Factor Analysis for Anonymization

Aida Calviño
Dept. of Statistics and O.R. III
Complutense University of Madrid
Madrid, Spain
aida.calvino@ucm.es

Palmira Aldeguer
Dept. of Statistics and O.R. III
Complutense University of Madrid
Madrid, Spain
paldegue@ucm.es

Josep Domingo-Ferrer
Dept. of Comp. Engineering and Maths
Universitat Rovira i Virgili
Tarragona, Catalonia, Spain
josep.domingo@urv.cat

***Abstract*—In this paper we propose a new method to anonymize (share relevant and detailed information while *not naming names*) and protect data sets (minimize the utility loss) based on *Factor Analysis*. The method basically consists of obtaining the factors, which are uncorrelated, protecting them and undoing the transformation in order to get interpretable protected variables.**

We first show how to proceed when all variables in the data set need protection and, then, we focus on the case where only a subset of variables has to be protected. Finally, we perform a simulation study to compare the proposed method with two alternative techniques: Microaggregation plus noise addition (which has been recognized as a very powerful method) and one anonymization method recently proposed based on Principal Components Analysis.

***Index Terms*—Statistical Disclosure Control, Privacy Protection, Dimensionality Reduction techniques.**

I. Introduction

Society's privacy requirements are increasing in recent years and, thus, new laws and agreements are developed to assure data protection of individuals (a good example of this kind is the *EU-US Privacy Shield Framework* [1]). However, there is also a movement regarding sharing publicly data (e.g., Open Data portals [2]). Therefore, the conundrum is how to accomplish both objectives at once (publish data while keeping privacy rights), as, the more one shares, the less privacy remains. Regarding this matter, anonymization techniques become the perfect mechanism to ensure the balance between these goals.

Anonymization techniques permit avoiding the risk of (direct or indirect) identification and disclosure of sensitive information. They are commonly used to publish data so as to make it impossible to an observer to get new knowledge on an individual. They are also referred to as Statistical Disclosure Control (SDC) methods as they are based on the modification of the initial data or creation of a new data set that preserves most of the statistical properties of the original.

SDC techniques are generally applied to create new data that resembles the original one, but minimizing the risk of disclosure of sensitive information or re-identification risk. Statistical algorithms regarding this problem are based on the idea of maximizing the utility of a data set but minimizing the risk of revealing sensitive information. Therefore, a key point of SDC is the definition of utility and disclosure risk.

Several methods have been proposed in the literature (see [12] or [20] for a summary). Some of them are based on the alteration of the structure while a second alternative is to only alter the content, that is, to modify the values rather than the structure. Examples of the first type are: (i) selecting a subsample of the initial data (see [10]); (ii) deleting extreme observations, (iii) creating summary variables, etc. Some examples of the second kind include: adding noise to values (see [4]); swapping values' observations with or without rank restrictions (see [16]); or microaggregation (see [7]).

According to [17], a very interesting possibility is to jointly apply some of the techniques to create a combination that improves protection and utility by making the most of the advantages of the combined techniques. A good example of this type, as acknowledged by the authors, is microaggregation plus noise addition.

Despite the many alternatives to protect data that have been proposed in the literature, none of them is *perfect* for every possible scenario as the result may vary upon different circumstances. Besides, it is very difficult to previously determine the best method for a specific situation because it depends on the original data (see [3] or [8] for references on this issue), as well as on the intended use of the protected data.

In this article we use two benchmark methods to compare our results with: anonymization based on Principal Analysis Components (PCA) [6] and microaggregation plus noise addition (M-AGG) [17]. The first benchmark method is selected because of its similarity to the proposed method (PCA and FA are both used for dimensionality reduction). Regarding the second benchmark method, it was previously used for comparison with PCA in [6] and it has been selected for continuity reasons.

The first alternative we are considering in this paper is based on Principal Components and was proposed by [6]. The authors propose to alter the principal components instead of the original variables so that it becomes possible to protect different variables at once by just changing a component (as they represent several variables with different weights). Hence, one can maximize the effectiveness of a change by modifying the component that represents a large number of variables or the maximum variance. As a consequence, the alterations can be done univariately, i.e., considering each

This work was partly supported by the Spanish Government through project TIN2014-57364-C2-R "SmartGlacis" and by the European Commission through project H2020-644024 "CLARUS".

variable separately, so one can apply a broad range of methods to modify the components since univariate methods are simpler and more efficient than their multivariate versions. According to [6] and [13], an algorithm to alter components that leads to competitive results is swapping (altering the order of elements without position restrictions). Finally, note that this benchmark method, as well as the method proposed in this paper, falls into the broader category of *spectral anonymization* defined by [13].

The second method, microaggregation plus noise addition, was proposed by [17] and has been recognized as a very powerful method (see [21]). The authors propose to aggregate close observations and add multivariate random noise with the same correlation matrix as the loss of variance given by the microaggregation phase to recover the lost variance. Better results are obtained with small groups of observations.

In this paper, we propose a new method to anonymize data based on Factor Analysis. Basically, the method consists of obtaining the factors of the data set, anonymizing them, instead of the original variables, and undoing the transformation in order to get interpretable protected variables. The reason is twofold: a) it is possible to protect several variables at the same time by just protecting one factor; and b) the factors are uncorrelated and, thus, univariate anonymization methods can be used, while still preserving the correlation structure of the original variables.

