A Conceptual Framework for Assessing Anonymization-Utility Trade-Offs Based on Principal Component Analysis

Giuseppe D'Acquisto [*], Maurizio Naldi [*1]

[*] *Dept. of Civil Engineering and Computer Science*
University of Rome Tor Vergata.
Via del Politecnico 1, 00133 Rome, Italy
dacquisto@ing.uniroma2.it, maurizio.naldi@uniroma2.it
[1] *Dept. of Law, Economics, Politics and Modern languages*
LUMSA University, Via Marcantonio Colonna 19, 00192 Rome, Italy
m.naldi@lumsa.it

***Abstract* - An anonymization technique for databases is proposed that employs Principal Component Analysis. The technique aims at releasing the least possible amount of information, while preserving the utility of the data released in response to queries. The general scheme is described, and alternative metrics are proposed to assess utility, based respectively on matrix norms; correlation coefficients; divergence measures, and quality indices of database images. This approach allows to properly measure the utility of output data and incorporate that measure in the anonymization method.**

Keywords - Anonymization; Privacy; Principal Component Analysis; Databases

I. INTRODUCTION

ANONYMIZATION may be employed in databases to achieve privacy protection. Through anonymization, personally identifiable information is removed, or obfuscated so that the people whom the data describe remain anonymous. Several techniques have been proposed in the literature to achieve anonymization (see, e.g., [1] for a survey).

While those techniques may exhibit several degrees of robustness against re-identification attacks (see, e.g., Bayesian attacks for differential privacy schemes [2], [3], or attacks based on the exploitation of externalities [4], or the use of recolouring techniques borrowed from signal processing [5]), they often fail to take into account the resulting utility of released data. For example, in randomization-based techniques, the amount of added noise needed to obscure the true data may be so large as to make the query output useless. In differential privacy, this may require a careful choice of the mechanism parameters [6], [7]. The contrasting wishes to minimize the risk of re-identification while providing useful data to queriers gives rise to an anonymization-utility trade-off, which triggers the need for proper utility measures and utility-aware anonymization techniques [8].

In this paper, we introduce a technique that would provide a utility-aware anonymization for databases. Our technique is based on a novel use of Principal Component Analysis (PCA), a statistical technique to achieve dimensionality reduction that has already been employed to protect privacy [9]. The technique is typically employed in signal processing applications to reduce the information as little as possible, while maintaining the quality of the output data. Instead, here we strive to achieve the maximum possible loss of information (and then potential individuals' identification), while preserving the utility of the output data. For that purpose, we introduce metrics of utility and incorporate them in the anonymization technique. The new approach we propose can therefore boast a built-in utility-aware mechanism, in contrast to existing anonymization technique, where data utility is not explicitly considered or is not properly measured.

Here we do not fully explore the technique, but rather set up a conceptual framework and propose several alternatives

for its implementation. Our contributions are:

- incorporating utility measurement in the proposed PCA-based anonymization scheme (Section III).
- proposing four categories of metrics to assess the residual utility of released data (Section IV).

II. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a statistical technique that applies linear transformations to a set of variables so as to arrive at a new set of variables that are statistically uncorrelated. It is typically employed to reduce the number of variables (the principal components, i.e. a subset of the new variables mentioned above) that retain most of the information contained in the original set. An extensive treatment of PCA is contained, e.g., in [10].

The PCA technique considers a set of n observations of p variables. Each observation represents a point in Rp. Actually, we are dealing with databases, where data

concerning subjects may be seen as matrices where each row is a record representing the data concerning a specific individual, and the columns are the fields representing the attributes of the data subject. For the time being, we consider just databases made of numerical entries: we defer the case of categorical attributes to a later implementation of our method.

