Recent Results On The Noise Addition Method For Database Security

Patrick Tendick and Norman S. Matloff, University of California, Davis
Norman S. Matloff, Dept. of Computer Science, UCD, Davis, CA 95616

Abstract: Matloff[1] has shown that the noise addition method for protecting the privacy of individual records can introduce serious biases into query responses. This paper summarizes bias correction methods for both the multivariate normal and nonparametric cases. Methods for obtaining standard errors for population parameters estimated from query responses will also be discussed.

Key Words: Statistical database, regression function, deconvolution, minimum distance estimation.

1. The Noise Addition Method of Preserving Individual Privacy

In a *statitistical database,* records containing information about individuals sampled from a population are stored for the purpose of studying *population characteristics.* For example, a user might want to know the average income of all individuals in a population who are over fifty years old, and might then estimate this using the average income of all individuals in the database over fifty. On the other hand, a devious user of the database might try to obtain information about specific individuals. For this reason, noise is often added to the attributes of the records in order to preserve individual privacy[2]. That is, if X is an attribute stored in the database, X is replaced by

$$Z = X + \delta \quad , \tag{1.1}$$

where δ is a zero mean random variable independent of X. The reasoning behind this is that since the added noise has mean zero, it averages out when an attribute is averaged over a large number of records. Thus, individual attributes are changed drastically, but sample means of attributes are altered only slightly. (Here we are treating X as a random variable, since it is assumed that the data comprise a random sample from some population.)

Unfortunately, if the modified observations are used to estimate certain population parameters, a bias results. For example, if $X \sim N(20,4)$ and $\delta \sim N(0,4)$, then

$$P\{X_i > 24\} = 0.159 \quad , \tag{1.2}$$

whereas

$$P\{Z_i > 24\} = 0.239 \quad . \tag{1.3}$$

Here X might be the income of a randomly chosen individual, in thousands of dollars; adding noise has artificially increased the income of some people, pushing a disproportionate number of their incomes above $24,000. Noise addition can also bias the regression function of one variable upon another[1]. For example, if (X,Y) is bivariate normal, then the regression function of Y on X is

$$m(X) = \alpha + \beta X \quad , \tag{1.4}$$

where

$$\beta = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \quad . \tag{1.5}$$

In order to insure privacy, the database manager might replace (X,Y) with (Z,W), where

$$Z = X + \delta \tag{1.6}$$

and

$$W = Y + \epsilon \quad , \tag{1.7}$$

where δ and ϵ are zero mean normal random variables that are independent of (X,Y) and each other. The regression function of Z on W is then

$$m'(Z) = \alpha + \beta' Z \quad , \tag{1.8}$$

where

$$\begin{aligned} \beta' &= \frac{\mathrm{Cov}(Z,W)}{\mathrm{Var}(Z)} \\ &= \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X) + \mathrm{Var}(\delta)} \quad . \end{aligned} \tag{1.9}$$

For example, if $\mathrm{Var}(\delta)=\mathrm{Var}(X)$, then a considerable bias is introduced. This is the well known *errors in variables regression problem.*

Both types of bias mentioned above may be thought of as constituting sampling bias. That is, the bias is due to the fact that the population we are sampling from is not the true target population. This type of bias is, of course, very serious, since it does not subside as the sample size increases. Thus, means of correcting or avoiding this problem are needed if the noise addition method is to be useable.

For the noise addition method to be useable, it is also necessary to have a *measure of security,* that is, a quantity that indicates the degree of protection provided by adding noise with a certain variance. In the univariate case, such a measure is the squared correlation coefficient between Z and X, namely

$$\rho^2 = \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + \mathrm{Var}(\delta)} \quad . \tag{1.10}$$

This is, of course, the proportion of Var(X) that a devious user can account for using a linear predictor based upon Z. If ρ^2 is close to zero, the level of protection is high, whereas if ρ^2 is close to 1, the amount of protection provided is small. Note that ρ^2 approaches zero as the variance of the added noise increases.

2. Bias Correction In The Multivariate Normal Case

Let us now examine the effect of adding noise to a (p+1)-dimensional vector $(\mathbf{X}^T, Y)^T$ of attributes. In order to protect individual privacy, $(\mathbf{X}^T, Y)^T$ might be replaced by

$$\begin{pmatrix} \mathbf{Z} \\ W \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ Y \end{pmatrix} + \begin{pmatrix} \delta \\ \epsilon \end{pmatrix} \quad , \tag{2.1}$$

where $(\delta^T, \epsilon)^T$ is a zero mean random vector independent of $(\mathbf{X}^T, Y)^T$. If $(\mathbf{X}^T, Y)^T$ is multivariate normal with mean vector $(\mu_X^T, \mu_Y)^T$ and covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \sigma_{XY} \\ & \sigma_{YY} \end{pmatrix} \tag{2.2}$$

then the regression function of Y on $\mathbf{X}$ is[3]

$$m(\mathbf{x}) = \mu_Y + \sigma_{XY}^T \Sigma_{XX}^{-1}(\mathbf{x} - \mu_X) \quad . \tag{2.3}$$

