Least-squares approximation of a space distribution for a given covariance and latent sub-space

José Camacho a,⁎, Pablo Padilla a, Jesús Díaz-Verdejo a, Keith Smith b, David Lovett b

a Departamento de Teoría de la Señal, Telemática y Comunicaciones, Universidad de Granada, 18071, Granada, Spain
b Perceptive Engineering Ltd., United Kingdom

A R T I C L E   I N F O

Article history:
Received 22 July 2010
Received in revised form 21 December 2010
Accepted 22 December 2010
Available online 29 December 2010

Keywords:
Covariance matrices
Constrained least squares
Principal component analysis
Partial least squares

A B S T R A C T

In this paper, a new method to approximate a data set by another data set with constrained covariance matrix is proposed. The method is termed Approximation of a Distribution for a given COVariance (ADICOV). The approximation is solved in any projection subspace, including that of Principal Component Analysis (PCA) and Partial Least Squares (PLS). Given the direct relationship between covariance matrices and projection models, ADICOV is useful to test whether a data set satisfies the covariance structure in a projection model. This idea is broadly applicable in chemometrics. Also, ADICOV can be used to simulate data with a specific covariance structure and data distribution. Some applications are illustrated in an industrial case of study.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Projection methods are intimately related to covariance matrices. For instance, the loading vectors of Principal Component Analysis (PCA) can be extracted either from a mean centered1 data set X or from the corresponding covariance matrix Cx = (X′X)/(Nₓ − 1), with Nₓ being the number of observations in X. To identify the loadings from X, the NIPALS algorithm [1] or the Singular Value Decomposition (SVD) [2], among others, can be employed. At the same time, the loading vectors are the eigenvectors of Cx and thus can be computed from the eigendecomposition (ED) of this matrix. In addition, the eigenvectors remain the same when multiplying Cx by any non-zero scalar k. Therefore, the loading vectors in PCA can be computed directly from X′X. Similarly, the loadings and weights in Partial Least Squares (PLS) regression can be identified either from the original data sets X and Y [1,3], or from cross-product matrices X′X and X′Y using the kernel algorithm [4–6].

The identification of projection models from covariance matrices presents one main advantage and one main drawback. The advantage is that when the number of observations in the data matrices is very large, it is more efficient to fit a model from the covariance matrices [4]. This is because the size of the covariance matrix depends only on the number of variables of the data set, and not on the number of observations. Moreover, the computation of the covariance can be performed in an iterative manner which is appropriate for very large data sets, since the complete set of observations does not need to be considered at once. The limitation is that covariance matrices do not give any information about the distribution of the data in the model subspace: the scores. Notice that two different data sets with a completely different distribution of the scores and even with different numbers of observations may yield the same covariance matrix and so the same projection model. The visualization of the distribution of the scores is needed in some applications. For exploratory data analysis (EDA), it is highly important to investigate the distribution of the data together with the model structure [7,8]. Also, in multivariate statistical process monitoring (MSPM), the score distribution helps us determine if the process is under statistical control. Anyway, although scores are not obtained when projection models are calibrated from cross-product matrices, they can be computed subsequently using the original data.

In this paper, the parallelism between original data and covariance in the calibration of projection models is exploited to propose a method to approximate a data set by another data set with a given covariance structure. The method is termed Approximation of a Distribution for a given COVariance (ADICOV). The approximation is solved in any projection subspace. Given the direct relationship between covariance matrices and projection models, ADICOV may be useful in two types of applications:

• To yield data with a specific covariance structure which approximates a given space distribution.
• To test whether a data set satisfies the covariance structure in a model. This idea can be valuable in a high number of applications,
including process monitoring, the choice of the method for missing value estimation and observation-wise compression.

The paper is organized as follows. Section 2 introduces ADICOV. Section 3 performs some experiments with random data. From the results of these experiments, Section 4 discusses some applications of ADICOV. Section 5 presents an industrial case of study. In Section 6, the concluding remarks are drawn.

2. Approximation of a distribution for a given covariance (ADICOV)

Let us define \( \mathbf{X} \) as the original or calibration data matrix and \( \mathbf{L} \) as the processed or test data matrix. ADICOV computes the approximation matrix \( \mathbf{A} \) as the least squares approximation of \( \mathbf{L} \) constrained to the covariance of \( \mathbf{X} \), represented by \( \mathbf{C}_x \). Thus, \( \mathbf{A} \) has the same covariance matrix as \( \mathbf{X} \), and approximates the space distribution of \( \mathbf{L} \). The least squares approximation can be solved in the original space of the variables or in a subspace of interest. For instance, if ADICOV is used to test whether a data set is coherent with a PCA model, the least squares approximation should be performed on the PCA subspace.