Additionally, we compare the proposed method with the ones mentioned above and explore the scenarios where the proposal outperforms the other alternatives.

To our knowledge, the main original contributions of this paper are the following:

1) An anonymization method based on factor analysis is proposed for the first time.
2) Two different simulation studies are carried out to determine the best scenarios to apply the proposed method. The first one is referred to the general protection case, where all variables require protection; whereas the second one evaluates the selective protection one, where only a subset of variables (the sensitive ones) require protection.
3) A new measure to evaluate the goodness of a protection method in the selective protection scenario is proposed.

A. Measures to evaluate anonymization methods

To compare the effectiveness of an algorithm, measures have to be defined in order to evaluate them homogeneously. In anonymization, two opposed criteria arise to measure the results: disclosure risk and data utility. Several measures have been proposed in the literature to evaluate both aspects. We now briefly describe the ones that are used in this paper:

- Security/Disclosure risk: It evaluates the level of protection of the data. From a security point of view, the best outcome is obtained when there is no similarity between original and modified data sets. We make use of the measures defined in [9].
  - Distance based-record linkage (DBRL): Computation of the proportion of original observations whose closest protected observation coincides with its protected version.
  - Interval disclosure (ID): Computation of the proportion of original observations that lie in a narrow interval (obtained by means of the closest observations) defined around their protected versions.
- Data utility: Maintenance of the properties of the original data set and, therefore, its utility. Maximum utility is achieved when no modification is done to the original data set. We make use of the following measures:
  - Probabilistic Information Loss (PIL): It evaluates how much information one loses when using the protected data set instead of the original version, as defined in [15]. More precisely, it quantifies the difference between several statistical properties. In this paper, the following are selected: mean, variance, covariance, Pearson correlation, quantiles, kurtosis and skewness. Kurtosis and skewness were not included in [15] so we have extended the *pil* computation to consider them by means of the generalization of moments in [19].
  - Propensity scores (PS): The idea is to measure the ease/difficulty of discriminating between observations from the original and protected data sets based on their values through logistic regressions, as proposed by [21]. If it is not easy, then both data sets are similar and, thus, the utility of the protected version is high.

All the previous measures lie in the range $[0-1]$, where the lowest value represents the best outcome for each measure (maximum security or utility); thus, the objective is to minimize their value when obtaining protected data sets[1].

In order to consider the different aspects of the previous measures at the same time, we propose to aggregate them by means of the following formula:

$$summary = \frac{PIL + PS + DBRL + ID}{4}. \qquad (1)$$

All measures above are general-purpose, as they evaluate different aspects from a generic point of view. However, as previously discussed, for specific purposes new metrics can be defined or different weights can be given to the measures averaged in Expression (1) when security is more important than utility or *vice versa*.

The paper is organized as follows. In Section II, we provide background on factor analysis and explain the proposed anonymization method. Section III is devoted to the comparison of the three aforementioned methods in the general protection scenario, while Section IV deals with comparison in the selective protection scenario. Finally, in Section V we give some conclusions and future research lines.

[1] It is worth noting that applying no change results in a *summary* equal to 0.5 (total data utility but no protection) and a new data set with no relation with the original one (except its size) leads to the same value (no utility but total security).

II. Factor Analysis as an anonymization tool

A. Definition of factor analysis

Factor analysis (FA) is a statistical method proposed in the early 1900's by Charles Spearman. It is based on the assumption that observed variables measured on the same observations are generated by a smaller set of unobservable (and uncorrelated) variables (called factors). For example, let us assume that different tests are given to students regarding different skills, such as English, Mathematics, Arts, Physics, etc. In that case, one can assume (as Mr. Spearman did) that the students' grades are due to different types of intelligence that cannot be measured directly. The aim of FA is to find these factors (the types of intelligence in the example) assuming, as well, that, apart from the factors, there are some components that cannot be explained by them (called residuals). For more details on factor analysis, please refer to [14].

As a consequence, each standardized variable in a data set ($\boldsymbol{X}$) can be obtained by means of linear combinations of the k factors ($\boldsymbol{F}$) plus some residuals ($\boldsymbol{\varepsilon}$):

$$\boldsymbol{X}_{(m\times n)} = \boldsymbol{L}_{(m\times k)}\boldsymbol{F}_{(k\times n)} + \boldsymbol{\varepsilon}_{(m\times n)}, \tag{2}$$

where $\boldsymbol{L}$ is the loading matrix that contains the parameters of the linear combinations. Note that each column of $\boldsymbol{X}$, $\boldsymbol{F}$ and $\boldsymbol{\varepsilon}$ denotes values for one particular observation and matrix $\boldsymbol{L}$ does not vary across observations.

The aim of FA is to find matrices $\boldsymbol{L}, \boldsymbol{F}$ and $\boldsymbol{\varepsilon}$ such that: i) factors are uncorrelated, ii) factors and residuals are independent, and iii) the number of factors is smaller than or equal to the number of variables and needs to be fixed in advance.

