Research Article

Abraham Gutierrez and Sebastian Müller*

Quality Analysis in Acyclic Production Networks

https://doi.org/10.1515/eqc-2019-0014
Received June 27, 2019; accepted July 15, 2019

Abstract: The production network under examination consists of a number of workstations. Each workstation is a parallel configuration of machines performing the same kind of tasks on a given part. Parts move from one workstation to another and at each workstation a part is assigned randomly to a machine. We assume that the production network is *acyclic*, that is, a part does not return to a workstation where it previously received service. Furthermore, we assume that the quality of the end product is *additive*, that is, the sum of the quality contributions of the machines along the production path. The contribution of each machine is modeled by a separate random variable. Our main result is the construction of estimators that allow pairwise and multiple comparison of the means and variances of machines in the same workstation. These comparisons then may lead to the identification of unreliable machines. We also discuss the asymptotic distributions of the estimators that allow the use of standard statistical tests and decision making.

Keywords: Direct Acyclic Graphs, Production Networks, Quality Estimation, Anomaly Detection, Variability

MSC 2010: 90B30, 90B15, 62M02, 62M05

1 Introduction

In order to maintain competitiveness, industrial manufacturers have to pursue new sources to enhance process agility. In parallel production networks, the quality of the end product depends on various intermediate production steps. Therefore, a recent challenge is to achieve a level of visibility into the production flows that allow to optimize throughput by guaranteeing at the same time given quality standards. The challenge to maintain a visibility across all parallel workflows and to identify eventual sources of errors becomes particularly difficult in situations where the qualities of single machines are not observable.

In this note we discuss a new approach allowing to compare the impact of different machines of the same workstation, i.e. performing the same task in parallel, on the quality of the product. We assume that the quality of the product is only observable at the end of the process and that the qualities of the nodes along the path are latent variables. See Figure 1 for an illustration. Although the main interest is usually in the process mean, the size of the process variability is often crucial. Our approach allows to model the process mean and process variability in parallel under very general conditions. In particular, the only condition on the production network is that it is acyclic, i.e. a part does not return to a workstation where it previously received service. These kind of networks are modeled in mathematics and computer science as directed acyclic graphs (DAGs).

***Corresponding author: Sebastian Müller,** Aix Marseille Université, CNRS, Centrale Marseille, I2M, UMR 7373, Marseille, France, e-mail: sebastian.muller@univ-amu.fr
Abraham Gutierrez, Institute of Discrete Mathematics, Graz University of Technology, Steyrergasse 30, 8010 Graz, Austria, e-mail: a.gutierrez@math.tugraz.at

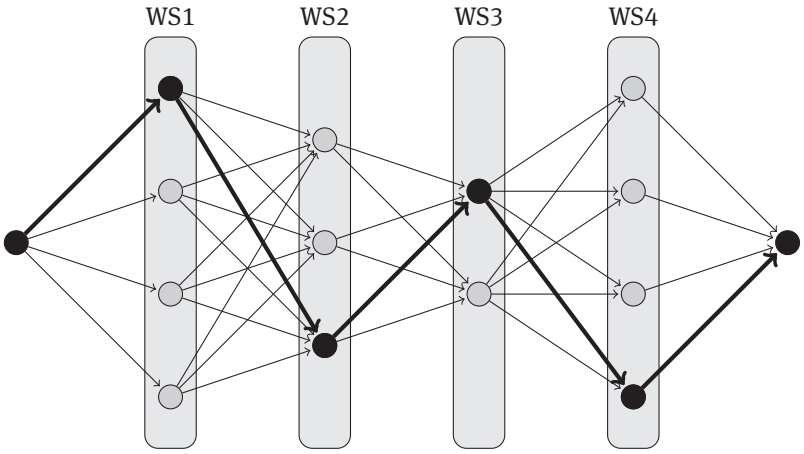

Figure 1: An illustration of a path passing through different workstations in a production network.

