Knockoff Procedure for False Discovery Rate Control in High-Dimensional Data Streams

Ka Wai Tsang, Fugee Tsung, Zhihao Xu

1 Introduction

Nowadays, high-dimensional data streams can be easily obtained in modern manufacturing and service industries. Together with the increasing low computational cost to handle big data, people are more willing to use high-dimensional data streams to build statistical models for monitoring complex systems and quality control. It makes the conventional tools for multivariate statistical process control (SPC) become difficult and even impossible, to apply in practice. The tasks of multivariate SPC basically include fault detection, which determines if there exist changes in the data streams, and fault identification, which isolating the components that are responsible for the changes. Beginning with acceptance sampling and Shewhart's control charts that use relatively simple univariate quality characteristics, multivariate quality control characteristics and more efficient SPC schemes such as multiple cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) have been developed and commonly used due to the advances in multivariate analysis and sequential analysis in the statistical literature. Because of the availability of "big data" for fault detection and identification and recent developments in the statistics literature on high-dimensional data analysis, the analysis for multivariate SPC has moved to a new direction in the past decade. For fault detection, the recent development for high-dimensional data streams can be found in Woodall and Montgomery [2014], Zou et al. [2015], Xian et al. [2018]. In this paper, we propose a novel knock-off procedure for fault identification for high-dimensional data streams.

Suppose we observe independent $X_{j,t}$ for the j^{th} data stream at time $t = 1, 2, \dots$ for $j = 1, \dots, m$, with the number of streams being m. After a change-point v, the distributions of the observations from some proportion of the streams get changed. Hence for the j^{th} data stream, the density function of $X_{j,t}$ is f_0^j for $t < v$, and is f_1^j for $t \geq v$. This leads to the null hypothesis $H_j : f_1^j = f_0^j$, or equivalently, the j^{th} data stream is not among those data streams that change their distributions at the change point v. Let $\mathbf{X_t} = [X_{1,t}, \dots, X_{m,t}]^T$, $\boldsymbol{\mu_0} = \mathrm{E}(\mathbf{X_t})$ for $t < v$ and $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X_t})$ for $t \geq v$. We assume the covariance matrix $\boldsymbol{\Sigma} = \mathrm{Var}(\mathbf{X_t})$ remains the same for all t and there is mean shift for the j^{th} stream if $f_1^j \neq f_0^j$. Therefore, assume $\boldsymbol{\mu_0}$ is a zero vector, our hypotheses are $H_j : \mu_j = 0$ versus $H_j^A : \mu_j \neq 0$ for $j = 1, \dots, m$ where μ_j is the mean of f_1^j. When m is small, a common test statistic used for fault detection and identification is Hotelling's T^2. To locate the mean shifts, different decompositions of Hotelling's T^2 statistics have been developed (Mason et al. [1995], Mason et al. [1997], Li et al. [2008]). Zhu and Jiang [2009] points out that such decomposition "depends on a prespecified order of variables and known a priori". As the shift structures may not be known in practices, Zhu and Jiang [2009] propose an adaptive T^2 chart, which does not require a priori knowledge of the potential shifts space, for fault detection and identification. However, when n is large, Li et al. [2020] comment that "the decomposition of the T^2 statistic consider $m!$ different decompositions of the T^2 statistic, and the standard step-down testing procedures still require the computation of $\binom{m}{k}$ test statistics of quadratic-term in the k^{th} step" and point out that "both methods are very time consuming." Another issue for multivariate SPC to the high dimensionality is the signals of those shifted streams, or out-of-control (OC) streams, can be buried in the large amount of unshifted streams, or in-control (IC) streams, and thus become difficult to detect. Wang and Jiang [2009] illustrate this issue by showing the detection power of a Hotelling's T^2 chart with dimension m and one shifted stream. They show that the shift can be detected with very high probability (0.982) when $n = 1$, but the detection probability drops to 0.018 when the dimension m increases from 1 to 20.

To handle the high dimensionality, various methods based on variable selection have been proposed recently. Wang and Jiang [2009] consider a penalized optimization problem to find a sparse estimator of μ such that a linear combination of the generalized distance with the sample mean and a LASSO-type penalty is minimum. Liang et al. [2016] extend their ideas of using LASSO penalty to non-normal situations. Instead of LASSO penalty, Jiang et al. [2012] and Li et al. [2017] consider Lo penalty of the $\Sigma_j \mathbb{1}_{\{|\mu_j| \neq 0\}} \leq s$ to select s data streams instead of selecting streams by penalized optimization, Mei [2011] proposes a top-r thresholding scheme to select r streams with largest local CUSUM statistics. Capizzi and Masarotto [2011] and Ing et al. [2017] model the main shifts as linear combinations of some explanatory variables and thus formulate the fault identification problem as variable selection problem in linear regression models. These variable selection methods are designed to select almost all significant signals (mean shifts). However, it is also important for these methods to avoid selecting too many false alarms. Otherwise, many resources will be wasted on checking those false alarms. Ing et al. [2017] propose an OGA procedure to control the family-wise error rate (FWER), which is the probability existing false alarms in the selection set. However, since it is usually not a matter to have a few, but not many, false alarms in practice, controlling FWER can be unnecessarily too conservative. Besides, their OGA procedure requires the number of OC observations large enough to asymptotically control FWER, which may not be possible in practice. As Li et al. [2020] points out that "In most SPC applications, a major goal is to detect a change as quickly as possible and, if this goal is attained, there may be only a few OC observations collected since the process is always shut down after a signal." Therefore, the performance of their OGA procedure can be poor for limited number of OC observation. A less conservative error rate that is commonly considered to control is the false discovery rate (FDR), which is the expected ratio of the false alarms to the number of all of the rejected hypotheses. Li and Tsung [2009] apply Benjamini and Hochberg's (BH) procedure to establish FDR-adjusted Shewhart chart and FDR-adjusted CUSUM chart. However, their proposed schemes do not guarantee the control of FDR. Du and Zou [2018] develop a dynamic multiple testing procedure for high dimensional multivariate SPC with the control of FDR at each time point, but not at the stopping time when signals are detected. These variable selection based schemes are mainly focus on online fault detection, but not fault identification. These methods select a set of data streams at each time point, and then conduct hypotheses testing or check criteria to detect signals. The number or the ratio of false discoveries in the selected set at the stopping time is not guaranteed to be controlled. For controlling FDR after variable selection, recently Barber et al. [2015] propose a variable selection procedure, knockoff filtering, that control FDR in the linear regression setting and their empirical results show that knockoff filtering has higher power than other selection methods with FDR controlled at the same level when the proportion of null hypotheses is high. In this paper, we apply their ideas to propose a knockoff procedure for fault identification with the control of FDR.

A brief introduction of knockoff filtering is presented in Section 2, in which we give our general algorithm for fault identification with theoretical justification for the control of FDR. Our general algorithms can be applied with various variable selection methods. Two particular variable selection methods are considered in Section 3 to illustrate the use of our general algorithm.