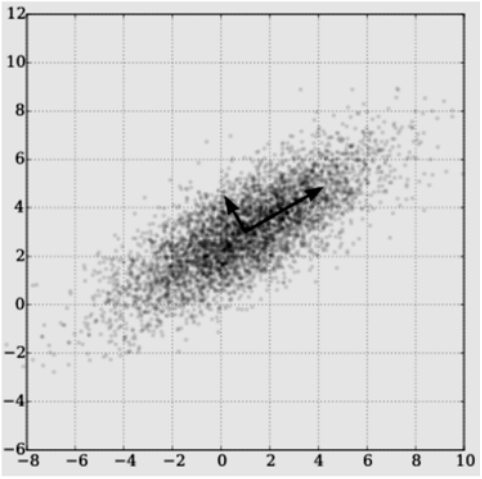

Fig. 1: Example of PCA for a multivariate Gaussian dataset

As recalled, each principal component is actually a linear transformation of the original variables. We can have at most the same number of principal components as the original variables, i.e. *p*, but the main aim of PCA is to reduce the dimensionality of the description: we wish to retain m principal components, with $m \leq p$.

The principal components are built starting from the most important one, i.e., that retaining the major portion of the information. In mathematical terms, that means building a linear combination of the p variables (i.e., setting its weights) so that it exhibits the maximum variance (under the constraint that the weight vector has unitary length).

A graphical representation of PCA in R2 is shown in Fig. 1, where the two axes representing the two principal components are shown for data following a multivariate Gaussian distribution.

The PCA technique has already been employed as an exploratory tool to describe large databases [11], [12]. Though the traditional way to employ PCA is to achieve dimensionality reduction by sacrificing as little information as possible (i.e., removing principal components starting from the least ones), for our purposes we wish instead to apply the PCA technique in a different way, to achieve as much information reduction as possible.

III. ANONYMIZATION BY PCA

After describing how PCA works, we now turn to its use to anonymize a database. In this section, we describe the overall method, starting with the rationale and then providing the step-by-step procedure.

PCA defines a new set of axes that represent linear combinations of the original variables. The typical application of PCA includes removing some principal components, by retaining the largest ones. Such a procedure has two limitations for our purpose:

- removing the least principal components achieves a reduction of information as small as possible;
- the principal components do not represent physically meaningful variables (e.g. a linear combination of height and income).

We overcome those limitations through the following modification of the typical application:

- we start removing principal components starting from the largest one;
- after removing the largest principal components, we project the resulting data onto the original axes, so that we revert to the original attributes of the subjects (e.g., height and income separately).

The resulting procedure goes through the following steps:

1) Form an array corresponding to the full database, represented as a collection of vectors in Rp;

2) Apply a PCA transformation to the array;

3) Remove the principal component corresponding to the largest eigenvalue;

4) Project the resulting collection of vectors on the original set of axes in Rp;

5) Apply a utility metric;

6) If the utility is still large enough then go back to step 2;

7) Transform the collection of vectors back into a database.

Steps 1-4 and 7 are quite straightforward and anyway well described in several textbooks. Instead, the choice of utility metrics and its employment are dealt with in the next section. The check embodied by Step 6 allows us to see if further information can be removed: if the utility is still large enough, that means that we can further reduce it and achieve a higher degree of anonymization.

IV. UTILITY METRICS

In the procedure we have sketched in Section III, we have mentioned the use of utility metrics. Those metrics allow us to assess how far the anonymized version of the database is from the true database or, in other terms, if the anonymized version can still be useful for queriers. In Section III we have not provided further details about such metrics. We fill the gap in this section, by providing four different proposals, i.e., four metrics that can serve that purpose. Those metrics are not intended to be alternative, since they could be employed in conjunction.

The four metrics (or classes of metrics, since we may have some alternative possibilities for each of them) we propose are:

1) Matrix norms;

2) Correlation;
3) Divergence measure;
4) Database image quality.

