Confidence Regions in Multivariate Calibration: A Proposal

Diego Zappa[1] and Silvia Salini[2]

[1] Istituto di Statistica,
Università Cattolica del Sacro Cuore, Italy
diego.zappa@unicatt.it
[2] Dipartimento di Economia Politica e Aziendale,
Università di Milano, Italy
silvia.salini@unimi.it

Abstract. Most of the papers on calibration are based on either classic or bayesian parametric context. In addition to the typical problems of the parametric approach (choice of the distribution for the measurement errors, choice of the model that links the sets of variables, etc.), a relevant problem in calibration is the construction of confidence region for the unknown levels of the explanatory variables. In this paper we propose a semiparametric approach, based on simplicial depth, to test the hypothesis of linearity of the link function and then how to find calibration depth confidence regions.

1 Introduction

Statistical calibration, broadly used in chemistry, engineering, biometrics and potentially useful in several practical applications, deals with the inference on the unknown values of explanatory variables given a vector of response variables. This is generally done using a model identified through a preliminary calibration experiment (general references on calibration are Brown (1993), Sundberg (1999)).

In Section [3] 2 we describe the multivariate calibration problem and in particular the difficulties in the construction of confidence regions in a parametric context; in Section 3, using a semi parametric approach, it is proposed a new methodology based on simplicial depth, able to overcome some problems of the parametric approach.

[3]The contents of this paper have been shared by both Authors. In particular Section 2 is due to S.Salini and Section 3 is due to D.Zappa.

2 Multivariate Calibration: the parametric approach

In univariate calibration the properties of the classical and the inverse estimators are known. Most of these results may be extended also to the multivariate context where the main and relevant problem is the construction of multivariate confidence regions (Salini 2003, II).

Following Brown (1993), we consider two steps.

1) *The calibration step.* We run an experiment of n observations on q response variables $Y_1, Y_2, ..., Y_q$ and p explanatory variables $X_1, X_2, ..., X_p$ in order to identify the transfer function that links the two sets of variables. Suppose that the transfer function is a linear model. Let E be a matrix of random variables (r.v.s) to represent the measurement errors. Then the calibration model is:

$$Y_1 = \mathbf{1}\alpha^T + \mathbf{X}\mathbf{B} + E_1 \tag{1}$$

where is $\mathbf{1}(n \times 1)$ the unit vector, $\mathbf{B}(p \times q)$ and $\alpha(q \times 1)$ a matrix and a vector of parameters respectively.

2) *The prediction step.* Analogously to the previous step, suppose that a matrix $Y_2(m \times q)$ of response variables is available, the prediction model is

$$Y_2 = \mathbf{1}\alpha^T + \mathbf{1}\xi^T\mathbf{B} + E_2 \tag{2}$$

where we are interested on the unknown values $\xi(p \times 1)$ of $\mathbf{X}$.

Let E_{1i} and E_{2j} be the i-th and the j-th column of E_1 and E_2, respectively. It will be assumed that $E(E_{1i}) = E(E_{2j}) = \mathbf{0}$, $E\left(E_{1i}E_{1i}^T\right) = E\left(E_{2j}E_{2j}^T\right) = \mathbf{\Gamma}, E_{1i}, E_{2j} \sim N(\mathbf{0}, \mathbf{\Gamma})$, and that the errors E_{2j} are not correlated with E_{1i}.

To find the confidence region for ξ, the most favorable situation is when p=q. Supposing that the variables $\mathbf{X}$ are standardized, it may be shown that

$$\left(\hat{\alpha} + \hat{\mathbf{B}}^T\xi\right) \sim N\left(\alpha + \mathbf{B}^T\xi, \mathbf{\Gamma}\left(\frac{1}{n} + \xi^T(\mathbf{X}^T\mathbf{X})^{-1}\xi\right)\right) \tag{3}$$

where $\left(\hat{\alpha}, \hat{\mathbf{B}}\right)$ are the maximum likelihood estimators of $(\alpha, \mathbf{B})$.

As the log-likelihood function of the mean sample vector $\bar{\mathbf{y}}_2$, conditional to ξ is:

$$l(\bar{\mathbf{y}}_2|\xi) \propto \left(\bar{\mathbf{y}}_2 - \alpha - \mathbf{B}^T\xi\right)^T \mathbf{\Gamma}^{-1}\left(\bar{\mathbf{y}}_2 - \alpha - \mathbf{B}^T\xi\right) m,$$

replacing $\alpha, \mathbf{B}, \mathbf{\Gamma}$ by their maximum likelihood estimate $\widehat{\alpha}$,$\widehat{\mathbf{B}}$,$\mathbf{S}$ respectively, we have the maximum likelihood estimator for ξ as :