2 Knockoff Procedure

Barber et al. [2015] introduce a powerful variable selection procedure, called knockoff filtering, controlling the FDR in the statistical linear model whenever there are at least as many observations as hypotheses. They consider the linear model $y_t = \sum_{j=1}^{p} X_{tj}\beta + \varepsilon_t, t = 1, \ldots, n$, where ε_t are error terms independent of X_{tj} with mean 0. They are interested in testing $H_j : \beta_j = 0$ for $j = 1, \ldots, p$ and restrict their attention to the case where $n \geq p$. Such restriction is released in Barber et al. [2019] who have extended knockoff filtering to the high-dimensional setting. The core part of knockoff filtering is the construction of knockoff variables that are defined to be pairwise exchangeable with the originals $\boldsymbol{X}_j = [X_{1j}, \ldots, X_{nj}]^T, j = 1, \ldots, p$, but not related to $\boldsymbol{y} = [y_1, \ldots, y_n]^T$. Let $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}$ be a knockoff variable for $\boldsymbol{X}_j$, then if H_j is true, the joint distribution of $[\boldsymbol{X}, \tilde{\boldsymbol{X}}] = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p, \tilde{\boldsymbol{X}}_{\boldsymbol{1}}, \ldots, \tilde{\boldsymbol{X}}_{\boldsymbol{p}}]$ does not change when $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}$ are swapped. More generally, for any subset T of the set of null hypotheses H_0, swapping columns of $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}$ for any $j \in T$ does not change the distribution of $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$. With such exchangeable knockoff variables, a set of variable importance statistics $[\boldsymbol{Z}, \tilde{\boldsymbol{Z}}] = [Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p]$, which reflects the importance of $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}$, can then be computed. For example, if we apply a forward stepwise selection algorithm, e.g. orthogonal greedy algorithm in Ing et al. [2017], on $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ and $\boldsymbol{y}$, then $Z_j(\tilde{Z}_j)$ can be defined as the order of selection of $\boldsymbol{X}_j(\tilde{\boldsymbol{X}}_{\boldsymbol{j}})$. Note that if H_j is true, $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}$ are of the same importance. Otherwise, $\boldsymbol{X}_j$ is more likely to be selected before $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}$ as $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}$ is not related to $\boldsymbol{y}$, and thus Z_j is expected to be smaller than $\tilde{Z}_j$ in such case. And then we can construct knockoff statistics $W_j = W(Z_j, \tilde{Z}_j)$ such that W_j satisfies (i) flip-sign property $W_j = W(Z_j, \tilde{Z}_j) = -W(\tilde{Z}_j, Z_j)$ and (ii) if H_j is true, $P(W_j > t) = P(W_j < t)$ for all $t > 0$ and $P(W_j > t) > P(W_j < t)$ if H_j does not hold. In the previous forward stepwise example, we can take $W_j = \tilde{Z}_j - Z_j$. Note that conditioning on $|W_j|$, the signs of the W_j's corresponding to null H_j are distributed as i.i.d. fair coin flips and W_j's are expected to be positive if H_j is that depends on the number of W_j's being positive if H_j is false. It suggests rejecting H_j for those W_j passing same threshold that depends on the number of W_j's being positive and negative. Barber et al. [2015] define the knockoff estimate of false discovery proportion (FDP) as

$$\widehat{FDP}(t) = \frac{1 + \#\{j : W_j \leq -t\}}{\max\{1, \#\{j : W_j \geq t\}\}} \tag{1}$$

and reject H_j if $W_j \geq t_{\min}$ where $t_{\min} = \min\{t : \widehat{FDP}(t) \leq \alpha\}$ to control the FDR at α level.

For the multivariate SPC problem, suppose we observe p data streams $\boldsymbol{X}_{\boldsymbol{j},\boldsymbol{\cdot}} = [X_{j,1}, \ldots, X_{j,\tau^{obs}}]^T$ after the change-point v, which is assumed to be known or can be well estimated by some change-point methods, see, for example, Zamba and Hawkins [2006]. The observed stopping time after v is supposed to be small since the process should be stopped after a signal is detected. We assume $\tau^{obs} = \tau(\{X_{j,1}, X_{j,2}, \ldots\}, j = 1, \ldots, p)$ is a function of a set of data streams. If $X_{j,t}$ with the null density function f_0^j are independent, then we can simply generate a knockoff copy for $\boldsymbol{X}_{j,\cdot}$ by generating $\tilde{\boldsymbol{X}}_{\boldsymbol{j},\boldsymbol{\cdot}} = [X_{j,1}, \ldots, X_{j,\tau^{obs}}]^T$ from the known density f_0^j. Let $\tau^{KF} = \tilde{\tau}(\{X_{j,1}, X_{j,2}, \ldots\} \cup \{\tilde{X}_{j,1}, \tilde{X}_{j,2}, \ldots\}, j = 1, \ldots, p)$ be the new stopping time for p original data streams and their knockoff copies. The stopping rule $\tilde{\tau}$ is chosen such that $\tau^{KF} \leq \tau^{obs}$. For example, for top-r procedure, the original stopping time τ is the sum of largest r CUSUM statistics being greater than or equal to some threshold a. In such case, we can choose $\tilde{\tau} = \tau$ since the sum of largest r CUSUM statistics of the original data streams and knockoff copies must be greater than or equal to a. In general, we can select a set of p data streams at each time $t \leq \tau^{obs}$ such that the stopping time τ is most likely to be satisfied, e.g. choosing a set of p data streams with largest absolute sample mean or with largest CUSUM statistics, or with smallest p-value etc. By defining $\tilde{\tau}$ to be first selecting the m data streams with most significant signals, and then apply τ on the selected set, we can guarantee $\tau^{KF} \leq \tau^{obs}$. For a data stream $X_{j,t}$ with $f_0^j = f_1^j$, i.e. H_j is true, $X_{j,t}$ and the corresponding knockoff copy $\tilde{X}_{j,t}$ have the same distribution. Therefore, we can compute the corresponding variable importance statistics Z_j and $\tilde{Z}_j$ (e.g. the sample means of $X_{j,t}$ and $\tilde{X}_{j,t}, t = 1, \ldots, \tau^{KF}$) such that Z_j and $\tilde{Z}_j$ have the same distribution, and hence the knockoff procedure can be applied to control FDR.

General knockoff procedure for multivariate SPC.

1. Generate knockoff copies $\tilde{X}_{j,t}$ of observed $X_{j,t}$ for $j = 1, \ldots, p, t = 1, \ldots, \tau^{obs}$
2. Compute the new stopping time $\tau^{KF} = \tilde{\tau}(\{X_{j,1}, X_{j,2}, \ldots\} \cup \{\tilde{X}_{j,1}, \tilde{X}_{j,2}, \ldots\}, j = 1, \ldots, p)$. The stopping rule $\tilde{\tau}$ is chosen such that $\tau^{KF} \leq \tau^{obs}$
3. Compute $Z_j = Z_{j,\tau^{KF}}$ and $\tilde{Z}_j = \tilde{Z}_{j,\tau^{KF}}$, where
$$Z_{j,t} = \max\{Z_{j,t-1} + X_{j,t}, 0\} \text{ and } \tilde{Z}_{j,t} = \max\left\{\tilde{Z}_{j,t-1} + \tilde{X}_{j,t}, 0\right\}$$
with $Z_{j,0} = \tilde{Z}_{j,0} = 0$. Set $W_j = Z_j - \tilde{Z}_j$ for $j = 1, \ldots, p$.
4. Reject H_j if $W_j \geq t_{\min}$ where $t_{\min} = \min\{t : \widehat{FDP}(t) \leq \alpha\}$.

