Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 1–27, http://dx.doi.org/10.1515/JOS-2017-0002

# A Simple Method for Limiting Disclosure in Continuous Microdata Based on Principal Component Analysis

*Aida Calviño*[1,2]

In this article we propose a simple and versatile method for limiting disclosure in continuous microdata based on Principal Component Analysis (PCA). Instead of perturbing the original variables, we propose to alter the principal components, as they contain the same information but are uncorrelated, which permits working on each component separately, reducing processing times. The number and weight of the perturbed components determine the level of protection and distortion of the masked data. The method provides preservation of the mean vector and the variance-covariance matrix. Furthermore, depending on the technique chosen to perturb the principal components, the proposed method can provide masked, hybrid or fully synthetic data sets. Some examples of application and comparison with other methods previously proposed in the literature (in terms of disclosure risk and data utility) are also included.

*Key words:* Statistical disclosure control; microdata protection; hybrid microdata; masking method; propensity score.

## 1. Introduction and Motivation

Limiting disclosure risk is a very important and hard task that agencies that publish collected data must deal with. The objective is to prevent users from being able to learn personal information about a certain individual (data can refer to people, enterprises, etc.) from any published data product. Statistical Disclosure Control (SDC) methods provide means of protecting the providers' confidential data, and approaches range from the simplest methods such as noise addition (see Brand (2002) for a survey and comparison of different SDC methods based on noise addition) to more complex ones such as synthetic data generation based on multiple imputation (see Rubin 1993).

Roughly speaking, SDC methods can be divided into three main categories: masking methods, (fully or partially) synthetic data generators, and hybrid data generators (for a full review on SDC methods see Hundepool et al. 2012). Masking refers to the process of producing a modified safe data set from the original, whereas synthetic data generators replace the original data for all or some variables with modeled (synthetic) data designed to

[1] Department of Computer Science and Mathematics, Universitat Rovira i Virgili, 43007 Tarragona, Spain
[2] Department of Statistics and Operations Research III, Complutense University of Madrid, 28040 Madrid, Spain
Email: aida.calvino@ucm.es

**Acknowledgments:** The author is indebted to Prof. Josep Domingo-Ferrer for his insightful comments, and to the four referees and the Associate Editor for their interesting comments that allowed her to improve the article substantially. This work was partly supported by the Spanish Government through project TIN2014-57364-C2-R (SmartGlacis), the European Commission through project H2020 "CLARUS" and the Government of Catalonia under grant 2014 SGR 537.

© Statistics Sweden

preserve specified properties of the original data. Finally, hybrid data generators combine original and synthetic data in order to obtain protected data sets closer to the original one.

Regarding masking methods, Duncan and Pearson (1991) showed that many of them, such as noise addition or microaggregation, fall into the broader category of “matrix masking”, which consists of masking a matrix $X$ as: $\tilde{X} = M_1 X M_2 + M_3$, where $\tilde{X}$ is the masked matrix, $M_1$ is a record-transforming mask, $M_2$ is a variable-transforming mask, and $M_3$ is a displacing mask. For instance, simple noise addition is a particular case of matrix masking where $M_1$ and $M_2$ are identity matrices and $M_3$ is a matrix containing realizations of a specific random vector. Another example of matrix masking is microaggregation (which groups similar records together and releases the average record of each group), where $M_2$ and $M_3$ are the identity and zero matrices, respectively, and $M_1$ is a block diagonal matrix such that the elements in a block can be considered similar and, therefore, are aggregated together.

An interesting method that is worth mentioning, and which does not belong to the previous category, is data swapping, which consists of swapping the values of the records univariately. So as to maintain the variance-covariance matrix as close as possible to the original one, rank swapping, a particular case of data swapping proposed by Moore (1996), limits the range of swapping values to the ones whose rank does not differ more than a prespecified threshold. As noted by Muralidhar et al. (2014), data swapping has a high level of user acceptance, as the values themselves do not suffer any modification and the unidimensional distributions are preserved.

As for synthetic data generators, the first and simplest method in this category was proposed by Liew et al. (1985) and consists of releasing a random sample from the statistical distribution underlying the original data set. However, it is not always possible to find the underlying distribution. Alternatively, Rubin (1993) proposed to generate synthetic data sets by means of the multiple imputation methodology, considering the records to be protected as missing and imputing them. Although this method provides data sets with very high utility, it is quite complex (see Raghunathan et al. (2003) or Drechsler (2011) for more information on multiple imputation-based SDC methods). Therefore, simpler alternatives to generate synthetic data have been explored, such as the Information Preserving Statistical Obfuscation (IPSO) proposed by Burridge (2003). The IPSO method basically sustitutes the original confidential variables with draws from a multivariate normal distribution with parameters obtained conditional on the non-confidential variables. Another alternative includes the synthetic data by bootstrap in Fienberg (1994), which has some similarities with the methods in Liew et al. (1985) and Rubin (1993), and consists of releasing a random draw from the smoothed empirical cumulative distribution function of the original data.

When this kind of SDC methods are applied, it is possible to release only synthetic variables (fully synthetic data) or to replace only the most sensitive or identifying variables with synthetic ones and release the remaining original ones (partially synthetic data). The main advantage of synthetic data is that, at first glance, no respondent re-identification seems possible as the data are artificial. However, this is not always true, as synthetic values can be very close to original ones if overfitting takes place. Nevertheless, as stated by Hundepool et al. (2012), the intruder never knows if a unit in the released data was actually in the original data set.

Some authors, as noted in Muralidhar and Sarathy (2008), argue that the utility of synthetic data sets is limited to the properties initially selected by the generating method. For this reason, and as an attempt to obtain synthetic data sets closer to the original ones, hybrid data generators have been proposed. Hybrid data methods combine original and synthetic data and, depending on how the combination is computed, they can lead to data closer to the original data or to the synthetic data (the definition of "combination" is different depending on the author). Interesting references of this kind of methods are Muralidhar and Sarathy (2008) and Domingo-Ferrer and González-Nicolás (2010), both of whom proposed methods that exactly preserve the mean vector and the covariance matrix of the original data and use the IPSO method in Burridge (2003) as a special case.

The method in Muralidhar and Sarathy (2008) (named MS in the sequel) essentially proposes to substitute the confidential variables by a convex linear combination of the original variables and the synthetic ones given by the IPSO method. Regarding the method in Domingo-Ferrer and González-Nicolás (2010), which is known as MicroHybrid (MH), the idea is to obtain the same groups as with the classical microaggregation but, instead of releasing an aggregated value of the records in the group, they propose to apply the IPSO method independently in each group and release the resulting synthetic values. Note that, for both cases, the extreme cases imply releasing the original data set or a partially synthetic one (as only the confidential variables are protected when applying these methods).

The quality of an SDC method depends on the data utility and the disclosure risk of the output data set. However, no general measure exists to evaluate those features as, depending on the posterior use, preservation of some data characteristics may be more or less important (see, for example, Sarathy and Krishnamurty 2002). Mateo-Sanz et al. (2005) proposed the probabilistic information loss index in order to measure the preservation of certain statistics, such as the mean or the variance. Similarly, Woo et al. (2009) proposed to use propensity scores as a measure of global data utility. From the disclosure risk perspective, a very common measure is distance-based record linkage, which was first proposed in Pagliuca and Seri (1999). All these methods will be further explainedand used in Section 4.

In this article we propose a simple method for limiting disclosure in continuous microdata based on Principal Component Analysis (PCA). The use of PCA is not new in the SDC literature. Banu and Nagaveni (2009) proposed a privacy-preserving clustering method that released masked data obtained from projecting the original data onto a transformation matrix built from the principal component loading matrix (which is obtained from a subset of the original data). However, the utility of this method is limited to cluster analysis, as it permits obtaining similar clusters but does not preserve any other statistic.

Instead of perturbing the original variables directly, the proposed method alters the principal components, as they contain the same information but are uncorrelated, which permits working on each component separately. Due to this last fact, the computation cost can be reduced, as it is less computationally intensive to work univariately than multivariately. However, it is important to highlight that the method is only applicable to continuous data, as it makes use of the classic PCA, which also requires continuous variables. Extending the methodology to categorical variables is still an open question.

The main advantages of the proposed method are:

1. Along the lines of the IPSO method in Burridge (2003) and the one proposed by Muralidhar and Sarathy (2008), the proposed method aims to preserve the mean vector and covariance matrix, as they are sufficient statistics for the multivariate normal distribution. Moreover, even in the absence of normality, many parametric statistical analyses (such as linear regression) will lead to the same result if those statistics are preserved.
2. The proposed method is very flexible, as it permits choosing any (univariate or multivariate) SDC method on the principal components as long as they preserve (at least asymptotically) the mean and variance of the principal components. In this sense, the method can provide masked, hybrid, or fully synthetic data sets depending on the choice of method. This is a great advantage as it gives the user the freedom to choose the kind of data set they need.
3. The proposed method is also fast, as the protection can be applied univariately, thus reducing computation times and making the protection process easier and more effective.
4. If one of the variables is a linear combination of the others, the number of components is less than the number of the variables and this variable is not involved in the protection process. This, however, does not represent a problem, because a protected version of this variable can still be obtained by adding the corresponding protected variables.
5. Some of the methods proposed in the literature (such as the IPSO method proposed by Burridge 2003) impose the same level of perturbation in all the variables of the data set. The proposed method, on the other hand, allows a choice of different levels of perturbation by means of the weights of the original variables in the principal components.
6. The proposed method is very simple and, therefore, more accessible to the public than more complex alternatives. Nevertheless, it still leads to a good balance between data utility and disclosure risk, as will be shown in Section 4. Finally, its computational effort is linear in the number of records, making it suitable for large data sets (see Subsection 3.4).

