Inferring the Interactions in Complex Manufacturing Processes Using Graphical Models

Li Zeng

Department of Industrial and Systems Engineering
University of Wisconsin
Madison, WI 53706
(*lzengl@wisc.edu*)

Shiyu Zhou

Department of Industrial and Systems Engineering
University of Wisconsin
Madison, WI 53706
(*szhou@engr.wisc.edu*)

The new trends in manufacturing toward modularity and flexibility result in a larger number of interdependent operations in a process, leading to complex multistage manufacturing processes. Identifying the variation flow and implementing quality control in such processes is very challenging because of the complex interactions among different stages. This article presents a systematic model building methodology to identify the underlying interactions among stages through the integration of advanced statistical techniques in graphical models and engineering insights to manufacturing processes. A statistical testing procedure is developed to efficiently construct the chain graph of the key product characteristics in a process, making use of identified relationships at previous stages. A case study validating the effectiveness of the proposed procedure is also presented.

KEY WORDS: Chain graph; *d*-separation; Graphical models; Multistage process; Partial correlation.

1. INTRODUCTION

Various new manufacturing paradigms, including modular production systems, cellular manufacturing, and reconfigure manufacturing, have been developed and adopted in recent years. In general, these new paradigms are aimed toward increasing modularity, flexibility, and self-sufficiency at production floor levels (Ashley 1997). Thus an emerging scenario that is becoming increasingly popular in manufacturing is that the complex operations in the process are divided and grouped into multiple stages, which are interconnected and easily reorganized to provide the production capability for a product family. Figure 1 illustrates a typical car body assembly process as an example of a multistage manufacturing process.

The final product of the car body assembly process is the structural frame of a car, as shown in Figure 1(a). Figure 1(b) shows a simplified diagram of the process, in which a set of subassemblies, including dash, underbody, left and right body sides, and so on, are welded together to form the physical frame, after which closure panels, including roof, doors, hood, and fenders, are mounted on the frame. Due to the product's complexity, there are close to 100 stages in a typical car body assembly process. To monitor the product quality, many key product characteristics (KPCs) on the car body, represented by the deviations of dimensional characteristics from their nominal values as shown in Figure 1(a), are often measured in the process. Again, due to the product's complexity, several hundred KPCs are often measured in a typical auto body assembly process.

Complex multistage manufacturing processes like car body assembly raise significant challenges in process quality control and variation reduction. The main challenge lies in the complex interactions among the KPCs at different stations/stages. At each assembly stage, certain features of the subassembly formed at preceding stages are often used as references to locate the subassemblies at the current stage. Thus the positional/dimensional errors generated at previous stages will influence the quality of the present stage. One can imagine that in a complex manufacturing process, process variation will propagate along the physical process topology and form a network of variation flow. To determine on which stage and what factor on that stage causes the excessive variation in certain KPCs requires identification of the interactions among KPCs at different stages.

Some research efforts have been directed at identifying the interactions among stages of a manufacturing process. These methods can be roughly classified as data-driven techniques, which are based on the statistical analysis of historical process data, and analytical methods, which are based on offline physical model of the process. Existing data-driven techniques include the cause-selecting control charts for a two-step process (Zhang 1985; Wade and Woodall 1993), variation analysis using linear regression and analysis of variance (ANOVA) tools (Lawless, Mackay, and Robinson 1999; Agrawal, Lawless, and Mackay 1999; Fong and Lawless 1998), and the procedure for measuring the influence of each stage's performance on the output quality of subsequent stages (Zantek, Wright, and Plante 2002). The analytical method is represented by recently developed 'stream of variation (SoV)' methodologies that focus on dimensional variation analysis of machining processes (e.g., Huang, Zhou, and Shi 2000; Djurdjanovic and Ni 2001; Zhou, Huang, and Shi 2003b) and assembly processes (e.g., Mantripragada and Whitney 1999; Ding, Ceglarek, and Shi 2000; Jin and Shi 1999). These aforementioned techniques either deal with simple systems with a limited number of stages or require a thorough physical understanding of the process, which might not be generally available. There is a lack of general methodologies for identifying the interactions among the stages of a large-scale, complex process.

TECHNOMETRICS, NOVEMBER 2007, VOL. 49, NO. 4
DOI 10.1198/004017007000000317

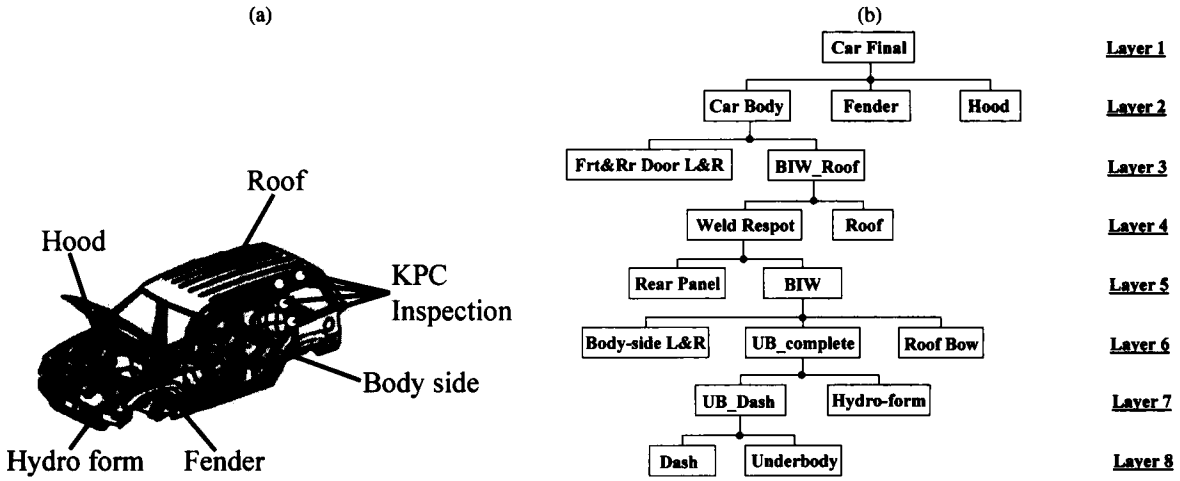

Figure 1. The car body assembly process. (a) The car body. (b) Process tree.

This article proposes a general methodology inferring the underlying interactions among KPCs. In this article a chain graphical model is used to describe the direct influences among KPCs at different stages. The ith KPC directly influences the jth KPC if it uniquely contributes to the variation of the jth KPC. To tackle the issue of high dimensionality, a chain graph (CG) building technique is developed by taking the known process physical layout into consideration and using the relationships identified at previous stages. The proposed methodology is an efficient way to conquer the interstage complexity in manufacturing processes.

The article is organized as follows. Section 2 presents the problem formulation and a review of current methods for constructing chain graphs. Section 3 presents the proposed CG building technique, introducing a theorem on conditioning set simplification that provides the theoretical basis for the proposed technique. Section 4 illustrates the application of the technique through a case study on the car body assembly process presented in Section 1, and Section 5 concludes.

