IDENTIFYING NONLINEAR VARIATION PATTERNS

IN MULTIVARIATE MANUFACTURING PROCESSES

A Dissertation

by

FENG ZHANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2004

Major Subject: Industrial Engineering

IDENTIFYING NONLINEAR VARIATION PATTERNS

IN MULTIVARIATE MANUFACTURING PROCESSES

A Dissertation

by

FENG ZHANG

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Daniel W. Apley	Yu Ding
(Co-Chair of Committee)	(Co-Chair of Committee)
Georgia-Ann Klutke	Way Kuo
(Member)	(Member)
Dean W. Wichern	Brett A. Peters
(Member)	(Head of Department)

December 2004

Major Subject: Industrial Engineering

ABSTRACT

Identifying Nonlinear Variation Patterns in Multivariate Manufacturing Processes.

(December 2004)

Feng Zhang, B.S., Tsinghua University;

M.S., Tsinghua University

Co-Chairs of Advisory Committee: Dr. Daniel W. Apley
                                    Dr. Yu Ding

 This dissertation develops a set of nonlinear variation pattern identification methods that are intended to aid in diagnosing the root causes of product variability in complex manufacturing processes, in which large amounts of high dimensional in-process measurement data are collected for quality control purposes. First, a nonlinear variation pattern model is presented to generically represent a single nonlinear variation pattern that results from a single underlying root cause, the nature of which is unknown a priori. We propose a modified version of a principal curve estimation algorithm for identifying the variation pattern. Principal curve analysis is a nonlinear generalization of principal components analysis (PCA) that lends itself well to interpretation and also has theoretically rich underpinnings. The principal curve modification involves a dimensionality reduction step that is intended to improve estimation accuracy by reducing noise and improving the robustness of the algorithm with the high-dimensional data typically encountered in manufacturing. An effective visualization

technique is also developed to help interpret the identified nonlinear variation pattern and aid in root cause identification and elimination. To further improve estimation robustness and accuracy and reduce computational expense, we propose a local PCA based polygonal line algorithm to identify the nonlinear patterns.

We also develop an approach for separating and identifying the effects of multiple nonlinear variation patterns that are present simultaneously in the measurement data. This approach utilizes higher order cumulants and pairwise distance based clustering to separate the patterns and borrows from techniques that are used in linear blind source separation. With the groundwork laid for a versatile flexible and powerful nonlinear variation pattern modeling and identification framework, applications in autobody assembly and stamping processes are investigated. The pattern identification algorithms, together with the proposed visualization approach, provides an effective tool to aid in understanding the nature of the root causes of variation that affect a manufacturing process.

ACKNOWLEDGMENTS

First of all I would like to express my very sincere gratitude to Dr. Daniel Apley, the chairman of my advisory committee, for his guidance, support, and advice during my research work. It was a great pleasure to work with Dr. Apley. I learned a great deal from him by observing his philosophy and way of handling both technical and non-technical aspects of life. I would also like to express my appreciation to Dr. Yu Ding, my co-chair, who has offered much help on my dissertation work. I would like to thank Dr. Georgia-Ann Klutke, Dr. Way Kuo, and Dr. Dean W. Wichern for serving on my committee and providing insightful comments on my dissertation, and to Dr. Simpson as the graduate council representative (GCR).

I would like to show my deepest appreciation to my parents, for their endless love and support and also to my sisters for their encouragement. I would like to thank all the past and current students in our department who provided advice and encouragement. My special thanks go to my friends and colleagues, David Weng, Hoyoung Lee, Changho Chin, Xuemei Shan, for their friendship and helpful advice for this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE   Page

FIGURE																							Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

I.1 Motivation

Recent significant advances in measurement and sensing technologies have allowed large quantities of information-rich measurement data to be recorded in databases for potential use in diagnosing and controlling product quality. Autobody assembly is an example of such modern manufacturing processes, where laser-optical measurement stations are built into the assembly line at various stages, providing a wealth of diagnostic information buried in the multivariate measurement data (Apley and Shi, 2001; Ding *et al*., 2002; Ceglarek and Shi, 1996; Jin and Shi, 1999). The dimensional integrity of an autobody is one of the key quality characteristics since poor fit can impact not only the downstream production processes, but also customer's perception of vehicle quality (due to wind noise, water leaking, etc.). For example, in a vehicle door assembly process a small variation in the door hinge produces a large variation at the top edge of the door. The primary objective of autobody assembly process is then to assemble body components together such that minimal dimensional variation will result. Recent developments in measurement systems, such as optical coordinate measuring machines (OCMMs), can measure 100 to 150 points on each major assembly with a 100 percent sample rate. These inspected measurement points are

This dissertation follows the style and the format of *IIE Transactions*.

located along the individual parts of the autobody. In modern manufacturing it is not uncommon to have hundreds, or even thousands (e.g., the stamping process), of different key dimensional characteristics.

Printed circuit board (PCB) assembly is another typical manufacturing process, in which the advanced measurement and data collection technologies are prevalent. For example, the laser-optical, X-ray and vision systems are commonly used to obtain detailed dimensional characteristics of the wet solder paste, after it is deposited onto the board during the screen printing stage (Glass and Thomsen, 1993). Usually the solder paste volumes in PCB assembly are measured in-process with a 100% sample rate.

Each manufacturing process is usually affected by numerous random variation sources, the effects of which will appear in the measurement data. Therefore, the collected data contain valuable diagnostic information concerning the variation sources that contribute to overall levels of product variability. Each variation source typically results in a distinct variation pattern in the multivariate measurement data, which is defined in terms of how variation source causes the different measured variables to vary with respect to each other. In this sense a variation pattern is viewed as the systematic relationships among a number of measured product characteristics on a specific part, describing the interdependencies among the variables caused by a specific source or ultimately a physical root cause.

For the purpose of diagnosing and controlling product quality, an effective methodology would be highly advantageous to utilize all quality-related information from the multivariate measurement data to identify the variation patterns and diagnose

the nature of their corresponding variation sources (or root causes). Proper identification of variation patterns and their root causes is necessary in order to eliminate the root causes, thereby reducing process variability and improving product quality. As in any quality control application, a primary objective of this dissertation is to extract quality-related information from the raw data in a form such that each potential variation pattern can be identified and interpreted precisely to an engineer or operator of the manufacturing process. Equipped with a clearer understanding of the nature of variation patterns present in the measurement data, the manufacturing engineers can utilize their engineering knowledge to identify and eliminate the underlying root causes of process variability. Hence, the variation pattern analysis will serve as diagnostic aids facilitating the ultimate goal of reducing product variability.

This dissertation focuses on nonlinear variation pattern representation and estimation, which is a natural extension of existing linear approaches. The nonlinear approach developed in this dissertation provides a more accurate description of many variation patterns and further identification of unobserved variation sources. A class of nonlinear methods will be investigated and developed for modeling and identifying manufacturing variation patterns.

In the remainder of the introduction, prior work on linear variation pattern analysis will be briefly reviewed in Section I.2, followed by a discussion on the limitations of the linear model and the corresponding nonlinear extension for manufacturing variation pattern identification in Section I.3. Section I.4 presents an outline of the remaining chapters of this dissertation.

I.2 Prior Work on Variation Pattern Analysis

There has been considerable work on identifying variation patterns from large sets of manufacturing measurement data. This section presents the background on these variation pattern analysis approaches, which assume a linear structured model.

Let $x = [x_1, x_2, \ldots, x_d]^T$ be a $d \times 1$ random vector that represents a set of d measured characteristics from the product, and x_n ($n = 1, 2, \ldots, N$) a sample of N observations of x. In autobody assembly, for example, x would represent the vector of all measured dimensional characteristics across a given autobody, and N would be the number of inspected bodies in the sample. The linear variation pattern analysis methods assume that x obeys the model

$$x = \mathbf{C}v + w, \tag{1.1}$$

where $\mathbf{C} = [c_1\ c_2\ \ldots\ c_p]$ is a $d \times p$ constant matrix usually assumed to have linearly independent columns. The vector $v = [v_1\ v_2\ \ldots\ v_p]^T$ is a $p \times 1$ zero-mean random vector with independent components, each scaled (without loss of generality) to have unit variance. The vector w is a $d \times 1$ zero-mean random vector that is independent of v (Apley and Shi, 2001; Apley and Lee, 2003).

The interpretation of the model is that there are p variation sources $\{v_i: i = 1, 2, \ldots, p\}$ that affect the measurement vector x. Each source v_i has a linear effect on x through the corresponding c_i. The vector c_i indicates the spatial nature of the variation caused by the ith source. Specifically, it indicates how the source v_i causes the different measured features to vary with respect to each other. In this sense c_i is referred to as a linear

variation pattern vector. On the other hand, since the elements of v are scaled to be of unit variance, c_i also indicates the severity of the ith source contributing to overall level of process variability. The random vector w represents the aggregated effects of measurement noise and any inherent variation not attributed to the variation sources. Throughout this dissertation, the covariance matrix of noise vector w is assumed to be $\Sigma_w = \sigma^2 I$ where I is a $d \times d$ identity matrix.

We illustrate the linear variation pattern model using an example from autobody assembly (Apley and Lee, 2003). Fig. 1 shows a rear quarter panel subassembly of an autobody. The measurement vector $x = [x_{1y} \; x_{1z} \; \ldots \; x_{5y} \; x_{5z}]^T$ represents the y and z-direction deviations of five separate features (i.e., $d = 10$) that are numbered 1 through 5 in Fig. 1. The measurements are taken after the quarter panel subassembly is joined to the bodyside, which is not shown in the figure.
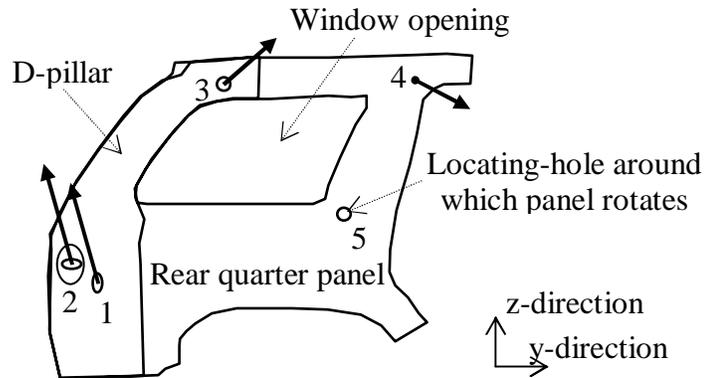
Fig. 1. Illustration of a linear variation pattern in autobody assembly: a rotation of the quarter panel subassembly around locating-hole 5.

For illustration purpose, we present two variation patterns in this example. The root cause of the first variation pattern c_1v_1 is a loose locating element that failed to properly constrain the quarter panel subassembly when it is placed into a fixture and welded to the bodyside. The geometry of the fixture and the position of the loose locating element are such that the quarter panel subassembly is free to rotate by small amounts about Feature 5 (a hole that mates with a pin rigidly attached to the fixture). The variation source v_1 is then a random variable that is proportional to the angle of rotation of each quarter panel subassembly. The vector $c_1 = [-.15\ .52\ -.15\ .53\ .28\ .25\ .29\ -.15\ 0\ 0]^T$ is determined by the fixture design and measurement deployment, and plotted as arrows in Fig. 1 to represent the rotation of the entire subassembly about Feature 5. In order to provide a clear understanding, the y-z coordinates of each feature in c_1 have been combined into a single arrow at the locations of the features to which they correspond.

The root cause of the second variation pattern c_2v_2 is an elongated hole in the D-pillar that is due to improper stamping. The elongated hole allows the D-pillar to translate by small amounts in the z-direction, relative to the rest of the quarter panel. Hence, the variation source v_2 is a random variable that is proportional to the amount of translation of each D-pillar. The variation pattern vector $c_2 = [0\ -.44\ 0\ 0\ 0\ -.44\ 0\ 0\ 0\ 0]^T$ is illustrated in Fig. 2 to represent the z-direction (up/down) translation of the D-pillar with respect to the rest of the quarter panel subassembly. The two variation patterns c_1v_1 and c_2v_2, together with the measurement noise w encountered in manufacturing process, are present in the measurement data x that is represented by the linear model (1.1) for variation pattern analysis.
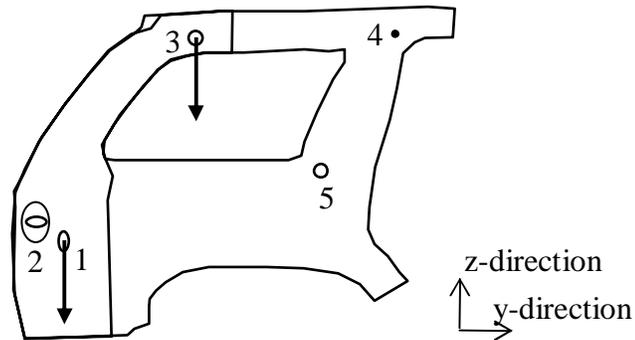
Fig. 2. Illustration of the second linear variation pattern in autobody assembly: a translation of the D-pillar in the z-direction.

As illustrated by the autobody assembly example, a common objective of linear variation pattern analysis methods is to develop a linear model linking product quality measurements to variation sources and extract variation pattern information based on the model for variation diagnosis (Ceglarek and Shi, 1996; Apley and Shi, 2001; Ding *et al*., 2002; Zhou *et al*., 2003). A fundamental distinction among these approaches is whether or not these variation patterns in model (1.1) are pre-modeled. Note that this dissertation focuses on the un-modeled case. However, since most of the prior work is on the pre-modeled case, the next section provides a brief review of that.

I.2.1 The Pre-modeled Variation Pattern Classification Methods

One class of approaches to identify and diagnose the root cause of variation patterns involves off-line analysis based on adequate understanding of the manufacturing processes. These diagnostics require the pattern vectors to be obtained through extensive analytical modeling. When a priori knowledge about the process is provided, it may be possible to accurately model corresponding potential variation patterns

(Ceglarek and Shi, 1996; Apley and Shi, 1998; Jin and Shi, 1999; Ding *et al*., 2002; Zhou *et al*., 2003).

If all of the hypothetical variation patterns are pre-modeled based on the CAD (Computer Aided Design) data for the process configuration and measurement location, principal component analysis (PCA) was used to classify the pre-modeled patterns that are actually present in the on-line measurement data (Ceglarek and Shi, 1996). This pre-modeled pattern classification method recognizes variation patterns effectively by assuming that only a single variation source is present in manufacturing. When there are multiple variation patterns present simultaneously, however, it is unable to detect and classify these different patterns.

Apley and Shi (1998) proposed a method that is capable of detecting and classifying multiple linear variation patterns. From geometric fixture information and measurement layout of the manufacturing process, a multiple variation pattern model is first constructed off-line. A least squares algorithm then estimates the severity of each potential variation source, followed by the detection and classification of multiple variation sources using a form of *F*-test.

The pre-modeled linear variation pattern methods were also applied to multistage manufacturing processes based on a state-space model, which considered the variation propagation along the production stages (Jin and Shi, 1999; Ding *et al*., 2002; Zhou *et al*., 2003).

All of the pre-modeled variation patterns classification approaches assumed that a complete set of potential variation patterns can be pre-modeled off-line. On the other

hand, if there is variation pattern present in the measurement data that is not pre-modeled, they are likely to miss this pattern and produce erroneous interpretations. Due to the complexity of the modern manufacturing processes, however, pre-modeling every variation pattern is not possible when the number of potential variation sources is too large to analytically model. The necessity for off-line pre-modeling therefore limits their applicability in practical dimensional variation pattern diagnostics.

I.2.2 The Un-modeled Variation Pattern Identification Methods

Some other linear variation pattern analysis methods are developed to overcome the aforementioned limitations by estimating the pattern vectors c_i blindly from the on-line data, with no prior knowledge about the nature of potential fault patterns. After blindly estimating the patter vectors, they can be graphically illustrated to lend insight into the root causes, as shown in the above autobody assembly example. Due to the similarity in model structure as in Equation (1.1), one may consider using factor analysis to estimate variation matrix $\mathbf{C}$ from data x. Factor analysis is to extract unobservable common factors from covariance matrix of multivariate data (Jackson, 1980; 1981; Johnson and Wichern, 2002).

Let $\mathbf{\Sigma}_x$ denote the covariance matrix of x. From the assumptions on model (1.1)

$$\mathbf{\Sigma}_x = E[(\mathbf{C}v + w)(\mathbf{C}v + w)^{\mathrm{T}}] = \mathbf{C}\mathbf{C}^{\mathrm{T}} + \sigma^2\mathbf{I}. \tag{1.2}$$

Most factor analysis methods are based on PCA, which involves analyzing the eigenvectors and eigenvalues of $\mathbf{\Sigma}_x$. Denote $\{\lambda_i: i = 1, 2, \ldots, d\}$ as the eigenvalues of

$\boldsymbol{\Sigma}_x$ arranged in descending order, and z_i as the corresponding eigenvectors. Each observation $\boldsymbol{x}$ can be written as a linear combination of eigenvectors, i.e.,

$$\boldsymbol{x} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}_1)\boldsymbol{z}_1 + (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}_2)\boldsymbol{z}_2 + \ldots + (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}_d)\boldsymbol{z}_d,$$

$$= y_1\boldsymbol{z}_1 + y_2\boldsymbol{z}_2 + \ldots + y_d\boldsymbol{z}_d, \qquad (1.3)$$

where $y_i = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}_i$ are uncorrelated PCA scores on eigenvectors $\boldsymbol{z}_i$. Define $\boldsymbol{\Lambda}_p = \mathrm{Diag}\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$ and $\mathbf{Z}_p = [\boldsymbol{z}_1 \ \boldsymbol{z}_2 \ \ldots \ \boldsymbol{z}_p]$ as the p largest eigenvalues and their associated eigenvectors. The PCA decomposition on covariance matrix of $\boldsymbol{x}$ yields

$$\boldsymbol{\Sigma}_x = \sum_{i=1}^{d} \lambda_i \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}} = \sum_{i=1}^{p} \left(\lambda_i - \sigma^2\right)\boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}} + \sigma^2 \sum_{i=1}^{d} \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}} = \mathbf{Z}_p[\boldsymbol{\Lambda}_p - \sigma^2 \mathbf{I}]\mathbf{Z}_P^{\mathrm{T}} + \sigma^2 \mathbf{I}. \qquad (1.4)$$

In order for the covariances in Equations (1.2) and (1.4) to be consistent, $\mathbf{C}$ has the form $\mathbf{Z}_p[\boldsymbol{\Lambda}_p - \sigma^2 \mathbf{I}]^{1/2}\mathbf{Q}$ for a $p{\times}p$ orthogonal matrix $\mathbf{Q}$. Thus, PCA simplifies the problem of estimating the $d{\times}p$ matrix $\mathbf{C}$ to one of estimating the $p{\times}p$ matrix $\mathbf{Q}$.

The standard factor analysis methods usually apply the varimax rotation to find $\mathbf{Q}$ such that each column of resulting $\mathbf{C} = \mathbf{Z}_p[\boldsymbol{\Lambda}_p - \sigma^2 \mathbf{I}]^{1/2}\mathbf{Q}$ consists of elements that are either very large or very small in magnitude with as few moderate-sized elements as possible (Jackson, 1981; Johnson and Wichern, 2002). However, the estimate of $\mathbf{C}$ from the varimax rotation is not necessarily the true matrix in model (1.1) that is caused by the physical characteristics of the manufacturing processes.

The method presented in Apley and Shi (2001) can be viewed as a modified form of standard factor analysis that attempts to rotate the estimate of $\mathbf{C}$ so that it is as close as possible to the true $\mathbf{C}$. To accomplish this, they assume that $\mathbf{C}$ has a ragged lower

triangular structure in order that the measurements in x can be grouped and each subgroup is affected by a single variation source. Although the method in (Apley and Shi, 2001) produces a more accurate identification of the variation sources and clearer understanding of their root causes, the certain structural constraints on the true $\mathbf{C}$ limit its applicability to some extent. To overcome this limitation, Apley and Lee (2003) introduced the blind source separation (BSS) method to identify multiple linear variation patterns over measurement data for manufacturing variation diagnosis.

Blind source separation is a term used to describe a problem in which there is an array of spatially distributed sensors, each of which receives signals from some unknown sources (Hyvarinen and Oja, 2000; Hyvarinen *et al.*, 2001). The term "blind" refers to the situation when no prior knowledge of the relationship between x and v is available other than the linear model structure (1.1). Due to the similarity of the model format, the BSS methods thus are applicable to manufacturing variation pattern analysis. Apley and Lee (2003) estimated the matrix $\mathbf{C}$ and individual variation sources solely from the data sample for subsequent root cause diagnosis. Blind source separation based methods have broader applicability than the preceding linear variation pattern analysis approaches in that they make no assumptions regarding the structure of matrix $\mathbf{C}$ other than on the variation source distributions. Apley and Lee (2003) also proposed a combinational BSS method with weaker requirements on the distributions of variations sources.

The pre-modeled variation pattern classification methods require that all of the potential variation patterns should be pre-modeled through intensive off-line analytical

modeling based on engineering knowledge of the manufacturing process. In reality, the presence of too many potential variation sources and the complexity of advanced manufacturing processes often make off-line modeling of all potential patterns impossible and, therefore, prevent the wide applicability of the classification methods. In this sense, the un-modeled variation pattern analysis methods have broader applicability in that they make no assumptions regarding the structure of matrix **C** and no off-line analytical modeling is required.

I.3 Nonlinear Variation Patterns

The linear variation pattern analysis methods are effective and have been widely used in manufacturing variation reduction. Linear models are attractive for their simplicity and amenability to analysis and visualization, as illustrated in the above autobody assembly example, however, they are inadequate for modeling the nonlinear interdependencies in the multivariate measured data, which are common in the complicated manufacturing processes. Therefore, for the purpose of diagnosing all potential variation patterns in real manufacturing processes, linear methods are incapable of capturing all the diagnostic related information and new methodology is expected for nonlinear variation faults analysis. To illustrate such a nonlinear variation pattern problem, consider another example from autobody assembly. Fig. 3 shows schematically the rear liftgate opening of a sports utility vehicle and indicates the locations at which six cross-car (left/right) dimensional measurements on the left and right bodysides are taken (denoted x_1 through x_6). The measurements are obtained

automatically via in-process laser measurement, so that 100% of the automobile bodies are measured. Although almost 200 different dimensional features were measured for each autobody, for simplicity we illustrate with only the six measurements shown in Fig. 3.
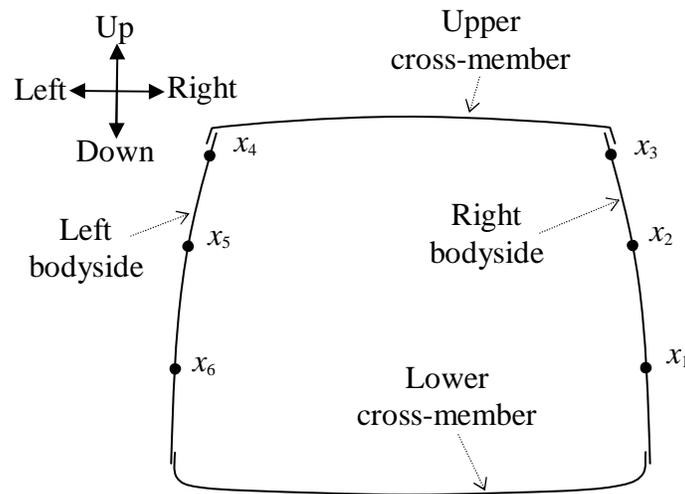
Fig. 3. Measurement layout on the liftgate opening of an autobody. The fore direction is pointing into the page.

As discussed in Section I.2, the assembly process is relatively complex and involves many different locating elements. The malfunction of a locating element often results in a distinct variation pattern in the dimensional measurement data. Fig. 4 illustrates this with scatter plots of pairs of the six variables over a sample of 100 measured autobodies. The measurements are deviations from nominal, in units of mm. A positive measurement represents deviation to the right. Although the relationship between x_2 and x_3 appears linear, the scatter plots for x_2/x_5 and x_3/x_4 clearly illustrate that the variation

pattern is nonlinear. Moreover, this nonlinear pattern appears to be approximately piecewise linear with only two segments (i.e., pieces).
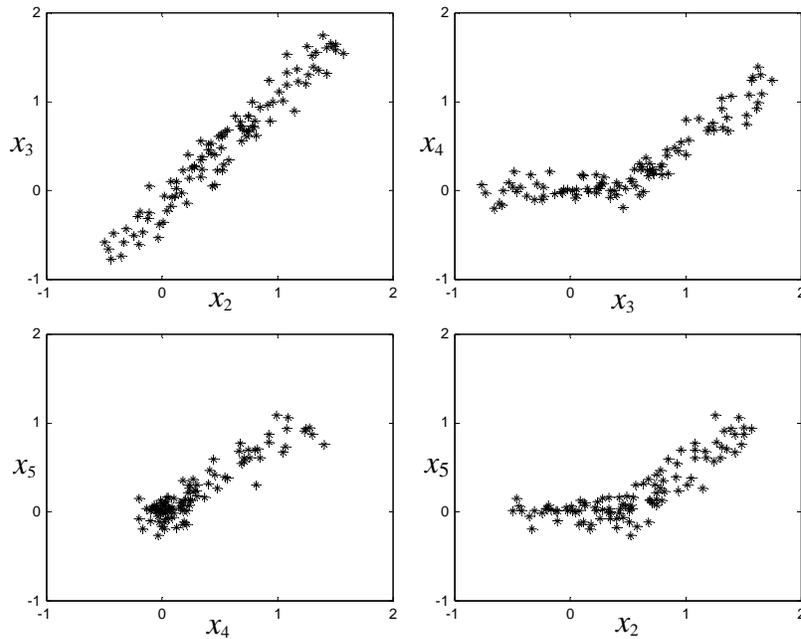
Fig. 4. Scatter plots of data from 100 measured automobile bodies exhibiting a nonlinear variation pattern.

I.4 Outline of Dissertation

In situations like that depicted in Fig. 4, linear models are inadequate for representing the nonlinear relationship between the different variables. To accommodate such situations, this dissertation will develop an approach for modeling and identifying nonlinear variation patterns that provides a more accurate way for diagnosing the manufacturing variation patterns. The objective considered in this dissertation is to identify as precisely as possible the nature of any variation pattern that

happens to be present in the data. The identification is "blind" in the sense that we are not attempting to recognize the presence of pre-modeled or pre-trained patterns. Rather, we seek to identify the nature of the patterns based only on a sample of data, with no prior training or modeling required. After blindly identifying the nature of a variation pattern, the results can be graphically illustrated in order to facilitate root cause identification.

The rest of the dissertation is organized as follows. The single nonlinear variation pattern modeling and identification problem is discussed in Chapter II, in which a data pre-processing approach is proposed to improve the pattern estimation accuracy. This pre-processing step can also modify the exiting nonlinear pattern identification method such that it is applicable for the high dimensional manufacturing problems when the nonlinear variation pattern lies in a linear variety (a translated linear subspace) of dimension $r < d$. For applications such as two-dimensional image processing, in which the dimension d of the original data is low, the proposed method offers no advantages over the current nonlinear pattern estimation. The utility of the method lies primarily in applications such as identifying and diagnosing variation patterns in large multivariate datasets, in which d is large and $r \ll d$. To help interpret the identified nonlinear variation pattern results, an interactive visualization method was developed to illustrate how the spatial characteristics of measurement points will evolve under the influence of the variation source.

It is also shown in Section II.4 that applying nonlinear pattern identification algorithm to only the first r PCA scores produces results that are equivalent to applying

the identification algorithms to the filtered d-dimensional data. This allows one to work in r dimensional space, as opposed to the full d dimensional space, which substantially reduces the computational expense of the algorithms when r is much less than d.

As will be discussed in Chapter II, there are many situations where this is the case, in particular when the nonlinear pattern can be closely approximated as piecewise linear. The nonlinear variation pattern will be approximated by polygonal lines that achieve the minimum of the sum of squared distance from the sample data to their projections on the piecewise linear curve. Inspired by the result in Chapter II, the number of line segments can be automatically determined by PCA information of the local linear space, as discussed by the proposed algorithm in Chapter III. The problem of polygonal line estimation is then reduced to determining the optimal positions of vertices that connects these line segments. The experimental results will demonstrate the robustness of the proposed polygonal line algorithm to a variety of nonlinear variation patterns.

The single nonlinear variation pattern identification methods in Chapter II and III cannot accommodate to the multiple variation sources problem in practical manufacturing applications. Chapter IV presents a method of identifying un-modeled multiple linear and nonlinear variation patterns by utilizing information of high order statistics (see Section IV.2.1). The proposed method can blindly determine the number of variation sources from on-line measurement data for subsequent identification of the individual linear pattern and nonlinear patterns (see Section IV.2.2). This multiple variation patterns identification algorithm has wider applicability for separating and

identifying nonlinear patterns, since there are no requirements on the differentiability of the nonlinear patterns or prior knowledge of the number of variation sources. To help separate and identify the variation patterns when only sample data are available, a clustering method is developed based on a pairwise similarity quantity taken as an independence measure (see Section IV.3). By utilizing this multiple variation patterns model and identification algorithm, the method proposed in Chapter IV is capable of estimating un-modeled variation patterns more effectively, which will be demonstrated by an autobody assembly example.

The dissertation ends with a conclusion in which the main results of this work are summarized and directions of future work are indicated.

CHAPTER II

SINGLE NONLINEAR VARIATION PATTERN IDENTIFCATION

AND VISUALIZATION USING PRINCIPAL CURVES METHOD

II.1 Representing Single Nonlinear Variation Pattern

In order to transform large sets of manufacturing measurement data into knowledge by discovering the nature of major variation patterns, a number of techniques have been recently developed (e.g., Apley and Shi, 2001; Apley and Lee, 2003; Ding *et al.*, 2002; Jin and Shi, 1999; Barton and Gonzalez-Barreto, 1996), as discussed in Section I.2. All of these approaches assume the variation patterns can be represented using linear model (1.1). If there is only a single variation source present in the manufacturing data, the linear model takes the form

$$x = ct + w, \tag{2.1}$$

where c is a $d \times 1$ constant vector and t is a scalar random variable that is scaled to have unit variance. The vector c indicates the spatial nature of the variation pattern, in terms of the resulting interrelationships between the different measured variables.

Although linear models of this form are relatively common for representing manufacturing variation when one variation pattern occurs, there are many situations in which nonlinear models are required to represent the nonlinear relationship between the different variables, such as the autobody assembly example introduced in Section I.3,

where the scatter plots in Fig. 4 clearly illustrate the presence of a nonlinear variation pattern.