The article is organized as follows. In Section 2, we provide background on Principal Components Analysis, including specific details on procedures. Section 3 is devoted to the proposed model and analysis of its characteristics. Some examples of application and comparison with other methods previously proposed in the literature (in terms of disclosure risk and data utility) are shown in Section 4. Section 5 includes some guidelines for the application of the proposed method. Finally, in Section 6 we give some conclusions and future research lines.

## 2. Principal Component Analysis

In this section, we review PCA and how it is performed. PCA is a classic statistical technique designed to identify the causes of data variability and to order them by importance. PCA builds a linear transformation that chooses a new orthogonal coordinate system and changes the data coordinates such that the largest variance is captured by the first axis, the second largest by the second, and so on. For a more comprehensive description of PCA, see Jolliffe (2002).

Mathematically speaking, the principal components are found as follows. Let $\boldsymbol{X}$ be the $n \times m$ matrix containing $n$ observations of $m$ random variables with column-wise zero empirical mean. PCA seeks to find a linear transformation of $\boldsymbol{X}$ as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}, \tag{1}$$

where $\boldsymbol{Y}$ is the transformed data and $\boldsymbol{W}$ is the transforming matrix, whose columns ($\boldsymbol{w}_i$) are the unit loading vectors.

The variance of the first principal component ($\boldsymbol{y}_1$) is given by:

$$\boldsymbol{y}_1 = \boldsymbol{w}'_1\boldsymbol{X}'\boldsymbol{X}\boldsymbol{w}_1. \tag{2}$$

As we look for the most informative component, $\boldsymbol{w}_1$ must lead to a vector $\boldsymbol{y}_1$ with maximum variance, that is,

$$\boldsymbol{w}_1 = \arg \max_{t/\|\boldsymbol{t}\|=1} \{\boldsymbol{t}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{t}\} = \arg \max_{\boldsymbol{t}} \left\{\frac{\boldsymbol{t}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{t}}{\boldsymbol{t}'\boldsymbol{t}}\right\}. \tag{3}$$

Note that the right-most quotient corresponds to the renowned Rayleigh quotient, which reaches its maximum value $\lambda_{max}$ (the highest eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$) when $\boldsymbol{t}$ is the corresponding eigenvector. Therefore, $\boldsymbol{y}_1$ has maximum variance when $\boldsymbol{w}_1$ equals the eigenvector associated with the highest eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$.

Following the same idea, it can be shown that the remaining loading vectors are given by the remaining eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ sorted by importance (variability) according to its corresponding eigenvalue.

To sum up, the principal components of an $n \times m$ matrix $\boldsymbol{X}$ are given by the columns of $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}$ such that the columns of $\boldsymbol{W}$ are the eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$.

It is important to highlight the fact that the variables need to be standardized prior to applying PCA, if they are not in the same unit (i.e., height and weight) or if, although being in the same unit, they have greatly varying sizes (i.e., state level population and number of employed persons), as the components are obtained as sums of the original variables.

## 3. Proposed Method

The proposed method consists of obtaining the principal components of the original data, to later perturb them. The reason to choose to work with the principal components is twofold: first, they are uncorrelated, which permits modifying them independently without perturbing the variance-covariance structure; and second, the components can be sorted by importance, which permits us to choose what components to alter, based on this information. Furthermore, the components are obtained by linear combinations of the original variables and, therefore, when perturbing a certain component, we are perturbing mainly the original variables with corresponding higher weights. There might be cases where the data owner is interested in perturbing a very sensitive variable more than others. In that case, we can analyze the scores (or weights) of the original variables on the components and decide to perturb only those with highest weights for the sensitive variable.

By using this method, all the variables in the data set need to be considered as, otherwise, the correlation structure with other variables could be destroyed. However, as mentioned above, it is possible to alter the sensitive variables more than the nonsensitive ones by means of a careful selection of the components to be altered. Note that it is not possible to leave any variables unchanged (unless no correlation exists between the confidential and nonconfidential variables).

### *3.1. Illustrative Example*

In this section we show an example of the PCA-based method, in order to better illustrate the following sections. Consider the data set in Table 1, which consists of three continuous variables $X$, $Y$, and $Z$, such that $X$ and $Y$ are highly correlated, and $Z$ is negatively correlated with both $X$ and $Y$. Table 1 also contains the principal components (PCs) of the data set. The weights of the variables on the principal components are:

$$\begin{array}{c|ccc} & \mathrm{PC1} & \mathrm{PC2} & \mathrm{PC3} \\ \hline \mathrm{X} & -0.6183 & 0.4407 & 0.6508 \\ \mathrm{Y} & -0.6656 & 0.1468 & -0.7317 \\ \mathrm{Z} & 0.4180 & 0.8856 & -0.2026 \end{array} \tag{4}$$

As already stated, the basis of the method is to perturb the principal components. This perturbation can be made univariately because of the uncorrelation of the principal components (note that the correlation matrix of the PCs in Table 1 is the identity matrix). In order to illustrate the method, we apply data swapping to the first components (the one with highest variance) and undo the transformation given by the matrix in Equation (4). The data swapping process consists of randomly sorting the observations, which is done by generating a random vector whose elements range from 1 to $n$ and by rearranging the observations according to this new order.

The results are shown in Table 2, where the perturbed components are on the left and the resulting variables are on the right.

As can be seen, compared with Table 1, the preservation of means is exact, while standard deviations and correlations are very close to the original values. Regarding protection, there are two records that remain unchanged, while the remaining ones get very different values. The reason why two records have remained unchanged is that there are only ten records altogether and, therefore, the swapping can lead to no changes with a relatively high probability taking into account that we have perturbed only the first component. If more components are perturbed, this probability decreases, as the probability of not swapping an observation equals $\frac{1}{n}$.

Table 3 shows the results obtained if PC1 and PC2 are swapped and when all the three components are swapped. Note that in both cases the level of protection and data utility is high. Moreover, no record is left unchanged and the contribution of PC3 is minimal as it contains less than four percent of the total variability (the records in the right table are not very different from those in the left one).

*Table 1. Illustrative example: data and principal components.*

| | X | Y | Z | | PC1 | PC2 | PC3 |
|---|---|---|---|---|---|---|---|
| | 26.20 | 26.13 | 85.98 | | 2.3095 | 0.4611 | −0.0835 |
| | 98.61 | 124.85 | 74.25 | | −1.1954 | 1.2161 | 0.4253 |
| | 43.97 | 60.01 | 55.26 | | 0.4223 | −1.2144 | 0.3255 |
| | 55.67 | 126.72 | 68.07 | | −0.3816 | 0.0625 | −0.6000 |
| | 82.51 | 147.21 | 58.13 | | −1.6351 | −0.0870 | −0.0773 |
| | 33.52 | 87.71 | 52.91 | | 0.2127 | −1.4687 | −0.3291 |
| | 72.80 | 118.89 | 85.34 | | −0.1328 | 1.5062 | −0.3122 |
| | 40.26 | 32.56 | 80.02 | | 1.6889 | 0.3184 | 0.2680 |
| | 38.14 | 43.15 | 70.48 | | 1.2865 | −0.3333 | 0.1997 |
| | 97.50 | 165.31 | 48.00 | | −2.5749 | −0.4609 | 0.1835 |
| Mean | 58.92 | 93.25 | 67.84 | Mean | 0.0000 | 0.0000 | 0.0000 |
| Std Dev | 26.96 | 50.24 | 13.75 | Std Dev | 1.5261 | 0.9445 | 0.3351 |
| Correlation | | | | Correlation | | | |
| X | 1.0000 | 0.8664 | −0.2417 | PC1 | 1.0000 | 0.0000 | 0.0000 |
| Y | 0.8664 | 1.0000 | −0.4639 | PC2 | 0.0000 | 1.0000 | 0.0000 |
| Z | −0.2417 | −0.4639 | 1.0000 | PC3 | 0.0000 | 0.0000 | 1.0000 |

