Journal of Official Statistics, Vol. 22, No. 3, 2006, pp. 487–506

Outlier Detection and Editing Procedures for Continuous Multivariate Data

B. Ghosh-Dastidar[1] *and J.L. Schafer*[2]

In large datasets, outliers may be difficult to find using informal inspection and graphical displays, particularly when there are missing values. We present a semi-automatic method of outlier detection for continuous, multivariate survey data that is designed to identify outlying cases and suggest potential errors on a case-by-case basis, in the presence of missing data. Our method relies on an explicit probability model for the data. The raw data with outliers is described by a contaminated multivariate normal distribution, and an EM algorithm is applied to obtain robust estimates of the means and covariances in the presence of missing values. Mahalanobis distances are computed to identify potential outliers and offending variables. The procedure is implemented in a software product, which detects outliers and suggests edits to remove offending values. We apply the algorithm to preliminary body-measurement data from the Third National Health and Nutrition Examination Survey, Phase I (1988–1991). This method works quite generally for continuous survey data, and is particularly useful when inter-variable correlations are strong.

Key words: Contaminated normal; EM algorithm; NHANES III; outliers; posterior probability.

1. Introduction

1.1. What Is an Outlier?

Outliers are observations that deviate from the specified data model. Generally speaking, outliers are all of those observations that appear to be extreme or unusual with respect to the rest of the observed data, and to prior subject-matter knowledge about what values are plausible. The presence of outliers may indicate that the data model does not have sufficiently heavy tails, or that there are misreporting and misrecording errors in the data. Outliers, for either reason, may exert undue influence on the results of statistical analyses, so they need to be identified prior to performing data analyses.

When we encounter a potential outlier, our first suspicion is that the observation resulted from a mistake or extraneous effect, and therefore should be discarded. However, if the outlier is an extreme value, it may be conveying important information about the underlying population of actual values. Thus, nonjudicious removal of observations that appear to be outliers may result in underestimation of the uncertainty present in the data.

[1] RAND, Arlington, VA 22202-5050, U.S.A. Email: bonnieg@rand.org
[2] Department of Statistics, The Pennsylvania State University, University, Park, PA 16802, U.S.A. Funding support was provided by Cooperative Agreement 43-3AEU-3-80087 between the National Agricultural Statistics Service, U.S.D.A. and Penn State, and by RAND Methods Funding.

As a consequence, estimated standard errors and p-values will be smaller than they should be, leading to potentially false findings of significance. In this article, we demonstrate how potential outliers may be identified, but we do not say what should be done with them. Caution and good judgment should be exercised when rendering decisions about whether outliers should be removed.

1.2. Informal Methods of Outlier Detection and Editing

Univariate displays such as histograms, boxplots and dot diagrams may be used to inspect a dataset one variable at a time. When examining univariate distributions, we can flag all observations beyond some range of plausibility as outliers. Figure 1 presents a dot diagram of a single variable X_1 with one outlier, which is much larger than the other observations, indicated by the bold point. Univariate techniques are useful, but a data point that passes all univariate tests may still be an outlier if it violates plausible relationships among variables. Consider the situation of Figure 2, which shows a scatterplot of two variables X_1 and X_2. The bold point is clearly an outlier because it lies outside the cloud of other points. However, each variable for this observation lies within its range of plausible values, making it impossible to detect the point by univariate methods alone.

Bivariate graphical displays can reveal outliers like the one seen in Figure 2. But bivariate plots are limited in the sense that they will not necessarily help identify, for any outlying case, which of the two variables is more likely to be erroneous (if indeed one is erroneous); this type of judgment must involve additional covariates and thus requires analysis in three or more dimensions. Moreover, bivariate analyses of a multivariate dataset can be tedious. A thorough inspection of multivariate data may require the construction of all possible bivariate plots. Each outlier may show up in multiple plots, making it difficult to match points in different plots. Finally, when the multivariate dataset contains missing values, outliers may not show up on bivariate scatterplots because standard plotting routines will typically omit an observation if either of the variables is missing.

Survey agencies often adopt data editing procedures to "clean up" their data by removing gross errors and then analyze the edited data as if it were the truth (e.g., Granquist and Kovar 1997). The edit procedures typically include data-specific edit rules developed by subject-matter specialists to determine whether an observation is reliable

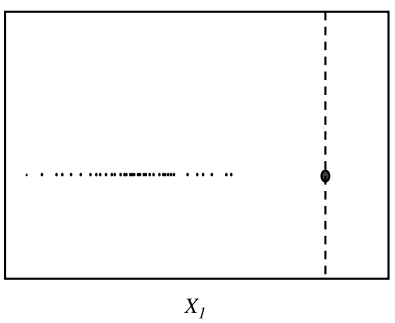

Fig. 1. Extreme value in a marginal distribution

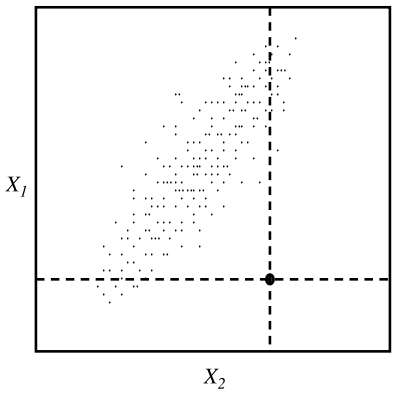

Fig. 2. Extreme value in a bivariate plot

(Barcaroli and Venturi 1993; Thompson and Sigman 1999), followed by deletion and imputation. Automated systems such as the U.S. Census Bureau's Structured Program for Economic Editing and Referrals (Winkler and Draper 1997) and Statistics Canada's Generalized Edit and Imputation System (Kovar and Whitridge 1990) also exist. These rely on the editing principles proposed by Fellegi and Holt (1976). While these methods work well, their approach is "rule based" and strictly deterministic while a more statistically motivated approach ought to have a probabilistic component.

1.3. Motivating Example: NHANES III

The motivating example for this article is the Third National Health and Nutrition Examination Survey (NHANES III), conducted by the National Center for Health Statistics (1994), designed to collect information about the health and diet of people in the United States. Multivariate surveys such as NHANES III are subject to nonresponse and outliers. Initial exploration of the preliminary Phase I (1988–1991) data by Ezzati-Rice, Khare, and Schafer (1993) revealed a substantial number of outliers, particularly in the body-measurement variables. In NHANES III, medical staff obtained a variety of physical measurements (e.g., height, weight, waist and hip circumference, skinfolds). When the standard protocol was followed, these characteristics were measured accurately with negligible error. Occasionally, mistakes or deviations from the intended protocol introduced gross errors of large magnitude in one or more of the variables. Ezzati-Rice et al. applied univariate and bivariate plots to edit the dataset and found that the informal methods were inadequate.

The NHANES III exercise motivated our search for a fast, reliable approach to outlier detection in multivariate datasets. In general, outliers in survey data require a multivariate detection approach. Further, although detecting outliers by visual inspection of scatterplots is straightforward, identifying which variables are at fault is hard. We believe that our procedure is an efficient and sound approach to identifying outliers, which also flags potentially erroneous variables. It is quite general and easily applied to other continuous, multivariate data with strong inter-variable correlations.

