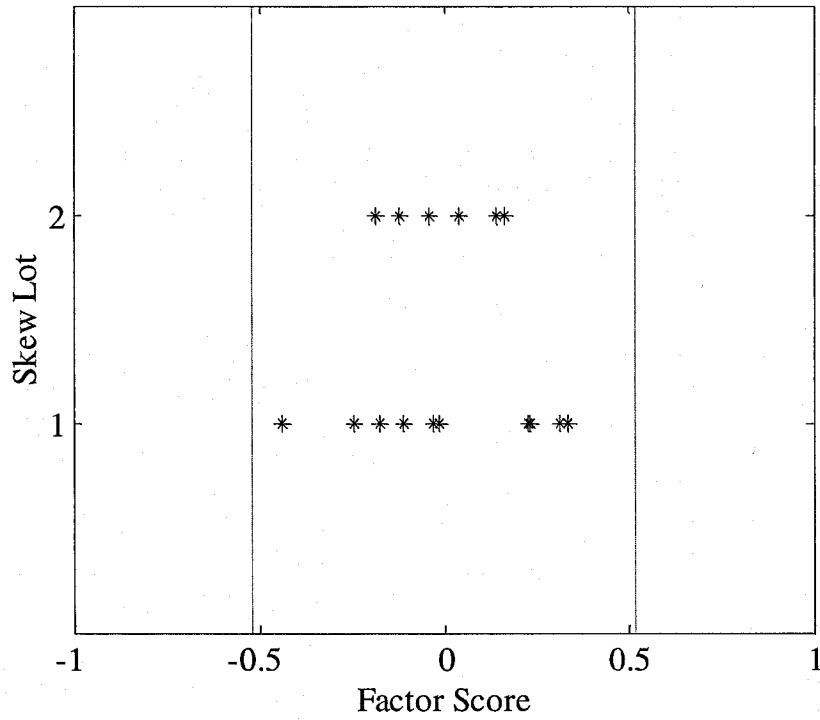


Figure 5.21: Chart for additional lots, Lithography score wafer mean.

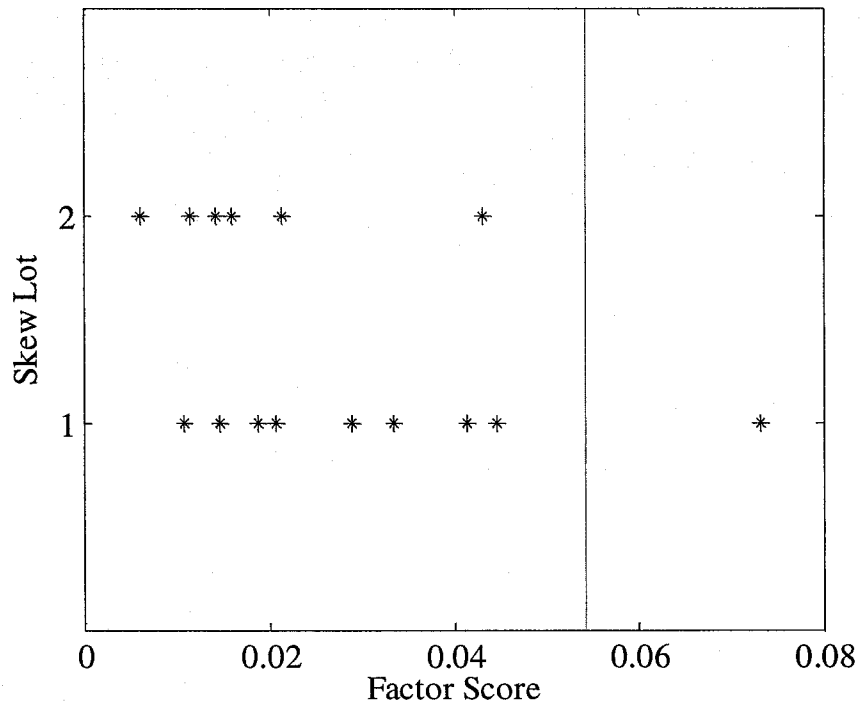


Figure 5.22: Chart for additional lots, N implant score wafer variance.

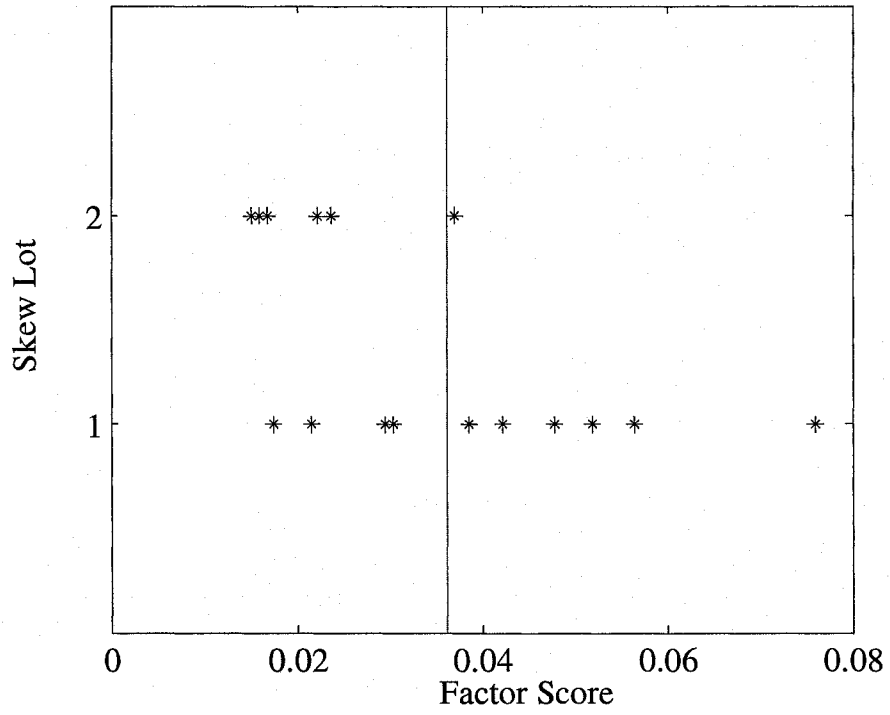


Figure 5.23: Chart for additional lots, P implant score wafer variance.

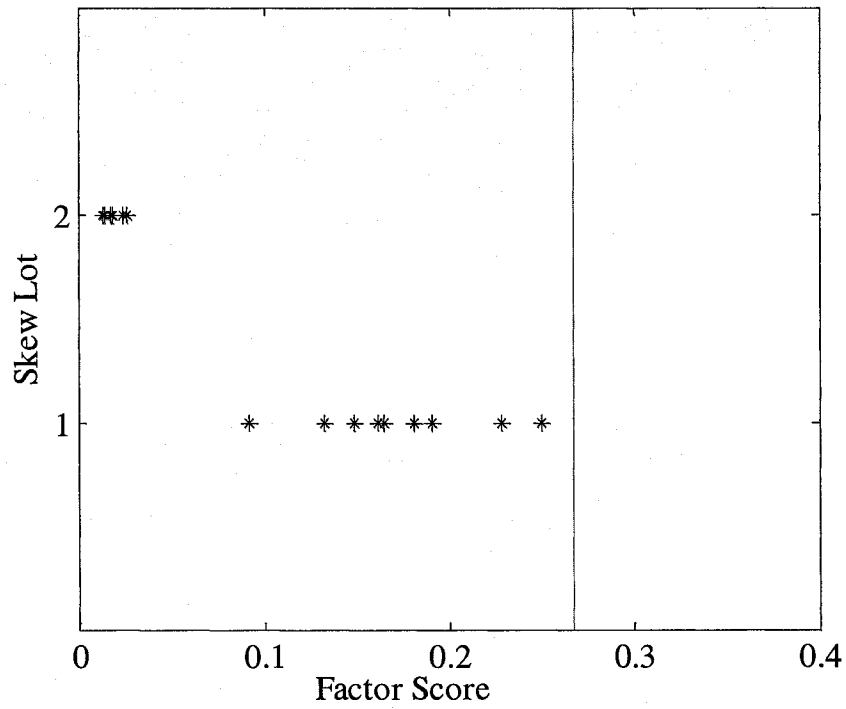


Figure 5.24: Chart for additional lots, Gate Oxide score wafer variance.