2. PROBLEM FORMULATION AND REVIEW OF GRAPHICAL MODELS

2.1 Problem Formulation

If we define KPCs as nodes, then a general multistage manufacturing process can be described by the layout shown in Figure 2, with q KPCs distributed at n stages. Because our focus is on identifying variation propagation in the process, the means of these KPC variables are assumed to be 0. We use variable X_j to represent the jth KPC in the process and further define $\mathcal{P}_j$ as the set of all KPCs in preceding stages of j. For example, $\mathcal{P}_5$ is $\{1, 2, 3\}$, whereas $\mathcal{P}_q$ includes all of the KPCs except the qth KPC in Figure 2. The physical characteristics of manufacturing processes make it reasonable to assume the following:

(A1) The KPCs at the same stage do not influence each other, and we are concerned only with the identifying the interstage relationships in this article. This assumption is reasonable because KPCs at the same stage are often generated and/or determined simultaneously, and thus they cannot influence each other. In some special cases where some KPCs at the same physical stage are generated sequentially, we can split the physical stage into multiple "artificial" stages and the assumption will still hold.

(A2) The variation of X_j consists of associated local variation, which often is not directly measurable, and the propagated variation contributed by some preceding KPCs to j.

(A3) Let $\mathbf{X} = (X_1, \ldots, X_q)' \in \mathbf{R}^q$ be a random vector including all of the q KPCs in the process; then $\mathbf{X}$ follows a multivariate normal distribution $\mathrm{N}_q(\mathbf{0}, \mathbf{\Sigma})$ with nonsingular covariance matrix $\mathbf{\Sigma}$.

For such a process, the key issue is to identify which preceding KPCs contribute variation to or interact with X_j based on known samples of X_j and X_i, $i \in \mathcal{P}_j$.

To tackle this problem, we define *direct influence* as follows: If X_i, $i \in \mathcal{P}_j$, uniquely contributes to the variation of X_j, then we claim that X_i directly influences X_j. Let $\mathbf{X}_{\mathcal{P}_j} = \{X_k, k \in \mathcal{P}_j\}$ be the vector including all of the KPCs in $\mathcal{P}_j$. To interpret the definition of direct influence mathematically, we split the vector $\mathbf{X}_{\mathcal{P}_j}$ into $\{X_i\}$ and the rest, denoted as $\mathbf{X}_{\mathcal{P}_j} \backslash \{X_i\}$. The linear least squares theory shows that (Whittaker 1990)

$$\operatorname{var}\big(\hat{X}_j(\mathbf{X}_{\mathcal{P}_j})\big) = \operatorname{var}\big(\hat{X}_j(\mathbf{X}_{\mathcal{P}_j} \backslash \{X_i\})\big) + \operatorname{var}\big(\hat{X}_j\big(X_i - \hat{X}_i(\mathbf{X}_{\mathcal{P}_j} \backslash \{X_i\})\big)\big), \quad (1)$$

where $\hat{X}_j(\mathbf{X}_{\mathcal{P}_j}) = \operatorname{cov}(X_j, \mathbf{X}_{\mathcal{P}_j}) \operatorname{var}(\mathbf{X}_{\mathcal{P}_j})^{-1} \mathbf{X}_{\mathcal{P}_j}$ is the linear least squares predictor of X_j from the elements of the set $\mathbf{X}_{\mathcal{P}_j}$ and "var" designates variation. Equation (1) indicates that the

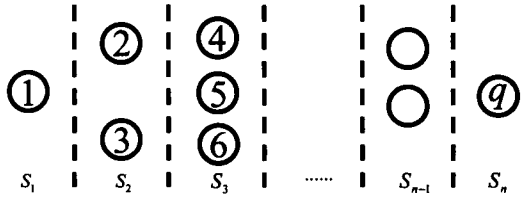

Figure 2. The layout of a general multistage process.

amount of variation in X_j explained by all elements of $\mathcal{P}_j$ can be decomposed into the variation explained by $\mathbf{X}_{\mathcal{P}_j}\backslash\{X_i\}$, the elements in $\mathcal{P}_j$ other than X_i, and $X_i - \hat{X}_i(\mathbf{X}_{\mathcal{P}_j}\backslash\{X_i\})$, X_i adjusted for $\mathbf{X}_{\mathcal{P}_j}\backslash\{X_i\}$. So actually, $\text{var}(\hat{X}_j(X_i - \hat{X}_i(\mathbf{X}_{\mathcal{P}_j}\backslash\{X_i\})))$ represents the unique contribution of X_i to the variation of X_j. Furthermore, $\text{var}(\hat{X}_j(X_i - \hat{X}_i(\mathbf{X}_{\mathcal{P}_j}\backslash\{X_i\})))$ can be simplified to

$$\text{cov}\big(X_j, X_i|\mathbf{X}_{\mathcal{P}_j}\backslash\{X_i\}\big)\,\text{var}\big(X_i|\mathbf{X}_{\mathcal{P}_j}\backslash\{X_i\}\big)^{-1} \times \text{cov}\big(X_i, X_j|\mathbf{X}_{\mathcal{P}_j}\backslash\{X_i\}\big), \quad (2)$$

where $\text{cov}(X_i, X_j|\mathbf{X}_{\mathcal{P}_j}\backslash\{X_i\})$ is also referred to as the *partial covariance* between X_i and X_j given $\mathbf{X}_{\mathcal{P}_j}\backslash\{X_i\}$ (Whittaker 1990). Clearly, if $\text{cov}(X_i, X_j|\mathbf{X}_{\mathcal{P}_j}\backslash\{X_i\})$ is nonzero, then (2) also is nonzero, meaning that X_i has a unique contribution to the variation of X_j. But because $\mathbf{X}$ is normally distributed [(A3)], $\text{cov}(X_i, X_j|\mathbf{X}_{\mathcal{P}_j}\backslash\{X_i\})$ or $\text{corr}(X_i, X_j|\mathbf{X}_{\mathcal{P}_j}\backslash\{X_i\})$ (called *partial correlation*) is 0 iff X_i and X_j are independent given all of the remaining variables in $\mathcal{P}_j$, written as $i \perp\!\!\!\perp j|\mathcal{P}_j\backslash\{i\}$ (Whittaker 1990). In other words, if X_i has no direct influence on X_j, then it also is conditionally independent with X_j, and vice versa.

In statistical science, graphical models, also called *conditional independence graphs*, are often used to describe the conditional independent relationships among a set of random variables (e.g., Whittaker 1990; Cox and Wermuth 1993, 1996; Lauritzen 1996; Jordan 1998). Based on the foregoing analysis, it is clear that identifying the direct influential relationships of KPCs in a process is equivalent to constructing a CG, which is a special kind of graphical model. The next section provides a brief review of the theories and notations of graphical models for the sake of completeness. Readers familiar with graphical models can jump to Section 2.3 directly.