Depending on the application, matrices \( \mathbf{X} \) and \( \mathbf{L} \) may or may not be independently generated. For instance, in process monitoring, \( \mathbf{X} \) contains the calibration data used for model fitting and \( \mathbf{L} \) contains posterior observations of the process, collected during actual monitoring. In that case, \( \mathbf{X} \) and \( \mathbf{L} \) are independently generated. On the other hand, the difference between \( \mathbf{X} \) and \( \mathbf{L} \) may be a data set with missing values and \( \mathbf{L} \) the same matrix where the missing elements have been estimated with a given method. ADICOV in this context is useful to determine whether the estimated values introduce a change in the covariance structure of the original data.

Once the approximation by ADICOV is performed, the interest may be specifically on the approximation matrix \( \mathbf{A} \) or on the difference found between \( \mathbf{L} \) and its approximation. The interest is in \( \mathbf{A} \) when the aim is data simulation. The interest is in the difference between \( \mathbf{L} \) and \( \mathbf{A} \) when the aim is to measure to what extent \( \mathbf{L} \) satisfies the covariance of \( \mathbf{X} \). This difference should be computed on the (sub)space of interest. Thus, if the original space of the variables is considered in ADICOV, the difference is computed between \( \mathbf{A} \) and \( \mathbf{L} \). Otherwise, the difference is computed between the projections of \( \mathbf{A} \) and \( \mathbf{L} \) on the (sub)space of interest, \( \mathbf{T}_a \) and \( \mathbf{T}_l \), respectively. Mathematically, it can always be stated that the difference is computed between \( \mathbf{T}_a \) and \( \mathbf{T}_l \) since these matrices are, in fact, \( \mathbf{A} \) and \( \mathbf{L} \) when no subspace is considered.

The difference between \( \mathbf{T}_a \) and \( \mathbf{T}_l \) may be assessed for different purposes. When \( \mathbf{L} \) is obtained after some processing of \( \mathbf{X} \), the difference between \( \mathbf{T}_a \) and \( \mathbf{T}_l \) can be computed to evaluate to what extent the processing performed involved the introduction of artifacts in the data. This is only sensible when the processing is not supposed to change the covariance structure. This is the case, for instance, of the estimation of missing values commented before. On the other hand, when \( \mathbf{X} \) and \( \mathbf{L} \) are independently generated, the difference between \( \mathbf{T}_a \) and \( \mathbf{T}_l \) gives an idea of the degree of stability of the stochastic process under analysis. This is useful in process monitoring and to assess the robustness of a calibration model. A similar philosophy can be applicable in model selection, model validation and classification.

2.1. Projection onto the subspace of interest

ADICOV constrains the approximation matrix \( \mathbf{A} \) to satisfy the covariance specified in \( \mathbf{C}_x \) in a certain subspace. Thus, the equivalence in covariance between \( \mathbf{X} \) and \( \mathbf{A} \) can be imposed in, say, the subspace corresponding to the first pair of Principal Components (PCs) or to the last three Latent Variables (LV) in PLS. The projection of covariance matrix \( \mathbf{C}_x \) on a specific subspace, referred as \( \mathbf{G}_x \), follows:

\[
\mathbf{G}_x = \mathbf{R}^\top \mathbf{C}_x \mathbf{R}.
\]  

with \( \mathbf{R} \) as the corresponding projection matrix. Also, data matrix \( \mathbf{L} \) is projected on the subspace yielding \( \mathbf{T}_l \):

\[
\mathbf{T}_l = \mathbf{L} \mathbf{R}.
\]

Let us define \( \mathbf{T}_a \) as the least squares approximation of \( \mathbf{T}_l \) with \( \mathbf{G}_x \) as the covariance matrix. Then, \( \mathbf{A} \) is obtained from the following operation:

\[
\mathbf{A} = \mathbf{T}_a \mathbf{Q}^{-\top}.
\]

where \( \mathbf{Q} \) performs the linear transformation from the subspace to the original space of the variables.

If the subspace of interest is the PCA subspace being \( \mathbf{P} \) as the loading matrix [9], then \( \mathbf{R} \) in Eqs. (1) and (2) and \( \mathbf{Q} \) in Eq. (3) are set to \( \mathbf{P} \). If the subspace of interest is the PLS subspace being \( \mathbf{W} \) as the loading matrix of the x-block and \( \mathbf{W} \) as the weight matrix [10], then \( \mathbf{R} = \mathbf{W}^{-\top} (\mathbf{P}^\top \mathbf{W})^{-1} \) in Eqs. (1) and (2) and \( \mathbf{Q} = \mathbf{P} \) in Eq. (3).

It should be noted that the applications presented later on in this paper are illustrated with PCA, so that both \( \mathbf{C}_x \) and \( \mathbf{R} = \mathbf{Q} = \mathbf{P} \) are computed from \( \mathbf{X} \). This implies that \( \mathbf{G}_x \) is a diagonal matrix with the eigenvalues of \( \mathbf{X} \). Nonetheless, from a general point of view \( \mathbf{G}_x \) does not need to be diagonal.