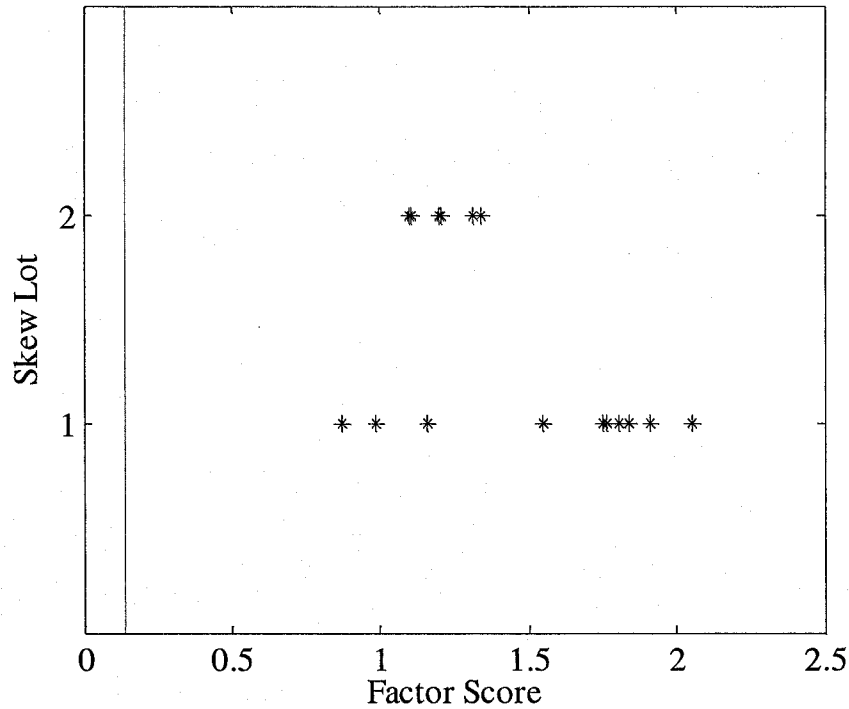


Figure 5.25: Chart for additional lots, Lithography score wafer variance.

5.6 Validation Results

We validate our characterization of the estimated independent components by evaluating the performance of the control charts based on our knowledge of the design disturbances. If the control charts correctly classify which wafers were processed with which disturbances we can be confident our characterization and methodology is working as intended. Given that the designed disturbances were wafer-wide step functions for n implant, p implant, and oxide thickness, the wafer mean \bar{X} charts for those three component scores should classify those disturbances, while the σ^2 chart for lithography should classify the across wafer lithography disturbances.

5. ICA Disturbance Modeling

The detection results for wafers included in the ICA estimation are given in Table 5.3 and those for the wafers not included in ICA are given in Table 5.4. For the wafer mean charts we get zero false positives and zero false negatives for 100% correct classification of all 241 wafers. The σ^2 chart for lithography has four false positives and zero false negatives for a classification rate of 98.4%. This is very good considering that the classification rate was not used as a criterion during data preprocessing and model estimation. *Based on these results we can confidently claim that our independent component scores are latent variable indicators for the designed disturbances in the experimental lots.*

Table 5.3: Detection results for wafers included in ICA.

Disturbance	True	False	True	False
N Implant	160	0	65	0
P Implant	160	0	65	0
Gate Oxide	18	0	207	0
Lithography	192	4	29	0

Table 5.4: Detection results for wafers not included in ICA.

Disturbance	True	False	True	False
N Implant	8	0	8	0
P Implant	8	0	8	0
Gate Oxide	0	0	16	0
Lithography	16	0	0	0

5.7 Discussion

Examinations of the compressed Shewart charts for *p*-implant and gate-oxide wafer variances, reveal a significant number of alarm signals. These charts do not correspond to any of the designed disturbances nor do the alarms correlate with the disturbances. While no process cause can be attributed to these alarms, their most likely methodological causes are the variability in the sampling pattern and number of samples taken from each wafer and the fact that we did no product specific normalization of variance when we pooled our data. The variability of how the wafers were sampled can be remedied in practice by standardizing testing across products. The issue of how to normalize variance by product with sparse data is an appropriate subject for future research.

As a check of the robustness of our technique with respect to the assumption that the designed disturbances are the major sources of variance in the experimental data used for ICA, we repeated the analysis keeping the first five principal components. The results show that our four factor-linked components are stable. The explanation of variance for the five component model is given in Table 5.5. The additional factor appears to be linked to narrow channel phenomena unrelated to the designed disturbances. Classification of disturbances is identical to that with the four component model.

As we noted in the discussion of Hotelling T^2 tests in chapter three, the optimal test for a shift in the mean of a multivariate distribution in a known direction is a

5. ICA Disturbance Modeling

univariate test. One interpretation of our independent components is that they provide a set of orthogonal alternatives for these tests. An alternative approach to developing these directions would be to do multi-dimensional clustering of known faults from a historical data base. The cluster centers could be used to generate such test. A difficulty of this approach would be keeping the cluster centers orthogonal from each other so that the test statistics remain uncorrelated. Our method guarantees this orthogonality.

Table: 5.5 Explanation of variance of five independent components.

	Gate	N	P	Diffusion	Narrow	Total
TOX N	99.3%	0.3%	0.1%	0.0%	0.1%	99.7%
VT NLW	4.4%	93.5%	0.5%	0.2%	0.1%	98.6%
ID NLW	8.7%	89.1%	0.3%	0.0%	0.2%	98.3%
VT NMW	3.0%	60.7%	0.0%	32.4%	1.4%	97.5%
ID NMW	5.9%	31.4%	0.0%	55.2%	1.2%	93.7%
VT NSW	1.9%	43.1%	0.0%	50.4%	1.6%	96.9%
ID NSW	4.3%	24.1%	0.0%	65.6%	1.3%	95.2%
VT NMN	5.1%	55.6%	0.2%	12.3%	7.2%	80.3%
ID NMN	14.3%	22.4%	0.1%	11.9%	42.6%	91.4%
VT NLN	7.1%	89.1%	0.4%	0.1%	0.1%	96.8%
ID NLN	10.3%	54.1%	0.4%	1.9%	16.2%	82.8%
TOX P	99.3%	0.3%	0.1%	0.0%	0.0%	99.7%
VT PLW	1.1%	0.4%	94.3%	2.2%	0.2%	98.3%
ID PLW	2.5%	0.4%	92.8%	1.9%	0.3%	97.9%
VT PMW	0.5%	0.0%	73.0%	23.5%	1.0%	98.0%
ID PMW	1.5%	0.0%	44.4%	50.7%	1.3%	97.9%
VT PSW	0.2%	0.0%	53.1%	42.9%	1.2%	97.4%
ID PSW	1.0%	0.0%	34.1%	61.5%	1.4%	98.0%
VT PMN	1.7%	0.1%	71.1%	8.0%	6.4%	87.3%
ID PMN	8.7%	0.3%	47.0%*	14.7%	21.1%	91.8%
VT PLN	3.5%	0.3%	89.3%	2.2%	0.4%	95.5%
Total	13.5%	26.9%	28.6%	20.8%	5.0%	94.9%

