THE USE OF MEASUREMENT ERROR TO AVOID DISCLOSURE

Gary Sullivan and Wayne A. Fuller, Iowa State University
Wayne A. Fuller, Statistical Laboratory, ISU, Ames, IA 50011

ABSTRACT

The data perturbation technique of masking each data vector by adding a random error vector is considered. After describing the general technique, we consider the approach an intruder might use in attempting to determine an individual's confidential attributes. It is shown that the conditional expected value of the attributes given the masked data and the public data is the best predictor of the unknown attributes. We present a masking algorithm designed to preserve the moments and univariate distribution functions of masked variables, while providing disclosure protection. The procedure is designed so that the covariance structure of the masked data is similar to that of the original data.

1. INTRODUCTION

The statistical community became concerned about maintaining respondent confidentiality in the late 1960s when files of linked records, data banks, and statistical file systems were initially requested by researchers. Steinberg and Pritzker (1967) suggested that files of linked records be created in a manner that would maintain confidentiality. They advised record linkers to "expunge all individual identifiers at the instant of creation." Bachi and Baron (1969) give a summary of the confidentiality problems faced in setting up data banks or linking records between data files. Duncan and Lambert (1986) provide a good review of the federal statutes dealing with confidentiality.

Mugge (1983) discusses confidentiality measures taken at the National Center for Health Statistics. Cox, et al. (1985) provide a good discussion of Census Bureau data products and the techniques used to mask them before release.

A microdata record contains detailed information about an individual respondent. A microdata file is a valuable asset in economic modeling, statistical analysis and general research. Unfortunately, public access to raw microdata records poses a direct threat to confidentiality. Even after identifiers such as name and address are stripped from the records, an indirect threat may still exist if the remaining information is abundant.

Paass (1985) investigates the case of an intruder attempting to determine the identity of a record in a microdata release by matching to a record in a public use file. Paass believes the intruder's biggest problem in attempting to disclose identity is caused by measurement error in the records. The problem defined by Paass is similar to matching masked records to the unmasked originals.

In response to Paass' findings, Kim (1986) proposed a masking scheme which combines the addition of error with a linear transformation that adds an additional layer of protection. Kim's objective was to create a new data set with the same correlation structure as the original data.

Others discussing microdata confidentiality problems include include Fellegi (1975), Gates (1988), McGuckin and Nguyen (1988), Spruill (1983) and Wolf (1988).

Building upon the ideas for the masking of microdata, this research is directed toward finding effective ways to preserve respondent confidentiality by masking data with added error.

2. MODEL AND RESULTS FOR THE NORMAL DISTRIBUTION

Let $S(x) = [\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)]$ represent the N data records belonging to a confidential sample. Assume $\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)$ are independent $N[\boldsymbol{\mu}, \Sigma(xx)]$ random vectors, where $\mathbf{x}(j) = [x(j1), x(j2), \ldots, x(jk)]'$.

A microdata release is to be formed from the confidential sample and will contain m different records $(m \le N)$. The released microdata file is denoted by

$$MD_X = (\mathbf{X}_{n_1}, \mathbf{X}_{n_2}, \ldots, \mathbf{X}_{n_m})'$$

where $\mathbf{X}'_{n_j} = (X_{j1}, X_{j2}, \ldots, X_{jk})$,

$$\mathbf{X}_{n_j} = \mathbf{x}_{n_j} + \mathbf{u}_{n_j} \quad , \quad j = 1, 2, \ldots, m \ ,$$

$\mathbf{u}$ are independent $N[\mathbf{0}, \Sigma(uu)]$ random vectors, and are independent of the $\mathbf{x}$ for all i,j . Hence, $X(n1), X(n2), \ldots, X(nm)$ are independent $N[\boldsymbol{\mu}, \Sigma(XX)]$ where $\Sigma(XX) = \Sigma(xx) + \Sigma(uu)$.

A confidentiality problem arises when a record from an independent private data source, having an identification variable (e.g., name), is known by an intruder. From the intruder's perspective, the objective is to predict the values of the confidential variables of the target individual. Consider the following as an intruder's approach to this problem.