2.2. Imposing a covariance matrix

\( \mathbf{T}_a \) should be defined to have a covariance matrix equal to \( \mathbf{G}_x \). This can be performed as follows. The SVD of \( \mathbf{T}_a \) is:

\[
\mathbf{T}_a = \mathbf{U}_a \mathbf{S}_a \mathbf{V}_a^\top.
\]

and its covariance matrix:

\[
\mathbf{G}_x = \frac{1}{N_a - 1} \left( \mathbf{V}_a \mathbf{S}_a^2 \mathbf{S}_a \mathbf{V}_a^\top \right).
\]

with \( N_a = N_l \), being \( N_{l} \) and \( N_{a} \) as the number of observations in \( \mathbf{T}_a \) and \( \mathbf{T}_l \), respectively.

On the other hand, the ED of \( \mathbf{G}_x \) follows:

\[
\mathbf{V}_a = \mathbf{D}_a \mathbf{S}_a \mathbf{V}_a.
\]

where \( \mathbf{V}_a \) contains the eigenvectors of \( \mathbf{G}_x \) and \( \mathbf{D}_a \) denotes the corresponding eigenvalues in its diagonal. Let us define matrix \( \mathbf{S}_a \) as the diagonal matrix where each of the elements equals the square root of the corresponding element in \( \mathbf{D}_a \). \( \mathbf{S}_a \) has only real values since the elements in \( \mathbf{D}_a \) cannot be negative.

To satisfy \( \mathbf{G}_x = \mathbf{G}_x \), the following equalities need to hold:

\[
\mathbf{S}_a = \sqrt{\frac{N_a}{N_a - 1}} \mathbf{S}_a.
\]

Using the derivation above, the procedure to obtain the approximation matrix \( \mathbf{A} \) is as follows: matrices \( \mathbf{S}_a \) and \( \mathbf{V}_a \) are obtained from \( \mathbf{G}_x \) in order to compute \( \mathbf{T}_a \) from Eqs. (7), (8) and then (4), and in turn \( \mathbf{A} \) from Eq. (3). Still, matrix \( \mathbf{U}_a \) needs to be set in Eq. (4). This is explained in the next section. Provided that matrix \( \mathbf{U}_a \) in Eq. (4) is orthonormal, \( \mathbf{A} \) will have the same covariance as \( \mathbf{X} \) in the subspace of interest.

2.3. Least squares approximation

\( \mathbf{U}_a \) can be selected so that \( \mathbf{T}_a \) approaches a specific spatial distribution of the observations. From Eq. (3), this is the same as stating that \( \mathbf{A} \) approaches a specific spatial distribution of the
observations in the projected subspace. In particular, this distribution should be as close as possible to that of \( T_L \). This makes it possible to evaluate to what extent the transition from \( X \) to \( L \) involved the introduction of artifacts that affect the subspace of interest.

The problem of finding \( U_a \), so that \( T_L \) approximates \( T_L \) in the least squares sense is equivalent to finding the least squares solution for:

\[
T_L = U_L \cdot S_L \cdot V_L^T.
\]

where \( T_L \), \( S_L \) and \( V_L \) are given as explained in the previous sections. This, in turn, is an optimization problem where the solution, \( U_L \) is constrained to be orthogonal. In particular, this optimization problem is known as the Orthogonal Procrustes Problem [11]. This problem has an analytic solution, therefore avoiding the computational burden of optimization algorithms like [12]. The solution is the polar decomposition [13] of matrix \( M \), where:

\[
M = T_L \cdot (S_L \cdot V_L)^T.
\]

The polar decomposition of \( M \) can be computed from the SVD as follows:

\[
M = U_m \cdot S_m \cdot V_m^T.
\]

then the solution holds:

\[
U_a = U_m \cdot V_m.
\]

2.4. The ADICOV algorithm

According to the previous discussion, the ADICOV algorithm is computed as follows:

Inputs:
- \( R, Q \): subspace of interest.
- \( X \): original data set.
- \( L \): current data set.

Algorithm:
- \( N_r \) ← countObservations(\( X \))
- \( C_r = \frac{1}{N_r-1}X^T \cdot X \)
- \( N_r \) ← countObservations(\( L \))
- \( C_r = R^T \cdot C_r \cdot R \)
- \( T_L = L \cdot R \)
- \([V_x, D_x] \) ← ED(\( G_x \))
- \( S_x = \sqrt{\frac{N_r-1}{N_r}} \cdot S_x \)
- \( M = T_L \cdot (S_x \cdot V_x)^T \)
- \([U_m, S_m, V_m] \) ← SVD(\( M \))
- \( U_a = U_m \cdot V_m \)
- \( T_L = U_a \cdot S_a \cdot V_a^T \)
- \( A = T_L \cdot Q^T \)

