

**Office of Population Research Princeton University**  
**WORKING PAPER SERIES**

Outlier Detection and Editing Procedures for Continuous  
Multivariate Data

Working Paper No. 2003-07

Bonnie Ghosh-Dastider  
RAND, Santa Monica  
OPR, Princeton University

J. L. Schafer  
Department of Statistics  
Pennsylvania State University

This research was completed as dissertation work and supported in part by Cooperative Agreement 43-3AEU-3-80087 between the National Agricultural Statistics Service, U.S. Department of Agriculture and Pennsylvania State University, awarded to Dr. J. L. Schafer.

Papers published in the OPR Working Paper Series reflect the views of individual authors. They may be cited in other publications, but are intended to be work-in-progress. Comments are welcome. Additional copies are available by writing to the Office of Population Research Working Paper Series, Princeton University, Wallace Hall, Second Floor, Princeton, NJ 08544 Fax (609) 258-1039. Or on the web, <http://www.opr.princeton.edu/>

# Outlier Detection and Editing Procedures for Continuous Multivariate Data

Bonnie Ghosh-Dastidar and J.L. Schafer <sup>1</sup>

September 2003

<sup>1</sup>Bonnie Ghosh-Dastidar is Statistician, RAND, Santa Monica, CA 90407-2138 (E-mail: *bonnieg@rand.org*). Joseph L. Schafer is Associate Professor, Department of Statistics, The Pennsylvania State University, University Park, PA 16802 (E-mail: *jls@stat.psu.edu*). This research was completed as dissertation work and supported in part by Cooperative Agreement 43-3AEU-3-80087 between the National Agricultural Statistics Service, U.S. Department of Agriculture and The Pennsylvania State University, awarded to Dr. J.L.Schafer.

## **Abstract**

We present a semi-automatic method of outlier detection for continuous, multivariate survey data. In large datasets, outliers may be difficult to find using informal inspection and graphical displays, particularly when there are missing values. 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. Mahalanobis distances are computed to identify potential outliers. The procedure is implemented in a software product which detects outliers and suggests edits to remove offending values. We apply the algorithm to body-measurement data from the Third National Health and Nutrition Examination Survey. 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; outliers; posterior probability.

# 1 INTRODUCTION

## 1.1 What is an outlier?

Outliers are observations that appear to be extreme or unusual with respect to the rest of the data and to prior knowledge about what values are plausible. Outliers may be “erroneous” or “real” in the following sense. “Real” outliers are observations whose actual values are, in fact, very different than those observed for the rest of the data and violate plausible relationships among variables. “Erroneous” outliers are observations that are distorted due to misreporting or misrecording errors in the data-collection process.

Outliers of either type may exert undue influence on the results of statistical analyses, so they should be identified using reliable detection methods prior to performing data analyses. When we encounter a potential outlier, our first suspicion may be that the observation resulted from a mistake or other extraneous effect, and should be discarded. However, if the outlier is “real” rather than “erroneous,” it may be conveying important information about the underlying population of real values. Non-judicious removal of observations that appear to be outliers may result in underestimation of the uncertainty present in the data. As a consequence, estimated standard errors and p-values may be smaller than they should be, possibly leading to false findings of significance. In this paper, we demonstrate how potential outliers may be identified, but we don’t say what should be done with them. Caution and good judgment should always be exercised when rendering decisions about whether outliers should be removed.

## 1.2 Informal methods of outlier detection

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, indicated by the bold point. It is an outlier because it is much larger than the other observations.

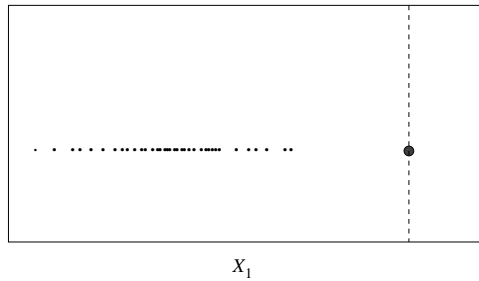


Figure 1: Extreme value in a marginal distribution

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.

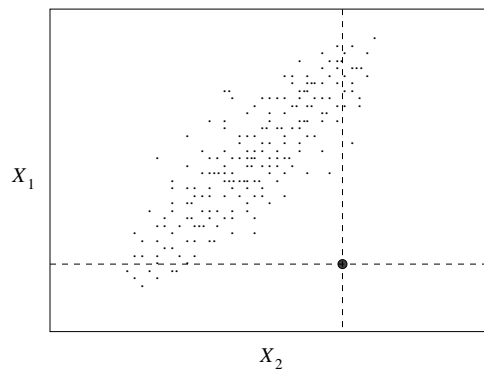


Figure 2: Outlier in a bivariate plot

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, and it may be difficult to match points in different plots. Finally, when the multivariate dataset contains missing values as well, outliers may not show up on bivariate scatterplots, because standard plotting routines will typically omit a unit from a scatterplot if either of the variables is missing.

### 1.3 Our method

In this paper we develop a semi-automatic method for outlier detection with 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 being 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  $\theta$ . 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 (1987).

This procedure yields robust estimates of the unknown model parameters, namely the means and covariances of the theoretical uncontaminated population. Mahalanobis distances relative to the center of the uncontaminated population are then calculated for all

observations. Points whose distances exceed a pre-determined 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.

