Rathindra Sarathy

TITLE: Sufficiency-based Non-Synthetic Perturbation Approach

ABSTRACT: A recently developed method of noise added perturbation makes it possible to maintain the mean vector and covariance matrix of the masked data to be exactly the same as that of the original data. The mean vector and covariance matrix are sufficient statistics when the underlying distribution is multivariate normal. Many type of statistical analyses used in practice rely on the assumption of multivariate normality (Gaussian model). For these analyses, maintaining the sufficient statistics of the masked data to be the same as that of the original data guarantees that the results of such analyses using masked data will be the same as that using the original data. However, as it is currently proposed, the perturbed values from this method are considered synthetic because they are generated without considering the values of the confidential variables (and are based only on the non-confidential variables). Some researchers argue that synthetic data results in information loss. In this study, we provide several suggestions for generating non-synthetic perturbed data that maintains the mean vector and covariance matrix of the masked data to be exactly the same as the original data while offering a selectable degree of similarity between original and perturbed data. We also provide a methodology for assessing the trade-off between disclosure risk and information loss for this approach.