6. Possibilistic Independent Component Analysis

6.1 Possibility Theory

Possibility theory is that sub-branch of evidence theory that concerns itself with consonant (nested) bodies of evidence (De Cooman 1997, Klir and Wierman 1998). This mode of dealing with uncertainty is commonly encountered both in the form of crisp logical constraints and in the form of fuzzy membership functions for linguistic values. Formally, a possibility distribution is a mapping from a domain of interest, Ω , to the real interval $[0,1]$

$$\mu : \Omega \rightarrow [0,1],$$

with

$$\sup_{\omega \in \Omega} \mu(\omega) = 1.$$

For our purposes, we are interested in possibility distributions defined on real intervals, i.e. subsets of \mathbb{R}^n . The possibility measure Π associated with a possibility distribution μ is defined by

$$\Pi(\emptyset) = 0,$$

and for $A \subseteq \Omega$

$$\Pi(A) = \max_{x \in A} \mu(x).$$

Note that

$$\Pi(\Omega) = 1,$$

and for any ample field of subsets $C = \{A_i | A_i \subseteq \Omega\}$

$$\Pi\left(\bigcup_i A_i\right) = \sup_i \Pi(A_i),$$

and

$$\Pi\left(\bigcap_i A_i\right) \leq \inf_i \Pi(A_i).$$

The possibilistic equivalent of the factorization of probability distributions is the *noninteraction* constraint

$$\mu(x, y) = \min\{\mu(x), \mu(y)\}.$$

This implies that the multivariate possibility distribution is the cylindric extension of the marginal distributions. Independence of possibilistic variables implies noninteraction of the marginal distributions (Klir 2006) and is a stronger condition than noninteraction.

6.2 Possibilistic Measures of Uncertainty

To perform BSS with possibilistic data, we need to construct an estimator of the transmission between continuous possibilistic variables (an appropriate information measure analogous to KL divergence in the probabilistic case). Following a proposal of Klir and Yuan (Klir and Yuan 1995), such a measure would be

$$T_{UL}(x_1, x_2) = UL(x_1) + UL(x_2) - UL(x_1, x_2),$$

where

$$UL(\mathbf{x}) = \int_0^{h(\mathbf{x})} HL(\alpha \mathbf{x}) d\alpha + (1-h(\mathbf{x}))HL(\Omega),$$

and

$$HL(\alpha \mathbf{x}) = \min_{t \in T} \ln \left[\prod_{i=1}^n [1 + \mu_L(\alpha x_{it})] + \mu_L(\alpha \mathbf{x}) - \prod_{i=1}^n [\mu_L(\alpha x_{it})] \right],$$

μ_L denotes Lebesgue measure, $h(\mathbf{x})$ is the height of the possibility density, T denotes the set of orthonormal transformations, αx_{it} is the i th projection of $\alpha \mathbf{x}$ with respect to the coordinate system t , and Ω and the α -cuts $\alpha \mathbf{x}$ are convex subsets of \mathbb{R}^n . While originally proposed for convex sets, this measure has been extended to nonconvex sets in (Ramer and Padet 2001).

Two variables are said to be noninteractive if

$$\mu(x, y) = \min\{\mu(x), \mu(y)\}.$$

The variables are independent iff

$$\mu(x|y) = \mu(x) \text{ and } \mu(y|x) = \mu(y),$$

in which case

$$T_{UL}(x_1, x_2) = 0.$$

6.3 Possibilistic Independence: Example

Suppose that two variables x_1 and x_2 are jointly observed, and their values always fall within the square region with vertices at $(0, \sqrt{2})$, $(\sqrt{2}, 0)$, $(0, -\sqrt{2})$, $(-\sqrt{2}, 0)$ illustrated in Figure 6.1. If nothing else is known about the frequency with which values appear, we can characterize the uncertainty of the values of the variables by

6. Possibilistic ICA

assigning a possibility of 1 to values within the square region and a possibility of 0 to those outside the region. This is a crisp possibilistic constraint on x_1 and x_2 .

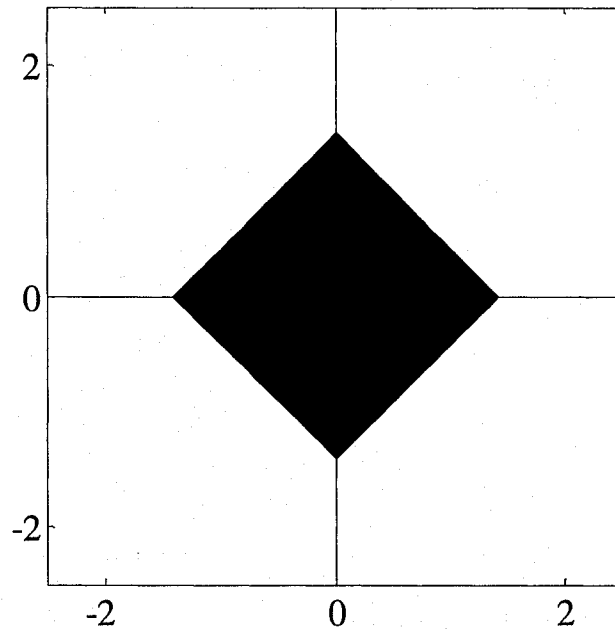


Figure 6.1: Observed possibility distribution.

Using the UL measure of uncertainty, the uncertainty of joint observations is 3.17 bits, the marginal uncertainty of either x_1 or x_2 is 1.94 bits and the transmission between x_1 and x_2 is 0.70 bits. Intuitively, a large absolute value for an observation of x_1 implies a reduced range of possible values for x_2 and vice versa. If a 45° rotation is applied to the variables, Figure 6.2 is obtained. The joint uncertainty is still 3.17 bits, but the marginal uncertainties are now 1.585 bits and the transmission is 0. The rotated variables are independent.

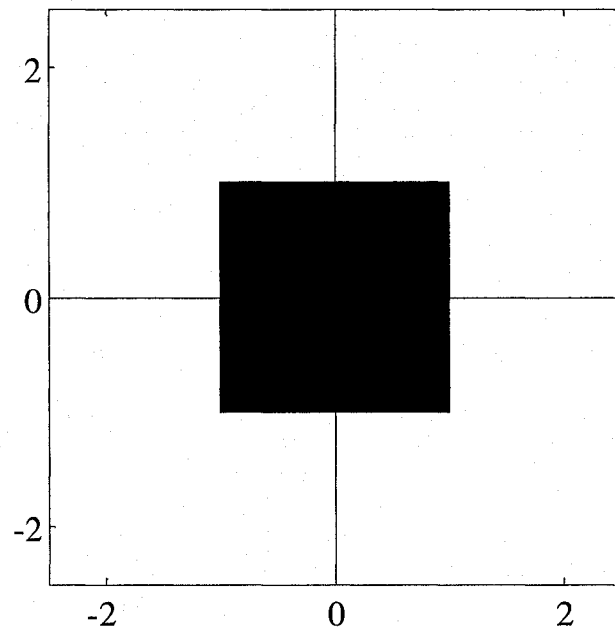


Figure 6.2: Demixed possibility distribution.

6.4 Proposed Method

The general recipe for PICA assuming a linear mixing model is:

1. Center the observed possibility distribution.
2. Find the principal axes of the distribution.
3. Scale the principal axes.
4. Find the rotation that minimizes the transmission between variables.