The target record is partitioned as

$$\mathbf{x}'_0 = (\mathbf{x}'_{0,1}, \mathbf{x}'_{0,2})_{1\times k}$$

where $\mathbf{x}(0,1)$ $(\ell \times 1)$ is known and $\mathbf{x}(0,2)$ is unknown. The records in the microdata release are partitioned in the same way,

$$\mathbf{X}_{n_i} = \begin{pmatrix} \mathbf{X}_{n_i,1} \\ \mathbf{X}_{n_i,2} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{n_j,1} \\ \mathbf{x}_{n_j,2} \end{pmatrix} + \begin{pmatrix} u_{n_i,1} \\ u_{n_i,2} \end{pmatrix} , \qquad (2.1)$$

where

$$\mathbf{X}_{n_i} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{XX11} & \Sigma_{XX21} \\ \Sigma_{XX12} & \Sigma_{XX22} \end{pmatrix}\right) . \qquad (2.2)$$

Also,

$$\boldsymbol{\Sigma}_{uu} = \begin{pmatrix} \boldsymbol{\Sigma}_{uu11} & \boldsymbol{\Sigma}_{uu12} \\ \boldsymbol{\Sigma}_{uu21} & \boldsymbol{\Sigma}_{uu22} \end{pmatrix}$$

and

$$\boldsymbol{\Sigma}_{xx} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx11} & \boldsymbol{\Sigma}_{xx12} \\ \boldsymbol{\Sigma}_{xx21} & \boldsymbol{\Sigma}_{xx22} \end{pmatrix}$$

are known positive definite matrices.

Assuming $\boldsymbol{\mu} = 0$, the conditional density of [**X**(n1), **X**(n2), ..., **X**(nm), **x**(0,1)] given that **x**(nj) corresponds to the target is

$$f[(\mathbf{X}_{n_1}, \mathbf{X}_{n_2}, \ldots, \mathbf{X}_{n_m}, \mathbf{x}_{0,1}) | \mathbf{x}_{n_j} = \mathbf{x}_0]$$

$$= [(2\pi)^{-\frac{mk+\ell}{2}} |\boldsymbol{\Sigma}_{XX}|^{-m/2} |\mathbf{A}|^{-1/2}]$$

$$\times \exp\{-1/2 [\sum_{t=1}^{m} \mathbf{X}'_{n_t} \boldsymbol{\Sigma}_{XX}^{-1} \mathbf{X}_{n_t}$$

$$+ (\mathbf{x}_{0,1} - \mathbf{B}\mathbf{X}_{n_j})'\mathbf{A}^{-1}(\mathbf{x}_{0,1} - \mathbf{B}\mathbf{X}_{n_j})]\}$$

$$= \text{constant} \times [\exp\{-1/2 \sum_{t=1}^{m} \mathbf{X}'_{n_t} \boldsymbol{\Sigma}_{XX}^{-1} \mathbf{X}_{n_t}\}]\gamma_{j0} , \qquad (2.3)$$

where

$$\mathbf{B} = \boldsymbol{\Sigma}_{x_1X} \boldsymbol{\Sigma}_{XX}^{-1} ,$$

$$\boldsymbol{\Sigma}_{x_1X} = E\{\mathbf{x}_{n_j,1} \mathbf{X}'_{n_j}\} ,$$

$$\mathbf{A} = \boldsymbol{\Sigma}_{xx11} - \boldsymbol{\Sigma}_{x_1X} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{Xx_1}$$

and

$$\gamma_{j0} = \exp\{-1/2 (\mathbf{x}_{0,1} - \mathbf{B}\mathbf{X}_{n_j})'\mathbf{A}^{-1}(\mathbf{x}_{0,1} - \mathbf{B}\mathbf{X}_{n_j})\} ,$$

$$\text{for } j = 1,2,\ldots,m . \qquad (2.4)$$

We now use γ(10), γ(20), ..., γ(m0) , to define the conditional probabilities, p(10), p(20), ..., p(m0) , that each record in the microdata release corresponds to the target record. Let

$$p_{j0} = (\sum_{t=1}^{m} \gamma_{t0})^{-1} \gamma_{j0} , \quad \text{for } j = 1,2,\ldots,m . \qquad (2.5)$$

Thus, p(j0) is the conditional probability that the j-th record in the microdata release corresponds to the target record, given that the target record is contained in the released file and given [**X**(n1), ..., **X**(nm), **x**(0,1)] . These probabilities are now used as weights in the construction of a predictor for the confidential variables of the target record.

