Received: 12 August 2019 | Revised: 6 September 2019 | Accepted: 12 September 2019
DOI: 10.1002/sta4.252

WILEY

ORIGINAL ARTICLE

Which principal components are most sensitive in the change detection problem?

Martin Tveten

Department of Mathematics, University of Oslo, Oslo, Norway

Correspondence
Martin Tveten, Department of Mathematics, University of Oslo, Niels Henrik Abels hus, Moltke Moes vei 35, 0851 Oslo, Norway.
Email: martintv@math.uio.no

Funding information
Norges Forskningsråd, Grant/Award Number: 237718

Principal component analysis (PCA) is often used in anomaly detection and statistical process control tasks. For bivariate normal data, we prove that the minor projection (the least varying projection) of the PCA-rotated data is the most sensitive to distributional changes, where sensitivity is defined as the Hellinger distance between the projections' marginal distributions before and after a change. In particular, this is almost always the case if only one parameter of the bivariate normal distribution changes, that is, the change is sparse. Simulations indicate that the minor projections are the most sensitive for a large range of changes and pre-change settings in higher dimensions as well, including changes that are very sparse. This motivates using only a few of the minor projections for detecting sparse distributional changes in high-dimensional data.

KEYWORDS
machine learning, quality control, statistical process control

1 | INTRODUCTION

It is popular to use principal component analysis (PCA) for anomaly detection and stochastic process control (SPC). Using PCA in SPC goes back to the work of Jackson and Morris (1957) and Jackson and Mudholkar (1979), and its various extensions (see Ketelaere et al., 2015 and Rato et al., 2016, for an overview) have been successfully applied to many real data situations. Within the machine learning literature on anomaly detection, Mishin et al. (2014) use PCA for temperature monitoring at Johns Hopkins, Harrou et al. (2015) apply PCA-based anomaly detection to find segments with abnormal rates of patient arrivals at an emergency department, and Camacho et al. (2016) relate PCA-based monitoring in SPC to modern anomaly detection in statistical networks. PCA has also been studied in the setting of change detection in multivariate functional data with the aim of detecting faulty profiles in a forging manufacturing process (Wang et al., 2018). Pimentel et al. (2014) provide an extensive review of novelty detection techniques and applications, and it is pointed to PCA being very useful for detecting outliers in this setting, for a large range of real world examples, covering industrial monitoring, video surveillance, text mining, sensor networks, and IT security. Moreover, many authors (Huang et al., 2007; Lakhina et al., 2004; Pimentel et al., 2014) acknowledge that it is most often the residual subspace of PCA that is most useful for outlier detection. On a similar note, Kuncheva and Faithfull (2014) offer an interesting alternative way to use PCA for change detection problems.

Most PCA-based methods utilize PCA in the intended way of creating a model based on retaining a small number of the most varying projections onto eigenvectors of the covariance matrix. As a consequence, the data are split into a model subspace that explains most of the variance in the data and a residual subspace. It is not self-evident that this is the best way to use PCA as a dimension reduction tool for change detection, so Kuncheva and Faithfull (2014) pose the question of which projections are the most sensitive to distributional changes in the data. Sensitivity is measured by a statistical divergence between the marginal distributions of projections before and after a change. They give a brief two-dimensional theoretical example that motivates monitoring the minor projections (the least varying projections) to detect anomalies that manifest in the form of sustained changes in the distribution of the data. An important feature of such an approach is that it can potentially be used to choose a subspace based on criteria linked to change detection, rather than on retaining data variance, hopefully yielding a better change and anomaly detection methods. The goal of this article is to give a more complete treatment of and extend the bivariate problem of Kuncheva and Faithfull (2014) in order to better understand the projections' sensitivity to changes under a simple setup and then study how these results carry over to higher dimensions by simulations.

There are three main differences between our approach and the approach of Kuncheva and Faithfull (2014). First, we express the projections' sensitivity to changes as functions of the parameters of the original data rather than of the parameters of the projections. The reason for this choice is that the original data are the object of the main interest, whereas the projections are ancillary. Our approach allows one to change individual parameters of the original data independently and see how this affects the marginal distributions of the projections as a consequence. We argue that this is more informative. Second, we study a much larger space of possible changes, including changes in only one parameter at a time. Such change scenarios where only a few of the dimensions change are called *sparse changes*, and they are the subject of much current interest (Chan, 2017; Liu et al., 2017; Wang et al., 2018; Wang & Samworth, 2018; Xie & Siegmund, 2013). Third, we measure sensitivity by the normal Hellinger distance between the marginal distributions of projections before and after a change, whereas Kuncheva and Faithfull (2014) use the normal Bhattacharyya distance. See Section 2 for an explanation of this choice.