As with most data analysis techniques, the first step for possibilistic BSS is to center the data with an affine translation, i.e. shift the coordinate axis so that the origin is at

ξ , where

$$\xi_i = \frac{\int x_i d\mu_i}{\int d\mu_i},$$

so that

$$\int (x_i - \xi_i) d\mu_i = 0.$$

Then we find the principal axis of the possibility distribution by diagonalizing

\mathbf{I} given by

$$I_{i,j} = \int x_i x_j d\mu,$$

so that

$$\mathbf{I} = \mathbf{E}\mathbf{D}\mathbf{E}^T.$$

We can then decompose the demixing transformation as

$$\mathbf{W} = \mathbf{B}\mathbf{D}^{-1/2}\mathbf{E}^T = \mathbf{B}\mathbf{V},$$

where $\mathbf{B} \in \text{SO}(n)$.

We then find a linear transformation that minimizes the HL transmission, i.e.

given n variables find

$$\begin{aligned} \mathbf{B} &= \underset{\mathbf{B} \in \text{SO}(n)}{\operatorname{argmin}} T_{UL}(\mathbf{W}x_1, \dots, \mathbf{W}x_n), \\ &= \underset{\mathbf{B} \in \text{SO}(n)}{\operatorname{argmin}} T_{UL}(\mathbf{B}\mathbf{V}x_1, \dots, \mathbf{B}\mathbf{V}x_n), \\ &= \underset{\mathbf{B} \in \text{SO}(n)}{\operatorname{argmin}} \left[\sum_{j=1}^n UL(\mathbf{B}\mathbf{V}x_j) - UL(\mathbf{B}\mathbf{V}\mathbf{x}) \right], \\ &= \underset{\mathbf{B} \in \text{SO}(n)}{\operatorname{argmin}} \left[\sum_{j=1}^n UL(\mathbf{B}\mathbf{V}x_j) - UL(\mathbf{V}\mathbf{x}) \right], \\ &= \underset{\mathbf{B} \in \text{SO}(n)}{\operatorname{argmin}} \left[\sum_{j=1}^n UL(\mathbf{B}\mathbf{V}x_j) \right], \end{aligned}$$

where each uncertainty term is calculated separately. If we assume the possibility distribution is normal, i.e. $h(x) = 1$, we have

$$\begin{aligned}
\mathbf{B} &= \operatorname{argmin}_{\mathbf{B} \in SO(n)} \left[\sum_{j=1}^n UL(\mathbf{B}\mathbf{V}x_j) \right], \\
&= \operatorname{argmin}_{\mathbf{B} \in SO(n)} \left[\sum_{j=1}^n \int_0^1 HL({}^\alpha \mathbf{B}\mathbf{V}x_j) d\alpha \right], \\
&= \operatorname{argmin}_{\mathbf{B} \in SO(n)} \left[\sum_{j=1}^n \int_0^1 \ln(1 + \mu_L({}^\alpha \mathbf{B}\mathbf{V}x_j)) d\alpha \right], \\
&= \operatorname{argmin}_{\mathbf{B} \in SO(n)} \left[\int_0^1 \ln \left(\prod_{j=1}^n (1 + \mu_L({}^\alpha \mathbf{B}\mathbf{V}x_j)) \right) d\alpha \right].
\end{aligned}$$

This final expression is advantageous in that no evaluation of the multivariate HL function is required. The minimization of the transmission can be performed calculating with the univariate marginals of the α -cuts. Given a representation of a possibility distribution as finite set of k discrete α -cuts, this becomes

$$\mathbf{B} = \operatorname{argmin}_{\mathbf{B} \in SO(n)} \left[\sum_{i=1}^k \alpha_i \ln \left(\prod_{j=1}^n (1 + \mu_L({}^i \mathbf{B}\mathbf{V}x_j)) \right) \right].$$

Numerical approximations of the natural gradient with respect to the generators of $SO(n)$ can be derived for this minimization within a particular computational framework. Thus no nested optimization problems need be solved and reasonably efficient standard computational techniques can be used.

6.5 Crisp Example

Consider a case with two independent sources characterized by a joint crisp possibility distribution as in Figure 6.3. The value of either source variable can be

6. Possibilistic ICA

anywhere in the interval $[0, 1]$ with possibility 1. Let these source variables be mixed together via a linear mixing process characterized by the mixing matrix

$$\mathbf{M} = \begin{pmatrix} 2.5 & -1.5 \\ 0.5 & 0.9 \end{pmatrix}.$$

The possibility distribution of the observed mixture will then appear as in Figure 6.4; centering the distribution results in Figure 6.5. Adopting the principal axes as the working coordinate system results in Figure 6.6, and scaling by the reciprocal of the square root of the eigenvalues produces Figure 6.7.

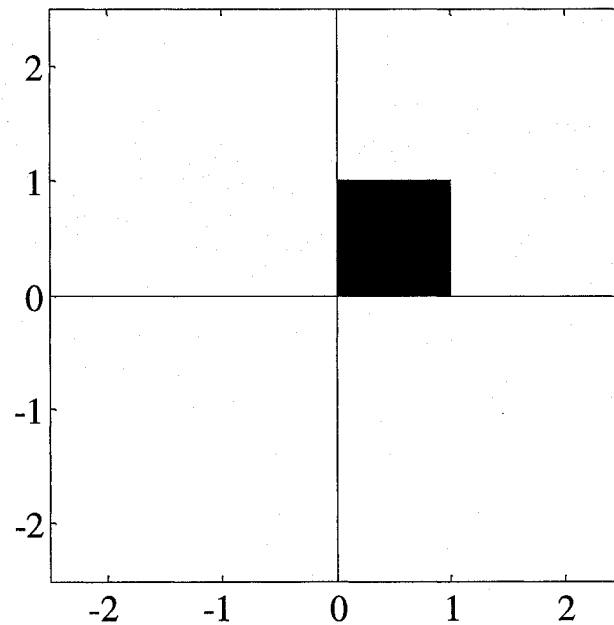


Figure 6.3: Joint possibility distribution of two noninteractive sources.

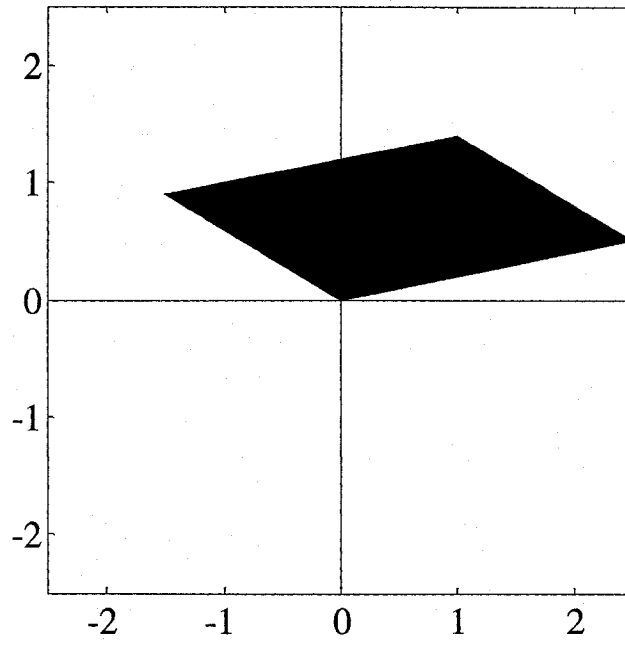


Figure 6.4: Mixed (observed) possibility distribution.