In order to represent nonlinear variation patterns in analogy with the linear model (2.1), we propose the following model:

$$x = f(t) + w, \tag{2.2}$$

where t and w are as in the linear model, and $f(\cdot)$ is some general vector-valued nonlinear function that represents the spatial nature of the nonlinear variation pattern. As assumed in Chapter I, the random vector w represents the measurement noise and has an isotropic covariance matrix $\Sigma_w = \sigma^2 I$. Apley and Lee (2003) discussed approaches for dealing with more general noise covariance structure, which also apply to the present case of nonlinear patterns.

For the nonlinear variation pattern in the autobody assembly example as introduced in Section I.3, its root cause is a fixturing problem when the right bodyside was clamped in the framing station (a major assembly station in which the left and right bodysides are joined to the underbody and a set of upper cross-members). When the right bodyside deviates by only a small amount to the right, it has no effect on the left bodyside. When the right bodyside deviates by a larger amount to the right, however, it begins to interfere with the upper cross-member. The upper cross-member then interferes with the left bodyside, pulling the left bodyside to the right also.

Therefore, the nonlinear variation pattern $f(\cdot)$ would be a piecewise linear curve with two pieces, as shown in Fig. 5. And t is a random variable that is scaled as the arc length along the curve f. Note that in this example, t does not have a clear physical

meaning. In fact, the random variable *t* can be arbitrarily rescaled without changing the curve that passes through the data. This does not present a problem, because it is the nature of the function $f(\bullet)$ that aids in interpreting the root cause.
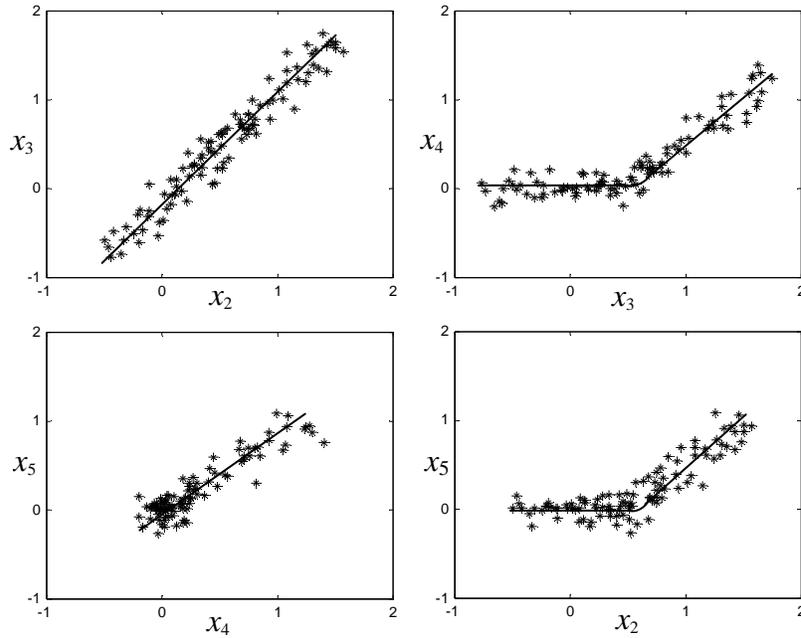
Fig. 5. Scatter plots of data and the components f_i (solid line) of nonlinear variation pattern *f*.

Sometimes, however, *t* does have a physical meaning. To further clarify the nonlinear variation pattern representation (2.2), consider an example from the metal flanging process, which forms the edges of the panels into a flange to add stiffness to the panel or create a mating surface. In the flanging process, however, wrinkling often appears in the formed flange surface, which not only mars the appearance of the part,

but may also interfere with subsequent sealing and welding operations. Fig. 6 shows the wrinkling encountered in a flange operation, in which the panel is formed into a concave shape at the same time the flange is formed.
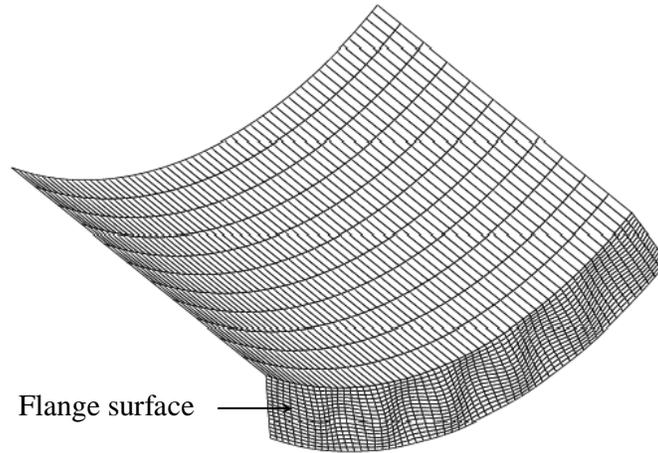
Flange surface

Fig. 6. Illustration of a panel in shrink flanging exhibiting wrinkles on the flange surface.

An analytical model of wrinkling for the flanging process shown in Fig. 6 is (refer to Wang *et al*., 2001 for details on the derivation of the model):

$$g(\theta, r) = g_0 \sin(\frac{m\pi\theta}{\theta_0}) \sin(\frac{n\pi(r - R_f)}{f_l \sin\beta}), \qquad (2.3)$$

where g_0 is an amplitude constant of wrinkle height g, m and n are the wave numbers in the tangential and radial direction. The other parameters are denoted in Fig. 7.

Front View   Side View
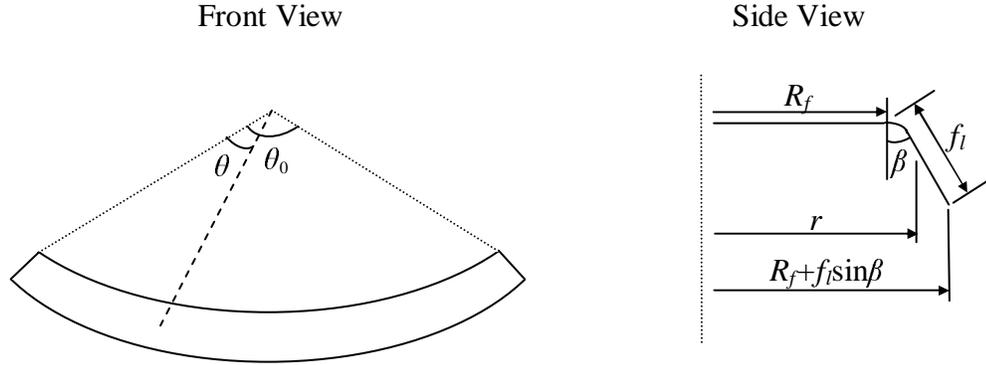
Fig. 7. Schematic of a flanged sheet model.

In practice, as boundary conditions (e.g., binding force) vary from part to part, the characteristics of the wrinkles on flange surface will change. In this example, for illustration purpose, the variation pattern f denotes the position of the wrinkles varying along the length of the flange and is defined as a function of random variable t, i.e.,

$f(t) = [g(\theta_1 - t, r_1)\, g(\theta_2 - t, r_2) \ldots g(\theta_d - t, r_d)]^T$, where $(\theta_i, \quad r_i)$ denotes the ith measurement location distributed across the panel ($i = 1, 2, \ldots, d$), and t is taken as a phase angle to represent the variation on the boundary conditions with respect to each flanged part. The vector x measures the height of the wrinkles (normal to the flange surface) over an array of the d locations along the flange surface. Note that in real manufacturing, the measurement data could represent point cloud data from a laser scan of the panel.

The plot of nonlinear variation pattern $f(\bullet)$ shown in Fig. 8 changes dynamically as we move the slide bar control back and forth. The position of the slide bar represents

the value of t for which nonlinear pattern is currently plotted. Fig. 8 illustrates in a graphical way that how the wrinkles shift along the length of flange from part to part with respect to the varying values of t. Therefore, in this example, the random variable t in model (2.2) is equipped with a physical understanding as the variation on the boundary conditions, which can facilitate the interpretation of nonlinear variation pattern $f(\cdot)$ for gaining insight into the behavior of wrinkling.
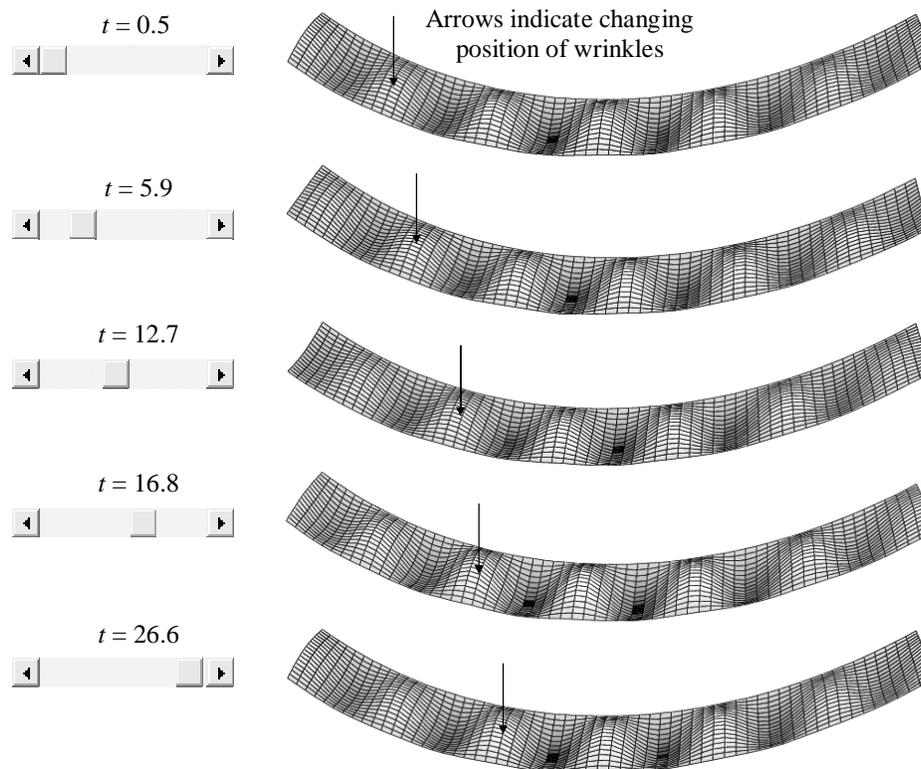
Fig. 8. Illustration of the variation pattern in the flange for five different frames given different values of t. The slide bar control in the left panel shows the values of t for the five frames.

In the linear situations, the objective is to blindly estimate c, given a sample of data. Apley and Shi (2001) and Apley and Lee (2003) provide examples of this and discuss how graphical illustrations of the estimated c can aid in identifying the root cause of the variation. The above wrinkling example also illustrates the importance of visualizing nonlinear variation pattern for facilitating the identification of the root causes. Graphical displays of high dimensional variation patterns, such as Fig. 8, are often illuminating, especially when enhanced by interaction (e.g. changing the value of t by moving the slider bar control) for diagnosing the nature of f and describing how the dimensional variation is governed by the random variable t. Therefore, the objective of nonlinear variation pattern analysis is to blindly estimate the entire function f and develop a suitable illustration for aiding in root cause identification. The remainder of this chapter focuses on how to estimate f.

II.2 PCA and Principal Curves

For linear variation patterns, PCA has been used to provide an estimate of c. It is well known that when a single linear variation pattern is present, the dominant eigenvector z_1 is a scaled version of c (Ceglarek and Shi, 1996; Apley and Shi, 2001). Therefore, the dominant eigenvector of the sample covariance matrix can be used as an estimate of c.

Denote μ_x as the mean of the d-dimensional random vector x. One related, fundamental property of PCA is that the line $\mu_x + tz_1$, with t a scalar, provides the best one-dimensional linear approximation to the distribution of x, in the sense of

minimizing the mean squared (orthogonal) distance between x and the approximating line. Principal curves are a natural nonlinear generalization of this concept. Loosely speaking, a principal curve is defined as a one-dimensional curve that passes through the middle of the distribution of multivariate data. Principal curve estimation has been applied to a variety of nonlinear data analysis problems (Delicado, 2001; Chang and Ghosh, 2001; Tibshirani, 1992; Dong and McAvoy, 1995). Because $f(t)$ in Equation (2.2) satisfies the definition of a principal curve stated below, we use the same notion for a principal curve.

The concept of principal curve was first proposed in (Hastie and Stuetzle, 1989) as a non-parametric nonlinear extension of the first principal component. According to the definition of Hastie and Stuetzle (hereafter referred to as HS) a one-dimensional curve $f(t) = [f_1(t), f_2(t), \ldots, f_d(t)]^{\mathrm{T}}$ (i.e., $f: \Re \rightarrow \Re^d$) in d-dimensional space is a principal curve of (the distribution of) x if:

1. $f(t)$ does not intersect itself, or $t_1 \neq t_2 \Rightarrow f(t_1) \neq f(t_2) \; \forall t_1, t_2 \in \Re$;

2. $f(t)$ has finite length inside any bounded subset of $\Re^d$; and

3. $f(t) = E(x/t_f(x) = t)$, where $E(\bullet|\bullet)$ denotes a conditional expectation, and the projection index $t_f(x)$ is defined as

$$t_f(x) = \sup\{v : \|x - f(v)\| = \inf_u \|x - f(u)\|\}. \tag{2.4}$$

HS showed that no infinitesimally small smooth perturbation to a principal curve will decrease $E\|x - f(t_f(x))\|^2$. In this sense principle curves generalize the minimum mean squared distance property of the linear PCA approximation to the

distribution of x. In the hypothetical case that the distribution of x is known, the HS algorithm for constructing principal curves iterates over the following steps:

Step 0. Initialize $f^{(0)}(t) = \mu_x + tz_1$, and set $j = 1$.

Step 1 (Expectation). Define $f^{(j)}(t) = E[x \mid t_{f^{(j-1)}}(x) = t]$.

Step 2 (Projection). For $x \in \Re^d$, set $t_{f^{(j)}}(x) = \max\{v : \|x - f^{(j)}(v)\| = \min_u \|x - f^{(j)}(u)\|\}$.

Step 3. Compute $\Delta(f^{(j)}) = E\|x - f^{(j)}(t_{f^{(j)}}(x))\|^2$. If $|\Delta(f^{(j)}) - \Delta(f^{(j-1)})| < threshold$,

then stop. Otherwise, let $j = j + 1$ and go to Step 1.

In practice, the theoretical distribution of x is unknown and the HS algorithm must be implemented on a specific sample. The actual HS algorithm starts with $\hat{f}^{(0)}(t) = \hat{\mu}_x + t\hat{z}_1$ and the expectations in Steps 1 and 3 are replaced by some form of locally weighted sample averages, as shown in Fig. 9. Throughout, the overscore symbol "^" will be used to denote an estimate of a quantity.

(a) $f(t) = E[x \mid t_f(x) = t]$ (b) $\hat{f}(t) = \hat{E}[x \mid t_f(x) \approx t]$
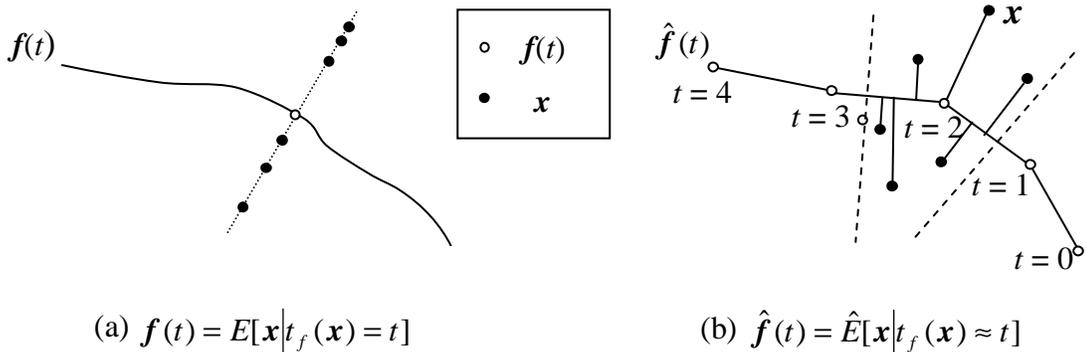
Fig. 9. (a) Ideally, with infinite data, every point $f(t)$ on the principal curve is defined as the average of all data points x projecting exactly on t. (b) In practice, with limited data, a point on the curve is the sample average of all points x projecting within a neighborhood of t.

II.3 Principal Curves in Lower Dimensional Linear Varieties

As one potential method for treating nonlinear variation pattern in model (2.2), principal curve method is found to be intuitive and interpretable for variation pattern analysis. However, most applications of principal curves have been for low dimensional data, especially for two-dimensional image processing (e.g., Kégl *et al.*, 1999; Chang and Ghosh, 2001; Banfield and Raftery, 1992). For high dimensional data encountered in manufacturing, principal curve estimation becomes inefficient.

In order to effectively use principal curve concepts for identifying nonlinear variation patterns in the high dimensional data collected from manufacturing processes, we propose a pre-processing step that filters out much of the noise and reduces the dimensionality of the data. Technically, this will require that $f(t)$ lie in a linear variety of dimension $r < d$. There are many practical situations where nonlinear variation patterns lie in lower dimensional subspaces, in particular when $f(t)$ can be closely approximated as a piecewise linear curve. This was the case in the autobody example introduced in Section I.3, because (as will be shown as below) a piecewise linear curve with p segments lies in a linear variety of dimension $r \leq p$. This has important implications for high-dimensional data, since any nonlinear principal curve can be approximated arbitrarily closely by a piecewise linear curve. As another example, suppose that only $r < d$ elements of $f(t)$ vary with t and that the other elements remain constant. In other words, $d - r$ elements of x are unaffected by the variation pattern. The principal curve would also lie in an r-dimensional linear variety in this case.

Let $\mathbf{M}$ denote some r-dimensional subspace of $\mathfrak{R}^d$, and let $\boldsymbol{a}_0 + \mathbf{M}$ denote the r-dimensional linear variety that results from translating $\mathbf{M}$ by some constant vector $\boldsymbol{a}_0$. In situations where $\boldsymbol{f}(t)$ lies in $\boldsymbol{a}_0 + \mathbf{M}$, the following theorem relates $\boldsymbol{a}_0 + \mathbf{M}$ to the results of standard linear PCA.

Theorem 2.1: *Suppose that $\boldsymbol{f}(t)$ lies in the r-dimensional linear variety $\boldsymbol{a}_0 + \mathbf{M}$ for all t and that the noise covariance is $\boldsymbol{\Sigma}_w = \sigma^2 \mathbf{I}$. Suppose also that no other linear variety in which $\boldsymbol{f}(t)$ lies has dimension smaller than r. Then*

$$\lambda_1 \geq \lambda_2 \geq \ ... \ \geq \lambda_r > \sigma^2 = \lambda_{r+1} = \lambda_{r+2} ... = \lambda_d \text{, and} \tag{2.5}$$

$$\boldsymbol{a}_0 + \mathbf{M} = \boldsymbol{\mu}_x + \operatorname{span}\{z_1, z_2, \ ..., z_r\}. \tag{2.6}$$

Proof: First note that because $\boldsymbol{f}(t) \in \boldsymbol{a}_0 + \mathbf{M}$ for all t, it must be the case that its mean $\boldsymbol{\mu}_f$ also lies in $\boldsymbol{a}_0 + \mathbf{M}$. Also, because w is zero-mean, $\boldsymbol{\mu}_x = \boldsymbol{\mu}_f$. Hence, $\boldsymbol{\mu}_x \in \boldsymbol{a}_0 + \mathbf{M}$, so that the linear variety $\boldsymbol{a}_0 + \mathbf{M}$ is the same as $\boldsymbol{\mu}_x + \mathbf{M}$. By the independence of w and t,

$$\boldsymbol{\Sigma}_x = E[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)^{\mathrm{T}}] = E[(\boldsymbol{f}(t) - \boldsymbol{\mu}_x + w)(\boldsymbol{f}(t) - \boldsymbol{\mu}_x + w)^{\mathrm{T}}]$$

$$= E[(\boldsymbol{f}(t) - \boldsymbol{\mu}_x)(\boldsymbol{f}(t) - \boldsymbol{\mu}_x)^{\mathrm{T}}] + \sigma^2 \mathbf{I}.$$

Now consider any unit-norm vector $z \in \mathbf{M}^{\perp}$ (the orthogonal complement of $\mathbf{M}$). Because $\boldsymbol{f}(t) - \boldsymbol{\mu}_x$ lies in $\mathbf{M}$ and z lies in the orthogonal complement of $\mathbf{M}$, $(\boldsymbol{f}(t) - \boldsymbol{\mu}_x)^{\mathrm{T}}z = 0$ for all t. Therefore, $\boldsymbol{\Sigma}_x z = E[(\boldsymbol{f}(t) - \boldsymbol{\mu}_x)(\boldsymbol{f}(t) - \boldsymbol{\mu}_x)^{\mathrm{T}}z] + \sigma^2 \mathbf{I}z = \sigma^2 z$. Hence, z is an eigenvector of $\boldsymbol{\Sigma}_x$ with eigenvalue σ^2. Because $\mathbf{M}^{\perp}$ is a $d - r$ dimensional subspace, there exists an orthonormal set of $d - r$ such eigenvectors, which we denote $\{z_{r+1}, z_{r+2}, \ ..., z_d\}$. Because the remaining eigenvectors $\{z_1, z_2, \ ..., z_r\}$ are orthogonal to $\{z_{r+1}, z_{r+2},$

…, z_d}, they must all lie in the r-dimensional subspace **M**. Consequently, **M** = span{z_1, z_2, …, z_r}, which completes the proof of Equation (2.6).

Because $z_j \in$ **M** for $1 \le j \le r$, its corresponding eigenvalue is

$$\lambda_j = z_j^T(\lambda_j z_j) = z_j^T(\Sigma_x z_j) = z_j^T\{E[(f(t) - \mu_x)(f(t) - \mu_x)^T] + \sigma^2 \mathbf{I}\}z_j$$

$$= E[z_j^T(f(t) - \mu_x)(f(t) - \mu_x)^T z_j] + \sigma^2$$

$$= E[(z_j^T(f(t) - \mu_x))^2] + \sigma^2 > \sigma^2.$$

That $E[(z_j^T(f(t) - \mu_x))^2]$ is strictly greater than zero follows from the condition that no other linear variety in which $f(t)$ lies has dimension smaller than r. Indeed, if $E[(z_j^T(f(t) - \mu_x))^2] = 0$ for some $1 \le j \le r$, this would imply that $f(t)$ lies in the $r - 1$ dimensional linear variety μ_x + span{$z_1, z_2, …, z_{j-1}, z_{j+1}, …, z_r$}. ■

The significance of the theorem is that when $f(t)$ lies in an r-dimensional linear variety, PCA can be used to identify the linear variety. The linear variety is given by the span of the first r eigenvectors, translated by μ_x, where r is equal to the number of dominant eigenvalues. Although Equation (2.5) applies to the theoretical covariance matrix Σ_x, in practice, PCA is conducted on a sample covariance matrix. In this case, it may not be clear what value should be chosen for r. There are, however, a number of statistical methods for estimating the number of dominant eigenvalues in this situation (see Apley and Shi, 2001).

The following corollary relates to how the results of PCA can be used to filter the data in order to reduce the effects of the noise in principle curve estimation.

Corollary 2.2: *Consider the $d \times r$ matrix* $\mathbf{Z} = [z_1\; z_2\; \ldots\; z_r]$ *and the $d \times d$ matrix* $\mathbf{P}_z =$ $\mathbf{ZZ}^T$, *which is the projection operator onto* $\mathbf{M}$. *Define the "filtered" version* $x_z = \mathbf{P}_z(x$ $- \boldsymbol{\mu}_x) + \boldsymbol{\mu}_x$ *of the random vector* x. *Then*

$$x_z = f(t) + w_z ,$$

where $w_z = \mathbf{P}_z w$ *is the filtered version of the noise, obtained by projecting* w *onto the r-dimensional subspace* $\mathbf{M}$.

Proof: $\quad x_z = \mathbf{P}_z(x - \boldsymbol{\mu}_x) + \boldsymbol{\mu}_x = \mathbf{P}_z(f(t) - \boldsymbol{\mu}_x + w) + \boldsymbol{\mu}_x$

$$= f(t) - \boldsymbol{\mu}_x + \mathbf{P}_z w + \boldsymbol{\mu}_x = f(t) + w_z. \qquad\qquad (2.7)$$

The third equality follows from the fact that $f(t) - \boldsymbol{\mu}_x \in \mathbf{M}$. ∎

The corollary implies that by filtering the data using the results of PCA, the principal curve component $f(t)$ of the model (2.2) is unchanged, whereas the noise component w is reduced by an amount that depends on d and r. To quantify the extent to which the noise is reduced, consider the total variance $E[w^T w]$ as a measure of the noise level. Before filtering, the total noise variance is $E[w^T w] = \text{trace}\{E[ww^T]\} = \text{trace}\{\sigma^2 \mathbf{I}\} = d\sigma^2$. After filtering, the total variance of the filtered noise is $E[w_z^T w_z] = \text{trace}\{E[w_z w_z^T]\} = \text{trace}\{\mathbf{P}_z \boldsymbol{\Sigma}_w \mathbf{P}_z\} = \sigma^2 \text{trace}\{\mathbf{P}_z \mathbf{P}_z\} = r\sigma^2$. Thus, the ratio of total noise variance before and after filtering is d/r. When $f(t)$ lies in a linear variety of dimension much smaller than d, the reduction in noise variance will be substantial. The simulation results presented in Section II.6 demonstrate that this improves the principal curves estimation accuracy.

In the autobody assembly example introduced in Section I.3, the principal curve $f(t)$ appears to be piecewise linear with two linear segments. More generally, suppose $f(t)$ is piecewise linear with p linear segments, which is illustrated in Fig. 10 for the case that $p = 4$ and $d = 3$. Although we illustrate this with $p > d$, the primary utility of the approach will be for situations in which $p \ll d$. The following arguments show that a piecewise linear $f(t)$ lies in a linear variety whose dimension is equal to the number of linearly independent pieces.
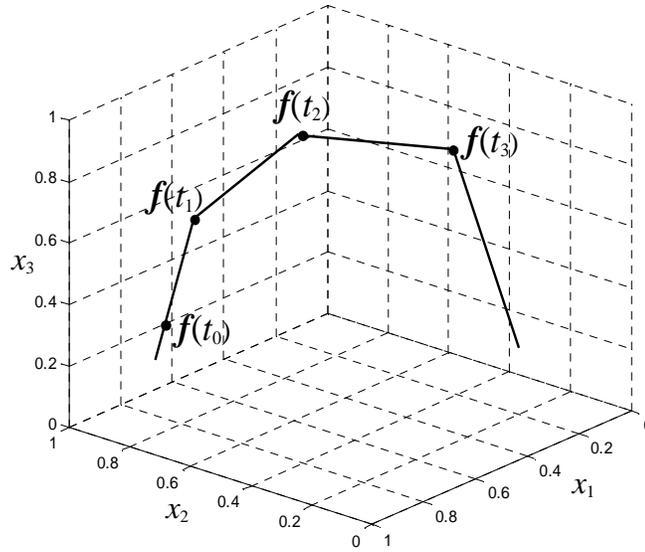
Fig. 10. A piecewise linear principal curve.

For $j = 1, 2, \ldots, p$, define $\mathbf{\Omega}_j = \{t: f(t)$ lies on the jth segment$\}$ and $c_j = \partial f(t)/\partial t\big|_{t \in \Omega_j}$, which is proportional to the direction of the jth segment. Without loss of generality,

assume that t is scaled so that $\left. \partial f(t)/\partial t \right|_{t \in \Omega_j}$ is constant over each Ω_j. For each fixed t,

$f(t)$ can be expressed as a linear combination of $\{c_j: j = 1, 2, \ldots, p\}$ added to $f(t_0)$,

where $f(t_0)$ is an arbitrary point on $f(t)$. In other words, $f(t)$ can be expressed as

$$f(t) = f(t_0) + \mathbf{C}v(t), \tag{2.8}$$

where $\mathbf{C} = [c_1\ c_2\ \ldots\ c_p]$, and $v(t) = [v_1(t)\ v_2(t)\ \ldots\ v_p(t)]^T$ for some set of p random

variables $v_j(t)$ that are functions of t alone. For example, suppose $t_0 \in \Omega_1$, and $\{t_j: j = 1,$

$2, \ldots, p-1\}$ are defined such that $f(t_j)$ is the intersection between the jth and the $(j+1)^{\text{st}}$

segments. Then for each t, $v(t) = [t_1-t_0, t_2-t_1, \ldots\ t_{k-1}-t_{k-2}, t-t_{k-1}, 0, \ldots, 0]^T$ where k is

such that $t \in \Omega_k$.

From Equation (2.8), it is clear that a piecewise linear curve $f(t)$ lies in the r-dimensional linear variety $f(t_0) + \text{span}\{c_1, c_2, \ldots, c_p\}$, where $f(t_0)$ is an arbitrary point on $f(t)$ and $r = \text{rank}\{\mathbf{C}\}$. Hence, the preceding theorem and corollary are applicable to this situation. This has important practical implications with the high dimensional manufacturing data, because any nonlinear principal curve can be approximated arbitrarily closely by a piecewise linear one. If only a small (relative to d) number of linear pieces are needed to adequately approximate a principal curve, the preceding results imply that r will be small relative to d and the level of noise reduction achieved by filtering the data will be substantial.

When $f(t)$ lies entirely in an r-dimensional linear variety, the results of PCA provide an estimate of the linear variety and its dimension. On the other hand, suppose that $f(t)$ does not lie *entirely* in any low-dimensional linear variety, but that it can be reasonably

approximated by a piecewise linear curve. If the approximation is close, then $f(t)$ will have only a small component that falls outside the linear variety in which the piecewise linear approximation lies, which may be of much lower dimension than d. In this case, the results of PCA provide an estimate of a linear variety in which $f(t)$ *approximately* lies. The methods of estimating the number of dominant eigenvalues (i.e., the dimension of the linear variety) discussed in (Apley and Shi, 2001) are attractive in that the size of an eigenvalue is measured relative to σ^2. If the size of all components of $f(t)$ that lie outside a particular linear variety is small relative to σ^2, it would typically be the case that these components could be neglected.