In short, we find the following. For bivariate data, we prove that if only one of the two components' means changes in any direction, one component's variance increases, or the correlation between the components changes, the minor projection is the most sensitive. The principal projection is the most sensitive if one of the components' variance decreases and the correlation is not too close to 1. Lastly, if both means change, which projection is the most sensitive depends on the relative directions and sizes of change, and when both variances change by an equal amount, both projections are equally sensitive. Thus, on average (with all change scenarios up to a certain size equally likely), the minor projection is the most sensitive, mainly due to the sparse change scenarios. Our simulations confirm that the trend of the minor projections being more sensitive on average also holds for higher dimensions. Moreover, and most importantly, the minor projections seem to be quite sensitive even to very sparse changes. This knowledge carries large potential for creating more efficient change or anomaly detection methods.

The rest of the article is organized as follows: Section 2 formulates the problem precisely, Section 3 contains the theoretical results about sensitivity to changes in two dimensions, and in Section 4, we explore sensitivity in higher dimensions by simulations. The proofs are found in Appendix A.

2 | PROBLEM FORMULATION

Consider independent observations $x_t \in \mathbb{R}^D$, $t = 1, \ldots, n$, and let $\kappa \in \{1, \ldots, n-1\}$ be a change-point. For $t \leq \kappa$, the observations have mean $\boldsymbol{\mu_0}$ and covariance matrix Σ_0, whereas for $t > \kappa$, the data have mean μ_1 and covariance matrix Σ_1. Assume without loss of generality that the data are standardized with respect to the pre-change parameters, so that $\boldsymbol{\mu_0} = \mathbf{0}$ and Σ_0 is a correlation matrix with correlation parameter ρ. For $D = 2$, the changed mean is given by $\boldsymbol{\mu_1} = (\mu_1, \mu_2)^t$, and the changed covariance matrix can be expressed in terms of Σ_0 and parameter-wise multiplicative change factors as

$$\Sigma_1 = \begin{pmatrix} a_{11}^2 & a_{11}a_{22}a_{12}\rho \\ a_{11}a_{22}a_{12}\rho & a_{22}^2 \end{pmatrix},$$

where

$$-1 < \rho, a_{12}\rho < 1 \text{ and } \rho \neq 0. \tag{1}$$

For example, if $a_{11} = 2$, it means that the standard deviation of the first component has doubled compared with what it was originally in Σ_0. Similarly, $a_{12} = 0.5$ means that the correlation is half as strong after the change. Note that we exclude the degenerate cases of correlations equal to -1 and 1.

Next, let $\{\lambda_j, v_j\}_{j=1}^D$ be the normalized eigensystem of Σ_0, ordered by $\lambda_1 \geq \ldots \geq \lambda_D$. The orthogonal projections $y_{j,t} = v_j^t x_t$, with progressively decreasing variances λ_j, are our main objects of interest.

The general problem is to find out which of the D projections are the most sensitive to different distributional changes defined by $(\boldsymbol{\mu_1}, \Sigma_1)$, for each pre-change correlation matrix Σ_0. In the bivariate case, $(\Sigma_0, \boldsymbol{\mu_1}, \Sigma_1)$ is fully specified by $(\rho, \mu_1, \mu_2, a_{11}, a_{12}, a_{22})$. Note that a collection of the most and least varying $y_{j,t}$'s is referred to as the *principal projections* and *minor projections*, respectively.

We define sensitivity to changes as the normal Hellinger distance between the marginal distribution of a projection before and after a change. The squared Hellinger distance between two normal distributions $p(x) = N(x|\xi_1, \sigma_1^2)$ and $q(x) = N(x|\xi_2, \sigma_2^2)$ is given by

$$H^2(p, q) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp\left\{-\frac{1}{4}\frac{(\xi_1 - \xi_2)^2}{\sigma_1^2 + \sigma_2^2}\right\}.$$

The formal definition of sensitivity to changes is contained in Definition 1.

Definition 1. For $j = 1, \ldots, D$, let p_j and q_j denote the marginal pre- and post-change density functions of $y_{j,t}$, respectively, given by

$$p_j(y) = N(y \mid v_j^T \mu_{\mathbf{0}}, v_j^T \Sigma_0 v_j) = N(y|0, \lambda_j),$$
$$q_j(y) = N(y \mid v_j^T \mu_{\mathbf{1}}, v_j^T \Sigma_1 v_j).$$

The sensitivity of the jth projection based on Σ_0 to the change specified by $(\boldsymbol{\mu_1}, \Sigma_1)$ is defined as $H(p_j, q_j)$, abbreviated by H_j or $H_j(\Sigma_0, \boldsymbol{\mu_1}, \Sigma_1)$.

Our aim in the next section is to determine which pre-change parameters and changes the inequality $H_2 > H_1$ holds for when $D = 2$ in light of Definition 1.