Given the previous assumptions, the values in matrix $\boldsymbol{L}$ are the correlations[2] between the corresponding original variables and factors:

$$Cov(\boldsymbol{X},\boldsymbol{F}) = E(\boldsymbol{X}\boldsymbol{F}') = E((\boldsymbol{L}\boldsymbol{F}+\boldsymbol{\varepsilon})\boldsymbol{F}') = E(\boldsymbol{L}\boldsymbol{F}\boldsymbol{F}'+\boldsymbol{\varepsilon}\boldsymbol{F}) = \boldsymbol{L}$$

An interesting feature of factor analysis is that its solutions are not unique. If $\boldsymbol{T}_{(k\times k)}$ is an orthogonal matrix (i.e., $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{I}$) then Expression (2) can be rewritten as:

$$\boldsymbol{X} = \boldsymbol{L}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{F} + \boldsymbol{\varepsilon},$$

leading to a set of new factors $\boldsymbol{T}'\boldsymbol{F}$ whose loading matrix is given by $\boldsymbol{L}\boldsymbol{T}$. Therefore, given a solution to FA, it is possible to find new ones by means of an orthogonal matrix. This feature of factor analysis is commonly referred to as *rotation*.

Rotation aims at reducing the number of factors with which the variables are highly correlated, taking into account that the values of the loading matrix represent the correlation between original variables and factors and, thus, their square is the Pearson correlation index (that lies between 0 and 1). The perfect scenario would be one where factors and variables have correlations indices of 0 and 1 only. In that case, the variability of the values in the loading matrix would be maximum (as opposed to the case where all correlations are 0.5). Several types of rotation have been proposed in the literature but the most frequent one is *varimax*, where the orthogonal matrix $\boldsymbol{T}$ that maximizes the variability of the loading values is selected.

B. Proposed anonymization method

FA is interesting for data protection because it allows protecting the factors instead of the original variables, in such a way that one can resort to simple univariate modifications (because factors are orthogonal) affecting several original variables at the same time without perturbing the relations among them. Therefore, if one modifies factors by using a method that does not alter means and variances of original factors, the mean vector and the variance-covariance matrix of the original data are preserved in the protected version (the formal proof of this property can be found in [6]).

Moreover, rotation is ideal for selective protection, that is, when only a subset of variables need protection, as one can rotate the factors in order to capture sensitive variables and protect only those factors associated with them (more details are given below in Section IV).

On a technical note, the algorithm associated with the proposed anonymization method is as follows:

1) Take the original data set and standardize it.
2) Obtain the factors and residuals given by a FA[3].
3) Protect the factors by means of a univariate method (in our case, swapping for the sake of concreteness, although other methods can be used).
4) Calculate the anonymized observations using Expression (2) with the protected factors, instead of the original ones.

Figure 1 shows the previous steps on an illustrative simulated data example, where we have used a correlation level 0.8. Note that factors have been swapped (without rank restrictions) and means are totally preserved.

A similar procedure has been recently proposed by [6] using Principal Components Analysis (PCA), instead of factor analysis. The main advantage of using FA instead of PCA is the fact that factors can be rotated allowing for a stronger or milder modification of the variables involved in each factor. Additionally, residuals can be modified (or not) independently of the factors, which gives more flexibility to the protection process[4].

As previously expounded, correlation among variables determines the quantity of information kept by the factors and the residuals (take into account that factors contain the *common* information of the variables, while the residuals contain the *specificities*). As long as only factors are modified[5], the more information represented by them, the better the potential change of the data. Thus, for highly correlated variables, as it

[2] As already mentioned, a common practice in FA is to standardize the variables previous to its application, which results in equal covariance and correlation marices.

[3] Without loss of generality, in the sequel, we take the same number of factors and original variables in order to minimize the quantity of information contained in the residuals.

[4] Note that this is possible due to the fact that factors and residuals are independent and, thus, can be modified separately.

[5] Residuals can also be modified but, as they are correlated, multivariate perturbative methods that preserve the correlation structure are required. For that reason, without loss of generality, in the sequel we assume that only factors are modified.