1.4. The Proposed Method

We propose a semi-automatic method for outlier detection and editing in continuous multivariate survey data. The technique discussed here relies on an explicit probability model for the data. First, the raw data with outliers and missing values are described by a contaminated multivariate normal distribution. This contaminated multivariate normal is a mixture of two multivariate normal distributions with the same mean but different covariance matrices, one proportionately larger than the other. The uncontaminated population is described by the distribution with smaller covariances, and the outliers by the distribution with larger covariances. The mixing probability describes the proportion of observations expected to be outliers.

Maximum likelihood estimation of parameters requires maximizing the loglikelihood function. In many statistical problems, this is done by setting first derivatives equal to zero, i.e., $l'(\theta|X) = 0$; for our problem, however, the solution to $l'(\theta|X) = 0$ does not exist in closed form. Moreover, gradient methods such as Newton-Raphson are difficult to apply because the second derivative of the loglikelihood is a complicated function of the elements of θ. The situation is complicated even more by the presence of missing data. To maximize the loglikelihood, we apply the version of the EM algorithm (Dempster, Laird, and Rubin 1977) described in Little and Rubin (1987).

This procedure yields robust estimates of the unknown model parameters, namely the means and covariances of the theoretical uncontaminated population. Mahalanobis distances (Marcoulides and Hershberger 1997, p. 105) relative from the center of the uncontaminated population are then calculated for all observations. Points whose distances exceed a predetermined cutoff are flagged as possible outliers. Then, for each potential outlier, Mahalanobis distances of subvectors are computed to identify one or more offending variables and suggest plausible edits. Although we discuss how to identify potential outliers only under the contaminated multivariate normal, other plausible probability models for outliers are discussed. In addition, whether an outlier should be deleted or not is a subjective decision and should be made by subject-matter experts with knowledge of the data collection process. Thus, the method discussed here is semi-automatic.

1.5. Scope of the Rest of the Article

Section 2 discusses the implementation of the outlier detection and editing method. The contaminated multivariate normal model is presented, along with EM algorithms for parameter estimation with and without missing data, and strategies for identifying the outliers and suggesting plausible edits. Section 3 illustrates the performance of this method when applied to a subset of NHANES III.

2. Methodology

2.1. Contaminated Multivariate Normal Distribution

The probability model chosen to describe the observed data is a contaminated multivariate normal distribution. This distribution is a mixture of two multivariate normals centered at

the same mean but with different covariance matrices, one being proportionately larger than the other. Let $\{\boldsymbol{x}_i : i = 1, \ldots, n\}$ be a random sample of values subject to contamination, where each $\boldsymbol{x}_i$ is a vector of length k. We assume that $\boldsymbol{x}_i \sim N_k(\boldsymbol{\mu}, \Psi/\lambda)$ with probability δ and $\boldsymbol{x}_i \sim N_k(\boldsymbol{\mu}, \Psi)$ with probability $(1 - \delta)$, where λ is a positive scalar less than 1. The parameters $\boldsymbol{\mu}$ and Ψ are unknown, whereas δ and λ are assumed known. The quantity δ specifies what proportion of the observations are contaminated. The quantity λ is the variance inflation factor that indicates the magnitude of the errors leading to contamination; for example, if $\lambda = .5$ the contamination is regarded as inflating the variances by a factor of 2. In practice, it may be of interest to explore a grid of plausible values of δ and λ, as the values of these parameters may affect determinations about outliers.

The contaminated multivariate normal is not the only model that one might consider to describe a dataset with outliers. For example, the multivariate t-distribution with few degrees of freedom can also be used to model heavy-tailed datasets. The multivariate t, however, generates a continuum of unusual values rather than only a few erratic observations and thus seems less suitable for modeling data containing gross measurement or recording errors. The contaminated multivariate normal, on the other hand, allows for a modest number of gross errors. Moreover, its parameters $\boldsymbol{\mu}$ and Ψ have an attractive interpretation as the moments of the uncontaminated component of the population.

Let $\boldsymbol{x}_i$ be an observation from a dataset with k variables expressed as a $(1 \times k)$ vector. Under the contaminated normal model, the probability density function of $\boldsymbol{x}_i$ is

$$p(\boldsymbol{x}_i|\theta) = (1 - \delta)|2\Pi\Psi|^{-\frac{1}{2}} \exp\{ -(\boldsymbol{x}_i - \boldsymbol{\mu})\Psi^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})^T/2\} + \delta|(2\Pi\Psi)/\lambda|^{-\frac{1}{2}} \exp\{ -\lambda(\boldsymbol{x}_i - \boldsymbol{\mu})\Psi^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})^T/2\} \quad (1)$$

where $\theta = (\boldsymbol{\mu}, \Psi)$ is unknown, while δ and λ are regarded as fixed and known. Let $d_i^2 = (\boldsymbol{x}_i - \boldsymbol{\mu})\Psi^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})^T$ denote the squared Mahalanobis distance from $\boldsymbol{x}_i$ to the mean $\boldsymbol{\mu}$ with respect to Ψ, the covariance matrix of the uncontaminated population. The probability density function can be written in terms of the d_i^2 as

$$p(\boldsymbol{x}_i|\theta) = (2\Pi)^{-\frac{k}{2}}|\Psi|^{-\frac{1}{2}}\left\{(1 - \delta)\exp(-d_i^2/2) + \delta\lambda^{\frac{k}{2}}\exp(-\lambda d_i^2/2)\right\} \quad (2)$$

Consider n independent, identically distributed (i.i.d.) observations from a k-variate contaminated normal distribution, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. The loglikelihood function of θ given X is

$$l(\theta|X) = -\frac{nk}{2}\log(2\Pi) - \frac{n}{2}\log|\Psi| + \sum_{i=1}^{n}\log\left\{(1 - \delta)\exp(-d_i^2/2) + \delta\lambda^{\frac{k}{2}}\exp(-\lambda d_i^2/2)\right\} \quad (3)$$

2.2. Augmenting the Data

Maximum likelihood (ML) estimation of parameters requires maximizing the loglikelihood function. As described in Section 1.4, the loglikelihood (3) is difficult to maximize by gradient methods. We simplify the problem by augmenting the data–i.e., by introducing an imaginary unobserved variable into the dataset which, if it were seen, would lead to ML estimates in closed form. Associate with each observation $\boldsymbol{x}_i$ a dichotomous variable q_i indicating whether or not the $\boldsymbol{x}_i$ comes from the uncontaminated component of the population; that is, $q_i = \lambda$ if $\boldsymbol{x}_i$ is distributed as $N_k(\boldsymbol{\mu}, \Psi/\lambda)$ and $q_i = 1$ if $\boldsymbol{x}_i$ is distributed as $N_k(\boldsymbol{\mu}, \Psi)$. Each q_i takes the values λ and 1 with probabilities δ and $(1 - \delta)$, respectively. The marginal distribution of q_i is $p(q_i = \lambda) = \delta$, and $p(q_i = 1) = 1 - \delta$. The conditional distribution of $\boldsymbol{x}_i$ given q_i is then

$$\boldsymbol{x}_i|\theta, q_i \sim N_k(\boldsymbol{\mu}, \Psi/q_i) \tag{4}$$

We will use the term "observed data" to refer to X alone, and "augmented data" to refer to both X and $Q = (q_1, \ldots, q_n)$. The augmented-data loglikelihood function–i.e., the loglikelihood function that we would get if Q was observed–is

$$\begin{aligned} l(\theta|X, Q) = &- \frac{nk}{2} \log(2\Pi) - \frac{n}{2} \log|\Psi| + \frac{k}{2} \sum_{i=1}^{n} \log(q_i) \\ &- \frac{1}{2} \sum_{i=1}^{n} [(\boldsymbol{x}_i - \boldsymbol{\mu}) \Psi^{-1} q_i (\boldsymbol{x}_i - \boldsymbol{\mu})^T] \end{aligned} \tag{5}$$