The concept of DAGs is becoming increasingly important having various applications in different fields of science and engineering. For instance DAGs have many applications not only in modeling production networks, e.g. [7], but also in process networks and distributed computing, e.g. [5], scheduling for systems of tasks, e.g. [16], DAG-networks for deep learning, e.g. [17], and DAG-based alternatives of the blockchain technology, e.g. [13]. Moreover, directed acyclic graphs (DAGs) are more and more often used to represent causal relationships among random variables in complex systems, e.g. [12]. We speak in general about production networks, but due to the wide applications of DAGs we want to emphasize here that our approach can be applied to any kind of acyclic network where effects along a path are additive and where measurements can only be performed at the end of the process.

Parameter estimation is not only important for the decision making process, but it is also an essential prerequisite in order to obtain meaningful simulations. However, complex production networks depend on a high number of parameters and their estimation is challenging. Since simulations in manufacturing plays an increasingly important role, e.g. [9, 11], we believe that our approach is an interesting contribution to this topic as well.

Previous Work

Estimation for the mean can also be conducted using a linear regression with categorical covariates, e.g. [10]. However, since in a linear regression homoscedasticity, i.e. the homogeneity of the variances, is a crucial condition, this approach naturally does not allow to compare differences of the variances. In fact, linear regression models are often plagued by different variabilities or heteroskedasticity. We refer to [8, Chapter 4] for an overview on how to detect and control heteroskedasticity. In contrast to these problems, our approach naturally allows different variabilities between variables and is able to detect differences in variability of values of a given variable.

Multifactor experimental designs, [4, 15] are also alternatives to estimate the mean differences but also rely on homoscedasticity. They are mostly used in the context of statistically planned experiments, which consists of a few experimental runs to obtain data on the product characteristics. If the number of observations for each setting (or path in our notation) is sufficiently high and under further conditions described in [4, Section 4], these methods allow a comparison of the variances, too. While this may offer a feasible, however not direct, way to identify differences in variability if the number of machines is small, it seems not practical in more complex networks.

There is also a connection to critical paths analysis, see e.g. [3, 14]. While these methods allow to find critical paths in acyclic networks they are not suited to compare nor estimate differences in mean and variances of given tasks.

2 The Model

A directed acyclic graph (DAG) is a finite directed graph with no directed cycles. It consists of a finite vertex set V and a finite set of directed edges $E = \{(v, w) : v, w \in V,\ v \neq w\}$. In our setting the DAG contains two special vertices: a source s and a sink t. We are interested in the paths from the source to the sink in this graph. We denote a path $\vec{p}$ in the DAG as $\vec{p} = (p_0, p_1, \dots, p_c, p_{c+1})$, where $p_0 = s$ and $p_{c+1} = t$ and $(p_i, p_{i+1}) \in E$. We define $\vec{p}[j] := p_j$. We refer to Figure 2 for an illustration and to [1] for more details on directed graphs.

We assume that at each step $1 \le i \le c$ the path $\vec{p}$ has $r_i \le r$ different choices and the nodes in each column are always numerated starting with 1. The possible choices of a path can therefore be modeled through an $r \times c$ matrix. More precisely, given a path $\vec{p}$, we associate an $r \times c$ binary matrix $V_{\vec{p}}$ that has 1's only in the nodes visited by the path:

$$V_{\vec{p}} := (V_{\vec{p}}(i, j))_{i\in[r],\, j\in[c]},$$

where we denote $[k] := \{1, 2, \dots, k\}$ for an integer k. We call $V_{\vec{p}}$ the *indicator matrix* of the path $\vec{p}$.

Each path contains exactly one node of each column. The aim of this paper is to study differences among nodes of the same columns. We think of nodes in the same columns as different possibilities for a given task, as different persons performing the same job, as different machines in the same workstation or as variations of the same kind of treatment.