Remark

(i) Kuncheva and Faithfull (2014) also define sensitivity as a divergence between distributions before and after a change but use the Bhattacharyya distance. The closely related Hellinger distance was chosen here because it turns out to be simpler to prove the sensitivity propositions because of Lemma 1 (see Appendix A). It is also an advantageous feature of the Hellinger distance that it is a true metric and takes values in $[0, 1]$. That it is a true metric implies for instance that a change in variance from 1 to $a > 1$ is an equally large change as from 1 to $1/a$ for the normal distribution. We find this an appealing feature because it is also a property of the generalized likelihood ratio test for a change in the mean and/or variance of normal data (see Hawkins & Zamba, 2005, for the corresponding test statistic).

(ii) One of the differences between our approach and the work of Kuncheva and Faithfull (2014) can now be stated more precisely. Our aim is to study the sensitivity of the $y_{j,t}$'s as functions of parameters of the original data x_t. Kuncheva and Faithfull (2014), on the other hand, study (additive) changes in the parameters of y_t directly; for instance, λ_j changing to $\lambda_j + a$ for all j, but without relating this a back to which Σ_1's this change corresponds to.

3 | BIVARIATE RESULTS

This section contains all the bivariate results about sensitivity to changes. The detailed proofs are given in Appendix A.

For changes in the mean in two-dimensional data, Proposition 1 gives the condition for determining which projection is the most sensitive, as well as the results for some special cases.

Proposition 1. *Let $a_{11} = a_{22} = a_{12} = 1$ and $\mu_1, \mu_2 \in \mathbb{R}$ while not both being 0 simultaneously (only the mean changes). $H_2 > H_1$ if and only if $(\mu_1 - \mu_2)^2/(\mu_1 + \mu_2)^2 > (1 - |\rho|)/(1 + |\rho|)$.*

In particular, for all $|\rho| \in (0, 1)$,

1. *$H_2 > H_1$ if one of μ_1 and μ_2 is 0 whereas the other is not (one mean changes).*
2. *$H_2 > H_1$ if $\mu_1 = -\mu_2 = \mu \neq 0$ (equal changes in opposite directions).*
3. *$H_2 < H_1$ if $\mu_1 = \mu_2 = \mu \neq 0$ (equal changes in the same direction).*

When both variances change by the same amount, Proposition 2 tells us that both projections are equally sensitive no matter what the pre-change correlation or size of the change is.

Proposition 2. *Let $\mu_1 = \mu_2 = 0$, $a_{12} = 1$ and $a_{11} = a_{22} = a \neq 1$ (both variances change equally). For any $|\rho| \in (0, 1)$ and $a > 0$, $H_2 = H_1$.*

The picture becomes more complicated when only one variance changes (Proposition 3). If the variance increases, the minor projection is always the most sensitive. On the other hand, if the variance decreases, the principal projection is mostly the most sensitive but not always if the pre-change correlation is high (greater than $\sqrt{3}/2$). In total, this gives a slight edge to the minor projection.

Proposition 3. *Let $\mu_1 = \mu_2 = 0$, $a_{12} = 1$, and either $a_{11} = 1$ and $a_{22} = a \neq 1$, or $a_{11} = a$ and $a_{22} = 1$, where $a > 0$ (one variance changes).*

1. *For any $|\rho| \in (0, 1)$ and $a > 1$ (variance increase), $H_2 > H_1$.*
2. *When $|\rho| \in (0, 1)$ and $a \in (0, 1)$ (variance decrease), $H_2 < H_1$ in most cases. The only exception is if $|\rho| \in (\sqrt{3}/2, 1)$ and $a \in (0, \sqrt{4\rho^2 - 3})$, where $H_2 > H_1$.*

Finally, for a change in correlation, the minor projection is the most sensitive in most cases (Proposition 4). Only if the correlation changes direction and becomes stronger is the principal projection more sensitive.

Proposition 4. *Let $\mu_1 = \mu_2 = 0$, $a_{11} = a_{22} = 1$ and $a_{12} = a \neq 1$ such that (1) holds (the correlation changes). Then $H_2 > H_1$ for any $|\rho| \in (0, 1)$ and $a > -1$.*

4 | EXPLORING HIGHER DIMENSIONS

In the two-dimensional case, we saw that which projection is the most sensitive depends both on the change $(\boldsymbol{\mu_1}, \Sigma_1)$ and on the pre-change correlation matrix Σ_0. For a higher dimension D, solving inequalities like above for all the parameters in $(\Sigma_0, \boldsymbol{\mu_1}, \Sigma_1)$ quickly becomes tedious and uninformative. Therefore, we use simulation to obtain Monte Carlo estimates $E[H_j(\Sigma_0, \boldsymbol{\mu_1}, \Sigma_1)]$ instead, where we vary which parameters that change, the size of the changes, and the sparsity of the change (the number of dimensions that change). Let $\rho_{i,d}$ for $i \neq d$ denote the off-diagonal elements of Σ_0, μ_d be the dth element of $\boldsymbol{\mu_1}$, and σ_d be the dth diagonal element of Σ_1. Then our simulation protocol to get such estimates is as follows:

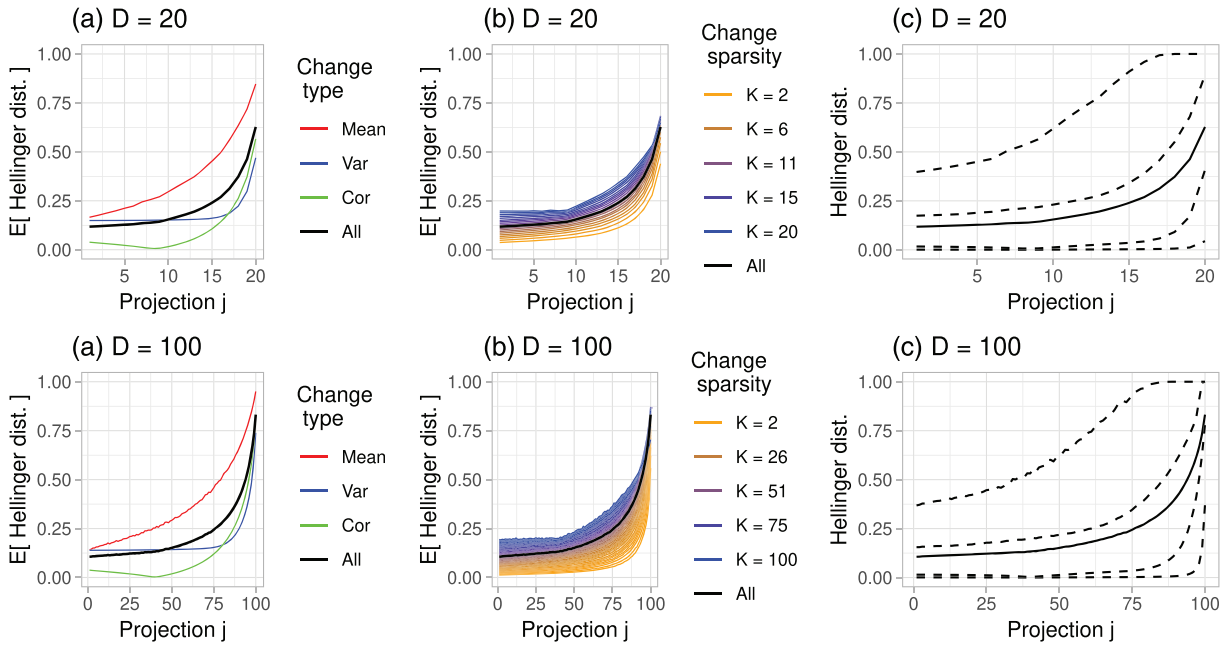

FIGURE 1 A summary of the sensitivity results obtained by the simulation protocol for $D = 20$ for $D = 100$. (a) Monte Carlo estimates of $E[H_j]$ for uniformly drawn changes in the mean, variance, and (decreases in) correlation, as well as uniformly drawn pre-change correlation matrices Σ_0. (b) Same as (a), but now the average sensitivity is conditional on the sparsity of the change, rather than the type of parameter. (c) 0.05, 0.25, 0.75, and 0.95 percentiles (the dashed lines, from bottom to top) of the distribution of H_j, together with $E[H_j]$ (solid line). Note that the percentiles are over Σ_0, $\boldsymbol{\mu_1}$, and Σ_1 simultaneously

1. Draw a correlation matrix Σ_0 uniformly from the space of correlation matrices by the method of Joe (2006) (clusterGeneration::rcorrmatrix in R).
2. Draw a change sparsity $K \sim \text{Unif}\{2, \dots, D\}$.
3. Draw a random subset $\mathcal{D} \subseteq \{1, \dots, D\}$ of size K.
4. Draw an additive change in mean $\mu \sim \text{Unif}(-3, 3)$, and set $\mu_d = \mu$ for $d \in \mathcal{D}$, whereas $\Sigma_1 = \Sigma_0$.
5. Draw a multiplicative change in standard deviation $\sigma \sim \frac{1}{2}\text{Unif}(1/3, 1) + \frac{1}{2}\text{Unif}(1, 3)$ (equal probability of decrease and increase in standard deviation) and set $\sigma_d = \sigma$ for $d \in \mathcal{D}$, keeping the remaining parameters constant.
6. Draw a multiplicative change in correlation $a \sim \text{Unif}(0, 1)$ and change $\rho_{i,d}$ to $a\rho_{i,d}$ for all $i \neq d \in \mathcal{D}$. The other parameters are kept constant.
7. For each of the three change scenarios 4–6, calculate $H_j(\Sigma_0, \boldsymbol{\mu_1}, \Sigma_1)$, $j = 1, \dots, D$.
8. Repeat 2–7 10^3 times.
9. Repeat 1–8 10^3 times.