Two comments to this algorithm are in due. First, if \( X \) is the input of the algorithm, the computation of \( C_r \) is not strictly necessary, since \( S_x \) and \( V_x \) can be directly obtained from the SVD of \( X \). Second, there are situations in which \( X \) cannot be used as the input to the algorithm, for instance because the number of observations is too large. Under these circumstances, \( C_r \) and \( N_r \) have to be used as alternative inputs to \( X \) and the first two steps of the algorithm are not performed. As discussed in the Introduction, \( C_r \) can be computed in an iterative manner which is suitable for very large data sets:

\[
C_r(t) = C_r(t-1) + x(t)^T \cdot x(t).
\]

where \( C_r(t) \) is the covariance matrix after the \( t \)-th observation \( x(t) \) has been considered.

2.5. ADICOV similarity index

ADICOV is useful to simulate data with a given covariance and space distribution and to check whether \( L \) satisfies the covariance of \( X \) in a given subspace. For the latter purpose, a similarity index between \( L \) and \( A \) is defined. This index is computed as the normalized form of the square of the Frobenius norm [14] of the difference between both matrices. Again, this should be performed on the sub-space of interest:

\[
E = (L - A) \cdot R.
\]

(14)

\[
f^A = \frac{1}{N_r \cdot D_r} \cdot |E|^2 = \frac{1}{N_r \cdot D_r} \cdot Tr(F^2 \cdot E).
\]

(15)

where \( f^A \) is the ADICOV similarity index, \( Tr() \) stands for the trace of a matrix, \( N_r \) is the number of observations in \( L \) and \( A \) and \( D_r \) is the dimension of the projection subspace. \( f^A \) is the normalized element-wise Euclidean distance between \( L \) and \( A \) in the projection subspace. In some contexts, the Mahalanobis distance may be more appropriate than the Frobenius distance. For this, \( R \) is substituted in Eq. (14) by the corresponding matrix. For instance, to compute the ADICOV index using the Mahalanobis distance in the PCA subspace with eigenvectors in \( P \) and singular values in the diagonal matrix \( S \), Eq. (14) is substituted by:

\[
E = (L - A) \cdot P \cdot S^{-1}. \]

(16)

3. Some experiments with random data

To understand the behavior of ADICOV in detail, its performance for different features of the input matrices should be studied. An experimental design with random data is performed for this purpose. Recall that \( X \) is the original data matrix with covariance matrix \( C_r \). \( L \) is the matrix to be approximated with ADICOV, and \( R \) and \( Q \) are the corresponding transformation matrices of the subspace of interest. The factors considered in the experiment are the following:

- The number of observations (\( N_r \)) in \( L \)
- The number of variables (\( M \)) for each observation.
- The dimension of the projection subspace (\( D_r \)).
- The similarity between \( X \) and \( L \).

Taken these previous starting considerations, the following experiment with random data matrices is processed. Firstly, \( X \) is randomly generated for a given dimension specified by \( N_r \) and \( M \), as follows:

\[
X = U_x \cdot S_x \cdot V_x^T.
\]

(17)

where \( U_x \) and \( V_x \) are obtained from the SVD of random data generated according to a multinormal distribution with 0 mean and unit standard deviation, and \( S_x \) contains a fixed number \( S \) of non-zero singular values. These eigenvalues are generated according to the square of a normal distribution with 0 mean and unit standard deviation. In previous investigations, it was found that the number and dispersion of the non-zero eigenvalues had an effect on the approximation by ADICOV. In turn, these two features indirectly depend on the dimension of \( X \). For instance, \( X \) matrices with a larger number of observations are prone to have more non-zero eigenvalues.
Following the simulation approach in Eq. (17), the number of non-zero singular values does not depend on \( N_l \) and \( M \). Furthermore, data sets with strong structural relationships among variables can be simulated.

From matrix \( X \), matrix \( L \) is simulated as follows:

\[
L = (1-k)X + kN,
\]

where \( N \) is multinormal data with 0 mean and unit standard deviation which represents white noise and \( k \) is a constant between 0 and 1. Finally, \( L \) is approximated with ADICOV in the subspace corresponding to the first \( D_r \) PCs of a PCA model computed from \( X \), yielding matrix \( A \), and the similarity index \( I^A \) is computed. This experiment is repeated for the following values:

- \( N_l = \{10, 100, 1000\} \).
- \( M = \{10, 100, 1000\} \).
- \( D_r = \{2, 5, 10\} \).
- \( k = \{0.02, 0.05, 0.10\} \) in Eq. (18).

In all the cases, the number of non-zero singular values in \( X \) is \( S = 20 \). Ten replicates are simulated for each combination of the different factors.