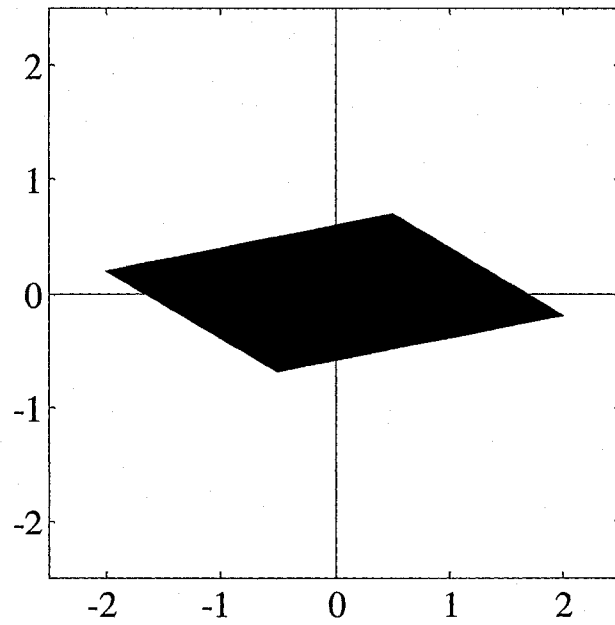


Figure 6.5: Centered distribution.

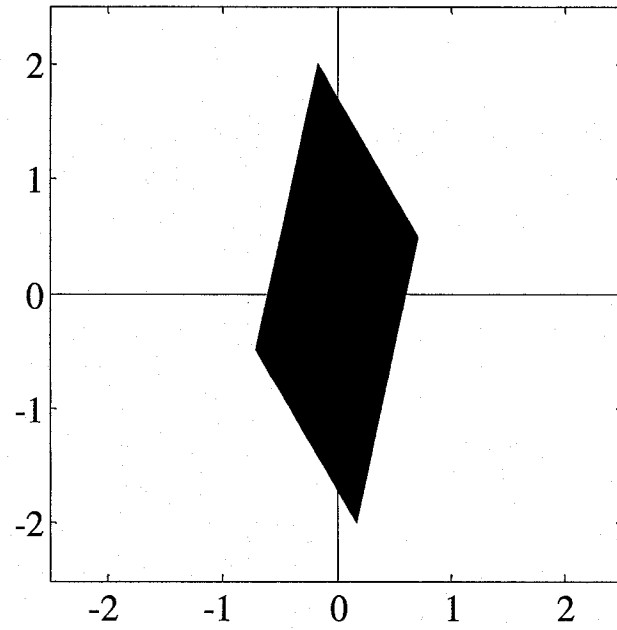


Figure 6.6: Centered distribution with principal axes aligned.

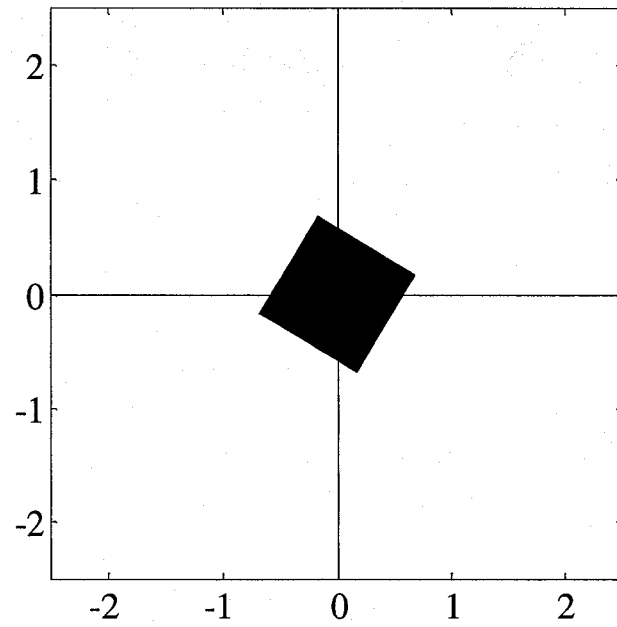


Figure 6.7: Distribution with principal axes scaled.

Finally, a counterclockwise rotation of ~ 0.54 radians produces noninteractive variables as shown in Figure 6.8. The demixing matrix found through BSS is

$$\hat{\mathbf{W}} = \begin{pmatrix} 0.1667 & -0.8333 \\ -0.3 & -0.5 \end{pmatrix},$$

the actual inverse of the mixing matrix is

$$\mathbf{W} = \begin{pmatrix} 0.3 & 0.5 \\ -0.1667 & 0.8333 \end{pmatrix}.$$

The reordering of the variables and the change of sign for the measurement scales is an artifact of the ordering and directions chosen for the eigenvectors when aligning the principal axes.

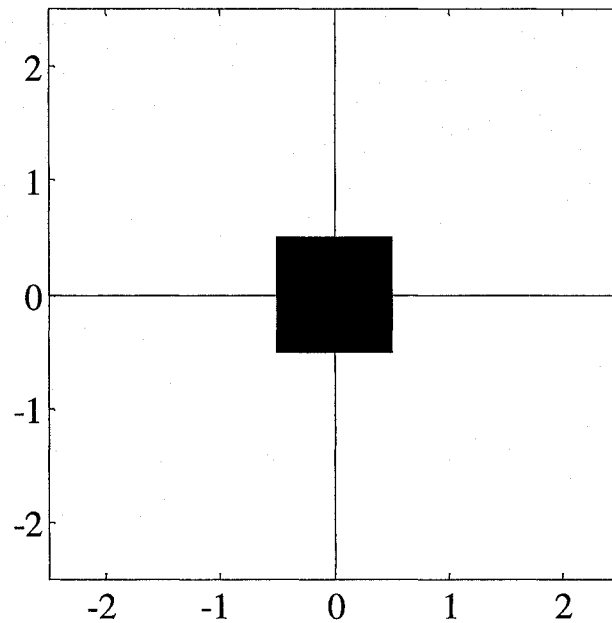


Figure 6.8: Distribution after final separating rotation.

6.6 Discussion

This chapter has outlined the potential for using PICA to perform BSS with variables characterized by possibility distributions. While the theoretical examples used to illustrate the general procedure were crisp, it should be clear that the same basic procedure could be applied to fuzzy membership functions or any possibility distribution of interest. Applications for this technique may include factor analysis as well as machine learning applications. As an example, one could merge data from a variety of sampling plans or measurement scales into a single body of evidence and then estimate a factor model based on the assumption of independent sources. Application of this technique could also lead to a reduction in complexity of a fuzzy inference system through the restructuring of variables to minimize multiple dependences.

A primary limitation of this proposal is the problem of developing practical representations of the level sets or α -cuts of the joint possibility distribution from empirical data, e.g. (Joslyn 1997). Coincident with this issue is that of numerical implementation of the uncertainty calculations. How to efficiently implement the necessary calculations depends on the representation chosen for the possibility distribution, which is related to how the distribution is to be empirically estimated. The above concerns have been omitted in this sketch and are probably best dealt with on an application specific basis. Obviously, an implementation based on a gradient

6. Possibilistic ICA

optimization method would be preferable to using a line search or simplex technique, and should be feasible in some contexts.

7. Conclusion

7.1 Summary

The first contribution of this work is the use of ICA to provide an explanation for the variance observed in a multivariable data set. The explanation of e-test variance attributed to underlying physical mechanisms quantified by the independent component scores is a novel application of ICA. The generally held view that ICA cannot be used to investigate variance is a misperception. While ICA cannot determine the difference between the magnitude of source variance and gain in the mixing process, it can be used to explore how observed variance is generated from its sources. This result was first published in (Shannon, *et al.* 2003a)

The primary contribution offered here is a new methodology for multivariate statistical process control that has the advantages of using univariate control charts, providing direct diagnosis of detected disturbances, and making no assumptions about the observed measurements other than that they are nongaussian. Our method is demonstrated on E-test data from a semiconductor manufacturing process. No measurements or experiments other than those normally performed in the operation of the process are required. Validation of our method in the form of classification of known disturbances produced excellent results. The preliminary results on process monitoring contained in chapter four appeared in (Shannon *et al.* 2003a) and formed the basis for a patent application (Abercrombie *et al.* 2003). The development of these

7. Conclusion

results into control charts began with (Shannon *et al.* 2003b, Shannon *et al.* 2004); the final results appear in (Shannon and McNames 2007).