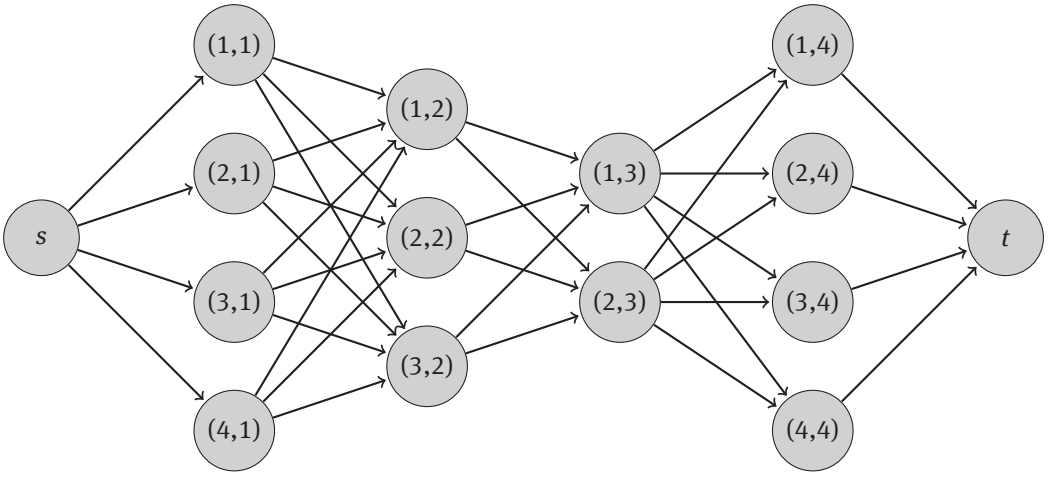

Figure 2: An illustration of a DAG with $c = 4$ and $r_1 = 4$, $r_2 = 3$, $r_3 = 2$ and $r_4 = 4$. Every node in column i has outgoing edges to every node in column $i + 1$, $i = 1, \dots, c - 1$.

The given data consists of the list of paths $\{\vec{p}_i\}_{i=1,\dots,n}$ in the DAG and the list of outputs $\{b(\vec{p}_i)\}_{i=1,\dots,n}$ for each path.

We consider the *quality matrix* S, which is a random matrix of size $r \times c$ with real entries

$$S := (s(i, j))_{i\in[r],\, j\in[c]}, \quad s(i, j) \in \mathbb{R}.$$

We model the paths with a random vector $\vec{\mathcal{P}} := (P_1, \dots, P_c)$, where the components P_i are random variables over the set $[r]$.

Throughout the paper we work under the following standing assumptions:

Assumption 1. We assume that:
(i) all entries of S have finite second moments,
(ii) the random variables $S(i, j)$, $i \in [r]$, $j \in [c]$ are (jointly) independent,
(iii) the paths $\vec{\mathcal{P}}_1, \vec{\mathcal{P}}_2, \dots$ are chosen independently and uniformly.

Note that we do not assume the entries of S to be identically distributed nor having the same variance.

Let $\vec{p} = (p_0, p_1, \dots, p_c, p_{c+1})$ be a realization of $\vec{\mathcal{P}}$, where $p_0 = s$ and $p_{c+1} = t$ almost surely. Then the *quality of the construction path* $\vec{p}$ is defined as

$$b(\vec{p}) = \sum_{j=1}^{c} S(p_j, j).$$

We also can think of $b(\vec{p})$ as the quality (or error) cumulated along the path $\vec{p}$.

Let us make precise where the randomness enters in our model. We choose a random path $\vec{\mathcal{P}}$ and random matrix S. The corresponding probability measure is denoted by $\mathbb{P}$. The random choice of $\vec{\mathcal{P}}$ and S induces a random variable $b(\vec{\mathcal{P}})$ and allows to generate a sequence of i.i.d. random variables $(\vec{\mathcal{P}}_1, b(\vec{\mathcal{P}}_1))$, $(\vec{\mathcal{P}}_2, b(\vec{\mathcal{P}}_2))$, etc.

The goal of our study is to give estimates on the law of S by observing the paths $\vec{\mathcal{P}}$ and its cumulated qualities $b(\vec{\mathcal{P}})$. Note that $(\vec{\mathcal{P}}_n, b(\vec{\mathcal{P}}_n))_{n\in\mathbb{N}}$ is in general not a sufficient statistic for S, i.e. we cannot recover the distribution of S by only observing realizations of $(\vec{\mathcal{P}}, b(\vec{\mathcal{P}}))$, as we see in the next remark.

Remark 1. Let us consider the case $r = 1$ and $c = 2$. Let $S(1,1) \sim \mathcal{N}(0,1)$ and $S(1,2) \sim \mathcal{N}(1,1)$ and define $\tilde{S}(1,1) := S(1,1) + 1$ and $\tilde{S}(1,2) := S(1,2) - 1$. Then for any given path $\vec{p}$ we have $\sum_{j=1}^{2} S(p_j, j) = \sum_{j=1}^{2} \tilde{S}(p_j, j)$. Hence, the statistic $(\vec{p}_n, b(\vec{p}_n))_{n\in\mathbb{N}}$ does not allow us to distinguish between S and $\tilde{S}$.