2.2 Brief Review of Graphical Models

An independence graph is a pair $\mathcal{G} = (V, E)$ in which V, a set of nodes, denotes a set of random variables $\mathbf{X}$ (with the jth random variable represented as X_j) and E is a set of edges between these nodes. There is no edge between two nodes i and j, $i, j \in V$, iff X_i and X_j are conditionally independent given others, written as $i \perp\!\!\!\perp j|V\backslash\{i,j\}$. The edge between i and j can be either directed from one to the other or undirected. In particular, a CG is a class of graphs in which the node set V is partitioned into numbered subsets, called blocks, such that all edges between nodes in the same subset are undirected and all edges between different subsets are directed, pointing from the set with the lower number to the one with the higher number. If we treat the stages as blocks and a directed edge as a direct influence, then a CG exactly contains the relationships of interest. Also note that the graph built in this article is a special kind of CG that, according to (A1), contains no edge within blocks (stages). Thus it also can be viewed as a special *directed graph*, a class of graphs in which all edges are directed.

We next introduce the following notations related to directed graphs. In a directed graph, if there is an edge from node i to j, then we say that i is a *parent* of j and j is a *child* of i. A *path* of length n from i to j is a sequence, $i_0 = i, i_1, \ldots, i_n = j$, of distinct nodes such that $i_{k-1} \to i_k$ for all $k = 1, \ldots, n$. A *cycle* of length n is a path in which the first and last nodes are identical (i.e., $i_0 = i_n$). A *trail* of length n from i to j is a sequence, $i_0 = i$, $i_1, \ldots, i_n = j$, of distinct nodes such that $i_{k-1} \to i_k$ or $i_k \to i_{k-1}$ for all $k = 1, \ldots, n$. Thus moving along a trail could go against the direction of the arrows, in contrast to the case of a path.

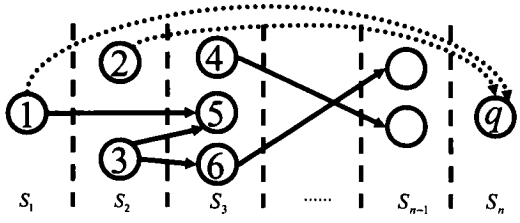

Figure 3. Available relationships identified at $S_1 \sim S_{n-1}$.

A directed graph that contains no cycles is called a *directed acyclic graph* (DAG), designated $\mathcal{D}$. The graph that we want to build is actually a DAG according to the physical characteristics of a manufacturing process. In a DAG, if there is a path from i to j, then i is called an *ancestor* of j and j is called a *descendant* of i.

The global Markov property and *d-separation* criterion are very important properties of DAGs (Lauritzen 1996). Letting A, B, and Q be disjoint subsets of nodes in a DAG, according to this criterion, $A \perp\!\!\!\perp B|Q$ iff Q d-separates A and B, and Q d-separates A and B if there is no active trail from A to B given Q or if all trails from A to B are blocked by Q. (Here a trail from A to B means a trail from any node in A to a node in B.) A trail is active given Q if (a) for every node γ on this trail at which arrows of the trail do not meet head-on, $\gamma \notin Q$, and (b) for every node γ at which arrows of the trail meet head-on, either γ or any descendant of γ is in Q. A trail that is not active given Q is also said to be blocked by Q. For example, considering the trail from node 1 to node 3 in Figure 3 and conditioning set $Q = \{5\}$, arrows of this trail meet head-on at 5, the only (intermediate) node on the trail, and $5 \in Q$, meaning that both (a) and (b) are satisfied. Consequently, we claim that this trail is active, and thus 1 and 3 are not independent given 5.

2.3 Challenges in Chain Graph Construction for Manufacturing Processes

The commonly used technique for CG construction is a stepwise model selection procedure that identifies the dependence of X_j, $j = 1, \ldots, q$, on its preceding KPCs sequentially (e.g., Edwards 2000). Specifically, in the jth step, the conditional independent relationship between X_j and each X_i, $i \in \mathcal{P}_j$, is tested, and if the result shows that X_i and X_j are not conditionally independent, then a directed edge is drawn from i to j; otherwise, no edge exists between them. In this way, the CG is built gradually from the leftmost stage to the last stage. It is also worth mentioning that the test for conditional independence in the procedure can be based on maximum likelihood estimates in submodels (e.g., Neapolitan 2004), as well as on sample partial correlations.

Another method was recently proposed in which all of the relationships in the process are tested simultaneously (Drton and Perlman 2005; Andersson, Madigan, and Peerlman 2001). According to this method, constructing CGs of the process is equivalent to conducting simultaneous hypothesis tests

$$\begin{aligned} &H_{ij}: \rho_{ij|\mathcal{P}_j\backslash\{i\}} = 0 \qquad \text{versus} \\ &K_{ij}: \rho_{ij|\mathcal{P}_j\backslash\{i\}} \neq 0, \qquad 1 \le i < j \le q \text{ and } k(i) < k(j), \end{aligned} \quad (3)$$

where $k(i)$ and $k(j)$ are the indices of the stages that contain KPC i and j. [Here $\rho_{ij|\mathcal{P}_j\backslash\{i\}}$ is short for $\text{corr}(X_i, X_j|X_{\mathcal{P}_j}\backslash\{X_i\})$.] Similarly, if H_{ij} is rejected, then a directed edge is drawn from i to j.

The foregoing methods do not fully use the identified relationships at previous stages. In both methods, all of the elements in $\mathcal{P}_j\backslash\{i\}$, regardless of their relationships that have been identified through previous tests, are included in the test of the conditional independence between i and j. Thus possible redundancy may result in the conditioning set, because not every element in $\mathcal{P}_j\backslash\{i\}$ contains information explaining the dependence between X_i and X_j. This redundancy may lead to a large number of variables involved in the testing and consequently not only add more difficulty to the procedure, but also degrade the power of the tests (Drton and Perlman 2005). This point becomes especially clear when complex manufacturing processes comprising numerous variables are considered.

To solve this problem, we develop a methodology that can reduce the redundancy in the conditioning set through fully exploiting relationships that have been identified at previous stages. Basically, we follow the similar sequential method as in the stepwise procedure to examine the relationships; that is, starting from the leftmost stage, the relationships between X_i, $i \in S_1 \sim S_{k-1}$, and X_j, $j \in S_k$, for $k = 2, \ldots, n$, are identified in sequence. But unlike in the stepwise procedure, at each step, because relationships among the variables at $S_1 \sim S_{k-1}$ and those between X_t, $t < i$, and X_j are already known, they are used to reduce the complexity in the subsequent testing. For example, in the situation illustrated in Figure 3, with the relationships of the variables at $S_1 \sim S_{n-1}$ that are already known, we do not need to condition on any variable when testing the relationship between X_2 and X_q. Similarly, when testing the relationship between X_1 and X_q, we need only condition on X_3 and X_5 instead of all of the remaining $q - 2$ variables. The simplified conditioning set has a much smaller dimension but contains all of the variables needed to explain the dependence between X_i and X_j.