II.4 The PCA-filtered Principal Curve Algorithm

The HS algorithm for estimating a principal curve based on a sample of data $\{x_n: n = 1, 2, \ldots, N\}$ seeks an estimate $\hat{f}(\bullet)$ that minimizes $\sum_{n=1}^{N}\left\| x_n - \hat{f} \right\|^2$ under certain smoothness constraints. Here, $\left\| x_n - \hat{f} \right\|^2 = \inf_t \left\| x_n - \hat{f}(t) \right\|^2$. After conducting PCA on the sample covariance matrix, a set of filtered observations $x_{z,n}$ ($n = 1, 2, \ldots, N$) could be generated in analogy with Equation (2.7) via $x_{z,n} = \hat{\mathbf{P}}_z(x_n - \hat{\boldsymbol{\mu}}_x) + \hat{\boldsymbol{\mu}}_x$. Here, $\hat{\mathbf{P}}_z$ would be formed from the first $\hat{r}$ eigenvectors $\{\hat{z}_j: j = 1, 2, \ldots, \hat{r}\}$ of the sample covariance matrix, and $\hat{\boldsymbol{\mu}}_x$ is taken to be the sample average of the observations. One might consider directly applying the principal curve estimation algorithm to the filtered observations by minimizing $\sum_{n=1}^{N}\left\| x_{z,n} - \hat{f} \right\|^2$. As was discussed in Section II.3 and will

be demonstrated in Section II.6, this reduces the effects of the noise and improves the accuracy of the principal curve estimation.

Aside from the PCA step, the computational expense of this approach would be the same as if the nonlinear variation pattern identification algorithm were applied to the original, unfiltered observations. If $\hat{r}$ is much less than d, however, considerable savings in computational expense can be achieved as follows. Note that since each $x_{z,n}$ lies in the linear variety $\hat{\boldsymbol{\mu}}_x + \hat{\mathbf{M}}$, where $\hat{\mathbf{M}} = \text{span}\{ \hat{z}_1, \hat{z}_2, \ldots, \hat{z}_r \}$, minimizing $\sum_{n=1}^{N} \left\| x_{z,n} - \hat{f} \right\|^2$ must always result in an $\hat{f}(\bullet)$ that also lies in $\hat{\boldsymbol{\mu}}_x + \hat{\mathbf{M}}$. Consequently, there is no loss of generality in restricting our search for $\hat{f}(t)$ to this $\hat{r}$-dimensional linear variety. This can be accomplished by working with the $\hat{r}$-dimensional vectors of PCA scores

$$y_n = \hat{\mathbf{Z}}^{\mathrm{T}}(x_n - \hat{\boldsymbol{\mu}}_x), \tag{2.9}$$

where $\hat{\mathbf{Z}} = [\hat{z}_1 \ \hat{z}_2 \ \ldots \ \hat{z}_r]$. Note that y_n consists of the coefficients of $x_{z,n} - \hat{\boldsymbol{\mu}}_x$ using $\{ \hat{z}_j : j = 1, 2, \ldots, \hat{r} \}$ as a basis for the $\hat{r}$-dimensional subspace $\hat{\mathbf{M}}$. Similarly define $\hat{h}(t) = \hat{\mathbf{Z}}^{\mathrm{T}}(\hat{f}(t) - \hat{\boldsymbol{\mu}}_x)$ to be the coefficients of $\hat{f}(t) - \hat{\boldsymbol{\mu}}_x$ in $\hat{\mathbf{M}}$. Because we are restricting $\hat{f}(t) - \hat{\boldsymbol{\mu}}_x$ to lie in $\hat{\mathbf{M}}$, it follows that $\hat{f}(t) - \hat{\boldsymbol{\mu}}_x = \hat{\mathbf{Z}}\hat{h}(t)$, or

$$\hat{f}(t) = \hat{\mathbf{Z}}\hat{h}(t) + \hat{\boldsymbol{\mu}}_x. \tag{2.10}$$

Using Equations (2.9) and (2.10) and the definition of $x_{z,n}$, for any t we have

$$x_{z,n} - \hat{f}(t) = \hat{\mathbf{Z}}\hat{\mathbf{Z}}^{\mathrm{T}}[x_n - \hat{\boldsymbol{\mu}}_x] + \hat{\boldsymbol{\mu}}_x - [\hat{\mathbf{Z}}\hat{h}(t) + \hat{\boldsymbol{\mu}}_x] = \hat{\mathbf{Z}}[y_n - \hat{h}(t)]$$

so that $\quad \left\| \boldsymbol{x}_{z,n} - \hat{\boldsymbol{f}}(t) \right\|^2 = \left\| \hat{\mathbf{Z}}[\boldsymbol{y}_n - \hat{\boldsymbol{h}}(t)] \right\|^2 = [\boldsymbol{y}_n - \hat{\boldsymbol{h}}(t)]^{\mathrm{T}} \hat{\mathbf{Z}}^{\mathrm{T}} \hat{\mathbf{Z}}[\boldsymbol{y}_n - \hat{\boldsymbol{h}}(t)]$

$$= [\boldsymbol{y}_n - \hat{\boldsymbol{h}}(t)]^{\mathrm{T}}[\boldsymbol{y}_n - \hat{\boldsymbol{h}}(t)] = \left\| \boldsymbol{y}_n - \hat{\boldsymbol{h}}(t) \right\|^2 .$$

Therefore, choosing $\hat{\boldsymbol{f}}(\bullet)$ to minimize $\sum_{n=1}^{N} \left\| \boldsymbol{x}_{z,n} - \hat{\boldsymbol{f}} \right\|^2$ is equivalent to choosing

$\hat{\boldsymbol{h}}(\bullet)$ to minimize $\sum_{n=1}^{N} \left\| \boldsymbol{y}_n - \hat{\boldsymbol{h}} \right\|^2$ and then recovering $\hat{\boldsymbol{f}}(\bullet)$ from $\hat{\boldsymbol{h}}(\bullet)$ via Equation

(2.10).

The advantage of working with the $\hat{r}$-dimensional vectors $\boldsymbol{y}_n$, as opposed to the d-dimensional vectors $\boldsymbol{x}_{z,n}$, is that computational expense and convergence speed are substantially improved.

The PCA-filtered principal curve algorithm is summarized as follows:

1). Linear Variety Identification

Conduct PCA on the sample data and estimate the dimension of the linear variety as the number $\hat{r}$ of dominant eigenvalues.

2). PCA Filtering

Form $\hat{\mathbf{Z}}$ from the first $\hat{r}$ eigenvectors and generate $\boldsymbol{y}_n = \hat{\mathbf{Z}}^{\mathrm{T}}(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_x)$.

3). Principal Curves Estimation

Use a standard principal curve estimation algorithm (e.g., the HS algorithm) to find the curve $\hat{\boldsymbol{h}}(\bullet)$ in $\hat{r}$-dimensional space minimizing $\sum_{n=1}^{N} \left\| \boldsymbol{y}_n - \hat{\boldsymbol{h}} \right\|^2$.

4). Principal Curve Recovery

Given $\hat{\boldsymbol{h}}(\bullet)$, recover the d-dimensional principal curve $\hat{\boldsymbol{f}}(\bullet)$ using Equation (2.10).

We illustrate the variation pattern identification algorithm by continuing the autobody assembly example introduced in Section I.3. An example with much higher dimensional data is considered in the following section. After applying PCA to a sample of $N = 100$ autobodies, it was concluded that there were two dominant eigenvalues. Hence, the principal curve lies approximately in a two-dimensional subspace, which is consistent with piecewise linear appearance of the variation pattern in Fig. 4. Fig. 11(a) is a plot of all the six eigenvalues. The two corresponding eigenvectors are $\hat{z}_1 = [.25\ .53\ .65\ .36\ .29\ .14]^T$ and $\hat{z}_2 = [-.21\ -.28\ -.34\ .62\ .56\ .26]^T$. Fig. 11(b) is a scatter plot of the two PCA scores y_1 and y_2 for the 100 autobodies, along with the estimated principal curve $\hat{h}(\bullet)$ in the two-dimensional subspace.

(a) (b)
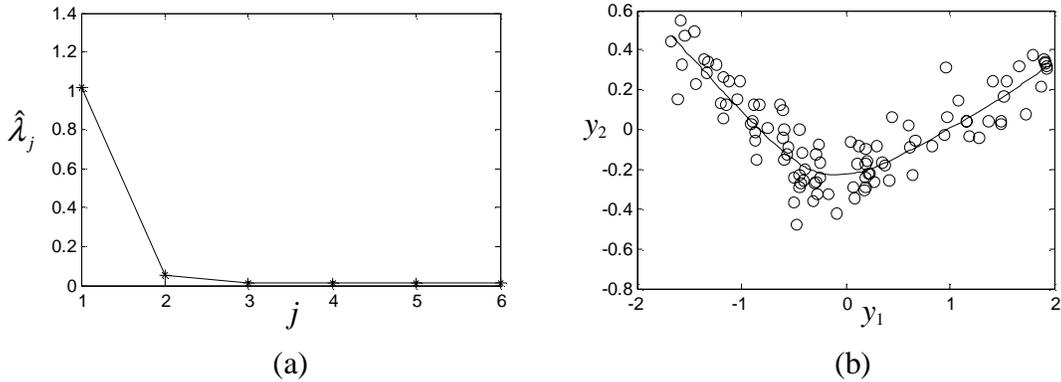
Fig. 11. Eigenvalues (a) and scatter plot of the two dominant PCA scores (b) for the autobody example. Panel (b) also shows the estimated principal curve $\hat{h}(\bullet)$ in the 2-dimensional subspace of dominant PCA scores.

In order to interpret the variation pattern, we should transform the lower dimensional principal curve $\hat{h}(\bullet)$ back to the principal curve $\hat{f}(\bullet)$ in the original 6-dimensional data space via Equation (2.10). The six elements of $\hat{f}(\bullet)$ (corresponding to the six variables $x_1, \ldots, x_6$) are plotted versus t in Fig. 12. Although the scaling of t is somewhat arbitrary, the scaling in Fig. 12 was such that t coincides with the arc length along the principal curve. Compared with the scatter plots in Fig. 4, the plots in Fig. 12 provide a clearer, less noisy, visualization of the nature of the nonlinear variation pattern.

It can be seen from Fig. 12 that only the measurement features 1, 2, and 3 on the right bodyside deviate from their nominal positions. Each deviation is proportional to the increasing t before it reaches the value of t_0. When t is greater than t_0, the pairs of corresponding measurement features on both bodysides will change simultaneously (x_3/x_4 deviate to the largest amount, x_2/x_5 deviate to the next largest, and x_1/x_6 to the least), which exhibit a linear match boxing pattern.
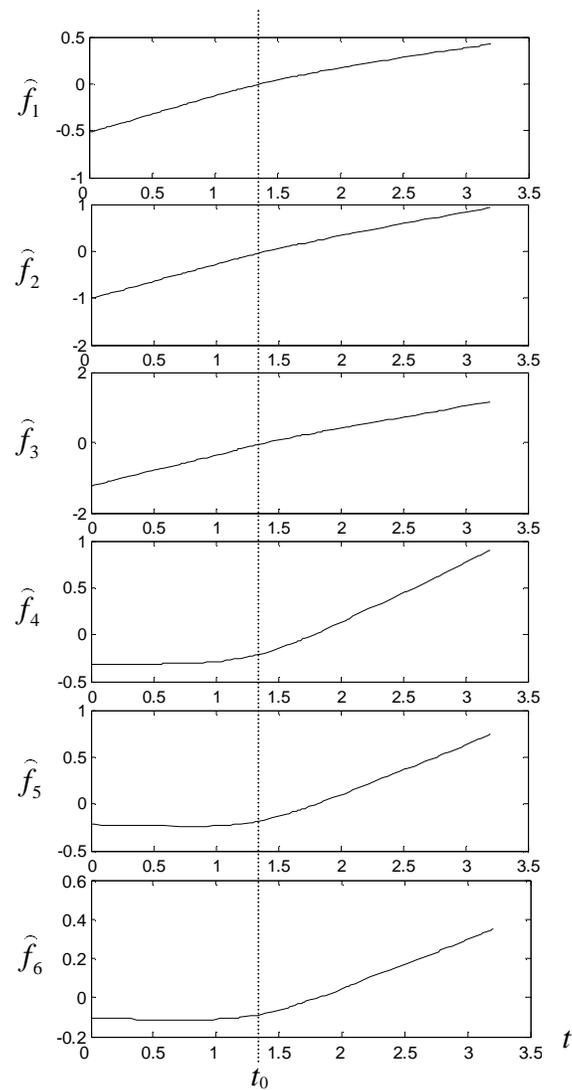
Fig. 12. Plots of the elements of $\hat{f}(\bullet)$ versus t, illustrating the characteristics of the nonlinear variation pattern.

II.5 Visualizing High Dimensional Variation Patterns

The purpose of identifying variation patterns is to serve as an aid in identifying and eliminating major root causes of manufacturing variation. In order for a process operator or engineer to effectively interpret the pattern identification results, graphical

visualization techniques are critical, especially with high dimensional data. In this section, we illustrate a method for visualizing variation patterns in high dimensional data with the example from sheet metal flanging as discussed in Section II.1, in which the random variable t is introduced in model (2.3) to represent the variation on the boundary condition. The variation pattern f is then expected to describe the changing of position of the wrinkles along the length of the flange due to varied boundary conditions.

In this example we defined $d = 1080$ points uniformly distributed over the panel. The variation pattern was simulated over a sample of $N = 70$ parts, with random noise added, and the results of PCA indicated that there were $\hat{r} = 2$ dominant eigenvalues. Similar to Fig. 11, Fig. 13 shows the first eleven eigenvalues and a scatter plots of the PCA scores corresponding to the two dominant eigenvalues, as well as the fitted principal curve $\hat{h}(\bullet)$ in two-dimensional space. The strong nonlinear (circular) pattern results from what can be viewed as phase shifting as the wrinkles change position along the flange.

After transforming $\hat{h}(\bullet)$ back to the principal curve $\hat{f}(\bullet)$ in the original 1080-dimensional space, the nature of the variation pattern could be visualized by graphically illustrating how $\hat{f}(\bullet)$ varies as t is varied. Fig. 14 shows a MATLAB$^{®}$ graphical user interface that was developed for this purpose. The plot of $\hat{f}(\bullet)$ shown in Fig. 14 changes dynamically as the user moves the slide bar control back and forth such that the position of the slide bar represents the value of t for which $\hat{f}(\bullet)$ is currently plotted.
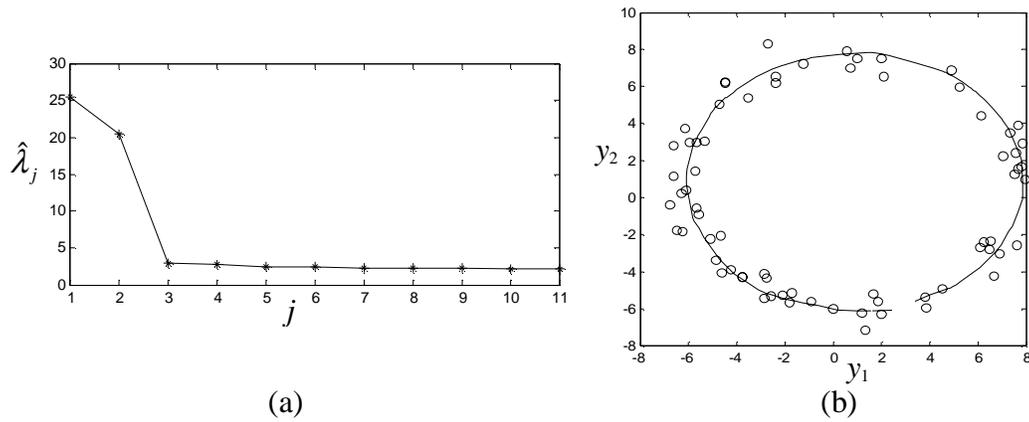
Fig. 13. The first eleven eigenvalues (a) and scatter plot of the dominant PCA scores (b) for the simulated shrink flanging example. Panel (b) also shows the estimated principal curve $\hat{h}(\bullet)$ in the 2-dimensional subspace of dominant PCA scores.
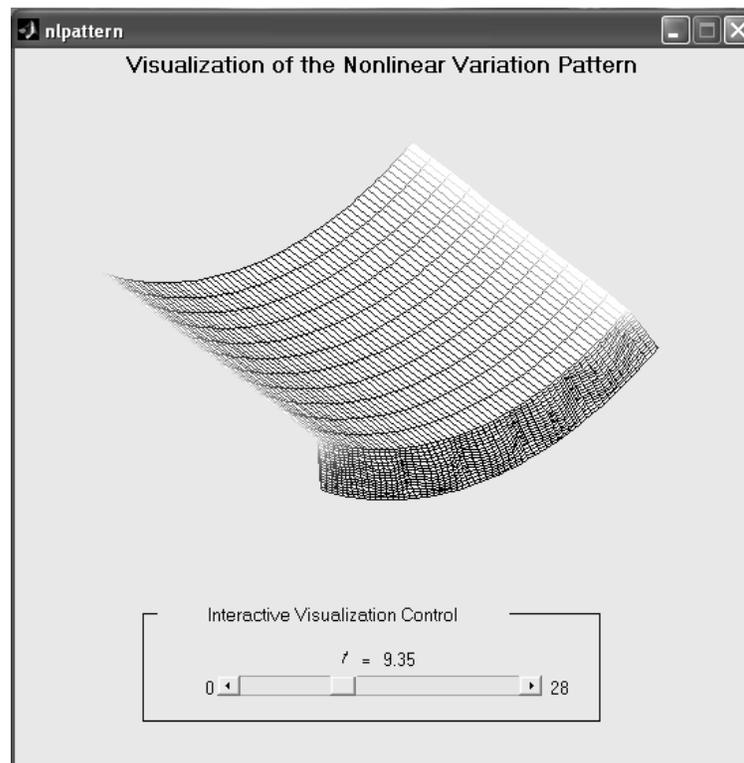
Fig. 14. Graphical user interface for interactively visualizing the nature of a nonlinear variation pattern that represents wrinkling in a shrink flanged panel.

As discussed in Section II.1, Fig. 8 shows five frames (for five different values spanning the range of t) produced from Fig. 14. Only the flange part of the panel is shown. Although it is much clearer when viewing an animation, it can still be seen from the five static frames that the variation pattern represents the wrinkles changing position from part to part. The effect of proposed principal curve algorithm for facilitating variation pattern visualization and diagnosis can be further illustrated by comparing the plots in Fig. 8 with that in Fig. 15, which is plotted from a sample observation before the noise-filtered algorithm was applied. As shown in Fig. 15, since the systematic variation pattern present in the data was distorted by the noise to a great extent, it is hard to identify and diagnose the nature of nonlinear pattern $f(\bullet)$ using the set of sample data only. In this sense, the proposed noise-filtered algorithm not only reduces the computational expense in principal curve estimation, but also improves the accuracy in nonlinear variation pattern identification and interpretation.
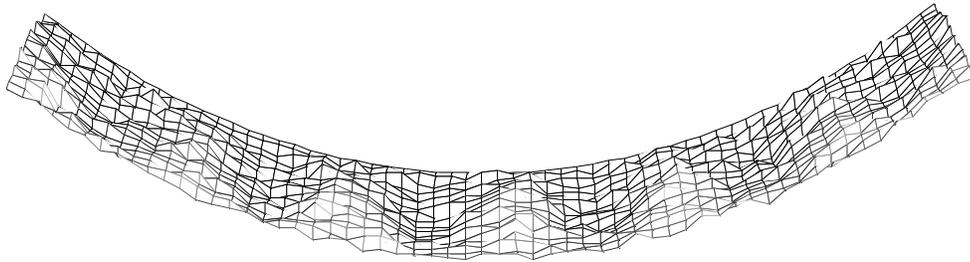
Fig. 15. Illustration of the wrinkling in the flange surface from a sample observation.

II.6 Monte Carlo Performance Comparison

In this section, Monte Carlo simulation is used to compare the performances of the standard HS algorithm and the PCA-filtered HS algorithm. In all simulations, the data were generated via model (2.2). The noise component of x was generated from a multivariate Gaussian distribution with covariance matrix $\sigma^2\,\mathbf{I}$. The principal curve component of x was generated as a uniformly distributed point along the specified principal curve $f(t)$. Denoting the principal curve estimates from the HS algorithm and the PCA-filtered HS algorithm as $\hat{f}_{HS}(\bullet)$ and $\hat{f}_{HS\text{-}PCA}(\bullet)$, respectively, consider the mean squared distances

$$J_{HS} = \int \left\| \hat{f}_{HS} - f(t) \right\|^2 g(t)dt \;\text{ and} \tag{2.11}$$

$$J_{HS\text{-}PCA} = \int \left\| \hat{f}_{HS\text{-}PCA} - f(t) \right\|^2 g(t)dt \tag{2.12}$$

as measures of closeness between the estimated and true principal curves. Here, $g(t)$ denotes the probability density of t, which was uniform in all cases. The variable t was scaled so that $\|\partial f(t)/\partial t\|$ was constant along the entire length of the curve. In each of the Monte Carlo simulations for the examples discussed below, J_{HS} and $J_{HS\text{-}PCA}$ were averaged over 10,000 Monte Carlo replicates to compare the accuracy of the two principal curve estimation methods.

A quadratic principal curve lying in a two-dimensional linear variety of $\mathfrak{R}^d$ was used for all simulations. The principal curve was constructed by first generating the curve $x_2 = x_1^2$ for $x_1 \in [-1, 1]$, which is illustrated in Fig. 16 for $d = 3$. The curve was

then multiplied by a Householder mirror reflection matrix in d-dimensional space to transform it without changing its shape or size so that it still fell in a linear variety of dimension $r = 2$.
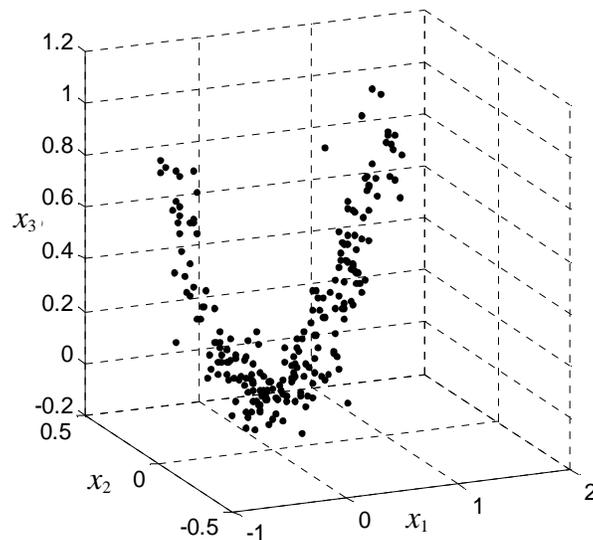
Fig. 16. Illustration of data along a quadratic principle curve with $d = 3$, $r = 2$ and $\sigma = 0.1$.

Table 1 compares the performance of the HS algorithm and the PCA-filtered HS algorithm for various r, d, and σ, with a sample size of $N = 200$. Note that the value of σ in Fig. 16 corresponds to the midrange value of 0.1 in Table 1. The extent to which PCA filtering improves the accuracy (as measured by $J_{HS}/J_{HS\text{-}PCA}$) varied between 1.48 and 6.96, with larger relative improvement for larger d (higher dimensional data) and/or large σ (larger noise-to-signal ratio). Simulations were also run for $N = 100$ and $N =$

400. The individual performance measures J_{HS} and $J_{HS\text{-}PCA}$ were roughly inversely proportional to N, and their ratio was virtually the same for all values of N.

Table 1. Performance comparison of the HS and PCA-filtered HS algorithms for quadratic principal curves ($N = 200$).

d	σ	J_{HS}	$J_{HS\text{-}PCA}$	$J_{HS}/J_{HS\text{-}PCA}$
8	0.05	0.0031	0.0021	1.48
8	0.1	0.021	0.0089	2.36
8	0.2	0.43	0.16	2.69
16	0.05	0.0034	0.0019	1.79
16	0.1	0.056	0.021	2.67
16	0.2	0.63	0.19	3.32
64	0.05	0.0065	0.0025	2.60
64	0.1	0.24	0.059	4.07
64	0.2	3.56	0.68	5.24
100	0.05	0.022	0.0052	4.23
100	0.1	0.56	0.098	5.71
100	0.2	5.64	0.81	6.96

II.7 Chapter Summary

This chapter develops a methodology for representing and blindly identifying nonlinear variation patterns in high dimensional measurement data. The nonlinear variation pattern model is an extension of previous linear models that provides more accurate representation of complex nonlinear phenomena. To overcome computational and accuracy problems caused by high dimensionality, thereby making the approach more suitable for large-scale manufacturing databases, a pre-filtering step for enhancing principal curve estimation was developed. The Monte Carlo simulation

results demonstrate that the pre-filtering, which involves the use of linear PCA, can substantially improve the pattern estimation.

This chapter also describes the use of graphical visualization methods for aiding in interpreting the blindly identified variation patterns. This allows potentially valuable information to be extracted from large volumes of recorded measurement data. Together, the pattern identification and visualization techniques provide an effective tool for diagnosing major root causes of manufacturing variation.

CHAPTER III

NONLINEAR VARIATION PATTERN IDENTIFICATION BY A POLYGONAL LINE ALGORITHM

III.1 Introduction

Chapter II discusses the applicability of principal curve method for single nonlinear variation pattern modeling and identification. When principal curves lie in lower dimensional linear varieties, the noise-filtered algorithm proposed in Section II.4 not only improves the accuracy of pattern identification for root cause diagnostics, but overcomes the problem of dimensionality in high dimensional manufacturing application. As the nonlinear generalization of principal component analysis, a principal curve provides a summarization of the data in terms of a one-dimensional manifold non-linearly embedded in the high data space, which serves as a nonlinear pattern identification and dimension reduction technique.

Several principal curve based nonlinear methods have been proposed subsequent to Hastie and Stuetzle's original work, many of which provided various algorithms thereupon owing to the ambiguity in the formulation of data model (Delicado, 2001). Banfield and Raftery (1992) modified the HS algorithm to reduce its estimation bias when estimating closed curves and used the results to identify ice floe outlines from satellite images. Another definition of principal curve with constrained length was proposed in (Kégl *et al.*, 1999; Verbeek *et al.*, 2001) to estimate principal curves that

minimizes the average squared distance between the curve and the observations. Tibshirani (1992) defines principal curves as curves minimizing a penalized log-likelihood measure, where the conditional distribution of x on projection index t is a spherical Gaussian.

Alternative methods have been given to tackle this and closely related problems. Mulier and Cherkassky (1995) applied self-organizing maps (SOM) to identify nonlinear patterns by incorporating some prototype-vectors. Finite mixture models were proposed to learn complex nonlinear data structure as a combination of "local models" (Hinton *et al.*, 1995; Tipping and Bishop, 1999).

Several problems exist with the aforementioned methods. When the sample data are concentrated around a complex curve or the model assumptions are not satisfied, these methods exhibit poor performance. This is due to the fixed model structures or to bad initialization and smoothness constraint (Tibshirani, 1992; Verbeek *et al.*, 2001). Also, often one does not know a priori how many "local models" one needs and one has to make a guess for the number of local models.

In this chapter, we propose a polygonal line based principal curve estimation algorithm for nonlinear variation pattern identification, where the nonlinearities among the multivariate measurement data can be summarized by a combination of local linear models. Recognizing the limitations of existing principal curve approaches, this chapter integrates the linear principal component analysis with the polygonal line algorithm to represent complicated nonlinear patterns. Such nonlinear variation pattern identification method for multivariate data is capable of recognizing the underlying principal curves

without loss of information. The experimental results in Section III.5 will demonstrate that the polygonal line algorithm can produce robust and accurate estimation for a variety of nonlinear patterns.

The complexity of principal curve estimation is another implementation factor when the size of sample is large. For N data points, the HS algorithm actually produces polygonal lines with N segments and the computational complexity is mostly determined by the sample size. In this chapter, we take a new approach by estimating principal curves as continuous curves which minimize the average squared distance between the curve and data points. This learning scheme chooses a curve from a class of polygonal lines with K segments by minimizing the distance measure, the computational complexity of which depends on K. The number of segments, K, when optimally chosen, is usually much less than N. Therefore, the proposed polygonal line algorithm can be implemented with reduced computational expense.

Piecewise linear principal curves have been investigated in Chapter II in that any nonlinear principal curve can be approximated arbitrarily closely by a polygonal line (piecewise linear curve) with finite line segments. The presumption throughout this chapter is that the proposed polygonal line algorithm will be applied to the PCA scores $\boldsymbol{y}$ projected in the r-dimensional space, where r is the dimension of the lower linear varieties in which the original d-dimensional variation pattern or its approximated piecewise linear curve lies. The identified polygonal line in $\Re^r$ is then transformed back to the original data space, yielding the estimate of the d-dimensional nonlinear pattern for variation pattern diagnosing use, as described in Section II.5.

The theoretical results in Chapter II will be utilized to construct a polygonal line by determining the number of line segments. As a compromise between linear and nonlinear pattern identification methods, the polygonal lines enjoy the simplicity and analyzability of linear models while remaining robust enough to model various nonlinear multivariate data.

III.2 Identifying Principal Curves with Polygonal Lines

A curve in r-dimensional Euclidean space is a continuous function $\boldsymbol{h}: \boldsymbol{I} \to \mathfrak{R}^r$, where $\boldsymbol{I}$ is a closed interval on the real line. Denote the expected squared distance between $\boldsymbol{y}$ and $\boldsymbol{h}$ by

$$\Delta(\boldsymbol{h}) = E[\inf_v \|\boldsymbol{y} - \boldsymbol{h}(v)\|^2] = E\|\boldsymbol{y} - \boldsymbol{h}(t_h(\boldsymbol{y}))\|^2, \tag{3.1}$$

where the projection index $t_h(\boldsymbol{y})$ is defined in Equation (2.4). Let $\boldsymbol{h}$ be a smooth (infinitely differentiable) curve and, for $\lambda \in \mathfrak{R}$, consider the perturbation $\boldsymbol{h} + \lambda\boldsymbol{g}$ of $\boldsymbol{h}$ by a smooth curve $\boldsymbol{g}$ such that $\sup_v \|\boldsymbol{g}(v)\| \le 1$ and $\sup_v \|\boldsymbol{g}'(v)\| \le 1$. HS proved that when $\boldsymbol{h}$ is a principal curve, it is a critical point of the distance function in the sense that, for all such $\boldsymbol{g}$,

$$\left.\frac{\partial\Delta(\boldsymbol{h} + \lambda\boldsymbol{g})}{\partial\lambda}\right|_{\lambda=0} = 0. \tag{3.2}$$

An analogous result holds for linear principal component if the perturbation $\boldsymbol{g}$ is a straight line. In this sense, the HS principal curve definition is a natural generalization

of principal components. Also, it is easy to check that principal components coincide with principal curves if the distribution of y is elliptical (Hastie and Stuetzle, 1989).