The final contribution of this research is a theoretical framework for performing an ICA-like separation of sources characterized by possibilistic data. This framework extends an analytic tool from the domain of probability distributions to the generalized uncertainty domain that includes classical possibility theory, fuzzy sets and systems, and Dempster-Schafer Evidence Theory. This generalization may have myriad uses in generalized information theory, but also may be useful in the advanced process control domain for estimating control charts for rare or costly disturbances that cannot be properly characterized by probability distributions. This framework was first presented in (Shannon 2005).

7.2 Future Research

A number of methodological questions could be investigated. How much data and how many examples of each disturbance type are required for ICA model estimation for a given size problem? While it is customary to need (and rare to have) a large data set for SPC design, the issues of how much is enough and what to do when there is not quite enough need to be addressed.

What is the best method for pooling data from multiple products or recipes that effectively removes product-to-product variation but leaves intact disturbance examples that may only be present in some products? One of the potential advantages of pooling products is that examples of all disturbances need not be generated for each

7. Conclusion

product. Techniques for product-to-product normalization should ensure disturbance specific information is not lost in the data pooling process while also removing all the nominal product-to-product variation.

How replicable must the known disturbances be for reliable estimation? Can historical data from naturally occurring disturbances that have been fully diagnosed be used for ICA modeling, or does one need purposely induced disturbances? Disturbances contained in a historical database would be desirable in that they represent faults actually encountered in process operation, but they may not represent a disturbance type in a crisp manner. It might be the case that many more examples would be needed for model estimation.

Additional case studies implementing independent component charts are needed. In the E-test context, the next research step would be to expand the framework to include more parameter measurements and disturbance types. It would also be appropriate to implement the charts in the actual production environment and compare their performance to the monitoring systems now in use. One would expect a variety of insights to emerge from implementation in a production environment.

A logical field to apply our method to would be batch process monitoring in the chemical industry. Processes in this field involve multiple segments in each of which the time course of the process is monitored. It is one of the most fertile fields for the development of APC methods, as it includes continuous dynamics with discrete

7. Conclusion

transitions between process stages. The amount of data available for analysis tends to be immense and the payoff for process improvement is significant.

Possibilistic ICA development is an entirely separate research area with many potential milestones. The first step would be to implement numerical algorithms for specific problems. The field of generalized uncertainty theory is undeveloped compared to the theories of probability and information theory. Many basic issues of how to represent data numerically and how to calculate with possibility distributions are unresolved. Fuzzy sets are the most commonly encountered examples of graded possibility distributions, and it is in the fuzzy systems context that possibilistic ICA is likely to be implemented.

References

- Abercrombie, D., T.T. Shannon and J. McNames, (2003) "Method of isolating sources in parametric data", U.S. Patent Application #20050060336.
- Alt, F.B. (1985) "Multivariate Quality Control," in *Encyclopedia of Statistical Sciences*, v. 6, N.L. Johnson and S. Kotz, eds., Wiley, NY, 1985.
- Amari, S.-I., A. Cickocki and H.H Yang (1996) "A new learning algorithm for blind source separation," in *Advances in Neural Information Processing 8*, pp. 757-763, MIT Press.
- Amari, S.-I. and J.-F. Cardoso (1997) "Blind Source Separation – Semiparametric Statistical Approach," *IEEE Trans. on Signal Processing*, vol.45, #11, pp. 2692-2700.
- Amari, S.-I., A. Cickocki and H.H Yang (2000) "Blind signal separation and extraction: Neural and information-theoretic approaches," in Haykin, S. Ed., *Unsupervised Adaptive Filtering*, vol. 1, pp. 63-138, Wiley.
- Amari, S. & H. Nagaoka (2000) *Methods of Information Geometry*, AMS Translations of Mathematical Monographs Vol. 191, Oxford Univ. Press.
- Bell, A.J. and T.J. Sejnowski (1995) "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129-1159.

References

- Belouchrani, A., K. Abed Meriam, J.F. Cardoso, and E. Moulines (1997) "A blind source separation technique based on second order statistics," *IEEE Transactions on Signal Processing*, vol. 45, # 2, pp. 434-444.
- Brockwell, P.J. and R.A. Davis (1996) *Introduction to Time Series and Forecasting*, Springer, New York.
- Cardoso, J.-F. (1989) "Blind identification of independent signals," in *Proc. of the Workshop on Higher-Order Spectral Analysis*, Vail, Colorado.
- Cardoso, J.-F. (1989) "Blind identification of independent signals," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'89)*, Glasgow, UK.
- Cardoso, J.-F. (1990) "Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'90)*, Albuquerque, NM.
- Cardoso, J.-F. and A. Souloumiac (1993) "Blind beamforming for non Gaussian signals," *IEE Proceedings-F*, vol. 140, # 6, pp. 362-370.
- Cichocki, A. and L. Moshynski (1992) "A new learning algorithm for blind separation of sources," *Electronics Letters*, vol. 28, # 21, pp. 1986-1987.
- Cichocki, A. and R. Unbehauen (1996) "Robust neural networks with on-line learning for blind identification and blind separation of sources," *IEEE Trans. on Circuits and Systems*, vol. 43, # 11, pp. 894-906.

References

- Common, P. (1989) "Separation of stochastic processes," in *Proc. of the Workshop on Higher-Order Spectral Analysis*, pp. 174-179, Vail, Colorado.
- Common, P. (1994) "Independent Component Analysis – A New Concept?" *Signal Processing*, vol.36, pp. 287-314.
- Cover, T.M. and J.A. Thomas (1991) *Elements of Information Theory*, J. Wiley, New York.
- Crosier, R.B. (1988) "Multivariate Generalizations of Cumulative Sum Quality Control Schemes," *Technometrics*, v. 30, #3, pp. 291-303.
- Davison, A.C. and D.V. Hinkley (1997) *Bootstrap Methods and their Applications*, Cambridge Univ. Press.
- De Cooman, G., (1997) "Possibility Theory", *International Journal of General Systems*, v. 25, #4.
- Godard, D.N. (1980) "Self-recovering equalization and carrier tracking in two-dimensional data communication systems," *IEEE Transactions on Communications*, vol. 28, pp. 1867-1875.
- Harman, H.H. (1967) *Modern Factor Analysis*, University of Chicago Press, Chicago, 2nd Ed.
- Hastie, T., R. Tibshirani, and J. Friedman (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.
- Hawkins, D.M., and D.H. Olwell, (1998) *Cumulative Sum Charts and Charting for Quality Improvement*, Springer, NY.