The determination of the contribution of each factor under consideration to the ADICOV index \( I^A \) and possible interactions between pairs of factors, is assessed by means of Analysis of Variance (ANOVA). Factors \( N_l, D_r \) and \( k \) are found to be statistically significant (p-value \( \leq 0.01 \)). Fig. 1 shows the Least Significant Difference (LSD) plots for those factors. The results let us propose some considerations for the proper usage of ADICOV:

- The ADICOV index \( I^A \) decreases with the number of observations, that is to say, \( L \) can be approximated for a given covariance with smaller variations as \( N_l \) increases.
- The number of variables has not a statistically significant effect on the index \( I^A \), since both ADICOV and the index are computed in the subspace of interest, instead of the original space. On the contrary, the number of constraints on the least squares approximation problem does depend on the number of PCs, that is the dimension of the subspace. The ADICOV index \( I^A \) increases with the dimension of the subspace of interest.
- Discrepancies between \( X \) and \( L \) yield a scenario with a higher index \( I^A \).

Two interactions between factors are also found statistically significant (p-value \( \leq 0.01 \)). The interaction plots are shown in Fig. 2. These interactions are the ones between the number of observations \( N_l \) with the number of PCs \( D_r \) and the number of observations \( N_l \) with the noise factor \( k \). The first interaction reveals that a low number of observations is more limiting than a high number of PCs. The second interaction shows that the dissimilarity in covariance between \( X \) and \( L \) is the most limiting factor. For high enough values of \( k \), a higher number of observations \( N_l \) does not lead to a reduction of \( I^A \). This conclusion can be also illustrated with the example shown in Fig. 3: consider two simple matrices \( X \) and \( L \) (\( N_l \) observations and 2 variables) with completely different covariance matrices. No matter the number of observations generated according to the two different covariance structures, the approximation of \( L \) constrained to the covariance of \( X \) using ADICOV yields a very different distribution of the points. Thus, the distribution in \( A \) (Fig. 3(c)) is very different to that of \( L \) (Fig. 3(b)) and so a very high \( I^A \) index is obtained.

4. Applications of ADICOV

As mentioned in the Introduction, ADICOV can be applied to generate data sets with a specific covariance structure or to test whether a data set meets a specific covariance structure. For the first objective, the interest is in the output matrix \( A \). For the second objective, the interest is in the ADICOV index \( I^A \) in Eq. (15).

The applications for which \( I^A \) is useful follow a similar pattern. There is a set of matrices \( \{L_1,...,L_r\} \) which are to be tested whether they approximate the covariance matrix of \( X \) in a certain subspace. ADICOV is applied on each of the matrices and the indices \( I^A(X,L_i) \) are computed. In this section, several applications of ADICOV following this pattern are introduced.

4.1. Simulating data

ADICOV can be applied to simulate data with a specific covariance matrix or projection model, and specific data distribution. To illustrate this application, in Fig. 4, three completely different data distributions which provide the same PCA model are shown: a multinormal distribution, a distribution with one outlier and a distribution with two clusters. The ADICOV algorithm was employed to yield these data sets using the same covariance matrix \( C_\alpha \) and different \( L \) matrices: a multinormal one, a multinormal one with an outlier, and a
distribution with two multinormal distributions with an offset between them. This example clearly illustrates that the same projection model may be the result of very different data distributions.

4.2. Avoiding the introduction of artifacts in a data set

In data analysis applications, all the steps performed on the data from collection to the actual analysis should be considered for a proper interpretation of the results [16]. In many cases, these steps are not performed in the same place and/or by the same person. Thus, there is a potential risk of misinterpretation of the analysis results, giving relevance to artifacts in the data which, far from being the trace of the phenomenon of interest, are determined by the data acquisition mechanism or the data transcription procedure. There may be many examples of this problem. Those considered here are data compression and missing data estimation.

4.2.1. Interval-wise compression

In the industrial environment, it is customary to collect a high number of variables at fast rates to perform control and monitoring tasks. This collection activity produces tons of data, with tens to thousands of variables and tens to millions of observations. Typically, these data are compressed to reduce the volume of data, for instance using interval-wise averages. Thus, instead of the actual reading of the sensors, an average per time interval is considered in the posterior analysis. This operation is similar to a low-pass filtering, and an expected side effect of this compression is a reduction of measurement noise.

Clearly, the compression operation may distort the covariance structure in the data if it is not properly performed. As a consequence, the PCA or the PLS model for applications such as data investigation, monitoring and control is actually affected by the compression. To avoid this, the parameters of the compression, for instance the interval length, should be properly defined. This can be carried out by choosing a number of possible intervals and computing the compression from the original matrix of data $X$ according to these intervals, yielding matrices $\{L_1, \ldots, L_f\}$. Then, ADICOV and $F^A$ can be used.
to compare the resulting compressed data sets in order to choose an interval length which does not change the covariance to a large extent.

4.2.2. Missing data estimation

Missing data methods based on projection models have been deeply studied in the literature. Most contributions focus on the estimation of missing data once the model is built [17–21]. Recently, the problem of model building with missing elements has also been considered [22]. In both cases, the estimates of missing elements should be distinguished from actual data for posterior analysis.