If $(\delta^T,\epsilon)^T$ is multivariate normal with mean zero and covariance matrix

$$\Sigma_D = \begin{pmatrix} \Sigma_{\delta\delta} & \sigma_{\delta\epsilon} \\ & \sigma_{\epsilon\epsilon} \end{pmatrix} \quad , \tag{2.4}$$

then $(\mathbf{Z}^T,W)$ is multivariate normal with mean $(\mu_X^T,\mu_Y)^T$ and covariance matrix $\Sigma+\Sigma_D$. Thus, the regression function of W on $\mathbf{Z}$ is

$$m'(\mathbf{x}) = \mu_Y + (\sigma_{XY}^T + \sigma_{\delta\epsilon}^T)(\Sigma_{XX} + \Sigma_{\delta\delta})^{-1}(\mathbf{x} - \mu_X) \quad , \tag{2.5}$$

which does not have the same regression coefficients as the regression function of Y on $\mathbf{X}$ in general. However, since Σ_D is known and $\Sigma+\Sigma_D$ (and hence Σ) can be estimated from the distorted data, the database software can correct this bias in the regression function.

Suppose that $\Sigma_D=d\Sigma$, that is, the added noise has the same covariance structure as the original data. Then the regression coefficients are not biased by noise addition. This is because the noise that is being added does not water down the relationship between $\mathbf{X}$ and Y. For example, suppose that X is age and Y is income, and let us assume that the relationship between X and Y is positive. If the noise added to X is positive, that is, the individual is made older by noise addition, then the noise added to Y is probably also positive since the relationship between δ and ϵ is also positive, so that the person is also made richer. In this way, the regression relationship between X and Y is preserved.

One might expect that since

$$m(\mathbf{x}) = E(Y\,|\mathbf{X}=\mathbf{x}) \tag{2.6}$$

is not biased if $\Sigma_D=d\Sigma$, then neither is

$$m(A) = E(Y\,|\mathbf{X}\in A) \tag{2.7}$$

$$= E(m(\mathbf{X})\,|\mathbf{X}\in A) \quad . \tag{2.8}$$

Unfortunately, this is not the case, because although $m(\mathbf{x})$ is not biased by adding this type of noise, the distribution of $\mathbf{X}$ over A is changed. However, the bias is easy to correct, and it can be shown that

$$m(A) = \frac{1}{d_1}m'(d_1A - d_2\mu_X) + \frac{d_2}{d_1}\mu_Y \quad , \tag{2.9}$$

where $d_1=\sqrt{1+d}$ and $d_2=d_1-1$, and where for any scalar a, set C, and vector $\mathbf{v}$ in $\mathbf{R}^p$, the set $aC+\mathbf{v}$ is defined to be $\{a\mathbf{c}+\mathbf{v}:\mathbf{c}\in C\}$. A similar bias correction may be found in the case where Σ_D is arbitrary.

3. Simulation Results

It can be argued that the bias correction (2.9) should work reasonably well even if the population is not multivariate normal. A simulation study of the bivariate case has supported this[4]. The bias correction was used to calculate a number of quantities of the form $E(Y\,|X\in A)$ based upon distorted data, and the bias corrected quantities where then compared with the true conditional expectations. This was done for eight different nonnormal bivariate densities and for three different values of the noise parameter d. The bias correction performed well in all cases, even those in which it was expected to fail.

4. Improved Protection From Adding Correlated Noise

Since the bias in $m(\mathbf{x})$ or $m(A)$ can be corrected for multivariate added noise with *any* covariance structure, what is the advantage of adding noise with the same covariance structure as the original population? One slight advantage is that the bias correction formulas are simpler. The most important difference, however, is that using $\Sigma_D=d\Sigma$ provides much better protection. Recall that we proposed using as a measure of security in the univariate case the correlation coefficient of the distorted attribute with the undistorted attribute. Suppose that the user wants to determine the value of a linear function of the data

$$U = \mathbf{c}^T\mathbf{X} \quad . \tag{4.10}$$

A multivariate generalization of the squared correlation is the squared multiple correlation coefficient[3] between U and $\mathbf{Z}$, namely

$$\rho_{UZ}^2 = \frac{\mathbf{c}^T\Sigma(\Sigma+\Sigma_D)^{-1}\Sigma\mathbf{c}}{\mathbf{c}^T\Sigma\mathbf{c}} \quad . \tag{4.11}$$