Averaging the simulated H_js yields estimates of $E[H_j]$, and we can condition on the type of parameter that changes and the change sparsity to see what the sensitivity is expected to be for different classes of changes. (Note that we only consider decreases in correlation. This is to avoid getting too many indefinite Σ_1's. If indefinite Σ_1's still occur, we find the closest positive-definite one by Higham's algorithm (Higham, 2002), implemented in the Matrix::nearPD-function in R.

Figure 1 shows that the trend of the minor components being the most sensitive continues for $D = 20$ and $D = 100$. This holds for changes in the mean, variance, and correlation (a) as well as all the different change sparsities (b). From the quantile plots (c), however, observe that a lot of variation is hidden in these averages, meaning that which projection is the most sensitive will depend on the specific Σ_0 and change $(\boldsymbol{\mu_1}, \Sigma_1)$, as in the bivariate case.

5 | CONCLUDING REMARKS

We have presented bivariate theory demonstrating that the minor projection of PCA-rotated data is usually the most sensitive to changes, especially if the change is sparse. Simulations confirm this to be the case on average for higher dimensions as well, but, in general, the sensitivity strongly varies with the pre-change correlation matrix and the specific change.

In future work, we aim to exploit these insights for creating computationally efficient change detection methods for high-dimensional data. The most promising and surprising part of our results is that even very sparse changes seem to be quite noticeable in the minor projections. This is important for change detection in high-dimensional data because a change rarely affects all dimensions or parameters at once. Most often, only a few parameters among many will change, and therefore, the problem of sparse changes will be the most relevant. One interpretation of

the results presented here is that for detecting sparse changes in the mean vector and/or covariance matrix of a high-dimensional data set or of a sequentially arriving data stream, it is potentially sufficient to search for changes in a few selected minor projections. This might lead to major improvements, not only computationally but also in terms of detection accuracy or speed. Choosing which minor projections to use for a specific change detection problem is the subject of ongoing work.

This work is funded by the Norwegian Research Council centre Big Insight, Project 237718. The author would also like to thank Ingrid Glad for useful input on the presentation of the material.

SUPPORTING INFORMATION AND DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available as part of the supporting information for this online article: **R code** .R file with the code for reproducing (and easily extending) the simulation study and Figure 1.

ORCID


Martin Tveten https://orcid.org/0000-0002-4236-633X

REFERENCES

Camacho, J., Pérez-Villegas, A., García-Teodoro, P., & Maciá-Fernández, G. (2016). PCA-based multivariate statistical network monitoring for anomaly detection. *Computers & Security*, *59*, 118–137. https://doi.org/10.1016/j.cose.2016.02.008

Chan, H. P. (2017). Optimal sequential detection in multi-stream data. *The Annals of Statistics*, *45*(6), 2736–2763. https://doi.org/10.1214/17-AOS1546

Harrou, F., Kadri, F., Chaabane, S., Tahon, C., & Sun, Y. (2015). Improved principal component analysis for anomaly detection: Application to an emergency department. *Computers & Industrial Engineering*, *88*, 63–77. https://doi.org/10.1016/j.cie.2015.06.020

Hawkins, D. M., & Zamba, K. D (2005). Statistical process control for shifts in mean or variance using a changepoint formulation. *Technometrics*, *47*(2), 164–173. https://doi.org/10.1198/004017004000000644

Higham, N. J. (2002). Computing the nearest correlation matrix—A problem from finance. *IMA Journal of Numerical Analysis*, *22*(3), 329–343. https://doi.org/10.1093/imanum/22.3.329

Huang, L., Nguyen, X., Garofalakis, M., Jordan, M. I., Joseph, A., & Taft, N. (2007). In-network PCA and anomaly detection. In Schölkopf, B., Platt, J. C., & Hoffman, T. (Eds.), *Advances in Neural Information Processing Systems 19*. MA, USA: MIT Press, pp. 617–624.

Jackson, J. E., & Morris, R. H. (1957). An application of multivariate quality control to photographic processing. *Journal of the American Statistical Association*, *52*(278), 186–199.