References

- Haykin, S. (1999) *Neural Networks: A Comprehensive Foundation*, Prentice Hall, Upper Saddle River, NJ.
- Hayter, A.J. and K.L. Tsui (1994) "Identification and Quantification in Multivariate Quality Control Problems," *Journal of Quality Technology*, v. 26(3), pp.197-208.
- Himberg, J., and A. Hyvärinen (2001) "Independent Component Analysis for Binary Data: An Experimental Study," in *Proceedings of ICA 2001*, San Diego, CA.
- Himberg, J., A. Hyvärinen, and F. Esposito (2004) "Validating the Independent Components of Neuro-imaging Time-series Via Clustering and Visualization," *NeuroImage* 22(3):1214-1222.
- Hotelling, H. (1947) "Multivariate quality control illustrated by the attesting of sample bomb sights" *Techniques of Statistical Analysis*, C. Eisenhart, M.W. Hastay and W.A. Wallis, eds., McGraw Hill, NY.
- Hyvärinen, A. (1997) "A family of fixed-point algorithms for independent component analysis," in *Proc. of IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP'97)*, pp.3917-3920, Munich, Germany.
- Hyvärinen, A. and E. Oja (1997) "A fast fixed point method for independent component analysis," *Neural Computation*, vol. 9, # 7, pp.1483-1492, 1997.
- Hyvärinen, A. (1998) "New approximations of differential entropy for independent component analysis and projection pursuit," *Advances in Neural Information Processing Systems*, vol. 10, MIT Press, pp. 273-279.

References

- Hyvärinen, A. (1999) "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. on Neural Networks*, vol. 10, #3, pp. 626-634.
- Hyvärinen, A. and P. Pajunen (1999) "Nonlinear Independent Component Analysis: Existence and Uniqueness Results," *Neural Networks*, vol. 12, pp. 429-439.
- Hyvärinen, A. and E. Oja (2000) "Independent Component Analysis: Algorithms and Applications," *Neural Networks*, vol. 13, pp.411-430.
- Hyvärinen, A., J. Karhunen and E. Oja (2001) *Independent Component Analysis*, Wiley-Interscience, New York.
- Jackson, J.E. (1980) "Principal Components and Factor Analysis: Part 1 – Principal Components," *Journal of Quality Technology*, v. 12(4), pp. 201-213.
- Joslyn, C. (1997) "Measurement of Possibilistic Histograms from Interval Data," *International Journal of General Systems*, v. 26, #1-2.
- Jutten, C. and J.Hérault (1991) "Blind separation of sources, Part I: An adaptive architecture based on neuromimetic architecture," *Signal Processing*, vol. 24, pp. 1-10.
- Kim, J.-O., and C.W. Mueller (1978) *Introduction to Factor Analysis*, Quantitative Applications in the Social Sciences #13, Sage Publications, Newbury Park, CA.
- Klir, G. (1985) *Architecture of Systems Problem Solving*, Plenum Publishing, New York.

References

- Klir, G.J. and B.Yuan (1995) "On nonspecificity of fuzzy sets with continuous membership functions," *Proceedings of the 1995 International Conference on Systems, Man, and Cybernetics*, IEEE, pp. 25-29.
- Klir, G.J. and Wierman, M.J. (1998) *Uncertainty-Based Information*, Physica-Verlag, Heidelberg.
- Klir, G. (2001) *Facets of Systems Science*, 2nd ed., Kluwer Academic/Plenum Publishing, New York.
- Klir, G.J. (2006) *Uncertainty and Information: Foundations of Generalized Information Theory*, Wiley Inter-Science, Hoboken, NJ.
- Kourti, T. (2002) "Process Analysis and Abnormal Situation Detection: From Theory to Practice", *Control Systems Magazine*, v. 22, # 5.
- Kourti, T. and J.F. MacGregor (1996) "Recent developments in multivariate SPC methods for monitoring and diagnosing process and product performance," *Journal of Quality Technology*, v. 28(4), pp. 409-428.
- Lane, S., E.B. Martin, R. Kooijmans and A.J. Morris (2001) "Performance monitoring of of a multi-product semi-batch process," *Journal of Process Control*, v. 11(1), pp. 1-11.
- Lo, S.-H., D.A. Buchanan, Y. Taur, and W. Wang (1997) "Quantum Mechanical Modeling of Electron Tunneling Current from Inversion Layer of Ultra-Thin Oxide nMOSFETs," *Electron Device Letters*, IEEE, pp. 209-211, May 1997.

References

- Mansour, A., C. Jutten, and P. Loubaton (2000) "Adaptive Subspace Algorithm for Blind Separation of Independent Sources in Convolutional Mixture," *IEEE Trans. on Signal Processing*, vol. 48, #2, pp. 583-586.
- Martin, E.B., J.A. Morris and S. Lane (2002) "Monitoring process manufacturing performance," *Control Systems Magazine*, v. 22(5), pp. 26-39.
- Meinecke, F., A. Ziehe, M. Kawanabe, and K.-R. Müller (2001) "Assessing Reliability of ICA Projections – A Resampling Approach," *Proc. of ICA 2001*, San Diego, CA.
- Molgedey, L. and H.G. Schuster (1994) "Separation of a mixture of independent signals using time delayed correlations," *Physical Review Letters*, vol. 72, pp. 3634-3636.
- Montgomery, D.C. (2005) *Introduction to Statistical Quality Control*, 5th edition, John Wiley & Sons, Hoboken, NJ.
- Moustakides, G.V. (1986) "Optimal stopping times for detecting changes in distributions," *Annals of Statistics*, v. 14, pp. 1379-1387.
- Ng, K.K. and J.R. Brews (1990) "Measuring the Effective Channel Length of MOSFETs," *Circuits and Devices*, IEEE, pp. 33-38, Nov. 1990.
- Norman, G., and D. Streiner (2000) *Biostatistics: The Bare Essentials*, B.C. Decker, Hamilton, Ontario.
- Page, E.S. (1954) "Continuous Inspection Schemes", *Biometrika*, v. 41, pp. 100-115.

References

- Parsons, R. (1978) *Statistical Analysis: A Decision Making Approach*, Harper & Row, NY.
- Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling (1986) *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, New York.
- Ramer, A. and C. Padet, (2001) "Nonspecificity in R", *International Journal of General Systems*, v. 30, #6, pp.661-680.
- Runger, G.C., F.B. Alt and D.C. Montgomery (1996) "Contributors to a Multivariate Statistical Process Control Signal," *Communications in Statistics – Theory and Methods*, v. 25(10), pp. 2203-2213.
- Scranton, R., G.C. Runger, J.B. Keats and D.C. Montgomery (1996) "Efficient Shift Detection Using Exponential Weighted Moving Average Control Charts and Principal Component Analysis," *Quality and Reliability Engineering International* , v. 12(3), pp. 165-172.
- Shannon, T.T. (2005) "Blind Source Separation with Possibilistic Variables," *Proc. of the North American Fuzzy Information Processing Society Conference*, IEEE.
- Shannon, T.T., D. Abercrombie, J. McNames (2003a) "Modeling CMOS gate characteristics using independent component analysis," *Proc. of the International Joint Conference on Neural Networks*, IEEE, pp. 809-813.

References

- Shannon, T.T., D. Abercrombie, J. McNames (2003b) "Process monitoring via independent components," *Proc. of the IEEE International Conference on Systems, Man and Cybernetics*, IEEE, pp. 3496-3500.
- Shannon, T.T., D. Abercrombie, J. McNames, B. Whitefield (2004) "Improved process monitoring with independent components," *Proceedings of the Advanced Semiconductor Manufacturing Conference (ASMC)*, IEEE.
- Shannon, T.T., and J. McNames (2007) "ICA based disturbance specific control charts," in submission to *IEEE International Conference on Information Reuse and Integration*.
- Singhal, A. and D.E. Seborg (2002) "Pattern matching in historical batch data using PCA," *Control Systems Magazine*, v. 22(5), pp. 53-63.
- Shewart, W.A. (1931) *Economic Control of Quality of Manufactured Product*, Van Nostrand, NY.
- Sorouchyari, E. (1991) "Blind separation of sources, Part III: Stability analysis," *Signal Processing*, vol. 24, pp. 21-29.
- Taleb, A. and C. Jutten (1999) "Source separation in post-nonlinear mixtures," *IEEE Transactions on Signal Processing*, vol. 47, # 10, pp. 2807-2820.
- Tong, L., R.-W. Liu, V.C. Soon, and Y.-F. Huang (1991) "Indeterminacy and identifiability of blind identification," *IEEE Transactions on Circuits and Systems*, vol. 38, pp. 499-509.