In this paper, we only show how to identify potential outliers under a particular probability model. Other plausible probability models for outliers are discussed. In addition, whether an outlier should be deleted or not is a subjective decision and should always be made by experts with knowledge of the subject matter and the process of data collection. Thus, the method discussed here is a semi-automatic method of outlier detection.

## 1.4 Motivating example: NHANES III

We are interested in body-measurement data from the Third National Health and Nutrition Examination Survey (NHANES III), conducted by the National Center for Health Statistics. Multivariate surveys such as NHANES III are subject to non-response and outliers. Initial explorations of the preliminary Phase I data by Schafer, Khare and Ezzati-Rice (1993) revealed a substantial number of outliers, particularly in the body-measurement variables. For the most part, these outliers reflected gross errors in the data recording and capture process. Schafer *et al.* then proceeded with informal methods and graphical displays to edit the dataset and found that the informal methods were inadequate.

The NHANES III exercise motivated our search for a fast, reliable method for outlier detection with multivariate datasets. We believe that our procedure is an efficient and sound approach to data editing. It is also quite general and can be easily applied to other continuous, multivariate data. It is particularly useful when inter-variable correlations are strong. Other useful applications of this method might include establishment surveys which consist of highly correlated economic variables, and longitudinal datasets in which measurements for units at different points in time are highly correlated.

## 1.5 Scope of the rest of the paper

Section 2 discusses the implementation of the outlier detection 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  $\{x_i : i = 1, \dots, n\}$  be a random sample of values subject to contamination, where each  $x_i$  is a vector of length  $k$ . We assume that  $x_i \sim N_k(\mu, \Psi/\lambda)$  with probability  $\pi$  and  $x_i \sim N_k(\mu, \Psi)$  with probability  $(1 - \pi)$ , where  $\lambda$  is a positive scalar less than 1. The parameters  $\mu$  and  $\Psi$  are unknown, whereas  $\pi$  and  $\lambda$  are assumed known. The quantity  $\lambda$  indicates the magnitude of the errors leading to contamination; for example, if  $\lambda = .1$  the contamination is regarded as inflating the covariances by a factor of 10. In practice, it may be of interest to explore a grid of possible values of  $\pi$  and  $\lambda$ , 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 low 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  $\mu$  and  $\Psi$  have an attractive interpretation as the moments

of the uncontaminated component of the population.

Let  $x_i$  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  $x_i$  is

$$p(x_i|\theta) = (1 - \pi)|2\pi\Psi|^{-\frac{1}{2}} \exp\{-(x_i - \mu)\Psi^{-1}(x_i - \mu)^T/2\} + \pi|(2\pi\Psi)/\lambda|^{-\frac{1}{2}} \exp\{-\lambda(x_i - \mu)\Psi^{-1}(x_i - \mu)^T/2\}, \quad (1)$$

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

$$p(x_i|\theta) = (2\pi)^{-\frac{k}{2}}|\Psi|^{-\frac{1}{2}} \left\{ (1 - \pi) \exp\left(-d_i^2/2\right) + \pi\lambda^{\frac{k}{2}} \exp\left(-\lambda d_i^2/2\right) \right\}. \quad (2)$$

Consider  $n$  independent, identically distributed (i.i.d.) observations from a  $k$ -variate contaminated normal distribution,  $X = (x_1, \dots, x_n)$ . The loglikelihood function of  $\theta$  given  $X$  is

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

## 2.2 Augmenting the data

Maximum likelihood (ML) estimation of parameters requires maximizing the loglikelihood function. As described in Section 1.3, 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.

Let us associate with each observation  $x_i$  a dichotomous variable  $q_i$  indicating whether or not the  $x_i$  comes from the uncontaminated component of the population; that is,  $q_i = \lambda$  if  $x_i$  is distributed as  $N_k(\mu, \Psi/\lambda)$  and  $q_i = 1$  if  $x_i$  is distributed as  $N_k(\mu, \Psi)$ . Each  $q_i$  takes

the values  $\lambda$  and 1 with probabilities  $\pi$  and  $(1 - \pi)$ , respectively. The marginal distribution of  $q_i$  is  $p(q_i = \lambda) = \pi$ , and  $p(q_i = 1) = 1 - \pi$ . The conditional distribution of  $x_i$  given  $q_i$  is then

$$x_i | \theta, q_i \sim N_k(\mu, \Psi/q_i). \quad (4)$$

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

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

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

$$\begin{aligned} S_0 &= \sum_{i=1}^n q_i, \\ S_1 &= \sum_{i=1}^n q_i x_i, \\ S_2 &= \sum_{i=1}^n q_i x_i^T x_i. \end{aligned} \quad (6)$$

If  $X$  and  $Q$  are observed, ML estimates of the parameters  $\mu$  and  $\Psi$  can be found by a weighted least squares method in which the observations with larger variances are down-weighted. In our model, the covariance matrix of  $x_i$  given  $q_i$  is  $\Psi/q_i$  where  $q_i$  equals  $\lambda$  or 1, so if  $q_i$  were known we would apply a weight to  $x_i$  proportional to  $q_i$ . The weighted estimates are

$$\begin{aligned} \hat{\mu} &= \frac{\sum_{i=1}^n q_i x_i}{\sum_{i=1}^n q_i} = S_1/S_0, \\ \hat{\Psi} &= \frac{\sum_{i=1}^n q_i (x_i - \hat{\mu})^T (x_i - \hat{\mu})}{n} = \frac{S_2 - S_1^T S_1/S_0}{n}. \end{aligned} \quad (7)$$