$$\widehat{\xi}_C = \left(\widehat{\mathbf{B}}\mathbf{S}_1^{-1}\widehat{\mathbf{B}}^T\right)^{-1}\widehat{\mathbf{B}}\mathbf{S}_1^{-1}\left(\bar{\mathbf{y}}_2 - \alpha\right) \tag{4}$$

where $\mathbf{S}_1 = \left(\mathbf{Y}_1 - \mathbf{X}\widehat{\mathbf{B}}\right)^T\left(\mathbf{Y}_1 - \mathbf{X}\widehat{\mathbf{B}}\right)$. To find a confidence region for ξ, using (3) and (4), the $100(1-\gamma)\%$ prediction ellipsoid for the unknown levels ξ is the volume

$$\xi : T^2 = \left(\overline{\mathbf{y}}_2 - \widehat{\alpha} - \widehat{\mathbf{B}}^T\xi\right)^T \mathbf{S}^{-1} \left(\overline{\mathbf{y}}_2 - \widehat{\alpha} - \widehat{\mathbf{B}}^T\xi\right) \leq c\,(\xi)\, K \qquad (5)$$

where $K = \frac{q}{\nu} F_{1-\gamma,q,\nu}$ and $F_{1-\gamma,q,\nu}$ is the upper $100(1-\gamma)\%$ point of the standard F distribution on q and ν degrees of freedom, $c\,(\xi) = \frac{1}{m} + \frac{1}{n} + \xi^T(\mathbf{X}^T\mathbf{X})^{-1}\xi$ and $\mathbf{S}\,(q \times q)$ is the *pooled* matrix of $\mathbf{S}_1$ and $\mathbf{S}_2$. It may be shown that the volume (5) is convex only when the matrix $\mathbf{C} = \hat{\mathbf{B}}\hat{\mathbf{S}}^{-1}\hat{\mathbf{B}}^T - K(\mathbf{X}^T\mathbf{X})^{-1}$ is positive definite and even so it may collapse to a point, the estimate (4).

When $q > p$, the ML estimator is a function of (4) and of a quantity that depends on an inconsistency diagnostic statistic. It may be shown that the left part of (5) may be decomposed in

$$T^2 = \left(\widehat{\xi}_C - \xi\right)^T \widehat{\mathbf{B}}\mathbf{S}_1^{-1}\widehat{\mathbf{B}}^T \left(\widehat{\xi}_C - \xi\right) + \\ + \left(\overline{\mathbf{y}}_2 - \widehat{\alpha} - \widehat{\mathbf{B}}^T\xi\right)^T \mathbf{S}^{-1} \left(\overline{\mathbf{y}}_2 - \widehat{\alpha} - \widehat{\mathbf{B}}^T\xi\right) = V + R \qquad (6)$$

where R is a measure of the consistency of $\overline{\mathbf{y}}_2$ to estimate ξ, while V may be used to find confidence region for ξ. Note that $R = 0$ when $p = q$, because $\widehat{\xi}_C$ is the solution of the system of equations $\overline{\mathbf{y}}_2 = \widehat{\alpha} + \widehat{\mathbf{B}}\widehat{\xi}_C$.

Williams conjectured that Q and R have approximate F distribution as follows

$$\frac{n-p-q}{p} Q_\gamma \sim c\,(\xi)\, F_{p,n-p-q} \qquad \frac{n-p-q}{q-p} R_\gamma \sim c\,(\xi)\, F_{q-p,n-p-q}$$

and then the statistical significance of R may be tested. The confidence region (5) may have an anomalous behavior with respect to R (Brown 1993, pag. 89): the width of the region increases as R decreases and decreases as R increases. Alternative techniques to find a calibrating confidence region are based on profile likelihood. The resulting regions have the desirable property to be expanded as R increases and to be reduced as R decreases. Unfortunately even in this case, we may obtain boundless confidence regions. Some very recent parametric proposals are due to Bellio (2002) and Mathew and Sharma (2002). In these papers accurate confidence regions are reported and in the latter the problem of finding joint confidence regions is treated only when the response and the explanatory variables have the same dimensions, or when the explanatory variable is one-dimensional. Another recent proposal is based on Kalman filter theory. Under certain hypothesis on the error measurement correlation matrix (Salini (2003)) Kalman filter may be used to upgrade the statistical information relative to the classical estimator so that it can be dynamically adjusted to give an update posterior estimate.

3 A proposal: semiparametric depth calibration regions

Most of the statistics reported in the previous paragraph have distributional properties mainly based on the assumption of multinormality. The problems

connected to this assumption (or more generally to any parametric assumption) are well known and additionally in multivariate calibration it has been shown that the problem of finding an empty calibration confidence region may exists. Some of these problems may be overcome by a nonparametric approach. Our proposal will exploit the results of the data depth method proposed by Liu and Singh (1993). References and some applications of data depth may be found e.g. in Zappa (2002). For the sake of readability of the rest of the paper some preliminaries, comments and description of notation are needed.