The ideal situation for the intruder attempting to predict the values of the confidential variables of the target record is to have p(k0) = 1 for some k ϵ {1,2,...,m} and p(j0) = 0 for all other j ϵ {1,2,...,m} . In this case, record **x**(nk) is the target record, where

$$\mathbf{X}_{n_k} = \mathbf{x}_{n_k} + \mathbf{u}_{n_k} . \qquad .$$

Consider the partitioning of **X**(nk) into non-confidential and confidential sub-vectors as in (2.1). Then, assuming **x**(0) = **x**(nk) , the minimum mean square error predictor of **x**(0,2) is the conditional expectation of **x**(nk,2) given [**X**(nk), **x**(0,1)] . Hence, the best predictor of **x**(nk,2) is

$$\hat{\mathbf{x}}_{n_k,2} = E\{\mathbf{x}_{n_k,2} | (\mathbf{X}_{n_k}, \mathbf{x}_{0,1}) \text{ and } \mathbf{x}_0 = \mathbf{x}_{n_k}\}$$

$$= \mathbf{X}_{n_k,2} - E\{\mathbf{u}_{n_k,2} | (\mathbf{X}_{n_k}, \mathbf{x}_{0,1})$$

$$\text{and } \mathbf{x}_0 = \mathbf{x}_{n_k}\}$$

$$= \mathbf{X}_{n_k,2} - \hat{\mathbf{u}}_{n_k,2} \qquad (2.6)$$

where

$$\hat{\mathbf{u}}_{n_k,2} = \begin{pmatrix} \boldsymbol{\Sigma}_{uu12} \\ \boldsymbol{\Sigma}_{uu22} \\ \mathbf{0} \end{pmatrix}' \begin{pmatrix} \boldsymbol{\Sigma}_{XX11} & \boldsymbol{\Sigma}_{XX12} & \boldsymbol{\Sigma}_{xx11} \\ \boldsymbol{\Sigma}_{XX21} & \boldsymbol{\Sigma}_{XX22} & \boldsymbol{\Sigma}_{xx21} \\ \boldsymbol{\Sigma}_{xx11} & \boldsymbol{\Sigma}_{xx12} & \boldsymbol{\Sigma}_{xx11} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_{n_j,1} \\ \mathbf{X}_{n_j,2} \\ \mathbf{x}_{0,1} \end{pmatrix} \qquad (2.7)$$

is the best predictor of **u**(nk,2) . Therefore, assuming **x**(nk) is the target record, **x**(nk,2) is the minimum mean square error predictor of **x**(0,2) .

The situation with p(k0) = 1 for some k will rarely occur in a real data situation. If the p(j0)'s are nearly equal, information from all records in the microdata release should enter into the prediction of **x**(0,2) . We now construct such a predictor.

The knowledge base for prediction of **x**(0,2) is **Σ**(uu) , **Σ**(xx) , **x**(0,1) , and the records in the microdata release. The conditional probability that **x**(nj) is the target, **x**(0) , given [**X**(n1), ..., **X**(nm), **x**(0,1)] is the p(j0) in (2.5). The conditional distribution of **x**(0,2) given [**X**(n1), ..., **X**(nm), **x**(0,1)] and **x**(nj) = **x**(0) is

$$f[\mathbf{x}_{0,2} | (\mathbf{X}_{n_1}, \ldots, \mathbf{X}_{n_m}, \mathbf{x}_{0,1}) \text{ and } \mathbf{x}_{n_j} = \mathbf{x}_0]$$

$$= (2\pi)^{-\frac{(k-\ell)}{2}} |\boldsymbol{\Sigma}_{\epsilon\epsilon}|^{-1/2}$$

$$\times \exp\{-1/2 [\mathbf{x}_{0,2} - \mathbf{C}(\mathbf{X}'_{n_j}, \mathbf{x}'_{0,1})']'$$

$$\times \boldsymbol{\Sigma}_{\epsilon\epsilon}^{-1} [\mathbf{x}_{0,2} - \mathbf{C}(\mathbf{X}'_{n_j}, \mathbf{x}'_{0,1})']\} \qquad (2.8)$$

where

$$\mathbf{C} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx12} \\ \boldsymbol{\Sigma}_{xx22} \\ \boldsymbol{\Sigma}_{xx12} \end{pmatrix}' \begin{pmatrix} \boldsymbol{\Sigma}_{XX11} & \boldsymbol{\Sigma}_{XX12} & \boldsymbol{\Sigma}_{xx11} \\ \boldsymbol{\Sigma}_{XX21} & \boldsymbol{\Sigma}_{XX22} & \boldsymbol{\Sigma}_{xx21} \\ \boldsymbol{\Sigma}_{xx11} & \boldsymbol{\Sigma}_{xx12} & \boldsymbol{\Sigma}_{xx11} \end{pmatrix}^{-1} \qquad (2.9)$$

and

$$\mathbf{\Sigma}_{\epsilon\epsilon} = \mathbf{\Sigma}_{xx22} - \mathbf{C}(\mathbf{\Sigma}_{xx12}, \mathbf{\Sigma}_{xx22}, \mathbf{\Sigma}_{xx12})' . \qquad (2.10)$$