## 2.3 Basic EM algorithm

Without observing  $Q$ , we cannot use (8) to calculate ML estimates for the parameters. Rather, we apply the EM algorithm, a general iterative procedure for ML estimation in

incomplete 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 (7). 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  $(\mu^{(t)}, \Psi^{(t)})$ , the parameter estimates from the previous iteration. Because  $S_0$ ,  $S_1$  and  $S_2$  are linear functions of the  $q_i$ 's, the E-step reduces to finding the conditional expectations of the  $q_i$ 's,

$$w_i^{(t)} = E(q_i | x_i, \mu^{(t)}, \Psi^{(t)}). \quad (8)$$

For the contaminated normal model, it can be shown that

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

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

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

The estimates of  $\mu$  and  $\Psi$  obtained through this procedure are more robust than the usual estimates, because potential outliers have large values of  $d_i^2$ , which are downweighted. 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  $X = (X_{obs}, X_{mis})$ , where  $X_{obs}$  and  $X_{mis}$  denote the observed and missing parts of  $X$ , respectively. The augmented data still consist of  $X = (X_{obs}, X_{mis})$  and  $Q = (q_1, \dots, q_n)$ , of which only  $X_{obs}$  is observed;  $X_{mis}$  and  $Q$  are missing. Let  $x_{obs,i}$  denote the observed portion of  $x_i$ , and  $x_{mis,i}$  denote the missing portion of  $x_i$ , so that  $X_{obs} = \{x_{obs,i} : i = 1, \dots, n\}$  and  $X_{mis} = \{x_{mis,i} : i = 1, \dots, n\}$ .

Let us assume that the missing data are missing at random (MAR) — i.e. the missing-data mechanism does not depend on  $X_{mis}$  (Rubin 1976, p. 582). ML estimates of  $\mu$  and  $\Psi$  can now be computed by applying a modified EM algorithm which treats  $X_{mis}$  and  $Q$  as missing data. The E-step requires a few modifications from the previous section because portions of  $x_i$  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 is the same function (6) as before.

The  $(t + 1)^{st}$  iteration of the modified EM algorithm proceeds as follows (Little 1987).

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

$$\begin{aligned} E(S_0 | X_{obs}, \theta^{(t)}) &= E\left(\sum_{i=1}^n q_i \mid x_{obs,i}, \theta^{(t)}\right) \\ &= \sum_{i=1}^n E(q_i \mid x_{obs,i}, \theta^{(t)}) \\ &= \sum_{i=1}^n w_i^{(t)}. \end{aligned} \tag{10}$$

The weights  $w_i^{(t)}$  are a simple modification of (10), calculated as follows:

- Replace  $k$  by  $k_i$ , the length of  $x_{obs,i}$ .
- Compute the squared distances  $d_i^2$  using the subvector of observed variables  $x_{obs,i}$ , and the corresponding portions of  $\mu^{(t)}$  and  $\Psi^{(t)}$ .

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

where  $\hat{\mu}_{obs}$  and  $\hat{\Psi}_{obs,obs}$  are the portions of the estimated mean vector and covariance matrix corresponding to the observed variables.

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

$$\begin{aligned} E\left(\sum_{i=1}^n q_i x_{ij} \mid X_{obs}, \theta^{(t)}\right) &= \sum_{i=1}^n E\left\{q_i E(x_{ij} \mid x_{obs,i}, \theta^{(t)}, q_i) \mid x_{obs,i}, \theta^{(t)}\right\} \\ &= \sum_{i=1}^n w_i^{(t)} \hat{x}_{ij}^{(t)}, \end{aligned} \quad (12)$$

where  $\hat{x}_{ij}^{(t)} = E(x_{ij} \mid x_{obs,i}, \theta^{(t)})$ , because the conditional mean of  $x_{ij}$  given  $x_{obs,i}$ ,  $\theta^{(t)}$  and  $q_i$  does not depend on  $q_i$ .

Finally, the  $(j, k)^{th}$  element of  $E(S_2 | X_{obs}, \theta^{(t)})$  is

$$\begin{aligned} E\left(\sum_{i=1}^n q_i x_{ij} x_{ik} \mid X_{obs}, \theta^{(t)}\right) &= \sum_{i=1}^n E\left\{q_i E(x_{ij} x_{ik} \mid x_{obs,i}, \theta^{(t)}, q_i) \mid x_{obs,i}, \theta^{(t)}\right\} \\ &= \sum_{i=1}^n \left(w_i^{(t)} \hat{x}_{ij}^{(t)} \hat{x}_{ik}^{(t)} + \Psi_{jk.obs,i}^{(t)}\right), \end{aligned} \quad (13)$$

where

$$\Psi_{jk.obs,i}^{(t)} = \begin{cases} 0 & \text{if } x_{ij} \text{ or } x_{ik} \\ & \text{are observed,} \\ q_i \text{ Cov}(x_{ij}, x_{ik} \mid x_{obs,i}) & \text{if } x_{ij} \text{ and } x_{ik} \\ & \text{are both missing.} \end{cases} \quad (14)$$

The quantities  $\hat{x}_{ij}^{(t)}$  and  $\Psi_{jk.obs,i}^{(t)}$  are the means and covariances, respectively, of the conditional distribution of  $x_{mis,i}$  given  $x_{obs,i}$ . The conditional means and covariances come from a multivariate regression of the response  $x_{mis,i}$  on the predictors  $x_{obs,i}$ . These parameters may be computed from the assumed values of  $\mu$  and  $\Psi$  by applying the sweep operator, as described by Little and Rubin (1987, p. 112-119).