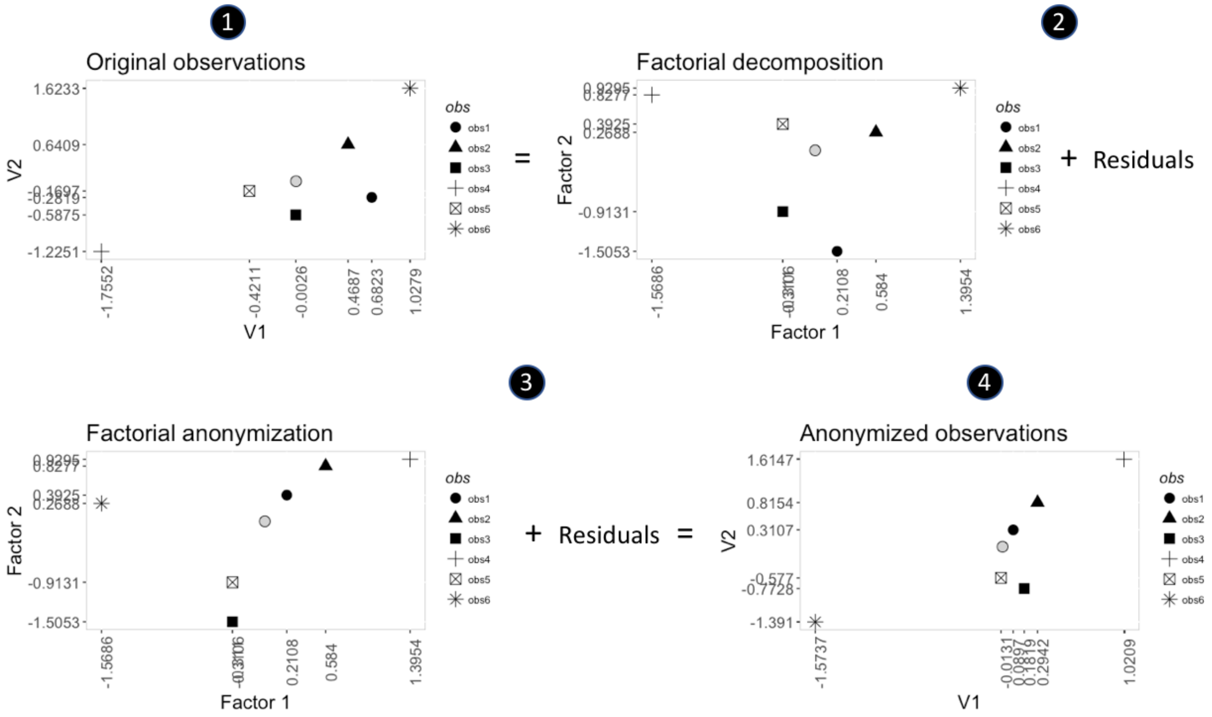

Fig. 1. Process to apply FA to anonymize a data set. Grey dots represent the mean of the values in each graphic.

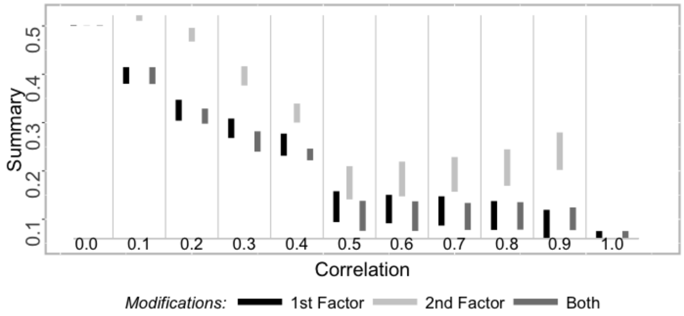

Fig. 2. Performance summary based on Expression (1) when applying the FA method to a two-variable data set, as a function of the correlation between the two variables. 95% bootstrap confidence intervals for the summary are given, and three different situations are considered: (i) anonymization of the first factor only; (ii) of the second factor only; (iii) of both factors. Lower values of the summary indicate better performance.

will be later shown, altering factors is more effective in terms of changes, as the residuals contain little information.

Figure 2 shows the performance of the method, in terms of the measure defined in Expression (1), when anonymizing a two-variable data set, for different correlation levels between the two variables. Note that, for each anonymization scenario, we are representing the 95% bootstrap confidence interval of the *summary* in order to show the mean performance, as well as its variance[6]. Moreover, we evaluate the effect of modifying only the first factor (in black), which contains the largest quantity of information, only the second (in light gray) or both (in dark gray).

As it can be seen, better results are attained for high correlation levels. Moreover, modifying only the second factor leads to worse results, as much of the original information remains unchanged. Note that when the variables are uncorrelated (which is extremely rare in reality), factors contain no information and altering them equals no protection. In that case, one should apply a univariate method to each original variable independently. Finally, note that, in general, narrower confidence intervals are obtained when protecting both factors.

According to the previous figure, *it is better to only modify the factors that represent more variance of a data set*. For the sake of brevity and clarity, we have represented a very simple illustrative example, but the previous statement holds for larger data sets[7].

III. General protection

As previously stated, in this paper we compare the proposed method with two alternative ones under two scenarios: general and selective protection. The first scenario refers to the case where all variables in the data set need protection, whereas the second scenario focuses on protecting only a subset of them. In this section, we analyze the first case.

Classical factor and principal components analyses look at the correlation structure of a data set. For that reason,

[6] Note that Figures 3 and 5 in this paper also show 95% bootstrap confidence intervals for the corresponding measures.

[7] Note that the authors of [13] stated that "*larger dimensions are worth more attention than smaller ones*" (p. 444).

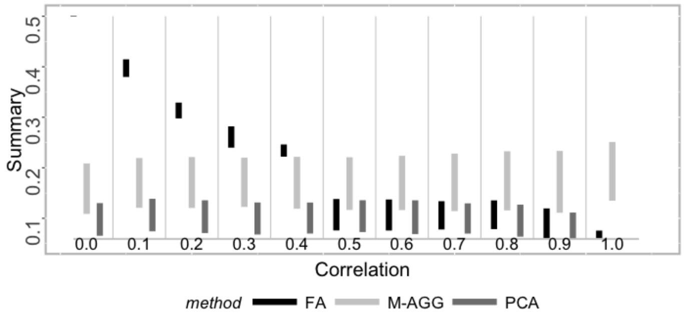

Fig. 3. Performance summary of different methods on a two-variable data set, based on Expression (1), as a function of the correlation between the two variables. 95% bootstrap confidence intervals for the summary are given, and three different anonymization methods are considered: (i) the proposed FA anonymization; (ii) microaggregation with noise addition (M-AGG); (iii) PCA anonymization. Lower values of the summary indicate better performance.