Hence, the conditional distribution of $\mathbf{x}(0,2)$ given $[\mathbf{X}(n1), \mathbf{X}(n2), \ldots, \mathbf{X}(nm), \mathbf{x}(0,1)]$ and given that $\mathbf{x}(0)$ corresponds to some record in the microdata release, is

$$f[\mathbf{x}_{0,2} | (\mathbf{X}_{n_1}, \ldots, \mathbf{X}_{n_m}, \mathbf{x}_{0,1}) \text{ and } \mathbf{x}_0 = \mathbf{x}_{n_j} \text{ for some } j \in \{1,2,\ldots,m\}]$$

$$= (2\pi)^{-\frac{m(k-\ell)}{2}} |\mathbf{\Sigma}_{\epsilon\epsilon}|^{-\frac{m}{2}} \sum_{j=1}^{m} p_{j0}$$

$$\times \exp\{-\tfrac{1}{2} [\mathbf{x}_{0,2} - \mathbf{C}(\mathbf{X}'_{n_j}, \mathbf{x}'_{0,1})']'$$

$$\times \mathbf{\Sigma}_{\epsilon\epsilon}^{-1}[\mathbf{x}_{0,2} - \mathbf{C}(\mathbf{X}'_{n_j}, \mathbf{x}'_{0,1})']\} \qquad (2.11)$$

where $\mathbf{C}$ and $\mathbf{\Sigma}(\epsilon\epsilon)$ are defined in (2.9) and (2.10), respectively. The best predictor of $\mathbf{x}(0,2)$ is then the mean of the conditional distribution of $\mathbf{x}(0,2)$, given $[\mathbf{X}(n1), \ldots, \mathbf{X}(nm), \mathbf{x}(0,1)]$,

$$\hat{\mathbf{x}}_{0,2} = \sum_{j=1}^{m} p_{j0}\mathbf{C}(\mathbf{X}'_{n_j}, \mathbf{x}'_{0,1})' . \qquad (2.12)$$

If we let $J = \{1,2,\ldots,m\}$, the variance of the predictor error can be expressed as

$$V\{(\mathbf{x}_{0,2} - \hat{\mathbf{x}}_{0,2}) | (\mathbf{X}_{n_1}, \ldots, \mathbf{X}_{n_m}, \mathbf{x}_{0,1}) \text{ and } \mathbf{x}_{n_j} = \mathbf{x}_0 \text{ for some } j \in J\}$$

$$= \mathbf{\Sigma}_{\epsilon\epsilon} + \sum_{j=1}^{m} p_{j0}\mathbf{W}'_j\mathbf{W}_j - (\sum_{j=1}^{m} p_{j0}\mathbf{W}_j)'(\sum_{j=1}^{m} p_{j0}\mathbf{W}_j) \qquad (2.13)$$

where $\mathbf{W}(j) = [\mathbf{X}'(nj), \mathbf{x}'(0,1)]\mathbf{C}'$.

We now investigate the problem of selecting an error covariance matrix to use in masking normal data. We consider the problem in the context of protection against an intruder with a data record of non-confidential variables, $\mathbf{x}(0,1)$, from a private source. Again, the microdata release consists of $\mathbf{X}(n1), \ldots, \mathbf{X}(nm)$ with

$$\mathbf{X}'_{n_j} = (\mathbf{X}'_{n_j,1}, \mathbf{X}'_{n_j,2})$$
$$= (\mathbf{x}'_{n_j,1}, \mathbf{x}'_{n_j,2}) + (\mathbf{u}'_{n_j,1} + \mathbf{u}'_{n_j,2})$$

for $j=1, 2, \ldots, m$.