- M-Step: Remains unmodified from Section 2.3.

## 2.5 Identifying potential outliers

The three diagnostic tools we use to identify possible outliers are

1. weights,

2. posterior probabilities, and
3. p-values for testing the hypothesis that observation  $x_i$  comes from the uncontaminated part.

The *weight* of observation  $i$  is the conditional expectation of  $q_i$  given the observed data and parameter estimates,  $w_i^{(\infty)} = E(q_i | x_{obs,i}, \hat{\mu}_{obs}, \hat{\Psi}_{obs,obs})$ . From the marginal density of  $q_i$  given by (4), we see that  $\lambda \leq w_i^{(\infty)} \leq 1$ . Observations that are located far from the mean of the distribution have large values of  $d_i^2$  and therefore low values of  $w_i^{(\infty)}$ , because  $w_i^{(\infty)}$  is a decreasing function of  $d_i^2$ . Given an appropriate cutoff value  $w^c$ , any observation  $i$  for which  $w_i^{(\infty)} < w^c$  can be flagged as a potential outlier.

The *posterior probability* of an observation originating from the contaminated population conditional upon the observed data and parameter estimates is given by (16).

$$\pi_i^* = P(q_i = \lambda | x_{obs,i}, \hat{\mu}_{obs}, \hat{\Psi}_{obs,obs}). \quad (15)$$

Using (4) and (9),

$$w_i^{(\infty)} = \lambda P(q_i = \lambda | x_{obs,i}, \hat{\mu}_{obs}, \hat{\Psi}_{obs,obs}) + P(q_i = 1 | x_{obs,i}, \hat{\mu}_{obs}, \hat{\Psi}_{obs,obs}), \quad (16)$$

and thus

$$\pi_i^* = \frac{1 - w_i^{(\infty)}}{1 - \lambda}. \quad (17)$$

A cutoff of .5 seems to work well in many applications, so that we flag observation  $i$  if  $\pi_i^* > .5$ . Thus, all observations which are more likely to have come from the contaminated population are treated as potential outliers.

The final approach utilizes *p-values* to test the hypotheses that the  $i^{th}$  observation comes from the contaminated part. This is a valid  $\alpha$ -level test for a single observation  $x_i$ . The null hypothesis states that the observation originates from the uncontaminated population while the alternative is that it comes from the contaminated part. The test statistic is  $d_i^2$  given by (12). It can be shown that asymptotically  $d_i^2 \sim \chi_{k_i}^2$ , so approximate p-values can be obtained from chi-square distributions. If the p-value for observation  $i$  is less than or equal to a cutoff value  $\alpha$ , we may flag the observation  $x_i$  as a potential outlier. The value of  $\alpha$

should be chosen in relation to the overall sample size  $n$ , because even in an uncontaminated population we expect  $n\alpha$  points to be flagged as potential outliers purely by chance.

## 2.6 Identifying influential variables

For each potential outlier, we can identify influential variables to suggest plausible edits. Influential variables are the ones with the most extreme values; hence, they 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  $x_{obs,i}$  with variable  $j$  omitted.
- (ii.) Find  $j_1$  such that  $d_i^{(j_1)2} < d_i^{(j)2}$  for all observed  $j$ . Variable  $j_1$  is then the most influential variable in the  $i^{th}$  case.

The suggested delete for outlying case  $i$  is variable  $j_1$ .

If the p-value for  $d_i^{(j_1)2}$  is still significant, there must be more than one offending variable. A procedure to identify the next  $m$  deletes after removing  $j_1$  from  $x_{obs,i}$  is:

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

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

### 3 APPLICATION

#### 3.1 Results

This method is illustrated here with an application to seven raw body-measurement variables (Table 1) from NHANES III for 1262 children 2–3 years of age. All of these variables were measured in centimeters. Preliminary diagnostics (Figure 3) such as histograms and normal quantile plots showed that most of the variables are approximately normally distributed. Transformations were not very useful for the skewed variables, therefore the untransformed variables were considered in this example. Bivariate scatterplots (Figure 4) of the body-measurement variables indicate moderate to high inter-variable correlations; they also reveal possible outliers.