Unfortunately, in most situations missing data are estimated and then, for simplicity, treated as actual data with the consequent risk of the introduction of artifacts. This problem has been treated by modelling the uncertainty generated in the estimations of missing elements after model building [23].

In a practical situation, the choice of a missing value method is an unsupervised problem in the sense that the actual missing elements are not available. Therefore, the analyst cannot rely on prediction error but on different selection criteria. An alternative criterion is to avoid the introduction of artifacts in the data due to estimation. ADICOV and $P^L$ can be used to select the missing data approach which, among a number of considered methods, introduces the lowest distortion in the covariance of the data.

4.3. Process monitoring

The performance of the monitoring system in an industrial process is very important from an economical point of view. An accurate monitoring system saves time in the detection of production problems [24] and so, it saves money. Projection models are nicely combined with multivariate statistical process monitoring (MSPM) techniques. Commonly, two complementary charts are used for monitoring: the Q-statistic, which compresses the residuals; and the D-statistic or Hotelling’s $T^2$ statistic [25], computed from the scores. There are two phases in the construction of the monitoring charts in MSPM [26,27]. The first phase is used to check whether the process is in control and to identify data useful for the calibration of the monitoring system and the second phase in which actual monitoring is performed. During the first phase, the whole data set is used to fit the projection model and build the monitoring charts. Abnormalities are detected and potent causes are identified. This may lead to a reduction in the variability of the process prior to the design of the actual monitoring system. The same monitoring statistics are used in both phases, although control limits are computed with slightly different equations [26,28–30]. Also, making the most of the nature of projection models, the contribution of the variables to an abnormality can be investigated using contribution plots [31,32].

The traditional statistics are based on observation-wise distances to the mean, both in the model and residual subspaces. The D-statistic can be identified as a Mahalanobis distance in the model sub-space and the Q-statistic as a Euclidean distance in the residual sub-space. In most situations, deviations in the Q-statistic indirectly point out to a change of covariance. Nonetheless, it should be noticed that this is an indirect measure, and deviations are not always caused by a covariance breakage. For instance, a data set which satisfies the relationship structure among variables in the model may present an abnormal separation from the mean. This would cause high values in both the D-statistic and the Q-statistic without a real breakage of structure. Typically, a high Q-statistic value may lead to the conclusion that the model is not adequately representing the new data, while in this case this conclusion would be wrong. Nevertheless, the most important drawback of the indirect association distance-fault is when a breakage of covariance structure shows a low deviation from the mean, and so goes unnoticed in traditional charts. For instance, in [33] it was shown that when the number of variables is high, changes in the covariance with small deviations may be difficult to detect.

In this context, the use of ADICOV is different than in the previous cases. In the first phase, $X$ is the matrix of calibration data, used for model fitting and, optionally, for the control limits design. Then, $X$ is divided in a number of sub-intervals of observations, yielding matrices $\{L_{1},...,L_{f}\}$. In the second phase, $X$ is the calibration data – only data in statistical control – and intervals of incoming observations are collected in matrices $\{L_{1},...,L_{f}\}$. Two types of indices are generated for monitoring. First, ADICOV is performed on the $L$ matrices in the subspace corresponding to the PCA (or PLS) model, yielding approximation matrices $\{A_{1}^{f},...,A_{L}^{f}\}$. These matrices are
compared to the original L matrices using the Mahalanobis distance, in turn yielding indices \( I_{1D}, \ldots, I_{fD} \). These indices are similar in nature to the D-statistic. Nonetheless, unlike the D-statistic, the ADICOV indices do not detect an abnormal deviation of the scores in the model subspace, but a breakage of the covariance structure. A second set of ADICOV matrices is generated for the residual subspace: \( \{A_{1}, \ldots, A_{k}\} \). These matrices are compared to the original L matrices using the Euclidean distance, yielding indices \( I_{1}, \ldots, I_{f} \). Again, these indices are similar to the Q-statistic, but are focused on detecting a breakage in the covariance rather than an abnormal deviation.

5. Case study: industrial data from continuous process

The data set under study was collected during a period of more than 4 days of continuous operation of a fluidized bed reactor fed with four reactants. The collection rate is 20 s. Data consist of 18,887 observations on 36 process variables. The process variables include feed flows, pressures, vent flow and steam flow.

The reaction generates multiple products; the operating objective is to maximize the yield of the desired products. The data do not contain any product quality data. Reactor temperatures and the mix of the four reactants are the key for achieving the desired product mix and yield. The reactant feed rates, reactor pressure and vaporization tubes are manipulated by the operator. There is no direct control of reactor temperature. The operator tends to move multiple process inputs at the same time, which causes a high degree of correlation in the reactor temperatures.