III.2.1 Piecewise Linear Curves

A straight line $s(t)$ is the first principal component, if and only if,

$$E[\|y - s\|^2] \leq E[\|y - \tilde{s}\|^2] \tag{3.3}$$

for any other straight line $\tilde{s}(t)$. We would like to generalize the property (3.3) of the first principal component and define principal curves so that they minimize the expected squared distance over a class of curves, rather than only being critical points of the distance function (3.1). To do this, it is necessary to consider the smoothness constraint on the curve since, otherwise, for any y with a density and any $\varepsilon > 0$, there exists a smooth curve h such that $\Delta(h) \leq \varepsilon$ and, thus, a minimizing h with respect to distance (3.1) has infinite length and usually produces over-fitting results. This problem can be resolved by restricting the curve as a piecewise linear curve with finite length that is determined by the number of line segments. On the other hand, if the distribution of y is concentrated on a polygonal line and is uniform there, the infimum of the squared distances $\Delta(h)$ is 0 over the class of smooth curves, but no smooth curve can achieve this infimum. For this reason, we relax the requirement that h should be differentiable and, instead, we define it to be a polygonal line. Note that, by the polygonal line definition of principal curves, h is still continuous. We give the following definition of polygonal line as the principal curve approximation:

Definition 3.1 *A curve **h** is called a principal curve with K line segments for **y** if **h** minimizes Δ(**h**) over all curves in the class of polygonal lines with K segments.*

A useful advantage of the new definition is that the principal curves with K line segments always exist, as the simulation results in Section III.5 demonstrate. Note that neither the HS nor our definition guarantees the uniqueness of principal curves. In our case, there might exist several principal curves for a given segment number, but each of these will have the same (minimal) square error.

In what follows, we consider the problem of polygonal line estimation based on a set of sample data. Suppose that N data points $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, …, $\boldsymbol{y}_N$ are drawn independently from the distribution of $\boldsymbol{y}$. The goal is to use the sample data to construct a curve with line segments at most K whose average squared loss is close to that of the unknown principal curve for $\boldsymbol{y}$. When the number of line segments is chosen appropriately, the estimated polygonal line can identify the nonlinear pattern with desired accuracy.

Define

$$\Delta(\boldsymbol{y}_n, \boldsymbol{h}) = \min_t \left\| \boldsymbol{y}_n - \boldsymbol{h}(t) \right\|^2 \tag{3.4}$$

as the squared distance between a point $\boldsymbol{y}_n \in \mathfrak{R}^r$ and the curve $\boldsymbol{h}$. For any $\boldsymbol{h} \in S_k$, where S_k is the class of piecewise linear curves with k line segments, the empirical squared error of $\boldsymbol{h}$ is the sample average of Equation (3.4), i.e.

$$J_h = \frac{1}{N} \sum_{n=1}^{N} J_{n,h} = \frac{1}{N} \sum_{n=1}^{N} \Delta(\boldsymbol{y}_n, \boldsymbol{h}), \tag{3.5}$$

where we have suppressed the dependence of J_h on the sample data.

For a k-segment polygonal line, $\boldsymbol{h}$ consists of $k+1$ vertices $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{k+1}\}$. Denote

$$\boldsymbol{h}_i(t) = \boldsymbol{v}_i + \frac{t - t_i}{t_{i+1} - t_i}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)$$ as the ith line segment for $t \in [t_i, t_{i+1}] \subset \Re$ where t_i is the

projection index of $\boldsymbol{v}_i$ along $\boldsymbol{h}$ ($i = 1, 2, \ldots, k$). The way the distance of point $\boldsymbol{y}$ and the

line segment $\boldsymbol{h}_i$ is measured depends on the value of $t_{h_i}(\boldsymbol{y})$. If $t_{h_i}(\boldsymbol{y}) = t_i$ or

$t_{h_i}(\boldsymbol{y}) = t_{i+1}$, the distance is measured as the distance of $\boldsymbol{y}$ and one of the vertices $\boldsymbol{v}_i$ or

$\boldsymbol{v}_{i+1}$ respectively. If $\boldsymbol{y}$ projects to $\boldsymbol{h}_i$ between its two vertices, the distance is measured as

if $\boldsymbol{h}_i$ were a line, as shown in Fig. 17, and the squared distance is then

$\left\| \boldsymbol{y} - \boldsymbol{v}_i \right\|^2 - ((\boldsymbol{y} - \boldsymbol{v}_i)^{\mathrm{T}} \boldsymbol{c}_i)^2$ where $\boldsymbol{c}_i = (\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)/(t_{i+1} - t_i)$. Formally,

$$\Delta(\boldsymbol{y}, \boldsymbol{h}_i) = \begin{cases} \left\| \boldsymbol{y} - \boldsymbol{v}_i \right\|^2 & t_{h_i}(\boldsymbol{y}) = t_i, \\ \left\| \boldsymbol{y} - \boldsymbol{v}_{i+1} \right\|^2 & t_{h_i}(\boldsymbol{y}) = t_{i+1}, \\ \left\| \boldsymbol{y} - \boldsymbol{v}_i \right\|^2 - ((\boldsymbol{y} - \boldsymbol{v}_i)^{\mathrm{T}} \boldsymbol{c}_i)^2 & \text{otherwise.} \end{cases} \qquad (3.6)$$

Given the definition (3.6), the distance in Equation (3.5) is $\Delta(\boldsymbol{y}_n, \boldsymbol{h}) = \min_i\{\Delta(\boldsymbol{y}_n, \boldsymbol{h}_i)\}$.
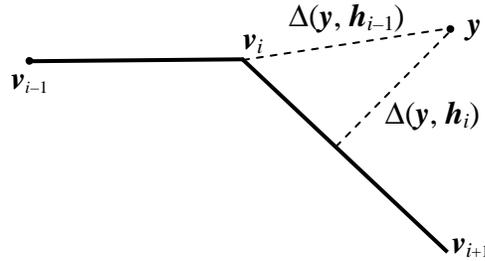
Fig. 17. Distance between a point and a line segment in two different geometric cases.

Given the number of line segments, the polygonal line learning algorithm will choose an **h** that minimizes the empirical distance (3.5).

III.2.2 Piecewise Linear Basis Functions

As discussed in Section II.3, the k-segment polygonal line **h**(t) as the principal curve approximation, can be written as **h**(t_0) plus a linear combination of $\{$**c**$_j$: $j = 1, 2,$ …, $k\}$, where **h**(t_0) is an arbitrary point on **h**(t) and **c**$_j$ is proportional to the direction of the jth segment.

Following the expression of Equation (2.8), a polygonal line can be readily written by the piecewise linear basis function representation (Hastie *et al.*, 2001):

$$\boldsymbol{h}(t) = \mathbf{W}\boldsymbol{b}_{\mathrm{w}}(t) \tag{3.7}$$

where $\mathbf{W} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k+1}]$ is an $r \times (k+1)$ weight matrix, and $\boldsymbol{b}_{\mathrm{w}}(t) = [b_{\mathrm{w},1}(t), \ldots, b_{\mathrm{w},k+1}(t)]^{\mathrm{T}}$ are the basis functions, defined by

$$b_{\mathrm{w},i}(t) = \begin{cases} (t - t_{i-1})/(t_i - t_{i-1}), & t_{i-1} < t \le t_i \\ (t_{i+1} - t)/(t_{i+1} - t_i), & t_i < t \le t_{i+1} \\ 0, & \textit{otherwise} \end{cases} \tag{3.8}$$

where $t \in \boldsymbol{I} = [t_1, t_{k+1}]$ being the so-called projection index along the polygonal line, and $t_1 < t_2 < \ldots < t_k < t_{k+1}$ are the projection indices of $\boldsymbol{v}_i$ ($i = 1, 2, \ldots, k+1$). As an example, in Fig. 18 a set of piecewise linear basis functions is plotted for illustration, when the polygonal line has $k = 2$ pieces. Thus, the piecewise linear basis functions (3.7) provide a parametric formulation for polygonal lines.
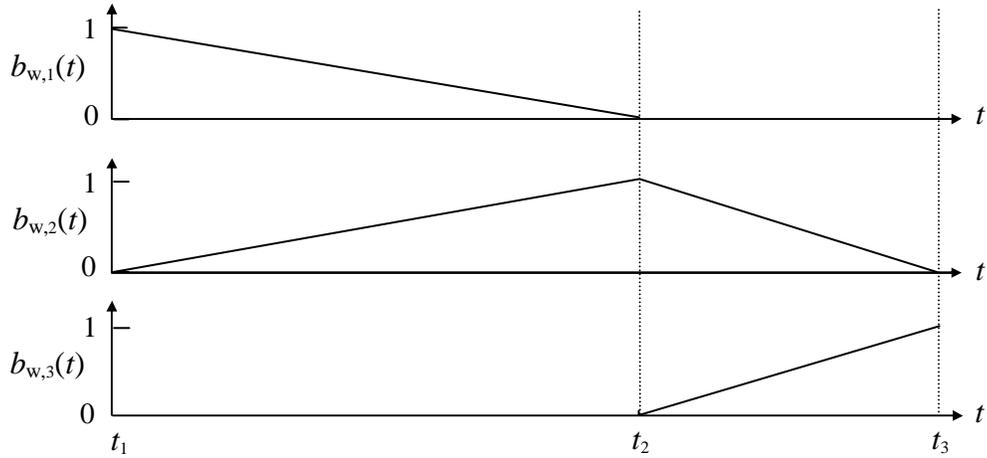
Fig. 18. Piecewise linear basis functions for $k = 2$.

Note that Definition (3.8) implies that the basis functions are non-negative and form a partition of unity

$$\sum_{i=1}^{k+1} b_{w,i}(t) = 1$$

for all t. The latter property obviates the need for an additional representation of a shift (bias) vector since

$$\mathbf{W}\boldsymbol{b}_w(t) + \boldsymbol{\mu}_0 = \mathbf{W}\boldsymbol{b}_w(t) + [\boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_0]\boldsymbol{b}_w(t) = \tilde{\mathbf{W}}\boldsymbol{b}_w(t).$$

Piecewise linear principal curves have received considerable attention in the literature in that any principal curve can be approximated arbitrarily closely by a piecewise linear one. If only a small number of line segments are needed to adequately represent the underlying principal curve, a substantial reduction on computational expenses is achieved. In addition, the curves can be parameterized in terms of arc length when the vertices are ordered along the polygonal line.

III.3 The Polygonal Line Algorithm

In this section we will introduce a local PCA based algorithm for approximate principal curves using polygonal lines. The accuracy of approximation depends on some factors such as the number of line segments and sample size. As will be discussed in Section III.3.3, each line segment defines an associated local partition of the data space such that the sample generated from the underlying principal curve is partitioned into distinct local regions. Each local partition can be represented by the first eigenvector if the number of dominant eigvenvalues is equal to one; otherwise, the data within this local partition involve nonlinearities and need a combination of multiple line segments rather than a straight line to approximate. For the latter case, we need to add new line segments to the current estimate of polygonal line and update all the vertices accordingly. The evolution of the polygonal line produced by the algorithm is illustrated in Fig. 19.

Similar to other nonlinear pattern identification methods (Kégl *et al*., 1999; Mulier and Cherkassky, 1995; Kambhatla and Leen, 1997), the proposed algorithm involves a two-step procedure: partition of the sample space and linear principal components analysis for each local partition. Iteratively, the proposed algorithm pieces together the local partitions by line segments and updates the optimal coordinates of their corresponding vertices by a global search method, where we take the Euclidean distance (3.5) as the objective function.
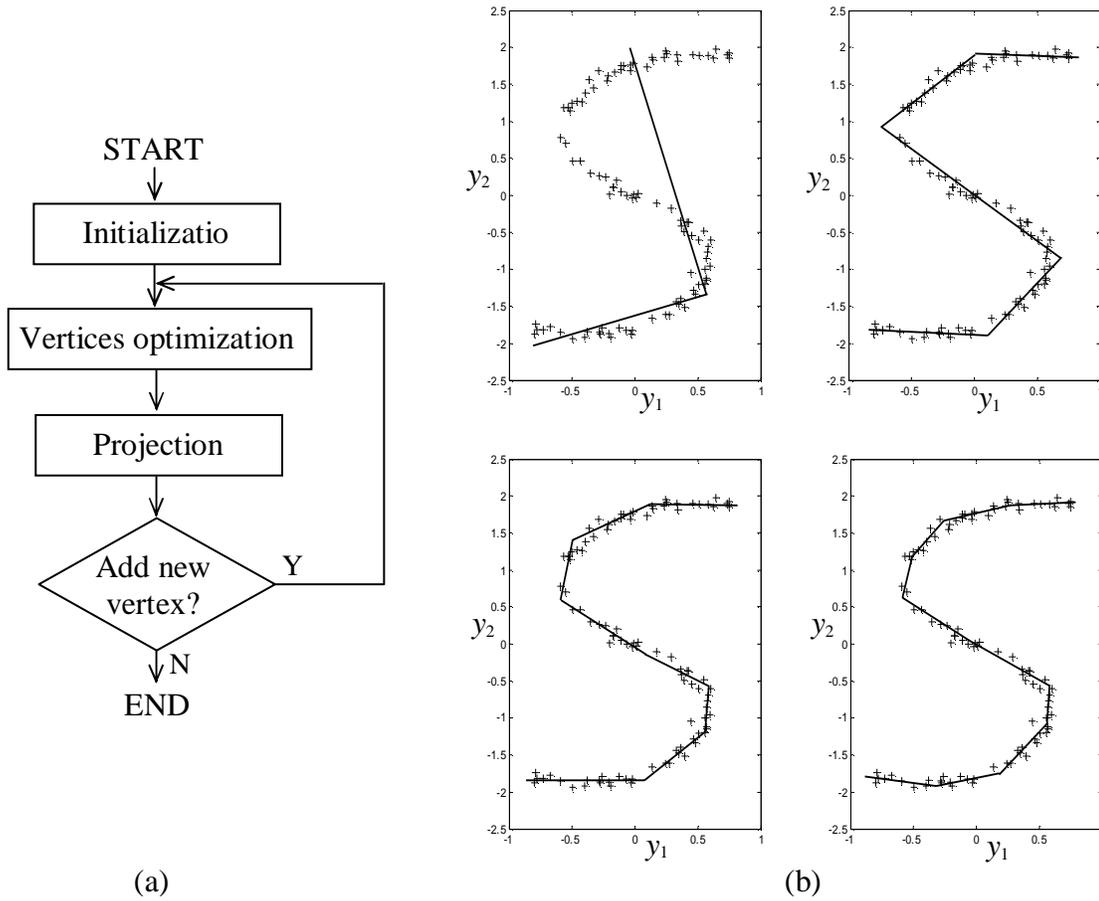
Fig. 19. (a) The flow chart of the polygonal line algorithm; (b) An example for illustration for an S-shaped curve that was produced by the algorithm with $K = 2, 5, 8$, and 10 for $N = 100$ data points with $\sigma = 0.05$.

As discussed in Chapter II, for the high dimensional data x in $\Re^d$, we first project them onto a lower dimensional linear subspace spanned by the first r dominant eigenvectors. Then the polygonal line algorithm is applied to the PCA scores y. After obtaining the optimal vertices of polygonal line h in $\Re^r$, we will transform h that corresponds to y back to the d-dimensional space via Equation (2.10). Thus, the nonlinear pattern present in the original data is identified by a polygonal line f in $\Re^d$.

III.3.1 Initialization of the Polygonal Line Algorithm

The theoretical results in Chapter II imply that a k-segment polygonal line lies in an r-dimensional linear variety with $k \geq r$. Therefore, after identifying this linear variety and projecting x onto it, we can always initialize the polygonal line for transformed PCA scores y with r line segments, followed by an optimal search for these $r+1$ vertices.

An intuitive and simple way to achieve the initialization is to project all the PCA scores y_n onto the first eigenvector $\hat{z}_1$ of $\hat{\Sigma}_x$, then start from one end point of the projection along the eigenvector and set the location of $r+1$ points such that the number of projected data within two consecutive points approximately equal to N/r. Now these $r+1$ points on $\hat{z}_1$ are taken as the initialized vertices for the polygonal line algorithm. Or, we can update these vertices by the projection-regression method introduced in Section III.3.4 to provide a better initialization.

III.3.2 Vertices Optimization

Given a data set $\{y_1, y_2, \ldots, y_N\}$, the goal of the vertices optimization step is to update $k+1$ vertices by minimizing the average squared distance function (3.5) subject to the constraint that the piecewise curve has k segments, where k is the number of current line segments.

The vertices optimization step first calculates the gradients of the objective function with respect to the $k+1$ vertices $\mathbf{W} = [v_1 \ v_2 \ \ldots \ v_{k+1}]$. Let y_n be an arbitrary sample data, and h_i be the closest line segment to y_n ($i = 1, 2, \ldots, k; \ n = 1, 2, \ldots, N$). For denotation simplicity, let $Vec(\mathbf{W})$ be an $r(k+1)$ vector containing the stacked columns of $\mathbf{W}$:

$$Vec(\mathbf{W}) = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{k+1} \end{bmatrix}.$$

For each $J_{n,h}$ in Equation (3.5), the gradient with respect to the unknown $k+1$ vertices, $dJ_{n,h} / dVec(\mathbf{W})$, is derived based on two geometric cases as shown in Fig. 17.

Case I: The closest point on line segment $\boldsymbol{h}_i$ to $\boldsymbol{y}_n$ is one of its vertices, denoted by $\boldsymbol{v}_i$, then

$$J_{n,h} = (\boldsymbol{v}_i - \boldsymbol{y}_n)^{\mathrm{T}} (\boldsymbol{v}_i - \boldsymbol{y}_n), \text{ and}$$

$$\frac{dJ_{n,h}}{dVec(\mathbf{W})} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ 2(\boldsymbol{v}_i - \boldsymbol{y}_n) \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}.$$

Case II: The closest point on line segment $\boldsymbol{h}_i$ to $\boldsymbol{y}_n$ is one other than the vertices $\boldsymbol{v}_i$ and $\boldsymbol{v}_{i+1}$, then

$$J_{n,h} = (\boldsymbol{v}_i - \boldsymbol{y}_n)^{\mathrm{T}} (\boldsymbol{v}_i - \boldsymbol{y}_n) - \frac{[(\boldsymbol{v}_i - \boldsymbol{y}_n)^{\mathrm{T}} (\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)]^2}{(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}} (\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}.$$

And the corresponding gradient vector is

$$\frac{dJ_{n,h}}{dVec(\mathbf{W})} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \partial J_{n,h} / \partial \boldsymbol{v}_i \\ \partial J_{n,h} / \partial \boldsymbol{v}_{i+1} \\ \vdots \\ \mathbf{0} \end{bmatrix},$$

where

$$\frac{\partial J_{n,h}}{\partial \boldsymbol{v}_i} = 2(\boldsymbol{v}_i - \boldsymbol{y}_n) - \frac{2(\boldsymbol{v}_i - \boldsymbol{y}_n)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}{(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}(\boldsymbol{v}_{i+1} + \boldsymbol{y}_n - 2\boldsymbol{v}_i)$$

$$- \frac{2[(\boldsymbol{v}_i - \boldsymbol{y}_n)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)]^2}{[(\boldsymbol{v}_i - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)]^2}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i), \text{ and} \qquad (3.9)$$

$$\frac{\partial J_{n,h}}{\partial \boldsymbol{v}_{i+1}} = \frac{2[(\boldsymbol{v}_i - \boldsymbol{y}_n)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)]^2}{[(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)]^2}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i) - \frac{2(\boldsymbol{v}_i - \boldsymbol{y}_n)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}{(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}(\boldsymbol{v}_i - \boldsymbol{y}_n). \quad (3.10)$$

Calculating the gradients for all N sample data, we have the gradient vector of

objective function J_h with respect to $Vec(\mathbf{W})$ as $\dfrac{dJ_h}{dVec(\mathbf{W})} = \dfrac{1}{N}\sum_{n=1}^{N}\dfrac{dJ_{n,h}}{dVec(\mathbf{W})}$, and the

optimization methods can apply to the search for the $k+1$ vertices.

Recall that the method of steepest descent method uses only the gradients in selecting a suitable search direction. This strategy is not always the most effective in the sense of convergence. If higher derivatives (e.g., Hessian) are incorporated, the resulting iterative optimization algorithm usually performs better than the gradient methods (Avriel, 2003; Bertsekas, 1995). Denote the Hessian matrix for $J_{N,h}$ as

$\mathbf{H} = \dfrac{1}{N}\sum_{n=1}^{N}\mathbf{H}_n$ where $\mathbf{H}_n$ is the Hessian with respect to $J_{n,h}$. Define

$$g_{n,i} = \frac{(\boldsymbol{v}_i - \boldsymbol{y}_n)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}{(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}.$$

The notation $g(\boldsymbol{v}_i, \boldsymbol{v}_{i+1})$ simplifies the gradients (3.9) and (3.10) to

$$\frac{\partial J_{n,h}}{\partial \boldsymbol{v}_i} = 2(\boldsymbol{v}_i - \boldsymbol{y}_n) - 2g_{n,i}(\boldsymbol{v}_{i+1} + \boldsymbol{y}_n - 2\boldsymbol{v}_i) - 2g_{n,i}^2(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i), \text{ and}$$

$$\frac{\partial J_{n,h}}{\partial \boldsymbol{v}_{i+1}} = 2g_{n,i}^2(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i) - 2g_{n,i}(\boldsymbol{v}_i - \boldsymbol{y}_n).$$

The Hessian matrix $\mathbf{H}_n$ is

$$\mathbf{H}_n = \frac{d^2 J_{n,h}}{dVec(\mathbf{W})dVec(\mathbf{W})^{\mathrm{T}}} = \begin{bmatrix} \dfrac{\partial(\dfrac{\partial J_{n,h}}{\partial \boldsymbol{v}_1})}{\partial \boldsymbol{v}_1^{\mathrm{T}}} & \cdots & \dfrac{\partial(\dfrac{\partial J_{n,h}}{\partial \boldsymbol{v}_1})}{\partial \boldsymbol{v}_{k+1}^{\mathrm{T}}} \\ \vdots & \vdots & \vdots \\ \dfrac{\partial(\dfrac{\partial J_{n,h}}{\partial \boldsymbol{v}_{k+1}})}{\partial \boldsymbol{v}_1^{\mathrm{T}}} & \cdots & \dfrac{\partial(\dfrac{\partial J_{n,h}}{\partial \boldsymbol{v}_{k+1}})}{\partial \boldsymbol{v}_{k+1}^{\mathrm{T}}} \end{bmatrix}$$

As the gradients calculation, the second derivatives of $J_{n,h}$ with respect to vertices ($\boldsymbol{v}_1$, $\boldsymbol{v}_2$, …, $\boldsymbol{v}_{k+1}$) also depend on the two geometric cases.

For case I, it is straightforward to calculate the Hessian matrix for $J_{n,h}$ as a diagonal matrix, i.e., $\mathbf{H}_n = Diag\{\mathbf{0}_{d\times d}, …, 2\mathbf{I}_{d\times d}, …, \mathbf{0}_{d\times d}\}$.

For case II, the Hessian matrix is of the following structure:

$$\mathbf{H}_n = \begin{bmatrix} \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \cdots & \cdots & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ \mathbf{0}_{d\times d} & & \dfrac{\partial(\dfrac{\partial J_{n,h}}{\partial \boldsymbol{v}_i})}{\partial \boldsymbol{v}_i^{\mathrm{T}}} & \dfrac{\partial(\dfrac{\partial J_{n,h}}{\partial \boldsymbol{v}_i})}{\partial \boldsymbol{v}_{i+1}^{\mathrm{T}}} & & \mathbf{0}_{d\times d} \\ & & \dfrac{\partial(\dfrac{\partial J_{n,h}}{\partial \boldsymbol{v}_{i+1}})}{\partial \boldsymbol{v}_i^{\mathrm{T}}} & \dfrac{\partial(\dfrac{\partial J_{n,h}}{\partial \boldsymbol{v}_{i+1}})}{\partial \boldsymbol{v}_{i+1}^{\mathrm{T}}} & & \\ \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots \\ \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \cdots & \cdots & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \end{bmatrix}, \tag{3.11}$$

where the four non-zero $d\times d$ sub-matrices in Equation (3.11) are

$$\frac{\partial(\frac{\partial J_{n,h}}{\partial \boldsymbol{v}_i})}{\partial \boldsymbol{v}_i^{\mathrm{T}}} = (2 + 4g_{n,i} + 2g_{n,i}^2)\mathbf{I} - \frac{2(\boldsymbol{v}_{i+1} + \boldsymbol{y}_n - 2\boldsymbol{v}_i)(\boldsymbol{v}_{i+1} + \boldsymbol{y}_n - 2\boldsymbol{v}_i)^{\mathrm{T}}}{(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}$$

$$- \frac{4g_{n,i}}{(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}[(\boldsymbol{v}_{i+1} + \boldsymbol{y}_n - 2\boldsymbol{v}_i)(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}} + (\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)(\boldsymbol{v}_{i+1} + \boldsymbol{y}_n - 2\boldsymbol{v}_i)^{\mathrm{T}}]$$

$$+ \frac{8g_{n,i}^2}{(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}};$$

$$\frac{\partial(\frac{\partial J_{n,h}}{\partial \boldsymbol{v}_i})}{\partial \boldsymbol{v}_{i+1}^{\mathrm{T}}} = (-2g_{n,i} - 2g_{n,i}^2)\mathbf{I} - \frac{2(\boldsymbol{v}_{i+1} + \boldsymbol{y}_n - 2\boldsymbol{v}_i)(\boldsymbol{v}_i - \boldsymbol{y}_n)^{\mathrm{T}}}{(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}$$

$$- \frac{4g_{n,i}}{(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}[(\boldsymbol{v}_{i+1} + \boldsymbol{y}_n - 2\boldsymbol{v}_i)(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}} - (\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)(\boldsymbol{v}_i - \boldsymbol{y}_n)^{\mathrm{T}}]$$

$$+ \frac{8g_{n,i}^2}{(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}};$$

$$\frac{\partial(\frac{\partial J_{n,h}}{\partial \boldsymbol{v}_{i+1}})}{\partial \boldsymbol{v}_i^{\mathrm{T}}} = \left[\frac{\partial(\frac{\partial J_{n,h}}{\partial \boldsymbol{v}_i})}{\partial \boldsymbol{v}_{i+1}^{\mathrm{T}}}\right]^{\mathrm{T}};$$

$$\frac{\partial(\frac{\partial J_{n,h}}{\partial \boldsymbol{v}_{i+1}})}{\partial \boldsymbol{v}_{i+1}^{\mathrm{T}}} = 2g_{n,i}^2\mathbf{I} - \frac{2(\boldsymbol{v}_i - \boldsymbol{y}_n)(\boldsymbol{v}_i - \boldsymbol{y}_n)^{\mathrm{T}}}{(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}$$

$$+ \frac{4g_{n,i}}{(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}[(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)(\boldsymbol{v}_i - \boldsymbol{y}_n)^{\mathrm{T}} + (\boldsymbol{v}_i - \boldsymbol{y}_n)(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}]$$

$$+ \frac{8g_{n,i}^2}{(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)}(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)^{\mathrm{T}}.$$

One advantage of the polygonal line algorithm is that the optimization step can simultaneously determine the optimal coordinates of all vertices, compared to some other approaches (Kégl *et al*., 1999; Verbeek *et al*., 2002) in which only a single vertex v_i is updated while all other vertices are kept fixed. In other words, the optimization procedure is iterated in a cyclic fashion (that is, after optimizing v_{k+1}, the procedure starts again for v_1) until convergence is achieved.

III.3.3 Local Linear Partitioning

In the proposed polygonal line algorithm, we need to partition the sample data into local regions for linear principal component analysis. Let Y be a set of N samples $\{y_n\}_{n=1}^N$ in $\mathfrak{R}^r$. Given the current k line segments h_i ($i = 1, 2, \ldots, k$) of polygonal line h, define the local regions $V_1, V_2, \ldots, V_k$ as

$$V_i = \{y_n \in Y \,|\, i = \arg\min_j \Delta(y_n, h_j)\}.$$

Hence, V_i contains all data points for which the ith line segment h_i is the closest. The way to construct the local partition by projecting the N sample data into k regions is built on Euclidean distance, and the mean vectors of the k local partitions are sample mean, i.e. $\hat{\mu}_i = \dfrac{1}{N_i} \sum_{y_n \in V_i} y_n$ where N_i is the data size of the ith local partition.

III.3.4 Adding a New Vertex

Given the local partitions defined above, for each V_i, we apply PCA to its sample covariance matrix $\hat{\Sigma}_i = \dfrac{1}{N_i} \sum_{y_n \in V_i} (y_n - \hat{\mu}_i)(y_n - \hat{\mu}_i)^{\text{T}}$ to determine r_i, the number of

dominant eigenvalues that is great than $\hat{\sigma}^2$ (In the algorithm, we choose $\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots > \hat{\lambda}_r \geq 2\hat{\sigma}^2$). If there is a partition with r_i greater than one, it implies that some nonlinear relationships are involved in the data within V_i. Therefore, a combination of multiple line segments rather than the first principal component of $\hat{\Sigma}_i$ will produce more accurate data representation. In case of multiple local partitions with $r_i > 1$, we can pick up one whose second dominant eigenvalue contributes larger to its total variance within V_i compared to other partitions. The midpoint of the chosen line segment then serves as the initialization of the new vertex for $\boldsymbol{h}$. This incremental scheme that increases the number of line segments by one will help avoid the over-fitting problem in polygonal line estimation.

Thus, in an iterative way the algorithm will increase the number of segments by adding a new vertex to the current polygonal line estimate. After adding a new vertex, the positions of all vertices are updated by the nonlinear optimization method in Section III.3.2.

Newton's method in the vertices optimization step uses the first and second derivatives and indeed performs better than the steepest descent method on the convergence property when the starting point is near the optimal solution (Bertsekas, 1995). To provide a good initialization for the vertices when a new vertex point is added, we propose the following projection-regression scheme to obtain better initial values of vertices. This initialization scheme arises from the piecewise linear basis function formulation of the polygonal lines, which iterates between two steps:

Projection Step: Given the current vertices as the columns of the weight matrix **W** in Equation (3.7), the projection step estimates the projection index t along the polygonal line constructed by **W**, as discussed in III.2.2. The basis function b_w is then obtained by Equation (3.8).

Regression Step: Given the basis $b_{w,n}$ for each data point, the regression-step produce the initialization for vertices that minimizes $\sum_{n=1}^{N} \left\| y_n - \mathbf{W} b_{w,n} \right\|^2$. The standard linear regression method applies to obtain:

$$\mathbf{W} = [y_1 \ y_2 \quad \cdots \quad y_N]\mathbf{B}_t^{\mathrm{T}} (\mathbf{B}_t \mathbf{B}_t^{\mathrm{T}})^{-1},$$

where $\mathbf{B}_t = [b_{w,1} \ b_{w,2} \ \ldots \ b_{w,N}]$.

Given the polygonal line representation, it is straightforward to see that the distance error (3.5) is non-increasing after each regression step. Hence the proposed projection-regression scheme always converges and produces a good initialization for the subsequent vertices optimization.