Example 2 (Binary Errors). The matrix S consists of independent Bernoulli random variables $S(i,j)$. The value 1 of this Bernoulli may encode a defect and hence $b(\vec{p})$ counts the number of defects of the end product.

Example 3 (Gaussian Quality). The matrix S consists of independent Gaussian random variables $S(i,j)$. The end quality $b(\vec{p})$ is then distributed as a mixture of Gaussian random variables.

Given a sequence of realizations $(\vec{p}_k)_{k\in[n]}$ of $\vec{\mathcal{P}}$, we define the following matrices that at are the core of our analysis:

$$B^{(n)} := \sum_{k=1}^{n} b(\vec{p}_k) V_{\vec{p}_k}, \quad V^{(n)} := \sum_{k=1}^{n} V_{\vec{p}_k}, n \geq 1.$$

The value $B^{(n)}(i,j)$ is the sum of all cumulated qualities of paths containing node (i,j), whereas $V^{(n)}(i,j)$ just counts the number of times node (i,j) was used. We define the *sample mean matrix* as the sample mean quality matrix:

$$T^{(n)} := (T^{(n)}(i,j))_{i\in[r],\, j\in[c]}, \quad \text{where } T^{(n)}(i,j) := \begin{cases} \dfrac{B^{(n)}(i,j)}{V^{(n)}(i,j)}, & \text{if } V^{(n)}(i,j) \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

The corresponding *sample variance matrix* $\Sigma^{(n)}$ is defined by

$$\Sigma^{(n)}(i,j) := \begin{cases} \dfrac{1}{V^{(n)}(i,j)} \displaystyle\sum_{k=1}^{n} \left(b(\vec{p}_k) V_{\vec{p}_k}(i,j) - T^{(n)}(i,j)\right)^2, & \text{if } V^{(n)}(i,j) \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

3 Results

Denote $D^{(n)} = (\vec{\mathcal{P}}_i)_{i\in[n]}$ the multi-set[1] or sequence of all paths up to time n and let

$$D^{(n)}_{(i,j)} := \{\vec{p} \in D^{(n)} : \vec{p}[j] = i\}$$

be the multi-set of all paths up to index n that go through the node (i,j); we note that we can recover $V^{(n)}(i,j)$ through $D^{(n)}_{(i,j)}$ by $V^{(n)}(i,j) = |D^{(n)}_{(i,j)}|$.

1 We use a multi-set since we need to keep track of the multiplicity of the paths.

In general, it is not possible to estimate the mean quality matrix $\mathbb{E}[S]$, see Remark 1. However, it is possible to identify nodes with higher or lower quality mean or variance in each column.

Theorem 4. *Let* $(i,j),(i',j) \in [r]\times[c]$. *Then*

$$T^{(n)}(i,j) - T^{(n)}(i',j) \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[S(i,j)] - \mathbb{E}[S(i',j)]$$

and

$$\Sigma^{(n)}(i,j) - \Sigma^{(n)}(i',j) \xrightarrow[n\to\infty]{a.s.} \mathbb{V}[S(i,j)] - \mathbb{V}[S(i',j)].$$

Proof. Using twice the law of large numbers and the continuous mapping theorem, we obtain

$$\begin{aligned} T^{(n)}(i,j) &= \frac{1}{|D^{(n)}_{(i,j)}|}\sum_{\vec{p}\in D^{(n)}_{(i,j)}} b(\vec{p}) \\ &= \frac{n}{|D^{(n)}_{(i,j)}|}\frac{1}{n}\sum_{\vec{p}\in D^{(n)}_{(i,j)}} b(\vec{p}) \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{\mathbb{P}(\vec{\mathcal{P}}[j]=i)}\mathbb{E}[b(\vec{\mathcal{P}});\vec{\mathcal{P}}[j]=i] = \mathbb{E}[b(\vec{\mathcal{P}})\mid\vec{\mathcal{P}}[j]=i]. \end{aligned}$$