Matrix norms. The first metric is based on the matrix representation of the database. Our database of n records and p fields is represented as a matrix with n rows and p columns. In the following, we indicate the number of columns by m, since we assume to act after PC removal. We can first apply a standardisation procedure, so that the results are not influenced by the different ranges of the variables: for example, if the two fields were height and income, measured respectively in meters and euros, the differences in the income field would dominate. If we call A the matrix representing the true database and B the matrix representing the anonymised database (with elements aij and bij respectively), the standardisation field by field would require the computation of the mean and the standard deviation for each field:

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$$
$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n}(a_{ij}-\mu_j)^2}{n}} \tag{1}$$

By acting on the standardized values $\hat{a}_{ij} = \frac{a_{ij}-\mu_j}{\sigma_j}$ and $\hat{b}_{ij} = \frac{b_{ij}-\mu_j}{\sigma_j}$, we form the difference matrix D, whose generic element is $d_{ij} = |\hat{a}_{ij} - \hat{b}_{ij}| = \frac{|a_{ij}-b_j|}{\sigma_j}$, $1 \le i \le n$ and $1 \le j \le m$ (we are not interested in the sign of the distance between the true database and the anonymized one)

Having defined the difference matrix, we can now adopt a metric M for the size of its contents. Among the most widespread metrics for this purpose, we can consider for example the sum of all its elements:

$$M = \sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij} \tag{2}$$

Alternatively, we can consider one of the norms of D (see Chapter 2.6.5 of [13]), e.g., the L1 norm:

$$M = \max_{1\le i\le n}\sum_{j=1}^{m} d_{ij} \tag{3}$$

which we have adapted to our case, since we are more interested in picking the record showing the largest distance from its true value (hence summing along columns rather than along rows as in the textbook formulation of that norm).

Another norm of possible use is the Frobenius norm:

$$M = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij}^2} \tag{4}$$

Whatever the norm we employ, the larger it is, the less useful the anonymized database is.

Correlation. The second metric we consider is correlation. Though a correlation measure is not defined for matrices, we can convert the matrices of interest into vectors, by reading matrices row-by-row, and then apply the usual correlation definition. In our case, since we have two n _ p matrices A and B, representing respectively the true database and the PCA-treated one, we can convert them into two vectors v and w of length nm, by defining a row index r and a column index c and relating them to the vector index k as follows:

$$r = \left\lceil \frac{k-1}{m} \right\rceil + 1$$
$$c = k - (r-1)m \tag{5}$$

We can then apply the classical definition, so that the correlation is:

$$\rho = \frac{\sum_{i=1}^{nm}(v_i-\overline{v})(w_i-\overline{w})}{\sqrt{\sum_{i=1}^{nm}(v_i-\overline{v})^2\sum_{i=1}^{nm}(w_i-\overline{w})^2}} \tag{6}$$

Contrary to the use of the norm metric, now the larger this correlation coefficient, the more useful the anonymized database is.

Divergence measure. The values observed for the m fields of the n records can be considered as a realization of a multivariate random variable. When comparing the two databases (the true one and the anonymized one), we can then compare the two associated multivariate random variables.

We can then adopt a measure of distance between the two distributions as a utility metric for the anonymization process. A possible metric is Kullback-Leibler (KL) divergence [14]. For our case, we have the two multivariate random variables a and b, pertaining respectively to the true database and the anonymized one, with probability density function _(a) and (b) respectively, so that the KL divergence is:

$$D_{KL} = \int_a \phi(x)\log\frac{\phi(x)}{\psi(x)}dx \tag{7}$$

Though this is not properly a metric (it is not symmetric), we are going to employ it by keeping the true database as a reference, i.e., we always apply it in the same direction, so that the lack of symmetry is not relevant.

If we assume that the two random variables follow a multivariate normal distribution, the KL divergence takes the form [15], [16]:

$$D_{KL} = \frac{1}{2}\left[\mathrm{tr}(R_a R_b^{-1}) - \log\frac{|R_a|}{|R_b|} - m\right] + \frac{1}{2}\left[(\bar{a}-\bar{b})^T R_b^{-1}(\bar{a}-\bar{b})\right], \quad (8)$$

where R_a and R_b are the auto-covariance matrices of the two variates a and b.

Under this assumption, the KL divergence can be computed by estimating the auto-covariance matrices and plugging them into Equation (8). The larger the divergence is, the less useful is the anonymized database.

Database image quality. In order to assess how much utility is retained after principal components are removed, we can resort to the representation of a database as an image. In fact, our database can be viewed as a color-scale image, where the level of color of the pixel of indices (x; y) is the value of the attribute y of the record x (the same procedure applies to a greyscale image).