III.3.5 Computational Complexity

The complexity of the polygonal line algorithm is dominated by the complexity of the projection step in local linear partitioning, which is $O(Nk)$ and k is the current number of line segments. Increasing the number of line segments one at a time (as described in Section III.3.4), the complexity of the algorithm to obtain the final polygonal line with K segments is $O(NK^2)$. As the experimental studies showed, the optimal number of line segments K is smaller compared to the sample size of data N

(usually $K < N^{1/2}$). This is slightly better than the $O(N^2)$ complexity of the HS principal curve algorithm.

The number of segments K is an important factor that controls the approximation errors. If the noise variance is relatively small, we can keep the approximation error low by allowing a larger K; on the other hand, increasing K would not improve the overall performance, so, a smaller number of segments is chosen.

The complexity can be dramatically decreased in certain situations. One possibility is to add more than one vertex at a time. For example, instead of adding only one vertex, new vertices are placed at the midpoint of these segments h_i for which $r_i > 1$ ($i =$ 1, 2, …, k). Then we can reduce the computational complexity for producing h. One can also set K to be a constant since increasing K beyond a certain threshold brings only diminishing returns, giving $O(NK)$ computational complexity. Also, the number of initial line segments for polygonal line algorithm can be naturally set to a constant in certain applications if we have a priori knowledge about the data. These simplifications work well in certain situations, but the original algorithm is more robust.

The polygonal line algorithm can be summarized as:

1) Project data x in $\Re^d$ onto PCA score space in $\Re^r$ by Equation (2.8).

2) Initialize the polygonal line $h^{(0)}$ with r line segments.

3) Update vertices $\{v_1, …, v_k\}$ and line segments.

4) Partition sample data into k local regions $\{V_1, …, V_k\}$.

5) Determine the local partition with multiple dominant eigenvalues, add a new vertex and go to step 3), until the algorithm converges when the improvement on objective function is below some threshold.

6) Transform back the polygonal line to original data space via Equation (2.10).

III.4 A Nonlinear Variation Pattern Identification Example

In this section, we present an example from sheet metal stamping process, in which a sheet metal is first clamped by the binders around the periphery of the die cavity and is subsequently drawn into the die cavity by a moving punch to form its final shape (Kim and Kim, 2000; Hsu *et al*., 2000). For purpose of illustration, a round cup constructed by CAD data is shown in Fig. 20, which is of desired shape from the metal forming design. Fig. 21 shows its simplified stamping process with some process variables: F_p is the punch force, F_b is the blank holder force, and F_r is the restraining force within the blank.

In spite of the designed shape as Fig. 20 illustrates, in the real manufacturing, however, due to the variation on some process variables such as blank holder force (BHF), sheet thickness, and frictions, there is often some wrinkling appearing on the formed parts after the stamping operation. As discussed in Chapter II, wrinkling is one of the main considerations regarding the quality of stamped parts, which is critical to avoiding problems in subsequent assembly and in the final product performance. In the present study, we will study how the wrinkling evolves with respect to the varying blank holder force through a set of simulations. Numerical simulation using finite

element analysis (FEA) method has become a prime tool to predict the wrinkling for the complicated sheet metal operations. The commercial finite element code Dynaform is used in this case study for the round cup forming.
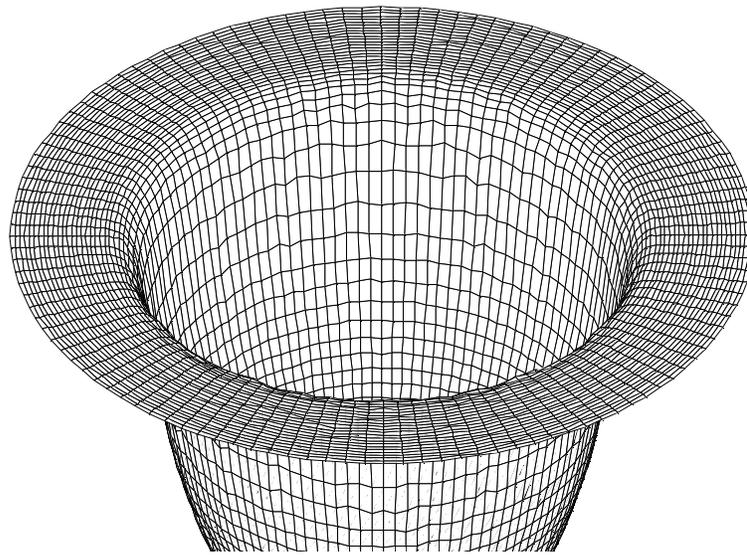
Fig. 20. The plot of a round cup constructed by CAD data from a stamping process.
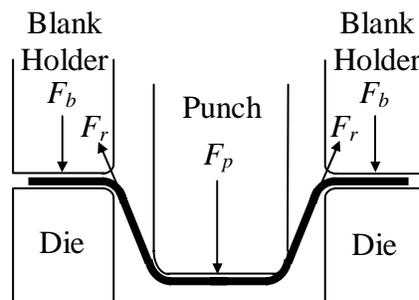
Fig. 21. Schematic of a simplified stamping process for a round cup.

In this case study, the value of BHF was varied in each FEA simulation that drawn randomly from a normal distribution with the standard blank holder force level as the mean. The measurement vector x consists of $d = 2457$ elements representing the height of the wrinkles over an array of locations distributed across the part, which could be obtained from laser scanned parts in real deep drawing stamping practice. As discussed in (Hsu *et al.*, 2000), the wrinkles on the surface are expected to be reduced as BHF increases until there is almost no wrinkling when blank holder force reaches an appropriate level. If BHF continues to increase, there will be a ring around the punch radius which may cause tearing fracture when the force is greater than some specific threshold. Based on a sample of $N = 48$ simulations, there were $\hat{r} = 3$ dominant eigenvalues as shown in Fig. 22. Fig. 23 shows the scatter plot of the first three PCA scores corresponding to the dominant eigenvalues, as well as the fitted principal curves $\hat{h}(\bullet)$ in 3-dimensional space using the proposed polygonal line algorithm and HS algorithm. Comparing the curves in Fig. 23, it can be seen that the proposed algorithm is capable of identifying the nonlinear pattern present in data y.
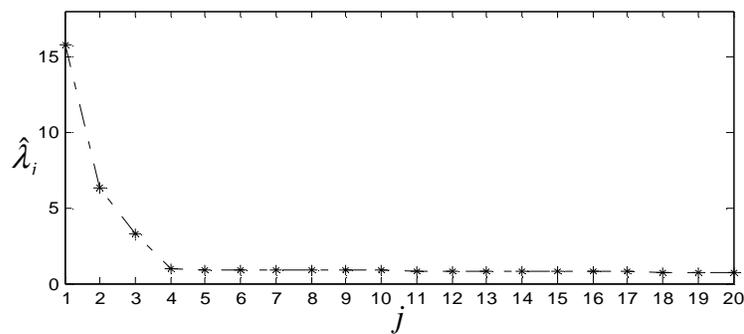
Fig. 22. The plot of the fist 20 eigenvalues from the sample covariance matrix of x.
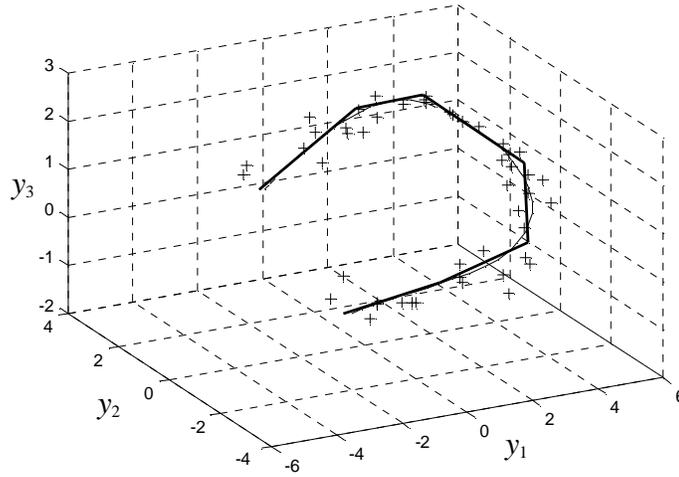
Fig. 23. The estimated principal curves from polygonal line algorithm (solid line) and HS algorithm (dashed line) in the 3-dimensional subspace of dominant PCA scores.

Transforming $\hat{h}(\bullet)$ back to the principal curve $\hat{f}(\bullet)$ in the original 2457-dimensional space, we can identify the nature of the variation pattern by graphically illustrating how $\hat{f}(\bullet)$ varies as t is varied as discussed in Section II.5. Fig. 24 shows four frames (for different values spanning the range of t) produced from the principal curve $\hat{f}$ with different values of t for which $\hat{f}(\bullet)$ is currently plotted. Investigating the plots in Fig. 24 reveals that lower BHF (corresponding to the smaller projection index t) caused wrinkles on the surface of the formed part (see Fig. 24a). Increasing the blank holder force will reduce the wrinkles (see Fig. 24b and 24c). However, inappropriately excessive BHF also restricted the material flow and the high restraining force would produce the ring around the punch radius that moves up (see Fig. 24d). In the worst case, a high BHF applied on the binder would lead to tearing failure near the punch radius. These distinct frames from variation pattern identification results then illustrate

how the wrinkles changing from part to part when BHF varies in the FEA simulations. In this example, the physical meaning of projection index t is explained as a monotonic function of the blank holder force.
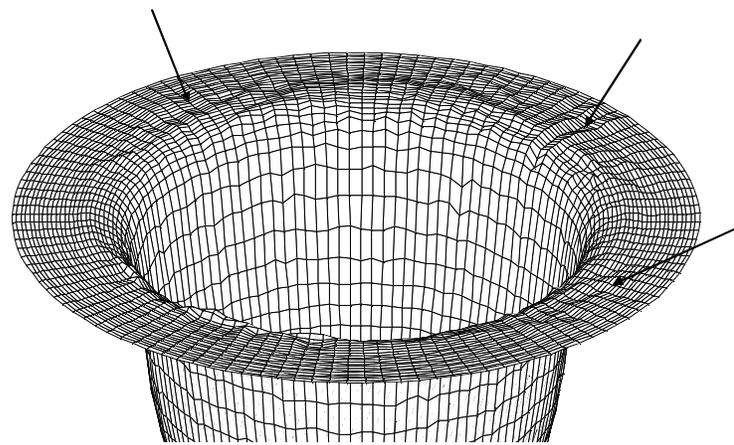
(a) $t = 0.29$
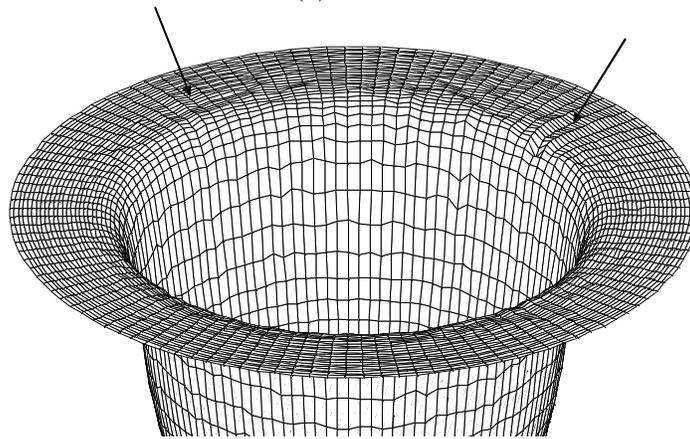
(b) $t = 4.08$

Fig. 24. Illustration of the nonlinear variation pattern with four different frames, in which the wrinkles are indicated by the arrows.
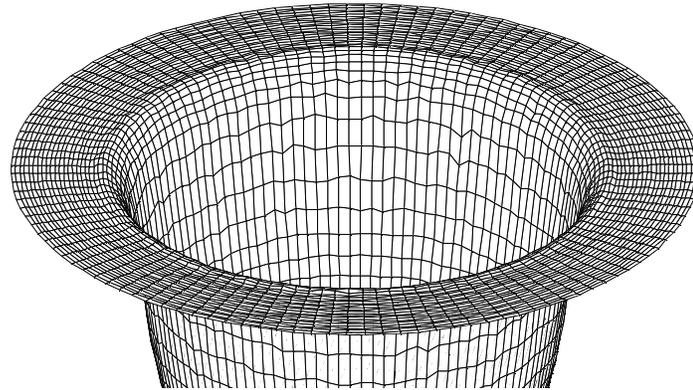
(c) $t = 12.95$
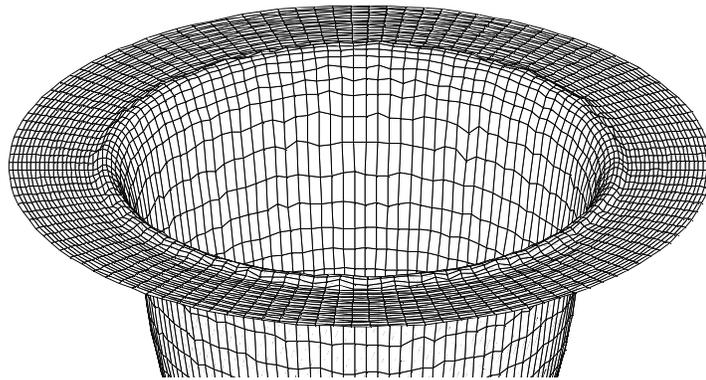
(d) $t = 23.76$

Fig. 24. Continued.

In real manufacturing variation reduction situations, the visualization result from the identified wrinkling variation pattern provides an intuitive way for the engineers or operators to gain insight of the nature of root cause of the wrinkles on the surface of formed part. For example, comparing the different plots of wrinkling in Fig. 24 with respect to various values of t in an ascending order, they will identify the variation

source as the variation on BHF by utilizing their engineering knowledge of the stamping process. And the blank holder force should be adjusted in such a manner that it suppresses the wrinkling tendency and produces the desired stamped parts in real manufacturing process.

III.5 Simulation Study

In this section, we will investigate the performance of the proposed polygonal line algorithm in different simulation cases. First recall the commonly used generative model (LeBlanc and Tibshirani, 1994), from which the simulation data are generated by decomposing the observable data into a systematic part and a noise part, i.e.

$$x_n = f(t_n) + w, \quad n = 1, \ldots, N.$$

If not stated otherwise, the noise w is assumed Gaussian, that is, $w \sim N_d(0, \sigma^2 I)$. Here, the latent variable t as a random variable is uniformly distributed on a curve (also called generating curve) and w is independent of t. In the following simulation examples, the principal curve was constructed from the generating curve on a lower dimensional data space. The curve was then multiplied by a Householder mirror reflection matrix to d-dimensional space ($d = 12$) so that it still fell in the r-dimensional linear variety, as discussed in Section II.6.

The performance of our algorithm is compared with HS principal curve algorithm on the basis of how closely the sample data points are concentrated on the estimated curves. For the cases of the different generating curves, we also evaluate, in a quantitative manner, how well the polygonal line algorithm summarizes the

multivariate data points as the data sizes and noise variance vary. 10,000 Monte Carlo replications were averaged in each case to compare the empirical square distance measures J_{HS} and J_{PL} in Equation (3.5), which denote the HS algorithm and the proposed polygonal line algorithm that applied to the PCA scores. We use different nonlinear patterns, noise parameters, and data sizes to investigate the robustness of the polygonal line algorithm.

III.5.1 A Half-circle Curve

In this case, the generating curve is a smooth half circle of radius one (that is, $r = 2$), as illustrated in Fig. 25. Three values of σ are used to represent the various noise variance considered in the generative data model.
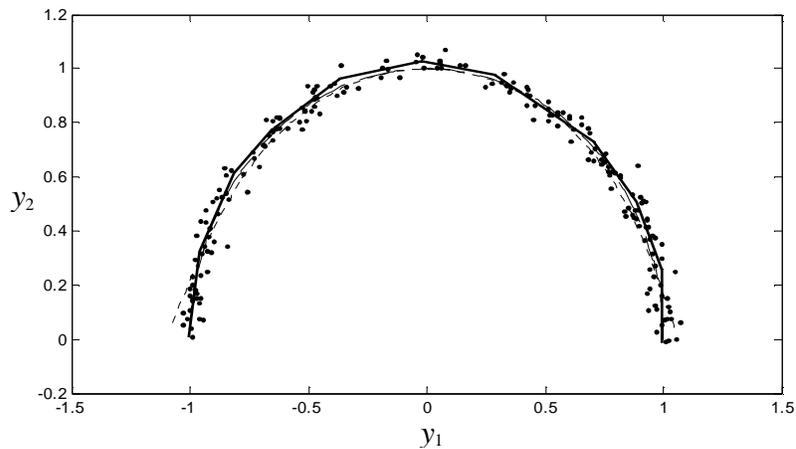
Fig. 25. The piecewise linear curve (solid line) produced by the polygonal line algorithm with $K = 10$, compared with HS principal curve (dotted line). Both from the sample data ($N = 200$ and $\sigma = 0.05$) generated from a half-circle curve (dashed line).

For various N and σ, the corresponding simulation results in Table 2 showed that the performances of two algorithms (HS and the polygonal line algorithm) are

comparable. The HS and our algorithm behave similarly in approximating the generating curve given the set of sample data.

If the distribution is concentrated on a curve and the noise variance is small, we expect both algorithms to approximate the generating curve very closely, as demonstrated in Table 2. The second situation that estimation bias will be reduced is when the sample size is large. Such an improvement on algorithm performance is justified by the results in Table 2. For example, when the other parameters were fixed, the estimation errors for both algorithms were reduced from $N = 100$ to $N = 400$.

Table 2. Performance comparison between HS algorithm and our polygonal line algorithm for the half-circle curves ($K = 10$).

N	σ	J_{HS} ($\times 10^{-3}$)	J_{PL} ($\times 10^{-3}$)
100	0.025	2.1	2.9
100	0.05	5.6	6.3
100	0.1	12.9	13.4
200	0.025	1.9	2.8
200	0.05	5.5	6.2
200	0.1	12.7	13.2
400	0.025	1.8	2.7
400	0.05	5.3	6.1
400	0.1	12.6	13.1

III.5.2 An S-Shaped Curve

In this case, we generate data from an S-shaped curve with varying sample size, noise variance to demonstrate the robustness of the polygonal line algorithm. Fig. 26 compares the piecewise curve produced by our algorithm ($K = 10$) and that from the HS

algorithm, in which the polygonal line algorithm produces curves that fit the data set closely while the HS algorithm exhibits more bias. This observation can be further illustrated with the quantitative comparisons in Table 3. The HS principal curve estimate did not follow the shape of the distorted S-shaped curve very well due to the smoothness constraint applied to the HS algorithm.
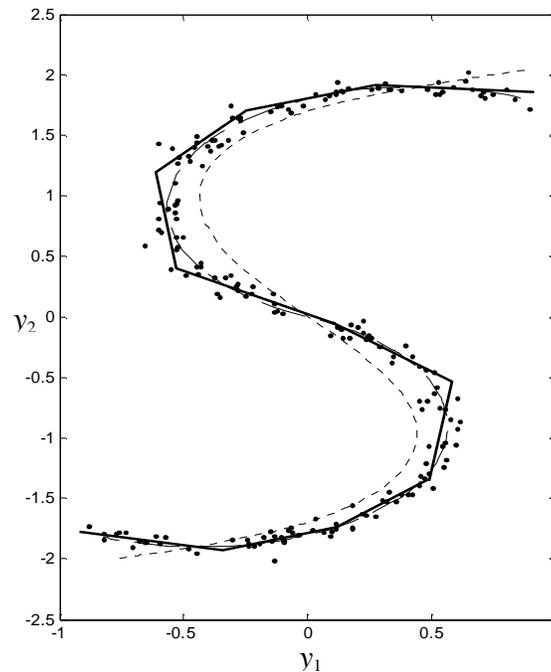
Fig. 26. The piecewise linear curve (solid line) produced by the polygonal line algorithm with $K = 10$, compared with HS principal curve (dotted line). Both from the sample data ($N = 200$ and $\sigma = 0.05$) generated from an S-shaped curve (dashed line).

The simulation results in Table 3 also justify the earlier conclusion that the estimation accuracy was improved as the data size grows and/or the variance of the noise decreases.

Table 3. Performance comparison between HS algorithm and our polygonal line algorithm for the S-shaped curves ($K = 10$).

N	σ	J_{HS} ($\times 10^{-3}$)	J_{PL} ($\times 10^{-3}$)
100	0.025	8.9	2.4
100	0.05	11.2	4.9
100	0.1	23.6	14.5
200	0.025	8.8	2.3
200	0.05	10.9	4.6
200	0.1	22.8	13.9
400	0.025	8.6	2.2
400	0.05	10.8	4.5
400	0.1	22.4	13.7

In some cases, however, there may be a preference for smooth curves over piecewise linear curves. To accommodate such situations we assign the projection index t_i to each vertex point v_i (implying an ordering on the data) from the polygonal line estimation. Then one may apply general regression methods to find a "good" smooth curve for the pairs of t_i and v_i ($i = 1, 2, \ldots, K$). In this sense the estimated polygonal line can serve as an initialization of other methods that seek smooth curves.

III.5.3 Failure Modes for HS Algorithm

We present two specific situations when the HS principal curve algorithm fails to recover the generating curve. In the first scenario, we take a zigzag curve as the generating curve in the data model, which consists of 6 segments of equal line length such that the two consecutive segments join at a right angle. The sample data were evenly distributed along each line segment in all simulation cases.

Fig. 27 shows the curves produced by the HS and the polygonal line algorithm when data size $N = 200$ and noise variance parameter $\sigma = 0.05$, which schematically illustrates the robustness and identification accuracy of the polygonal line algorithm using local PCA interferences for highly bended nonlinear patterns. As in the previous simulation cases, we introduced noise with different variances (i.e., $\sigma = 0.025, 0.05$, and 0.1) to the generating curve to compare the performance between the two algorithms. For the varying noise variances and data sizes, the polygonal line algorithm always identifies the generating nonlinear pattern. The comparison in Table 4 also provides a quantitative justification for Fig. 27 that the polygonal line algorithm performs much better than the HS algorithm.
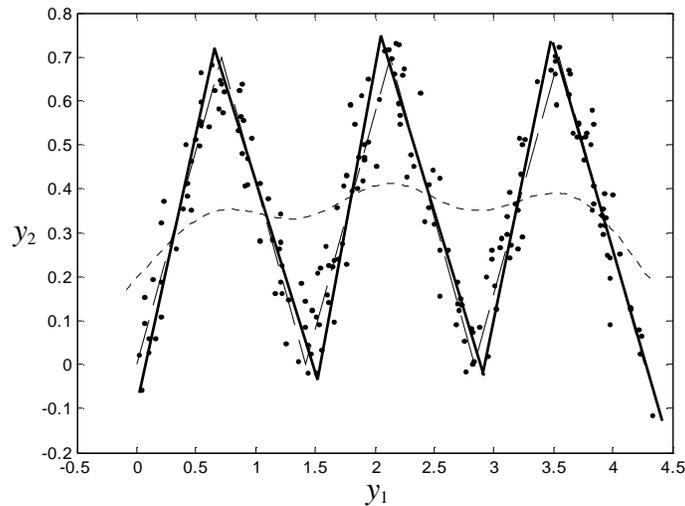
Fig. 27. The polygonal line (solid line) by our algorithm with $K = 6$, compared with HS principal curve estimation (dotted line). Both from the sample data ($N = 200$ and $\sigma = 0.05$) generated from a zigzag curve (dashed line).

Table 4. Performance comparison between HS algorithm and our polygonal line algorithm for zigzag curves.

N	σ	J_{HS} ($\times 10^{-3}$)	J_{PL} ($\times 10^{-3}$)
100	0.025	18.6	0.4
100	0.05	22.1	1.5
100	0.1	37.4	4.6
200	0.025	18.1	0.3
200	0.05	21.9	1.4
200	0.1	36.8	4.5
400	0.025	17.5	0.3
400	0.05	21.4	1.3
400	0.1	36.3	4.3

In this scenario, the polygonal line algorithm achieved high estimation accuracy and outperformed the HS algorithm. The abrupt changes in the direction of the generating zigzag curve causes the HS algorithm to oversmooth the estimated principal curve, even when the data is concentrated on the generating curve. On the other hand, the polygonal line algorithm can identify the piecewise linear curve with the number of line segments determined automatically.

The second scenario when the HS algorithm fails to produce a meaningful result is that the generating curve is more complicated. In other words, the estimated principal curve from HS algorithm can not identify the global nonlinear pattern from the sample of data. The spiral-shaped curves were chosen to investigate the robustness of the polygonal line algorithm for such complex nonlinear patterns, as shown in Fig. 28. The variance of the noise was set to $\sigma = 0.025$ to compare the performances of the two algorithms for recovering the shape of generating curves in the least noise situation. Fig. 28 shows the results estimated by the HS and the polygonal line algorithm for two

cases represented by curves with different levels of complexity. The polygonal line algorithm is capable of identifying the nonlinear patterns of both generating curves, while the HS algorithm either oversmoothes the curve (Fig. 28a) or fails to recover the shape of the spiral curve (Fig. 28b).
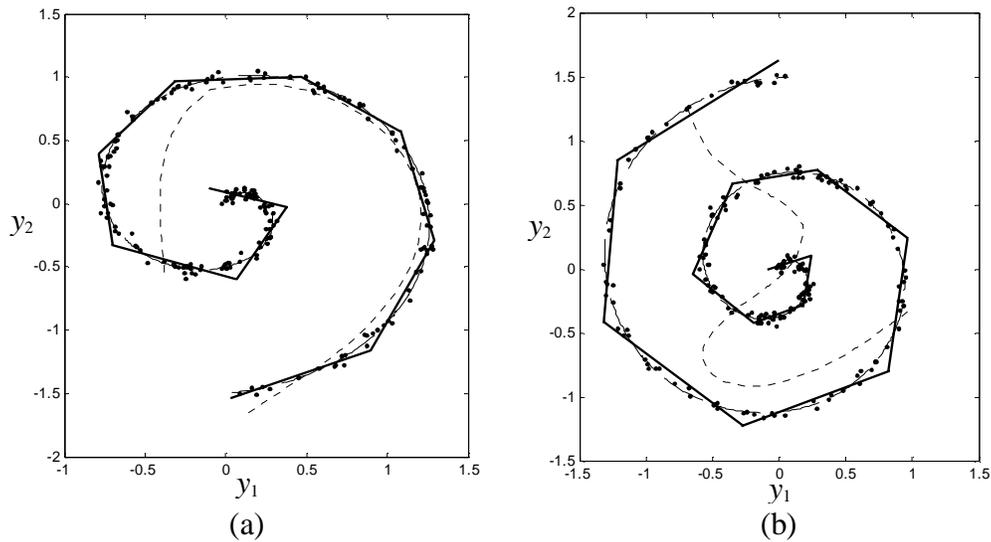
Fig. 28. The polygonal lines (solid line) by our algorithm, compared with HS principal curve estimations (dotted line). Both from the sample data ($N = 200$ and $\sigma = 0.025$) generated from spiral curves (dashed line) (a) a simple curve; (b) a more complex curve.

The failure here is due to the fact that the HS algorithm is stuck in a local minimum between the initial curve (the first principal component) and the desired solution (the generating curve). If this likely to occur in an application, one possible way to improve estimation is to replace the initialization step of HS algorithm by a more sophisticated routine that approximately captures the nonlinear pattern of the data. On the other hand, the proposed polygonal line algorithm identifies the nonlinear pattern as a collection of

line segments, each of which is viewed as a linear model for data in the corresponding local partition. Therefore, the complex global data structure can be extracted by the combination of good representations for local partitions using the polygonal line algorithm, which is free of the smoothness constraint applied to the HS algorithm.

III.6 Chapter Summary

In this chapter, we propose a polygonal line algorithm for principal curve estimation and nonlinear variation pattern identification, in which the local PCA inferences are utilized to determine the number of line segments automatically. Since most nonlinear curves may be approximated arbitrarily close by piecewise linear curves, the piecewise linear curves produced by the proposed algorithm can identify a variety of nonlinear patterns when the number of line segments is chosen optimally. The proposed algorithm was applied to a deep drawn stamping process for nonlinear variation pattern identification illustration. The Monte Carlo simulations have demonstrated that the polygonal line algorithm can produce robust and accurate estimates for varying nonlinear patterns. In the situations when the data points are concentrated on a curve with high curved or complicated shape, the polygonal line algorithm outperforms the HS principal curve algorithm in recovering the shape of nonlinear generating patterns. In addition, the polygonal line learning algorithm is helpful in reducing the computation complexity when the sample size is large.

CHAPTER IV

NONLINEAR BLIND SOURCE SEPARATION FOR IDENTIFYING MULTIPLE MANUFACTURING VARIATION PATTERNS

IV.1 Representing Multiple Nonlinear Variation Patterns

The methods for modeling and identifying single nonlinear variation pattern have been discussed in Chapter II and III, which appear to be effective in facilitate the diagnostics and elimination of root causes contributing to the process variability. Due to the complex physical characteristics, however, it is not uncommon to have various variation patterns simultaneously present in the manufacturing over measurement data. Often one or more of the variation patterns are nonlinear. In light of this, it will be more generic to extending the single nonlinear variation pattern identification methods by considering multiple potential variation faults that occur in the manufacturing processes. In this chapter, a nonlinear model is proposed to represent the possible multiple variation sources in high dimensional manufacturing processes. The separability of this model will be theoretically verified and an effective pattern identification algorithm is developed accordingly.