It should be pointed out that some previous attempts have been made to reduce the dimension of the conditioning set. The reduction is considered in the stepwise procedure, making use of the identified relationships between X_t, $t < i$, and X_j in testing the conditional independence between X_i and X_j (e.g., Edwards 2000). Clearly, the identified relationships among the variables at $S_1 \sim S_{k-1}$ are not used, accounting for the major part of the available relationships. In contrast, the methodology proposed in this article can use all of the available relationships identified in previous tests. The rationale for the reduction is thoroughly studied, and a systematic procedure for conditioning set simplification is developed. Some authors (e.g., Drton and Perlman 2005) also have considered conditioning set simplification in a simultaneous procedure by incorporating previous knowledge about the presence or absence of edges. In this method, a subset $\mathcal{P}'(i,j) \subseteq \mathcal{P}_j\backslash\{i\}$ that satisfies $\rho_{ij|\mathcal{P}_j\backslash\{i\}} = \rho_{ij|\mathcal{P}'(i,j)}$ is identified based on previous knowledge, and thus testing $\rho_{ij|\mathcal{P}_j\backslash\{i\}} = 0$ is equivalent to testing $\rho_{ij|\mathcal{P}'(i,j)} = 0$. In this way, the conditioning set in the testing can be simplified if $\mathcal{P}'(i,j)$ is smaller than $\mathcal{P}_j\backslash\{i\}$. In the present article, however, we identify a subset $\mathcal{R}(i,j) \in \mathcal{P}_j\backslash\{i\}$ based on test results at previous stages where $\mathcal{R}(i,j)$ satisfies the condition that $\rho_{ij|\mathcal{P}_j\backslash\{i\}} = 0$ is equivalent to $\rho_{ij|\mathcal{R}(i,j)} = 0$. Knowing $\mathcal{R}(i,j)$, we can likewise transform the null hypothesis $\rho_{ij|\mathcal{P}_j\backslash\{i\}} = 0$ to $\rho_{ij|\mathcal{R}(i,j)} = 0$. Because we do not require $\rho_{ij|\mathcal{P}_j\backslash\{i\}} = \rho_{ij|\mathcal{R}(i,j)}$, more aggressive reduction in the conditioning set often can be achieved. Furthermore, the reduction is based on testing results at previous stages, not on the physical knowledge that might not be always available. In the next section we present the proposed methodology for constructing the CG of a process with detailed procedure to identify the simplified conditioning set $\mathcal{R}(i,j)$.

3. CHAIN GRAPH BUILDING AND CONDITIONING SET SIMPLIFICATION FOR MANUFACTURING PROCESSES

To build the CG of a process, our proposed procedure starts at the leftmost stage and examines the direct influential relationships between node j, $1 < j \le q$, and each $i \in \mathcal{P}_j$ in the order of stages. Using the process in Figure 2 as an example, in the first step, the relationships between nodes 1 and 2, and nodes 1 and 3 are examined. Then node 4 is taken into consideration, examining the relationships among nodes 4 and 1, nodes 4 and 2, and nodes 4 and 3, and so on, up to the last node in the process, q. The relationship between node i and j is determined by testing the hypothesis

$$\begin{aligned} &H'_{ij}: \rho_{ij|\mathcal{R}\{i,j\}} = 0 \qquad \text{versus} \\ &K'_{ij}: \rho_{ij|\mathcal{R}(i,j)} \neq 0, \qquad 1 \le i < j \le q, \end{aligned} \tag{4}$$

and a directed edge is drawn from i to j if H'_{ij} is rejected. In (4), the simplified conditioning set $\mathcal{R}(i,j)$ is a subset of $\mathcal{P}_j\backslash\{i\}$ such that $\rho_{ij|\mathcal{P}_j\backslash\{i\}} = 0$ is equivalent to $\rho_{ij|\mathcal{R}(i,j)} = 0$.

3.1 Identifying the Simplified Conditioning Set $\mathcal{R}(i,j)$

One critical step in the foregoing procedure is identifying the set $\mathcal{R}(i,j)$. Using the tool of d-separation in graphical models, a theorem can be derived to provide the theoretical foundation for identifying $\mathcal{R}(i,j)$.

Theorem. Let A, B, and Q be disjoint subsets of the node set V of a directed, acyclic graph $\mathcal{D} = (V, E)$, and let Q_1 be a subset of Q. $\mathbf{X}_A$, $\mathbf{X}_B$, and $\mathbf{X}_Q$ are the random variables denoted by A, B, and Q, which have a positive joint density with respect to a product measure. If either of the following conditions is satisfied:

(I) Q_1 d-separates A and $Q\backslash Q_1$, and $\mathbf{X}_A$, $\mathbf{X}_B$, and $\mathbf{X}_Q$ follow a joint normal distribution, or

(II) $Q_1 \cup B$ d-separates A and $Q\backslash Q_1$,

then $A \perp\!\!\!\perp B|Q \Leftrightarrow A \perp\!\!\!\perp B|Q_1$.

The proof is given in the Appendix. This theorem implies that testing the conditional independent relationship between A and B given Q is equivalent to testing that given Q_1. In other words, in this testing, the conditioning set Q can be equivalently reduced to a subset of Q. Also note that the graphical condition (I) holds only for Gaussian distributions, whereas condition (II) holds for all distributions with positive joint density. Therefore, both conditions can be applied in the situation considered in this article.

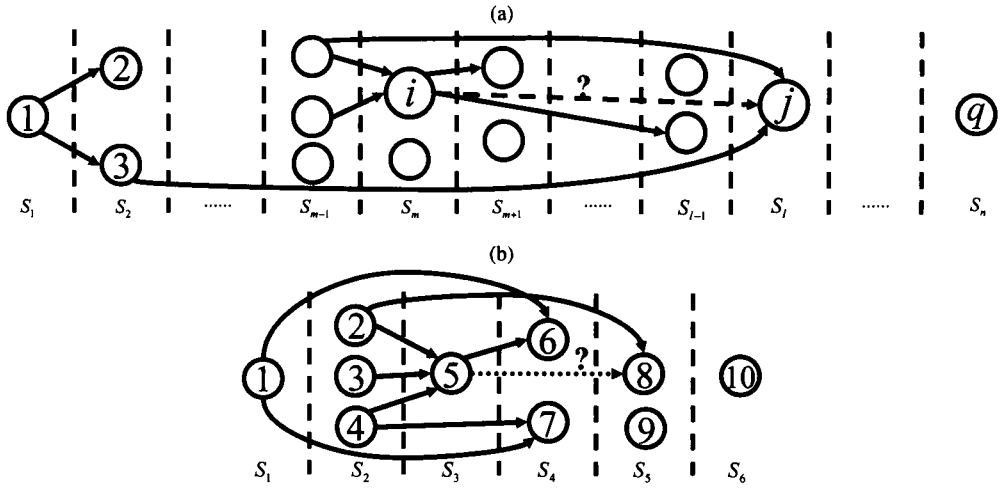

Figure 4. Available relationships in identifying the relationship between i and j. (a) The general case. (b) An example where $i = 5, j = 8$.

The two graphical conditions can be intuitively interpreted in terms of the reduced part of the conditioning set, $Q \backslash Q_1$. Essentially, this says that if there is no dependence between some variables of the conditioning set (i.e., $Q \backslash Q_1$) and A given the remaining variables in the conditioning set (i.e., Q_1) or all of the other variables (i.e., $Q_1 \cup B$), then these variables contain no information explaining the dependence between A and B, and thus they can be removed from the conditioning set. In other words, generally speaking, only those variables that have dependence with both A and B given others should be kept in the conditioning set. Based on this theorem, a heuristic procedure for the conditioning set simplification can be established as follows.