Here we consider CUSUM chart for computing Z_j and $\tilde{Z}_j$ in Step 3 as it has been proven optimal properties for a given size of mean shift, see Bagshaw and Johnson [1975].

If $\boldsymbol{X}_{j,\cdot} = [X_{j,1}, \ldots, X_{j,\tau^{obs}}]^T$ are independent, $\tilde{\boldsymbol{X}}_{\boldsymbol{j},\cdot} = [\tilde{X}_{j,1}, \ldots, \tilde{X}_{j,\tau^{obs}}]^T$ in Step 1 can be generated by the known density f_0^j. If not, the correlation among $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ needs to be considered in the construction of knockoff copies to ensure that the distributions of $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}$ are the same. If $\boldsymbol{X}_{j,\cdot}, j = 1, \ldots, p$, are not independent, we further assume that f_0^j is the density function of $\mathcal{N}(0, 1)$, the density f_1^j after the change-point v is the density function of $\mathcal{N}(\mu_j, 1)$, and the joint density of $\boldsymbol{X}_{\cdot,t} = [X_{1,t}, \ldots, X_{p,t}]^T$ after v is $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In such case, we follow the ideas in Barber et al. [2015] to construct $\tilde{\boldsymbol{X}}_{\cdot,\boldsymbol{t}} = [\tilde{X}_{1,t}, \ldots, \tilde{X}_{p,t}]^T$ that follows

$$\begin{bmatrix} \boldsymbol{X}_{\cdot,t} \\ \tilde{\boldsymbol{X}}_{\cdot,\boldsymbol{t}} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - diag(\boldsymbol{s}) \\ \boldsymbol{\Sigma} - diag(\boldsymbol{s}) & \boldsymbol{\Sigma} \end{bmatrix}\right), \tag{2}$$

where $\boldsymbol{s} = [s_1, \ldots, s_p]$ is chosen such that the covariance matrix in (2) is positive semi-definite. Equivalently, for each t with given observation $\boldsymbol{X}_{\cdot,t}$, $\tilde{\boldsymbol{X}}_{\cdot,\boldsymbol{t}}$ can be generated from the conditional distribution

$$\tilde{\boldsymbol{X}}_{\cdot,\boldsymbol{t}} | \boldsymbol{X}_{\cdot,t} \sim \mathcal{N}\left((\boldsymbol{\Sigma} - diag(\boldsymbol{s}))\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{\cdot,t} - \boldsymbol{\mu}), \boldsymbol{\Sigma} - (\boldsymbol{\Sigma} - diag(\boldsymbol{s}))\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - diag(\boldsymbol{s})\right) \tag{3}$$

Barber et al. [2015] show that the generated knockoff copies $\tilde{\boldsymbol{X}}_{\boldsymbol{j},\cdot}$ satisfy the pairwise exchangeable property with $\boldsymbol{X}_{j,\cdot}$. For the choice of $\boldsymbol{s}$, we simply consider $s_j = \min(1, 2\lambda_{\min}(\boldsymbol{\Sigma}))$ for all j, which is suggested in Barber et al. [2015]. Another suggested choice is to choose $\boldsymbol{s}$ such that the average correlation between $X_{j,t}$ and $\tilde{X}_{j,t}$ is minimum for higher power. It can be constructed by minimizing $\sum_{j=1}^{p} |1 - s_j|$ with the constraints $s_j \geq 0$ and $2\boldsymbol{\Sigma} - diag(\boldsymbol{s})$ is positive semi-definite. In practice, the mean shift $\boldsymbol{\mu}$ is unknown and hence we can only use an estimate $\hat{\boldsymbol{\mu}}$ to replace $\boldsymbol{\mu}$ in the $\tilde{\boldsymbol{X}}_{\cdot,\boldsymbol{t}}$ generation equation (3). An intuitive choice of μ_j estimate is the sample mean $\hat{\boldsymbol{\mu}} = \sum_{t=1}^{\tau^{obs}} X_{j,t}$. However, as we assume that most of the μ_j are zeros, it is reasonable to estimate μ_j by 0 if $|\hat{\mu}_j|$ is small. If we have the prior information to give a better estimate of μ_j. The bias of using $\hat{\mu}_j$ (or other estimators) instead of μ_j in (3) makes the distribution of $\tilde{X}_{j,t}$ different from that of $X_{j,t}$ even if H_j is true. Such bias can lead to the failure of FDR control of knockoff procedure. However, we can show that FDR can still be controlled for some suitable μ_j estimates under certain settings. Suppose the off-diagonal entries of the inverse of $\boldsymbol{\Sigma}$ are all non-positive and $\mu_j \geq 0$ for $j = 1, \ldots, p$. Then using 0 to estimate μ_j gives an under-estimator for μ_j. If $\boldsymbol{s}$ is chosen such that the diagonal entries of $diag(\boldsymbol{s})\boldsymbol{\Sigma}^{-1}$ are all less than or equal to 1 so that $(\boldsymbol{\Sigma} - diag(\boldsymbol{s}))\boldsymbol{\Sigma}^{-1} = (\boldsymbol{I} - diag(\boldsymbol{s})\boldsymbol{\Sigma}^{-1})$ have all entries non-negative, then replacing $\boldsymbol{\mu}$ by an under-estimator $\boldsymbol{0}$ increases the conditional mean of $\tilde{\boldsymbol{X}}_{\cdot,\boldsymbol{t}}$ in (3) and has no effect on the conditional covariance. It implies that $\tilde{Z}_j$ generated in Step 3 of the generated knockoff procedure is stochastically greater than Z_j, and hence $P(W_j > 0) = P(Z_j > \tilde{Z}_j) < \frac{1}{2}$ under H_j. Theorem 1 shows that under such situation, and other more general situations, the knockoff procedure can still control FDR.

Theorem 1 *Let V^+ be the number of W_j under H_j in Step 3 of general knockoff procedure being positive. Let $W_j^0, j = 1, \ldots, p$ and V_0^+ be the corresponding values of W_j and V^+ using the oracle true $\boldsymbol{\mu}$ in (3) to generate $\tilde{X}_{j,t}$ in Step 1. If V_0^+ is stochastically greater than V^+, i.e. $P(V_0^+ > t) \geq P(V^+ > t)$ for all $t \in \{0, 1, 2, \ldots, p_0\}$ where p_0 is the total number of null hypotheses, then the general knockoff procedure controls FDR at α level, i.e.*

$$FDR = \mathrm{E}\left(\frac{\#\{j \in H_0 : W_j \leq -t_{\min}\}}{\max\{1, \#\{j : W_j \geq t_{\min}\}\}}\right) \tag{4}$$

where $H_0 = \{j : H_j \text{ is true}\}$ is the set of true null hypotheses.

3 Application of Knockoff Procedure

3.1 Top-r Procedure

Suppose that there are m sensors in a network system and each sensor observes a time series signal $\{X_{j,t}\}_{t=1}^{\infty}$. And under global null, the stopping time T follows

$$\mathbb{E}^{(\infty)}(T) \geq \gamma$$

where $\gamma > 0$ is a pre-specified constant.