Consider a d-sensor manufacturing process that provides the measurement vector $\boldsymbol{x}$. Let v_i be the ith unit-variance random source characterized by a linear pattern vector $\boldsymbol{c}_i$ ($i = 1, 2, \ldots, p$), and $\boldsymbol{f}_j(t_j)$ be the nonlinear variation pattern caused by source t_j ($j = 1, 2,$

…, q). In the nonlinear variation pattern model as below, all variation sources v_i or t_j are assumed to be statistically independent, as well as of the noise vector $\boldsymbol{w}$,

$$\boldsymbol{x} = v_1\boldsymbol{c}_1 + v_2\boldsymbol{c}_2 + \ldots + v_p\boldsymbol{c}_p + \boldsymbol{f}_1(t_1) + \boldsymbol{f}_2(t_2) + \ldots + \boldsymbol{f}_q(t_q) + \boldsymbol{w}$$

$$= \boldsymbol{Cv} + \sum_{j=1}^{q} \boldsymbol{f}_j(t_j) + \boldsymbol{w}. \tag{4.1}$$

The nonlinear variation patterns $\boldsymbol{f}_j$ in model (4.1) are assumed to lie in one-dimensional manifolds embedded in the complete data space, as illustrated in Fig. 29. As demonstrated by the examples in previous chapters, this assumption is realistic for manufacturing variation analysis, which allows us to separate and identify the distinct linear or nonlinear variation patterns under a later proposed independence criterion.
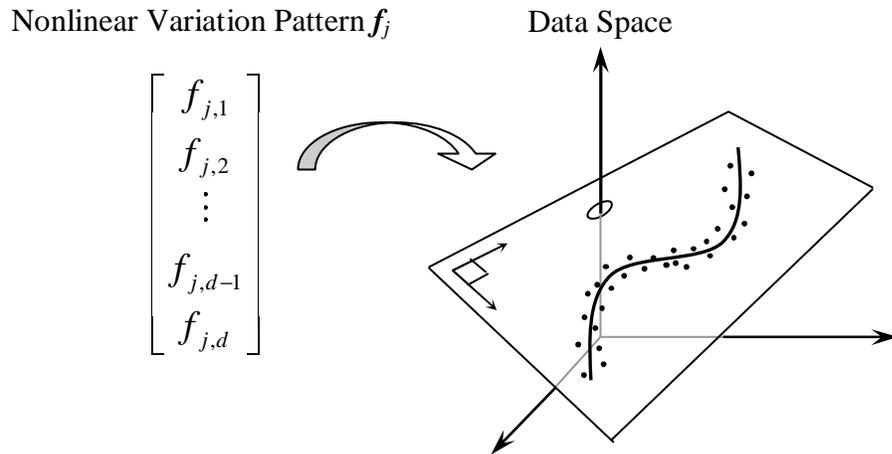
Fig. 29. The nonlinear variation pattern $\boldsymbol{f}_j$ lies in a low-dimensional manifold embedded in the data space.

As discussed in Chapter II, the single nonlinear variation pattern model takes the pattern $\boldsymbol{f}$ as a principal curve which is non-parametric and represented by a list of

ordered projection indices. Each nonlinear variation pattern $\boldsymbol{f}_j$ in model (4.1) is also a one-dimensional curve embedded in $\mathfrak{R}^d$ and can be identified by principal curve estimation methods. Therefore, the nonlinear model (4.1) provides flexibility for modeling a variety of variation patterns that are free of parametric or differentiable requirements. Since the model (4.1) does not require any prior knowledge of the nonlinear relationship among the different variables of $\boldsymbol{x}$ or the distribution of t_j other than the model structure, it can be applied to a more general problem setting of blind variation patterns analysis.

The objective of the nonlinear blind variation pattern identification problem based on model (4.1) aims to separate the effects of multiple variation sources mixed together in the data, which is similar to that of the nonlinear BSS problem. A nonlinear BSS model, in its most general form, is a mapping

$$\boldsymbol{x} = \boldsymbol{F}(\boldsymbol{t}) + \boldsymbol{w}, \tag{4.2}$$

where $\boldsymbol{x}$ and $\boldsymbol{t}$ denote the observation and independent sources, respectively. Then one needs to blindly estimate the nonlinear mapping $\boldsymbol{F}$ and sources $\boldsymbol{t}$ given the sample data only. However, in such a general model, independence assumption on sources does not imply source separation (Jutten and Karhunen, 2003; Taleb and Jutten, 1999). Hence, for general nonlinear BSS problem, a totally blind solution is not possible and one must assume some extra information, e.g., about the structure of the nonlinear mapping or the distribution of the sources (Taleb, 2002; Archard and Pham, 2001).

Up to now, most existing nonlinear BSS methods are developed based on conditioning the nonlinear mappings in a parametric model (e.g., approximating the

nonlinearities in model (4.2) by sigmoidal functions for instance) or distribution of the sources (e.g., assuming uniform distribution on the sources) (Valpola *et al.*, 2003; Yang, *et al.*, 1998; Harmeling *et al.*, 2003; Lappalainen and Honkela, 2000; Haritopoulos *et al.*, 2002; Pajunen and Karhunen, 1997). However, even the aforementioned simplifications of the nonlinear BSS problem can not lead to the theoretical justification for their separability of multiple independent sources (Theis and Lang, 2001). Moreover, the existing nonlinear BSS methods are typically only applied to the low dimensional ($d = 2$ or 3) applications.

As a generalization of single nonlinear variation pattern identification methods, the proposed nonlinear model (4.1) will deal with the multiple variation patterns identification problem that: 1) automatically selects the number of unknown independent sources; 2) provides theoretical analysis on separability of variation patterns; 3) avoids the differentiable requirement of the nonlinear mappings, or the pre-defined distribution on sources. The main technical challenges of this chapter are to determine the number of variation patterns and develop an effective algorithm to identify c_i and f_j over the sample of data. Proper identification of these patterns can provide insight into the nature of the phenomena that caused the variation faults.

Before deriving the variation patterns identification algorithm for model (4.1), we present some concepts from linear BSS methods that we need for multiple nonlinear patterns separation and identification. Consider the linear BSS model

$$x_{d \times 1} = \mathbf{C}_{d \times p} v_{p \times 1} + w \tag{4.3}$$

where the independent sources v are assumed to have unit-variance. Rather than working on the d-dimensional data x, linear blind source separation methods usually apply to the uncorrelated PCA score vector $y = \mathbf{W}^{-1}x$, where the whiten matrix $\mathbf{W}^{-1} = (\mathbf{\Lambda}_p - \sigma^2\mathbf{I})^{-1/2}\mathbf{Z}_p^{\mathrm{T}}$. Recall that $\mathbf{C} = \mathbf{Z}_p(\mathbf{\Lambda}_p - \sigma^2\mathbf{I})^{1/2}\mathbf{Q} = \mathbf{WQ}$ in Section I.2, we have $y = \mathbf{Q}v + \mathbf{W}^{-1}w$ where $\mathbf{Q}$ is a $p{\times}p$ orthogonal matrix.

When there is at most one variation source v_i is Gaussian, the linear BSS methods apply the properties of fourth order cumulants to estimate the unique matrix $\mathbf{Q} = [q_1 \ q_2 \dots q_p]$ and corresponding $\mathbf{C}$ in model (4.3). The fourth order cumulant for y is defined as $Cum_{i,j,k,l}(y) = E[y_iy_jy_ky_l] - E[y_iy_j]E[y_ky_l] - E[y_iy_k]E[y_jy_l] - E[y_iy_l]E[y_jy_k]$ ($1 \le i, j, k, l \le p$). We enumerate the most interesting properties of cumulants in Appendix A based on their definition (Stuart and Ord, 1987).

Given y and $p{\times}p$ matrix $\mathbf{M}^{kl} = e_ke_l^{\mathrm{T}}$, where e_k is a column vector with a 1 in k^{th} position and 0's elsewhere ($1 \le k, l \le p$), we define the cumulant matrix $\mathbf{C}_{\mathrm{Y}}(\mathbf{M}^{kl})$ component-wise (Cardoso, 1998) ($1 \le i, j \le p$)

$$[\mathbf{C}_{\mathrm{Y}}(\mathbf{M}^{kl})]_{ij} = Cum_{i,j,k,l}(y).$$

Following the assumptions of independence on v and Gaussian distribution on w, it is straightforward to establish the structure of $\mathbf{C}_{\mathrm{Y}}(\mathbf{M}^{kl})$ by properties A.1 ~ A.6

$$\mathbf{C}_{\mathrm{Y}}(\mathbf{M}^{kl}) = \mathbf{Q}\mathbf{\Delta}(\mathbf{M}^{kl})\mathbf{Q}^{\mathrm{T}}, \tag{4.4}$$

where $\mathbf{\Delta}(\mathbf{M}^{kl}) = Diag(k(v_1)q_{1k}q_{1l},\dots,k(v_p)q_{pk}q_{pl})$. In the factorization of $\mathbf{C}_{\mathrm{Y}}(\mathbf{M}^{kl})$ in Equation (4.4), the unknown kurtosis $k(v_i) = E[v_i^4] - 3E^2[v_i^2]$ enter only in the

diagonal matrix $\Delta(\mathbf{M}^{kl})$. Then the problem reduces to estimating $\mathbf{Q}$ as the diagonalizer of $\mathbf{C}_Y(\mathbf{M}^{kl})$. Note that the eigenvectors are uniquely determined if and only if the eigenvalues are all distinct. If $\mathbf{M}^{kl}$ is randomly chosen, then the eigenvalues are distinct with probability 1 (Cardoso, 1998). However, the definition of $\mathbf{M}^{kl}$ makes it not a random matrix. A reasonable way to alleviate this problem is to jointly process several cumulant matrices. Let $\boldsymbol{M} = \{\mathbf{M}^{kl}\}$ ($1 \le k, l \le p$) be a set of $p \times p$ matrices, and Off($\mathbf{F}$) $= \sum_{i \ne j} (f_{ij})^2$ the sum of the squares of the off-diagonal elements of matrix $\mathbf{F}$, which is a measure of diagonality of $\mathbf{F}$. The estimate of $\mathbf{Q}$ is taken as the orthonormal matrix $\mathbf{U}$ that minimizes the joint diagonality criterion:

$$\boldsymbol{\Phi}_{\boldsymbol{M}}(\mathbf{U}) \equiv \sum_{k,l} \mathrm{Off}\,(\mathbf{U}^{\mathrm{T}}\mathbf{C}_Y(\mathbf{M}^{kl})\mathbf{U}). \tag{4.5}$$

As an estimate of $\mathbf{Q}$, $\mathbf{U}$ is the rotation matrix that simultaneously brings the cumulants matrices $\mathbf{C}_Y(\mathbf{M}^{kl})$ generated by $\boldsymbol{M}$ to be as diagonal as possible. Define $\tilde{\boldsymbol{y}} = \mathbf{U}^{\mathrm{T}}\boldsymbol{y}$ as the rotated PCA scores, and $\mathbf{CM}$ as the pairwise cumulant matrix in a component-wise form, i.e., $[\mathbf{CM}]_{ij} = Cum^2_{i,i,i,j}(\tilde{\boldsymbol{y}}) + Cum^2_{i,i,j,j}(\tilde{\boldsymbol{y}}) + Cum^2_{i,j,j,j}(\tilde{\boldsymbol{y}})$ ($1 \le i, j \le p$).

IV.2 Multiple Nonlinear Variation Patterns Identification Algorithm

Given a set of sample data, the proposed nonlinear variation pattern identification algorithm starts with a whiten step that transforms the original data $\boldsymbol{x} \in \mathfrak{R}^d$ to uncorrelated PCA scores $\boldsymbol{y} \in \mathfrak{R}^r$. Next, the PCA scores are rotated by an orthogonal matrix $\mathbf{U}$ such that the components of $\boldsymbol{y}$ are as independent as possible. As shown in the

theoretical result below, the pairwise cumulant matrix **CM** of rotated PCA scores will have a diagonal block structure. Each block in **CM** corresponds to a single variation pattern. The final step of the algorithm is to separate these blocks and identify their corresponding linear or nonlinear variation patterns individually. Note that the core component of the algorithm is the rotation step that will produce a block structured matrix **CM** under the joint diagonality criterion (4.5).

IV.2.1 The Diagonalized Pairwise Cumulant Matrix

The nonlinear model (4.1) represents multiple potential variation patterns present in manufacturing. Assume that each f_j lies in an r_j-dimensional ($r_j < d$) linear variety, and no other linear variety in which f_j lies has dimension smaller than r_j, then the eigenvalues $\{ \lambda_{f_j,k} : k = 1, \ldots, d \}$ of covariance matrix of f_j, denoted as Σ_{f_j}, will have

$$\lambda_{f_j,1} \geq \lambda_{f_j,2} \geq \ldots \geq \lambda_{f_j,r_j} > 0 = \lambda_{f_j,r_j+1} = \lambda_{f_j,r_j+2} \ldots = \lambda_{f_j,d} . \tag{4.6}$$

The nonlinear variation pattern f_j can be represented by its PCA decomposition, i.e.

$$\boldsymbol{f}_j = s_{f_j,1}\boldsymbol{e}_{f_j,1} + \ldots + s_{f_j,r_j}\boldsymbol{e}_{f_j,r_j} = \mathbf{E}_{f_j}\boldsymbol{s}_{f_j} , \quad j = 1, 2, \ldots, q$$

where $\mathbf{E}_{f_j} = [\boldsymbol{e}_{f_j,1}\ \boldsymbol{e}_{f_j,2}\ \ldots\ \boldsymbol{e}_{f_j,r_j}]$ are the r_j eigenvectors associated with the dominant eigenvalues defined in Equation (4.6), and $\boldsymbol{s}_{f_j} = [s_{f_j,1}\ldots s_{f_j,r_j}]^{\mathrm{T}}$ are the PCA scores of f_j. Scaling the PCA scores such that they have unit-variances yields

$$\boldsymbol{f}_j = (s_{f_j,1}\big/\sqrt{\lambda_{f_j,1}})(\sqrt{\lambda_{f_j,1}}\,\boldsymbol{e}_{f_j,1}) + \ldots + (s_{f_j,r_j}\big/\sqrt{\lambda_{f_j,r_j}})(\sqrt{\lambda_{f_j,r_j}}\,\boldsymbol{e}_{f_j,r_j})$$

$$= \tilde{\mathbf{E}}_{f_j}\tilde{\boldsymbol{s}}_{f_j} . \tag{4.7}$$

Theorem 4.1: *Suppose that $f_j(\bullet)$ lies in r_j-dimensional linear variety and* **C** *has full rank. If* $\boldsymbol{\Sigma}_x$ *has* $r = p + \sum_j^q r_j$ *dominant eigenvalues greater than* σ^2, *then the pairwise cumulant matrix* **CM** *resulting from rotation matrix* **U** *under criterion* $\Phi_M(\bullet)$ *will have a diagonal block structure with p 1×1 blocks and q $r_j\times r_j$ blocks, each of which contains the sum of squared cross-cumulants of linearly transformed PCA scores of a single variation pattern.*

Proof: First, it follows from the assumption and source independence that

$$\boldsymbol{\Sigma}_x = \mathbf{CC}^\mathrm{T} + \tilde{\mathbf{E}}_{f_1}\tilde{\mathbf{E}}_{f_1}^\mathrm{T} + \cdots + \tilde{\mathbf{E}}_{f_q}\tilde{\mathbf{E}}_{f_q}^\mathrm{T} + \sigma^2\mathbf{I}. \tag{4.8}$$

The PCA decomposition provides another form for $\boldsymbol{\Sigma}_x$ as

$$\boldsymbol{\Sigma}_x = \mathbf{Z}_r(\boldsymbol{\Lambda}_r - \sigma^2\mathbf{I})\mathbf{Z}_r^\mathrm{T} + \sigma^2\mathbf{I}, \tag{4.9}$$

where $\boldsymbol{\Lambda}_r$ and $\mathbf{Z}_r$ are the r dominant eigenvalues and eigenvectors of $\boldsymbol{\Sigma}_x$.

To make the two forms of $\boldsymbol{\Sigma}_x$ in Equations (4.8) and (4.9) consistent, we have

$$[\mathbf{C}\tilde{\mathbf{E}}_{f_1}...\tilde{\mathbf{E}}_{f_q}] = \mathbf{Z}_r(\boldsymbol{\Lambda}_r - \sigma^2\mathbf{I})^{1/2}\mathbf{Q} \tag{4.10}$$

where **Q** is an $r\times r$ orthogonal matrix. Multiplying $\boldsymbol{x}$ by the whiten matrix yields

$$\boldsymbol{y} = \mathbf{W}^{-1}\boldsymbol{x} = (\boldsymbol{\Lambda}_r - \sigma^2\mathbf{I})^{-1/2}\mathbf{Z}_r^\mathrm{T}(\mathbf{C}\boldsymbol{v} + \tilde{\mathbf{E}}_{f_1}\tilde{\boldsymbol{s}}_{f_1} + ... + \tilde{\mathbf{E}}_{f_q}\tilde{\boldsymbol{s}}_{f_q} + \boldsymbol{w})$$

$$= \mathbf{Q}[\boldsymbol{v}^\mathrm{T}\,\tilde{\boldsymbol{s}}_{f_1}^\mathrm{T}...\tilde{\boldsymbol{s}}_{f_q}^\mathrm{T}]^\mathrm{T} + \mathbf{W}^{-1}\boldsymbol{w}$$

$$= [\mathbf{Q}_v\,\mathbf{Q}_{s,1}...\mathbf{Q}_{s,q}][\boldsymbol{v}^\mathrm{T}\,\tilde{\boldsymbol{s}}_{f_1}^\mathrm{T}...\tilde{\boldsymbol{s}}_{f_q}^\mathrm{T}]^\mathrm{T} + \mathbf{W}^{-1}\boldsymbol{w}$$

$$= \mathbf{Q}_v\boldsymbol{v} + \sum_{j=1}^q\mathbf{Q}_{s,j}\tilde{\boldsymbol{s}}_{f_j} + \mathbf{W}^{-1}\boldsymbol{w}$$

$$\equiv \boldsymbol{y}_v + \sum_{j=1}^q\boldsymbol{y}_{t,j} + \boldsymbol{y}_w \tag{4.11}$$

where $\mathbf{Q}_v$ and $\mathbf{Q}_{s,j}$ are $r \times p$ and $r \times r_j$ orthogonal matrices respectively ($j = 1, 2, \ldots, q$).

Consider the cumulant matrix $\mathbf{C}_Y(\mathbf{M}^{kl})$ for $\mathbf{y}$ and $r \times r$ matrix $\mathbf{M}^{kl}$. Due to the fact that $\mathbf{y}_v$, $\mathbf{y}_{t,j}$ and $\mathbf{y}_w$ in Equation (4.11) are independent, cumulant properties A.5 and A.6 decompose $\mathbf{C}_Y(\mathbf{M}^{kl})$ into a sum of cumulant matrices:

$$\mathbf{C}_Y(\mathbf{M}^{kl}) = \mathbf{C}_{Y_v}(\mathbf{M}^{kl}) + \sum_{j=1}^{q} \mathbf{C}_{Y_{t,j}}(\mathbf{M}^{kl}). \qquad (4.12)$$

Because $\mathbf{y}_v$ is a linear mixture of independent source $\mathbf{v}$ plus Gaussian noise, its cumulant matrix has the same structure as Equation (4.4) (Cardoso, 1998)

$$\mathbf{C}_{Y_v}(\mathbf{M}^{kl}) = \mathbf{Q}_v \mathbf{\Delta}_{M,v} \mathbf{Q}_v^T, \qquad (4.13)$$

where $\mathbf{\Delta}_{M,v} = Diag(k(v_1) q_{v,1k} q_{v,1l}, \ldots, k(v_p) q_{v,pk} q_{v,pl})$.

Each square matrix $\mathbf{C}_{Y_{t,j}}(\mathbf{M}^{kl})$ in Equation (4.12) by definition is a symmetric matrix, and represented by its eigen-decomposition, i.e., $\mathbf{C}_{Y_{t,j}}(\mathbf{M}) = \mathbf{P}_{M,j} \mathbf{\Delta}_{M,\tilde{s}_{f_j}} \mathbf{P}_{M,j}^T$ where $\mathbf{\Delta}_{M,\tilde{s}_{f_j}}$ is an $r_j \times r_j$ diagonal matrix. Since $\mathbf{y}_{t,j}$ is a linear mixture of non-independent elements of $\tilde{s}_{f_j}$, the matrix $\mathbf{P}_{M,j}$ resulting from the factorization of $\mathbf{C}_{Y_{t,j}}(\mathbf{M})$ not only depends on $\mathbf{Q}_{s,j}$ but on $\mathbf{M}^{kl}$. Such dependence of $\mathbf{P}_{M,j}$ on $\mathbf{M}^{kl}$ can be readily written as $\mathbf{P}_{M,j} = \mathbf{Q}_{s,j} \mathbf{T}_{M,j}$ where $\mathbf{T}_{M,j}$ is an $r_j \times r_j$ matrix and determined by $\mathbf{M}^{kl}$ and cumulants of $\tilde{s}_{f_j}$. Actually $\mathbf{T}_{M,j}$ is a rotation matrix in that both $\mathbf{P}_{M,j}$ and $\mathbf{Q}_{s,j}$ are orthogonal matrices by definition. The $r \times r$ matrix $\mathbf{C}_Y(\mathbf{M}^{kl})$, as a sum of the cumulant matrices in Equation (4.12), is written as

$$\mathbf{C}_{Y}(\mathbf{M}^{kl}) = [\mathbf{Q}_{v}\ \mathbf{Q}_{s,1}\mathbf{T}_{M,1}\cdots\mathbf{Q}_{s,q}\mathbf{T}_{M,q}]\begin{bmatrix}\boldsymbol{\Delta}_{M,v} & & & \\ & \boldsymbol{\Delta}_{M,\tilde{s}_{f_1}} & & \\ & & \ddots & \\ & & & \boldsymbol{\Delta}_{M,\tilde{s}_{f_q}}\end{bmatrix}\begin{bmatrix}\mathbf{Q}_{v}^{T} \\ \mathbf{T}_{M,1}^{T}\mathbf{Q}_{s,1}^{T} \\ \vdots \\ \mathbf{T}_{M,q}^{T}\mathbf{Q}_{s,q}^{T}\end{bmatrix}. \quad (4.14)$$

For an arbitrary matrix $\mathbf{M}^{kl}$ and its cumulant matrix $\mathbf{C}_{Y}(\mathbf{M}^{kl})$, denote the corresponding rotation matrix $\mathbf{U}_{M} = [\mathbf{U}_{M,v}\,\mathbf{U}_{M,1}\,\ldots\,\mathbf{U}_{M,q}]$, where the dimensions of sub-matrices $\mathbf{U}_{M,v}$ and $\mathbf{U}_{M,j}$ are chosen as those of $\mathbf{Q}_{v}$ and $\mathbf{Q}_{s,j}$. Regarding the matrix structure in Equation (4.14), the matrix $\mathbf{U}_{M}$ that diagonalizes $\mathbf{C}_{Y}(\mathbf{M}^{kl})$ should be $\mathbf{U}_{M} = [\mathbf{Q}_{v}\,\mathbf{Q}_{s,j}\mathbf{T}_{M,1}\,\ldots\,\mathbf{Q}_{s,q}\mathbf{T}_{M,q}]$.

Consider the joint diagnoality criterion $\Phi_{\boldsymbol{M}}(\bullet)$ that is equivalent to diagonalzing each $\mathbf{C}_{Y}(\mathbf{M}^{kl})$ ($\mathbf{M}^{kl}\in\boldsymbol{M}$), and define $\mathbf{U} = [\mathbf{U}_{v}\,\mathbf{U}_{1}\,\ldots\,\mathbf{U}_{q}]$ as the rotation matrix under the joint diagonality criterion for $\boldsymbol{M}$. Since $\mathbf{Q}_{v}$ is independent of matrix $\mathbf{M}^{kl}$, we can always have $\mathbf{U}_{v} = \mathbf{Q}_{v}$ to diagonalize $\mathbf{Q}_{v}\boldsymbol{\Delta}_{M_{m},v}\mathbf{Q}_{v}^{T}$ in Equation (4.13). Due to the fact that $\mathbf{P}_{M_{m},j}$ depends on matrix $\mathbf{M}^{kl}$, however, the matrix sought to jointly diagonalize the cumulant matrices $\mathbf{P}_{M_{m},j}\boldsymbol{\Delta}_{M_{m},\tilde{s}_{f_{j}}}\mathbf{P}_{M_{m},j}^{T}$'s is not unique concerning different $\mathbf{M}^{kl}$. In other words, for the set of matrices $\boldsymbol{M}$, there is no orthogonal matrix $\mathbf{U}_{j}$ that can simultaneously achieve the diagonality for all $\mathbf{C}_{Y_{t,j}}(\mathbf{M}^{kl}) = [\mathbf{Q}_{s,j}\mathbf{T}_{M,j}]\boldsymbol{\Delta}_{M_{m},\tilde{s}_{f_{j}}}[\mathbf{Q}_{s,j}\mathbf{T}_{M,j}]^{T}$ ($\mathbf{M}^{kl}\in\boldsymbol{M}$). Therefore, the joint diagonality criterion $\Phi_{\boldsymbol{M}}(\bullet)$ will produce an optimal matrix $\mathbf{U}_{j}$ to diagonalize each cumulant matrix as much as possible in the sense of minimizing the sum of squared off-diagonal elements in $\mathbf{C}_{Y_{t,j}}(\mathbf{M}^{kl})$. The matrix $\mathbf{U}_{j}$ resulting from $\Phi_{\boldsymbol{M}}(\bullet)$, as a linear transformation of orthogonal matrix $\mathbf{Q}_{s,j}$, is determined by the set of matrices

$\{\mathbf{T}_{\mathrm{M},j}: 1 \leq j \leq q\}$. Since $\mathbf{U}_j$ is not equal to $\mathbf{Q}_{s,j}$, denote $\mathbf{Q}_{s,j} = \mathbf{U}_j\mathbf{T}_j$ where $\mathbf{T}_j$ is an unknown $r_j \times r_j$ rotation matrix. Then the rotated PCA scores resulting from $\mathbf{U}$ are

$$\tilde{\boldsymbol{y}} = [\mathbf{Q}_v\, \mathbf{U}_1 \ldots \mathbf{U}_q]^{\mathrm{T}}\{[\mathbf{Q}_v\, \mathbf{Q}_{s,1} \ldots \mathbf{Q}_{s,q}][\boldsymbol{v}^{\mathrm{T}}\, \tilde{\boldsymbol{s}}_{f_1}^{\mathrm{T}} \ldots \tilde{\boldsymbol{s}}_{f_q}^{\mathrm{T}}]^{\mathrm{T}} + \mathbf{W}^{-1}\boldsymbol{w}\}$$

$$= \begin{bmatrix} \boldsymbol{v} + \mathbf{Q}_v^{\mathrm{T}}\mathbf{W}^{-1}\boldsymbol{w} \\ \mathbf{T}_1\tilde{\boldsymbol{s}}_{f_1} + \mathbf{U}_1^{\mathrm{T}}\mathbf{W}^{-1}\boldsymbol{w} \\ \vdots \\ \mathbf{T}_q\tilde{\boldsymbol{s}}_{f_q} + \mathbf{U}_q^{\mathrm{T}}\mathbf{W}^{-1}\boldsymbol{w} \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{v} + \mathbf{Q}_v^{\mathrm{T}}\mathbf{W}^{-1}\boldsymbol{w} \\ \boldsymbol{y}_{s,1} + \mathbf{U}_1^{\mathrm{T}}\mathbf{W}^{-1}\boldsymbol{w} \\ \vdots \\ \boldsymbol{y}_{s,q} + \mathbf{U}_q^{\mathrm{T}}\mathbf{W}^{-1}\boldsymbol{w} \end{bmatrix}, \tag{4.15}$$

where $\boldsymbol{y}_{s,k}$ is a rotated version of $\tilde{\boldsymbol{s}}_{f_k}$ and independent of $\boldsymbol{v}$ and $\boldsymbol{y}_{s,l}$ ($1 \leq k \neq l \leq q$).

Given the independence between $\boldsymbol{v}$, $\boldsymbol{y}_{s,k}$ and $\boldsymbol{w}$ and the properties of cumulants (note that the fourth order cumulants of Gaussians are zero), it is straightforward to establish the pairwise cumulant matrix $\mathbf{CM}$ for $\tilde{\boldsymbol{y}}$ as a block structured matrix, as illustrated in Fig. 30, in which the p 1×1 blocks contains the sum of squared cross-cumulants of v_i ($i = 1, 2, \ldots, p$) and the q $r_j \times r_j$ blocks contains those of $\boldsymbol{y}_{s,j}$ ($j = 1, 2, \ldots, q$). ∎

IV.2.2 Variation Patterns Identification

For convenience, we denote the 1×1 diagonal entries in Fig. 30 corresponding to sources v_i as "linear block", and the $r_j \times r_j$ sub-matrices as "nonlinear blocks". Under the separability condition as in linear BSS methods, the orthogonal matrix $\mathbf{Q}_v$ is unique and its estimate $\mathbf{U}_v$ is given by the theoretical result in Section IV.2.1. Then, the linear variation pattern matrix $\mathbf{C}$ is

$$\mathbf{C} = \mathbf{Z}_r(\boldsymbol{\Lambda}_r - \sigma^2\mathbf{I})^{1/2}\mathbf{U}_v \tag{4.16}$$

and the estimate for variation source $\boldsymbol{v}$ is

$$\hat{\boldsymbol{v}} = \mathbf{U}_v^{\mathrm{T}} \boldsymbol{y} . \qquad (4.17)$$
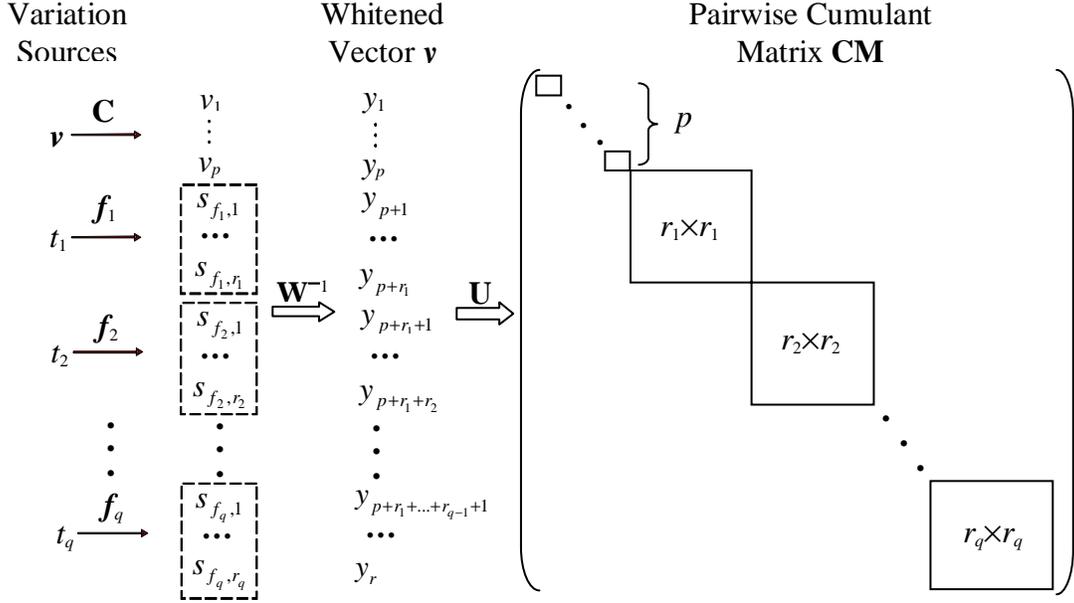
Fig. 30. The diagonal block structured matrix **CM** resulted from an optimal rotation matrix **U**.