*Table 2. Illustrative example: perturbed principal components (left) and masked data (right).*

| | PC1′ | PC2 | PC3 | | X′ | Y′ | Z′ |
|---|---|---|---|---|---|---|---|
| | −2.5749 | 0.4611 | −0.0835 | | 103.45 | 181.07 | 59.35 |
| | −1.1954 | 1.2161 | 0.4253 | | 98.61 | 124.85 | 74.25 |
| | −0.1328 | −1.2144 | 0.3255 | | 52.75 | 77.62 | 52.23 |
| | 2.3095 | 0.0625 | −0.6000 | | 13.11 | 41.35 | 82.74 |
| | 0.2127 | −0.0870 | −0.0773 | | 53.29 | 88.59 | 68.20 |
| | −1.6351 | −1.4687 | −0.3291 | | 62.74 | 146.33 | 42.84 |
| | 0.4223 | 1.5062 | −0.3122 | | 64.02 | 101.28 | 88.37 |
| | 1.6889 | 0.3184 | 0.2680 | | 40.26 | 32.56 | 80.02 |
| | −0.3816 | −0.3333 | 0.1997 | | 64.52 | 96.06 | 61.39 |
| | 1.2865 | −0.4609 | 0.1835 | | 36.43 | 42.82 | 69.05 |
| Mean | 0.0000 | 0.0000 | 0.0000 | Mean | 58.92 | 93.25 | 67.84 |
| Std Dev | 1.5261 | 0.9445 | 0.3351 | Std Dev | 27.19 | 48.06 | 14.21 |
| Correlation | | | | Correlation | | | |
| PC1′ | 1.0000 | 0.0583 | −0.1563 | X′ | 1.0000 | 0.8664 | −0.2857 |
| PC2 | 0.0583 | 1.0000 | 0.0000 | Y′ | 0.8664 | 1.0000 | −0.4970 |
| PC3 | −0.1563 | 0.0000 | 1.0000 | Z′ | −0.2857 | −0.4970 | 1.0000 |

*Table 3. Illustrative example: masked data when PC1 and PC2 have been swapped (left) and all the principal components have been swapped (right).*

|  | X'' | Y'' | Z'' |  | X''' | Y''' | Z''' |
|---|---|---|---|---|---|---|---|
|  | 93.05 | 174.62 | 48.71 |  | 97.50 | 165.31 | 48.00 |
|  | 90.10 | 119.57 | 65.53 |  | 87.48 | 125.05 | 65.94 |
|  | 83.42 | 96.65 | 83.66 |  | 72.52 | 119.48 | 85.38 |
|  | 8.65 | 38.58 | 78.17 |  | 25.72 | 2.83 | 75.46 |
|  | 57.86 | 91.43 | 72.89 |  | 57.75 | 91.65 | 72.90 |
|  | 65.61 | 148.11 | 45.77 |  | 74.41 | 129.66 | 44.38 |
|  | 47.75 | 91.18 | 71.69 |  | 42.96 | 101.22 | 72.45 |
|  | 20.12 | 20.06 | 59.38 |  | 10.46 | 40.29 | 60.91 |
|  | 67.30 | 97.79 | 64.23 |  | 62.69 | 107.45 | 64.96 |
|  | 55.34 | 54.55 | 88.42 |  | 57.70 | 49.60 | 88.04 |
| Mean | 58.92 | 93.25 | 67.84 | Mean | 58.92 | 93.25 | 67.84 |
| Std Dev | 27.94 | 47.38 | 14.00 | Std Dev | 26.78 | 48.73 | 14.24 |
| Correlation |  |  |  | Correlation |  |  |  |
| X'' | 1.0000 | 0.8258 | −0.2354 | X''' | 1.0000 | 0.8652 | −0.2847 |
| Y'' | 0.8258 | 1.0000 | −0.5947 | Y''' | 0.8652 | 1.0000 | −0.4964 |
| Z'' | −0.2354 | −0.5947 | 1.0000 | Z''' | −0.2847 | −0.4964 | 1.0000 |

### *3.2. Mathematical Formulation*

Let $\boldsymbol{X}$ be the $n \times m$ matrix containing the $n$ observations of the $m$ random variables such that $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the $\boldsymbol{X}$'s mean and standard deviation vectors, respectively.

In order to correctly apply PCA, we first need to standardize $\boldsymbol{X}$. In matrix form this is equivalent to:

$$\boldsymbol{X}_1 = (\boldsymbol{X} - \boldsymbol{M})\boldsymbol{S}, \tag{5}$$

where $\boldsymbol{X}_1$ is the standardized data set, $\boldsymbol{M}$ is an $n \times m$ matrix with rows equal to $\boldsymbol{\mu}$ and $\boldsymbol{S}$ is a $m$-diagonal matrix with non-zero elements equal to $s_{jj} = 1/\sigma_j$.

Next, we obtain the scores $\boldsymbol{X}_2$ of the observations on the principal components as:

$$\boldsymbol{X}_2 = \boldsymbol{X}_1\boldsymbol{A}, \tag{6}$$

where $\boldsymbol{A}$ is a matrix whose columns are the normalized (to unit vectors) eigenvectors of $\boldsymbol{X'}_1\boldsymbol{X}_1$.

As already stated, instead of perturbing the data directly, the basis of the proposed method consists of perturbing the scores $\boldsymbol{X}_2$. Mathematically speaking, the perturbation process can be written as:

$$\boldsymbol{X}_3 = \boldsymbol{X}_2\boldsymbol{B} + \boldsymbol{\epsilon}, \tag{7}$$

where $\boldsymbol{B}$ is an $m \times m$ diagonal matrix with ones in the rows corresponding to the $m_c$ components left unchanged and zeros otherwise and $\boldsymbol{\epsilon}$ is an $n \times m$ matrix with $m_c$ columns equal to zero and the remaining ($m_r$) ones contain the variables that are replacing the original principal components.

Finally, we obtain the masked data $\tilde{\boldsymbol{X}}$ by undoing the PCA transformation and recovering the original means, variances and covariances:

$$\tilde{\boldsymbol{X}} = (\boldsymbol{X}_3\boldsymbol{A}'\boldsymbol{S}^{-1}) + \boldsymbol{M}, \tag{8}$$

where $\boldsymbol{A}'$ is the transpose of $\boldsymbol{A}$ and $\boldsymbol{S}^{-1}$ is the inverse of $\boldsymbol{S}$. Recall that matrix $A$ is composed of normalized eigenvectors and, therefore, its inverse equals its transpose.

It is important to highlight that any SDC method can be applied as long as it preserves the properties of the original principal components, that is, the mean vector is equal to zero and it has the same diagonal variance-covariance matrix (note that the preservation of higher moments in the perturbed components leads to a better data utility of the final data set). The "protected" components obtained when applying the SDC method form matrix $\boldsymbol{\epsilon}$ in Equation (7). Some examples of possible SDC methods are random scores, noise addition, and swapping.

An interesting feature of this method is that depending on the method chosen to perturb the principal components, we can get masked, hybrid or fully synthetic data. In particular, if a masking method is chosen to alter the principal components, such as data swapping (Moore 1996), the resulting data set is masked, as the original values have been modified but not substituted. Similarly, if the components are substituted by random vectors (by means of the methodology in Liew et al. (1985) or Fienberg (1994)), we get fully synthetic data sets if no component is left unchanged and hybrid data sets, otherwise.

As previously noted, the principal components can be sorted by importance based on quantity of variability and, therefore, perturbing the first component does not lead to the same level of protection and data utility of the output data as perturbing the last component. The larger the total weight of altered components, the lower the data utility and disclosure risk.

Note that one remarkable difference with regard to this method compared to others previously suggested is that the level of perturbation is fixed, in the sense that we can only modify a certain number of components between 1 and $m$, whereas in other methods, such as noise addition, any quantity of noise can be added. It is important to take into account that, although the level of perturbation of the components is not limited, the effect on the final data set is, as the weight of the components is fixed. In other words, if only the "last" component is perturbed, even if the protected component has no resemblance to the original one, the effect on the final data set will be small, as only a small portion of the original variability has been changed.

### *3.3. Verifying the Preservation of the Mean Vector and the Variance-Covariance Matrix*

Preserving the mean vector and the variance-covariance matrix of the original data is a very important feature of a masking method which is very common in the SDC literature. For this reason, we will now show that the proposed method preserves both the mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}_X$.