A "trigger" event changes the density function of j-th sensor from f_0^j to f_1^j at time $v_k = v + \delta_k$. In order to discover when the event happened and stop the procedure as soon as possible to avoid potential costs, top-r involved a schema named Page's CUSUM. The Page's CUSUM statistics detect the signal distribution change, $P_{k,n}$ is computed as

$$\begin{aligned} P_{j,t} &= \max\left\{0, \max_{1 \leq v \leq n} \sum_{i=v}^{n} \log \frac{f_1^j(X_{j,i})}{f_0^j(X_{j,i})}\right\} \\ &= \max\left(P_{j,t-1} + \log \frac{f_1^j(X_{j,t})}{f_0^j(X_{j,t})}, 0\right) \end{aligned} \tag{5}$$

Notice that we need two known density functions f_0^j and f_1^j to calculate the statistic. f_0^j is the density function of default distribution. f_1^j is well estimated or known ahead of time.

For ordered decreasing Page's CUSUM statistics $P_{(1),n} \geq P_{(2),n} \geq \cdots \geq P_{(k),n}$, the top-$r$ scheme raises a global alarm at time

$$N_{top,r}(a) = \inf\left\{n : \sum_{k=1}^{r} P_{(k),n} \geq a\right\} \tag{6}$$

where a is the threshold and can be selected based on **asymptotic optimality theory**:

$$a = \log \gamma + (K - 1 + o(1)) \log \log \gamma \tag{7}$$

Here we assume all the changes are instantaneous and persistent if the sensor is affected.

To apply our knockoff procedure on Top-r scheme, the stopping time function $\tilde{\tau}$ here we choose is to apply Equation (6) $N_{top,r}(a)$ on $\{P_{j,1}, P_{j,2}, \ldots, P_{j,\tau^{obs}}\} \cup \{\tilde{P}_{j,1}, \tilde{P}_{j,2}, \ldots, \tilde{P}_{j,\tau^{obs}}\}, j =$

$1, \ldots, m$. And we also use a reasonable approach to do the estimation of $\boldsymbol{\mu}$. We generate 10^3 simulations under global null and use $(1-\alpha)\%$ quantile of each simulation's max value as the threshold. We use sample mean to estimate the distribution of $\boldsymbol{\mu}$. If the sample mean of target stream is less than the threshold, we threat it as 0, which is equivalent to $\boldsymbol{\mu}_e = ifelse(\bar{X} > threshold, \bar{X}, 0)$.

3.2 Simulation Result of Knockoff Procedure on Top-r

Here we do the simulation on the system with $m = 300$ data streams and there exists some # of OC streams. Similar to Mei [2011], we also assume $f_0^j \sim \mathcal{N}(0,1)$ and $f_1^j \sim \mathcal{N}(\mu_1, 1)$. And the parameters in top-r detection scheme is $r = 30$ and $a = 232.75$. Referring to Li et al. [2020], we also choose three certain representative correlation structure for illustration:

1. Case 1. $\boldsymbol{\Sigma} = \boldsymbol{I}_{m\times m}$
2. Case 2. $\boldsymbol{\Sigma}$ is a block-diagonal matrix with block size of 10. Within each block, the diagonal elements are 1 and the off-diagonal elements are 0.4.
3. Case 3. $\boldsymbol{\Sigma} = (\sigma_{ij}) = \rho^{|i-j|}$

In the following table, the Knockoff Estimate denotes the result of using above reasonable estimation approach and Knockoff Oracle denotes the result of using oracle true $\boldsymbol{\mu}$. All the FDR and Power in the following tables are estimated by 10^3 numerical simulations.

Table 1: Case 1: Independent

		Top-r		Knockoff		
μ_1	# of OC	FDR	Power	α	FDR	Power
0.5	20	35.45%	96.82%	0.1	8.11%	79.23%
				0.2	17.97%	89.90%
	40	4.20%	71.85%	0.1	8.70%	70.89%
				0.2	19.63%	83.99%
1	20	33.41%	99.88%	0.1	8.66%	95.78%
				0.2	17.90%	97.92%
	40	0.15%	74.89%	0.1	9.13%	92.08%
				0.2	19.14%	95.79%

Table 2: Case 2: Short-Range Correlation

		Top-r		Knockoff Estimate			Knockoff Oracle	
μ_1	# of OC	FDR	Power	α	FDR	Power	FDR	Power
0.5	20	35.75%	96.38%	0.1	8.91%	84.60%	8.24%	78.89%
				0.2	19.43%	93.62%	18.70%	89.77%
	40	4.10%	71.92%	0.1	4.88%	70.05%	8.72%	72.24%
				0.2	13.76%	85.08%	19.18%	83.35%
1	20	33.43%	99.86%	0.1	9.40%	96.18%	8.67%	95.60%
				0.2	18.51%	98.42%	18.17%	97.88%
	40	0.19%	74.86%	0.1	9.69%	92.62%	9.15%	92.01%
				0.2	19.96%	96.63%	19.23%	95.98%

Table 3: Case 3: Long-Range Correlation

			Top-r		Knockoff Estimate			Knockoff Oracle	
ρ	μ_1	# of OC	FDR	Power	α	FDR	Power	FDR	Power
0.50	0.5	20	35.64%	96.54%	0.1	6.08%	85.28%	8.32%	89.88%
					0.2	16.80%	95.56%	18.41%	96.01%
		40	4.25%	71.82%	0.1	2.08%	56.41%	8.51%	83.58%
					0.2	9.61%	82.78%	19.37%	91.62%
	1	20	33.41%	99.88%	0.1	8.70%	98.92%	9.09%	99.00%
					0.2	18.45%	99.70%	18.60%	99.60%
		40	0.17%	74.88%	0.1	8.70%	97.32%	9.23%	97.22%
					0.2	19.69%	98.87%	19.28%	98.85%
-0.50	0.5	20	35.51%	96.73%	0.1	9.98%	91.78%	8.70%	90.44%
					0.2	20.45%	97.12%	18.87%	96.24%
		40	4.37%	71.73%	0.1	13.98%	88.92%	8.71%	83.17%
					0.2	24.80%	94.21%	19.28%	91.66%
	1	20	33.39%	99.92%	0.1	9.02%	99.01%	8.56%	98.86%
					0.2	19.18%	99.71%	18.57%	99.69%
		40	0.13%	74.90%	0.1	9.66%	97.48%	8.58%	97.03%
					0.2	20.13%	98.95%	18.71%	98.86%

3.3 The False Discovery Rate–Adjusted Shewhart Chart

Knockoff can be used as a general method in censoring sensor network not only limited on Top-r. Here we provide an application on another model in sensor network. First, we are going to introduce the state-space model which can characterize a multistage process by incorporating physical laws and engineering knowledge.