Jackson, J. E., & Mudholkar, G. S. (1979). Control procedures for residuals associated with principal component analysis. *Technometrics*, *21*(3), 341–349. https://doi.org/10.1080/00401706.1979.10489779

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, *97*(10), 2177–2189. https://doi.org/10.1016/j.jmva.2005.05.010

Ketelaere, B. D., Hubert, M., & Schmitt, E. (2015). Overview of PCA-based statistical process-monitoring methods for time-dependent, high-dimensional data. *Journal of Quality Technology*, *47*(4), 318–335. https://doi.org/10.1080/00224065.2015.11918137

Kuncheva, L. I., & Faithfull, W. J. (2014). PCA Feature Extraction for Change Detection in Multidimensional Unlabeled Data. *IEEE transactions on neural networks and learning systems*, *25*(1), 69–80. https://doi.org/10.1109/TNNLS.2013.2248094

Lakhina, A., Crovella, M., & Diot, C. (2004). Diagnosing network-wide traffic anomalies. In *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ACM, New York, USA, pp. 219–230. https://doi.org/10.1145/1015467.1015492

Liu, K., Zhang, R., & Mei, Y. (2017). Scalable SUM-shrinkage schemes for distributed monitoring large-scale data streams. *Statistica Sinica*, *29*, 1–22. https://doi.org/10.5705/ss.202015.0316

Mishin, D., Brantner-Magee, K., Czako, F., & Szalay, A. S. (2014). Real time change point detection by incremental PCA in large scale sensor data. In *2014 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–6. https://doi.org/10.1109/HPEC.2014.7040959

Pimentel, M. A. F., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, *99*, 215–249. https://doi.org/10.1016/j.sigpro.2013.12.026

Rato, T., Reis, M., Schmitt, E., Hubert, M., & De Ketelaere, B. (2016). A systematic comparison of PCA-based statistical process monitoring methods for high-dimensional, time-dependent processes. *AIChE Journal*, *62*(5), 1478–1493. https://doi.org/10.1002/aic.15062

Wang, Y., Mei, Y., & Paynabar, K. (2018). Thresholded multivariate principal component analysis for phase I multichannel profile monitoring. *Technometrics*, *60*(3), 360–372. https://doi.org/10.1080/00401706.2017.1375993

Wang, T., & Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(1), 57–83. https://doi.org/10.1111/rssb.12243

Xie, Y., & Siegmund, D. (2013). Sequential multi-sensor change-point detection. *The Annals of Statistics*, *41*(2), 670–692. https://doi.org/10.1214/13-AOS1094

How to cite this article: Tveten M. Which principal components are most sensitive in the change detection problem?. *Stat*. 2019;8:e252. https://doi.org/10.1002/sta4.252

APPENDIX A: PROOFS

Before turning to the proofs of the propositions in Section 3, the expressions for the pre- and post-change means and variances of each projection are needed. The normalized eigenvectors (principal axes) and corresponding eigenvalues (variance in the data along a given principal axis) of Σ_0 are quickly verified to be

$$\begin{aligned} \lambda_1 &= 1+\rho, \quad \mathbf{v}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \\ \lambda_2 &= 1-\rho, \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \end{pmatrix}. \end{aligned} \tag{A1}$$

Note that which principal axis is the dominant one depends on the sign of ρ. If ρ is positive, $\mathbf{v}_1$ is the dominant one, but $\mathbf{v}_2$ is dominant if ρ is negative.

From the projections in (A1), the parameters of the projections before and after a change can be expressed as functions of the original correlation matrix and multiplicative change factors. For the principal component, the original and changed variances become as follows, respectively:

$$\begin{aligned} o_1^2 &= 1+\rho, \\ c_1^2 &= \frac{1}{2}a_{11}^2 + \frac{1}{2}a_{22}^2 + a_{11}a_{22}a_{12}\rho. \end{aligned} \tag{A2}$$

The expressions for the variances of the minor component are identical up to one switched sign:

$$\begin{aligned} o_2^2 &= 1-\rho, \\ c_2^2 &= \frac{1}{2}a_{11}^2 + \frac{1}{2}a_{22}^2 - a_{11}a_{22}a_{12}\rho. \end{aligned} \tag{A3}$$

Observe that if $\rho < 0$, then o_2 and c_2 would be equal to o_1 and c_1 with positive ρ, and vice versa. Thus, for $\rho \in (-1, 1)$, the general expressions are obtained by replacing ρ with $|\rho|$. Lastly, the changed mean components are given by

$$\begin{aligned} m_1 &= \frac{1}{\sqrt{2}}(\mu_1 + \mu_2), \\ m_2 &= \frac{1}{\sqrt{2}}(\mu_1 - \mu_2). \end{aligned} \tag{A4}$$

We first prove Proposition 1 for changes in the mean.