We can liken anonymization to image compression. Image compression aims at reducing the bit load associated to the image as much as possible while preserving the quality of the image, Reducing the bit load amounts to reducing the information content of the image, which is what we wish to achieve through anonymization. Since PCA has been applied to image compression as well [17], [18], [19], we can exploit the representation of databases as images to translate the results of image compression through PCA into database transformation through PCA as well.

Utility reduction can therefore be assessed through the tools that have been devised to assess the quality of an image after compression. Assessing the quality of an image is intrinsically difficult [20], and several approaches have been proposed, which may be classified into the two classes named subjective and objective approaches.

A major index employed under the subjective approach is the Mean Opinion Score (MOS). The MOS has first been conceived for audio signals [21], and then extended to images and videos [22]. Several possibilities exist for the testing methods (see, e.g., [23] for a comparison). The differences are mainly linked to the number of images that are shown to the human observer and the grading scale. As to to the first issue, the observer may be presented either just with the image whose quality is to be assessed (single stimulus approach) or with two images, i.e., the image under test and its reference (free of impairments). In the latter case we have the so-called double stimulus approach. As to grading, each observer can assign a mark to the image under test, by employing either a discrete scale (e.g., a five points scale) or a continuous scale (which means indicating a point on a line, where notches are shown corresponding to marks from 0 to 100). The overall procedure therefore goes through the following steps:

1) A set of human observers is selected, following the principles of experiment design;

2) each observer is presented with the image(s) to be assessed;

3) each observer assigns a grade on the scale indicated;

4) the arithmetic average of the grades assigned by the observers is computed to provide the Mean Opinion Score.

Despite its limitations [24], MOS remains a well established tool to assess the quality of images, and is considered as the ground truth assessment in many contexts, e.g. when assessing retargeting tasks [25]. In our case, we envisage opting for a double stimulus approach, with a continuous scale to avoid the quantization errors typically associated to the discrete scale.

Under the objective approach, two major indices are instead the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM) [26]. For an image of size n_m pixels, and a resolution of 8 bit/pixel, the PSNR for the reference image f and the (distorted) image g under test is:

$$\mathrm{PSNR} = 10\log_{10}\frac{255^2}{\mathrm{MSE}(f,g)}, \quad (9)$$

where the Mean Squared Error MSE is defined as:

$$\mathrm{MSE}(f,g) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(f_{ij}-g_{ij})^2 \quad (10)$$

The PSNR is always non negative but is not bounded; a high PSNR means that the database with image g is close to the database with image f (since high PSNR implies a low MSE). Instead, SSIM is defined as the product of three quality reduction factors, namely loss of correlation s, luminance distortion l, and contrast distortion c:

$$\mathrm{SSIM} = s(f,g)l(f,g)c(f,g) = \frac{\sigma_{fg}+C_3}{\sigma_f\sigma_g+C_3}\times\frac{2\mu_f\mu_g+C_1}{\mu_f^2+\mu_g^2+C_1}\times\frac{2\sigma_f\sigma_g+C_2}{\sigma_f^2+\sigma_g^2+C_2}, \quad (11)$$

where μ_f *and* μ_g are the average luminance of *f* and *g*, σ_f and σ_g the respective standard deviations, and C_1, C_2 and C_3 are three positive constants to avoid having a null denominator. In contrast to the PSNR, SSIM is bounded and takes value in the [0,1] range, with higher values representing images *g* closer to the reference image *f*.

Though adopting the image paradigm, we must however stop short of applying a full parallelism when employing PCA. In fact, images are typically correlated both horizontally and vertically. While we may assume that a horizontal correlation (among the attributes of a single record) exists in databases, it would be hazardous to assume that a vertical correlation (between different records) exists as well.

A possible approach would therefore consist in applying image compression through PCA to a set of well-known references pictures, e.g., Lena [27], by first removing the vertical correlation (e.g., by randomly shuffling the image rows) and then progressively removing the largest principal components, assessing the resulting image quality through

MOS, and proceeding till the image is still recognizable. We would then consider the number of removed principal components as the limit number of principal components that we can safely remove when applying PCA to a database. We expect the results to depend on the particular image we are considering, through the structure of eigenvalues. In order to reduce that bias, we envisage building sets of image libraries that exhibit a similar eigenvalues structure, e.g. by looking at the dominant eigenvalue or a parameter of the eigenvalue curve.