This loglikelihood function is a linear function of the following augmented-data sufficient statistics:

$$\begin{aligned} S_0 &= \sum_{i=1}^{n} q_i \\ S_1 &= \sum_{i=1}^{n} q_i \boldsymbol{x}_i \\ S_2 &= \sum_{i=1}^{n} q_i \boldsymbol{x}_i^T \boldsymbol{x}_i \end{aligned} \tag{6}$$

If X and Q are observed, ML estimates of the parameters $\boldsymbol{\mu}$ and Ψ can be found by a weighted least squares method in which the observations with larger variances are automatically downweighted. In our model, the covariance matrix of $\boldsymbol{x}_i$ given q_i is Ψ/q_i where q_i equals λ or 1, so if q_i were known we would apply a weight to $\boldsymbol{x}_i$ proportional to

q_i. The weighted estimates are

$$\hat{\boldsymbol{\mu}} = \sum_{i=1}^{n} q_i \boldsymbol{x}_i / \sum_{i=1}^{n} q_i = S_1/S_0$$

$$\hat{\boldsymbol{\Psi}} = \sum_{i=1}^{n} \frac{q_i(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})}{n} = \frac{S_2 - S_1^T S_1/S_0}{n} \tag{7}$$

2.3. Basic EM Algorithm

Without observing Q, we cannot use (7) to calculate ML estimates for the parameters. Rather, we apply the EM algorithm, a general iterative procedure for ML estimation in missing data problems. In EM, we start with an initial guess of the unknown parameters, and then iteratively perform the following two steps.

- E-step: Replace the sufficient statistics in the augmented-data loglikelihood function with their conditional expectations given the observed data and current parameter estimates.
- M-step: Calculate new parameter estimates based on the augmented data–that is, maximize the loglikelihood function obtained as a result of the E-step.

The augmented-data sufficient statistics S_0, S_1 and S_2 are given by (6). When the data matrix X is fully observed, the $(t+1)$st iteration of EM proceeds as follows.

- E-Step: Estimate S_0, S_1 and S_2 by their conditional expectations given X and $(\boldsymbol{\mu}^{(t)}, \Psi^{(t)})$, the parameter estimates from the previous iteration. S_0, S_1 and S_2 are linear functions of the q_i's (7). Therefore, the E-step reduces to finding the conditional expectation of q_i, which is w_i or the observation *weight*,

$$w_i^{(t)} = E(q_i | \boldsymbol{x}_i, \boldsymbol{\mu}^{(t)}, \Psi^{(t)}) \tag{8}$$

For the contaminated normal model, it can be shown that

$$w_i^{(t)} = \frac{1 - \delta + \delta\lambda^{k/2+1} \exp\{(1-\lambda)d_i^{2(t)}/2\}}{1 - \delta + \delta\lambda^{k/2} \exp\{(1-\lambda)d_i^{2(t)}/2\}} \tag{9}$$

(Little and Rubin 1987, p. 212). The calculation of d_i^2 at the tth step uses $(\boldsymbol{\mu}^{(t)}, \Psi^{(t)})$ in place of $\boldsymbol{\mu}$ and Ψ.

- M-Step: Compute new estimates $(\boldsymbol{\mu}^{(t+1)}, \Psi^{(t+1)})$ as in (7), replacing the sufficient statistics q_i with their expected values $w_i^{(t)}$ from the E-step.

The estimates of $\boldsymbol{\mu}$ and Ψ obtained through this procedure are more robust than the usual estimates, because potential outliers have large values of d_i^2, which are downweighted proportional to their weight w_i. This algorithm can be regarded as a special case of iteratively reweighted least squares (Rubin 1983).

2.4. Modifications to EM for Missing Data

The algorithm of Section 2.3 can be easily extended to situations where the data matrix X contains missing values. Suppose we partition X as (X_{obs}, X_{mis}), where X_{obs} and X_{mis} denote the observed and missing parts of X, respectively. The augmented data will consist of $X = (X_{obs}, X_{mis})$ and $Q = (q_1, \ldots, q_n)$, of which only X_{obs} is observed; X_{mis} and Q are missing. For each observation $\boldsymbol{x}_i$, let $\boldsymbol{x}_{obs,i}$ and $\boldsymbol{x}_{mis,i}$ denote the observed and missing portions, respectively, so that $X_{obs} = \{\boldsymbol{x}_{obs,i} : i = 1, \ldots, n\}$ and $X_{mis} = \{\boldsymbol{x}_{mis,i} : i = 1, \ldots, n\}$.

Let us assume that the missing data are missing at random (MAR) so that the missing-data mechanism does not depend on X_{mis} (Rubin 1976, p. 582). Under a MAR assumption, ML estimates of $\boldsymbol{\mu}$ and Ψ can be computed by applying a modified EM algorithm which treats both X_{mis} and Q as missing. The E-step requires a few modifications from the previous section because portions of $\boldsymbol{x}_i$ in the sufficient statistics S_0, S_1, and S_2 are now missing. The M-step, however, remains unchanged because the augmented-data loglikelihood (5) is the same function as before.

The $(t+1)$st iteration of the modified EM algorithm proceeds as follows (Little and Rubin 1987, pp. 212–213).

- E-Step: Estimate S_0, S_1 and S_2 by their conditional expectations given X_{obs} and $\theta^{(t)} = (\boldsymbol{\mu}^{(t)}, \Psi^{(t)})$, the parameter estimates from the previous iteration. The conditional expectations are

$$E(S_0|X_{obs}, \theta^{(t)}) = E\left(\sum_{i=1}^{n} q_i|\boldsymbol{x}_{obs,i}, \theta^{(t)}\right) \quad = \sum_{i=1}^{n} E(q_i|x_{obs,i}, \theta^{(t)}) \quad = \sum_{i=1}^{n} w_i^{(t)} \tag{10}$$

The weights $w_i^{(t)}$ are a simple modification of (9), calculated as follows:

(i) Replace k by k_i, the length of $\boldsymbol{x}_{obs,i}$
(ii) Compute the squared distances d_i^2 using $\boldsymbol{x}_{obs,i}$, and $\hat{\boldsymbol{\mu}}_{obs}$ and $\hat{\Psi}_{obs,obs}$, the portions of the estimated mean vector and covariance matrix corresponding to the observed variables in $\boldsymbol{x}_{obs,i}$

$$d_i^2 = (\boldsymbol{x}_{obs,i} - \hat{\boldsymbol{\mu}}_{obs})\hat{\Psi}_{obs,obs}^{-1}(x_{obs,i} - \hat{\boldsymbol{\mu}}_{obs})^T \tag{11}$$

The calculation of d_i^2 at the tth step uses $(\boldsymbol{\mu}^{(t)}, \Psi^{(t)})$ in place of $\boldsymbol{\mu}$ and Ψ.