Procedure for Identifying $\mathcal{R}(i,j)$. Figure 4(a) shows a general case of the available relationships when the partial correlation between nodes i and j, at S_m and S_l, is to be tested. In the figure, all of the relationships among the variables in $\mathcal{P}_j$ and those between j and $1, 2, \ldots, i-1$ are known according to previous tests, whereas those between j and $t \in \mathcal{P}_j$, $t \geq i$, have not been tested. Figure 4(b) is an example of a specific process where $q = 10$, $m = 3$, $l = 5$, $i = 5$, and $j = 8$. Later we use this example to demonstrate the procedure for identifying $\mathcal{R}(i,j)$. The four steps of the procedure are as follows:

Step 1: Obtain the saturated graph $\mathcal{G}(i,j)$. The saturated graph can be easily obtained by first removing other nodes in j's stage and nodes in later stages (thus keeping only nodes in $\mathcal{P}_j \cup \{j\}$) and then adding edges to all of the untested pairs. For the example shown in Figure 4(b), nodes 9 and 10 should be removed, and edges should be added to the three untested pairs, $\{5, 8\}$, $\{6, 8\}$, and $\{7, 8\}$. The resulting saturated graph, $\mathcal{G}(i,j)$, is shown in Figure 5. This graph is constructed for two reasons. First, because our focus is the relationships within $\mathcal{P}_j \cup \{j\}$, the nodes in the same stage as j and those in later stages can be ignored, because they will not influence these relationships in a manufacturing process. Second, because the edges between the untested pairs are unknown at the moment, it is natural and safe to assume that they exist.

Step 2: $\mathcal{P}_j \backslash \{i\} \rightarrow C \cup M$. In $\mathcal{G}(i,j)$, without considering node j and the edges pointing to it (as if j is removed from the graph), the subset M is defined as i's parents and those that have common children with i, whereas C is defined as i's children. Actually, the set $M \cup C$ is referred to as the *Markov blanket* of i (Lauritzen 1999). For example, Figure 5 readily shows that $M = \{1, 2, 3, 4\}$ and $C = \{6\}$, as shown in Figure 6(a). Here nodes 2, 3, and 4 in M are parents of 5, whereas node 1 has a common child, node 6, with 5.

It is known that the Markov blanket of i (i.e., $C \cup M$) d-separates i and $\mathcal{P}_j \backslash \{i\} \backslash (C \cup M)$ (Lauritzen 1999). According to (I) of the Theorem (let $A = \{i\}$, $B = \{j\}$, $Q = \mathcal{P}_j \backslash \{i\}$, and $Q_1 = C \cup M$), $\mathcal{P}_j \backslash \{i\}$ can be equivalently reduced to $C \cup M$. Note that variables in C always should be kept in the conditioning set, because directed edges exist between each of them and j. This is why we should only consider the division of M in the next step.

Step 3: $C \cup M \rightarrow C \cup M_1 \cup M_2$. In $\mathcal{G}(i,j)$, considering the trails from the nodes in M to j, M can be divided into three subsets, M_0, M_1, and M_2, as defined here. For each $k \in M$:

- If there is an edge from k to j, then $k \in M_1$, and we say that there is a *direct trail* from k to j.
- If there is no direct trail from k to j, but given $\{i\} \cup C \cup (M \backslash \{k\})$, there is an active trail from k to j, then $k \in M_2$ and we say that there is an *indirect trail* from k to j.
- If there is neither a direct nor an indirect trail from k to j, then $k \in M_0$ and we say that there is no active trail from k to j.

By the foregoing definitions, for the example, $M_1 = \{2\}$, $M_2 = \{1, 4\}$, and $M_0 = \{3\}$, as shown in Figure 6(b). Note

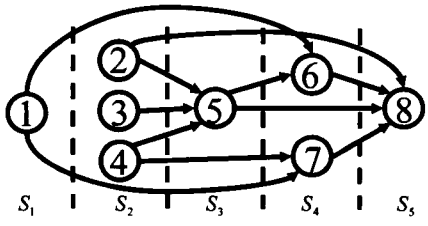

Figure 5. The identified saturated graph.

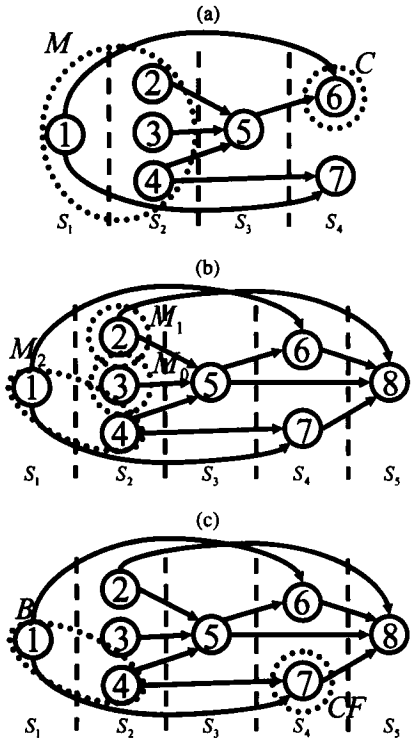

Figure 6. The identified important subsets in $\mathcal{G}(i,j)$: (a) M and C; (b) M_0, M_1, and M_2; (c) B and CF.

that the indirect trail from 1–8 is the trail containing only one intermediate node, 7, which is not in the conditioning set $\{5\} \cup C \cup (M \backslash \{1\})$ (i.e., $\{2, 3, 4, 5, 6\}$), and at which the arrows of the trail do not meet head-on. According to the definitions of M_0, M_1, and M_2, it is easy to get that $C \cup M_1 \cup M_2$ d-separates j and M_0. By (II) of the Theorem (let $B = \{i\}$, $A = \{j\}$, $Q = C \cup M$, and $Q_1 = C \cup M_1 \cup M_2$), $C \cup M$ can be equivalently reduced to $C \cup M_1 \cup M_2$.

Step 4: $C \cup M_1 \cup M_2 \rightarrow C \cup M_1 \cup (M_2 \backslash B) \cup CF$ whenever B and CF exist. In $\mathcal{G}(i,j)$, denoting the set of (intermediate) nodes on the indirect trails from M_2 to j as F, sometimes it also is possible to find a subset $B \subseteq M_2$ and a corresponding subset $CF \subseteq F$ that smaller than B (i.e., $|B| - |CF|$ is positive, where $|\cdot|$ is the number of nodes in a set), such that if CF is added into the conditioning set, then (a) the indirect trails from B to j are blocked [i.e., given $\{i\} \cup C \cup (M \backslash B) \cup CF$, there is no longer an active trail from B to j], and (b) the classifications of other nodes in M will not be changed [i.e., given $\{i\} \cup C \cup (M \backslash \{M_2 \backslash B\}) \cup CF$, there is still an active trail from $M_2 \backslash B$ to j, and given $\{i\} \cup C \cup (M \backslash M_0) \cup CF$, there is no active trail from M_0 to j]. (The direct trails from M_1 to j always exist.) When the satisfying pair of subsets, B and CF, is not unique, the one with maximized $|B| - |CF|$ will be chosen. In simple cases, B and CF can be identified through manual inspection. For example, in Figure 5, it is easy to get that $B = M_2 = \{1, 4\}$ and $CF = \{7\}$. The two sets are shown in Figure 6(c).