Fig. 2: Lena original image

TABLE I: QUALITY INDICES AFTER PRINCIPAL COMPONENTS REMOVAL

No. of component removed	SSIM	PSNR
1	0.9335	20.0081
2	0.9036	17.9902
5	0.8614	16.1854

Some sets would allow for just a limited number of components to be removed before the image is not recognizable, while other sets would allow for a more extensive removal before the degradation is excessive. This could be related to the identifiability of the processed image (and, in parallel, of the database to be anonymized), and be considered for properly training the MOS-based anonymizer.

We can report now an early example of the effect of principal component removal. In Fig. 2, we show the original picture of Lena in greyscale.

In Fig. 3, we report the image as it appears after removing respectively just the largest, the two largest, or the five largest principal components. The removal of the largest PC does not impact the quality, while we see some artefact when we remove two PCs, and the artefacts get more significant when we remove 5 PCs, though the image is still well recognizable.

The resulting objective quality indices for those cases are shown in Table I: similarly to what a quick visual analysis of the picture would tell us, those indices fall sharply as more PCs are removed.

As to the identifiability properties of the image as revealed by the eigenvalue structure, we show in Fig. 4 how the eigenvalues decay. The superimposed curve is the best fit (R2 = 0:9993) obtained with a symmetrical sigmoidal curve:

$$y = d + \frac{a-d}{1+\left(\frac{x}{c}\right)^b}, \tag{12}$$

(a) Removal of largest principal component

(b) Removal of the two largest principal components

(c) Removal of the five largest principal components

Fig. 3: Impact of principal component removal.

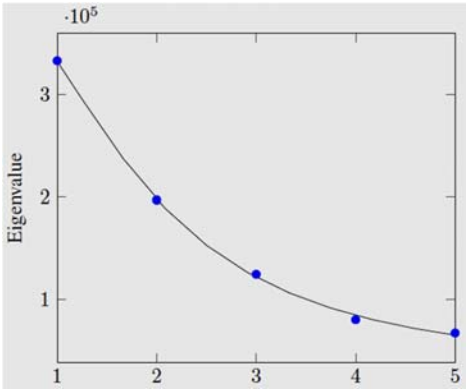

Fig. 4. Decay of dominant eigenvalues for Lena

where c and b respectively act as a scale factor and a shape factor, with the latter suitable to represent the speed at which eigenvalues decay. In our case, we have b = 2:21.

V. CONCLUSIONS

Previous proposals to anonymize data do not explicitly take into account the residual utility of data after the database has been processed to achieve privacy. We propose enhancing an approach based on Principal Component Analysis by incorporating a measure of utility and proposing four classes of metrics that can be used for that purpose. This approach represents a step forward in the direction of devising reliable privacy protection mechanisms that actually provide usable data.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Fabio Ricciato of Eurostat for the fruitful discussions on the subject, and Mr. Luca Della Gatta for contributing Fig. 3 and Table I.

REFERENCES

[1] J. Domingo-Ferrer, D. Sánchez, and J. Soria-Comas, "Database anonymization: privacy models, data utility, and micro aggregation based inter-model connections," Synthesis Lectures on Information Security, Privacy, & Trust, vol. 8, no. 1, pp. 1–136, 2016.

[2] M. Naldi and G. D'Acquisto, "Differential privacy for counting queries: can Bayes estimation help uncover the true value?" arXiv preprint arXiv:1407.0116, 2014.

[3] X. Zhang, J. Hamm, M. K. Reiter, and Y. Zhang, "Statistical privacy for streaming traffic," in Network and Distributed Systems Security (NDSS) Symposium 2019, 2019.

[4] M. Naldi and G. D'Acquisto, "Mr X vs. Mr Y: The Emergence of Externalities in Differential Privacy," in Annual Privacy Forum. Springer, 2017, pp. 120–140.