This measures the proportion of the variance of U that the user can account for by knowing $\mathbf{Z}$. We will use this as a measure of the security provided U by noise addition. If the components of the added noise vector δ are independent with variances proportional to the variance of the corresponding components of $\mathbf{X}$, i.e., $\Sigma_D=d\cdot\mathrm{diag}(\Sigma)$, then it can be shown that for $\mathbf{c}$ properly chosen, ρ_{UZ}^2 can be as large as

$$\rho_{*Z}^2 = \sup_{\mathbf{c}\in\mathbf{R}^p} \rho_{UZ}^2 \tag{4.12}$$

$$= \frac{\lambda_1}{\lambda_1 + d} \quad , \tag{4.13}$$

where λ_1 is the largest eigenvalue of the correlation matrix of $\mathbf{X}$. Since λ_1 can be as large as p, the dimension of $\mathbf{X}$, the amount of information that the user can obtain about certain linear functions of the data can be quite large for attribute vectors of high dimension.

On the other hand, if $\Sigma_D=d\Sigma$, then

$$\rho_{UZ}^2 = \frac{1}{1+d} \tag{4.14}$$

for all linear functions U. That is, *all* linear functions of the data receive the same amount of protection when the added noise has the same covariance structure as the original population.

5. Bias Avoidance In The Nonnormal Case

Now suppose that the distribution of the attribute X is not close to being normal. Our goal is to provide users with estimates of quantities of the form

$$\theta = E[g(X)] \quad , \tag{5.1}$$

where g is a continuous function provided by the user. The mean and variance are all of this form, and the proportion of the population falling in a given interval is well approximated by quantities of this form. In general, estimates of the form

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(Z_i) \quad , \tag{5.2}$$

are biased, and bias corrections like (2.9) are not applicable. We shall avoid the bias problem by *deconvolving,* that is, we will nonparametrically estimate the distribution function of the random variable X based upon a sample from the distribution of the distorted variable Z. The parameter estimate $\tilde{\theta}$ will then be computed from the estimated distribution function of X. To illustrate the method, we will consider only the univariate case, although the multivariate case is a simple extension. We will assume that the distributions of both X and Z place all mass within a finite interval [a,b].

If X has distribution function F_X and the added noise δ has distribution function F_δ, then the distorted attribute

$$Z = X + \delta \tag{5.3}$$

has distribution function

$$F_Z(z) = \int_R F_\delta(t)dF_X(z-t) \tag{5.4}$$

provided X and δ are independent. Nonparametric estimates of F_Z based upon a sample $Z_1,...,Z_n$ are easy to obtain. However, using such an estimate to get an estimate of F_X is not straightforward. This is because for an arbitrary distribution function F_Z, there need not exist a distribution function F_X for which (5.4) holds. We will use minimum distance methods to choose an estimate $\tilde{F}_X$ so that the corresponding estimate

$$\tilde{F}_Z(z) = \int_R F_\delta(z-t)d\tilde{F}_X(t) \tag{5.5}$$

of F_Z is close to the data in some way; minimum distance estimates have been used extensively in the estimation of finite mixtures[5,6]. One such method is to choose $\tilde{F}_X$ to minimize

$$s_1^2(\hat{F}_Z,\tilde{F}_Z) = \int_a^b [\hat{F}_Z(z) - \tilde{F}_Z(z)]^2 dt \tag{5.6}$$

subject to (5.5), where F_Z is the empirical distribution function of $Z_1,...,Z_n$; s_1^2 is similar to the Cramer-Von Mises distance. An alternative is to minimize the squared "distance" between the fitted moment generating function of Z and the sample moment generating function, that is, minimize

$$s_2^2(\overline{\psi}_Z,\tilde{\psi}_Z) = \int_c^d [\overline{\psi}_Z(t) - \tilde{\psi}_Z(t)]^2 dt \tag{5.7}$$

subject to

$$\tilde{\psi}_Z(t) = \psi_\delta(t)\int_R e^{tx}d\tilde{F}_X(x) \tag{5.8}$$

where

$$\overline{\psi}_Z(t) = \frac{1}{n}\sum_{i=1}^{n} e^{tZ_i} \tag{5.9}$$

is the *empirical moment generating function of* Z, ψ_δ is the moment generating function of δ, $\tilde{\psi}_Z$ is the fitted moment generating function of Z, and c and d are appropriate constants. The motivation behind these estimators is that we have a natural estimator of the distribution function or moment generating function of Z, but this estimator cannot be used directly to find an estimator of F_X. Rather, we choose our estimate $\tilde{F}_X$ of F_X in such a way that the corresponding estimate $\tilde{F}_Z$ of F_Z is close to the natural estimate of F_Z or ψ_Z.