3.3.1 Background of State-Space Model

Suppose that the multistage process involves m stages and $y_{j,t}$ represents the quality measurement at j-th stage of t-th product. The in-control state-space model of this process can be described as

$$\begin{aligned} y_{j,t} &= H_j x_{j,t} + \nu_{j,t} \\ x_{j,t} &= F_j x_{j-1,t} + \omega_{j,t} \end{aligned} \tag{8}$$

where $x_{j,t}$ denotes the quality information of t-th product at j-th stage, $\mathbf{F}_j$ denotes the transformation matrix from stage $j-1$ to stage j, and $\mathbf{H}_j$ denotes the observation matrix representing the relation between quality information $x_{j,t}$ and quality measurement $y_{j,t}$. Both $\mathbf{F}_j$ and $\mathbf{H}_j$ are assumed to be known, which can be derived from other information of the process. Here ω and ν are error terms. In this paper, we follow the assumption considered by Xiang and Tsung [2008], where $\omega_{j,t} \sim \mathcal{N}(0, \sigma^2_{\omega_t}), \nu_{j,t} \sim \mathcal{N}(0, \sigma^2_{\nu})$. The initial state follows $x_0 \sim \mathcal{N}(a_0, \sigma_0^2)$, where a_0 and σ_0 are assumed known in advance.

The out-of-control stage in this process can be modeled as an addition term on the quality information $x_{j,t}$. Suppose that after a specific product, there exist a mean shift of magnitude δ at stage ξ. Similar to the change-point mentioned above, this product is also assumed to be known. The out-of-control model can be described as

$$\begin{aligned} y_{j,t} &= H_j x_{j,t} + \nu_{j,t} \\ x_{j,t} &= F_j x_{j-1,t} + \omega_{j,t} + \mathbb{1}_{\{t=\xi\}}\delta \end{aligned} \tag{9}$$

Based on our in-control state-space model (8), the standardized one-step-head forecast error of $y_{j,t}$ given $y_{j-1,t}$, as $e_{j,t}$, which can be calculated through the following recursive model:

$$\begin{aligned} e_{j,t} &= v_{j,t} V_j^{-1/2} \\ v_{j,t} &= y_{j,t} - H_j u_{j,t} \\ u_{j,t} &= F_{j-1} u_{j-1,t} + G_{j-1,t} v_{j-1,t} \\ V_j &= H_j^2 Q_j + \sigma_v^2 \\ Q_j &= F_{j-1}^2 Q_{j-1} - F_{j-1} H_{j-1} Q_{j-1} G_{j-1} + \sigma_{\omega_j}^2 \\ G_j &= F_j H_j Q_j V_j^{-1} \end{aligned} \tag{10}$$

where the initial values are $v_{1,t} = y_{1,t} - a_0, e_{1,t} = v_{1,t}/V_1, u_{1,t} = 0, W_1 = \sigma_{\omega_1}^2 + \sigma_0^2$. Durbin and Koopman [2012] already showed that under the in-control state-space model, $e_{j,t}$ is independently and identically distributed as a standard normal distribution, $\mathcal{N}(0,1)$. When a particular stage ξ is out-of-control, $e_{\xi,t}$ will experience a mean shift while the variance remains the same. Thus the multi-stage process monitoring and diagnosis may be formulated into the following multiple hypotheses testing problem:

$$\begin{aligned} H_{0,j}: &\quad e_{j,t} \sim \mathcal{N}(0,1) \\ H_{1,j}: &\quad e_{j,t} \sim \mathcal{N}(\mu_j,1) \end{aligned} \tag{11}$$

$j = 1, \ldots, N$ where $\mu_j \neq 0$.

BHq procedure is proved to be able to control FDR at level $n_0 q/n$. And to improve the power of BHq procedure, Benjamini et al. [2006] proposed a two-stage linear step-up procedure, which can also control FDR at level q. Here is the procedure:

1. Use the BHq procedure at level $q' = q/(1+q)$. Let r_1 denotes the number of rejections. If $r_1 = 0$, reject nothing and stop; if $r_1 = n$, reject all m hypotheses and stop.
2. Let $\hat{n}_0 = n - r_1$.
3. Use the BHq procedure at level $q^* = q'n/\hat{n}_0$.

3.3.2 The FDR–Adjusted Shewhart Chart

When the process is in control, the standardized one-step-ahead forecast errors follow standard normal distributions. Hence, the p-value can be calculated as

$$p_{j,t} = 2\left(1 - \Phi(|e_{j,t}|)\right) \tag{12}$$

where Φ is the cumulative density function of the standard normal distribution.

The signaling time of the FDR-adjusted Shewhart chart, T , can be computed by

$$\begin{aligned} l &= \max\{n : p_{(n),t} \leq n\alpha/N, 1 \leq n \leq N\}, \\ T &= \inf\{t : l \geq 1\}. \end{aligned} \tag{13}$$

where $p_{(n),j}$ denotes the nth smallest p-value.

The step-by-step procedure of FDR-adjusted Shewhart chart can be described as follows:

1. Model a multistage process with N stages by a in-control state-space model as in (8), assuming that all of the parameters in the model are known.
2. Choose the in-control average run length (ARL), ARL_0. The significance level α can be determined later.
3. For product j, calculate the standardized one-step-ahead forecast error of each stage, $e_{1,t}, e_{2,t}, ..., e_{N,t}$ by (10), and model the multiple hypothesis testing problem as in (11).

4. Calculate the corresponding p-values of standardized one-step-ahead forecast error of each stage by (12).
5. Find the largest l using the two-stage linear step-up procedure.
6. If $l > 0$, the process is detected as out-of-control. The stages associated with the reject hypotheses are identified as faulty stages. Otherwise, repeat the step 3-6 for product $t+1$.

3.4 Knockoff Procedure on FDR-adjusted Shewhart chart

We notice that directly applying BHq procedure on the standardized one-step-ahead forecast errors cannot achieve FDR control. Here we apply our knockoff procedure on the FDR-Adjusted Shewhart Chart to control FDR. Referring to the setting in the FDR-adjusted Shewhart chart proposed by Li and Tsung [2009], we assume the values H_{ij} in observation matrix $\mathbf{H}_j$ are identical constants, which can be denoted as H. Here we introduced a new test statistic, the difference between quality measurement, denote as $d_{j,t}$:

$$d_{j,t} = y_{j,t} - F_j y_{j-1,t}, \tag{14}$$

and set $d_{1,t} = y_{1,t} = Hx_{1,t} + \nu_{n,t} = HF_1 x_{0,t} + H\omega_{1,t} + v_{1,t}$. This can be further simplified to be

$$d_{j,t} = H\omega_{j,t} + \nu_{j,t} - F_j \nu_{j-1} + H\mathbb{1}_{t=\xi}\delta \text{ , where } j \neq 1$$

The covariance matrix can also be computed, since

$$\begin{aligned} \mathrm{Var}(d_{j,t}) &= H^2\sigma^2_{\omega_j} + (1+F_j^2)\sigma^2_\nu \text{ , where } j \neq 1 \\ \mathrm{Cov}(d_{j,t}, d_{j-1,t}) &= -F_j\sigma^2_\nu \end{aligned}$$