So as to facilitate the computations, we first show the direct relation between $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$, following from (5)-(8):

$$\tilde{\boldsymbol{X}} = [(\boldsymbol{X} - \boldsymbol{M})\boldsymbol{SAB} + \boldsymbol{\epsilon}]\boldsymbol{A}'\boldsymbol{S}^{-1} + \boldsymbol{M} = (\boldsymbol{X} - \boldsymbol{M})\boldsymbol{SABA}'\boldsymbol{S}^{-1} + \boldsymbol{\epsilon A}'\boldsymbol{S}^{-1} + \boldsymbol{M}. \quad (9)$$

First, we deal with the mathematical expectation of the masked data set $\tilde{\boldsymbol{X}}$:

$$\begin{aligned} E[\tilde{\boldsymbol{X}}] &= E[(\boldsymbol{X} - \boldsymbol{M})\ \boldsymbol{SABA}'\boldsymbol{S}^{-1} + \boldsymbol{\epsilon A}'\boldsymbol{S}^{-1} + \boldsymbol{M}] \\ &= (E[\boldsymbol{X}] - E[\boldsymbol{M}])\ \boldsymbol{SABA}'\boldsymbol{S}^{-1} + E[\boldsymbol{\epsilon}]\boldsymbol{A}'\boldsymbol{S}^{-1} + E[\boldsymbol{M}] \qquad (10) \\ &= (\mu - \mu)\ \boldsymbol{SABA}'\boldsymbol{S}^{-1} + \boldsymbol{0}\boldsymbol{A}'\boldsymbol{S}^{-1} + \mu = \mu. \end{aligned}$$

Next, we focus on the variance of $\tilde{\boldsymbol{X}}$. From now on, $\boldsymbol{\Sigma}_X$ refers to the variance-covariance of data set $\boldsymbol{X}$.

$$\begin{aligned} \boldsymbol{\Sigma}_{\tilde{X}} &= (\boldsymbol{SABA}'\boldsymbol{S}^{-1})'\boldsymbol{\Sigma}_X(\boldsymbol{SABA}'\boldsymbol{S}^{-1}) + (\boldsymbol{A}'\boldsymbol{S}^{-1})'\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{A}'\boldsymbol{S}^{-1}) + \boldsymbol{0} \\ &= (\boldsymbol{S}^{-1})'\boldsymbol{A}\boldsymbol{\Sigma}_1\boldsymbol{A}'\boldsymbol{S}^{-1} + (\boldsymbol{S}^{-1})'\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\boldsymbol{A}'\boldsymbol{S}^{-1} \qquad (11) \\ &= (\boldsymbol{S}^{-1})'\boldsymbol{A}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\boldsymbol{A}'\boldsymbol{S}^{-1}. \end{aligned}$$

where $\boldsymbol{\Sigma}_1 = \boldsymbol{B}'\boldsymbol{A}'\boldsymbol{S}'\boldsymbol{\Sigma}_X\boldsymbol{SAB} = \boldsymbol{B}'\boldsymbol{\Sigma}_{X_2}\boldsymbol{B}$.

Without loss of generality, we assume that the components have been sorted in such a way that the ones that remain unaltered are the first ones and the altered ones are the last ones. Note that $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$have a special block structure as shown below. This is due to the fact that $\boldsymbol{\Sigma}_1$ is obtained by multiplying and premultiplying $\boldsymbol{\Sigma}_{X_2}$, which is a diagonal matrix as it refers to the variance-covariance matrix of the principal components, by matrix $\boldsymbol{B}$, which is a diagonal matrix with zeros in the rows associated to altered components. On the other hand, $\boldsymbol{\epsilon}$ has been defined to be an $n \times m$ matrix with $m_c$ columns equal to zero and the remaining ($m_r$) ones replacing the original principal components. Therefore, the variance and covariance associated with the "zero" columns are also zero and we have forced the altered components to maintain the variances of the original principal components. Then,

$$\boldsymbol{\Sigma}_1 = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{X_2}[m_c] & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{0} \end{array}\right), \tag{12}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \left(\begin{array}{c|c} \boldsymbol{0} & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{\Sigma}_{X_2}[m_r] \end{array}\right), \tag{13}$$

where, abusing notation, we have defined $\boldsymbol{\Sigma}_{X_2}[m_c]$ and $\boldsymbol{\Sigma}_{X_2}[m_c]$ to be the submatrices associated to the $m_c$ unaltered components and $m_r$ to the altered components, respectively.

Taking (12) and (13) into account, we have

$$\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{X_2}[m_c] & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{\Sigma}_{X_2}[m_r] \end{array}\right) = \boldsymbol{\Sigma}_{X_2}, \tag{14}$$

and, thus,

$$\boldsymbol{\Sigma}_{\tilde{X}} = (\boldsymbol{S}^{-1})'\boldsymbol{A}\boldsymbol{\Sigma}_{X_2}\boldsymbol{A}'\boldsymbol{S}^{-1} = \boldsymbol{\Sigma}_X. \tag{15}$$

The perturbing method chosen for the principal components determines whether the first and second moments are exactly or asymptotically preserved. For example, if we choose to substitute component $i$ with a realization of a random variable with expectation and variance equal to $\mu_i$ and $\sigma_i^2$, respectively, it is very unlikely that the observed mean and variance equals $\mu_i$ and $\sigma_i^2$. In that case, the original mean vector and variance-covariance matrix are asymptotically (but not exactly) preserved. On the other hand if we choose swapping – as the values are the same – the mean is preserved exactly. However, this is not the case with the variance-covariance matrix, as the covariance of the perturbed principal components is not numerically zero.

Although, as previously stated, many perturbing methods are valid, we suggest using data swapping and random draws from the smoothed empirical cumulative distribution functions (see Fienberg 1994) for masking and hybrid/synthetic data sets, respectively, as they permit maintaining not only the first and second moments but the whole distribution on the univariate components. Note that rank swapping is not necessary here, as the components are uncorrelated and, therefore, the covariance structure does not need to be preserved. This, in turn, helps to preserve the distribution of the original data. Moreover, note that when the number of altered components is high, the resulting records do not clearly represent any of the original records. These are obtained by means of the components' scores of other records, selected randomly and, thus (although the resulting data set is not strictly synthetic), it does bear some similarity to the synthetic data generation philosophy.

### 3.3.1. On the Preservation of the Third Moment

In this section we deal with the preservation of the third moment ($\mu_3$), which is related to the symmetry of the variables. We remind the reader that the third moment is given by:

$$\mu_3(X) = E[(X - E(X))^3], \tag{16}$$

and, if $\boldsymbol{X} = \boldsymbol{Y} + \boldsymbol{Z}$, then it holds that:

$$\begin{aligned}\mu_3(\boldsymbol{X}) &= \mu_3(\boldsymbol{Y}) + \mu_3(\boldsymbol{Z}) - 6Cov(\boldsymbol{Y},\boldsymbol{Z})(E(\boldsymbol{Y}) + E(\boldsymbol{Z})) \\ &\quad + 3(Cov(\boldsymbol{Y},\boldsymbol{Z}^2) + Cov(\boldsymbol{Y}^2,\boldsymbol{Z})).\end{aligned} \tag{17}$$

The proposed method essentially decomposes the original variables into some uncorrelated ones and then undoes the decomposition. Without loss of generality, let's assume that we want to preserve the third moment of an original variable $\boldsymbol{X}$, whose principal components are $\boldsymbol{Y}$ and $\boldsymbol{Z}$. As the principal components are uncorrelated, it holds that:

$$\mu_3(\boldsymbol{X}) = \mu_3(\boldsymbol{Y}) + \mu_3(\boldsymbol{Z}) + 3\,(Cov(\boldsymbol{Y},\boldsymbol{Z}^2) + Cov(\boldsymbol{Y}^2,\boldsymbol{Z})), \tag{18}$$

and, therefore, the third moment of $\boldsymbol{X}$ is preserved as long as the addends on the right-hand side are preserved. If a perturbation method is chosen such that the third moment of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are preserved (as univariate data swapping), then the preservation of $\mu_3(\boldsymbol{X})$ depends on the preservation of $Cov(\boldsymbol{Y},\boldsymbol{Z}^2)$ and $Cov(\boldsymbol{Y}^2,\boldsymbol{Z})$, which, in turn, depends on the number and weight of the perturbed components.

It is important to highlight that, if the principal components are independent (and not only uncorrelated), it holds that:

$$\mu_3(\boldsymbol{X}) = \mu_3(\boldsymbol{Y}) + \mu_3(\boldsymbol{Z}), \tag{19}$$

and, in that case, the preservation of the third moment of the original variables can be ensured, as the perturbed principal components are also independent. If multivariate normality holds, the principal components are also normal (as a consequence of the infinite divisibility of the normal distribution) and, in that case, uncorrelation implies independence. However, although generally uncorrelation does not imply independence, multivariate normality is not the only case.

To sum up, the preservation of the third moment depends on the preservation of the third moment of the perturbed principal components, as well as on the independence of the principal components.

## *3.4. Computational Effort*

As has been shown, the proposed method essentially consists of obtaining the eigenvectors of the correlation matrix (or equivalently the variance-covariance matrix of the standardized data set) and then applying products and/or sums to the original matrix data $\boldsymbol{X}$ and the transformation matrix $\boldsymbol{A}$. Finally, there might be a random number generation phase associated with the altering components phase. Therefore, the running time of the method is $O(nm^2)$, where $n$ is the number of records and $m$ is the number of variables.

Generally, the number of records is much larger than the number of variables and, therefore, the proposed method is suitable for large data sets.

## 4. Empirical Results

In this section a simulation study for the PCA-based method is shown. In particular, we evaluate its performance in terms of data utility and disclosure risk in two scenarios: a) when it is applied to get a masked data set that protects all the variables in the data set, and b) when only a subset of variables needs to be protected and the output is a hybrid data set.