In the same way as in Step 2, we can prove that $\mathcal{P}_j \backslash \{i\}$ also can be equivalently reduced to $C \cup M \cup CF$. Moreover, B and CF are defined such that $C \cup M_1 \cup (M_2 \backslash B) \cup CF$ d-separates j and $M_0 \cup B$. According to (II) of the theorem [let $B = \{i\}$, $A = \{j\}$, $Q = C \cup M \cup CF$, and $Q_1 = C \cup M_1 \cup (M_2 \backslash B) \cup CF$], $C \cup M \cup CF$, and thus $\mathcal{P}_j \backslash \{i\}$ can be equivalently reduced to $C \cup M_1 \cup (M_2 \backslash B) \cup CF$.

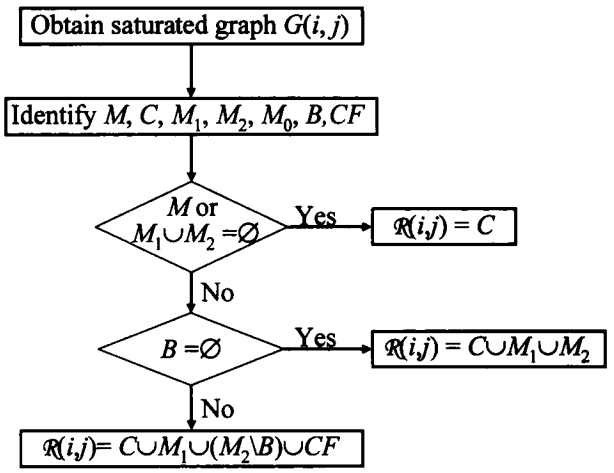

Figure 7. The procedure to identify $\mathcal{R}(i,j)$.

Based on these four steps, we obtain that generally $\mathcal{R}(i,j) = C \cup M_1 \cup (M_2 \backslash B) \cup CF$. The expression can be simplified under various special conditions:

- If $M = \emptyset$ or $M_1 \cup M_2 = \emptyset$, then $\mathcal{R}(i,j) = C$.
- If $M_1 \cup M_2 \neq \emptyset$ and $B = \emptyset$, then $\mathcal{R}(i,j) = C \cup M_1 \cup M_2$.
- If $B \neq \emptyset$, then $R(i,j) = C \cup M_1 \cup (M_2 \backslash B) \cup CF$.

It is worth pointing out that sometimes, if identifying B and CF proves difficult, then a simplified and faster procedure can be just identifying M, C, M_1, M_2, M_0 and take $C \cup M_1 \cup M_2$ as $\mathcal{R}(i,j)$. Clearly, the resulting $\mathcal{R}(i,j)$ might not be the smallest conditioning set, but it often is significantly more simplified than the original one, $\mathcal{P}_j \backslash \{i\}$. The entire procedure is summarized in the flow chart shown in Figure 7.

3.2 The Partial Correlation Test

After $\mathcal{R}(i,j)$ is identified, testing on $\rho_{ij|\mathcal{R}(i,j)}$ can be conducted. Assume that $r_{ij|\mathcal{R}(i,j)}$ is the corresponding sample partial correlation, N is the sample size, and $k(i,j)$ denotes the number of variables in $\mathcal{R}(i,j)$. Let

$$z_{ij|\mathcal{R}(i,j)} = \frac{1}{2} \log \frac{1 + r_{ij|\mathcal{R}(i,j)}}{1 - r_{ij|\mathcal{R}(i,j)}} \quad \text{and}$$

$$\varsigma_{ij|\mathcal{R}(i,j)} = \frac{1}{2} \log \frac{1 + \rho_{ij|\mathcal{R}(i,j)}}{1 - \rho_{ij|\mathcal{R}(i,j)}}.$$

We know that (Anderson 2003)

$$\sqrt{N - 3 - k(i,j)}\left(z_{ij|\mathcal{R}(i,j)} - \varsigma_{ij|\mathcal{R}(i,j)}\right) \sim \mathrm{N}(0,1), \quad \text{as } N \rightarrow \infty,$$

where $\mathrm{N}(0,1)$ is the standard normal distribution. For the test in (4), the null distribution is

$$\sqrt{N - 3 - k(i,j)}\, z_{ij|\mathcal{R}(i,j)} \sim \mathrm{N}(0,1).$$

Thus the following statistic is used:

$$w_{ij} = \sqrt{N - 3 - k(i,j)}\, z_{ij|\mathcal{R}(i,j)},$$

and H'_{ij} is rejected if $|w_{ij}| > \Phi^{-1}(1 - \alpha/2)$, where Φ^{-1} is the inverse cumulative distribution function of $\mathrm{N}(0,1)$ and α is the specified type I error.

4. CASE STUDY

Here we apply the proposed methodology to the car body assembly process described in Section 1. For this typical car body assembly process, approximately 40 stages and 390 KPCs are measured and inspected. Because of the process's complexity, it is especially important to identify the direct influences among the KPCs. For simplicity, we selected 14 KPCs to validate the proposed methodology. These KPCs denote the position deviations of some important features. The physical layout of these KPCs in the process and in the case study are illustrated in Figures 8(a) and 8(b). The assembly process is simulated in 3DCS, a commercially used dimensional variation simulation software for assembly and machining processes. This software is based on the governing physical laws in joining to simulate the assembly process and is widely used in practice for dimension management. The simulated measurement data set is available from the authors on request. In the simulation, the sample size was set to $N = 100$. The type I error for each test was set at $\alpha = .005$, and the critical value was calculated as 2.807 (see Sect. 3.2).

The constructed CG is shown in Figure 9. Table 1 lists the test results of the eight direct influential relationships identified among the 14 KPCs, with "corr" denoting partial correlation. It is clear that using the proposed methodology greatly simplifies the conditioning sets involved in the tests. Many of these sets include no variable, and, consequently, the corresponding partial correlations degrade to simple correlations.