We will not work directly with the estimators found by minimizing (5.6) or (5.7), but rather with numerical approximations to them. Consider the estimator which minimizes (5.6). We could approximate F_X with a distribution function which places mass on a grid on [a,b], that is, use

$$F_X(x) = \sum_{j=0}^{q} \pi_j I_{[x_j,\infty)}(x) \quad , \tag{5.10}$$

$$\sum_{j=0}^{q} \pi_j = 1 \quad , \tag{5.11}$$

$$\pi_j \geq 0 \quad \text{for all } j \quad , \tag{5.12}$$

where

$$x_j = a + \frac{j(b-a)}{q} \quad , \quad j = 0,...,q \quad . \tag{5.13}$$

Plugging this into (5.5) yields

$$F_Z(z) = \sum_{j=0}^{q} \pi_j F_\delta(z-x_j) \quad . \tag{5.14}$$

In other words, F_Z is treated as a finite mixture of known distributions. If the integral in (5.6) is then approximated using a rectangular rule, a grid of points $z_1,...,z_m$ is chosen on [a,b] so that

$$z_k = a + \frac{k(b-a)}{m} \quad , \quad k = 1,...,m \quad . \tag{5.15}$$

and the problem is to minimize

$$s_1^2 = \frac{1}{m}\sum_{k=1}^{m} \left[\hat{F}_Z(z_k) - \sum_{j=0}^{q} \pi_j F_\delta(z_k - x_j)\right]^2 \tag{5.16}$$

with respect to $\pi_0,...,\pi_q$ subject to (5.11) and (5.12). Ignoring the nonnegativity constraints (5.12), this approximation is just a linear least squares problem, solveable with any standard regression package.

The estimator obtained by minimizing (5.16) subject to (5.11) is a reasonable estimator in its own right. If F_X is of the form (5.10), then the estimates $\tilde{\pi}_1,...,\tilde{\pi}_q$ of $\pi_1,...,\pi_q$ obtained in this manner are unbiased and have covariance matrix

$$\text{Cov}(\tilde{\pi}) = \frac{1}{n}(\Gamma^T\Gamma)^{-1}\Gamma^T\Lambda\Gamma(\Gamma^T\Gamma)^{-1} \quad , \tag{5.17}$$

where

$$\Gamma_{ij} = F_\delta(z_i - x_j) - F_\delta(z_i - x_0) \tag{5.18}$$

and

$$\Lambda_{ij} = F_Z[\max(z_i,z_j)] - F_Z(z_i)F_Z(z_j) \quad ; \tag{5.19}$$

Γ is known and Λ is easily estimated from $Z_1,\ldots,Z_n$. Furthermore, θ may be estimated by

$$\tilde{\theta} = \sum_{j=0}^{q} \tilde{\pi}_j g(x_j) \quad , \tag{5.20}$$

which is unbiased (still assuming that the density of X places all of its mass on $x_0,\ldots,x_q$) and approximately normal with variance

$$\mathrm{Var}(\tilde{\theta}) = \frac{1}{n}\mathbf{g}^T(\Gamma^T\Gamma)^{-1}\Gamma^T\Lambda\Gamma(\Gamma^T\Gamma)^{-1}\mathbf{g} \quad , \tag{5.21}$$

where

$$\mathbf{g} = [g(x_1) - g(x_0),\ldots,g(x_q) - g(x_0)]^T \quad . \tag{5.22}$$

Thus, it is possible to add noise to the sensitive attribute X, estimate the distribution of X from the distorted attribute Z, and use this estimated distribution to find nearly unbiased estimators of quantities of the form (5.1) for any continuous function g provided by the user. The only bias is due to the fact that the distribution of X is taken to be discrete, and this bias subsides very rapidly as the grid used becomes "finer". The estimators are approximately normal, and standard errors are easily estimated. And privacy is protected because of the added noise.

References

1. N. Matloff, "Another Look at the Use of Noise Addition for Database Security", Proceedings of the 1986 IEEE Symposium on Security and Privacy.

2. J. Traub, H. Wozniakowski, and Y. Yemini, "Statistical Security of a Statistical Data Base, Technical Report", Columbia University Department of Computer Science, September 1981.

3. R. Johnson, and D. Wichern, *Applied Multivariate Statistical Analysis,* Prentice-Hall, 1982.

4. P. Tendick, and N. Matloff, "A Solution to the Bias Problem in a Database Security Mechanism", unpublished manuscript.

5. R. Quandt, and J. Ramsey, "Estimating Mixtures of Normal Distributions and Switching Regressions", Journal of the American Statistical Association, December 1978.

6. W. Woodward, W. Parr, W. Schucany, and H. Lindsey, "A Comparison of Minimum Distance and Maximum Likelihood Estimation of a Mixture Proportion", Journal of the American Statistical Association, September 1984.