Hence, $\mathbf{d} \sim \mathcal{N}(\boldsymbol{\mu}^d, \boldsymbol{\Sigma}^d)$, where

$$\boldsymbol{\mu}^d_i = \begin{cases} HF_1a_0 + H \times \mathbb{1}_{t=\xi}\delta & i = 1 \\ H \times \mathbb{1}_{t=\xi}\delta & i \neq 1 \end{cases}$$

$$\boldsymbol{\Sigma}^d_{ij} = \begin{cases} H^2F_1^2\sigma_0^2 + H^2\sigma^2_{\omega_1} + \sigma^2_\nu & i = j = 1 \\ H^2\sigma^2_{\omega_j} + (1+F_j^2)\sigma^2_\nu & i = j \neq 1 \\ -F_j\sigma^2_\nu & |i-j| = 1 \\ 0 & \text{Otherwise} \end{cases}$$

Similar to the knockoff on top-r, here we construct the knockoff dummy variable first. Let $\mathbf{d}_{\cdot,t}$ denotes the difference statistic of product t. The knockoff dummy variable $\tilde{\mathbf{d}}_\mathbf{j}$ can be construct by

$$\begin{bmatrix} \boldsymbol{d}_{\cdot,t} \\ \tilde{\boldsymbol{d}}_{\cdot,t} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}^d \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}^d & \boldsymbol{\Sigma}^d - diag(\boldsymbol{s}) \\ \boldsymbol{\Sigma}^d - diag(\boldsymbol{s}) & \boldsymbol{\Sigma}^d \end{bmatrix} \right)$$

Condition on the observed value, the dummy variable can be sampling from

$$\tilde{\boldsymbol{d}}_{\cdot,t} | \boldsymbol{d}_{\cdot,t} \sim \mathcal{N}\left((\boldsymbol{\Sigma}^d - diag(\boldsymbol{s}))(\boldsymbol{\Sigma}^d)^{-1}(\boldsymbol{d}_{\cdot,t} - \boldsymbol{\mu}^d), \boldsymbol{\Sigma}^d - (\boldsymbol{\Sigma}^d - diag(\boldsymbol{s}))(\boldsymbol{\Sigma}^d)^{-1}(\boldsymbol{\Sigma}^d - diag(\boldsymbol{s})) \right). \tag{15}$$

To apply general knockoff procedure on FDR-Adjusted Shewhart Chart, the new stopping time $\tilde{\tau}$ we used here is apply original stopping rule τ on the smallest p p-values among totally $2p$ p-values for the original data streams and knockoff copies, which can also guarantee that $\tau^{KF} \leq \tau^{obs}$. And, the true distribution of $\boldsymbol{\mu}^d$ is also unknown. Similar to the knockoff on top-r scheme, we can also use the $(1-\alpha)\%$ quantile of the maximum value for each data streams under global null as the threshold to estimate $\boldsymbol{\mu}_d$.

3.5 Simulation Result of Knockoff Procedure on FDR-adjusted Shewhart chart

Referring to the setting in Li and Tsung [2009], here we assume all the correlation among the process stages can be described by a simple state-space model with identity constants, $F_t = H_t = 1$ in (8). The other parameters in the state-space model are $\sigma_\nu = 1$, $\sigma_{\omega_n} = 1$, and, for the initial state, $a_0 = 0$ and $\sigma_0 = 1$. The process involve $N = 300$ operation stages and some the data streams are set to be OC streams. The in-control average run length (ARL_0) here is specified as 500. In the following table, we show the comparison of both FDR and power between FDR-adjusted Shewhart chart, knockoff with reasonable $\boldsymbol{\mu}^d$ estimate approach described above and knockoff with oracle true $\boldsymbol{\mu}^d$. All the FDR and Power in the following tables are estimated by 10^3 numerical simulations.

Table 4: Knockoff on FDR-adjusted Shewhart Chart

		FDR-adjusted (alpha=0.2%)			Kncokoff			Knockoff Oracle	
δ	# of OC	ARL	FDR	Power	α	FDR	Power	FDR	Power
0.5	10	462.81	94.25%	0.60%	0.10	5.72%	58.05%	6.03%	59.75%
					0.20	15.43%	78.45%	16.86%	80.25%
	20	418.91	81.00%	0.95%	0.10	9.12%	72.38%	8.09%	75.55%
					0.20	18.26%	83.58%	17.15%	84.58%
1	10	391.39	80.00%	2.00%	0.10	8.98%	91.60%	7.49%	88.65%
					0.20	16.84%	92.95%	17.21%	94.55%
	20	312.54	61.25%	2.08%	0.10	9.98%	91.25%	7.93%	90.40%
					0.20	19.56%	91.85%	18.92%	94.60%
1.5	10	251.69	49.00%	5.15%	0.10	8.49%	92.15%	7.84%	92.85%
					0.20	16.56%	95.90%	17.43%	93.30%
	20	162.45	41.75%	2.98%	0.10	8.52%	93.65%	8.15%	92.65%
					0.20	19.93%	95.15%	19.32%	94.73%
2	10	113.52	29.25%	7.20%	0.10	7.82%	84.05%	6.66%	88.90%
					0.20	17.99%	94.75%	18.01%	95.75%
	20	61.89	19.00%	4.15%	0.10	8.61%	88.20%	7.49%	88.35%
					0.20	19.65%	92.38%	17.82%	89.93%
5	10	1.61	0.92%	18.35%	0.10	6.64%	33.40%	5.33%	30.75%
					0.20	20.05%	64.80%	17.58%	70.00%
	20	1.095	1.10%	18.90%	0.10	11.25%	53.30%	6.71%	46.03%
					0.20	23.63%	73.45%	17.51%	73.40%
8	10	1	2.32%	82.50%	0.10	8.80%	67.70%	7.46%	73.30%
					0.20	17.79%	96.90%	19.54%	97.50%
	20	1	3.89%	89.05%	0.10	9.57%	96.28%	8.27%	97.18%
					0.20	19.26%	98.10%	19.13%	99.03%

4 Empirical Study

In this section, we apply the top-r thresholding scheme and our knockoff procedure to a real dataset from a semiconductor manufacturing process, which is under consistent surveillance via the monitoring of signals/variables collected from sensors and or process measurement points. This dataset is publicly available in the UC Irvine Machine Learning Repository. It contains

590 data streams and 1567 observations. Among the total 1567 observations, 1463 observations are classified as conforming (IC) based on the physical testing and remaining 104 observations are classified as nonconforming (OC).

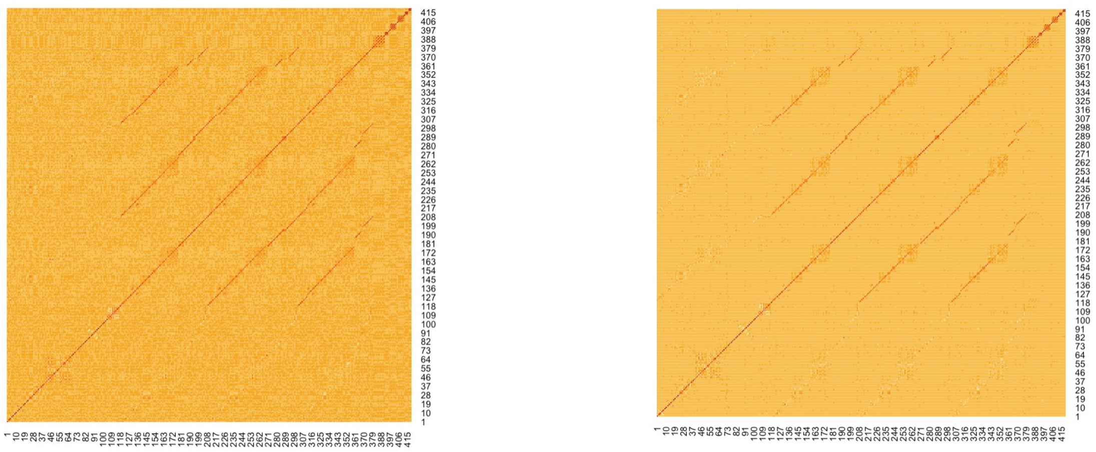
First, we do the data preprocessing to deal with the constant data streams and the missing values. Since the fraction of missing values is trivial in this dataset, we simply use mean imputation (replace each missing values with the mean of the observed values for that variable). Some data streams remain constant or are very discrete (a large proportion of the observations share the same constant), we simply remove these data streams. After that, there are $p = 419$ data streams remained and will be used for the following analysis.

