Robust Multivariate Calibration

Silvia Salini

Department of Economics, Business and Statistics
University of Milan, Italy
silvia.salini@unimi.it

Abstract. Multivariate calibration uses an estimated relationship between a multivariate response Y and an explanatory vector X to predict unknown X in future from further observed responses. Up to now very little has been written about robust calibration. An approach can be based on the outliers deletion methods. An alternative is to employ robust procedures. The purpose of this paper is to present multivariate calibration methods which are able to detect and investigate those observations which differ from the bulk of the data or to identify subgroups of observations. Particular attention will be paid to the *forward search* approach.

1 Introduction

Multivariate calibration uses an estimated relationship between a multivariate response Y (of dimension q) and an explanatory vector X (of dimension p) to predict unknown X in future from further observed responses. The purpose of this paper is to present multivariate calibration methods which are able to detect and investigate those observations which differ from the bulk of the data or, more generally, to identify subgroups of observations. We are concerned not only with the identification of atypical observations, but also with the effect that they have on parameter estimates, on inferences about models, and on their suitability. In this paper particular attention will be paid to the *forward search* approach (Atkinson, Riani and Cerioli, 2004). In this method we start with a fit to very few outlier-free observations and then successively fit larger subsets. We thus order the observations by closeness to the fitted model. As a result, not only are outliers and distinct subsets of the data discovered, but the influential effect of these observations is made clear. Section 2 gives more details about multivariate calibration. Section 3 presents some possible approach on robust calibration. In section 4 the forward search procedure is applied to real data set, and some comments and remarks are given.

2 Multivariate Calibraton

Statistical calibration, potentially useful in several practical applications, deals with the inference on unknown values of explanatory variables, given a

vector of response variables. Suppose for example that two different instruments for the measurement of the same phenomenon are considered. The first one $\mathbf{X}$ (*standard method*) is more difficult, accurate and expensive than the second one $\mathbf{Y}$ (*test method*). A sample of n units, in which both measures $\boldsymbol{x}$ and $\boldsymbol{y}$ are available, is considered. The set of values $(\boldsymbol{x}_i,\boldsymbol{y}_i)$ i=1,...,n is the *calibration experiment.* The statistical calibration problem arises when only the $\boldsymbol{y}_i$ obtained by the test method are known and the unknown $\boldsymbol{x}_i$ have to be estimated. The solution of this problem, *prediction experiment*, depends on the probabilistic model supposed to have generated the calibration experiment. In particular, it is assumed that the values $\boldsymbol{y}_i$ are realizations of a random variable (r.v) $\mathbf{Y}$ with known density function.

The assumptions on the values $\boldsymbol{x}_i$ may be of two types: i) the $\boldsymbol{x}_i$ are realizations of a r.v. and therefore $(\boldsymbol{x}_i,\boldsymbol{y}_i)$ are realizations of a multivariate r.v. (*random calibration*); ii) the $\boldsymbol{x}_i$ are chosen by the experimenter (*controlled calibration).*

In the classical parametric approach a linear multivariate model for both experiments is considered. Suppose that the calibration experiment is made of n observations, q response variables $Y_1, Y_2, \ldots, Y_q$ and p explanatory variables $X_1, X_2, \ldots, X_p$ with $q \geq p$, and suppose that $\mathbf{Y}_1 = \mathbf{1}\alpha^T + \mathbf{XB} + \mathbf{E}_1$, where $\mathbf{Y}_1(n \times q)$, $\mathbf{X}(n \times p)$, $\mathbf{1}(n \times 1)$ are known matrices; $\mathbf{E}_1(n \times q)$ is a matrix of random errors, whose i-th row is $\mathbf{E_{1i}} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Gamma})$; $\mathbf{B}(p \times q)$ and $\alpha(n \times 1)$ are unknown parameters. The model for the prediction experiment is given by: $\mathbf{Y}_2 = 1\alpha^T + 1\xi^T\mathbf{B} + \mathbf{E}_2$, where $\mathbf{Y}_2(m \times q)$, $\mathbf{E}_2(m \times q)$ whose j-th row is $\mathbf{E_{2j}} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Gamma})$ and $\xi\,(q \times 1)$ is the unknown vector of calibration measures.