Certainly a large error variance will lower the probability of matching a record. At the same time adding large error will distort the data. The data provider must balance the objectives of providing a file that resembles the original data as closely as possible and providing confidentiality protection for the respondents. It is always possible to transform the x-vectors so that

$$\mathbf{x}_{n_j} \sim N(\boldsymbol{\mu}, \delta\mathbf{I}_k) .$$

We assume in this section that the covariance matrix is $\delta\mathbf{I}(k)$ and that a decision has been made to fix the ratio of error variance to total variance at $1/(1 + \delta)$ for all variables. Hence,

$$\mathbf{u}_{n_j} \sim N(\mathbf{0}, \mathbf{\Sigma}_{uu}) ,$$

where

$$\Sigma_{uu} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{12} & 1 & \cdots & \rho_{2k} \\ \vdots & & & \vdots \\ \rho_{1k} & \rho_{2k} & \cdots & 1 \end{pmatrix} , \qquad (2.14)$$

and $\mathbf{\Sigma}(uu)$ is positive definite. We proceed to determine the optimal structure of $\mathbf{\Sigma}(uu)$ from the standpoint of the data provider.

Recall that the intruder computes the predictor defined in (2.12) as

$$\hat{\mathbf{x}}_{0,2} = \sum_{j=1}^{m} p_{j0}\hat{x}_{n_j,2} .$$

Suppose $x(0,1)$ corresponds to the k-th record in the microdata release. A high probability for $p(kk)$ means a more accurate predictor of the confidential variables. Hence, the data provider attempts to minimize $p(kk)$ thereby improving confidentiality protection. Minimizing $p(kk)$ is closely related to minimizing the log odds ratio, $\ell n[p(kk)p(jk)^{-1}]$, for a randomly chosen element j and fixed k. So, we consider the problem of choosing the correlations in $\mathbf{\Sigma}(uu11)$ to minimize the expected value of the log odds ratio, $E\{\ell n[p(kk)] - \ell n[p(jk)]\}$. The data dependence is removed by considering the expectation. We proceed to show that in order to minimize $E\{\ell n[p(kk)] - \ell n[p(jk)]\}$, the data provider should add vectors of error having a covariance matrix equal to a multiple of $\delta\mathbf{I}$, the covariance matrix of the unmasked data vectors. We first give some results which lead to the main theorem.

Theorem 2.1. Assume the microdata release consists of $\mathbf{X}(n1), \mathbf{X}(n2), \ldots, \mathbf{X}(nm)$ where $\mathbf{X}(nj)$ satisfies (2.1, 2.2) for $j=1, 2, \ldots, m$. Also, for all j,

$$\mathbf{x}_{n_j} \sim N(\mathbf{0}, \mathbf{\Sigma}_{xx}) ,$$
$$\mathbf{u}_{n_j} \sim N(\mathbf{0}, \mathbf{\Sigma}_{uu}) ,$$

where $\mathbf{\Sigma}(xx) = \delta\mathbf{I}(\ell)$ and $\mathbf{\Sigma}(uu)$ is given by (2.14). Assume the target record, $\mathbf{x}(0)$, corresponds to the k-th record in the microdata release. Then minimizing $E\{\ell n[p(kk)] - \ell n[p(jk)]\}$ $(j \neq k)$, where $p(jk)$ is defined in (2.5), is equivalent to minimizing $\mathrm{tr}\{\mathbf{A}^{-1}\}$, where

$$\mathbf{A} = \Sigma_{xx11} - \Sigma_{x_1X}\Sigma_{XX}^{-1}\Sigma_{Xx_1}$$

and

$$\Sigma_{Xx_1} = E\{\mathbf{X}_{n_j}\mathbf{x}'_{n_j,1}\} .$$

Proof. Omitted.

We now investigate $\mathrm{tr}\{\mathbf{A}^{-1}\}$. It follows from the definition of $\mathbf{A}$ that

$$\mathbf{A}^{-1} = \{\delta^{-1}[\mathbf{I}_\ell - \mathbf{E}^{-1}]^{-1}\}$$

where

$$\mathbf{E} = \delta(\Sigma_{XX11} - \Sigma_{XX12}\Sigma_{XX22}^{-1}\Sigma_{XX21}) .$$

Let $\lambda(1) \geq \lambda(2) \geq \ldots \geq \lambda(\ell)$ be the characteristic roots of

$$|(\mathbf{I}_\ell - \mathbf{E}^{-1}) - \lambda_i\mathbf{I}_\ell| = 0 .$$

If we define the roots of $\mathbf{E}$ as $\alpha(1) \geq \alpha(2) \geq \ldots \geq \alpha(\ell)$, then $\alpha(i) = 1/[1 - \lambda(i)]$ and $\lambda(i) = [\alpha(i) - 1]/\alpha(i)$, for $i = 1,2,\ldots,\ell$. So, minimizing

$$\mathrm{tr}\{[\mathbf{I}_\ell - \mathbf{E}^{-1}]^{-1}\} = \sum_{i=1}^{\ell} \lambda_i^{-1}$$

is equivalent to minimizing

$$\sum_{i=1}^{\ell} (\alpha_i - 1)^{-1}\alpha_i = \ell + \sum_{i=1}^{\ell} (\alpha_i - 1)^{-1} . \qquad (2.15)$$