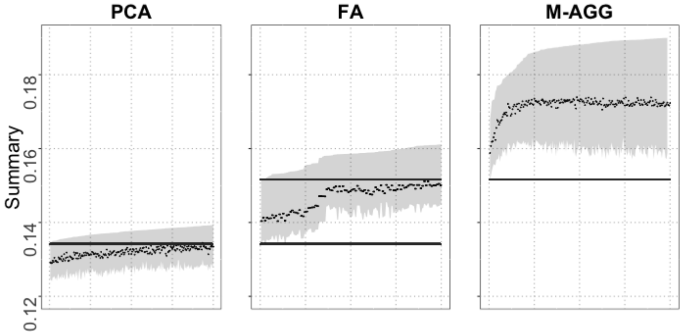

Fig. 4. *Overall* comparison among PCA, FA and M-AGG anonymization methods. 200-best results are shown for each method in ascending order by 97.5th quantile. Black horizontal lines show the 97.5th quantile value of the best PCA, the 2.5th and 97.5th quantiles of the best FA configuration and the 2.5th quantile of the best M-AGG configuration. Black dots represent the median value.

the correlation structure will affect the performance of the anonymization process, as previously shown in Figure 2.

Figure 3 compares, in terms of the measure defined in Expression (1), the performance when anonymizing a two-variable data set (for different correlation levels between the two variables) of the proposed method (FA), microaggregation plus noise (M-AGG) and anonymization based on PCA (PCA). As it can be seen, microaggregation and PCA are not heavily affected by correlation levels. However, this is not the case for FA, where better results are obtained for larger correlation levels.

Moreover, as it can be seen in the figure, FA leads to better and similar results than M-AGG and PCA, respectively, when the correlation index is larger than 0.5. Note that confidence intervals are narrower for FA and PCA than for M-AGG. Therefore, correlation among variables in a data set is a very important feature to be taken into account when selecting the anonymizing method to apply.

A. General Protection for the Tarragona data set

In this section we compare the three methods above on a real data set that is commonly used to test protection methods in the literature, i.e., the Tarragona data set in [5]. The factors of this data set represent 89% of its variance, so the Tarragona data set is a good example to test the method because it allows wider changes when modifying the factors[8].

Figure 4 shows the performance of the three methods for the Tarragona data set. In particular, we show the 200 best results of each method (without paying attention to the parameters that generated them) sorted in ascending order of the 97.5th quantile. When comparing FA with PCA and M-AGG in this highly correlated scenario, Figure 4 reveals that PCA is the best by a narrow margin, followed by FA and finally M-AGG, consistently with the findings in previous sections.

Moreover, the black horizontal lines in Figure 4 show that the 97.5th quantile value of the best PCA configuration is lower than the 2.5th quantile of the best FA configuration (a similar comparison can be made for FA and M-AGG). Therefore, as intervals do not overlap (and given the relation between confidence intervals and hypothesis tests), we can assert that some PCA configurations exist that are better than any FA possibility and that there are some FA combinations that are better than any M-AGG realization. Thus, there exists a strict order of methods regarding their performance on this data set: PCA, FA, microaggregation.

IV. Selective protection

Section III deals with the protection of a complete data set without paying special attention to the most sensitive variables. However, there are some scenarios where only a subset of variables needs protection and the remaining ones should stay intact. In this section, we focus on this latter case and test if PCA is still the best alternative for the selective protection scenario.

In this case, the metrics defined above are still useful but it is necessary to resort to new ones to take into account the fact that non-sensitive (public) variables should stay unaltered. This means that one has to select a method that maximizes the protection of sensitive variables and minimizes the alteration of the public ones, while preserving the statistical characteristics of the initial data set.

In order to evaluate the alteration level of public variables, we define a new measure, that we call *selectivity*, based on Pearson correlation indices in order to evaluate how close each variable in the new data set is to the original one[9]. The objective is to get results close to 0 for sensitive variables (no similarity between them) and results close to 1 for the

[8] For this comparison we have carried out a simulation study with the following characteristics: (i) 100 simulations for each algorithm and combination of parameters, (ii) the Euclidean distance (mdav) is used for M-AGG, (iii) all combinations of factors and components are tested for FA and PCA, while for M-AGG, the number of observations aggregated together is limited to be less than half of the number of observations. Other data sets also in [5], such as Census and EIA, have been tested in the simulation study but will be skipped here for the sake of brevity.

[9] As far as we know, no such measure has been previously proposed in the literature.

public variables (complete correlation, i.e., low level of modification)[10].