In the same way

$$T^{(n)}(i',j) = \frac{1}{|D^{(n)}_{(i',j)}|}\sum_{\vec{p}\in D^{(n)}_{(i',j)}} b(\vec{p}) \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}[b(\vec{\mathcal{P}})\mid\vec{\mathcal{P}}[j]=i'].$$

Using the assumption that the paths are chosen uniformly and the definition of $b(\vec{\mathcal{P}})$, we obtain the first part of the theorem from

$$\begin{aligned} \mathbb{E}[b(\vec{\mathcal{P}})\mid\vec{\mathcal{P}}[j]=i] - \mathbb{E}[b(\vec{\mathcal{P}})\mid\vec{\mathcal{P}}[j]=i'] &= \frac{\mathbb{E}[b(\vec{\mathcal{P}});\vec{\mathcal{P}}[j]=i] - \mathbb{E}[b(\vec{\mathcal{P}});\vec{\mathcal{P}}[j]=i']}{\mathbb{P}(\vec{\mathcal{P}}[j]=i)} \\ &= \frac{\mathbb{E}[S(i,j);\vec{\mathcal{P}}[j]=i] - \mathbb{E}[S(i',j);\vec{\mathcal{P}}[j]=i]}{\mathbb{P}(\vec{\mathcal{P}}[j]=i)} \\ &= \mathbb{E}[S(i,j)] - \mathbb{E}[S(i',j)]. \end{aligned}$$

For the second part of the theorem, we use the law of large numbers and the continuous mapping theorem to get that

$$\Sigma^{(n)}(i,j) = \Big(\frac{1}{|D^{(n)}_{ij}|}\sum_{\vec{p}\in D^{(n)}_{ij}} b^2(\vec{p})\Big) - \big(T^{(n)}(i,j)\big)^2 \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}[b(\vec{\mathcal{P}})^2\mid\vec{\mathcal{P}}[j]=i] - \mathbb{E}[b(\vec{\mathcal{P}})\mid\vec{\mathcal{P}}[j]=i]^2.$$

Using the definition of $b(\vec{p})$ we deduce with elementary calculations that

$$\mathbb{E}[b(\vec{\mathcal{P}})^2\mid\vec{\mathcal{P}}[j]=i] - \mathbb{E}[b(\vec{\mathcal{P}})\mid\vec{\mathcal{P}}[j]=i]^2 = A_j + \mathbb{V}[S(i,j)]$$

where A_j is a quantity that only depend on the column j. Applying this identity for (i,j) and (i',j), we obtain that

$$\Sigma^{(n)}(i,j) - \Sigma^{(n)}(i',j) \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{V}[S(i,j)] - \mathbb{V}[S(i',j)]. \qquad \square$$

3.1 Asymptotic Distribution

The sum

$$\sum_{\vec{p}\in D^{(n)}_{(i,j)}} b(\vec{p}) = \sum_{k=1}^{n} b(\vec{\mathcal{P}}_k)\mathbf{1}\{\vec{\mathcal{P}}_k[j]=i\}$$

can be interpreted as the sum of random variables appearing in an acception-rejection method. More precisely, we start with $k=1$ and consider $(\vec{\mathcal{P}}_k, b(\vec{\mathcal{P}}_k))$. If $\vec{\mathcal{P}}_k[j]=i$, we set $Y := b(\vec{\mathcal{P}}_k)$ and stop, otherwise we increase k and repeat until $\vec{\mathcal{P}}_K[j]=i$ for the first K. Now, for $y\in\mathbb{R}$,

$$\mathbb{P}(Y\le y) = \sum_{k=1}^{\infty}\mathbb{P}(b(\vec{\mathcal{P}}_k)\le y\mid K=k)\mathbb{P}(K=k) = \mathbb{P}(b(\vec{\mathcal{P}}_1)\le y\mid\vec{\mathcal{P}}_1[j]=i).$$