We now show that the minimum of

$$\sum_{i=1}^{\ell} 1/[\alpha(i) - 1]$$

is attained when $\rho(ij) = 0$ for all $i \leq j \leq \ell$. We begin by demonstrating that the roots of $\mathbf{E}$ are individually maximized when $\Sigma(XX21) = \mathbf{0}$.

Lemma 2.1. Assume $\mathbf{S}$ and $\mathbf{T}$ are real, symmetric $(n \times n)$ matrices. Define the characteristic roots of $\mathbf{S}$ and $\mathbf{T}$ to be $\nu(1,\mathbf{S}) \geq \ldots \geq \nu(n,\mathbf{S})$ and $\nu(1,\mathbf{T}) \geq \ldots \geq \nu(n,\mathbf{T})$, respectively. Then,

$$\nu(j,\mathbf{S}) + \nu(n,\mathbf{T}) \leq \nu(j,\mathbf{S+T}) \quad \text{for } j=1,2,\ldots,n .$$

Proof. See Bhatia (1987, p.34). □

Theorem 2.2. Let

$$\Sigma_{XX} = \begin{pmatrix} \Sigma_{XX11} & \Sigma_{XX12} \\ \Sigma_{XX21} & \Sigma_{XX22} \end{pmatrix}$$

be a real, symmetric $(k \times k)$ positive definite matrix, with $\Sigma(XX11)$ having dimension $(\ell \times \ell)$. Let $\beta(1) \geq \beta(2) \geq \ldots \geq \beta(\ell)$ and $\gamma(1) \geq \gamma(2) \geq \ldots, \geq \gamma(\ell)$ be the roots of $\Sigma(XX11)$ and $[\Sigma(XX11) - \Sigma(XX12)\Sigma(XX22)^{-1}\Sigma(XX21)]$, respectively. Then $\beta(i) \geq \gamma(i)$ for $i=1, 2, \ldots, \ell$.

Proof. Follows directly from Lemma 2.1. □

Therefore, the roots of $\mathbf{E} = \Sigma(XX11) - \Sigma(XX12)\Sigma(XX22)^{-1}\Sigma(XX21)$ are individually maximized when $\Sigma(XX21) = \mathbf{0}$ or when $\mathbf{E} = \Sigma(XX11)$.

We have now reduced the problem to minimizing (2.15) where $\alpha_1 \geq \ldots \geq \alpha_\ell > 1$ are the roots of

$$\Sigma_{XX11} = \delta\mathbf{I}_\ell + \Sigma_{uu11} ,$$

with $\delta > 0$ and $\Sigma(uu11)$ having the structure of a correlation matrix. Since the roots of $\Sigma(XX11)$ are equal to the roots of $\Sigma(uu11)$ each increased by δ , our objective is to minimize

$$\sum_{i=1}^{\ell} \phi_i^{-1} = \mathrm{tr}\{\Sigma_{uu11}^{-1}\} \qquad (2.16)$$

where $\phi(1) \geq \phi(2) \geq \ldots \geq \phi(\ell) > \max\{0, 1-\delta\}$ are the roots of $\Sigma(uu11)$. Theorem 2.3 states that the minimum is obtained when all correlations are set equal to zero.

Theorem 2.3. Assume $\Sigma(uu11)$ is a symmetric non-singular matrix of the form given in (2.14). The minimum of (2.16) is attained when $\Sigma(uu11) = \mathbf{I}(\ell)$.

Proof. Omitted.

We have shown that, for the identity covariance matrix and the ratio of error variance to total variance fixed at $1/(1 + \delta)$, the correct match probability, $p(kk)$, is minimized on the average when $\Sigma(uu11) = \delta\mathbf{I}(\ell)$. Therefore, when creating a microdata file, there is a sense in which a data provider affords respondents maximum protection against disclosure by adding error vectors which have a covariance matrix equal to a multiple of the covariance matrix of the original data vectors.