When $q = p$ the multivariate classical estimator for ξ is

$$\hat{\xi}_C = \left(\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathbf{T}}\right)^{-1}\hat{\mathbf{B}}\mathbf{S}^{-1}\left(\bar{\mathbf{y}}_2 - \hat{\alpha}\right) \tag{1}$$

where $\hat{\mathbf{B}}$ and $\hat{\alpha}$ are least-squares estimators, $\bar{\mathbf{y}}_2$ is the mean of the observations in the predicted experiment and $\mathbf{S}$ is the pooled covariance matrix. (1) is also the maximum likelihood (ML) estimator of ξ.

When $q > p$, the ML estimator is a function of $\hat{\xi}_C$ and a quantity that depends on an inconsistency diagnostic statistic R, a measure of the consistency of $\mathbf{y}_2$ to estimate (more details in Zappa and Salini, 2004).

3 Prediction Diagnostics

Detecting outliers is an important aspect in the process of statistical modeling. Outliers, with respect to statistical models, are those observations that are inconsistent with the chosen model. Once a multivariate calibration model is built, it is used to predict a characteristic (e.g. standard measure) of new samples. Developing robust calibration procedures is important because standard regression procedures are very sensitive to the presence of atypical observations; furthermore, very little is known about robust calibration.

There are two basic approaches to robust calibration: use robust regression methods or perform classic estimators on data after rejecting outliers.

Robust estimates work well even if the data are contaminated. Several robust regression estimation methods have been proposed (Rousseeuw and Leroy, 1987): the M-estimator is the most popular; the R-estimator is based on the ranks of the residuals, the L-estimator is based on linear combination of order statistics; the Least Median of Squares (LMS) estimator minimizes the median of the squares of the residuals; the S-estimator is based on the minimization of a robust M-estimate of the residual scale; the Generalized M-estimator (GM) attempts to down-weigh the high influence points as well as large residual points; the MM-estimator is a multistage estimator which combines high breakdown with high asymptotic efficiency. In the calibration literature, generalized M-estimation techniques have been applied to the controlled calibration problem and orthogonal regression on the measurement-error model to the random calibration problem. These techniques give robust calibration estimators (Cheng and Van Ness, 1997) but extensions of these methods when $p > 1$ and $q > 1$ lead to difficulties. In addition, although robust estimators can sometimes reveal the structure of the data, they do so at the cost of down-weighting or discarding some observations. Finally, if the calibration experiment is made up of different subsets, the use of robust estimators will tend to produce a centroid which lies in between different groups. In this last case prediction will be strongly determined by the size of the subsets which make up the calibration experiment.

A second approach on robust multivariate calibration consists in performing a classical multivariate estimator on data after rejecting outliers. There are many methods to detect outliers. A single outlier can easily be detected by the methods of deletion diagnostics in which one observation at a time is deleted, followed by the calculation of new parameter estimates and residuals. With two outliers, pairs of observations can be deleted and the process can be extended to the deletion of several observations at a time. This is the basic idea of multiple deletion diagnostics. A difficulty both for computation and interpretation is the explosion of the number of combinations to be considered. A similar approach is based on the repeated application of single deletion methods (backward methods). However, such backwards procedures can fail due to masking.

The forward search appears to be more effective than the other approaches especially in the presence of multiple outliers (Atkinson and Riani, 2000). Also in the calibration context, the problem can be formulated as searching for the outlier-free data subset, the basic idea of forward search method (see next section). Genetic algorithms are proposed as a reasonable tool to select the optimum subset (Walczak, 1995). The results obtained with this genetic approach are compared with classical robust regression method of least median of squares (LMS).

Another approach, a third one, for the prediction of diagnostics in multivariate calibration problem, could be based on the inconsistency diagnostic R, mentioned in the previous section; the statistic R is central to diagnostic checking, whether or not it influences confidence intervals and point estimators (Brown and Sundberg, 1989).