In order to be able to compute a general measure for a data set, we define an index $\boldsymbol{IC}$ in Expression (3) that represents the individual correlation index for each variable between its original and protected versions:

$$\boldsymbol{IC}(x) = \begin{cases} |Cor(x, x')|, & \text{if } \; x \text{ requires protection} \\ 1 - |Cor(x, x')|, & \text{otherwise,} \end{cases} \tag{3}$$

where x and x' represent the original and protected version of a variable.

Then, the *Selectivity* metric in Expression (4) is given by the arithmetic mean of the $\boldsymbol{IC}$ indices of all the variables in the data set.

In summary, to evaluate selective protection methods, we make use of the following sets of metrics:

- *Security*: Based on DBRL and interval disclosure (ID), as previously defined.
- *Utility*: Based on PIL and propensity score (PS), as previously defined.
- *Selectivity*: Based on the proposed metric, to evaluate to what extent the public variables are kept unaltered.

Again, we propose to aggregate the metrics in order to obtain a single measure to evaluate the methods:

$$overall = \frac{PIL + PS + DBRL + ID + 2 * Selectivity}{6}, \tag{4}$$

where we have attached a weight 2 to the *Selectivity* metric in order to assign the same weight to the three sets of metrics.

We next explain how the methods studied in this paper can be adjusted for the previously defined selective protection scenario.

1) PCA: When using PCA for selective protection, one needs to carefully decide which components must be modified. The objective is to select components that maximize the level of modification of sensitive variables while minimizing the modification of public variables. We propose to determine whether a component has to be changed or not based on a threshold to measure if the component relies on sensitive variables or not. The algorithm is as follows:

i) Select first sensitive variable to be protected.
ii) Sort components in decreasing order based on the loadings for that variable.
iii) Select a component if the loading for that variable reaches the threshold.
iv) Repeat steps (ii) and (iii) with the remaining variables to be protected until all of them have been evaluated.

In general, PCA protection methods behave worse for selective protection than for general protection. This is due to the fact that it is not possible to completely isolate sensitive and public variables in two groups using their principal components. Based on this behavior, when protecting highly correlated variables, changes will be less *selective* (public variables will also be highly modified) and, hence, *overall* results will not be as good as for the general protection scenario.

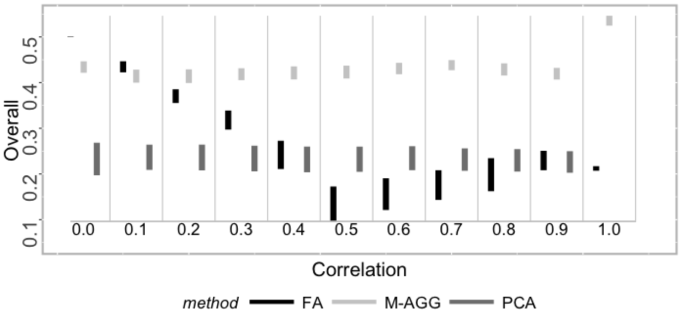

Fig. 5. Performance summary of different methods on a two-variable data set, based on Expression (4), as a function of the correlation between the two variables. 95% bootstrap confidence intervals for the summary are given, and three different anonymization methods are considered: (i) the proposed FA anonymization; (ii) microaggregation with noise addition (M-AGG); (iii) PCA anonymization. Lower values of the summary indicate better performance.

2) Microaggregation Plus Noise: When using microaggregation plus noise for selective protection, the key difference is that changes are only applied to sensitive variables. The only parameter to be set is k (number of elements included in a group) which, according to the previous sections and the literature, leads to better results for low values (in the following, we restrict the value of k to the range [2-5]). However, even for protected variables, changes are not very aggressive (because of the small k), so original and protected values will be very similar (thus, highly correlated).

3) Factor Analysis: When using Factor Analysis for selective protection, there are two different choices to be made: (i) how to rotate the factors and, (ii) which ones to modify. The proposed strategy is as follows:

- Rotation: As previously explained, factors can be rotated without altering the result (the rotation should be done before the protection of the factors (between Steps 2 and 3 in Figure 1)). The rotation algorithm is designed in order to accomplish one of the following goals[11]:
  - group sensitive variables in few factors in order to maximize changes ;
  - group public variables in few factors in order to minimize changes.

  The procedure to group sensitive (or public [12]) variables follows the steps:

  1) select the first sensitive variable and rotate all factors until only the first factor represents it;

[10] Since Pearson correlations lie in the range $[-1, 1]$, we use instead their absolute value to create a symmetric function around 0. The rationale for choosing absolute values rather than squares is to avoid concentration near zero produced by squaring the values.

[11] The usual rotation methods, such as the *Varimax* previously explained, were developed in order to group variables (without restrictions) and, thus, are not suitable from an anonymization point of view where one should group similar variables (from a protection perspective) together.

[12] For public variables, the process is analogous to the one proposed for sensitive variables, but selecting public variables instead of the sensitive ones.

2) take the second variable and rotate all remaining factors until only the first and second factor represent that variable;
3) continue until all sensitive variables are processed.

- Factor selection: Different criteria can be used to select factors:
  - modify factors whose loadings with sensitive variables are over a threshold;
  - do not modify factors whose loadings with public variables are over a threshold;
  - modify only those factors whose loadings are higher for sensitive variables than for public ones.