The results have been obtained using R project (R Core Team 2014) and, more specifically, package sdcMicro (Templ 2008) when possible. Some ad hoc functions and programs also needed to be developed. Data sets *Tarragona* and *Census* have also been provided by this package. Regarding computation times, anonymizing a data set (of up to 55,000 records and 35 variables) by means of the proposed method takes less than one second on a Toshiba satellite L50-B-11W laptop, Intel Core i7-870 1.8GHz, 4MB, RAM: 8GB.

### *4.1. Fully Masked Data Set*

In this case, the proposed method is applied to the *Tarragona* data set in Brand et al. (2002). It consists of 13 quantitative variables associated to 832 real companies in the province of Tarragona in 1995.

For the sake of completeness, we compare the results derived from the proposed method with two well-known masking methods that have been identified as well-performing in terms of data utility and disclosure risk: rank swapping (see Domingo-Ferrer and Torra (2001) or Jiménez et al. (2014)) and microaggregation plus noise addition (see Oganian and Karr (2006), or Woo et al. (2009)).

Following the ideas of Domingo-Ferrer and Torra (2001) or Jiménez et al. (2014), we compute a score as the mean average of disclosure risk and data utility in order to be able to compare the three methods. The disclosure risk and data utility measures also consist of a score made of the mean average of two different criteria. In particular, disclosure risk is evaluated by means of distance-based record linkage and interval disclosure (see Domingo-Ferrer and Torra 2004), while data utility is computed based on the Probabilistic Information Loss (PIL) measure proposed by Mateo-Sanz et al. (2005) and the propensity scores proposed by Woo et al. (2009). In the following we briefly described these four measures:

- ***Distance-based record linkage (DBRL)***: DBRL is one of the most common methods for evaluating the disclosure risk of a masked data set. It consists of obtaining the closest masked record (in terms of normalized euclidean distance) to all original records and determining how many of them were generated by the corresponding original record. As noted in Domingo-Ferrer and Torra (2004), variables should be standardized when using distance-based record linkage in order to avoid scaling problems. This index can take values between zero percent (no record linkage) and 100% (total record linkage).
- ***Interval disclosure (ID)***: It consists of determining the proportion of original records that lay in an interval whose center is the corresponding masked record. The extremes

of the interval are given by the two masked values whose ranks differ ± $p$ percent of the total number of records. The measure associated with ID is obtained by averaging this proportion for $p$, taking values from one percent to ten percent with one percent increments. In the simulation study that follows, we have substituted the corresponding masked record by the closest record in order to be able to analyze the attribute disclosure, that is, how much the intruder can learn if re-identification takes place. Moreover, as noted previously, when the number of altered components is large, masked records have a very weak connection with the corresponding original ones (this also happens for microaggregation plus noise addition when the parameter is large). In this way, we can work with an homogeneous measure independently of the parameters of the method. This index also takes values between zero percent (no attribute disclosure) and 100% (total attribute disclosure).

- ***Probabilistic Information Loss (PIL)***: It consists of evaluating, from a probabilistically point of view, the information loss suffered from the masking processes based on the observed difference between some statistics obtained from the original and the masked data set. Given a certain parameter $\theta$ and its masked value $\hat{\theta}$, the probabilistic information loss can be measured as the standardized sample discrepancy as follows:

$$pil(\theta) = 2 \cdot P\left(0 \leq N(0,1) \leq \frac{|\theta - \hat{\theta}|}{\sqrt{Var(\hat{\theta})}}\right), \tag{20}$$

where $N(0,1)$ is a standardized normal distribution. The variances of the considered statistics are given in Mateo-Sanz et al. (2005). The final PIL measure is given by the mean average of the *pil* associated with the means, variances, covariances, Pearson'scorrelation coefficients and quantiles. The *pil* given by Equation (20) takes values between 0 (no information loss) and 1 (total information loss) and, therefore, the total PIL also takes values in that range.

- ***Propensity scores (PS)***: Propensity scores were adopted from the statistical literature by Woo et al. (2009) in an attempt to define new global measures of data utility. In the observational study literature, propensity scores are the probabilities of being assigned to a treatment, given other variables (covariates). When two large groups have the same distributions of propensity scores, the groups should have similar distributions on the covariates. Therefore, one can consider the masked data as the treatment and estimate the probability of being assigned to the treatment (the propensity scores) for both the masked and the original data sets. If the distributions of the propensity scores of both sets are similar, we can conclude that the distributions of the original and masked data are also similar and, therefore, the data utility should be relatively high. Woo et al. (2009) propose to evaluate the similarity of the propensity scores using the following formula:

$$PS = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - 0.5)^2, \tag{21}$$

where $N = 2n$ and $\hat{p}_i$ is the propensity score for unit $i$. Note that when the original and the masked data sets have similar distributions, it is difficult to distinguish them and, therefore, the propensity scores are close to 0.5 and *PS* equals 0. On the other hand, if the distributions are very different, they are perfectly distinguishable and the propensity scores for the original and masked data take approximately value 0 and 1, respectively. In that case, *PS* is close to 0.25. The main disadvantage of this method is that it relies on the choice of the model used to estimate the propensity scores. Nevertheless, the authors suggest to use a logistic regression using a second-order polynomial in all the variables, as well as in their interactions. In this article, we have considered the model suggested by the authors as well as a logistic regression with third-order polynomials in all the variables, and the same interactions.

It is important to highlight that we are considering all the variables at the same time in the disclosure risk phase, which is equivalent to assuming that the intruder has information about all the variables in the data set. Therefore, as this is rarely the case, the disclosure indexes should be taken as worst case ones. However, this is done with all the three methods and, thus, this fact will not affect the conclusions derived from the comparison.

On the other hand, all the indexes, except for the propensity scores, lay between 0 and 1 and, thus, when computing their mean average we get a score that also lies between 0 and 1. To overcome the inconvenience of the propensity score, we multiply it by a factor of four, thus obtaining an index that takes values between 0 and 1. Finally, note that the larger the score is, the worse the method is, as it has higher disclosure risk and less data utility.

We have decided to rely on several data utility and disclosure risk indexes, as each of them measures a different concept and permits obtaining a global score. Table 4 shows the results, including the four indexes explained previously (the propensity score is shown with its two variants) and the resulting score (two different scores arise because of the propensity score), of a simulation study performed in order to be able to evaluate the results of our method and compare it with microaggregation plus noise and rank swapping. As the three methods depend on random values in one way or another, and thus, very good or bad results can be obtained by chance and mask the real behavior of the method, we have to resort to a simulation study, which takes into account several realizations.

In order to be able to determine the better performance of a method, comparing the mean values of the score is not enough. For this reason, we have computed the confidence interval (CI) of the scores of the methods along with the mean average, based on 100 realizations. As the score does not belong to any known distribution, we resort to bootstrap techniques, in particular to the Percentile Bootstrap CI. The bootstrap method, which is one of a broader class of resampling methods, uses Monte Carlo sampling to generate an empirical sampling distribution of the estimate (see Efron and Tibshirani (1993) for more details on bootstrap methods).

For the PCA-based method, *parms* refers to the number of components that have been swapped starting by the one with more variance. The total proportion of altered variance is also shown in parentheses. With respect to microaggregation plus noise, *parms* refers to the number of records grouped together in the microaggregation phase. Finally, *parms* refers to the maximum relative rank difference allowed in rank swapping.

The skewness and kurtosis relative bias has been calculated as well, and, in addition, the mean average for all the variables and all the realizations is shown in Table 4.

*Table 4. Simulation study showing the parameters used ( parms), the skewness and kurtosis bias, Propensity Scores ($PS_2$ and $PS_3$ for second-degree and third-degree interactions, respectively), Probabilistic Information Loss (PIL), Distance-based Record Linkage (DBRL) and Interval Disclosure (ID) indexes, as well as the global scores for the PCA-based, "microaggregation plus noise" and rank swapping methods with different parameters. The smallest score for each of the methods has been boldfaced.*