4 Forward Search in Multivariate Calibration: An Example

The forward search is a general technique for robust estimation. The approach in calibration field considers the direct regression model and is based on the idea of forming a clean subset of the data, and then testing the outlyingness of the remaining points relative to the chosen clean subset. The algorithm combines robust estimation, diagnostics and computer graphics. The first step of the algorithm is based on the idea of elemental sets. The forward search starts by selecting an outlier free subset of p observations, where p is the number of parameters to be estimated in the model. To select this subset, a large number of subsets are examined, and the one with the smallest median residual is chosen - this is known as least median of squares (LMS) estimation. Having chosen this initial subset, the search moves from step p to step $p+1$ selecting $(p+1)$ units with the smallest least squares residuals. The model is re-fitted in this way until all units are included in the subset. Throughout the search, certain statistics such as the residuals, are monitored. Diagnostic plots are then constructed with the X-axis representing the subset size and the Y-axis representing the statistic of interest. In the case of calibration problem with $q > p$, q direct regression models are considered and the initial subset of dimension r, $S^{(r)}$, used to initialize the forward search, is found using the intersection of units, that have the smallest LMS residuals considering each response independently. In symbols for each response j, $S^{(p)}_{\mathbf{c}^*,j}$ satisfies

$$e^2_{[\text{med}],S^{(p)}_{\mathbf{c}^*,j}} = \min_{\mathbf{c}}[e^2_{[\text{med}],S^{(p)}_{\mathbf{c},j}}], \tag{2}$$

where $e^2_{[k],S^{(p)}_{\mathbf{c},j}}$ is the k th ordered squared residual among $e^2_{i,S^{(p)}_{\mathbf{c},j}}$, in the regression which considers the j-th variable as response, $i = 1, \ldots, n$, $\mathbf{c}$ is a collection of p units (the number of $\mathbf{c}$ collections is $\binom{n}{p}$) and med is the integer part of $(n+p+1)/2$. The initial subset is associated with the k units whose residuals at maximum have the r-th position ($r \leq n/2$) among $e^2_{[1],S^{(p)}_{\mathbf{c}^*,j}}, \ldots, e^2_{[n],S^{(p)}_{\mathbf{c}^*,j}}$, $j = 1, 2, \ldots, q$. The search progresses from subset size m to $m+1$ by selecting the smallest $(m+1)$ Mahalanobis distances (MD) (Atkinson, Riani and Cerioli, 2004, p. 66) from multivariate regression $d^*_{im} = (e^T_{im}\hat{\Sigma}e^{-1}_{um}e_{im})^{1/2}$ are scaled by the square root of the estimated covariance matrix, where $\hat{\Sigma}_u = (E^T E)/(m-p)$.

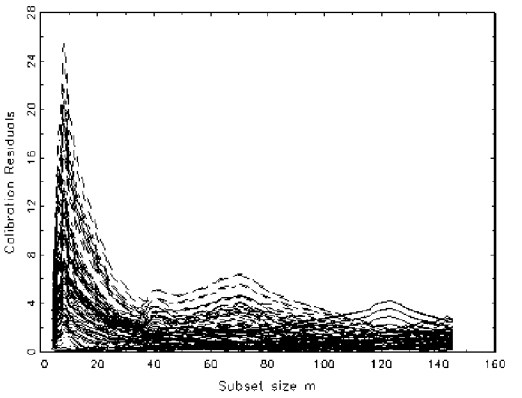

Fig. 1. Forward plot of scaled Mahalanobis distances based on residuals of calibration experiment.

This algorithm proceeds up to when all units are included in the subset $(m = k, k + 1, \ldots, n)$.

In order to illustrate how the method works we can use a dataset refereed to the noise of the traffic([1]). Calibration experiment has to determine the hourly equivalent level of the noise of traffic. Time sampling techniques differ for the size of the sample: surveying every second (87000 observations), surveying every minute (1450 observations), surveying every 10 minute (145 observations), surveying every hour (24 observations). As standard measure X is considered the hourly mean obtained by surveying every second, as test measure Y is considered the hourly mean of surveying every minute and the hourly mean of surveying every 10 minute. Therefore $q = 2$ and $p = 1$ and $n = 145$. without loss of generality only 1000 subsets are considered to select the initial subset. Fig. 1 shows the typical output of forward search, it monitors the calibration residuals at each steps of the forward search, every trajectory refers to one unit. The plot evidences the potential presence of groups, corresponding to different time slots. In particular forward plot in Fig. 2 and Fig. 3 show that unit 43 and unit 115 have a different trajectory than the others. The units correspond to time 7 AM and 6.50 AM, critical time for the city traffic.