5.1. Interval-wise compression

The data set considered contains a very high number of observations which may be difficult to handle. In addition, if data are compressed interval-wise, the measurement noise is expected to be reduced. Nonetheless, the size of the interval should be carefully chosen in order to avoid an important impact in the projection model due to the compression. This can be performed using ADICOV. In the experiment with random data in Section 3, it was shown that the reduction in the number of observations implies the increase of the ADICOV index, evidencing a higher difficulty for small data sets to approximate a given data distribution with a fixed covariance. However, the extent to which this occurs depends on other interrelated factors, such as the eigenvalues of the covariance matrix and, specially, the degree of similarity between matrices \( L_1 \) and \( A \). Thus, the correct size of the interval should be evaluated from the specific data of the process under analysis.

The selection of the interval size with ADICOV in this example is performed as follows. Initially, a number of possible intervals are considered: interval size (minutes) = \( 2^a \) for \( a = \{0, 1, \ldots, 9\} \). Computing the average observation of each interval, the corresponding compressed data matrices are obtained: \( \{L_1, \ldots, L_{512}\} \). Then, ADICOV is employed to find the least squares approximation of the \( L \) matrices, \( \{A_1, \ldots, A_{512}\} \), with the covariance of the non-compressed data set. This approximation is carried out on the original space of the variables, so that \( R \) and \( Q \) were set to the \( 36 \times 36 \) identity matrix \( I \). Finally, the ADICOV index (Eq. (15)) is computed for each of the 10 cases: \( \{I_{1}, \ldots, I_{512}\} \). The result is shown in Fig. 5. It can be seen that a compression interval of 1 h introduces a distortion of the covariance ten times higher than a 2 min compression interval. This information is useful to select an adequate interval.

According to Fig. 5, as expected, the lower the size of the interval used for compression, the lower the ADICOV index as a consequence of a lower distortion of the covariance matrix in \( L \) matrices. The index increase from 256 to 512 is specially abrupt. Fig. 6 illustrates the extent to which the distortion indicated by the index affects a PCA model fitted from the data. For this, the loading plot of the first PC and the eigenvalues corresponding to the original data matrix (X) and matrices \( L_1, L_{256} \) and \( L_{512} \) are compared. The highest distortion is found for \( L_{512} \) whereas \( L_1 \) shows almost no distortion. These results are coherent with Fig. 5.

<table>
<thead>
<tr>
<th># obs.</th>
<th>( I_{QRE} )</th>
<th>( I_{IM} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.0705</td>
<td>0.0846</td>
</tr>
<tr>
<td>18,878</td>
<td>0.0373</td>
<td>0.0272</td>
</tr>
</tbody>
</table>

Fig. 5. ADICOV evaluation of the interval size in interval-wise compression.

Fig. 6. Comparison of the loading vector corresponding to the first PC (a) and the eigenvalues (b) for the original matrix X and compressed matrices \( L_1, L_{256} \) and \( L_{512} \).
5.2. Missing data estimation

To assess the distortion due to the estimation of missing values, one third of the elements in the original data matrix is randomly discarded and considered to be missing values. Since the influence of estimations in the final model is also dependent on the size of the data set, two cases are considered: a) only the first 100 observations with missing values are available in \( X_1 \) (100×36) and b) the complete data set with missing values is available in \( X_2 \) (18,878×36). The covariance of each of these matrices is computed by considering only the available information.

For each of the two data sets, two techniques are employed for model building: the NIPLS algorithm [1] and the iterative Trimmed Score Regression (TSR) algorithm [22]. In both cases, a 2 PCs PCA model is used to perform the estimates. These estimates replace the missing elements in the \( X \) matrices, yielding matrices \( \{L_{\text{NIPLS}}^1 , L_{\text{NIPLS}}^2 \} \) for NIPLS and matrices \( \{L_{\text{TSR}}^1 , L_{\text{TSR}}^2 \} \) for TSR. Then, ADICOV is applied to find the least squares approximation of the \( L \) matrices on the original space of the variables. Thus, again \( R \) and \( Q \) are set to the 36×36 identity matrix \( I \). Finally, the ADICOV index (Eq. (15)) is computed for each of the cases. The result is shown in Table 1. As expected, the influence of estimations in the model is higher for a lower number of observations. Also, the estimation method with lower index varies for the two cases (100 and 18,878 observations). It was surprising to find that the estimation based on NIPLS, which is in general regarded to provide a bad estimation performance [22], outperformed TSR for the first case. This example shows that there is not a single “best” method for all the possible cases, and illustrates the convenience of using ADICOV.

Again, the extent to which the introduction of estimates in the data distorts the model can be visually assessed by looking at score plots and eigenvalue plots. This is shown for \( X_1 \) (first 100 observations) in Fig. 7. The distortion in the model due to the estimation of missing elements is especially high for the second PC onwards. Still, it is difficult to select between the two estimation procedures, NIPLS and TSR, by looking at this plot, whereas with ADICOV this selection is straightforward. The fact that the distortion affects from the second PC onwards could also be found with ADICOV by selecting the subspaces corresponding to the first PC, first 2 PCs, and so on.