In other words, the distribution of Y equals the distribution of $b(\vec{\mathcal{P}}_1)$ conditioned on $\vec{\mathcal{P}}_1[j] = i$. Iterating this acception-rejection method, we see that $|D^{(n)}_{(i,j)}|$ describes the number of acceptions using $(\vec{\mathcal{P}}_k, b(\vec{\mathcal{P}}_k))$, $1 \le k \le n$. Hence the estimator $T^{(n)}$ has the same distribution as

$$\frac{1}{|D^{(n)}_{(i,j)}|} \sum_{k=1}^{|D^{(n)}_{(i,j)}|} Y_k,$$

where Y_k, $k \in \mathbb{N}$, is a sequence of i.i.d. random variables distributed as $b(\vec{\mathcal{P}}_1)$ conditioned on $\vec{\mathcal{P}}_1[j] = i$. Finally, Anscombe's theorem, [6, Theorem 1.3.1], implies that

$$\sqrt{|D^{(n)}_{(i,j)}|}\,(T^{(n)}(i,j) - \mu_{i,j}) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \sigma^2_{i,j}),$$

where

$$\mu_{i,j} := \mathbb{E}[b(\vec{\mathcal{P}}) \mid \vec{\mathcal{P}}[j] = i] \quad \text{and} \quad \sigma^2_{i,j} := \mathbb{E}[b(\vec{\mathcal{P}})^2 \mid \vec{\mathcal{P}}[j] = i] - \mathbb{E}[b(\vec{\mathcal{P}}) \mid \vec{\mathcal{P}}[j] = i]^2.$$

For the variance we assume that the entries of S have finite forth moments. The estimator for the variance is

$$\Sigma^{(n)}(i,j) = \frac{1}{|D_{(i,j)}|} \sum_{\vec{p} \in D^{(n)}_{(i,j)}} \Big(b(\vec{p}) - \frac{1}{|D^{(n)}_{(i,j)}|} \sum_{\vec{p} \in D^{(n)}_{(i,j)}} b(\vec{p}) \Big)^2.$$

Since the distribution of $\Sigma^{(n)}(i,j)$ does not change if we replace $b(\vec{p})$ by $b(\vec{p}) - \mu_{i,j}$, we can assume that $\mu_{i,j} = 0$. Moreover,

$$\Sigma^{(n)}(i,j) = \frac{1}{|D_{(i,j)}|} \sum_{\vec{p} \in D^{(n)}_{(i,j)}} b(\vec{p})^2 - \big(T^{(n)}(i,j)\big)^2.$$

We have that $T^{(n)}(i,j)$ converges almost surely to 0 and $\sqrt{D^{(n)}_{(i,j)}}T^{(n)}(i,j)$ converges in distribution to $\mathcal{N}(0, \sigma^2_{i,j})$. Slutzky's theorem implies that $\sqrt{D^{(n)}_{(i,j)}}T^{(n)}(i,j)^2$ converges in distribution to 0 and therefore also in probability to 0. Using Slutzky's theorem and Anscombe's theorem as above, we obtain that

$$\sqrt{D^{(n)}_{(i,j)}}(\Sigma^{(n)}(i,j) - \sigma^2_{i,j}) = \sqrt{D^{(n)}_{(i,j)}}\Big(\frac{1}{|D_{(i,j)}|} \sum_{\vec{p} \in D^{(n)}_{(i,j)}} b(\vec{p})^2 - \sigma^2_{i,j} \Big) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \tau^2_{i,j}),$$

with $\tau^2_{i,j} = \mathbb{E}[(b(\vec{\mathcal{P}}) - \mu_{i,j})^4 \mid \vec{\mathcal{P}}[j] = i] - \mathbb{E}[(b(\vec{\mathcal{P}}) - \mu_{i,j})^2 \mid \vec{\mathcal{P}}[j] = i]^2$.

3.2 Pairwise and Multiple Comparison

The asymptotic distributions of the estimators $T^{(n)}$ and $\Sigma^{(n)}$ justify that pairwise comparison of the mean quality and the variance of quality can be done using standard tests, e.g. t-test, Fisher tests and Bartlett tests. We sketch only the construction of a test statistic to compare the means of two nodes under the hypothesis that they have the same variance. However, it is standard to extent this result to the case where the variances are not equal and to the estimators of the variance differences. More details on possible applications are given in Section 4. We have for n sufficiently large

$$\frac{\sqrt{|D^{(n)}_{(i,j)}|}}{\sigma_{i,j}}(T^{(n)}(i,j) - \mu_{i,j}) \sim \mathcal{N}(0,1) \quad \text{and} \quad \frac{\sqrt{|D^{(n)}_{i',j}|}}{\sigma_{i',j}}(T^{(n)}(i',j) - \mu_{i',j}) \sim \mathcal{N}(0,1).$$