Table 1: Select body-measurement variables from NHANES III

Variable	Description
sm5	survey-reported height
ht	standing height
sight	sitting height
recum	recumbent length *
headc	head circumference
waist	waist circumference
butto	buttocks circumference

\* Recumbent length=full body length measured when lying down

A contaminated normal model with a mixing probability ( $\pi$ ) of .04 and variance inflation constant ( $\lambda$ ) equal to .5 were assumed for this dataset. Based on the bivariate scatter plots and knowledge of the data, a low rate of contamination and a variance inflation factor of 2 seemed appropriate. Sensitivity of the results to choices of  $\pi$  and  $\lambda$  are discussed below. The EM algorithm for this model converged in 20 iterations giving estimates of  $\mu$  and  $\Sigma$  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 0–422.6, the  $w_i^{(\infty)}$ 's ranged from .5–.998, and the  $\pi_i^*$ 's were between .004 and 1.

Histograms of  $w_i^{(\infty)}$  and  $\pi_i^*$  (Figures 5 and 7) suggested cutoffs of  $w^c = .75$  and  $\pi^c = .5$  which flagged about 3% of the observations as potential outliers. The contaminated normal model downweights extreme observations so that their weights are close to  $\lambda$ , the lower bound. 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 7. Figures 6 and 8 show the distribution of  $w_i^{(\infty)}$  and  $\pi_i^*$ , respectively, for only those observations identified as outliers by  $w^c$  and  $\pi^c$ .

A number of potential outliers and suggested edits are shown in Table 3. The column for  $Df$  in this table indicates the number of observed variables. For example, case 38 has a  $Df$  of 7 which means that all 7 variables were observed. The corresponding  $d_i^2$  produces a significant p-value of .002, which increases to .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. However, the former seems to have more than one erroneous value whereas the latter has only one. 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.

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>.

### 3.2 Sensitivity to $\pi$ and $\lambda$

The known parameters,  $\pi$  and  $\lambda$ , of the contaminated normal may affect determinations about outliers, and therefore require careful selection. Typically, there is a range of sensible values for  $\pi$  and  $\lambda$  that we can start with. For our problem, values of  $\pi$  equal to .01, .04 and .10, and that of  $\lambda$  equal to .667, .50, .25 and .01 were considered. Repeated runs of EM were performed for all possible combinations of  $\pi$  and  $\lambda$ . The number of outliers (Table 2) and suggested edits from each run were printed out. Then graphical displays were created, using

the identify() function in S-Plus (Figure 9), for  $\lambda$  of .5 and  $\pi$  of .01, .04 and .10. Similar displays can be produced for different combinations of these parameters.

Table 2: Distribution of number of outliers

$\lambda$	$\pi$		
	.01	.04	.10
.667	28	36	40
.50	38	42	57
.25	42	51	67
.01	36	39	41

to see which of the observations were being flagged as outliers.

According to Figure 9, the gross outliers are identified by all three values of  $\pi$ . Higher values of .04 and .10 identify a larger number of observations that are on the periphery of the cloud point, and harder to distinguish as outliers. It is reassuring to see that reasonable values of  $\pi$  and  $\lambda$  produced almost identical results. However, some combinations, such as  $\pi = .10$ ,  $\lambda = .25$ , flagged additional observations which upon inspection seemed to be fine. In such situations, a conservative approach to data editing would advocate that the original data be left alone to retain all the important features of the data. Also, a combination of  $\pi$  and  $\lambda$  that flags too many points may be inadvisable because it requires unnecessary inspection of data, thus wasting valuable resources.

We recommend that  $\lambda$  be no larger than .50 to allow for enough separation between the uncontaminated and contaminated populations. When  $\lambda$  is close to 1, the covariances of the two populations are essentially equal so that it is hard to distinguish between the two. As a result, our method will be more sensitive to the choice of  $\pi$ . For the NHANES III dataset,  $\pi = .04$  and  $\lambda = .5$  were selected with all of these issues in mind, and also based upon information provided by the plots and data print-outs.

## 4 Discussion

This method of outlier detection based on the multivariate contaminated normal distri-

bution is a quick, efficient approach to editing that performs well for data with gross errors. For simplicity, we assume a contaminated normal to model the errors in the data although it does not fully capture the intermittent error mechanism. The contaminated normal model assumes that outliers are drawn from a distribution with larger variance, thus throwing off the covariance matrix for the entire vector of observations,  $x_i = (x_{i1}, \dots, x_{ip})$ . However, extreme observations usually result from gross errors in one or a few of the variables. Therefore, a possible extension of this method would be to model variables individually. Future research will consider alternative models that may more fully capture the mechanism underlying the gross response errors in survey data. Editing also requires good understanding of the data, and we recommend that methods such as these be used in consultation with subject-matter specialists.

## References

- [1]. 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, B39*, 1-38.
- [2]. Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York.
- [3]. 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.
- [4]. Little, R. J. A. (1987). Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values. *Applied Statistician, 37*, 23-38.
- [5] Rubin, D. B. (1976), Inference and Missing Data. *Biometrika, 63(3)*, 581-592.
- [6]. Rubin, D. B. (1983). Iteratively reweighted least squares, *Encyclopedia of the Statistical Sciences, Vol. 4*, John Wiley & Sons, Inc., New York, 272-275.
- [7]. Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.
- [8]. Schafer, J. L., Khare, Meena and Ezzati-Rice, T.M. (1993). Multiple Impuation of Missing Data in NHANES III. *Proceedings of the 1993 Annual Research Conference*, Bureau of the Census, Washington, DC.

