Information-Theoretic Risk and Utility Measures for Microdata

Josep Domingo-Ferrer[1]

Universitat Rovira i Virgili

Chair in Data Priv@cy

http://unescoprivacychair.urv.cat

November 20-21, 2008

[1] Based on discussions with David Rebollo-Monedero

Introduction

- Information loss measures in SDC of microdata are usually based on the relative discrepancy between some statistics or models computed on the original data X and on the masked/synthetic data X'[2]
- A critique to the above measures is that, for continuous attributes, relative discrepancies are unbounded and difficult to combine with disclosure risk, which is naturally bounded between 0 and 1[3]

---

[2] E.g. Domingo-Ferrer and Torra (2001) "Disclosure protection methods and information loss for microdata". In *Confidentiality, Disclosure and Data Access*, Elsevier, 91-110.

[3] Trottini (2003) *Decision Models for Data Disclosure Limitation*, Ph. Thesis, Carnegie-Mellon Univ.

Probabilistic information loss measures

- Probabilistic information loss measures yielding a figure between $[0, 1]$ which can be readily compared to disclosure risk have been proposed [4].
- Let θ be a population parameter (on X) and let $\hat{\Theta}$ be the corresponding sample statistic (on X').
- If the size n' of X' is large (> 100), then

$$Z = \frac{\hat{\Theta} - \theta}{\sqrt{\hat{\Theta}}}$$

can be assumed to follow a $N(0, 1)$ distribution.

[4] Mateo-Sanz, Domingo-Ferrer and Torra (2005) "Probabilistic information loss measures in confidentiality protection of continuous data", *Data Mining and Knowledge Discovery* 11(2):181-193.

Probabilistic information loss measures (II)

A probabilistic information loss measure $pil(\theta)$ for parameter θ is the probability that the absolute value of the discrepancy Z is $\leq$ the actual discrepancy in sample X':

$$pil(\theta) = 2 \cdot P(0 \leq Z \leq \frac{|\hat{\theta} - \theta|}{\sqrt{Var(\hat{\Theta})}})$$

Clearly, the more different is $\hat{\Theta}$ from θ, the greater is $pil(\theta)$.

Contribution and plan of this talk

- Motivation for information-theoretic measures
- Information-theoretic loss measures
- Information-theoretic risk measures
- Loss-risk optimization models for perturbation and synthetic data
- Conclusions

Motivation for information-theoretic measures

- Loss measures based on relative discrepancies are very easy to understand, but rather difficult to trade off against risk (unboundedness).
- Probabilistic loss measures have the following strong points:
  - They can be applied to the same usual statistics θ (means, variances, covariances, etc.) like measures based relative discrepancies.
  - They are bounded within $[0, 1]$, so they easily compare to disclosure risk.
- Both relative-discrepancy and probabilistic loss measures lack an underlying theory to allowing to optimize their trade-off with disclosure risk.

Mutual information

- The mutual information $I(X;Y)$ between two random variables X and Y measures the mutual dependence of the two variables and is measured in bits.
- Mutual information can be expressed as a function of Shannon's entropy:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

$$= H(X) + H(Y) - H(X,Y)$$

where $H(X), H(Y)$ are marginal entropies, $H(X|Y), H(Y|X)$ are conditional entropies and $H(X,Y)$ is the joint entropy of X and Y.

Universitat Rovira i Virgili

Mutual information and random Gaussian data

- If U and V are random, jointly Gaussian vectors, and U' is the best linear estimate of U from V, then U' is a sufficient statistic, that is, $I(U'; V) = I(U; V)$.
- If U and V are random, jointly Gaussian scalars with correlation coefficient ρ, then $I(U; V) = -\log\sqrt{1-\rho^2}$.
- If U and V are random, jointly Gaussian vectors with matrix correlation

  $$P = \Sigma_U^{-1/2}\Sigma_{UV}\Sigma_V^{-1/2}$$

  then

  $$I(U; V) = -1/2 \log\det(I - PP^t)$$

  where P^t is the transpose of P, I the identity matrix and $\det(\cdot)$ is the determinant.