Then, we are going to verify the multivariate normal assumption of this dataset. We conduct a Shapiro-Wilks GOF test for normality and conclude that many data streams are not normally distributed (i.e., the p-values are less than 0.01). To make sure that the normality assumption is approximately valid, we preform an inverse transformation on all the observations, say $\Phi^{-1}\left(\hat{F}_{nk}(X(k))\right), k = 1, \ldots, t$, where $X(k)$ is the kth data stream, $\hat{F}_{nk}(X(k))$ is the empirical distribution function based on the 1463 IC observations of the kth data streams. We apply this transformation on both IC observations and OC observations. It should be noted that this transformation cannot guarantee the normality assumption is valid.

Under the multivariate normal assumption, we estimate the sample covariance matrix by maximum likelihood estimation. Here the sample size 1463 may be smaller than ideal to properly determine the IC correlation structure by considering the number of data streams is as high as 419, and some diagonal entries are too small, which is not ideal for our knockoff procedure. We apply some adjustments to the covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})$.

1. For the off-diagonal entries σ_{ij}, if $\sigma_{ij} \leq 0.1$, $\sigma_{ij} = 0$.
2. Apply the eigen-decomposition to the $\mathbf{\Sigma}$, since $\mathbf{\Sigma}$ is symmetric, $\mathbf{\Sigma} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$.
3. For the diagonal entries of $\boldsymbol{D}$, denoted as D_{ii}, if $D_{ii} <= 0.2$, $D_{ii} = 0.2$. After this adjustment, the diagonal matrix denoted as $\hat{\boldsymbol{D}}$.
4. The adjusted covariance matrix is computed by $\hat{\mathbf{\Sigma}} = \boldsymbol{V}\hat{\boldsymbol{D}}\boldsymbol{V}^T$.

Here we provide the heapmap of the covariance matrix before and after the adjustment in Figure 1. We can see that there is no significant changes, the major correlation structure is remained.



(a) Sample Covariance Matrix $\mathbf{\Sigma}$

(b) Adjusted Covariance Matrix $\hat{\mathbf{\Sigma}}$

Figure 1: HeatMap of the Correlation Structure

We first apply the top-r thresholding scheme with $r = 25, 50, 75, 100$ on this dataset, and the thresholding parameter a is chosen given the fact that under IC data, the average stopping time is around 400. Table 5 shows the result of this application. The corresponding stopping time is 51, 36, 28 and 27. The stopping time decreases for r increases from 25 to 75 and remains steady when r increases from 75 to 100, which indicates that except from a few strong signals there exists several weak signals in this semiconductor manufacturing process.

Table 5: Simulation Result of Empirical Study

r	a	t	t_{kf}	# rej
25	302	51	50.52(0.70)	39.57(7.85)
50	375	36	31.97(2.08)	25.97(7.53)
75	430	28	26.71(0.46)	23.37(8.83)
100	480	27	24.92(0.37)	22.97(5.92)

Theoretically, our knockoff procedure just need to apply once in order to control the FDR. However, since there exits randomness in the knockoff procedure, the rejection set of the knockoff procedure might change dynamically in each execution. To make the result of knockoff procedure more robust, we just repeatedly execute our knockoff procedure 100 times and compute the rejection proportion probability of each data stream. We denote the average rejection number of our knockoff procedure as k, and reject the top k data streams by their rejection proportion probability. Under significant level $\alpha = 0.1$, the number of rejection by our knockoff procedure is 54, 32, 27 and 25. When the stopping time decreases, we can use less information, which would leads to the decrease in power. We would expect that our rejection set would contain majority of the strong signals of OC streams and few (around 10%) IC streams.