Under the hypothesis that $\mu_{i,j} = \mu_{i',j}$ and $\sigma = \sigma_{i,j} = \sigma_{i',j}$ one obtains that for n sufficiently large

$$\left(\frac{|D^{(n)}_{i,j}| \cdot |D^{(n)}_{i',j}|}{|D^{(n)}_{i,j}| + |D^{(n)}_{i',j}|} \right)^{\frac{1}{2}} \frac{T^{(n)}(i,j) - T^{(n)}(i',j)}{\sigma} \sim \mathcal{N}(0,1).$$

Replacing σ by its estimator

$$\hat{\sigma} = \left(\frac{1}{|D^{(n)}_{i,j}| + |D^{(n)}_{i',j}|} \big(|D^{(n)}_{i,j}| \Sigma^{(n)}(i,j) + |D^{(n)}_{i',j}| \Sigma^{(n)}(i',j) \big) \right)^{\frac{1}{2}},$$

we find that

$$\left(\frac{|D_{i,j}^{(n)}| \cdot |D_{i',j}^{(n)}|}{|D_{i,j}^{(n)}| + |D_{i',j}^{(n)}|}\right)^{\frac{1}{2}} \frac{T^{(n)}(i,j) - T^{(n)}(i',j)}{\hat{\sigma}}$$

can be approximated by a standard normal distribution for n sufficiently large.

The identification of nodes with bad quality boils down to multiple comparisons. For a given column j with r_j nodes we define a vector $X = (X_1, \dots, X_{r_j-1})$ of test statistics

$$X_k = T^{(n)}(k,j) - T^{(n)}(r_j,j), \quad k \in \{1, \dots, r_j - 1\}.$$

The vector X satisfies (asymptotically) the positive regression dependency. Therefore, we suggest the Benjamini–Yekutieli method, see [2], to control the false discovery rate.

4 Examples

4.1 Wafer Production

Our study was motivated by a root-cause analysis in the wafer fabrication. Wafer fabrication is in general a procedure of many repeated sequential processes. For instance, a simplified illustration consists of twelve subsequent fabrication steps, see [18], where intermediate measurement of qualities are not a feasible. In our concrete examples we treated up to 30 different steps and more than 90 machines. Unfortunately, since our industrial partner insists on the fulfillment of an NDA, we are not allowed to publish any more information about the project. Probably for the same reasons, it was impossible for us to find public available data on other industrial projects.

4.2 Simulations

We consider the DAG network as given in Figure 2 and consider Gaussian qualities as described in Example 3. The matrix S consists in this case of independent Gaussian random variables. In this simulation we consider the distribution that is characterized by

$$\mathbb{E}[S] := \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & * & 0 \\ 0 & * & * & 0 \end{bmatrix} \quad \text{and} \quad \mathbb{V}[S] := \begin{bmatrix} 1 & 1 & 4 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & * & 1 \\ 1 & * & * & 1 \end{bmatrix},$$

where the $*$ are placeholders for the machines that do not exist. We simulate $n = 200$ observations and obtain the following results for the estimators T and Σ (rounded to two decimals):

$$T = \begin{bmatrix} 0.60 & -0.03 & 0.75 & 0.70 \\ 0.25 & 2.06 & 0.40 & 0.89 \\ 0.74 & -0.10 & * & 0.38 \\ 0.69 & * & * & 0.14 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 6.46 & 5.62 & 8.92 & 7.12 \\ 8.76 & 6.13 & 5.38 & 4.89 \\ 4.97 & 6.58 & * & 8.41 \\ 7.37 & * & * & 8.14 \end{bmatrix}.$$

While the difference in mean in the second column seems to be obvious, the difference of the variances in the third column might be overlooked and differences in the variance in the first and last columns could be suspected.