Mutual information (II)

If mutual information can be used to express information loss or/and disclosure risk, then the machinery of information theory can be used to optimize the tradeoff between both quantities.

Information-theoretic loss measures

- Let X, Y be, respectively, the key and confidential attributes in the original microdata set.
- Let X' the key attributes in the masked microdata set (as in k-anonymization, we assume that only key attributes are masked).
- If we focus on the damage inflicted to key attributes[5], a possible information loss measure is the expected distortion $E(d(X, X'))$ where $d(x, x')$ is a distortion measure, *e.g.* $d(x, x') = ||x - x'||^2$.
- A probably better option is to focus on how masking affects the dependences between the key and confidential attributes.
- A possible measure for this is $I(X; Y) - I(X'; Y)$.

[5] Rebollo-Monedero, Forné and Domingo-Ferrer (2008) "From t-closeness to PRAM and noise addition via information theory", in *PSD 2008*, LNCS 5262, 100-112.

Mutual information vs MSE

- The MSE $E(d(X,X')) = E(||X - X'||^2)$ seems better adapted than $I(X;X')$ to measuring how well statistical properties are preserved.
- However, the MSE and the mutual information are not that different, both belonging to the family of so-called Bregman divergences[6]. [7]

[6] Rebollo-Monedero (2007), *Quantization and Transforms for Distributed Source Coding*, Ph. D. Dissertation, Stanford University.

[7] Bregman (1967), "The relaxation method of finding the common point of conves sets and its application to the solution of problems in convex programming", *USSR Comput. Math., Math. Phys.*, 7: 200-217

Mutual information vs correlations

- $I(X;Y) - I(X';Y)$ bears some resemblance to the relative discrepancy between correlation matrices proposed as a loss measure by Domingo-Ferrer and Torra (2001).
- However, mutual information measures the general dependence between attributes, while the correlation measures only the linear dependence, so the former is superior[8].
- It will be shown below that, under some assumptions, preserving mutual information preserves the covariance matrix up to a constant factor.

[8] Wentian Li (1990) "Mutual information functions vs correlation functions", *Journal of Statistical Physics*, 60: 823-837.

Information-theoretic risk measures

- The mutual information $I(X'; X)$ between the released and the original key attributes is a measure of identity disclosure [9].
- The mutual information $I(X'; Y)$ between the released key attributes and the confidential attributes is a measure of attribute disclosure.
- Measuring risk as $I(X'; Y)$ conforms to the t-closeness privacy property [10] requiring that the distance between the distribution of Y within records sharing each combination of values of X' and the distribution of Y in the overall dataset be no more than t.

[9] Note that $I(X'; X)$ was previously regarded as a possible information loss measure (which it is for key attributes).

[10] Li, Li and Venkatasubramanian (2007) "t-Closeness: privacy beyond k-anonymity and l-diversity". In *ICDE 2007*, 106-115.

Loss-risk optimization

- Several combinations of the above loss and risk measures can be used when trying to optimize the tradeoff of information loss and disclosure risk.
- Two approaches:
  - Place an upper-bound constraint on the loss D and minimize the risk R.
  - Place an upper-bound constraint on the risk R and minimize the loss D (more natural in SDC).

Universitat Rovira i Virgili

Model 1

$$\inf_{p_{X'|X}} R(D) = I(X'; Y)$$

$$\text{subject to } \; D = E(d(X, X')) \leq d$$

for a certain pre-specified maximum tolerable loss d.

Model 1 and perturbation

- Model 1 was related in Rebollo-Monedero, Forné and Domingo-Ferrer (2008) to the rate-distortion function optimization in information theory: the risk R was assimilated to the rate and the loss D to the distortion.
- An optimal random perturbation $p(X'|X)$ key attributes was obtained.
- For the case of univariate Gaussian, real-valued X and Y, a closed form of the minimum was obtained:

$$R_{inf} = -\frac{1}{2}\log(1-(1-d)\rho^2_{XY})$$

$$p^{opt}_{X'|X} = N(\mu_X, (1-d)\sigma^2_X)$$

Model 2 and perturbation

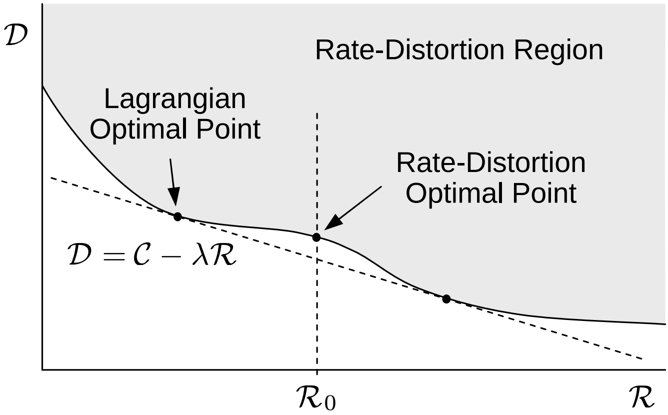
If we take the more natural approach of minimizing D for a maximum tolerable risk r, we get

$$\inf_{p_{X'|X}} D(R) = E(d(X, X'))$$

$$\text{subject to} \;\; R = I(X'; Y) \leq r$$

- This problem could be related to optimizing the distortion-rate function optimization in quantization (future work).
- This again yields an optimal perturbation $p_{X'|X}$, which can be heuristically computed.

Risk-loss as Lagrangian rate-distortion optimization



Models 3 and 4

Model 3

$$\inf_{p_{X'|X}} R(D) = I(X;X')$$

$$\text{subject to } \; D = I(X;Y) - I(X';Y) \leq d$$

Model 4

$$\inf_{p_{X'|X}} D(R) = I(X;Y) - I(X';Y)$$

$$\text{subject to } \; R = I(X;X') \leq r$$

UNIVERSITAT ROVIRA I VIRGILI

Model 4 and synthetic data generation

- Synthetic data generation can be viewed as a form of perturbation [11].
- If we want to generate synthetic key attributes X' in such a way that the connection between key attributes and confidential attributes is minimally affected, we can use Model 4 to compute $p_{X'|X}$.
- Synthetic X' can be generated by drawing from $p_{X'|X}$.

[11] Abowd and Vilhuber (2008) "How protective are synthetic data?", in *PSD 2008*, LNCS 5262, 239-246.

Mutual information vs covariance preservation

We justify that preserving mutual information (that is, achieving $D = 0$ in Model 4) preserves the covariance matrix (up to a constant factor):

- Let X and Y be zero-mean, jointly Gaussian r.v, $\mathbb{R}$- and $\mathbb{R}^n$-valued, respectively.
- Let $X' = a^T Y$ be the best linear MSE estimate of X given Y, for $a \in \mathbb{R}^k$.
- Then $a = \Sigma_{XY} \Sigma_Y^{-1}$

Mutual information vs covariance preservation (II)

- The covariance matrix is preserved when replacing X by X'

$$\Sigma_{X'Y} = \Sigma_{XY}\Sigma_Y^{-1}\Sigma_Y = \Sigma_{XY}$$

- At the same time, X' is a sufficient statistic for X given Y, that is, $I(X'; Y) = I(X; Y)$[12].

[12] Rebollo-Monedero, Rane, Aaron and Girod (2006), "High-rate quantization and transform coding with side information at the decoder", *Signal Processing* 86:3160-3179, Prop. 15.

Conclusions

- Information loss measures based on relative discrepancies are awkward to combine with risk measures in order to optimize the risk-loss tradeoff.
- Probabilistic loss measures are a step forward, but lack a theoretical framework.
- We have explored here loss and risk measures based on information theory, namely on mutual information.
- Models for optimizing the information-theoretic risk-loss tradeoff when perturbing data and generating synthetic data have been presented.
- It has been shown that preserving mutual information offers covariance matrix preservation.

Future work

- The information-theoretic measures and models are just a first step.
- In the context of synthetic data generation, information-theoretic loss measures should be devised whose minimization is equivalent to preserving a given model.
- Whenever possible, closed-form expressions for the optimal $p_{X'|X}$ transformations would be desirable.
- If a closed form expression is not possible, a convex optimization problem to be solved numerically is the next most attractive option.