5.3. Process monitoring

The last application of ADICOV considered in this paper is process monitoring. This example is restricted to the first phase of the construction of monitoring charts. A PCA model with 2 PCs is fitted using the original data set and Hotelling’s \( T^2 \) statistic and Q-statistics are computed. In Section 5.1, interval-wise compression was considered. To illustrate the consequence of such compression in monitoring, original data is averaged using a 64 min interval. The size of the interval was chosen to improve visualization of the results, while similar results were observed for lower intervals. From this compressed data, another PCA model with 2 PCs and corresponding statistics is computed. Finally, and using this same interval length, ADICOV-based monitoring statistics, that is \( \text{ID} \) and \( \text{IQ} \) indices, are computed.

After proper scaling, the resulting statistics of each of the three approaches are compared in Fig. 8, both for the model subspace (D-
In this paper, a new method to approximate a data set by another data set with a given covariance matrix is proposed. The method is termed Approximation of a Distribution for a given Covariance (ADICOV). The method performs a constrained least squares approximation which is solved for any projection subspace, including that of Principal Component Analysis (PCA) and Partial Least Squares (PLS).

ADICOV was deeply analyzed using random data to understand which factors are more influential in the approximation of a data set for an imposed covariance matrix. Relevant factors by order of importance are: the similarity between the actual covariance structure of the data set approximated and that used in the approximation, the number of observations rows in the data matrices and the dimension of the subspace of interest.

ADICOV may be applied with two different goals: a) to yield data with a specific covariance structure and space distribution and b) to test whether a data set satisfies the covariance structure in a projection model. To illustrate the latter, an industrial case study and two types of applications were considered:

- To select, among a number of possible processing operations or methods, the one which introduces less distortion in the covariance structure from original data. ADICOV-based statistics in the model subspace (Fig. 8(a)) are similar to traditional statistics, but a higher baseline is observed in the former. Also, those $I^D$ indices corresponding to observations under control seem to be more stable. In the residual subspace (Fig. 8(b)), clear differences between $I^S$ indices and $Q$-statistics are found in two periods: between observations 5,500 and 7,000, approximately, and the final period. In the first period, ADICOV shows abnormally high $I^S$ values in comparison to that of the rest of observations, while the $Q$-statistics remain normal. This period coincides with an abnormality detection in the $D$-statistic chart. ADICOV leads to conclude that this abnormality also affects the covariance of the residuals, even though a large deviation in the residual subspace ($Q$-statistic) is not found. Clearly, the conclusion regarding the abnormality in this period is different if ADICOV is not used. Fig. 9 shows the scores corresponding to the third PC of the original data, which was left in the residuals. The scores are moving through different operation points, which reflect that a tighter control may easily reduce process variability. In the period highlighted by ADICOV, the scores show a sharp change of value. This plot supports that the abnormality found in the first period in Fig. 8(a), which reflects a change of operation point in the process, is also present in the residuals. Notice that the first two PCs do present a more stable profile, exception made on the two abnormal periods highlighted. Regarding the last period of the process, the $Q$-statistics signal an abnormality which is smoothed in the ADICOV chart. This abnormality is also clear in the score plot of Fig. 9. Unlike what happened in the previous case, the abnormal behavior at the end of the process is also clearly manifesting in the fourth PC, among several others. This may cause that, while the deviation from the mean may be clear, the correlation structure in the data may not be so seriously affected.

Fault detection using traditional statistics has proven to be a very valuable tool. According to the results, ADICOV in this context may be seen as a complementary tool. Especially since ADICOV may be applied to monitor sets of observations instead of individual observations.² Future research, needed to derive statistically grounded control limits and contribution plots for this approach, will determine in which situations the ADICOV indices are more appropriate than the traditional statistics in the context of process monitoring.

6. Conclusion

Fault detection using traditional statistics has proven to be a very valuable tool. According to the results, ADICOV in this context may be seen as a complementary tool. Especially since ADICOV may be applied to monitor sets of observations instead of individual observations. Future research, needed to derive statistically grounded control limits and contribution plots for this approach, will determine in which situations the ADICOV indices are more appropriate than the traditional statistics in the context of process monitoring.

² Nonetheless, extensions of this approach may provide an ADICOV-based monitoring system for individual observations.
structure. Possible examples are selecting the method for missing data estimation and selecting the method for data compression.

• When the aim is to assess the stability of a process. Examples are process monitoring and assessing the robustness of calibration models in spectroscopy.

Other potential applications, not considered in this paper, are model selection, model validation and classification.

Acknowledgments

Research in this paper is partially supported by the Spanish Ministry of Science and Technology through grant TEC2008-06663-C03-02. The referees are acknowledged for their useful comments.

References