The jth component of $E(S_1|X_{obs}, \theta^{(t)})$ is

$$E\left(\sum_{i=1}^{n} q_i x_{ij}|X_{obs}, \theta^{(t)}\right) = \sum_{i=1}^{n} E\{q_i E(x_{ij}|\boldsymbol{x}_{obs,i}, \theta^{(t)}, q_i)|\boldsymbol{x}_{obs,i}, \theta^{(t)}\} = \sum_{i=1}^{n} w_i^{(t)} \hat{x}_{ij}^{(t)} \tag{12}$$

where $\hat{x}_{ij}^{(t)} = E(x_{ij}|\boldsymbol{x}_{obs,i}, \theta^{(t)})$, because the conditional mean of x_{ij} given $\boldsymbol{x}_{obs,i}$, $\theta^{(t)}$ and q_i

does not depend on q_i. Finally, the (j,k)th element of $E(S_2|X_{obs}, \theta^{(t)})$ is

$$E\left(\sum_{i=1}^{n} q_i x_{ij} x_{ik} | X_{obs}, \theta^{(t)}\right) = \sum_{i=1}^{n} E\{q_i E(x_{ij} x_{ik} | \boldsymbol{x}_{obs,i}, \theta^{(t)}, q_i) | \boldsymbol{x}_{obs,i}, \theta^{(t)}\}$$

$$= \sum_{i=1}^{n} (w_i^{(t)} \hat{x}_{ij}^{(t)} \hat{x}_{ik}^{(t)} + \Psi_{jk.obs,i}^{(t)}) \tag{13}$$

where

$$\Psi_{jk.obs,i}^{(t)} = \begin{cases} 0 & \text{if } x_{ij} \text{ or } x_{ik} \\ & \text{are observed} \\ q_i \operatorname{Cov}(x_{ij}, x_{ik} | \boldsymbol{x}_{obs,i}) & \text{if } x_{ij} \text{ or } x_{ik} \\ & \text{are both missing} \end{cases} \tag{14}$$

The quantities $\hat{x}_{ij}^{(t)}$ and $\Psi_{jk.obs,i}^{(t)}$ are the means and covariances, respectively, of the conditional distribution of $\boldsymbol{x}_{mis,i}$ given $\boldsymbol{x}_{obs,i}$. Note that because the joint distribution of $x = (\boldsymbol{x}_{mis,i}, \boldsymbol{x}_{obs,i})$ is normal, the conditional distribution of $\boldsymbol{x}_{mis,i}$ given $\boldsymbol{x}_{obs,i}$ is also normal. Thus, the conditional means and covariances come from a multivariate regression of the response $\boldsymbol{x}_{mis,i}$ on the predictors $\boldsymbol{x}_{obs,i}$. These parameters may be computed from the assumed values of $\boldsymbol{\mu}$ and Ψ by applying the sweep operator, as described by Little and Rubin (1987, pp. 112–119).

- M-Step: Compute new estimates $(\boldsymbol{\mu}^{(t+1)}, \Psi^{(t+1)})$ as in (7), replacing the sufficient statistics q_i with their expected values $w_i^{(t)}$ from the E-step. This step remains unmodified from Section 2.3.

2.5. *Identifying Potential Outliers and Erroneous Values*

The diagnostic tools we use to identify possible outliers are *posterior probabilities* and *weights*. The *posterior probability* of an observation originating from the contaminated population conditional upon the observed data and parameter estimates is given by

$$\delta_i^* = P(q_i = \lambda | \boldsymbol{x}_{obs,i}, \hat{\boldsymbol{\mu}}_{obs}, \hat{\Psi}_{obs,obs}) \tag{15}$$

and $P(q_i = 1 | \boldsymbol{x}_{obs,i}, \hat{\mu}_{obs}, \hat{\Psi}_{obs,obs}) = (1 - \delta_i^*)$. This quantity is estimated in the E-step of the EM algorithm because it goes into the computation of the weights (9). We will select a cutoff value δ^c and consider observations with δ_i^* exceeding δ^c as potential outliers. The most intuitive choice for the cutoff is .5. That is, if δ_i^* exceeds .5, the observation is more likely to have come from the contaminated population than the noncontaminated population. Thus, it is a potential outlier. Therefore we will set δ^c to .5 for our applications.

The *weight* of observation i is the conditional expectation of q_i given the observed data and parameter estimates (8). We will refer to the final weight obtained from the converged parameter estimates as $w_i^* = E(q_i | \boldsymbol{x}_{obs,i}, \hat{\boldsymbol{\mu}}_{obs}, \hat{\Psi}_{obs,obs})$. From the marginal density of q_i given in Section 2.2, we see that w_i^* must lie between λ and 1 because it is a weighted

average of the two values. Also, w_i^* is a decreasing function of d_i^2 (9); therefore, observations that are far from the observed mean $\boldsymbol{\mu}_{obs}$ or potential outliers will tend to have small weights. Thus, given an appropriate cutoff value w^c in the lower tail, any observation i for which $w_i^* < w^c$ will be flagged as a potential outlier.

While the choice of w^c may seem arbitrary, we will exploit the one-to-one relationship between weights and posterior probabilities to derive a meaningful w^c. Using the marginal density of q_i, (8) and (15), we see that

$$w_i^* = \lambda \times P(q_i = \lambda | \boldsymbol{x}_{obs,i}, \hat{\boldsymbol{\mu}}_{obs}, \hat{\boldsymbol{\Psi}}_{obs,obs}) + P(q_i = 1 | \boldsymbol{x}_{obs,i}, \hat{\mu}_{obs}, \hat{\boldsymbol{\Psi}}_{obs,obs}) \tag{16}$$

and thus

$$w_i^* = 1 + (1 - \lambda)\delta_i \tag{17}$$

Therefore, we can plug in $\delta^c = .5$ and the assumed value of λ from the data model to obtain the corresponding cutoff value w^c.

The one-to-one relationship between the weight and posterior probability of an observation implies that the two methods will produce the same results. After identifying potential outliers using either criterion, we will identify influential variables that may cause the observation to be outlying so that we can suggest plausible edits on a case-by-case basis. Influential variables are the ones that make the largest contributions toward the squared distance. There may be one or more such variables in each outlying case. First, we discuss the situation when there is only one influential variable (Little and Smith 1987).

(i) For each outlying case i and observed variable j, compute $d_i^{(j)2}$, the squared distance with variable j omitted. This distance is based on $\boldsymbol{x}_{obs,i}$ with variable j omitted (11).
(ii) Find j_1 such that $d_i^{(j_1)2} < d_i^{(j)2}$ amongst all observed j for case i. Thus, variable j_1 is the most influential variable in the ith case.

Therefore, the suggested delete for outlying case i is variable j_1.