It is also verified that all of these relationships can be interpreted in terms of their physical interactions. For example, the direct influences among the 4th KPC (X_4) the 9th KPC (X_9), and the 10th KPC (X_{10}) which denote the x-direction deviations of three points on the flush surface of the right body side, the right rear edge of the roof, and the outer surface of the rear door, can be clearly confirmed by the physical steps in the process, as shown in Figure 10. Figure 10(a) shows that to form the car

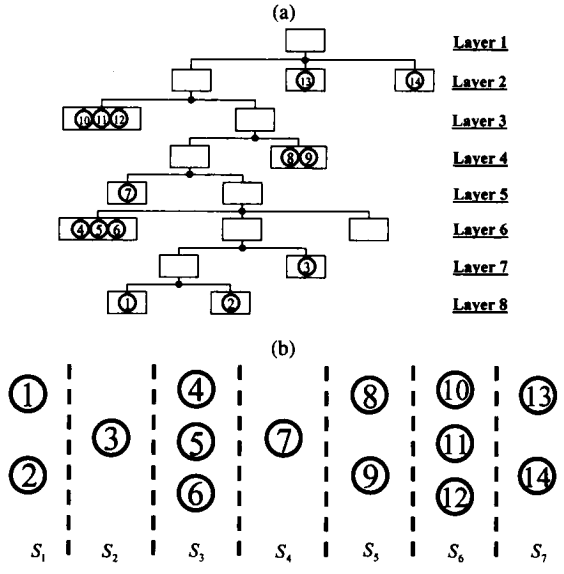

Figure 8. Physical layout of the selected KPCs. (a) Selected KPCs and layout in the process. (b) KPC layout in the case study.

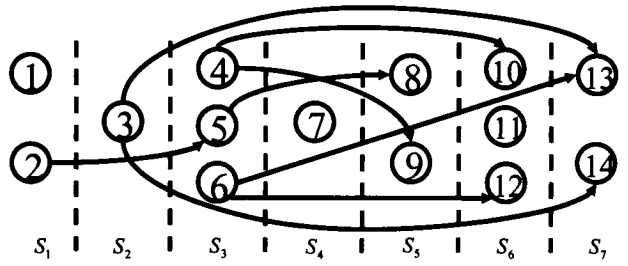

Figure 9. The constructed chain graph of the car body assembly process.

frame, the left and the right body sides are welded on the underbody, after which the roof bows are added by the fixtures located on the body sides. Once the car frame is finished, the roof is welded in place by the fixtures on roof bows, as illustrated in Figure 10(b), and rear door is installed on the frame by hinges attached on the body sides after roofing, as in Figure 10(c). Because of the critical role of body sides in deciding the positions of roof and rear door, it is easy to understand the two direct influences from X_4 to X_9 and X_{10}. The physical steps also explain the nonexistence of a direct influential relationship between X_9 and X_{10}, because it is clearly shown that the influence from X_9 to X_{10} is due to X_4.

5. SUMMARY AND DISCUSSION

In this article we have presented a methodology to conquer the interstage complexity in manufacturing processes with complex topologies. A statistical testing procedure has been developed to construct the CG representing the direct influential relationships among KPCs. The proposed procedure can significantly reduce the redundancy in testing and thus improve the detection power.

Several interesting open issues remain in the proposed methodology. First, we have developed the CG building procedure under assumption (A1). In practice, there are cases in which (A1) does not hold, that is, there are direct influential relationships among variables at the same stage. One way to handle this situation is to "artificially" separate the KPCs into different stages based on previous process knowledge, as described in (A1). When previous knowledge is not available, the proposed method also can be extended by assuming that there are undirected edges between any two nodes at the same stage and changing the definition of $\mathcal{P}_j$ to be all of the variables at j's stage and preceding stages. Moreover, because the graph

Table 1. Test results of the identified relationships

Test	Partial correlation	Statistic
1	corr(2, 5)	−3.0717
2	corr(5, 8)	−3.3176
3	corr(4, 9)	7.8978
4	corr(4, 10|9)	8.1556
5	corr(6, 12)	4.6126
6	corr(3, 13)	7.2648
7	corr(6, 13|12)	−3.3438
8	corr(3, 14)	3.7819

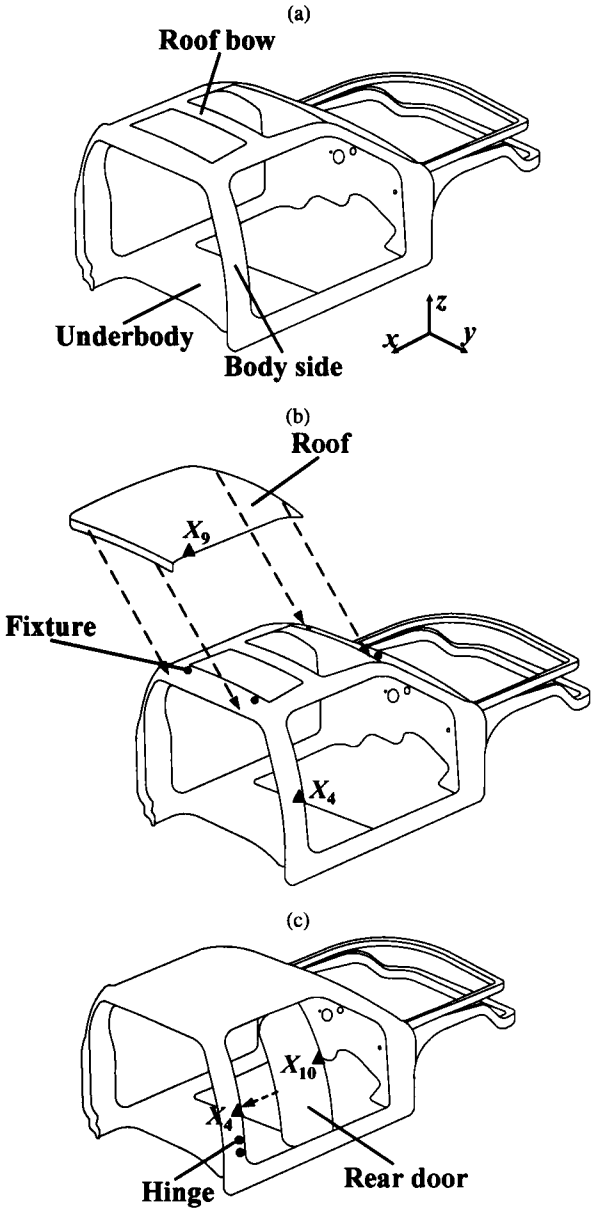

Figure 10. Interpretation of two identified direct influences. (a) The car frame. (b) Roofing process. (c) Rear door installation.

now includes both directed and undirected edges, the concept of *d*-separation defined for DAGs can no longer be used. Instead, some analogous separation criterion (Bouckaert and Studený 1995) or equivalent idea, such as *moral graph* (Lauritzen 1996), is used in identifying the simplified conditioning set. But this extension is not optimal considering the complexity of the resulting procedure, and further research is needed. The second open issue in the proposed method is the identification of sets B and CF presented in Section 3.1. In simple cases, these two sets can be identified by human inspection; however, when the set of M_2 and the corresponding set F are quite large, an efficient algorithm is needed to identify B and CF. Finally, the overall false-alarm rate and detection power of the proposed procedure are also worth examining. Due to the iterative nature of the proposed technique, the errors (especially type II errors corresponding to missing edges) will propagate in the iterative testing process and thus affect the overall accuracy of the procedure. It is interesting, yet challenging to investigate the major factors that influence the overall false-alarm rate and detection power. This is our current research, and we will report the results in the near future.

ACKNOWLEDGMENT

Financial support for this work is provided by National Science Foundation grant CMMI-0545600. The authors gratefully appreciate the editor and the referee for their valuable comments and suggestions.