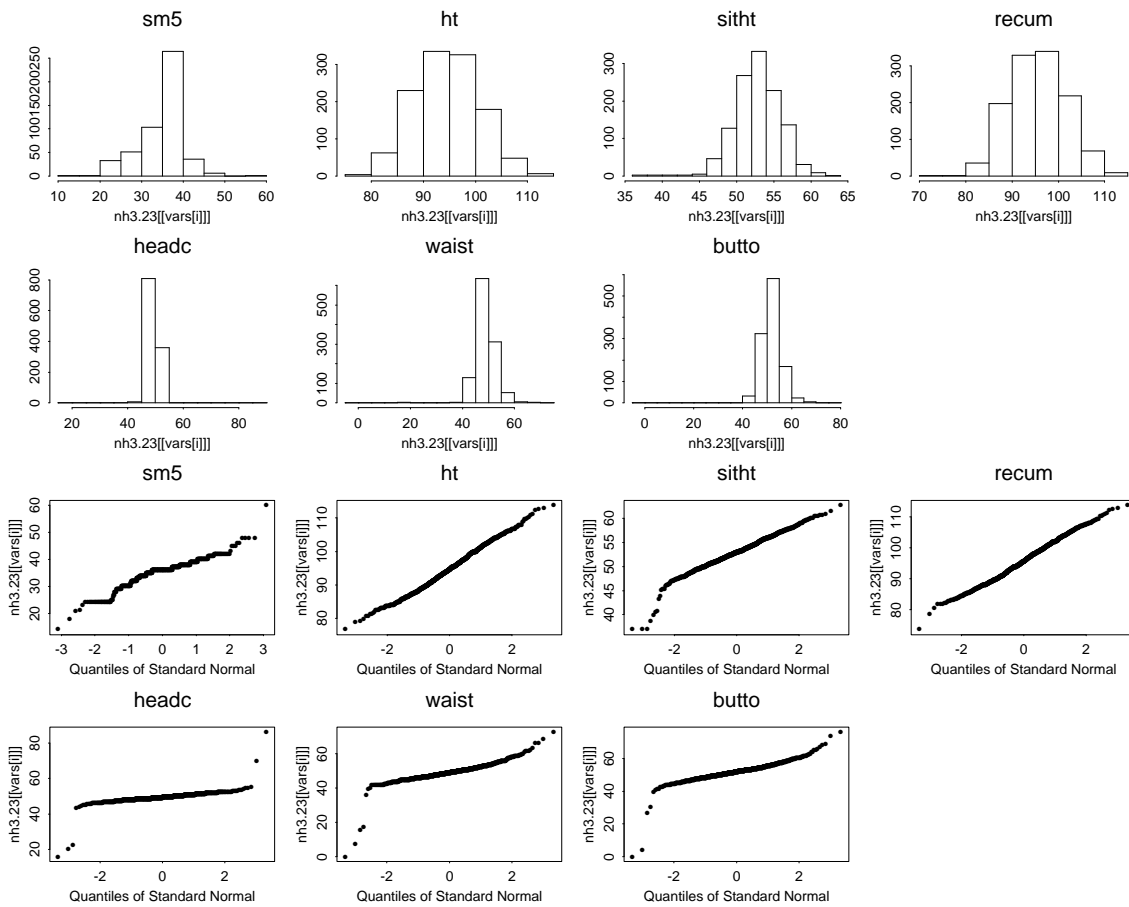


Figure 3: Marginal distributions of body-measurement variables

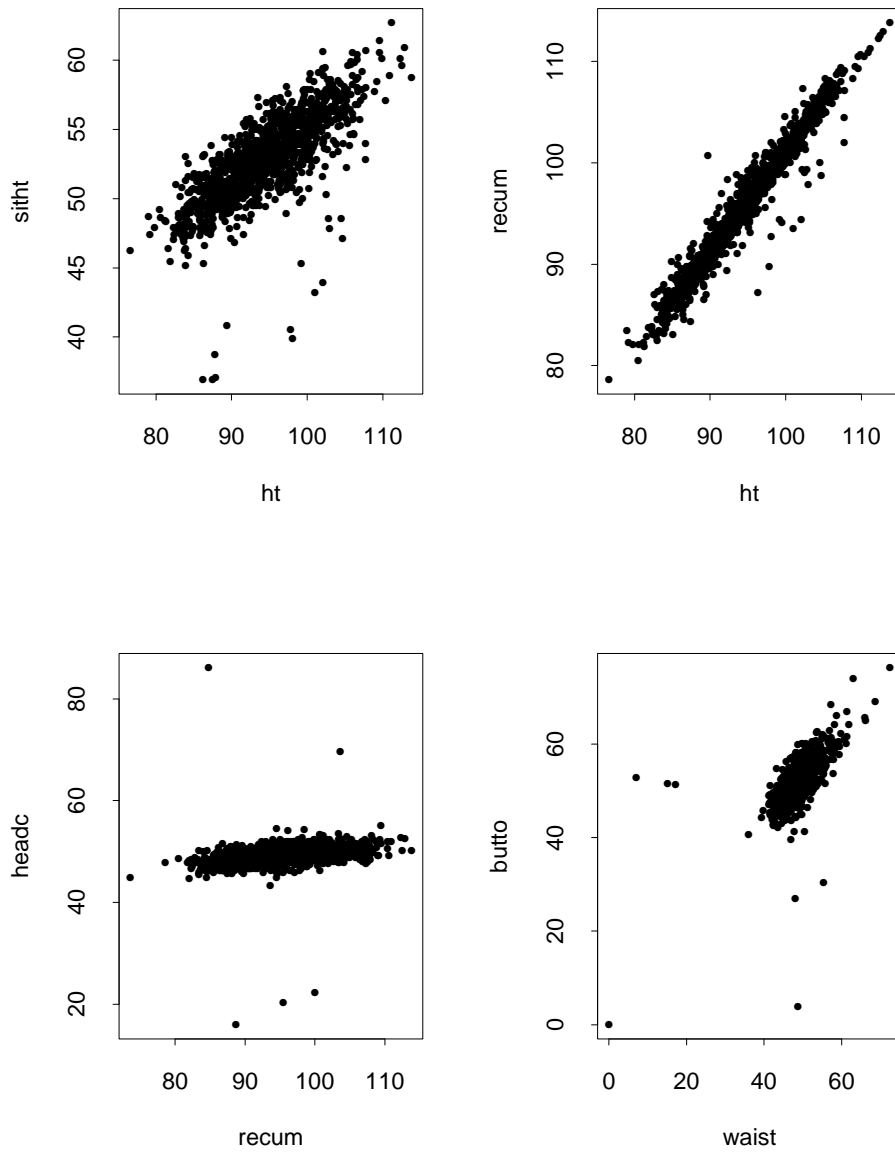


Figure 4: Scatterplots of body-measurement variables

Table 3: Parameter estimates from EM

Means						
sm5	ht	sight	recum	headc	waist	butto
35.3	94.7	52.9	95.7	49.2	48.8	51.8
Standard deviations						
sm5	ht	sight	recum	headc	waist	butto
4.67	6.08	2.93	6.05	2.04	3.85	4.07
Correlation matrix						
sm5	ht	sight	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