3. A MEASUREMENT ERROR ALGORITHM FOR CONFIDENTIALITY PROTECTION

In this chapter we outline a method of adding measurement error to the variables of a data set to protect the confidentiality of the respondents. The objectives of the procedure are to create a new data set such that:

1. The variables in the new data set are the sum of the original variables and a measurement error.
2. The covariance matrix of the new set of variables is nearly the same as the covariance matrix of the original variables.
3. The marginal sample cumulative distribution function of each of the created variables is nearly the same as the marginal cumulative distribution function of the corresponding original variable.
4. The probability that an intruder with some information on an individual can correctly identify the record of that individual is considerably less than one.

These objectives are competitive. The covariance matrix and the marginal distribution functions are maintained with small error, while larger error reduces the probability of correct identification.

In order to use the properties of the normal distribution, the masking procedure consists of several steps. The observed variables are transformed into pseudo normal random variables, normal error is added to the normal variables, and then the sum is back transformed to the original scale. We outline the procedure of transforming data to normality prior to adding normal measurement error. The treatment for each variable is:

1. Construct the sample cumulative distribution function (CDF). Reduce any extreme observations. The reduction of very extreme observations is required if measurement error of reasonable variance is to protect confidentiality.
2. Convert the sample CDF to the sample CDF of a uniform random variable.
   A. If the original variable is a continuous variable, the step function CDF is converted to a continuous piece-wise linear function using linear interpolation between the points that are half-way between the jump points.
   B. If the variable is a discrete variable, there is a proportion of the observations associated with each value of the discrete variable. A corresponding proportion of the interval (0, 1) is assigned to each value. For each observation a pseudo uniform observation is generated by making a random selection of a value within the assigned interval.
3. Convert the new uniform observation from step 1 or step 2 to a $N(0, 1)$ random variable.

A multinomial variable with k categories requires k-1 zero-one variables to identify the category into which an observation falls. These k-1 variables satisfy certain restrictions. If the number one is used to identify the occupied category, no more than one of the variables for a particular observation can take on the value one. Let

$$\begin{aligned} Z_{ti} &= 1 \quad \text{if observation } t \text{ falls in category } i \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

for t=1, 2, ..., n and i=1, 2, ..., k, where n is the number of observations and k is the number of categories.

The first step in transforming these variables is to create a new set of uncorrelated binomial variables. These are based on the conditional distribution of $Z(ti)$ given $Z(t,1)$, $Z(t,2)$, ..., $Z(t,i-1)$. To illustrate the procedure, assume there are 5 cells and let

$$\begin{aligned} \phi_1 &= P\{\text{cell } 1\} , \\ \phi_2 &= P\{\text{cell } 2|\text{Not } 1\} = (1 - P_1)^{-1}P_2 , \\ \phi_3 &= P\{\text{cell } 3|\text{Not } (1, 2)\} = (1 - P_1 - P_2)^{-1}P_3 , \\ \phi_4 &= P\{\text{cell } 4|\text{Not } (1, 2, 3)\} \\ &= (1 - P_1 - P_2 - P_3)^{-1}P_4 . \end{aligned}$$

To create the uncorrelated pseudo variables, let

$$\begin{aligned} W_1 &= Z_1 , \\ W_2 &= Z_2 \quad \text{if} \quad Z_1 = 0 \\ &= 1 \text{ with Prob. } \phi_2 \text{ if } Z_1 = 1 \\ &= 0 \text{ with Prob. } (1 - \phi_2) \text{ if } Z_1 = 1 , \\ W_3 &= Z_3 \quad \text{if} \quad Z_1 = Z_2 = 0 \\ &= 1 \text{ with Prob. } \phi_3 \text{ if } Z_1 + Z_2 = 1 \\ &= 0 \text{ with Prob. } (1 - \phi_3) \text{ if } Z_1 + Z_2 = 1 , \\ W_4 &= Z_4 \quad \text{if} \quad Z_1 = Z_2 = Z_3 = 0 \\ &= 1 \text{ with Prob. } \phi_4 \text{ if } Z_1 + Z_2 + Z_3 = 1 \\ &= 0 \text{ with Prob. } (1 - \phi_4) \text{ if } Z_1 + Z_2 + Z_3 = 1 . \end{aligned}$$

Note that the created W variables are binomial variables with mean equal to the conditional probabilities. The W's go into the transform operation for discrete variables. After the error masking operation, the masked W-variables are transformed back into Z-variables or, equivalently, into a variable that identifies the category.