APPENDIX: PROOF OF THE THEOREM

Let X, Y, Z, and W be random vectors. The following properties of conditional independence are used:

(P1) $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp W|Y, Z \Leftrightarrow X \perp\!\!\!\perp (W, Y)|Z$
(P2) $X \perp\!\!\!\perp Y|W, Z$ and $X \perp\!\!\!\perp W|Y, Z \Leftrightarrow X \perp\!\!\!\perp (W, Y)|Z$
(P3) $X \perp\!\!\!\perp (W, Y)|Z \Rightarrow X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp W|Z$
(P4) $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp W|Z \Rightarrow X \perp\!\!\!\perp (W, Y)|Z$.

Properties (P1)~(P3) hold whenever X, Y, Z, and W have a positive joint density with respect to a product measure. (P4) is true whenever X, Y, Z, and W are jointly normal (Drton and Perlman 2005; Whittaker 1990).

Let us first consider condition (I). Given $Q\backslash Q_1 \perp\!\!\!\perp A|Q_1$, if $A \perp\!\!\!\perp B|Q$, then, by (P1),

$$A \perp\!\!\!\perp \{B, Q\backslash Q_1\}|Q_1.$$

Applying (P3) leads to

$$A \perp\!\!\!\perp B|Q_1.$$

If, on the other hand, $A \perp\!\!\!\perp B|Q_1$, then by (P4), $A \perp\!\!\!\perp \{B, Q\backslash Q_1\}|Q_1$ will result once again. By (P2),

$$A \perp\!\!\!\perp \{B, Q\backslash Q_1\}|Q_1 \Rightarrow A \perp\!\!\!\perp B|Q\backslash Q_1, Q_1 = A \perp\!\!\!\perp B|Q.$$

Thus (I) holds.

Similarly, we can obtain that, given $Q\backslash Q_1 \perp\!\!\!\perp A|Q_1, B$,

$$A \perp\!\!\!\perp B|Q \overset{\text{P2}}{\Rightarrow} A \perp\!\!\!\perp \{B, Q\backslash Q_1\}|Q_1 \overset{\text{P3}}{\Rightarrow} A \perp\!\!\!\perp B|Q_1.$$

On the other hand,

$$A \perp\!\!\!\perp B|Q_1 \overset{\text{P1}}{\Rightarrow} A \perp\!\!\!\perp \{B, Q\backslash Q_1\}|Q_1 \overset{\text{P2}}{\Rightarrow} A \perp\!\!\!\perp B|Q.$$

Thus condition (II) holds.

It is worth noting that condition (I) depends on all four properties and thus holds only for normal joint distributions. Condition (II) needs only the properties (P1)~(P3) and thus can be applied to any joint distribution with positive density.

[Received July 2005. Revised May 2007.]

REFERENCES

Agrawal, R., Lawless, J. F., and Mackay, R. J. (1999), "Analysis of Variation Transmission in Manufacturing Processes, Part II," *Journal of Quality Technology*, 31, 143–154.

Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis* (3rd ed.), Hoboken, NJ: Wiley.

Andersson, S. A., Madigan, D., and Peerlman, M. D. (2001), "Alternative Markov Properties for Chain Graphs," *Scandinavian Journal of Statistics*, 28, 33–86.

Ashley, S. (1997), "Manufacturing Firms Face the Future," *Mechanical Engineering*, 119, 70–74.

Bouckaert, R. R., and Studený, M. (1995), "Chain Graphs: Semantics and Expressiveness," in *Symbolic and Qualitative Approaches to Reasoning and Uncertainty*, eds. C. Froideveaux and J. Kohlas, Berlin: Springer-Verlag, pp. 69–76.

Cox, D. R., and Wermuth, N. (1993), "Linear Dependencies Represented by Chain Graphs," *Statistical Science*, 8, 204–218.

——— (1996), *Multivariate Dependencies: Models, Analysis, and Interpretation*, London: Chapman & Hall.

Ding, Y., Ceglarek, D., and Shi, J. (2000), "Model and Diagnosis of Multistage Manufacturing Processes, Part I: State-Space Model," in *Proceedings of the 2000 Japan/USA Symposium on Flexible Automation*, Ann Arbor, MI, July 23–26.

Djurdjanovic, D., and Ni, J. (2001), "Linear State-Space Model of Dimensional Machining Errors," *Transactions of NAMRI/SME*, 29, 541–548.

Drton, M., and Perlman, M. D. (2005), "A SINful Approach to Gaussian Graphical Model Selection," *Journal of Statistical Planning and Inference*, in press.

Edwards, D. (2000), *Introduction to Graphical Models* (2nd ed.), New York: Springer.

Fong, D. Y. T., and Lawless, J. F. (1998), "The Analysis of Process Variation Transmission With Multivariate Measurements," *Statistica Sinica*, 8, 151–164.

Huang, Q., Zhou, N., and Shi, J. (2000), "Stream of Variation Model and Diagnosis of Multi-Station Machining Processes," in *Proceedings of the 2000 ASME International Mechanical Engineering Congress and Exposition*, Orlando, FL, November 5–10, MED-Vol. 11, pp. 81–88.

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

Jordan, M. I. (ed.) (1998), *Learning in Graphical Models*, Cambridge, MA: MIT Press.

Lauritzen, S. L. (1996), *Graphical Models*, Oxford, U.K.: Oxford University Press.

——— (1999), "Causal Inference From Graphical Models," in *Complex Stochastic Systems*, eds. O. E. Barndorff-Nielsen, D. R. Cox, and C. Klüppelberg, London/Boca Raton: Chapman and Hall/CRC Press, pp. 63–107.

Lawless, J. F., Mackay, R. J., and Robinson, J. A. (1999), "Analysis of Variation Transmission in Manufacturing Processes—Part I," *Journal of Quality Technology*, 31, 131–142.

Mantripragada, R., and Whitney, D. E. (1999), "Model and Controlling Variation Propagation in Mechanical Assemblies Using State Transition Models," *IEEE Transactions on Robotics and Automation*, 15, 124–140.

Neapolitan, R. E. (2004), *Learning Bayesian Networks*, Upper Saddle River, NJ: Prentice-Hall.

Wade, M. R., and Woodall, W. H. (1993), "A Review and Analysis of Cause-Selecting Control Charts," *Journal of Quality Technology*, 25, 161–169.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, Chichester, U.K.: Wiley.

Zantek, P. F., Wright, G. P., and Plante, R. D. (2002), "Process and Product Improvement in Manufacturing Systems With Correlated Stages," *Management Science*, 48, 591–606.

Zhang, G. X. (1985), "Cause-Selecting Control Charts: A New Type of Quality Control Chart," *The QR Journal*, 12, 221–225.

Zhou, S., Chen, Y., and Shi, J. (2004), "Root Cause Estimation and Statistical Testing for Quality Improvement of Multistage Manufacturing Processes," *IEEE Transactions on Automation Science and Engineering*, 1, 73–83.

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

Zhou, S., Huang, Q., and Shi, J. (2003b), "State-Space Model of Dimensional Variation Propagation in Multistage Machining Process Using Differential Motion Vectors," *IEEE Transactions on Robotics and Automation*, 19, 296–309.