Table 4: A few suggested deletes for NHANES III

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

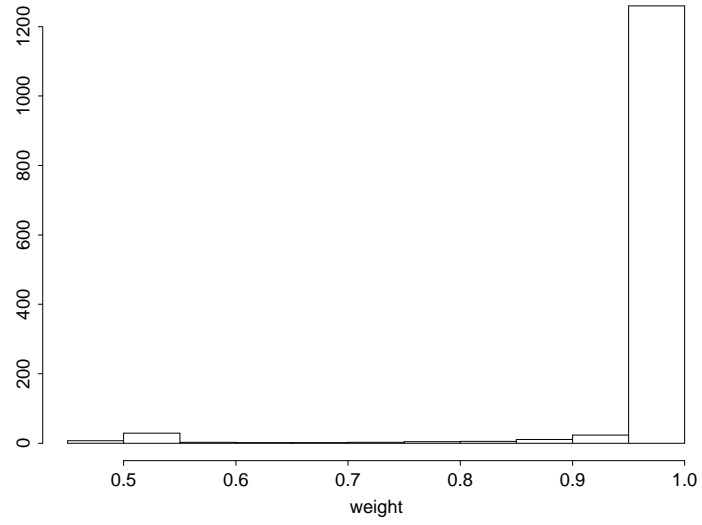


Figure 5: Histogram of  $w_i^{(\infty)}$

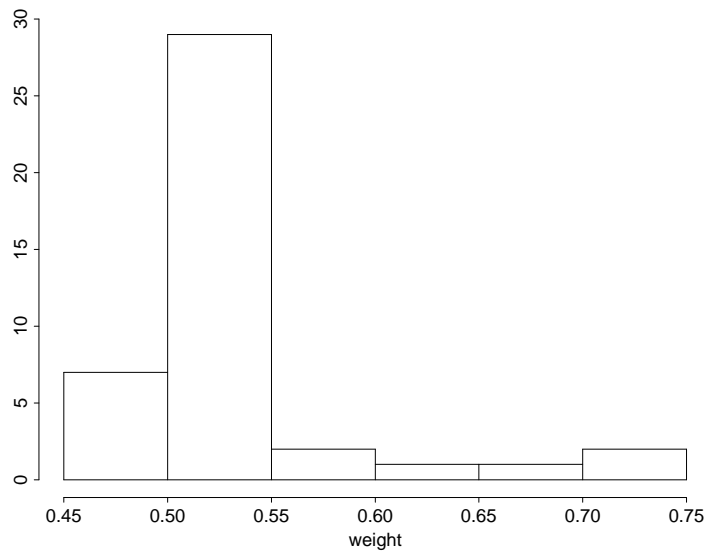


Figure 6: Histogram of  $w_i^{(\infty)}$  for potential outliers only