As it can be seen, the proposed method is more flexible when dealing with selective protection, as there are more configuration possibilities than for PCA or M-AGG. Therefore, it is possible to adjust its parameters to optimize different scenarios, leading, thus, to better results.

A. *Comparison*

As in the general scenario, the correlation structure of the data set influences the performance of the anonymization methods for selective protection. Figure 5 shows the *overall* performance of the three protection methods over a two-variable data set with different correlation levels when the objective is to protect only one variable. As it can be seen, correlation levels do not affect M-AGG or PCA in this case either. However, unlike in the general protection case, it can be seen that the FA method outperforms PCA from medium to high correlation levels.

When extrapolating to larger data sets, as more variables get involved, we need to define more concepts, that are required to analyze the correlation structure of the data set to be protected:

- Within-correlation ($\boldsymbol{w}$): the correlation among public variables.
- Between-correlation ($\boldsymbol{b}$): the correlation between the sensitive and public variables.

In order to be able to compare the method in a more realistic case, we have carried out a simulation study, whose results are shown in Figure 6. Without loss of generality, and for the sake of conciseness, we assume that only one variable out of N requires protection. The correlation matrix is as follows:

$$Cor(dataset) = \begin{pmatrix} 1 & \boldsymbol{b} & \boldsymbol{b} & \dots & \boldsymbol{b} \\ \boldsymbol{b} & 1 & \boldsymbol{w} & \dots & \boldsymbol{w} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{b} & \boldsymbol{w} & \boldsymbol{w} & \dots & 1 \end{pmatrix} \tag{5}$$

Figure 6 shows the *overall* mean performance (represented in gray scale) of the three methods for different values of $\boldsymbol{w}$ and $\boldsymbol{b}$ in Expression (5). As it can be seen, better results are obtained for PCA when the data set has high within-correlation levels and low between-correlation levels. In this case, it is easy to isolate sensitive and public variables in different components and, thus, PCA leads to the best possible results. Better results for M-AGG are also achieved in that scenario because low correlations between sensitive and public variables are easier to preserve.

Finally, the best results for FA are obtained when public variables are poorly correlated, as they will be represented basically in the residuals (and these remain unmodified). Thus, the ideal scenario for FA is: significant between-correlation levels, in order to have a good representation of sensitive variables in the factors, but low within-correlation levels, so factors would represent more variance from sensitive variables than from public ones (see Figure 6).

B. *Selective Protection for the Tarragona Data set*

The correlation matrix in Expression (5) is very unlikely to happen in real data sets. Therefore, we need to define new measures to evaluate between- and within-correlations:

$$Between(p) = \tfrac{1}{N-1} \textstyle\sum_{\substack{i=1\\ i\neq p}}^{N} |\boldsymbol{Cor}(x_p, x_i)|, \tag{6}$$

$$Within(-p) = \tfrac{2}{(N-2)(N-1)} \textstyle\sum_{\substack{i=1\\ i\neq p}}^{N} \sum_{\substack{j>i\\ j\neq p}}^{N} |\boldsymbol{Cor}(x_i, x_j)|, \tag{7}$$

where p refers to the variables that require protection and $-p$ to the public variables[13].

Using the above formulas, Figure 6 also shows, above the three grids, results for the Tarragona data set, where only one variable is selected each time[14]. Shapes represent which is the best algorithm for each protected variable.

The four points (**x**) further to the right correspond to protected variables with large between-correlation levels; while circular points (**o**) on the left are from lower between-correlation levels. This result is consistent with the above conclusions and simulations with synthetic data, where M-AGG showed worse results for large between-correlations.

V. Conclusions and Future Work

In this paper we have presented a new method to anonymize continuous variables based on factor analysis, which basically consists of obtaining the latent factors of a data set and anonymizing them instead of the original variables. Moreover, we have compared our proposal with two competitive alternative methods.

The advantage of our method is twofold: it is possible to make use of any univariate anonymization method and several variables can be protected by means of the modification of only one factor. Furthermore, as long as the anonymization methods used to protect the factors preserve their means and variances, the mean vector and the variance-covariance matrix of the original data set are also preserved.

In summary, the best algorithm to protect a complete data set is PCA (see [6]), but this method is not adequate when protecting selectively because it is impossible to guarantee full isolation of sensitive variables. In selective scenarios, M-AGG is very *selective* but lacks effectiveness (both regarding security and utility). Therefore, FA is a good alternative due

[13] Note that, using these formulas, the correlation matrix in Expression (5) gives a value of $\boldsymbol{b}$ and $\boldsymbol{w}$ for between- and within-correlations, respectively.

[14] It is worth noting that within-correlations are quite stable as they evaluate the "general" correlation and, thus, the effect of a single variable is negligible.