In the final part of this section we compare the forward approach with other robust estimators. It is important to notice that extensions of robust method (Cheng and Van Ness, 1997) when $p > 1$ and $q > 1$ using robust regression approach lead to difficult because the necessary robust multivariate regression theory has not been developed. In our case case $q = 2$, then two robust model are estimated. Some robust regression estimators (Huber 1981,

[1] I am grateful to G. Brambilla (Institute of Acoustic "O.M. Corbino" C.N.R. Roma) for providing the data.

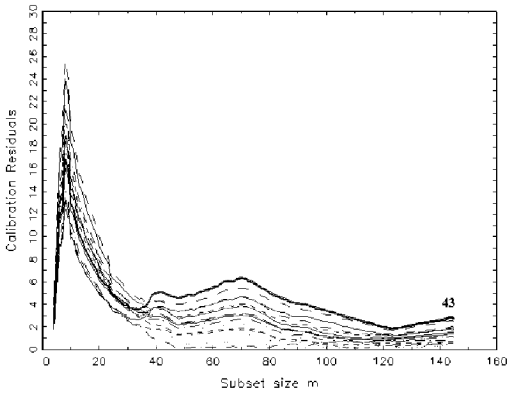

Fig. 2. Forward plot of scaled Mahalanobis distances at the step 60 of the search. Unit 43, evidenced in bold, has different trajectory than the others. The units correspond to time 7 AM, a critical time for the city traffic.

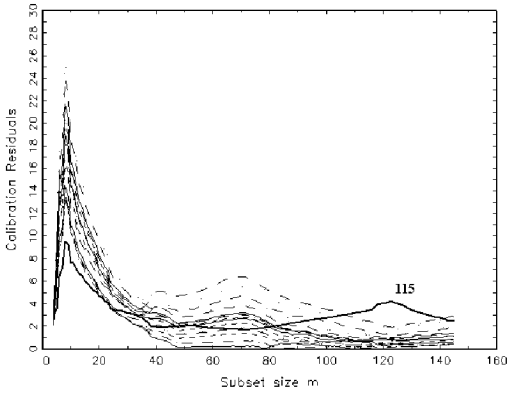

Fig. 3. Forward plot of scaled Mahalanobis distances at the step 60 of the search. Unit 115, evidenced in bold, has different trajectory than the others. The units correspond to time 6.50 PM, a critical time for the city traffic.

Hampel et al. 1986) are implemented. A combining method (Johnson and Krishnamoorthy, 1996) is applied to combine the univariate robust estimators, in fact the response variable is determined by two different measuring methods. The following equation shows the combining formula:

$$\hat{x}_c = \sum_{i=1}^{q} w_i \hat{x}_{ci} \tag{3}$$

where $w_i = (\hat{\beta}_i^2/S_i^2)/(\sum_{i=1}^{q} \hat{\beta}_i^2/S_i^2)$ in which $\hat{\beta}_i^2$ is a robust estimator and $S_i^2 = \sum_{j=1}^{n} (\hat{y}_{ij} - y_{ij})^2/(n-q)$. The estimator in (3) is a GLM estimator that minimizes $\sum_{i=1}^{q} (y_i - \hat{\alpha}_i - \hat{\beta}_i x)^2/S_i^2$ with respect to x. Table 1 shows the classical combined estimator and the robust ones in both cases of contaminated and non contaminated data.

Models	Data 1	Data 2
Classical	6.514	20.031
Huber	6.575	16.402
Tukey	6.687	15.106

Table 1. Standard Deviation of residuals for classical estimator and robust estimators for both not contaminated (Data 1) and contaminated (Data 2) data.