The Mahalanobis distance is asymptotically distributed as $d_i^2 \sim \chi_{k_i}^2$ (Mardia, Kent, and Bibby 1979). Thus, we will compare the $d_i^{(j_1)2}$ against the chi-squared reference statistic with degrees of freedom equal to the number of observed variables and $\alpha = .05$. If the p-value for $d_i^{(j_1)2}$ is significant, there must be more than one offending variable. A procedure to identify the next m deletes after removing j_1 from $\boldsymbol{x}_{obs,i}$ is:

(i) For each outlying case i and every possible combination of m remaining observed variables, compute $d_i^{(j_1, j_{k_1}, \ldots, j_{k_m})2}$, the squared distance with variables $j_1, j_{k_1}, \ldots, j_{k_m}$ omitted.
(ii) Find $j_2, \ldots, j_{m+1}$ such that $d_i^{(j_1, j_2, \ldots, j_{m+1})2} < d_i^{(j_1, j_{k_1}, \ldots, j_{k_m})2}$. Variables $j_2, \ldots, j_{m+1}$ are the m most influential variables in the ith case.

Variables $j_1, j_2, \ldots, j_{m+1}$ are then the suggested deletes for outlying case i. Perform the above process for $m = 1, 2, 3, \ldots$ until the p-value of the squared distance with the influential variables omitted is no longer significant.

3. Application

3.1. Data

In this section, we illustrate the method proposed with raw body measurements from NHANES III preliminary Phase I, made available by the NCHS. These data were collected during physical examinations of 1,262 children 2–3 years of age. The variables considered here are self-reported height (SM5), standing height (HT), sitting height (SITHT), recumbent length (RECUM), head (HEADC), waist (WAIST) and buttocks (BUTTO) circumference, all of which are continuous and exhibit moderate to strong correlations with one another. The only variable that may require an explanation is RECUM, which refers to a body-length measurement taken while the child is lying down. The unit of measurement is centimeter for all of these variables. Preliminary graphs (Figure 3) such as histograms and normal quantile plots showed that most of the variables are approximately normally distributed. Although a few of the variables had a slight skew, we left them untransformed because the standard transformations did not help. Bivariate scatterplots (Figure 4) indicated moderate to high inter-variable correlations amongst the body-measurement variables and possible outliers.

3.2. Selection of Parameter and Cutoff Values

This method assumes that the contamination parameters δ and λ are known, while the parameters of the normal distribution, μ and Ψ, are estimated. Reasonable choices of δ and λ can be made from preliminary exploration of the data including tabulations, univariate and bivariate scatter plots and from consultations with subject-matter experts. For example, a bivariate plot of our data (Figure 4) suggested that the proportion of outlying values is low while the variance of the gross errors is roughly twice that of the non-outlying observations. We picked the values that struck an optimal balance between identifying all of the gross errors but not too many of the data points that were part of the regular data cloud. We found that a mixing proportion of .04 and a variance inflation factor of .5 balanced these two objectives the best in the context of this application. For purposes of comparison, we ran our algorithm with these values and also three other pairs of parameter values and displayed the observations flagged by each set of parameter values using the *identify()* function in S-Plus. In the absence of sufficient information about the data, it is advisable to iterate through several pairs of plausible parameter values, make graphical displays and see which values work well. While selecting δ and λ is not an exact science, we found that similar reasonable values produced similar results.

We also required cutoff values δ^c and w^c to identify outliers. We will use a cutoff value of .5 for δ^c because it is intuitive to flag observations that are more likely to originate from the contaminated normal distribution as outliers. Using the one-to-one relationship between δ^c and w^c (17), we plugged in $\lambda = .5$ and $\delta^c = .5$ to get the corresponding cutoff $w^c = .75$. If desired, one can vary the cutoff values using the distributions of δ_i^* and w_i^* (Figure 5 and 6) and then inspect all of the observations identified as outliers on a case-by-case basis.

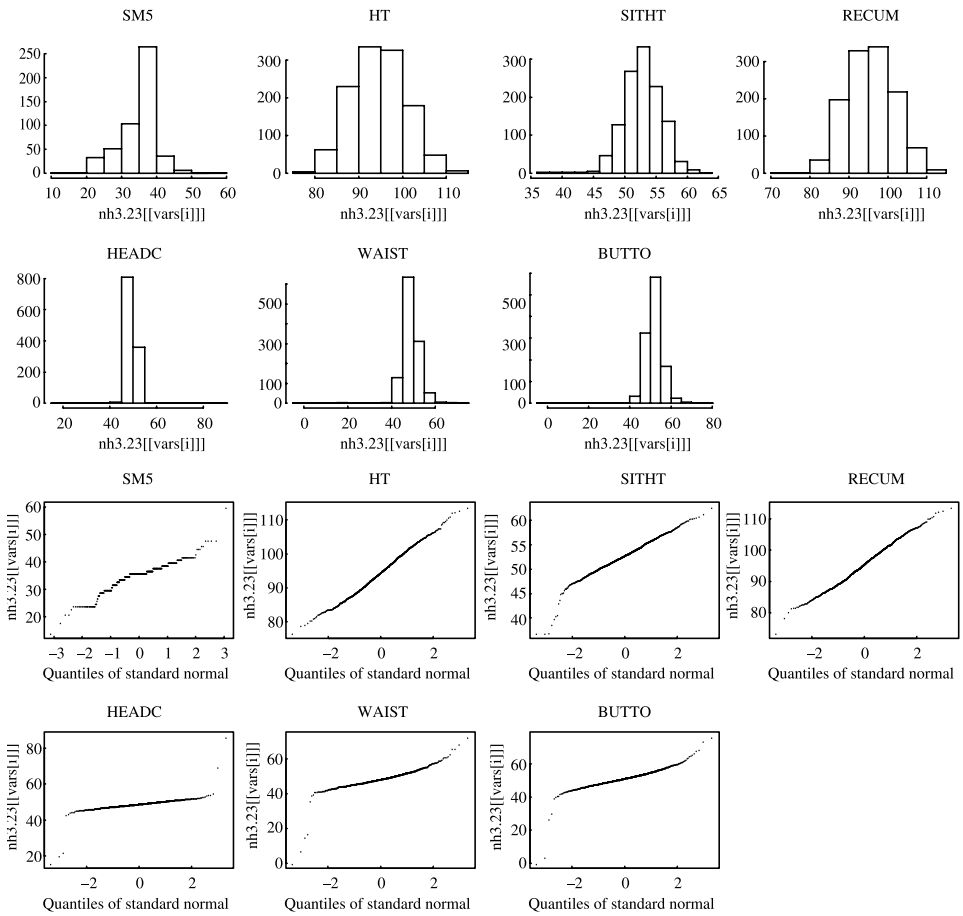

Fig. 3. Marginal distributions of body-measurement variables: SM5 = survey-reported height, HT = standing height, SITHT = sitting height, RECUM = recumbent length, HEADC = head circumference, WAIST = waist circumference and BUTTO = buttocks circumference

3.3. Results

The EM algorithm for this model converged in 20 iterations producing the robust estimates of μ and Σ displayed in Table 2. The estimated squared Mahalanobis distance, weight and posterior probability of contamination of each observation were calculated from these results. The d_i^2's ranged from 0 to 422.6; the w_i^*'s ranged from .5 to .998; the δ_i^*'s were between .004 and 1. The cutoffs of $\delta^c = .5$ and $w^c = .75$ flagged about 3% of the observations as potential outliers. The contaminated normal model downweights extreme observations so that their weights are close to the lower bound λ. Thus, in Figure 5, the potential outliers appear at the lower end of the histogram. On the other hand, outlying observations will have large posterior probabilities of contamination and are found in the upper tail of the histogram in Figure 6. Figure 7 and 8 show the distribution of the posterior probabilities and weights, respectively, for only those observations identified as outliers by w^c and δ^c.