References

- Turner, D. (2001) "Error Detection and Principal Component Analysis on a Large Semiconductor Data Set," ECE 457 – Learning From Data Final Report, Portland State University, June 2001.
- Ünday, C. and A. Çinar (2002) "Statistical monitoring of multistage, multiphase batch processes," *Control Systems Magazine*, v. 22(5), pp. 40-52.
- Whitfield, J. (1985) "A Modification on 'An Improved Method to Determine MOSFET Channel Length,'" *Electron Device Letters*, IEEE, 6(3), pp. 109.
- Wu, J.-M. and S.-J. Chiu (2001) "Independent component analysis using Potts models," *IEEE Transactions on Neural Networks*, vol. 12, # 2, pp. 202-211.

Appendix A: Information Theory

In this section, we briefly summarize the definitions and features of the information theoretic concepts of entropy, differential entropy, negentropy, and mutual information.

Definition: The *entropy* $H(x)$ of a discrete random variable x with probability mass function $p(x)$ is defined as

$$H(x) = -E\{\log p(x)\} = -\sum_{x \in X} p(x) \log p(x).$$

Definition: The *joint entropy* $H(x, y)$ of a pair of discrete random variables x and y , with joint probability mass function $p(x, y)$ is defined as

$$H(x, y) = -E\{\log p(x, y)\} = -\sum_{y \in Y} \sum_{x \in X} p(x, y) \log p(x, y).$$

Definition: The *conditional entropy* $H(x | y)$ of a discrete random variable x with respect to a discrete random variable y , with joint probability mass function $p(x, y)$ is defined as

$$H(x | y) = -E_{p(x,y)}\{\log p(x | y)\} = -\sum_{y \in Y} \sum_{x \in X} p(x, y) \log p(x | y).$$

Definition: The *mutual information* $I(x; y)$ between two discrete random variables x and y with joint probability mass function $p(x, y)$ is defined as

$$I(x; y) = \sum_{x \in X, y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}.$$

Proposition: The relationship between entropy and mutual information produces the following identities:

$$I(x, y) = H(x) + H(y) - H(x, y),$$

$$\begin{aligned} I(x, y) &= H(x) - H(y | x), \\ &= H(y) - H(x | y). \end{aligned}$$

Definition: The *differential entropy* $h(x)$ of a continuous random variable x with probability density $p(x)$ is defined as

$$h(x) = -E\{\log p(x)\} = -\int p(x) \log p(x) dx.$$

Proposition: For a continuous random variable x and scalar constant c

$$h(x + c) = h(x),$$

$$h(cx) = h(x) + \log |c|,$$

and for a vector valued continuous random variable \mathbf{x} and non-singular matrix \mathbf{A}

$$h(\mathbf{A}\mathbf{x}) = h(\mathbf{x}) + \log |\mathbf{A}|.$$

Proposition: The maximum entropy distribution for an n -dimensional, vector valued continuous random variable \mathbf{x} with covariance matrix \mathbf{C} is the multivariate Gaussian distribution. If \mathbf{x} is Gaussian with covariance matrix \mathbf{C} then

$$h(\mathbf{x}) = 0.5 \log(2\pi e)^n |\mathbf{C}|.$$

Definition: The *negentropy* $j(\mathbf{x})$ of a continuous random vector of dimension n , with covariance matrix \mathbf{C} is defined as

$$j(\mathbf{x}) = (0.5) \log(2\pi e)^n |\mathbf{C}| - h(\mathbf{x}).$$

Note that for all \mathbf{x}

$$j(\mathbf{x}) \geq 0,$$

and for any non-singular matrix \mathbf{A}

$$j(\mathbf{Ax}) = j(\mathbf{x}).$$

Definition: The *Kullback-Leibler divergence* or *distance* between two densities $p(x)$ and $p(y)$ is defined by

$$D(p(x) \| p(y)) = \int p(x) \log \frac{p(x)}{p(y)} dx dy.$$

Definition: The *mutual information* $I(x; y)$ between two continuous random variables x and y with joint density $p(x, y)$ is defined as

$$I(x; y) = \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy.$$

Clearly from the definitions

$$I(x; y) = D(p(x, y) \| p(x)p(y)).$$

Proposition:

$$D(p(x) \| p(y)) \geq 0,$$

with equality iff $p(x) = p(y)$ almost everywhere. This is a direct consequence of Jensen's inequality.

Corollary: $I(x, y) \geq 0$ with equality iff $p(x, y) = p(x)p(y)$, i.e. x and y are independent.

Appendix B: Cumulants

Definition: The j -th moment $m_j(x)$ of a continuous random variable x with density $p(x)$ is defined as

$$m_j(x) = E\{x^j\} = \int x^j p(x) dx.$$

Definition: The characteristic function $\Psi_x(\omega)$ of a continuous random variable x with density $p(x)$ is defined as

$$\Psi_x(\omega) = E\{e^{i\omega x}\} = \int e^{i\omega x} p(x) dx,$$

which is the Fourier transform of the density.

The characteristic function offers a straightforward approach to calculating the moments of a random variable via the identity

$$m_j(x) = i^{-j} \frac{d^j}{d\omega^j} \Psi_x(\omega),$$

where i is the square root of negative one. The moments of x are the Taylor series coefficients of $\Psi_x(\omega)$, therefore the characteristic function is also called the moment generating function.

Definition: The *second characteristic function* or *cumulants generating function*

$\Phi_x(\omega)$ of a continuous random variable x with density $p(x)$ is defined as

$$\Phi_x(\omega) = \ln(\Psi_x(\omega)) = \ln(E\{e^{i\omega x}\}).$$

Definition: The *j -th cumulant* $k_j(x)$ of a continuous random variable x with density

$p(x)$ is defined as

$$k_j(x) = i^{-j} \frac{d^j}{d\omega^j} \Phi_x(\omega).$$

General expressions for the first four cumulants are

$$k_1(x) = E\{x\},$$

$$k_2(x) = E\{x^2\} - [E\{x\}]^2,$$

$$k_3(x) = E\{x^3\} - 3E\{x^2\}E\{x\} + 2[E\{x\}]^3,$$

$$k_4(x) = E\{x^4\} - 3[E\{x^2\}]^2 - 4E\{x^3\}E\{x\} + 12E\{x^2\}[E\{x\}]^2 - 6[E\{x\}]^4.$$

Note that the first order cumulant is merely the mean or first moment and that the second order cumulant is the variance of the distribution. Also, note that for a zero mean variable, the formulas for the cumulants simplify significantly. Cumulants have the very nice property that for two independent random variables x and y , $k_j(x + y) = k_j(x) + k_j(y)$. Clearly, the cumulants of order three or higher are all zero for a Gaussian distribution.

For zero mean, vector valued random variables we have

$$k(x_i, x_j) = E\{x_i x_j\},$$

$$k(x_i, x_j, x_k) = E\{x_i x_j x_k\},$$

$$k(x_i, x_j, x_k, x_l) = E\{x_i x_j x_k x_l\} - E\{x_i x_j\}E\{x_k x_l\} - E\{x_i x_k\}E\{x_j x_l\} - E\{x_i x_l\}E\{x_j x_k\}.$$

These cumulants are tensors of their respective rank. The second order cumulant is the covariance matrix for the distribution.