| *parms* | Skewness Bias | Kurtosis Bias | $PS_2$ | $PS_3$ | PIL | DBRL | ID | $Score_2$ | $Score_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 (63.44%) | 0.4512 | 0.5018 | 0.0020 | 0.0988 | 0.2306 | 0.3642 | 0.4567 | 0.2633 (0.2506.0.2791) | 0.2876 (0.2757.0.2998) |
| 2 (73.26%) | 0.4863 | 0.5183 | 0.0060 | 0.1660 | 0.2436 | 0.1802 | 0.4326 | 0.2156 (0.2064.0.2308) | 0.2556 (0.2450.0.2716) |
| 3 (82.13%) | 0.4869 | 0.5421 | 0.0104 | 0.2200 | 0.2487 | 0.0974 | 0.4269 | 0.1958 (0.1873.0.2078) | **0.2482 (0.2354.0.2592)** |
| 4 (88.32%) | 0.5129 | 0.5688 | 0.0148 | 0.2792 | 0.2478 | 0.0555 | 0.4221 | 0.1850 (0.1755.0.1957) | 0.2511 (0.2372.0.2623) |
| 5 (91.60%) | 0.5192 | 0.5729 | 0.0172 | 0.3052 | 0.2494 | 0.0328 | 0.4174 | 0.1792 (0.1717.0.1884) | 0.2512 (0.2360.0.2645) |
| 6 (94.17%) | 0.5222 | 0.5725 | 0.0204 | 0.3268 | 0.2501 | 0.0178 | 0.4120 | 0.1751 (0.1666.0.1856) | 0.2517 (0.2342.0.2642) |
| 7 (96.25%) | 0.5232 | 0.5725 | 0.0228 | 0.3540 | 0.2549 | 0.0095 | 0.4095 | 0.1742 (0.1647.0.1883) | 0.2569 (0.2363.0.2759) |
| 8 (97.91%) | 0.5247 | 0.5754 | 0.0240 | 0.3548 | 0.2517 | 0.0052 | 0.4061 | 0.1718 (0.1653.0.1889) | 0.2545 (0.2293.0.2735) |
| 9 (98.78%) | 0.5234 | 0.5728 | 0.0268 | 0.3592 | 0.2547 | 0.0030 | 0.4006 | 0.1713 (0.1621.0.1851) | 0.2544 (0.2290.0.2713) |
| 10 (99.54%) | 0.5259 | 0.5740 | 0.0284 | 0.3628 | 0.2549 | 0.0018 | 0.3987 | 0.1709 (0.1619.0.1811) | 0.2545 (0.2373.0.2710) |
| 11 (99.82%) | 0.5232 | 0.5738 | 0.0280 | 0.3660 | 0.2550 | 0.0016 | 0.3971 | 0.1704 (0.1624.0.1817) | 0.2549 (0.2358.0.2756) |
| 12 (99.93%) | 0.5253 | 0.5754 | 0.0284 | 0.3612 | 0.2512 | 0.0016 | 0.3963 | 0.1694 (0.1608.0.1808) | 0.2525 (0.2367.0.2702) |
| 13 (100.00%) | 0.5268 | 0.5751 | 0.0280 | 0.3612 | 0.2501 | 0.0012 | 0.3973 | **0.1691 (0.1626.0.1783)** | 0.2525 (0.2275.0.2688) |
| 3 | 0.3615 | 0.5154 | 0.0296 | 0.3092 | 0.2864 | 0.0461 | 0.3426 | 0.1762 (0.1664.0.1863) | **0.2461 (0.2332.0.2630)** |
| 5 | 0.4429 | 0.6042 | 0.0320 | 0.4404 | 0.3058 | 0.0239 | 0.3283 | **0.1725 (0.1614.0.1866)** | 0.2746 (0.2563.0.2887) |
| 7 | 0.5189 | 0.6781 | 0.0316 | 0.6128 | 0.3156 | 0.0166 | 0.3288 | 0.1732 (0.1597.0.1885) | 0.3185 (0.2954.0.3367) |
| 10 | 0.6124 | 0.7665 | 0.0324 | 0.6544 | 0.3302 | 0.0111 | 0.3329 | 0.1766 (0.1634.0.1947) | 0.3321 (0.3170.0.3507) |
| 25 | 0.7844 | 0.8885 | 0.0340 | 0.7788 | 0.3480 | 0.0052 | 0.3392 | 0.1816 (0.1668.0.1977) | 0.3678 (0.3509.0.3928) |
| 50 | 0.8864 | 0.9328 | 0.0348 | 0.8432 | 0.3625 | 0.0028 | 0.3517 | 0.1880 (0.1712.0.2132) | 0.3901 (0.3693.0.4133) |
| 75 | 0.9300 | 0.9449 | 0.0340 | 0.8644 | 0.3670 | 0.0023 | 0.3611 | 0.1911 (0.1734.0.2103) | 0.3987 (0.3794.0.4204) |
| 0.05 | 0 | 0 | 0.1676 | 0.1784 | 0.3259 | 0.7645 | 0.6972 | 0.4888 (0.4785.0.4992) | 0.4914 (0.4817.0.5022) |
| 0.1 | 0 | 0 | 0.2864 | 0.2984 | 0.3725 | 0.4972 | 0.4441 | 0.4000 (0.3891.0.4112) | 0.4030 (0.3913.0.4140) |
| 0.15 | 0 | 0 | 0.3784 | 0.3856 | 0.3829 | 0.2599 | 0.3460 | 0.3418 (0.3309.0.3537) | 0.3436 (0.3318.0.3549) |
| 0.2 | 0 | 0 | 0.4588 | 0.4580 | 0.3872 | 0.1210 | 0.3169 | **0.3209 (0.3084.0.3323)** | **0.3207 (0.3037.0.3370)** |
| 0.5 | 0 | 0 | 0.7396 | 0.6788 | 0.3925 | 0.0094 | 0.2645 | 0.3515 (0.3404.0.3606) | 0.3363 (0.2972.0.3566) |
| 1 | 0 | 0 | 0.8572 | 0.7316 | 0.3930 | 0.0015 | 0.2362 | 0.3720 (0.3647.0.3782) | 0.3405 (0.2615.0.3706) |

The smallest score for each of the methods has been boldfaced. It can be seen that, independently of the chosen propensity score, the smallest scores are taken by the “microaggregation plus noise” and the proposed methods. Rank swapping, which has been recognized as a very well-performing technique by Domingo-Ferrer and Torra (2004) among others, leads to worse results, as its scores are higher and its confidence intervals do not overlap with those of the other methods (this is due to the fact that the variance-covariance matrix is not very well preserved). However, rank swapping is the only method that leads to zero bias on the skewness and kurtosis of the variables.

Regarding second-degree propensity scores, we see that the best results are around 0.17 for both the “microaggregation plus noise” and the proposed methods. For the PCA-based method, the best balance between data utility and disclosure risk is taken for a high number of altered components, in which case the DBRL gets very low values because of the “synthetic” aspect of the resulting masked data set. Furthermore, the data utility from the PS perspective does not decrease dramatically with increasing numbers of swapped components. However, the CIs of $\text{Score}_2$ almost coincide for more than 94% of the variability perturbed. This is due to the fact that the utility remains almost constant and the risk, although it decreases with the number of components, is already very small.

As for microaggregation plus noise, which has been recognized as a very well-performing method in Woo et al. (2009), the lowest score takes place for $parms = 5$, although the results are not statistically different up to $parms = 10$, as can be deduced from the CIs. It can be seen that this method provides better attribute disclosure protection, but worse record linkage compared with the PCA-based method, and that it has a slightly bigger score.

As regards third-degree propensity scores, we can see that the utility is worse, as both “microaggregation plus noise” and the PCA-based methods fail at preserving third moments. Note that the relative bias is similar for both methods for the best parameters and increases with the number of perturbed components and the size of the groups. For this reason, the proposed method now reaches its best value for $parms = 3$, as the bias is smaller. Nevertheless, it can be seen that very similar results (in fact, they are not statistically different) are obtained for any number of components larger than three.

Again, the best results are reached for small parameters for the “microaggregation plus noise” method. Its CI overlaps with those of the proposed method, meaning that the balance between data utility and risk disclosure achieved is similar for both methods.

All in all, the proposed method has led to statistically similar results to “microaggregation plus noise”. In spite of that, microaggregation techniques are usually slower than the proposed method and, thus, the proposed method is preferable as it can provide data sets with similar quality, but faster. For the computer specified above, the “microaggregation plus noise” method (using the “mdav” algorithm for the microaggregation phase), took between 5 and 30 times longer than the proposed method, depending on the data set.

### 4.2. *Hybrid Data for Partial Protection*

As stated in Section 3, the proposed method also permits perturbing some variables more than others by selecting the components more related to the target variables and just modifying them. This is an interesting feature, as it also allows preserving the whole original correlation structure.

We now show how to do it in the *Census* data set in Brand et al. (2002), which contains 1,080 records and 13 variables and has been previously used in Domingo-Ferrer and Torra (2004), Domingo-Ferrer and González-Nicolás (2010) or Jiménez et al. (2014). It is important to highlight that one of the 13 variables is a linear combination of other variables and, therefore, it is omitted from the analysis. However, after the masking is performed on the remaining variables, one can directly obtain its protected version by simply adding the corresponding protected variables.

In this example, we assume that only the first variable needs special protection. In order to proceed, we first need to obtain the weights of the first variable on the twelve components. The left plot in Figure 1 shows these weights. It can be seen that this variable has very little influence on many components, as their weights are close to zero, but has a great influence on the third component (represented by a gray circle). In fact, the weight of the third component on the first variable represents 59.91% of the total components' weights. Therefore, to protect the first variable, perturbing the third component is enough.

In order to check how the remaining variables can be affected with the perturbation of the third component, we can analyze the weights of the original variables on this component (shown in the middle plot of Figure 1). Note that the only significant weights are those associated with variables 1 and 8 (represented by a gray circle) and, therefore, only those variables will be significatively affected. Finally, looking at the weights of the eighth variable on all the components (see the right plot in Figure 1), we can observe that the third component is not the one with the highest weight and, thus, the effect of perturbing it will not be significant, as most of the information (around 87%) of this variable will be left unchanged.