Proof of Proposition 1. Let $p_1(x) = N(x \mid 0, o_1^2)$, $q_1(x) = N(x \mid m_1, o_1^2)$, $p_2(x) = N(x \mid 0, o_2^2)$, and $q_2(x) = N(x \mid m_2, o_2^2)$, where m_i, o_i are as in (A2), (A3), and (A4), with ρ replaced by $|\rho|$ as noted above. The Hellinger distances between the distributions before and after a change along each principal axis are given by for $j = 1, 2$

$$H_j^2 = H^2(p_j, q_j) = 1 - \exp\left\{-\frac{1}{8o_j^2}m_j^2\right\}.$$

Then some algebra results in the inequality we needed to prove:

$$\begin{aligned} H_2 &> H_1 \\ \frac{1}{8(1-|\rho|)}\frac{(\mu_1-\mu_2)^2}{2} &> \frac{1}{8(1+|\rho|)}\frac{(\mu_1+\mu_2)^2}{2} \\ \frac{(\mu_1-\mu_2)^2}{(\mu_1+\mu_2)^2} &> \frac{1-|\rho|}{1+|\rho|} \end{aligned}$$

From this inequality, the three special cases (i), (ii), and (iii) are immediately given. □

In the proofs concerning changes in the covariance matrix, we will make use of the following lemma. It reduces the inequality of Hellinger distances to a simpler inequality of ratios of variances.

Lemma 1. *Let p_1, q_1, p_2, q_2 be 0-mean normal distribution functions with variances $\sigma_{p_1}^2, \sigma_{q_1}^2, \sigma_{p_2}^2$, and $\sigma_{q_2}^2$, respectively. Furthermore, let*

$$\log r_j = \left| \log \frac{\sigma_{q_j}^2}{\sigma_{p_j}^2} \right|, \quad j = 1, 2.$$

Then $H(p_2, q_2) > H(p_1, q_1)$ if and only if $\log r_2 > \log r_1$.

Proof. First observe that when the means are 0, then we can write the Hellinger distance between two normal distributions as the following.

$$\begin{aligned} H^2(p, q) &= 1 - \left(\frac{2\sigma_p \sigma_q}{\sigma_p^2 + \sigma_q^2} \right)^{1/2} \\ &= 1 - \sqrt{2} \left(\frac{\sigma_p}{\sigma_q} + \frac{\sigma_q}{\sigma_p} \right)^{-1/2} \\ &= 1 - \sqrt{2} \left(\frac{\sigma_p^2}{\sigma_q^2} + \frac{\sigma_q^2}{\sigma_p^2} + 2 \right)^{-1/4}. \end{aligned}$$

This gives us the inequality

$$\begin{aligned} H(p_2, q_2) &> H(p_1, q_1) \\ \frac{\sigma_{p_2}^2}{\sigma_{q_2}^2} + \frac{\sigma_{q_2}^2}{\sigma_{p_2}^2} &> \frac{\sigma_{p_1}^2}{\sigma_{q_1}^2} + \frac{\sigma_{q_1}^2}{\sigma_{p_1}^2}. \end{aligned}$$

By setting $r_2 = \sigma_{p_2}^2 / \sigma_{q_2}^2$ and $r_1 = \sigma_{p_1}^2 / \sigma_{q_1}^2$, the inequality can be written as

$$r_2 + r_2^{-1} > r_1 + r_1^{-1}.$$

Now assume first that $r_1, r_2 > 1$, that is, $\sigma_{p_j}^2 > \sigma_{q_j}^2$. Then we see that

$$\begin{aligned} r_2 + r_2^{-1} &> r_1 + r_1^{-1} \\ r_2 - r_1 + \frac{r_1 - r_2}{r_1 r_2} &> 0 \\ (r_2 - r_1) \left(1 - \frac{1}{r_1 r_2} \right) &> 0. \end{aligned}$$

By the assumption that $r_1, r_2 > 1$, this inequality holds if and only if $r_2 > r_1$.

Finally, note that by interchanging $\sigma_{p_j}^2$ and $\sigma_{q_j}^2$, the same result is obtained when $\sigma_{q_j}^2 \geq \sigma_{p_j}^2$. Thus, to make the result hold in general, we can set

$$r_j = \exp \left\{ \left| \log \frac{\sigma_{q_j}^2}{\sigma_{p_j}^2} \right| \right\}, \quad j = 1, 2,$$

which is an expression for the ratio between variances where the largest of the variances is always in the numerator. Therefore, we get that $\log r_2 > \log r_1$ is equivalent to $H_2 > H_1$. □

The rest of this article contains the individual proofs of the remaining propositions in the main body of the text.

Proof of Proposition 2. Let $\log r_j$ for $j = 1, 2$ be defined as in Lemma 1. When assuming that $a_{12} = 1$ and $a_{11} = a_{22} = a \neq 1$, we get that

$$\log r_2 = \left| \log \frac{a^2/2 + a^2/2 - |\rho| a^2}{1 - |\rho|} \right| = |\log a^2|,$$

and

$$\log r_1 = \left| \log \frac{a^2/2 + a^2/2 + |\rho| a^2}{1 + |\rho|} \right| = |\log a^2|.$$

Hence, by arguments along the lines of the proof of Lemma 1, we see that $H_2 = H_1$ no matter what $|\rho|$ or a is. □