The above operations define transformations of continuous, discrete and categorical random variables into normal random variables. The sample covariance matrix of the set of normal variables is then computed. Let the column vector of normal observations be $\mathbf{Z}(t)$ and let $\mathbf{m}(ZZ)$ be the sample covariance matrix. By the method of construction, the diagonal elements of $\mathbf{m}(ZZ)$ will be approximately equal to one. Let $\xi(t)$ be a vector of normal independent (0, 1) random variables of the same dimension as $\mathbf{Z}(t)$. Then the masked variable is

$$\begin{aligned} \overset{*}{\mathbf{Z}}_t &= \mathbf{Z}_t + \tau\, \mathbf{m}_{ZZ}^{1/2}\, \xi_t \\ &= \mathbf{Z}_t + \mathbf{a}_t , \end{aligned}$$

where $\mathbf{a}(t) = \tau\, \mathbf{m}(ZZ)^{1/2}\, \xi(t)$ and τ is a constant chosen by the data provider. The larger τ, the greater the confidentiality protection gives to the respondents. The data provider can experiment with τ and matching programs to choose a τ that affords the desired level of protection. The τ will be a function of the number of observations, the number of variables and the distribution of the observations.

During the masking operation, it seems desirable to check that no generated error vector is "too close" to zero. The exact details of this restriction on the generated vector would probably be kept secret. The program developed at Iowa State University has a routine that rejects error vectors for which $\xi'(t)\xi(t)$ is too small.

The elements of $\overset{*}{\mathbf{Z}}(t)$ are mapped back to uniform random variables and, hence, back to the original scale using the inverse of the sample CDF. The masked variables are denoted by $\overset{*}{\mathbf{X}}(t)$. Because the sample CDF is used to map the variables back, the sample CDF of the $\overset{*}{\mathbf{X}}(t)$-

variables will be very close to that of the original $\mathbf{X}(t)$-variables. Also the covariance matrix of the $\overset{*}{\mathbf{X}}(t)$-variables will be close to that of the $\mathbf{X}(t)$-variables because the added error has the same covariance matrix as the transformed variables. If the original variables are normal, the covariance matrix of the transformed variables will differ from the original covariance matrix only because of random variation. For nonnormal variables, additional differences between the two covariance matrices are introduced by the transformations of the variables into normal variables and the associated back transformations of the masked variables.

REFERENCES

Bachi, R., and Baron, R. 1969. Confidentiality problems related to data banks. Bulletin of the International Statistical Institute 43:225-241.

Bhatia, R. 1987. Perturbation Bounds for Matrix Eigenvalues. John Wiley, New York.

Cox, L. H., Johnson, B., McDonald, S., Nelson, D., and Vazquez, V. 1985. Confidentiality issues at the Census Bureau. Paper presented at the First Annual Research Conference of the Bureau of the Census, Washington, D.C.

Duncan, G. T., and Lambert, D. 1986. Disclosure -limited data dissemination. Journal of the American Statistical Association 81:10-28.

Fellegi, I. P. 1975. Controlled random rounding. Survey Methodology 1:123-135.

Gates, G. W. 1988. Census Bureau microdata: Providing useful research data while protecting the anonymity of respondents. Paper presented at the annual meeting of the American Statistical Association, New Orleans, Louisiana.

Kim, J. 1986. A method for limiting disclosure of microdata based on random noise and transformation. In Proceedings of the Section on Survey Research Methods of the American Statistical Association, 370-374.

McGuckin, R. H., and Nguyen, S. V. 1988. Public use microdata: Disclosure and usefulness. Discussion paper, Center for Economic Studies, Bureau of the Census.

Mugge, R. H. 1983. Issues in protecting confidentiality in national health statistics. In Proceedings of the Section on Survey Research Methods of the American Statistical Association, 592-594.

Paass, G. 1985. Disclosure risk and disclosure avoidance for microdata. Sankt Augustin, Federal Republic of Germany.

Spruill, N. L. 1983. The confidentiality and analytic usefulness of masked business microdata. In Proceedings of the Section on Survey Research Methods of the American Statistical Association, 602-607.

Steinberg, J., and Pritzker, L. 1967. Some experiences with and reflections on data linkage in the United States. Bulletin of the International Statistical Institute, 786-808.

Wolf, M. K. 1988. Microaggregation and disclosure avoidance for economic establishment data. Paper presented at the annual meeting of the American Statistical Association, New Orleans, Louisiana.