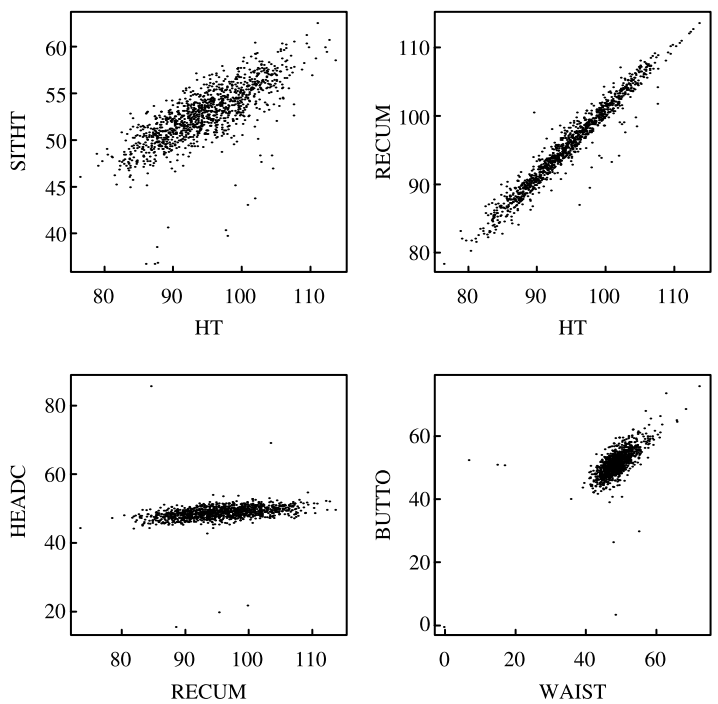

Fig. 4. Scatterplots of body-measurement variables: SM5 = survey-reported height, HT = standing height, SITHT = sitting height, RECUM = recumbent length, HEADC = head circumference, WAIST = waist circumference and BUTTO = buttocks circumference

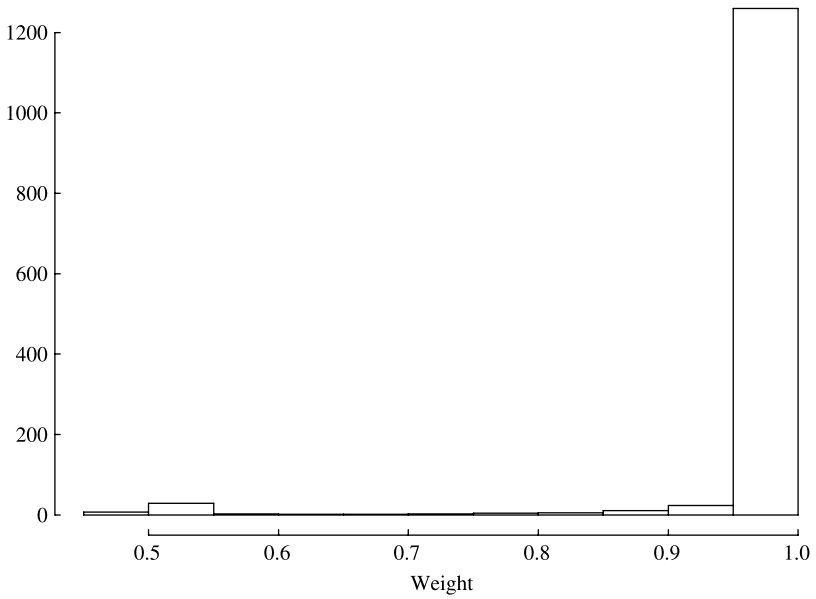

Fig. 5. Histogram of w_i^* for all observations

Table 1. Distribution of the number of outliers with $\delta =$ Proportion of contamination and $\lambda =$ Variance inflation factor

λ	δ		
	.01	.04	.10
.50	38	42	57
.25	42	51	67
.01	36	39	41

A select few potential outliers and their suggested edits are shown in Table 3. The first column gives an unidentifiable sample observation number, while the second column specifies the Mahalanobis distance for that observation before any variables are removed. The third column *Df* indicates the number of observed variables while the column labelled *Edits* lists the likely outlying values. For example, Case 38 has a *Df* of 7, which means that all seven variables in this example were observed. The corresponding d_i^2 of 23.0 produces a significant p-value of .002, with a suggested edit of SITHT. The p-value increases to a nonsignificant value of .07 after SITHT is deleted. Therefore, it would appear that SITHT was indeed erroneous. Now consider observations 53 and 72, which are equidistant from the center with the same number of observed values. The former seems to have more than one erroneous value because the p-value is still smaller than .05 after removing HT, while the latter has only one influential variable BUTTO. Another example is case 436, with very little change in its d_i^2 upon editing. It is most likely that this observation has several outlying variables, therefore further edits should be considered.

Table 2. Parameter estimates from EM algorithm for contaminated multivariate normal distribution with $\delta = .04$ and $\lambda = .5$

Means

sm5	ht	sitht	recum	headc	waist	butto
35.3	94.7	52.9	95.7	49.2	48.8	51.8

Standard deviations

sm5	ht	sitht	recum	headc	waist	butto
4.67	6.08	2.93	6.05	2.04	3.85	4.07

Correlation matrix

sm5	ht	sitht	recum	headc	waist	butto
1.0	.526	.480	.517	.247	.258	.351
	1.0	.787	.979	.352	.447	.535
		1.0	.797	.366	.479	.572
			1.0	.353	.464	.549
				1.0	.269	.273
					1.0	.689
						1.0