We perform pairwise t-tests for each column with Benjamini–Yekutieli adjustment. While in all but the second column no statistically significant difference is detected, the difference in mean of the second machine in column 2 is detected with a p-value of $9.8 \cdot 10^{-6}$. The Bartlett test does not find any differences in the variances in columns 1, 2, and 4. However, the difference of variance in column 3 seems to be statistically

significant with a p-value of 0.01. The number of observations of 200 is rather small compared to the number of different paths, which is 96. This explains the fact that the matrix Σ does not reflect the correct differences of variances; in particular, we have in the third column that $8.92 - 5.38 = 3.54 > 3 = 4 - 1$. In order to demonstrate the convergence of the estimators, we calculate the estimators for $n = 10000$:

$$T = \begin{bmatrix} 0.63 & 0.02 & 0.62 & 0.65 \\ 0.62 & 1.99 & 0.68 & 0.59 \\ 0.66 & -0.06 & * & 0.69 \\ 0.70 & * & * & 0.67 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 6.32 & 5.47 & 7.82 & 6.44 \\ 6.28 & 5.42 & 4.85 & 6.03 \\ 6.26 & 5.39 & * & 6.23 \\ 6.44 & * & * & 6.59 \end{bmatrix}.$$

Acknowledgment: The authors wish to thank Alessandro Chiancone, Herwig Friedl, Jérôme Depauw, and Marc Peigné for stimulating discussing during this project. Grateful acknowledgement is made for hospitality from TU-Graz where the research was carried out during visits of Sebastian Müller.

Funding: Abraham Gutierrez acknowledges financial support from the Austrian Science Fund project FWF P29355-N35.

References

[1] J. Bang-Jensen and G. Gutin, *Digraphs. Theory, Algorithms and Applications*, 2nd ed., Springer Monogr. Math., Springer, London, 2009.
[2] Y. Benjamini and D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, *Ann. Statist.* **29** (2001), no. 4, 1165–1188.
[3] D. Bohme, M. Geimer, F. Wolf and L. Arnold, Identifying the root causes of wait states in large-scale parallel applications, in: *39th International Conference on Parallel Processing*, IEEE Press, Piscataway (2010), 90–100.
[4] N. R. Draper and F. Pukelsheim, An overview of design of experiments, *Statist. Papers* **37** (1996), no. 1, 1–32.
[5] D. G. Feitelson, L. Rudolph, U. Schwiegelshohn, K. C. Sevcik and P. Wong, Theory and practice in parallel job scheduling, in: *Job Scheduling Strategies for Parallel Processing*, Lecture Notes in Comput. Sci. 1291, Springer, Berlin (1997), 1–34.
[6] A. Gut, *Stopped Random Walks. Limit Theorems and Applications*, 2nd ed., Springer Ser. Oper. Res. Financ. Eng., Springer, New York, 2009.
[7] S. T. Hackman and R. C. Leachman, A general framework for modeling production, *Manag. Sci.* **35** (1989), no. 4, 478–495.
[8] C. Kleiber and A. Zeileis, *Applied Econometrics with R*, Springer, New York, 2008.
[9] V. S. Kouikoglou and Y. A. Phillis, *Hybrid Simulation Models of Production Networks*, Springer, New York, 2001.
[10] D. C. Montgomery, E. A. Peck and G. G. Vining, *Introduction to Linear Regression Analysis*, 3rd ed., Wiley Ser. Probab. Stat. Texts References Pocketbooks Sect., Wiley-Interscience, New York, 2001.
[11] D. Mourtzis, M. Doukas and D. Bernidaki, Simulation in manufacturing: Review and challenges, *Procedia CIRP* **25** (2014), 213–229.
[12] J. Pearl, *Causality*, Cambridge University, Cambridge, 2009.
[13] S. Popov, The Tangle, version April 30, 2018.
[14] M. Schulz, Extracting critical path graphs from mpi applications, in: *IEEE International Conference on Cluster Computing*, IEEE Press, Piscataway (2005), 1–10.
[15] D. Selvamuthu and D. Das, *Introduction to Statistical Methods, Design of Experiments and Statistical Quality Control*, Springer, Singapore, 2018.
[16] S. S. Skiena, *The Algorithm Design Manual*, Springer, London, 2008.
[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, Going deeper with convolutions, in: *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Press, Piscataway (2015), 1–9.
[18] Wikipedia contributors, Wafer fabrication, 2019.