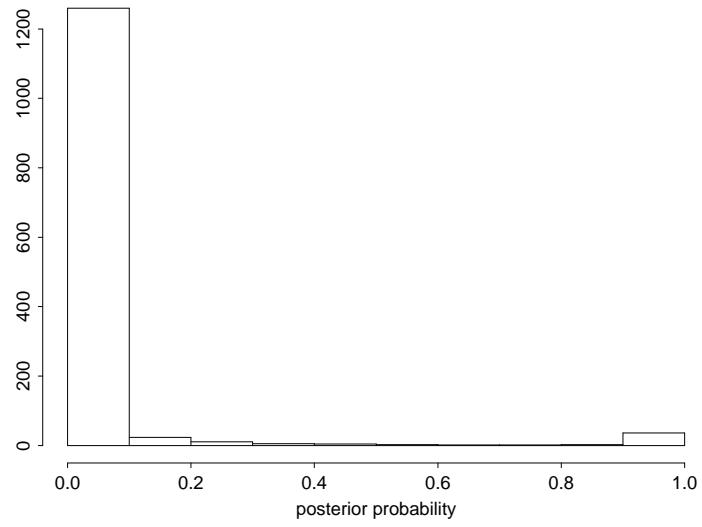


Figure 7: Histogram of  $\pi^*$

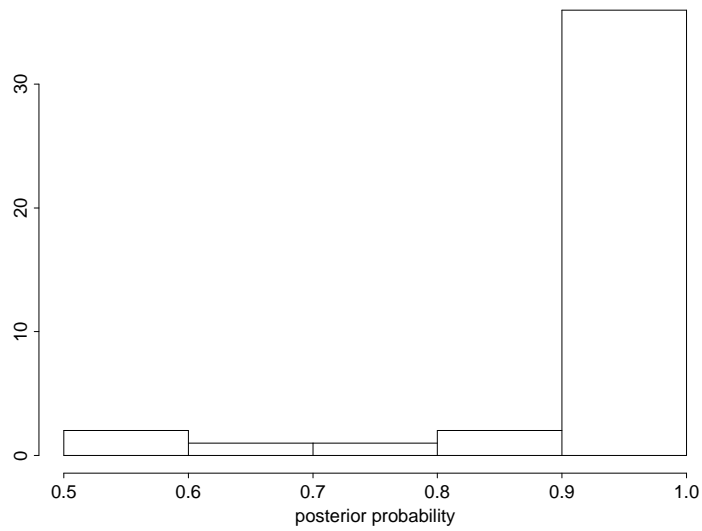


Figure 8: Histogram of  $\pi^*$  for potential outliers only

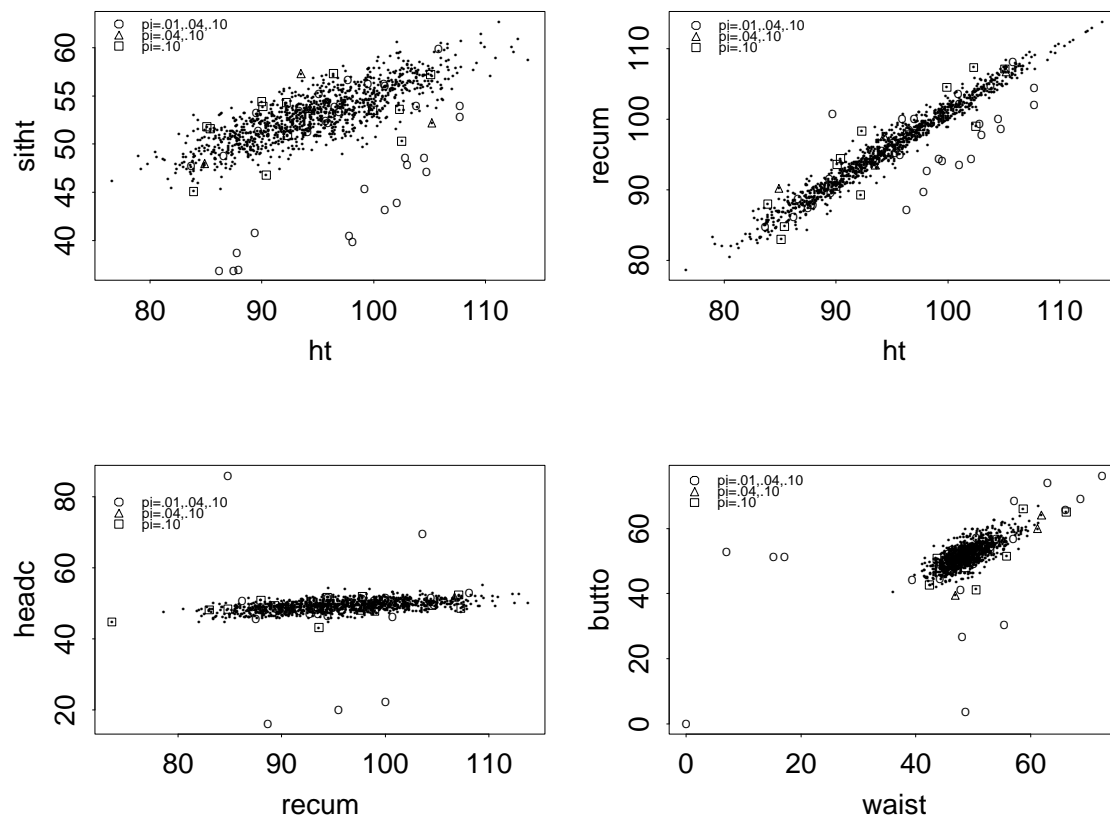


Figure 9: Identifying outliers for  $\lambda = .5$  and  $\pi = .01, .04, .10$