In the first case there are not outliers but only groups, as evidenced in forward plot in Fig. 1, robust and classical estimators perform in the same way. In presence of outliers robust estimators fit better than the classical one. Fig. 4 represents the true value of x versus the classical estimator, the Hubert and the Tukey estimator. As we expected robust estimators fit better than the classical one that it is very sensitive to the presence of atypical observations.

5 Conclusion

The problem of robust multivariate calibration is approached by the forward search method and by the classical robust regression procedures. In presence of groups the forward search performs better than classical robust procedures that are useful in presence of single outliers. It is important to notice that the combining method proposed in section 4 does not consider the robust multivariate regression theory (Rousseeuw et al., 2004) but refers to the cases in which the multivariate response variable is measured by different instruments or determined by various methods. Further, robust multivariate regression procedures can be applied on calibration problems. We want to study this extension and plan to report it elsewhere. We are currently investigating the behavior of the inconsistency diagnostic R mentioned in section 3 with forward search plots. We are interesting to create the envelopes for the R statistic, in this way we could be able to accept or reject the hypothesis that a new observation is inconsistent with the data.

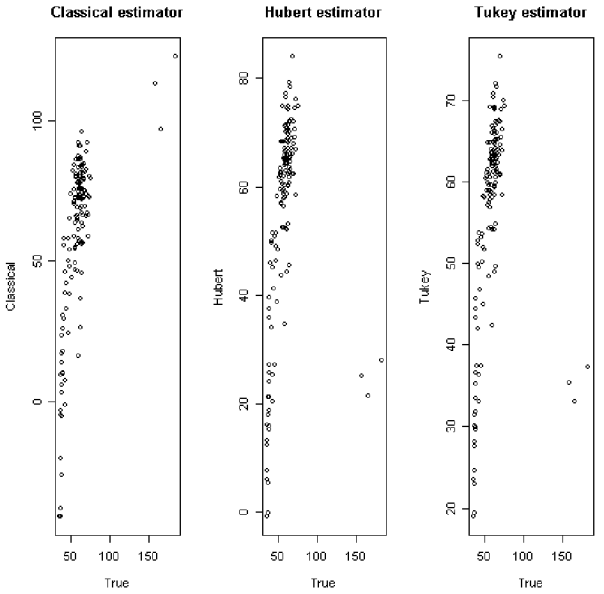

Fig. 4. True value versus classical and M-estimators in contaminated data

References

ATKINSON, A.C. and RIANI, M. (2000): *Robust Diagnostic Regression Analysis*, Springer, New York.

ATKINSON, A.C., RIANI, M. and CERIOLI, A. (2004): *Exploring Multivariate Data With the Forward Search*, Springer Verlag, New York.

BROWN, P.J. and SUNDBERG, R. (1989), Prediction Diagnostic and Updating in Multivariate Calibration, *Biometrika, 72, 349-361.*

CHENG, C.L. and VAN NESS, J.W. (1997), Robust Calibration, *Technometrics, 39, 401-411.*

HUBER P.J. (1981): *Robust Statistics.* Wiley.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J. and STAHEL, W.A. (1986): *Robust Statistics: The Approach based on Influence Functions.* Wiley.

JOHNSON, D.J. and KRISHNAMOORTHY, K. (1996): Combining Independent Studies in Calibration Problem, *Journl of the American Statistical Associations, 91, 1707-1715.*

ROUSSEEUW, P. and LEROY, A. (1987): *Robust Regression and Outlier Detection*, Wiley, New York.

ROUSSEEUW, P., VAV AELST, S., VAN DRISSEN, K., and AGULLÓ, J. (2004): Robust Multivariate Regression, *Technometrics, 46, 293-305.*

WALCZAK, B. (1995): Outlier Detection in Multivariate Calibration, *Chemiometrics and Intelligent Laboratory Systems, 28, 259-272.*

ZAPPA, D. and SALINI, S. (2004): Confidence Regions for Multivariate Calibration: a proposal. In: M. Vichi et al editors: *New Developments in Classification and Data Analysis.* Springer, Bologna, 225–233.