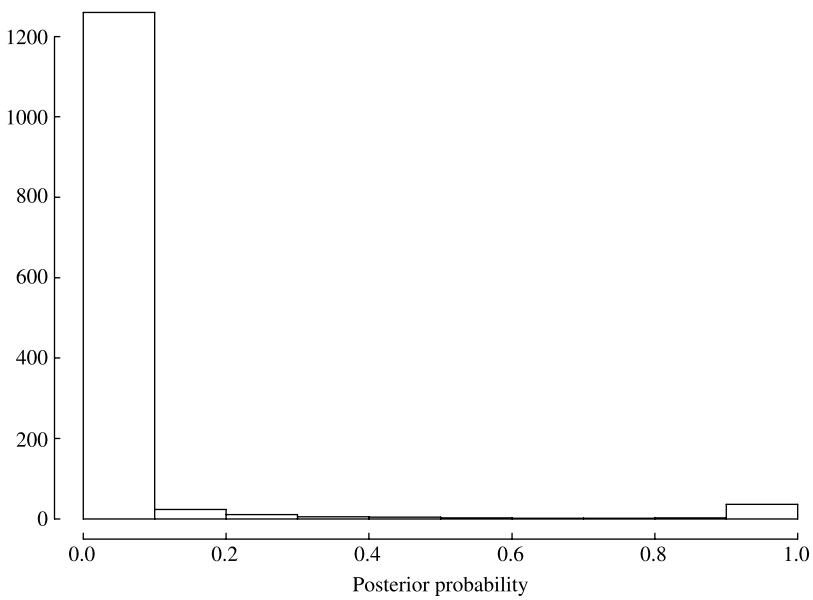

Fig. 6. Histogram of δ_i^* for all observations

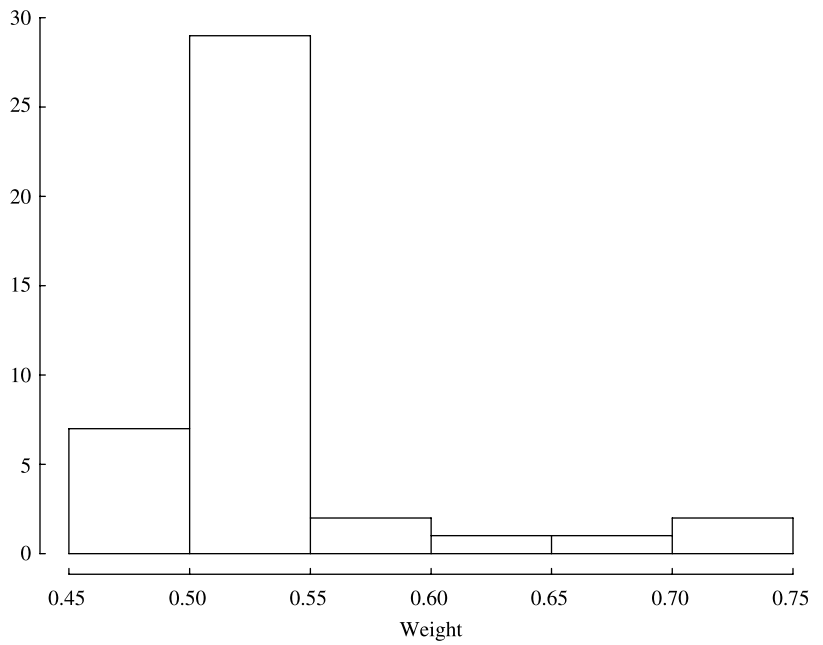

Fig. 7. Histogram of w_i^* for potential outliers only

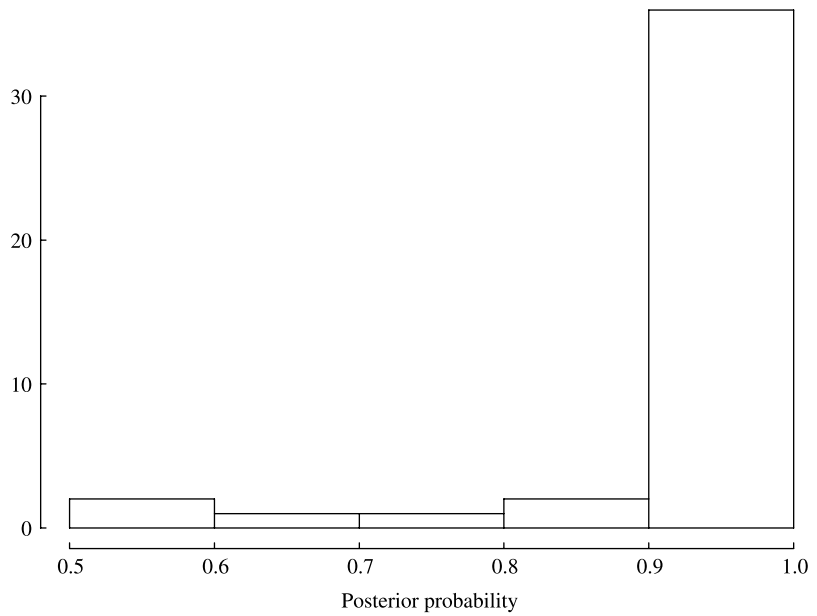

Fig. 8. Histogram of δ_i^* for potential outliers only

Computations were done in an S-Plus environment with calls to Fortran using simple modifications of code developed by Schafer (1997). Schafer's software for incomplete multivariate normal data may be downloaded from the website http://www.stat.psu.edu/~jls. The routines used in this article are available from the author upon request.

3.4. Sensitivity to δ and λ

The known parameters, δ and λ, of the contaminated normal distribution may affect determinations about outliers, and therefore require careful selection. Preliminary exploration of the data can suggest very sensible values of δ and λ (Section 3.2). We also explored the sensitivity of the results as we changed the values of δ and λ over a wide range. We set δ equal to .01, .04 and .10, and λ equal to .50, .25 and .01. The first two values of δ are likely contenders while .10 (or 10% contamination) seems too high for this dataset. Thus, it should flag a disproportionate number of regular observations as outliers. Similarly, the first two values of λ are within the range of plausibility while .01 is implausible because the variance of the contaminated population is not 100 times that of the uncontaminated population according to Figure 4. Repeated runs of EM were

Table 3. A few suggested deletes for NHANES III Preliminary Phase I Data

Case	d_i^2	*Df*	*p*-value	Edits	New d_i^2	*p*-value
38	23.0	7	1.72×10^{-3}	Delete sitht	11.7	6.96×10^{-2}
53	81.4	6	1.89×10^{-15}	Delete ht	25.9	9.18×10^{-5}
72	83.5	6	6.66×10^{-16}	Delete butto	0.6	9.86×10^{-1}
279	418.8	7	0.0×10^{0}	Delete headc	9.1	1.68×10^{-1}
436	48.3	7	3.06×10^{-8}	Delete butto	39.3	6.38×10^{-7}

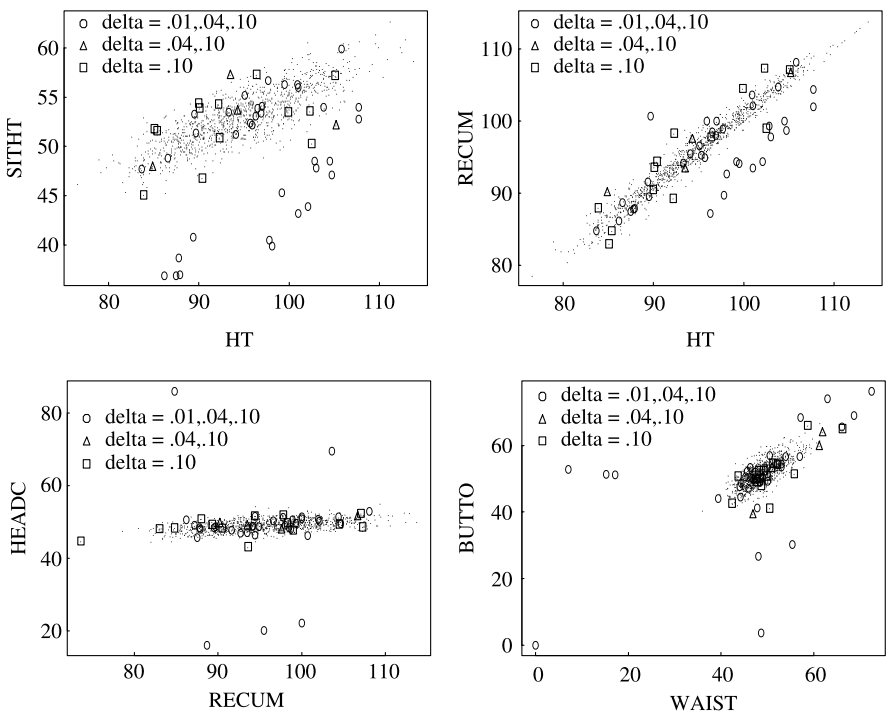

Fig. 9. Identifying outliers for $\lambda = .5$ and $\delta = .01, .04, .10$

performed for the nine possible combinations of δ and λ. The number of outliers (Table 1) and the number of suggested edits from each run were printed out. Then graphical displays were created using the identify() function in S-Plus for λ of .5 and δ of .01, .04 and .10 (Figure 9). Similar displays can be produced for different combinations of these parameters.