Following Equations (4.7) and (4.10) we have

$$\boldsymbol{f}_j = \widetilde{\mathbf{E}}_{f_j} \widetilde{\boldsymbol{s}}_{f_j} = \mathbf{Z}_r (\boldsymbol{\Lambda}_r - \sigma^2 \mathbf{I})^{1/2} \mathbf{Q}_{s,j} \widetilde{\boldsymbol{s}}_{f_j}$$

$$= \mathbf{Z}_r (\boldsymbol{\Lambda}_r - \sigma^2 \mathbf{I})^{1/2} \mathbf{U}_j \mathbf{T}_j \widetilde{\boldsymbol{s}}_{f_j}$$

$$= \mathbf{Z}_r (\boldsymbol{\Lambda}_r - \sigma^2 \mathbf{I})^{1/2} \mathbf{U}_j \boldsymbol{y}_{s,j} . \qquad (4.18)$$

The rotated PCA scores in Equation (4.15) produces the estimates of $\boldsymbol{y}_{s,j}$ ($j = 1, 2, \ldots, q$)

$$\mathbf{U}^{\mathrm{T}} \boldsymbol{y} = [\mathbf{U}_v \, \mathbf{U}_1 \ldots \mathbf{U}_q]^{\mathrm{T}} \{ [\mathbf{Q}_v \, \mathbf{Q}_{s,1} \ldots \mathbf{Q}_{s,q}] [\boldsymbol{v}^{\mathrm{T}} \, \widetilde{\boldsymbol{s}}_{f_1}^{\mathrm{T}} \ldots \widetilde{\boldsymbol{s}}_{f_q}^{\mathrm{T}}]^{\mathrm{T}} + \mathbf{W}^{-1} \boldsymbol{w} \}$$

$$= \begin{bmatrix} \boldsymbol{v} + \mathbf{U}_v^{\mathrm{T}} \mathbf{W}^{-1} \boldsymbol{w} \\ \mathbf{T}_1 \tilde{\boldsymbol{s}}_{f_1} + \mathbf{U}_1^{\mathrm{T}} \mathbf{W}^{-1} \boldsymbol{w} \\ \vdots \\ \mathbf{T}_q \tilde{\boldsymbol{s}}_{f_q} + \mathbf{U}_q^{\mathrm{T}} \mathbf{W}^{-1} \boldsymbol{w} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{v}} \\ \hat{\boldsymbol{y}}_{s,1} \\ \vdots \\ \hat{\boldsymbol{y}}_{s,q} \end{bmatrix}. \tag{4.19}$$

Given the optimal rotation matrix $\mathbf{U}$, we can estimate the principal curve $\hat{\boldsymbol{h}}_j$ to $\hat{\boldsymbol{y}}_{s,j}$

in Equation (4.19) and then recover the estimate of nonlinear variation pattern by

$\hat{\boldsymbol{f}}_j = \mathbf{Z}_r (\boldsymbol{\Lambda}_r - \sigma^2 \mathbf{I}) \mathbf{U}_j \hat{\boldsymbol{h}}_j$ using Equation (4.18).

IV.3 A Clustering Method for Separating Blocks

As shown in Fig. 30, all the off-block elements in the pairwise cumulant matrix **CM**

should be zero from the theoretical results. In practice, however, the proposed theorem

is applied to sample data, and no matrix **U** can produce the ideal matrix structure as Fig.

30. In order to identify the linear and nonlinear variation patterns from a specific

sample, we present a clustering method to separate the blocks in the estimate of **CM**

and determine the number of variation sources, p and q.

In the definition of pairwise cumulant matrix, the cross-cumulants are adopted as an

independence measure for the elements in $\tilde{\boldsymbol{y}}$: if $\tilde{y}_i$ and $\tilde{y}_j$ are generated from two

independent variation sources, then $Cum_{i,i,i,j}(\tilde{\boldsymbol{y}}) = 0$ (the same for $Cum_{i,i,j,j}(\tilde{\boldsymbol{y}})$ and

$Cum_{i,j,j,j}(\tilde{\boldsymbol{y}})$, $1 \leq i \neq j \leq r$); otherwise, the cross-cumulant takes non-zero values. In

other words, the larger the magnitude of cross-cumulant is, the more likely that the

tested components are from a single variation source. By constructing this pairwise

"independence distance" for elements in $\tilde{\boldsymbol{y}}$, a clustering method will help us separate all the linear and nonlinear blocks in Fig. 30 by classifying their corresponding PCA scores into distinct patterns. The values of p and q are determined accordingly.

Define $\hat{\mathbf{C}\mathbf{M}}$, the estimate of pairwise cumulant matrix over a sample of $\tilde{\boldsymbol{y}}$, as

$$[\hat{\mathbf{C}\mathbf{M}}]_{ij} = \hat{Cum}_{i,i,i,j}^{2}(\tilde{\boldsymbol{y}}) + \hat{Cum}_{i,i,j,j}^{2}(\tilde{\boldsymbol{y}}) + \hat{Cum}_{i,j,j,j}^{2}(\tilde{\boldsymbol{y}})$$

$$= \left[\frac{1}{N} \sum_{n=1}^{N} \tilde{y}_{i,n}(\tilde{y}_{j,n})^3 - \frac{3}{N^2} \sum_{n=1}^{N} \tilde{y}_{i,n}\tilde{y}_{j,n} \sum_{n=1}^{N} (\tilde{y}_{j,n})^2 \right]^2 + \left[\frac{1}{N} \sum_{n=1}^{N} (\tilde{y}_{i,n})^2 (\tilde{y}_{j,n})^2 \right.$$

$$\left. - \frac{1}{N^2} \sum_{n=1}^{N} (\tilde{y}_{i,n})^2 \sum_{n=1}^{N} (\tilde{y}_{j,n})^2 - 2\left(\frac{1}{N} \sum_{n=1}^{N} \tilde{y}_{i,n}\tilde{y}_{j,n} \right)^2 \right]^2$$

$$+ \left[\frac{1}{N} \sum_{n=1}^{N} (\tilde{y}_{i,n})^3 \tilde{y}_{j,n} - \frac{3}{N^2} \sum_{n=1}^{N} \tilde{y}_{i,n}\tilde{y}_{j,n} \sum_{n=1}^{N} (\tilde{y}_{i,n})^2 \right]^2 .$$

The goal of clustering analysis is to segment the components of $\tilde{\boldsymbol{y}}$ such that those within each block in $\hat{\mathbf{C}\mathbf{M}}$ are more likely related to a single variation source than components assigned to different blocks. Hierarchical clustering is one of the widely used clustering methods in practice, which produce hierarchical representations in which the clusters at each level of the hierarchy are created by merging clusters at the next lower level. At the lowest level, each cluster contains a single component (Johnson and Wichern, 2002; Hastie *et al*., 2001, Cordon, 1987).

The proposed clustering method begins with every component representing a singleton subgroup. At each of the remaining steps the closest two (least independent) subgroups merged into a single group, producing one less group at the next higher

level. Given the hierarchical representation, the clustering method will decide how to choose the number of clusters such that the components within each cluster are more similar in distance to each other than to those assigned to different clusters at that level. Clustering is perceived as an unsupervised procedure since there are no predefined classes (Halkidi and Vazirgiannis, 2001; Salvador and Chan, 2004; Milligan and Cooper, 1985).

In order to separate the clusters (or blocks) in $\hat{\mathbf{CM}}$, we apply Dunn validity index to identify a set of "compact and well separated clusters" (Halkidi *et al*., 2000). The validity index is defined for a specific number of clusters n_c ($2 \leq n_c \leq r$)

$$\gamma_{n_c} = \min_{i=1,\ldots,n_c} \left\{ \min_{j=i+1,\ldots,n_c} \left[\frac{d(C_i, C_j)}{\max_{k=1,\ldots,n_c} diam(C_k)} \right] \right\}$$

where $d(C_i, C_j) = \min_{u \in C_i, v \in C_j} d(u,v)$ is the dissimilarity between two clusters C_i and C_j, and $diam(C) = \max_{u,v \in C} d(u,v)$ is the diameter of a cluster used as a measure of its dispersion, where $d^{-1}(u, v)$ is the uv^{th} entry of sample pairwise cumulant matrix.

If the dataset contains compact and well-separated clusters, the distance (dissimilarity) between the clusters is expected to be large and the diameter of the clusters is expected to be small. Thus, based on the Dunn index definition, we may conclude that large values of γ_{n_c} indicate the presence of desired clusters. Index γ_{n_c} does not exhibit any trend with respect to the number of clusters. Therefore, the maximum in the plot of γ_{n_c} versus n_c can be an indication of the optimal number of

blocks in Fig. 30. The simulations in Section IV.5 demonstrate that this clustering method is effective in providing accurate separation of linear and nonlinear blocks caused by independent variation sources.

The proposed nonlinear variation patterns separation and identification algorithm can be summarized as

1) Estimate r, the number of dominant eigenvalues of Σ_x.

2) Transform x into PCA score vector y using Equation (4.11).

3) Rotate PCA scores and calculate the optimal rotation matrix $\mathbf{U}$ by minimizing the joint diagonality criterion (4.5).

4) Apply the clustering method to sample pairwise cumulant matrix. Determine the number of blocks $n_c^* = \arg\max \gamma_{n_c}$, and segment the components of $\tilde{y}$.

5) Identify the linear variation pattern vectors c_i and corresponding estimate of variation sources v via Equation (4.16) and (4.17).

6) Estimate the individual nonlinear variation patterns f_j by fitting the principal curve $\hat{h}_j$ to $\hat{y}_{s,j}$ and recovering it via $\hat{f}_j = \mathbf{Z}_r(\mathbf{\Lambda}_r - \sigma^2\mathbf{I})\mathbf{U}_j\hat{h}_j$.

IV.4 Autobody Assembly Variation Pattern Illustrative Example

In this section, the applicability of the multiple nonlinear variation pattern identification algorithm is illustrated with an example from autobody assembly. Fig. 31 shows the layout of 26 measurement points (labeled 1-26 in the figure) taken at the body-in-white (BIW) stage of the assembly process. At this stage, the cross-member,

roof header, and plenum join the left and right bodysides to form the BIW. All x and y-direction deviations are measured for each point, except for points 10 and 23, for which only the x-direction deviations are measured. Thus, the measurement vector x has a total of $d = 50$ elements.
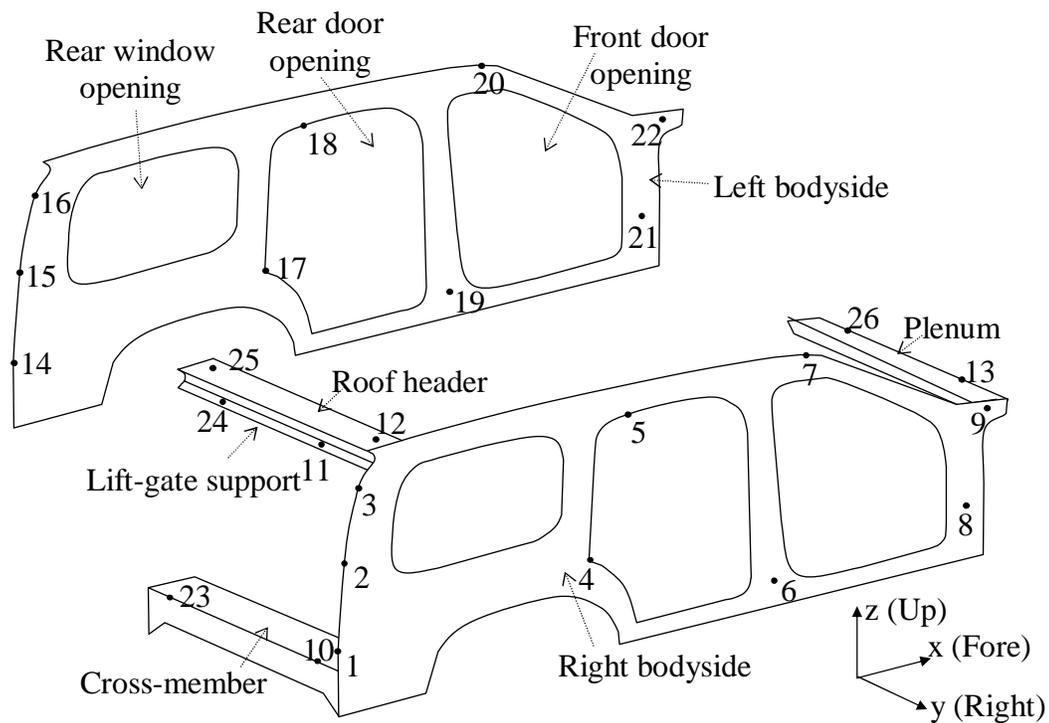
Fig. 31. Measurement layout at the BIW stage.

Before the bodysides and connecting members are welded together, they must be accurately located with respect to one another with pins rigidly attached to the fixtures (Apley and Shi, 2001). Each pin mates with either a hole or slot in the panel so that the bodyside position is constrained but not overly constrained. In practical assembly,

however, the repeated use makes it possible that the pins become loose or worn. As a consequence, the panel is no longer constrained in its desired position, and from autobody to autobody, the loose pin will cause a variation pattern in the BIW dimensions. As discussed in Chapter II, one of the keys to diagnosing dimensional variability in the assembly process is to estimate variation patterns based on a sample of data in order to gain insights into their root causes.

In section II.5 we discussed the example of a single nonlinear variation pattern (denoted f_1 here), which was identified based on the real measurement data collected from autobody assembly. To investigate the performance of the multiple variation patterns identification algorithm, in the present study, a linear variation pattern $c_1 v_1$ was added to f_1 such that it is suitable for the structure of model (4.1). This linear variation pattern is designed based on the case study in (Apley and Shi, 2001) and it represents a realistic dimensional variation fault in the autobody assembly process.

Given a sample of $N = 100$ data, there are $\hat{r} = 3$ dominant eigenvalues estimated from $\hat{\Sigma}_x$. The first 20 eigenvalues were plotted in Fig. 32.

Apply the proposed nonlinear variation patterns identification algorithm to the 3-dimensional PCA score vector y, and the sample pairwise cumulant matrix is

$$\hat{\mathbf{CM}} = \begin{bmatrix} 0.7168 & 0.3675 & 0.0054 \\ 0.3675 & 0.4693 & 0.0071 \\ 0.0054 & 0.0071 & 0.8532 \end{bmatrix}.$$

The visual inspection on the sample pairwise cumulant matrix indicated that there is a linear and nonlinear block resulting from the optimal rotation matrix, corresponding to

the linear and nonlinear variation pattern. The block clustering method proposed in Section IV.3 also produced $n_c^* = 2$ blocks (since $\gamma_2 = 7.18 > \gamma_3 = 0.86$) for the rotated PCA scores. Therefore, both the multiple nonlinear variation pattern identification result and the clustering approach identify the presence of two independent sources contributing to a linear and nonlinear variation pattern present in the sample data.
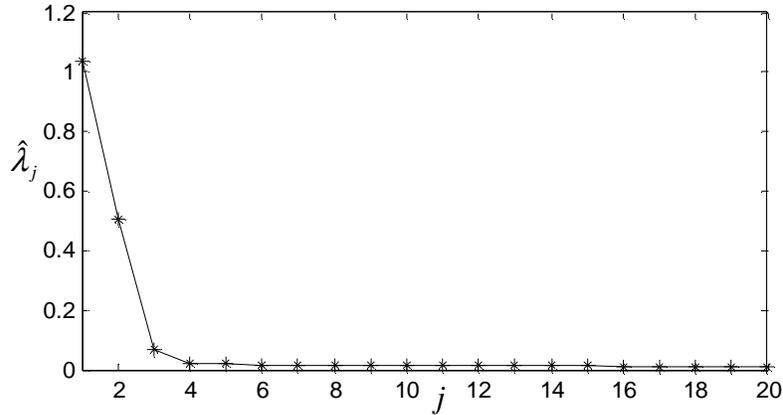
Fig. 32. The first 20 eigenvalues $\hat{\lambda}_j$ estimated from the sample data with $N = 100$.

After identifying the nonlinear block in the pairwise cumulant matrix, the corresponding nonlinear variation pattern f_1 was estimated via Equation (4.18). The nonlinearity among f_1 can be further illustrated by the scatter plot of its dominant PCA scores $\hat{s}_{f_1,1}$ and $\hat{s}_{f_1,2}$, as shown in Fig. 33.
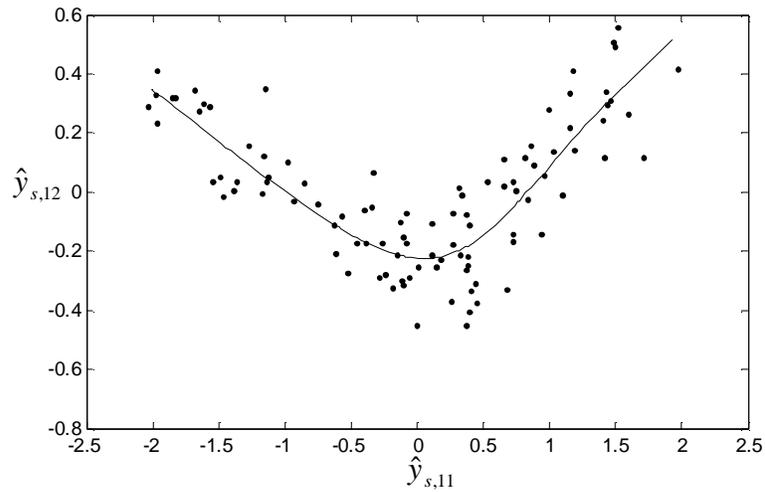
Fig. 33. The principal curve (solid line) estimated from the projections of nonlinear variation pattern $\hat{f}_1$ on the 2-dimensional dominant PCA score subspace.

As demonstrated in previous chapters, graphical displays of the estimated variation pattern generally facilitate identification of the root causes of variability. To gain diagnostic information from the nonlinear variation pattern, in Fig. 34 we plot the 6 elements of $\hat{f}_1$ (corresponding to the y-direction deviations of points 1, 2, 3, 14, 15, and 16) versus t, the projection index estimated from the principal curve method. The other elements of $\hat{f}_1$ are omitted due to their negligible magnitudes. From the comparison between Fig. 12 and Fig. 34, it can be seen that the identified nonlinear pattern $\hat{f}_1$ is almost indistinguishable from the single variation pattern estimate in Section II.4. This fact justified the performance of the proposed algorithm on identifying multiple linear and nonlinear variation patterns. The potential root cause for $\hat{f}_1$ is then a fixturing fault in locating the right bodyside in the framing station, as

discussed in Section II.1. When the right bodyside deviates to the right, it has no affect on the left bodyside, until the deviation amount is large enough to cause the interference between the upper cross-member and later with the left bodyside.
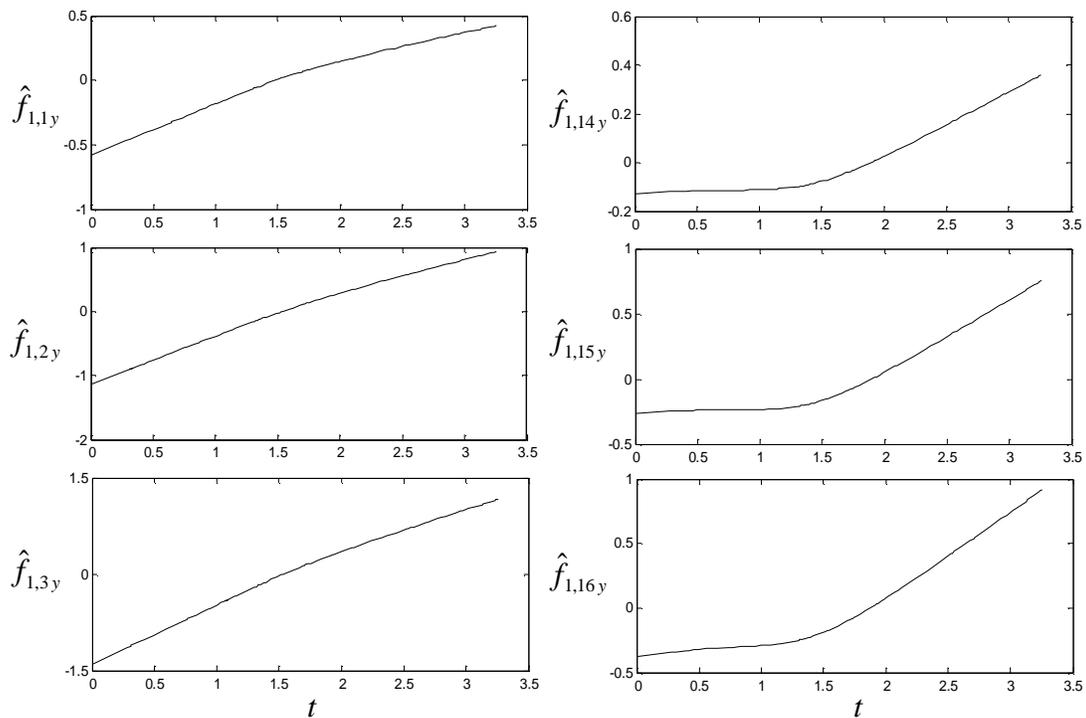
Fig. 34. Plots of the elements of $\hat{f}_1$ versus t, illustrating the characteristics of the nonlinear variation pattern with respect to the measurement points.

After identifying the linear and nonlinear blocks, the linear variation pattern vector c_1 and variation source v_1 were estimated and plotted in Fig. 35 and Fig. 36 respectively. The dominant 9 elements of $\hat{c}_1$ were {.241 .234 .246 .23 .228 .236 .244 .232 .251} that correspond to the x-direction deviation at points 1, 2, 3, 4, 5, 6, 7, 8, and

9. Each element of $\hat{\boldsymbol{c}}_1$ is plotted as an arrow at the location of measurement across the automobile body, as shown in Fig. 35. The length of the arrows is proportional to the magnitude of the elements in $\hat{\boldsymbol{c}}_1$ with the sign represented by the direction of the arrow. Many elements were negligibly small, in which case their arrows were omitted at the corresponding measure points. Note that the meaning of the pattern will not be affected by the direction of all arrows in Fig. 35.
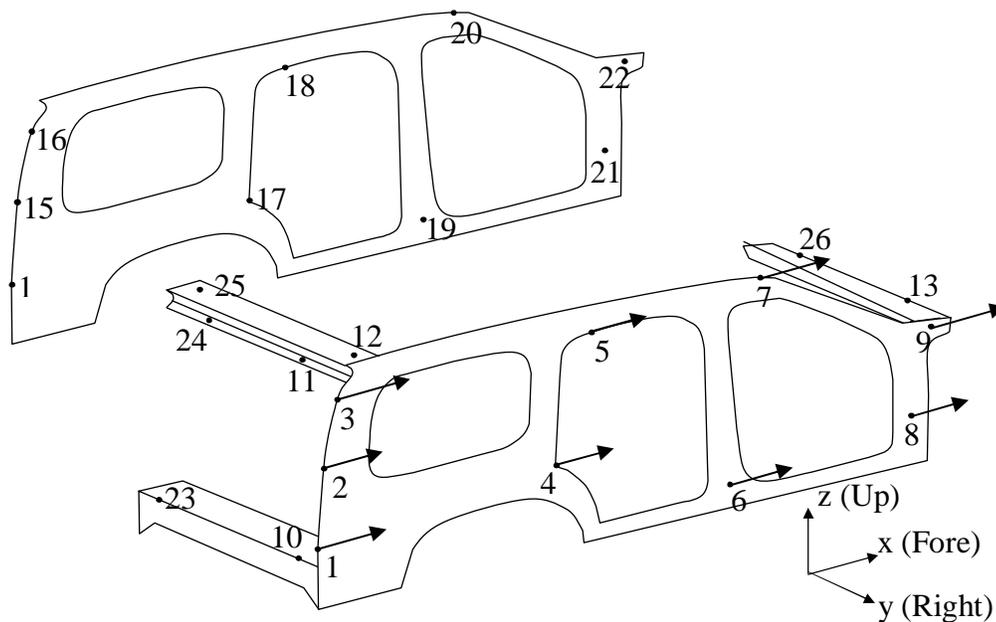
Fig. 35. Graphical illustration of the linear variation pattern for autobody assembly example.

It appears that the linear variation fault affects only points on the right bodyside, causing each point to translate by approximately the same amount. Thus, the variation source v_1 results in the right bodyside being incorrectly positioned with respect to the rest of the measured autobody sample. A plausible root cause for the linear variation

pattern is that the pin constraining the right bodyside in the x-direction becomes loose. When a bodyside is placed in the fixture, it may be positioned too far forward (in the positive x-direction). For the next automobile body, the right bodyside may be positioned too far aftward (in the negative x-direction). From autobody to autobody, the loose pin will cause the translation variation on the right bodyside. From the histogram of the estimated variation source $\hat{v}_1$ shown in Fig. 36, it is closely a Gaussian random variable, which implies the statistical characteristics of the root cause.
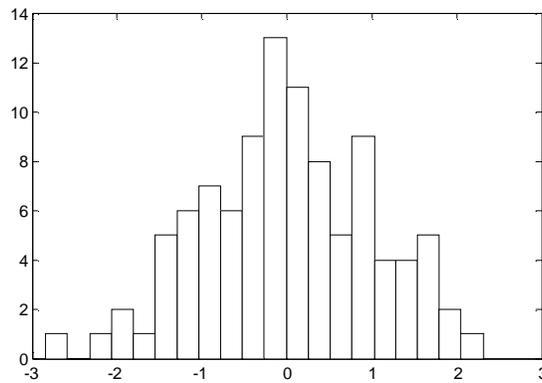
Fig. 36. Histogram of the estimated variation source v_1 associated with the linear variation pattern.

IV.5 Simulation Study

IV.5.1 Simulation Case 1

In this case, we demonstrate the performance of the proposed algorithm on a 18-dimensional data set from the generative model

$$x = c_1 v_1 + c_2 v_2 + f_1(t_1) + w$$

where the sources v_1, v_2 are a rectangle and cosine wave, and $\boldsymbol{f}_1$ is an S-shaped curve of t_1 that lies in a two-dimensional linear variety. The observation data $\boldsymbol{x}$ is distorted by Gaussian noise $\boldsymbol{w}$ with isotropic variance σ^2.

Since $\boldsymbol{f}_1$ in this example lies in $r_1 = 2$ dimensional subspace, the number of the dominant eigenvalues for $\boldsymbol{\Sigma}_x$ is $r = p + r_1 = 4$. The sources v_1, v_2 and their estimates $\hat{v}_1$ and $\hat{v}_2$ are shown in Fig. 37, which indicated the identification accuracy of the proposed algorithm in recovering the linear variation sources.
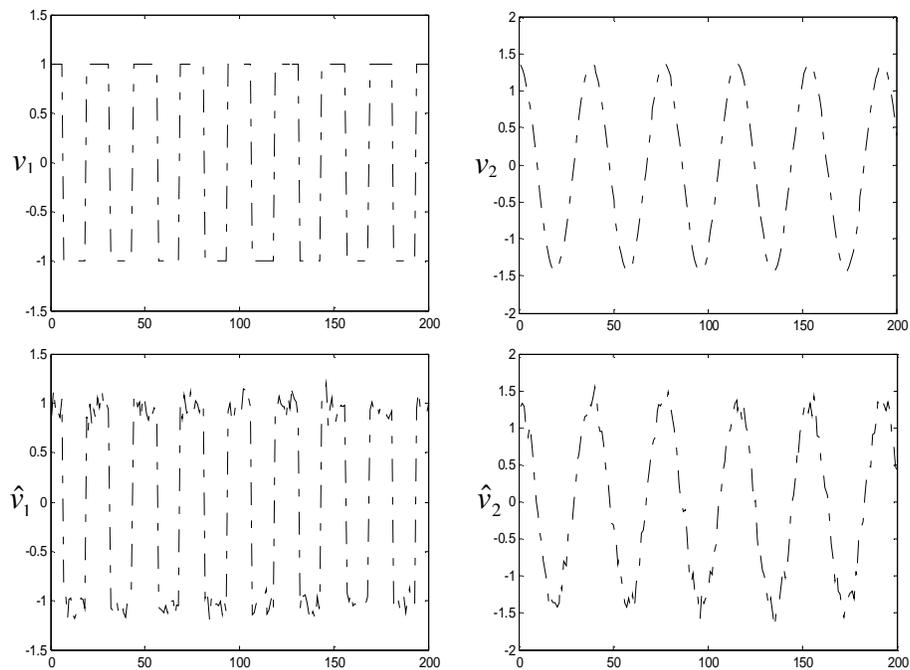
Fig. 37. The original sources v_1, v_2 and their estimates $\hat{v}_1$, $\hat{v}_2$ from the proposed algorithm when sample size $N = 200$ and $\sigma = 0.06$.

To visualize the performance of nonlinear variation pattern identification, Fig. 38 compared the estimated principal curve from the two-dimensional PCA scores of $\hat{f}_1$ to that from the generating data. The closeness between the two curves indicates that the proposed multiple nonlinear variation patterns identification algorithm is capable of separating and identifying the nonlinear pattern from the sample data.
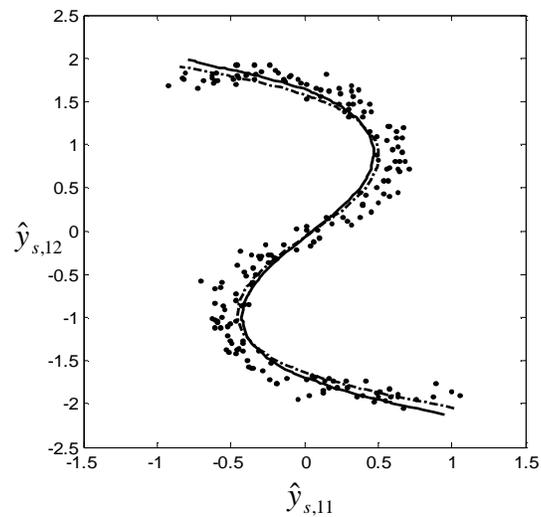
Fig. 38. The principal curve (dashed line) was estimated from the 2-dimensional PCA scores (dotted points) of $\hat{f}_1$, compared with the generating S-shaped curve (solid line) when $N = 200$ and $\sigma = 0.06$.