As already stated, with this example we aim to show how to obtain hybrid data sets. For this reason, instead of using swapping to alter the components, we substitute them with a random draw from its smoothed empirical cumulative distribution function (see Fienberg 1994). Furthermore, we compare the results with the ones obtained using the methods proposed in Muralidhar and Sarathy (2008) and Domingo-Ferrer and González-Nicolás (2010) (as previously noted, we will refer to them as MS and MH, respectively). Both

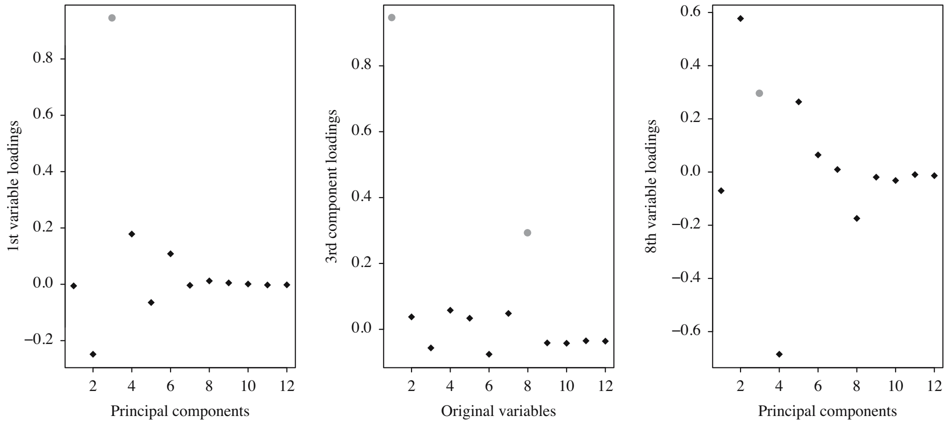

*Fig. 1. Illustrative example: representation of the weights of the first and eighth variables on all the components (left and right plots, respectively) and of the third component on all the variables (middle plot).*

methods provide very good results and require some sensitive and some nonsensitive variables. In our example, we assume that the first variable in the census data set is the sensitive one. Moreover, for the MS method we show the results associated with the following selected parameters:

- α equal to 0: In this case, the MS method equals the IPSO method in Burridge (2003).
- α equal to 0.4: In order to perform a fair comparison between the methods, we have selected this parameter as it leads to a result that takes 40% of the information from the original record, as it is the case with the PCA method, where we are only perturbing the third component and it represents almost 40% of the total variability of the first variable.
- α equal to 0.9: a case where the result is highly dependent on the original record.

Regarding MH, we show the results for group sizes of 54, 180, and 540, which correspond to 20, 6, and 2 microaggregated groups, respectively. Note that for the ease of brevity we only show a subset of the results obtained when comparing the methods.

After applying the method, we can analyze the characteristics of the resulting data set. In particular, Table 5 shows Pearson's correlation coefficients of the original and resulting variables after applying the PCA-based method. It can be seen that the coefficients are larger than 0.99 for all the variables except for the first one, which is low as a result of the masking process, and the eighth one, as we predicted. However, it is still higher than 0.9, so it has only suffered a mild perturbation. As for the MS and MH methods, the nonsensitive variables remain unchanged and Pearson's correlation coefficients of the original and resulting first variable are 0.0331, 0.4199, and 0.9033 for $\alpha$ equal to 0, 0.4, and 0.9 after applying the MS method, respectively, and 0.6528, 0.3877, and 0.0512 for $k$ equal to 54, 180, and 540, respectively.

With regard to the preservation of the original Pearson's correlation coefficients, Table 6 contains the original and resulting Pearson's correlation coefficients of the first and third variables for the proposed method (for the ease of brevity we do not show all variables, but the results are similar to those of the third variable). It can be seen that the coefficients are similar for the masked variable (the absolute difference is around four percent) and almost coincide for the remaining ones. We did not show these coefficients for the MS and MH methods, as both exactly preserve means, variances and covariances.

Furthermore, we have analyzed how close the original records are to the masked ones by means of the classic rank interval disclosure, defined in the previous subsection. We remind that it computes the proportion of records that lie in a narrow interval around its masked value. Again, we have obtained 1,000 hybrid data sets and we have computed this proportion. Table 7 shows the mean value of this index for the sensitive variable. Moreover, the centered 95% percentile bootstrap confidence interval of the proportion has been obtained.

*Table 5. Pearson's correlation coefficient between the original and the protected variables applying the PCA-based method.*

| Variable | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 0.1736 | 0.9986 | 0.9970 | 0.9968 | 0.9989 | 0.9945 | 0.9978 | 0.9181 | 0.9984 | 0.9983 | 0.9988 | 0.9988 |

*Table 6. Original and resulting Pearson's correlation coefficients of the first and third variables applying the PCA-based method.*

| Variable | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| original $V_1$ | 1.0000 | 0.0102 | 0.0490 | −0.0036 | 0.0182 | −0.0996 | 0.0038 | −0.0236 | 0.0311 | 0.0324 | 0.0386 | 0.0362 |
| resulting $V_1$ | 1.0000 | 0.0463 | 0.0869 | 0.0325 | 0.0572 | −0.0701 | 0.0364 | 0.0040 | 0.0684 | 0.0544 | 0.0713 | 0.0810 |
| original $V_3$ | 0.0490 | 0.4910 | 1.0000 | 0.3824 | 0.4952 | 0.2894 | 0.4359 | −0.0576 | 0.5597 | 0.5469 | 0.5594 | 0.5498 |
| resulting $V_3$ | 0.0869 | 0.4907 | 1.0000 | 0.3825 | 0.4946 | 0.2860 | 0.4361 | −0.0475 | 0.5580 | 0.5457 | 0.5580 | 0.5479 |

*Table 7. Disclosure risk DR (rank interval disclosure) and data utility DU (propensity score) of the output data sets.*

| | | PCA-based | IPSO | MS ($\alpha = 0.4$) | MS ($\alpha = 0.9$) | MH ($k = 54$) | MH ($k = 180$) | MH ($k = 540$) |
|---|---|---|---|---|---|---|---|---|
| DR | Mean | 0.1151 | 0.1088 | 0.1540 | 0.3597 | 0.2187 | 0.1592 | 0.1105 |
| | CI | (0.0998, 0.1304) | (0.0946, 0.1223) | (0.1380, 0.1704) | (0.3442, 0.3771) | (0.2026, 0.2349) | (0.1441, 0.1760) | (0.0976, 0.1245) |
| DU | Mean | 0.0035 | 0.0100 | 0.0089 | 0.0008 | 0.0004 | 0.0049 | 0.0092 |
| | CI | (0.0019, 0.0058) | (0.0068, 0.0137) | (0.0059, 0.0125) | (0.0003, 0.0015) | (0.0002, 0.0007) | (0.0032, 0.0066) | (0.0059, 0.0127) |

Based on the mean and CI disclosure risk values, it can be seen that the proposed method, the IPSO method (MS $\alpha = 0$ and MH $k = 1{,}080$) and MH ($k = 540$) lead to very similar levels of protection. However, this is not the case for large and small values of $\alpha$ and $k$, respectively, as their mean values are larger and their CIs do not overlap with the one associated with the PCA-based method.

Finally, although it is clear that the MS method preserves better the sufficient statistics (mean and variance-covariance matrix), the PCA-based method also provides good approximations. For this reason, it would be interesting to compare the results based on other indexes that take into account different features of the data sets, such as the propensity score. Table 7 shows the mean propensity score of 1,000 hybrid data sets, as well as its centered 95% bootstrap confidence interval.

As expected, for the MS method, the larger $\alpha$ is, the larger data utility is obtained with the MS method. It can be seen that the PCA-based method leads to more useful data sets if the parameter $\alpha$ is smaller than 0.4, as its mean propensity score is smaller than the others, and the CIs for $\alpha$ equal to 0 and 0.4 do not overlap with the PCA-based one. Similar conclusions can be drawn for the MH method: large values of $k$ lead to data sets with lower utility. Moreover, the MH method leads to worse, similar and better utility levels than the PCA-based method for $k$ equal to 540, 180, and 54, respectively.

However, in this case, we can conclude that the PCA-based method outperforms the MS and the MH methods in Muralidhar and Sarathy (2008) and Domingo-Ferrer and González-Nicolás (2010), respectively, as it provides data sets with a better balance between data utility and disclosure risk (as neither MS nor MH provide at the same time better disclosure risk and data utility than the proposed method).

It is important to highlight that the good or bad performance of the proposed method in the case of partial protection is data-dependent in the sense that depending on the correlation structure, more or fewer variables will be affected by the protection process. For example, if the confidential variables to be protected show low-medium correlation with the non confidential ones, the principal components with high weights on the confidential ones will then tend to show low weights on the nonconfidential ones and, thus, the perturbation process will affect them mildly. On the other hand, if the confidential variables are highly correlated with the nonconfidential ones, the perturbed principal components with high weights on the confidential variables will also have high weights on the nonconfidential one, and, therefore the nonconfidential variables will be highly affected. The example showed in this subsection is of the first type. Better results than in the proposed method are expected for the MS and MH methods for the second type of situation, as the nonconfidential variables are not perturbed and that leads to better overall utility (although similar levels of protection).