We found that reasonable combinations of δ and λ produced almost identical results. For example, we expected that $(\lambda, \delta) = (.5, .04)$ and $(\lambda, \delta) = (.25, .01)$ would produce similar values. That is, if the outliers have a larger variance (i.e., smaller value of λ), then we would expect a lower proportion of observations to be outliers. Running the algorithm with these two pairs of values identified 42 observations as outliers. However, upon closer inspection, we found that $(\lambda, \delta) = (.25, .01)$ missed a few of the observations outside the point cloud, so that the variance inflation factor should be larger than .25. When the variance inflation factor was set to .5, the δ values of .01 and .04 identified all of the gross outliers. Less appropriate combinations such as $(\lambda, \delta) = (.25, .10)$ flagged a larger number of observations ($n = 67$) but the additional observations were part of the regular data cloud, and not surprisingly looked fine upon inspection. In general, a combination of δ and λ that flags too many points is not advisable because it requires unnecessary inspection of nonoutlying cases, thus wasting valuable resources.

Having settled on $\lambda = .5$ as being a reasonable choice, we explored the effect of varying δ in Figure 9. We can see that the gross outliers were identified by all three values of δ. As expected, higher values of δ flagged a larger number of observations as potential outliers, but the additional cases were often on the periphery of the cloud and hard to distinguish as

outliers. For observations that are not clearly outliers, a conservative approach to data editing would advocate that the original data be left alone to retain all the important features of the data. In summary, we found that the values $\lambda = .5$ and $\delta = .04$ flagged all of the outlying values, but similar reasonable values worked just as well. Thus, the method was robust to the choices of δ and λ.

4. Discussion

The proposed outlier detection method is a simple and efficient approach to outlier detection that performs well for datasets with gross errors. Further, it works as a data editing tool, generating recommended edits on a case-by-case basis. However, decisions about what should be done with the potential outliers are deferred to subject-matter experts intimately familiar with the data, requiring resources and good judgement. For simplicity, we have assumed a multivariate contaminated normal to model the gross reporting and recording errors in the data. The contaminated normal model assumes that the observations are drawn from a mixture of two normal distributions, with the outliers being drawn from the one with larger variance, thus throwing off the covariance matrix for the entire vector of observations, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$. Our experience with the NHANES III suggests that extreme observations usually result from gross errors in one or more of the variables rather than all of the measured variables. Thus, one important and immediate extension of this method would be to model variables individually (Ghosh-Dastidar and Schafer 2003).

Further, there are alternative candidates for the data model, such as the multivariate t, that may be more suitable for other applications. While the contaminated normal assumes that the outliers arise from a distribution with larger variance than the regular observations, the multivariate t suggests that the data with errors arise from a heavy-tailed distribution. Both the contaminated normal and multivariate t are appropriate for continuous data alone. It is conceivable to extend this approach to other types of data (e.g., categorical) using categorical models such as the log-linear. An important limitation of the contaminated normal model is its lack of appropriateness for asymmetric or skewed data. While it is possible to transform the data and conduct outlier detection on the transformed scale, further simulations should be conducted to look at the behavior of this method when the data deviate from normality.

This algorithm requires the user to provide values for the mixing proportion and ratio of variances in the contamination model, which are estimated from the data. A sensitivity analysis conducted with different candidate values suggested that the method is robust to parameter assumptions, provided the values of λ and δ are reasonable. However, there are two suggestions for the practitioner when selecting a λ. When λ is close to 1, the covariances of the two populations are essentially equal so that it is hard to distinguish between the two. Therefore, we recommend that λ be set to .5 or less to allow for separation between the uncontaminated and contaminated populations. Also, combinations of δ and λ that flag too many points are not advisable because it requires unnecessary inspection of nonoutlying cases, thus wasting valuable resources.

Finally, survey data often originate from complex sample designs. Therefore, sample estimates should be weighted using the inverse sampling probabilities in order

to make population-level inferences. However, this outlier detection method does not attempt to produce estimates generalizable to the population. It is simply a data-driven editing tool that facilitates the identification of potential outliers and erroneous values in multivariate data. While we do not include sampling weights in our EM computations to derive the parameter estimates of the mean and covariance matrix, the contaminated normal automatically downweights influential observations to decrease their undue influence to produce robust estimates. However, we could easily incorporate the design effects within the EM algorithm to produce parameter estimates that are generalizable to the population.

5. References

Barcaroli, G. and Venturi, M. (1993). An Integrated System for Edit and Imputation of Data: An application to the Italian Labor Force Survey. Proceedings of the 49th Session of the International Statistical Institute, Florence, Italy.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1–38.

Ezzati-Rice, T.M., Khare, M., and Schafer, J.L. (1993). Multiple Imputation of Missing Data in NHANES III. Proceedings of the Annual Research Conference, U.S. Bureau of the Census, 459–487.

Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71, 17–35.

Ghosh-Dastidar, M. and Schafer, J.L. (2003). Multiple Edit Multiple Imputation for Multivariate Continuous Data. Journal of the American Statistical Association, 98, 807–817.

Granquist, L. and Kovar, J. (1997). Editing of Survey Data: How Much is Enough? In Survey Measurement and Process Quality, L.E. Lyberg, P.P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York, NY: Wiley.

Kovar, J.G. and Whitridge, P. (1990). Generalized Edit and Imputation System: Overview and Applications. Revista Brasileira de Estadistica, 51, 85–100, Rio de Janeiro.

Little, R.J.A. (1987). Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values. Applied Statistician, 37, 23–38.

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: John Wiley and Sons.

Little, R.J.A. and Smith, P.J. (1987). Editing and Imputation for Quantitative Survey Data. Journal of the American Statistical Association, 82, 58–68.

Marcoulides, G.A. and Hershberger, S.L. (1997). Multivariate Statistical Methods: A First Course. New Jersey: Lawrence Erlbaum Associates.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). Multivariate Analysis. New York: Academic Press.

National Center for Health Statistics (1994). Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988–1994. Vital and Health Statistics, Series. 1, No. 32.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika, 63, 581–592.

Rubin, D.B. (1983). Iteratively Reweighted Least Squares. Encyclopedia of the Statistical Sciences, 4, 272–275, New York: John Wiley and Sons.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

Thompson, K.J. and Sigman, R.S. (1999). Statistical Methods for Developing Ratio Edit Tolerances for Economic Data. Journal of Official Statistics, 15, 517–535.

Winkler, W.E. and Draper, L. (1997). The SPEER Edit System. In Statistical Data Editing, Vol. 2, J. Kovar and L. Granquist (eds). U.N. Commission for Europe, 51–55.

Received October 2003
Revised January 2006