Fig. 39 also verified the accuracy of the algorithm by displaying the scatter plots of v_1 and $\hat{v}_1$ under different noise situations. Even in the worse noise situation when $\sigma = 0.12$, the multiple nonlinear variation pattern identification algorithm still provides a

good estimate $\hat{v}_1$ for v_1. In addition, Fig. 39 provides a graphical way to observe that

the estimation accuracy will be naturally deteriorated by the increasing noise variance.
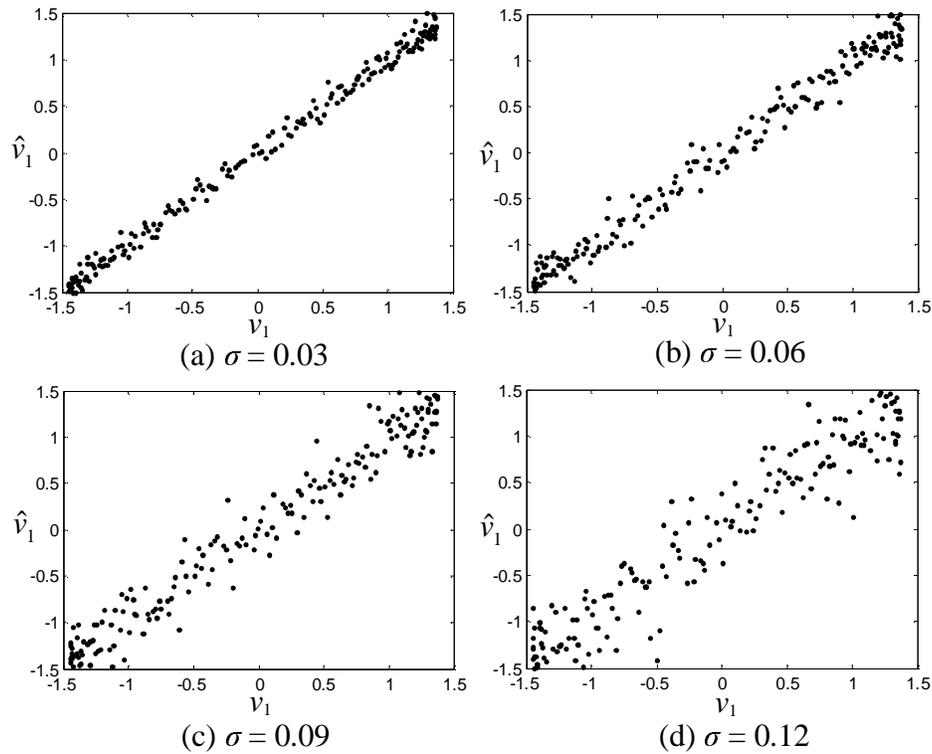
Fig. 39. Scatter plots of the variation sources v_1 and estimate $\hat{v}_1$ with sample size $N =$ 200 and varying σ from 0.03 to 0.12.

To study how the block clustering method perform in the synthetic data set, we use

various sample size and noise variance to investigate the correct clustering achieved by

the approach in Section IV.3. Fig. 40 shows the percentage (over 10,000 simulations)

of correct block separation. Although the clustering accuracy is usually deteriorated by

larger noise variances, the simulation results justify the performance of the clustering method on different cases, which always produces the correct block clustering results.
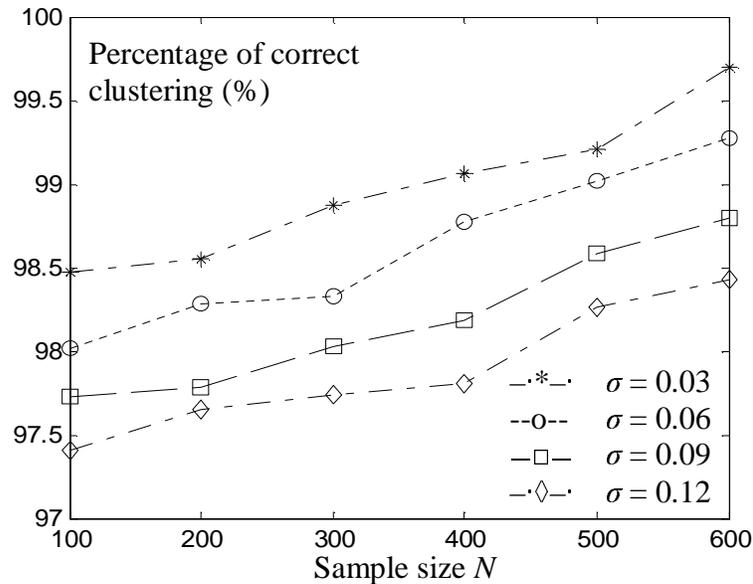
Fig. 40. Percentage of correct clustering of linear and nonlinear blocks with different sample sizes and noise variances for simulation case 1.

In order to evaluate the performance of the proposed multiple nonlinear variation pattern identification algorithm over the entire Monte Carlo simulations, consider the linear pattern identification accuracy measure $J_{Ci} = E\left[\left\|\hat{c}_i - c_i\right\|\left\|c_i\right\|^{-1}\right]$, where $\hat{c}_i$ denotes the estimate of the ith linear variation pattern vector c_i ($i = 1, 2$ in this case). The average value of $\left\|\hat{c}_i - c_i\right\|\left\|c_i\right\|^{-1}$ over 10,000 replicates is used to estimate J_{C1} and J_{C2}. The nonlinear variation pattern f_1 lies in a two-dimensional subspace, as well as the estimated curve $\hat{f}_1$. The projections of the original and estimated principal curves onto

the lower-dimensional space are illustrated in Fig. 38. In order to evaluate the estimation accuracy, we choose the distance measure $D_f = E\left\|\hat{\pmb{h}}_1 - \pmb{h}_1\right\|^2 l_{h_1}^{-1}$ to investigate the nonlinear variation pattern identification in a quantitative way, where $\pmb{h}_1$ and $\hat{\pmb{h}}_1$ are the nonlinear curves from the PCA scores of $\pmb{f}_1$ and $\hat{\pmb{f}}_1$ onto the two-dimensinal subspace. The weighted length of $\pmb{h}_1$ is $l_{h_1} = \int \| \pmb{h}_1'(t) \| g(t)dt$, where $g(t)$ is the density of t. In Table 5, the Monte Carlo simulations were averaged to summarize the performance of multiple linear and nonlinear variation patterns identification algorithm for this simulation case.

IV.6.2 Simulation Case 2

To investigate the performance of proposed algorithm on a more complicated nonlinear variation pattern model, the data in this case was generated from

$$\pmb{x} = \pmb{c}_1 v_1 + \pmb{c}_2 v_2 + \pmb{f}_1(t_1) + \pmb{f}_2(t_2) + \pmb{w},$$

where the independent sources v_1, v_2 and linear variation pattern vectors are the same as in the preceding example, nonlinear pattern $\pmb{f}_1$ is a quadratic function of t_1 that lies in the 2-dimensional linear variety, and $\pmb{f}_2$ is a piecewise linear curve lying in a 3-dimensional linear variety.

Given the linear and nonlinear variation patterns defined in the data model, the number of the dominant eigenvalues for $\pmb{\Sigma}_x$ is $r = p + r_1 + r_2 = 7$. The sources v_1, v_2 and their estimates $\hat{v}_1$ and $\hat{v}_2$ can be compared in the way as the first case. To demonstrate the estimation accuracy of nonlinear variation pattern identification, the estimated

principal curves from the PCA scores of $\hat{f}_1$ and $\hat{f}_2$ were compared with the curves from the generating data, as shown in Fig. 41 and 42.

Table 5. Performance results from the nonlinear BSS algorithm on multiple variation patterns identification for simulation case 1.

N	σ	J_{C1}	J_{C2}	D_f
100	0.03	0.029	0.043	0.00012
100	0.06	0.036	0.049	0.00041
100	0.09	0.041	0.053	0.00122
100	0.12	0.047	0.061	0.00216
200	0.03	0.028	0.041	0.00011
200	0.06	0.036	0.048	0.00036
200	0.09	0.040	0.052	0.00117
200	0.12	0.046	0.058	0.0021
300	0.03	0.027	0.039	0.00011
300	0.06	0.035	0.047	0.00036
300	0.09	0.039	0.051	0.00114
300	0.12	0.045	0.056	0.00207
400	0.03	0.026	0.039	0.00011
400	0.06	0.034	0.046	0.00033
400	0.09	0.038	0.05	0.00112
400	0.12	0.044	0.055	0.00203
500	0.03	0.024	0.038	0.00011
500	0.06	0.033	0.045	0.00029
500	0.09	0.038	0.049	0.00109
500	0.12	0.043	0.055	0.002
600	0.03	0.023	0.037	0.0001
600	0.06	0.032	0.043	0.00028
600	0.09	0.036	0.048	0.00106
600	0.12	0.041	0.054	0.00198
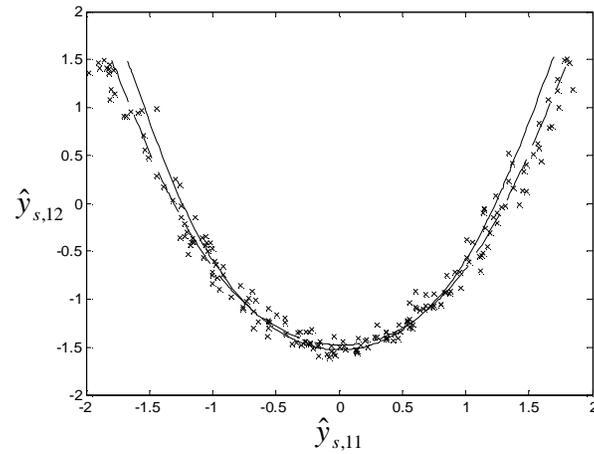
Fig. 41. The principal curve (dashed line) estimated from PCA scores (dotted points) of $\hat{f}_1$, compared with the generating quadratic curve (solid line) when $N = 200$ and $\sigma = 0.06$.
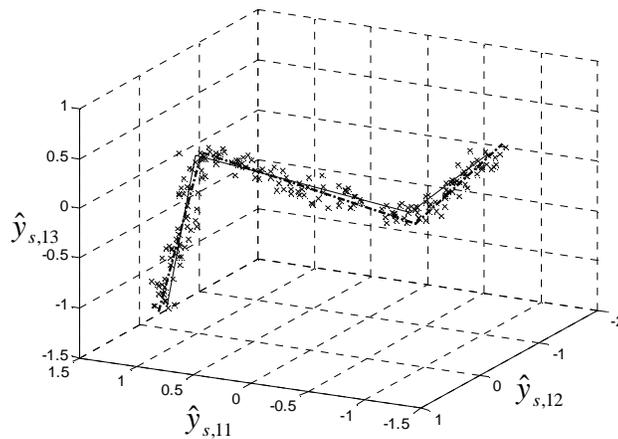
Fig. 42. The piecewise linear curve (dashed line) estimated from PCA scores (dotted points) of $\hat{f}_2$, compared with the generating piecewise linear curve (solid line) when $N = 200$ and $\sigma = 0.06$.

The closeness between the estimated curves and generating curves indicates that the proposed nonlinear algorithm is capable of separating and identifying each nonlinear

variation pattern. Due to the poor performance of HS algorithm on piecewise linear curve identification, the curves in Fig. 42 were estimated by the polygonal line algorithm proposed in Chapter III. To investigate the performance of the block clustering method, in Fig. 43 we plot the percentage of correct block clustering on the synthetic data set over various sample sizes and noise variances. As in the first simulation case, the observations are fairly intuitive that the success clustering percentages were improved as sample size N increases, and/or σ decreases.
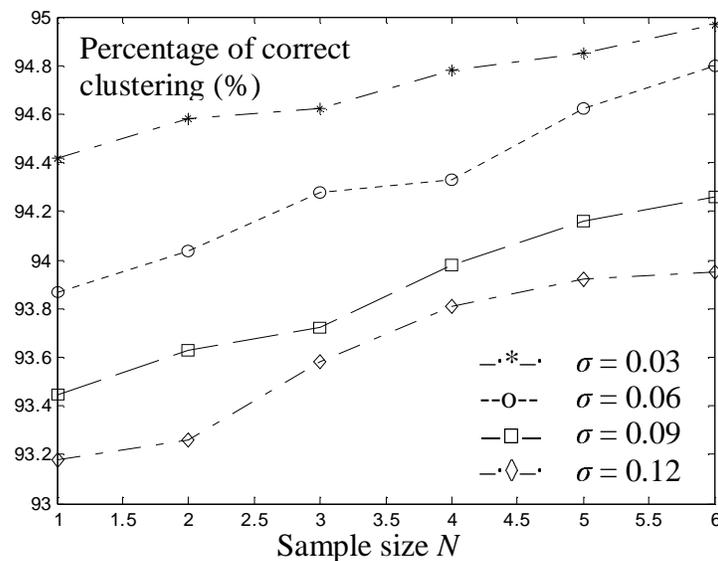
Fig. 43. Percentage of correct clustering of linear and nonlinear blocks with different sample sizes and noise variances for simulation case 2.

Similar Monte Carlo simulations were conducted to provide quantitative performance measure on the proposed nonlinear variation pattern identification algorithm in this case, except that an additional nonlinear variation pattern was

introduced lying in a three-dimensional linear variety. The performance of the proposed

multiple variation patterns identification algorithm was then summarized in Table 6.

The results in Table 5 and 6 agree with the fact that the linear and nonlinear variation

patterns identification accuracy is improved with respect to larger sample size N and/or

smaller noise variance σ.

Table 6. Performance results from the nonlinear BSS algorithm on multiple variation patterns identification for simulation case 2.

N	σ	J_{C_1}	J_{C_2}	D_{f_1}	D_{f_2}
100	0.03	0.039	0.054	0.00038	0.00019
100	0.06	0.047	0.061	0.00147	0.00073
100	0.09	0.056	0.069	0.00407	0.00176
100	0.12	0.068	0.081	0.00747	0.0034
200	0.03	0.039	0.053	0.00037	0.00018
200	0.06	0.046	0.061	0.00144	0.00069
200	0.09	0.055	0.067	0.004	0.00173
200	0.12	0.067	0.079	0.0074	0.00333
300	0.03	0.038	0.053	0.00037	0.00017
300	0.06	0.046	0.059	0.0014	0.00064
300	0.09	0.054	0.066	0.00397	0.00171
300	0.12	0.066	0.078	0.00733	0.00331
400	0.03	0.037	0.052	0.00037	0.00017
400	0.06	0.045	0.058	0.00137	0.00062
400	0.09	0.053	0.065	0.00393	0.00169
400	0.12	0.065	0.078	0.0073	0.00329
500	0.03	0.035	0.051	0.00036	0.00017
500	0.06	0.044	0.057	0.00131	0.00058
500	0.09	0.051	0.064	0.0039	0.00167
500	0.12	0.065	0.077	0.00726	0.00322
600	0.03	0.034	0.049	0.00035	0.00016
600	0.06	0.041	0.056	0.00126	0.00058
600	0.09	0.049	0.063	0.00386	0.00162
600	0.12	0.063	0.076	0.00719	0.0032

IV.6 Chapter Summary

In this chapter, we extended the single nonlinear variation pattern analysis problem by proposing a nonlinear model for representing multiple potential variation patterns that occur simultaneously in the manufacturing processes. A high order cumulants based identification algorithm was then derived to separate the distinct linear and nonlinear variation patterns. When each nonlinear function lies in a lower subspace of the complete observation data, the nonlinear variation pattern model and independence criterion produce a block structured pairwise cumulant matrix for PCA scores. The linear and nonlinear blocks can be separated by a clustering method over the sample data, developed based on the pairwise independence measure. The efficiency of the proposed algorithm on blind identification of multiple linear and nonlinear variation patterns has been demonstrated by some simulation cases and an autobody assembly example.

CHAPTER V

CONCLUSIONS AND FUTURE WORK

V.1 Conclusions

This dissertation has developed a set of related techniques for diagnosing nonlinear spatial variation patterns that exist in the complex manufacturing processes. The objective considered in this dissertation is to identify as precisely as possible the nature of each distinct variation pattern, based entirely on the data sample with no a priori knowledge of the nature of each pattern. Proper identification of the variation patterns can provide much insight into the interdependencies and correlations between the different variables and, ultimately, the phenomena that cause the variation patterns and product variability.

Chapters I reviewed the existing linear variation pattern analysis methods – classifying pre-modeled variation patterns and estimating un-modeled variation patterns. The limitations of linear models for representing the nonlinear manufacturing variations were also discussed through an example from the autobody assembly process. Chapter II and III extended the linear models and identification methods by considering a single nonlinear variation pattern to represent the nonlinearities present in the multivariate measurement data. Chapter II proposed an algorithm of improving accuracy and reducing computational expense in principal curves estimation when they lie in a lower dimensional subspace of the complete data space. The approach involves the use of

standard linear principal component analysis to identify the subspace. The observations are then projected onto this lower dimensional subspace before applying standard principal curve algorithm to identify the nonlinear multivariate variation pattern. It has been investigated that when the nonlinear variation patterns lie in some lower-dimensional linear varieties, the modified principal curves method can be used effectively for variation pattern identification and interpretation. Chapter III developed a polygonal line algorithm for principal curves approximation and nonlinear pattern identification. The main feature differs our approach from the HS algorithm is that it is more robust to the changes in the nonlinear variation patterns. Experimental results demonstrate that the polygonal line algorithm compares favorably to the HS method both in terms of nonlinear pattern identification performance and computational complexity.

To find a more generic method for manufacturing variation analysis, Chapter IV enhanced the single pattern identification approaches by representing and identifying multiple linear and nonlinear variation patterns through a nonlinear BSS model. A high order cumulants based blind pattern identification algorithm was developed to diagnose the nature of the various variation patterns. To help separate and identify the blocks from the sample pairwise cumulant matrix, we present a block clustering method to determine the number of variation sources automatically and estimate individual variation patterns.

Although this dissertation has focused on the statistical aspects of estimating variation patterns, visualization of the statistical results is no less important to

understanding the nature of the variation patterns. The simulation examples in Chapter II and III have been used to illustrate the concepts of variation patterns, in which visualization of the estimated patterns was accomplished with the aid of interactive graphical tools.

The methods developed in Chapters II through IV apply to a fairly broad class of real problems. This work, however, is certainly not complete. A few of these extensions are discussed below as "future work".

V.2 Future Work

Some possible future research directions are listed in the following:

- It is natural to extend the one-dimensional curves (i.e., principal curves) to higher-dimensional manifolds (e.g., principal surfaces) for representing the nonlinear variation patterns that are embedded in the complete data space. Such a more generic variation pattern diagnostic method requires the estimation of the intrinsic dimensionality of the manifolds, as well as the visualization of identified patterns, which will be a challenging problem for future research.

- The discussion on multiple nonlinear variation patterns identification method in this dissertation was restricted to a set of independent time-invariant variation sources. An obvious extension is the inclusion of autocorrelation into the nonlinear model, making use of the temporal information often present in the source signals. Integration of nonlinear model, which have the ability for multiple variation

patterns separation and identification, with sequence or time series models would be useful in improving the description and interpretation of the variation patterns.

- The blind identification methods estimate variation patterns from on-line measurement data, in which case the accuracy of the methods is affected by some factors (e.g., the number of variation source, the sample size). While the accuracy of the estimated variation patterns depends on a number of factors, the pre-modeled variation patterns are not affected by these factors. Pre-modeled variation patterns totally depend on engineering knowledge of the process. If the manufacturing process is understood adequately and pre-modeling of some potential variation patterns is possible, the pre-modeled variation patterns are the most accurate. The extension of the blind separation method in Chapter IV can be studied further. Estimating some linear un-modeled variation patterns with partially pre-modeled patterns may be one likely area for future study.

- Development of nonlinear models with locally varying measurement noise, i.e., where the parameters of the noise covariance matrix may depend on different spatial measurement features. This is desirable because using a fixed isotropic noise covariance does not fit the data well in some areas of data space.

REFERENCES

Apley, D.W. and Lee, H.Y. (2003) Identifying Spatial Variation Patterns in Multivariate Manufacturing Processes: A Blind Separation Approach. *Technometrics*, **45**, 220-235.

Apley, D. W. and Shi, J. (2001) A Factor-Analysis Method for Diagnosing Variability in Multivariate Manufacturing Processes. *Technometrics*, **43**, 84-95.

Apley, D. W. and Shi, J. (1998) Diagnosis of Multiple Fixture Faults in Panel Assembly. *ASME Journal of Manufacturing Science and Engineering*, **120**, 793-801.

Archard, S. and Pham, D.T. (2001) Blind Source Separation in Post Nonlinear Mixtures. In *Proceedings of ICA2001*, San Diego, CA, 295–300.

Avriel, M. (2003) *Nonlinear Programming: Analysis and Methods*, Dover Publications, NY.

Banfield, J.D. and Raftery, A.E. (1992) Ice Floe Identification in Satellite Images Using Mathematical Morphology and Clustering about Principal Curves. *Journal of the American Statistical Association*, **87**, 7-16.

Barton, R.R and Gonzalez-Barreto, D.R (1996) Process-oriented Basis Representation for Multivariate Process Diagnostics. *Quality Engineering*, **9**, 107-118.

Bertsekas, D.P. (1995) *Nonlinear Programming*, Athena Scientific, Belmont, MA.

Cardoso, J.F. (1998) Blind Signal Separation: Statistical Principles. *IEEE Proceedings*, **86**, 2009 –2025.

Ceglarek, D. and Shi, J. (1996) Fixture Failure Diagnosis for Autobody Assembly Using Pattern Recognition. *ASME Journal of Engineering for Industry*, **118**, 55–66.

Chang, K. and Ghosh, J. (2001) A Unified Model for Probabilistic Principal Surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**, 22-41.

Delicado, P. (2001) Another Look at Principal Curves and Surfaces. *Journal of Multivariate Analysis*, **77**, 84-116.

Ding, Y., Ceglarek, D. and Shi, J. (2002) Fault Diagnosis of Multistage Manufacturing Processes by Using State Space Approach. *Journal of Manufacturing Science and Engineering*, **124**, 313-322.

Dong, D. and McAvoy, T.J. (1995) Nonlinear Principal Component Analysis – Based on Principal Curves and Neural Networks. *Computers in Chemical Engineering*, **20**, 65-78.

Gordon, A.D. (1987) A Review of Hierarchical Classification. *Journal of the Royal Statistical Society*, **150**, 119-137.

Glass, S. and Thomsen, J. (1993) How SMT Boards Are Assembled. *Printed Circuit Fabrication*, **16**, 42-47.

Halkidi, M., Vazirgiannis, M. and Batistakis, I. (2000) Quality Scheme Assessment in the Clustering Process. In *Proceedings of PKDD*, Lyon, France.

Halkidi, M. and Vazirgiannis, M. (2001) Clustering Validity Assessment: Finding the Optimal Partitioning of A Data Set. In *Proceedings of ICDM*, San Jose, CA.

Haritopoulos, M., Yin, H. and Allison, N. (2002) Image Denoising using Self-organizing Map-based Nonlinear Independent Component Analysis. *Neural Networks*, **15**, 1085-1098.

Harmeling, S., Ziehe, A., Kawanabe, M. and Muller, K.R. (2003) Kernel-based Nonlinear Blind Source Separation. *Neural Computation*, **15**, 1089-1124.

Hastie, T. and Stuetzle, W. (1989) Principal Curves. *Journal of the American Statistical Association*, **84**, 502-516.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, NY.

Hinton, G.E., Revow, M. and Dayan, P. (1995) Recognizing Handwritten Digits using Mixtures of Linear Models. *Advances in Neural Information Processing Systems*, Chapter **7**, MIT Press, Cambridge, MA.

Hsu, C.W., Ulsoy, A.G. and Demeri, M.Y. (2000) An Approach for Modeling Sheet Metal Forming for Process Controller Design. *Journal of Manufacturing Science and Engineering*, **122**, 717-724.

Hu, S. and Wu, S.M. (1992) Identifying Root Causes of Variation in Automobile Body Assembly Using Principal Component Analysis. *Transactions of NAMRI*, **XX**, 311-316.

Hyvarinen, A. and Oja, E. (2000) Independent Component Analysis: Algorithms and Applications. *Neural Networks*, **13**, 411-430.

Hyvarinen, A., Karhunen, J. and Oja, E. (2001) *Independent Component Analysis*. Wiley, NY.

Jackson, J. E. (1980) Principal Components and Factor Analysis: Part I – Principal Components. *Journal of Quality Technology*, **12**, 201–213.

Jackson, J. E. (1981) Principal Components and Factor Analysis: Part II – Additional Topics Related to Principal Components. *Journal of Quality Technology*, **13**, 46–58.

Jin, J. and Shi, J. (1999) State Space Modeling of Sheet Metal Assembly for Dimensional Control. *ASME Journal of Manufacturing Science and Engineering*, **121**, 756-762.

Johnson, R. A. and Wichern, D. W. (2002) *Applied Multivariate Statistical Analysis (5th ed.)*, Prentice Hall, NJ.

Jutten, C. and Karhunen, J. (2003) Advanced in Nonlinear Blind Source Separation. In *Proceedings of ICA2003*, Nara, Japan, 245-256.

Kambhatla, N. and Leen, T.K. (1997) Dimension Reduction by Local PCA. *Neural Computation*, **9**, 1493-1510.

Kégl, B., Krzyzak, A., Linder. T. and Zeger, K. (1999) A Polygonal Line Algorithm for Constructing Principal Curves. In *Proceedings of Neural Information Processing Systems,* Cambridge, MA, 501-507.

Kim, D.J. and Kim, B.M. (2000) Application of Neural Network and FEM for Metal Forming Processes. *International Journal of Machine Tools & Manufacture*, **40**, 911-925.

Lappalainen, H. and Honkela, A. (2000) Bayesian Nonlinear Independent Component Analysis by Multi-layer Perceptrons. *Advances in Independent Component Analysis*, 93–121. Springer-Verlag, Berlin.

LeBlanc, M. and Tibshirani, R. (1994) Adaptive Principal Surfaces. *Journal of the American Statistical Association*, **89**, 53-64.

Milligan, G.W. and Cooper, M.C. (1985) An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, **50**, 159-179.

Mulier, F. and Cherkassky V. (1995) Self-Organization as an Iterative Kernel Smoothing Process. *Neural Computation*, **7**, 1165-1177.

Pajunen, P. and Karhunen, J. (1997) A Maximum Likelihood Approach to Nonlinear Blind Source Separation. In *Proceedings of ICANN*, Lausanne, Switzerland, 541–546.

Salvador, S. and Chan, P. (2004) Determining the Number of Clusters/Segments in Hierarchical Clustering Algorithms. submitted to *Data Mining and Knowledge Discovery*.

Stuart, A. and Ord, J. K. (1987) *Kendall's Advanced Theory of Statistics (5th ed.)*, Vol. 1, Wiley, NY.

Taleb, A. (2002) A Generic Framework for Blind Source Separation in Structured Nonlinear Models. *IEEE Transactions on Signal Processing,* **50**, 1819-1830.

Taleb, A. and Jutten, C. (1999) Source Separation in Post-Nonlinear Mixtures. *IEEE Transactions on Signal Processing*, **47**, 2807-2820.

Theis, F.J. and Lang, E.W. (2001) Maximum Entropy and Minimal Mutual Information in a Nonlinear Model. In *Proceedings of ICA2001*, San Diego, CA, 669–674.

Tibshirani, R. (1992) Principal Curves Revisited. *Statistics and Computing*, **2**, 183-190.

Tipping, M.E. and Bishop, C.M. (1999) Mixtures of Probabilistic Principal Component Analysers. *Neural Computation*, **11**, 443-482.

Valpola, H., Oja, E., Ilin, A. and Honkela, A. (2003) Nonlinear Blind Source Separation by Variational Bayesian Learning. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences,* **86**, 532 – 541.

Verbeek, J.J., Vlassis, K. and Kröse, B. (2001) A Soft k-segments Algorithm for Principal Curves. In *Proceedings of ICANN2001*, Vienna, Austria.

Wang, X., Cao, Ji., and Li, M. (2001) Wrinkling Analysis in Shrink Flanging. *Journal of Manufacturing Science and Engineering*, **123**, 426-432.

Yang, H.H., Amari, S. and Cichocki, A. (1998) Information-Theoretic Approach to Blind Separation of Sources in Nonlinear Mixture. *Signal Processing*, **64**, 291– 300.

Zhou, S., Ding, Y., Chen, Y. and Shi, J. (2003) Diagnosability Study of Multistage Manufacturing Processes Based on Linear Mixed-Effects Models. *Technometrics*, **45**, 312-325.

APPENDIX A

This appendix summarizes some important properties of high order cumulants, to be used in the linear BSS methods and the nonlinear variation patterns identification algorithm in Chapter IV.

Property A.1 (Scaling) *If y_1, y_2, … y_p are multiplied with constants a_1, a_2, … a_p, then we have*:

$$Cum(a_1 y_1, a_2 y_2, …, a_p y_p) = \left(\prod_{i=1}^{p} a_i \right) Cum(y_1, y_2, …, y_p).$$

Property A.2 (Sum) *Cumulants of a sum are the sum of the cumulants*:

$$Cum(z_1 + y_1, y_2, …, y_p) = Cum(y_1, y_2, …, y_p) + Cum(z_1, y_2, …, y_p),$$

in which z_1 is a real random variable.

Property A.3 (Symmetry) *Cumulants are symmetric on their arguments, i.e.*

$$Cum(y_1, y_2, …, y_p) = Cum(y_{P(1)}, y_{P(2)}, …, y_{P(p)}),$$

in which P *is an arbitrary permutation of* $(1, 2, …, p)$.

Property A.4 (Partitioning of Independent Variables) *If a proper subset of the p random variables y_1, y_2, …, y_p is independent of the other variables, then we have:*

$$Cum(y_1, y_2, …, y_p) = 0.$$

Note that cross-cumulants (cumulants involving at least two different variables) are zero if the elements of *y* are mutually independent. .

Property A.5 (Sum of Independent Variables*) If the random variable y_1, y_2, …, y_p are mutually independent of random variables z_1, z_2, …, z_p, then we have*

$$Cum(y_1 + z_1, y_2 + z_2, \ldots, y_k + z_k) =$$

$$Cum(y_1, y_2, \ldots, y_k) + Cum(z_1, z_2, \ldots, z_k), \ 1 < k \le p \ .$$

Property A.6 (Gaussianity) *Cumulants of order higher than two for Gaussian variable are zero.*

VITA

Name and Permanent Address

Feng Zhang

136 Dongxing Street, Dong Ying City

Shan Dong Province, P.R. China 257500

Chronology of Education

B.S., Automatic Control, Tsinghua University, 1996

M.S., Control Theory and Control Engineering, Tsinghua University, 1999

Major Fields

His research concentrates on developing and applying statistical learning algorithms to identify patterns from large amounts of high dimensional data that reflect the underlying root causes contributing to the overall product variation.

Awards

Wisenbaker Engineering Fellowship, Texas A&M University	1999 - 2000
The Best Thesis Award of Tsinghua University	1999
General Electric (GE) Foundation Fellowship	1997, 1998
Outstanding Tsinghua University Student Scholarship	1992-1997
The 3rd Prize in the 7th Physics Contest for Undergraduate Students	1992