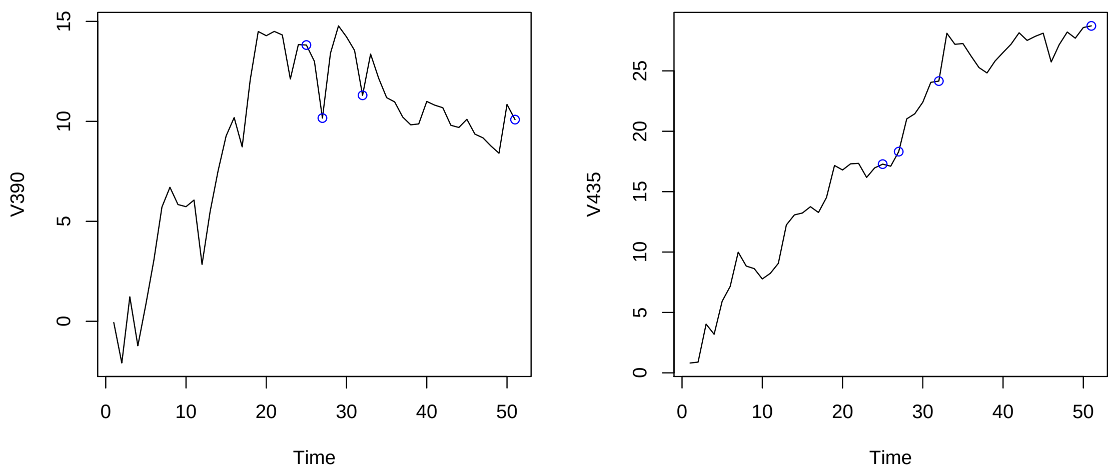


Figure 2: CUSUM Statistics of Stream V390 and V435

Table 6: Proportion of Rejection by Knockoff Procedure

Stream #	r=25	r=50	r=75	r=100
V39	1	0.94	0.97	0.92
V60	1	1	0.98	0.99
V161	1	1	0.98	0.98
V551	1	1	0.98	0.99
V554	1	1	0.93	0.95
V557	1	1	0.98	0.99
V206	1	0.98	0.05	0.03
V104	1	0.97	0.76	0.84
V113	1	0.95	0.49	0.56
V342	1	0.95	0.22	0.29
V248	1	0.93	0.82	0.84
V296	1	0.49	0.98	0.99
V432	1	0.47	0.98	0.98
V214	1	0.01	0.08	0.18
V352	1	0.01	0.04	0.13
V288	1	0	0.01	0.06
V156	0.99	0	0	0
V511	0.99	0.99	0.92	0.82
V386	0.99	0.98	0.98	0.99
V520	0.99	0.86	0.6	0.62
V429	0.99	0	0	0
V291	0.98	0	0.04	0.11
V125	0.95	0.14	0.05	0.2
V478	0.93	0.66	0	0
V349	0.92	0.3	0.35	0.36
V80	0.89	0.09	0.04	0.02
V38	0.88	0.67	0.35	0.27
V130	0.8	0.02	0	0.01
V22	0.79	0.58	0.92	0.87
V299	0.76	0.97	0.89	0.98
V34	0.7	0.33	0.28	0.62
V164	0.67	0.95	0.85	0.93
V211	0.65	0.46	0.32	0.42
V37	0.52	0.02	0.01	0
V57	0.52	0.01	0	0
V369	0.52	0.6	0.32	0.19
V358	0.46	0.04	0.03	0.03
V176	0.45	0.3	0.05	0.08
V64	0.43	0	0.01	0.02
V426	0.41	0	0.01	0.02
V435	0.01	1	0.98	0.99
V390	0.05	0.77	0.68	0.86
V272	0.19	0.43	0.38	0.35
V56	0.1	0.35	0.5	0.49

A Proof of Theorem 1

Let $S^+(t) = \{j : W_j \geq t\}, S^-(t) = \{j : W_j \geq -t\}$, then the selected set is $\hat{S} = S^+(t_{\min})$. Note that, by the construction of W_j, $P(W_j = 0) = 0$. Without loss of generality, we assume $|W_1| \geq |W_2| \geq \cdots \geq |W_p| > 0$, and then we can rewrite $\hat{S} = S^+\left(|W_{\hat{k}}|\right) = \left\{j \leq \hat{k} : W_j > 0\right\}$, where

$$\hat{k} = \max\left\{k : \frac{1 + |S^-\left(|W_k|\right)|}{\max\{1, |S^+\left(|W_k|\right)|\}} \leq \alpha\right\}.$$

Note that $|S^-\left(|W_{\hat{k}}|\right)| = \left\{j \leq \hat{k} : W_j < 0\right\}$ and hence

$$FDP = \frac{\#\{j \in H_0 : j \in |S^+\left(|W_{\hat{k}}|\right)|\}}{\max\{1, |S^+\left(|W_{\hat{k}}|\right)|\}\}} \leq \alpha \frac{V^+(\hat{k})}{1 + V^-(\hat{k})} \tag{A1}$$

where $V^+(k) = \#\{j \in H_0 : 1 \leq j \leq k, W_j > 0\}$ and $V^-(k) = \#\{j \in H_0 : 1 \leq j \leq k, W_j < 0\}$. By the same argument of the proof Lemma 4 in Barber et al. [2015], we have

$$\mathrm{E}\left(\frac{V^+(\hat{k})}{1 + V^-(\hat{k})}\right) \leq \mathrm{E}\left(\frac{V^+(p)}{1 + V^-(p)}\right) = \mathrm{E}\left(\frac{V^+}{1 + p_0 - V^+}\right). \tag{A2}$$

Since V_0^+ is stochastically greater than V^+, and $f(x) = x/(1 + p_0 - x)$ is nondecreasing, we have

$$\mathrm{E}\left(\frac{V^+}{1 + p_0 - V^+}\right) \leq \mathrm{E}\left(\frac{V_0^+}{1 + p_0 - V_0^+}\right) \leq 1. \tag{A3}$$

The last inequality comes from the proof Lemma 4 in Barber et al. [2015]. Combining (A1), (A2) and (A3) yields the result in Theorem 1.

References

M. Bagshaw and R. A. Johnson. The influence of reference values and estimated variance on the arl of cusum tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(3):413–420, 1975.

R. F. Barber, E. J. Candès, et al. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

R. F. Barber, E. J. Candès, et al. A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537, 2019.

Y. Benjamini, A. M. Krieger, and D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

G. Capizzi and G. Masarotto. A least angle regression control chart for multidimensional data. *Technometrics*, 53(3):285–296, 2011.

L. Du and C. Zou. On-line control of false discovery rates for multiple datastreams. *Journal of Statistical Planning and Inference*, 194:1–14, 2018.

J. Durbin and S. J. Koopman. *Time series analysis by state space methods*. Oxford university press, 2012.

C.-K. Ing, T. L. Lai, M. Shen, K. Tsang, and S.-H. Yu. Multiple testing in regression models with applications to fault diagnosis in the big data era. *Technometrics*, 59(3):351–360, 2017.

W. Jiang, K. Wang, and F. Tsung. A variable-selection-based multivariate ewma chart for process monitoring and diagnosis. *Journal of Quality Technology*, 44(3):209–230, 2012.

J. Li, J. Jin, and J. Shi. Causation-based t^2 decomposition for multivariate process monitoring and diagnosis. *Journal of Quality Technology*, 40(1):46–58, 2008.

W. Li, X. Pu, F. Tsung, and D. Xiang. A robust self-starting spatial rank multivariate ewma chart based on forward variable selection. *Computers & Industrial Engineering*, 103:116–130, 2017.

W. Li, D. Xiang, F. Tsung, and X. Pu. A diagnostic procedure for high-dimensional data streams via missed discovery rate control. *Technometrics*, 62(1):84–100, 2020.

Y. Li and F. Tsung. False discovery rate-adjusted charting schemes for multistage process monitoring and fault identification. *Technometrics*, 51(2):186–205, 2009.

W. Liang, D. Xiang, and X. Pu. A robust multivariate ewma control chart for detecting sparse mean shifts. *Journal of Quality Technology*, 48(3):265–283, 2016.

R. L. Mason, N. D. Tracy, and J. C. Young. Decomposition of t^2 for multivariate control chart interpretation. *Journal of quality technology*, 27(2):99–108, 1995.

R. L. Mason, N. D. Tracy, and J. C. Young. A practical approach for interpreting multivariate t^2 control chart signals. *Journal of Quality Technology*, 29(4):396–406, 1997.

Y. Mei. Quickest detection in censoring sensor networks. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2148–2152. IEEE, 2011.

K. Wang and W. Jiang. High-dimensional process monitoring and fault isolation via variable selection. *Journal of Quality Technology*, 41(3):247–258, 2009.

W. H. Woodall and D. C. Montgomery. Some current directions in the theory and application of statistical process monitoring. *Journal of Quality Technology*, 46(1):78–94, 2014.

X. Xian, A. Wang, and K. Liu. A nonparametric adaptive sampling strategy for online monitoring of big data streams. *Technometrics*, 60(1):14–25, 2018.

L. Xiang and F. Tsung. Statistical monitoring of multi-stage processes based on engineering models. *IIE transactions*, 40(10):957–970, 2008.

K. Zamba and D. M. Hawkins. A multivariate change-point model for statistical process control. *Technometrics*, 48(4):539–549, 2006.

Y. Zhu and W. Jiang. An adaptive t^2 chart for multivariate process monitoring and diagnosis. *IIE Transactions*, 41(11):1007–1018, 2009.

C. Zou, Z. Wang, X. Zi, and W. Jiang. An efficient online monitoring method for high-dimensional data streams. *Technometrics*, 57(3):374–387, 2015.