## 5. Final Considerations

In light of the previous examples, we can give some guidelines on how to apply the proposed method:

- The proposed method is expected to lead to better results than other proposed methods where all variables need protection. If there is only a subset of confidential variables, the performance of the method depends on the correlation structure of the confidential and nonconfidential variables. If this correlation is low, then the proposed method outperforms other methods previously proposed in the literature. Otherwise, it is preferable to resort to other methods.

- If all variables need protection, selection of the components to be perturbed depends on the desired results. If third or higher moments are not critical, then it is recommended to alter all of them, as the utility achieved is similar to that obtained with fewer components, but the record linkage gets reduced.
- On the other hand, if third or higher moments are critical, we recommend starting perturbing the first components (those with higher variance) until the desired protection level has been reached.
- If all variables need some kind of protection but we wish to perturb some more than the others, then the matrix of weights needs to be analyzed for a careful selection of the components.
- Perturbing only a few of the "last" components leads to very little protection, as its corresponding variance is very small.

The proposed method cannot guarantee that the protected variables lie in a predefined interval. For those cases, we suggest applying one of the methods proposed in Kim et al. (2015).

## 6. Conclusions and Future Work

In this article we have presented a simple and versatile method for limiting disclosure in continuous microdata based on PCA that preserves the mean vector and the variance-covariance matrix. The versatility of the method comes from the fact that it can provide masked, hybrid or fully synthetic protected data sets and it can be used to protect all or only some of the variables in a data set.

The method is very simple and, thus, does not require complex or very powerful software. We have not compared the method with more sophisticated techniques, such as multiple imputation, as we aim to provide an easy and efficient tool, in terms of good and fast results that can be widely applicable.

Some simulation studies have been performed to compare the proposed method with other well-performing techniques in terms of data utility and disclosure risk. Regarding the application of the proposed method to protect all variables at the same time, it has been shown that the PCA-based method offers a very good balance between data utility and disclosure risk and provides much better results than rank swapping and similar ones compared to "microaggregation plus noise".

As for what we call *hybrid data for partial protection*, the PCA-based method has provided better data sets than the methods proposed in Muralidhar and Sarathy (2008) and Domingo-Ferrer and González-Nicolás (2010), in the sense that, when comparing protected data sets with these three methods with similar data utility in terms of disclosure risk, the proposed method leads to safer data sets and the same happens with data sets with similar disclosure risk. As it has been already highlighted, the PCA-based method outperforms the MS and the MH method only if the confidential variables are not highly correlated with with the nonconfidential ones.

Regarding future work, the method could be extended to categorical variables by means of Categorical Principal Components Analysis. Moreover, other types of orthogonal transformations, such as the Independent Component Analysis, are to be explored in the future to check if they can lead to better results. Finally, in order to improve the usage of the method in the case of partial protection, we plan to use Factor Analysis instead of PCA, as it is possible to rotate the factors obtained, isolating the effect of the variables on the factors.

## 7. References

Banu, R. and N. Nagaveni. 2009. "Preservation of Data Privacy Using PCA Based Transformation." In *International Conference on Advances in Recent Technologies in Communication and Computing*, 439–443. Doi: http://dx.doi.org/10.1109/ARTCom.2009.159.

Brand, R. 2002. "Microdata Protection through Noise Addition." In *Inference Control in Statistical Databases*, edited by J. Domingo-Ferrer. Lecture Notes in Computer Science, 2316: 97–116. Berlin Heidelberg: Springer. Doi: http://dx.doi.org/10.1007/3-540-47804-38.

Brand, R., J. Domingo-Ferrer, and J. Mateo-Sanz. 2002. *Reference Data Sets to Test and Compare SDC Methods for Protection of Numerical Microdata*. Deliverable of European Project IST-2000-25069 CASC. Available at: http://neon.vb.cbs.nl/casc (accessed August 2016).

Burridge, J. 2003. "Information Preserving Statistical Obfuscation." *Statistics and Computing* 13: 321–327. Doi: http://dx.doi.org/10.1023/A:1025658621216.

Domingo-Ferrer, J. and U. González-Nicolás. 2010. "Hybrid Microdata Using Microaggregation." *Information Sciences* 180: 2834–2844. Doi: http://dx.doi.org/10.1016/j.ins.2010.04.005.

Domingo-Ferrer, J. and V. Torra. 2001. "A Quantitative Comparison of Disclosure Control Methods for Microdata." In *Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies*, edited by P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz. 111–133. Elsevier. Available at: https://www.iiia.csic.es/es/publications/quantitativecomparison-disclosure-control-methods-microdata (accessed August 2016).

Domingo-Ferrer, J. and V. Torra. 2004. "Disclosure Risk Assessment in Statistical Data Protection." *Journal of Computational and Applied Mathematics* 164: 285–293. Doi: http://dx.doi.org/10.1016/S0377-0427(03)00643-5.

Drechsler, J. 2011. *Synthetic datasets for statistical disclosure control: theory and implementation*, volume 201. Springer Science & Business Media.

Duncan, G. and R. Pearson. 1991. "Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future." *Statistical Science* 6: 219–239.

Efron, B. and R. Tibshirani. 1993. *An introduction to the Bootstrap*. New York: Chapman and Hall.

Fienberg, S. 1994. *A Radical Proposal for the Provision of Micro-Data Samples and the Preservation of Confidentiality*. Technical Report 611, Department of Statistics, Carnegie Mellon University.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Nordholt, K. Spicer, and P. de Wolf. 2012. *Statistical Disclosure Control*. Chichester, UK: John Wiley & Sons.

Jiménez, J., G. Navarro-Arribas, and V. Torra. 2014. “JPEG-Based Microdata Protection.” In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer. Lecture Notes in Computer Science, 8744: 117–129. Springer International Publishing. Doi: http://dx.doi.org/10.1007/978-3-319-11257-210.

Jolliffe, I. 2002. *Principal Component Analysis*. New York, USA: Springer.

Kim, H., A. Karr, and J. Reiter. 2015. “Statistical Disclosure Limitation in the Presence of Edit Rules.” *Journal of Official Statistics* 31: 121–138. Doi: http://dx.doi.org/10.1515/jos-2015-0006.

Liew, C., U. Choi, and C. Liew. 1985. “A Data Distortion by Probability Distribution.” *ACM Transactions Database Systems* 10: 395–411.

Mateo-Sanz, J., J. Domingo-Ferrer, and F. Sebé. 2005. “Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata.” *Data Mining and Knowledge Discovery* 11: 181–193. Doi: http://dx.doi.org/10.1007/s10618-005-0011-9.

Moore, R. 1996. *Controlled Data Swapping Techniques for Masking Public use Microdata Sets*. Technical report, U.S. Bureau of the Census, Washington, D.C. Available at: https://www.census.gov/srd/papers/pdf/rr96-4.pdf (accessed August 2016).

Muralidhar, K. and R. Sarathy. 2008. “Generating Sufficiency-Based Non-Synthetic Perturbed Data.” *Transactions on Data Privacy* 1: 17–33. Available: at http://www.tdp.cat/issues/tdp.a005a08.pdf (accessed August 2016).

Muralidhar, K., R. Sarathy, and J. Domingo-Ferrer. 2014. “Reverse Mapping to Preserve the Marginal Distributions of Attributes in Masked Microdata.” In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer. Lecture Notes in Computer Science, 8744: 105–116. Springer International Publishing. Doi: http://dx.doi.org/10.1007/978-3-319-11257-29.

Oganian, A. and A. Karr. 2006. “Combinations of SDC Methods for Microdata Protection.” In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer and L. Franconi. Lecture Notes in Computer Science, 4302: 102–113. Berlin Heidelberg: Springer. Doi: http://dx.doi.org/10.1007/1193024210.

Pagliuca, D. and G. Seri. 1999. *Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey*. Esprit SDC Project, Deliverable MI-3/D2.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Website: http://www.R-project.org/.

Raghunathan, T.E., J. Reiter, and D. Rubin. 2003. “Multiple Imputation for Statistical Disclosure Limitation.” *Journal of Official Statistics* 19: 1–16.

Rubin, D. 1993. “Statistical Disclosure Limitation.” *Journal of Official Statistics* 9: 461–468.

Sarathy, R. and M. Krishnamurty. 2002. “The Security of Confidential Numerical Data in Databases.” *Information Systems Research* 13: 389–403. Doi: http://dx.doi.org/10.1287/isre.13.4.389.74.

Templ, M. 2008. “Statistical Disclosure Control for Microdata Using the Rpackage sdcMicro.” *Transactions on Data Privacy* 1: 67–85. Doi: http://dx.doi.org/10.18637/jss.v067.i04.

Woo, M., J. Reiter, A. Oganian, and A. Karr. 2009. "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation." *Journal of Privacy and Confidentiality* 1: 111–124.

Received September 2015
Revised August 2016
Accepted August 2016