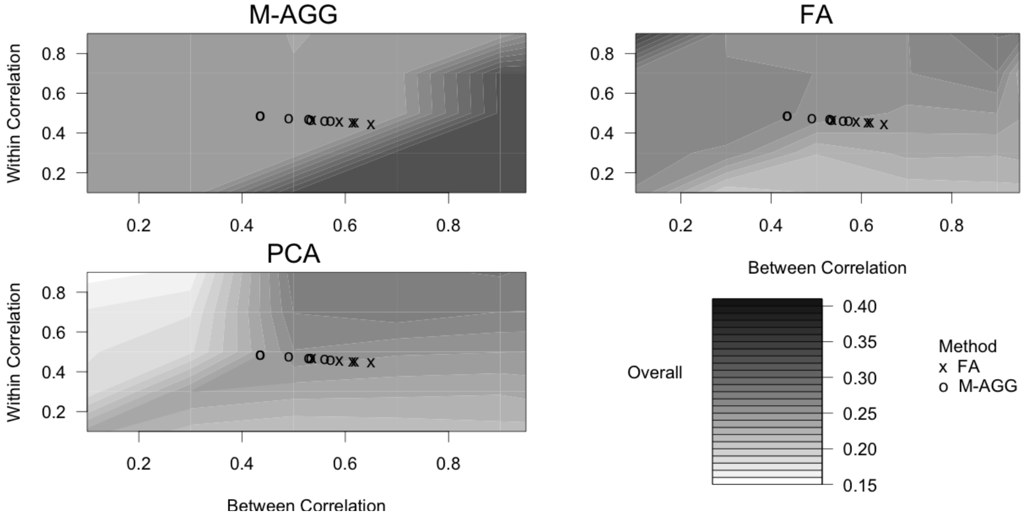

Fig. 6. Correlation effect for each method. The gray scale compares the *overall* measure among methods. Symbols in black represent protection of the Tarragona data set.

to its sharp modifications of sensitive variables and it offers better summary results (security and utility) than M-AGG in some scenarios.

A cookbook to recommend the algorithm to be selected for anonymization would be:

- *PCA: full protection or high within-correlation and low between-correlation.*
- *FA: low within-correlation and significant between-correlation.*
- *Microaggregation: low between-correlation.*

Regarding future work, some questions arise in this paper for further analysis:

- find an accurate index to evaluate which method to apply.
- conduct a general analysis on how to proceed when more than one variable requires protection.
- provide a formal security analysis, in the style of privacy models such as *k-anonymity* [18] or *differential privacy* [11] models.

References

[1] EU-US Privacy Shield Framework. https://www.privacyshield.gov/.

[2] US Governments Open Data Portal. https://www.data.gov.

[3] Opinion 05/2014 on anonymisation techniques. Technical Report 0829/14/EN WP216, The Working Party On The Protection Of Individuals With Regard To The Processing Of Personal Data, 2014. http://ec.europa.eu/.

[4] R. Brand. Microdata protection through noise addition. In Josep Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 97–116. Springer Berlin Heidelberg, 2002.

[5] R. Brand, J. Domingo-Ferrer, and J.M. Mateo-Sanz. Reference data sets to test and compare SDC methods for protection of numerical microdata. 2002. Deliverable of European Project IST-2000-25069 CASC, available at http://neon.vb.cbs.nl/casc/CASCrefmicrodata.pdf.

[6] A. Calviño. A simple method for limiting disclosure in continuous microdata based on principal component analysis. *Journal of Official Statistics*, 33(1):15–41, 2017.

[7] J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. on Knowl. and Data Eng.*, 14(1):189–201, 2002.

[8] J. Domingo-Ferrer and V. Torra. A quantitative comparison of disclosure control methods for microdata. In *Confidentiality, disclosure, and data access : Theory and practical applications for statistical agencies*, pages 111–133. Elsevier, 2001.

[9] J. Domingo-Ferrer and V. Torra. Disclosure risk assessment in statistical data protection. *Journal of Computational and Applied Mathematics*, 164:285 – 293, 2004.

[10] J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.

[11] C. Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, pages 1–12. Springer Berlin Heidelberg, 2006.

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

[13] T.A. Lasko and S.A. Vinterbo. Spectral anonymization of data. *IEEE Trans. on Knowl. and Data Eng.*, 22(3):437–446, 2010.

[14] D. N. Lawley and A. E. Maxwell. Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D*, 12(3):209–229, 1962.

[15] J.M. Mateo-Sanz, J. Domingo-Ferrer, and F. Sebé. Probabilistic information loss measures in confidentiality protection of continuous microdata. *Data Mining and Knowledge Discovery*, 11(2):181–193, 2005.

[16] R. Moore. Controlled data swapping techniques for masking public use microdata sets. Technical report, U.S. Bureau of the Census, Washington, D. C., 1996. Available at https://www.census.gov/srd/papers/pdf/rr96-4.pdf.

[17] A. Oganian and A.F. Karr. Combinations of SDC methods for microdata protection. In Josep Domingo-Ferrer and Luisa Franconi, editors, *Privacy in Statistical Databases*, volume 4302 of *Lecture Notes in Computer Science*, pages 102–113. Springer Berlin Heidelberg, 2006.

[18] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, 1998.

[19] A. Stuart and O. Keith. *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*. Wiley, 1994.

[20] M. Templ, B. Meindl, and A. Kowarik. Introduction to statistical disclosure control (SDC). Technical report, data-analysis OG, 2017. Available at https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf.

[21] M.J. Woo, J.P. Reiter, A. Oganian, and A.F. Karr. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1:111–124, 2009.