[5] M. Naldi, A. Mazzoccoli, and G. D'Acquisto, "Hiding alice in wonderland: A case for the use of signal processing techniques in differential privacy," in Annual Privacy Forum. Springer, 2018, pp. 77–90.

[6] M. Naldi and G. D'Acquisto, "Differential privacy: An estimation theory-based method for choosing epsilon," arXiv preprint arXiv:1510.00917, 2015.

[7] N. Kohli and P. Laskowski, "Epsilon voting: Mechanism design for parameter selection in differential privacy," in 2018 IEEE Symposium on Privacy-Aware Computing (PAC). IEEE, 2018, pp. 19–30.

[8] J. Domingo-Ferrer, S. Ricci, and J. Soria-Comas, "A methodology to compare anonymization methods regarding their risk-utility tradeoff," in Modeling Decisions for Artificial Intelligence. Springer, 2017, pp. 132–143.

[9] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang, "Analyze gauss: optimal bounds for privacy-preserving principal component analysis," in Proceedings of the forty-sixth annual ACM symposium on Theory of computing. ACM, 2014, pp. 11–20.

[10] I. Jolliffe, "Principal component analysis," in International encyclopedia of statistical science. Springer, 2011, pp. 1094–1096.

[11] R. R. Meglen, "Examining large databases: a chemometric approach using principal component analysis," Marine Chemistry, vol. 39, no.1-3, pp. 217–237, 1992.

[12] E. G. Emberly, R. Mukhopadhyay, C. Tang, and N. S. Wingreen, "Flexibility of _-sheets: Principal component analysis of database protein structures," Proteins: Structure, Function, and Bioinformatics, vol. 55, no. 1, pp. 91–98, 2004.

[13] D. Zwillinger, CRC standard mathematical tables and formulae. Chapman and Hall/CRC, 2002.

[14] S. Kullback and R. A. Leibler, "On information and sufficiency," The annals of mathematical statistics, vol. 22, no. 1, pp. 79–86, 1951.

[15] Y. Kakizawa, R. H. Shumway, and M. Taniguchi, "Discrimination and clustering for multivariate time series," Journal of the American Statistical Association, vol. 93, no. 441, pp. 328–340, 1998.

[16] A. R. Runnalls, "Kullback-Leibler approach to Gaussian mixture reduction," IEEE Transactions on Aerospace and Electronic Systems, vol. 43, no. 3, 2007.

[17] Q. Du and J. E. Fowler, "Hyperspectral image compression using jpeg2000 and principal component analysis," IEEE Geoscience and Remote sensing letters, vol. 4, no. 2, pp. 201–205, 2007.

[18] S. Lim, K. H. Sohn, and C. Lee, "Principal component analysis for compression of hyperspectral images," in Geoscience and Remote Sensing Symposium, 2001. IGARSS'01. IEEE 2001 International, vol. 1. IEEE, 2001, pp. 97–99.

[19] J. Wang and C.-I. Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," IEEE transactions on geoscience and remote sensing, vol. 44, no. 6, pp. 1586–1600, 2006.

[20] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in ICASSP, vol. 4, 2002, pp. 3313–3316.

[21] ITU-T, "Recommendation P.80: Methods for subjective determination of transmission quality," 1996.

[22] ITU-R, "Recommendation BT.500-11: Methodology for the subjective assessment of the quality of television pictures," 2002.

[23] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in Visual Communications and Image Processing 2003, vol. 5150. International Society for Optics and Photonics, 2003, pp. 573–583.

[24] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (mos) revisited: methods and applications, limitations and alternatives," Multimedia Systems, vol. 22, no. 2, pp. 213–227, 2016.

[25] L. Ma, W. Lin, C. Deng, and K. N. Ngan, "Image retargeting quality assessment: A study of subjective scores and objective metrics," IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 6, pp.626–639, 2012.

[26] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in 2010 20th International Conference on Pattern Recognition. IEEE, 2010, pp. 2366–2369.

[27] D. C. Munson, "A note on Lena," IEEE Transactions on Image Processing, vol. 5, no. 1, pp. 3–3, 1996.