Proof of Proposition 3. Using the formulas for the variances of the projections (A2) and (A3), the inequality we have to study according to Lemma 1 becomes the following:

$$\begin{aligned}\left|\log\frac{a^2-2a|\rho|+1}{2(1-|\rho|)}\right| &> \left|\log\frac{a^2+2a|\rho|+1}{2(1+|\rho|)}\right| \\ \left|\log\left[\frac{(1-a)^2}{2(1-|\rho|)}+a\right]\right| &> \left|\log\left[\frac{(1-a)^2}{2(1+|\rho|)}+a\right]\right|.\end{aligned} \tag{A5}$$

First, we have to find the sign of the expressions inside the absolute values for each a and $|\rho|$. For the left-hand side, we get

$$\begin{aligned}\frac{(1-a)^2}{2(1-|\rho|)}+a &= 1 \\ a = 1 \text{ and } a &= 2|\rho|-1.\end{aligned}$$

Thus, for $a > 1$ and $a < 2|\rho| - 1$, the left-hand side is positive, whereas negative in between. For the right-hand side, the expression inside the absolute value signs are positive for $a > 1$ and $a < -(1 + 2|\rho|)$. Because $a > 0$, however, the relevant root for the right-hand side is only $a = 1$. In total, this gives us three regions of $(a, |\rho|)$-values to check inequality (A5): $a > 1$ and $|\rho| \in (0, 1)$, $a \in (2|\rho| - 1, 1)$ and $|\rho| \in (0, 1)$, and $a \in (0, 2|\rho| - 1)$ and $|\rho| \in (1/2, 1)$.

$a > 1$ and $|\rho| \in (0, 1)$:

The absolute value signs can now be dissolved, so that inequality (A5) becomes

$$\frac{(1-a)^2}{(1-|\rho|)} > \frac{(1-a)^2}{(1+|\rho|)}.$$

Because $|\rho| \in (0, 1)$, we see that the inequality holds for any $a > 1$. Hence, $H_2 > H_1$ in this scenario, when the variance increases.

$a \in (2|\rho| - 1, 1)$ and $|\rho| \in (0, 1)$:

In this case, inequality (A5) becomes

$$\frac{(1-a)^2}{(1-|\rho|)} < \frac{(1-a)^2}{(1+|\rho|)}.$$

That is, it does not hold for any of the a's or $|\rho|$'s within the relevant region. Note that when $|\rho| < 1/2$, a is kept between $(0, 1)$.

$a \in (0, 2|\rho| - 1)$ and $|\rho| \in (1/2, 1)$:

Now we get the inequality

$$\frac{(1-a)^2}{2(1-|\rho|)}+a > \left(\frac{(1-a)^2}{2(1+|\rho|)}+a\right)^{-1},$$

which is equivalent to

$$a^4 - a^2(4\rho^2 - 2) + 4\rho^2 - 3 > 0. \tag{A6}$$

The roots of the function on the left-hand side are $a = \pm 1$ and $a = \pm\sqrt{4\rho^2 - 3}$, but the only relevant root for $a \in (0, 2|\rho| - 1)$ and $|\rho| \in (1/2, 1)$ is $a_0 := \sqrt{4\rho^2 - 3}$.

Next, for $|\rho| < \sqrt{3}/2$, the root a_0 moves into the complex plane, and the function on the left-hand side of (A6) is always less than 0 for the relevant a's. That is, $H_2 < H_1$ in this case. If $|\rho| > \sqrt{3}/2$, on the other hand, then (A6) holds for $a \in (0, a_0)$, but not for $a \in (a_0, 2|\rho| - 1)$.□

Proof of Proposition 4. In this scenario, the inequality to check due to Lemma 1 and expressions (A2) and (A3) is

$$\left|\log\frac{1-a|\rho|}{1-|\rho|}\right| > \left|\log\frac{1+a|\rho|}{1+|\rho|}\right|. \tag{A7}$$

To dissolve the absolute value signs, we first have to see for which values of a and $|\rho|$ the expressions inside are positive or negative. It is easily verified that the expression inside the left-hand side absolute value is positive for $a < 1$, whereas the right-hand side is positive if $a > 1$, both being negative otherwise.

First assume that $a < 1$. Then inequality (A7) becomes

$$\begin{aligned}\frac{1-a|\rho|}{1-|\rho|} &> \frac{1+|\rho|}{1+a|\rho|} \\ 1-(a\rho)^2 &> 1-\rho^2 \\ a^2 &< 1.\end{aligned}$$

Hence, $a \in (-1, 1)$ yields $H_2 > H_1$. On the other hand, if $a > 1$, we obtain

$$\frac{1 - |\rho|}{1 - a|\rho|} > \frac{1 + a|\rho|}{1 + |\rho|}$$
$$a^2 > 1,$$

which is always true. Thus, in total, $H_2 < H_1$ only if $a < -1$. □