Generally speaking a depth function, $D(\cdot,\cdot)$, is an application $D(\cdot,\cdot) : \mathbb{R}^k \times \mathcal{F} \rightarrow \mathbb{R}^1$, where $\mathcal{F}$ is a class of distributions on the Borel sets of $\mathbb{R}^k$. In a recent paper of Zuo and Serfling (2000) the basic properties that $D(\cdot,\cdot)$ should possess are reported. Among them probably the most relevant property is that $D(\cdot,\cdot)$ should be affine invariant, that is, for any non singular matrix $\mathbf{A}$ and any constant vector $\mathbf{b}$, $D(\mathbf{Ax}+\mathbf{b}; F_{\mathbf{Ax}+\mathbf{b}}) = D(\mathbf{x}, F_{\mathbf{x}})$.

There are several notions of depth functions. We will focus on the simplicial depth. Let $\{z_1,\ldots,z_n\} \subset \Xi_z \in \mathbb{R}^k$ a sample of n k-dimensional observations, with $n > k$. $S[z_{i_1},\ldots,z_{i_{k+1}}]$ will stand for the simplex with vertices $\{z_{i_1},\ldots,z_{i_{k+1}}\}$ for any i set of $k+1$ different points taken from n. Then, for any point z in $\mathbb{R}^k$, the sample simplicial depth at z, $SD(z)$, is defined as the number of simplexes that include z. In particular the relative rank, $r_{G_n}(z^*)$, of a new observation z^* with respect to the empirical distribution G_z that is

$$r_{G_{z_n}}(z^*) = \#\{z_i | SD(z_i) \leq SD(z^*) \text{ , for } i = 1,\ldots,n\}/(n+1), \qquad (7)$$

is a measure of how much outlying z^* is with respect to the data cloud Ξ_z. A relevant property of the simplicial depth is given by theorem 6.2 of Liu and Singh (1993) that will be used in the following . Synthetically, consider two samples, $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ from distribution G and $\mathbf{Y} = \{y_1, y_2, \ldots, y_m\}$ from distribution F. Let $Q(G_{X_n}, F_{Y_m}) = \frac{1}{m}\sum_{j=1}^{m} r_{GX_n}(y_j)$ where $r_{G_{X_n}}(y_j)$ is the proportion of x_i's having, with respect to the distribution G, $SD_G(x_i) \leq SD_G(y_i)$ and let $Q(G_{X_n}, G) = E[Q(G_{X_n}, F_{Y_m})|\mathbf{X}]$. Then

$$\sqrt{m}[Q(G_{X_n}, F_{Y_m}) - Q(G_{X_n}, G)] \xrightarrow{L} N\left(0, \frac{1}{12}\right) \quad \text{as} \quad n \rightarrow \infty,\ m \rightarrow \infty$$

Some additional notation, that will be used in the rest of the paper, must be presented. For every set $A_z \subseteq \Xi_z$, it will be defined by convex hull the intersection of all the possible convex subsets of Ξ_z containing A_z. Let $\{A^*_{z_1}, A^*_{z_2}, \ldots, A^*_{z_w}\}$ a set of subsets of Ξ_z such that $A^*_{z_1} \supset A^*_{z_2} \supset \ldots \supset A^*_{z_w}$. $co(A^*_z)$ will be the convex polytope connecting the vertices of the convex hull containing A^*_z such that $\forall z \subset A^*_z$

$$\exists\ S[z_{i_1},\ldots,z_{i_{k+1}}] \subseteq A^*_z : z = \sum_{j=1}^{k+1} \lambda_j z_{i_j} \quad \text{with} \quad \sum_{j=1}^{k+1} \lambda_j = 1, 0 \leq \lambda_j \leq 1$$

and $\mathcal{A}_z = \{A_{z_1}, A_{z_2}, ..., A_{z_w}\}$ will be the collection of sets of Ξ_z such that $co(A^*_{z_i}) = A_{z_i}$ for $i = 1, 2, ..., w$ with $A_{z_i} \cap A_{z_j} = \emptyset$ and $\cup_{i=1}^{w} A_{z_i} \subseteq \Xi_z$.

The most popular approach to multivariate calibration passes through two steps: 1) the application of data reduction techniques in order to reduce the complexity of the problem; 2) the implementation of parametric or non-parametric extrapolation methods to find a functional relationship between the set of variables. Criticism to this approach is mostly focused on the loss of information that the implementation of these techniques implies and on the assumption of linearity. These are the main two reasons that support the following proposal where: 1) all the information included in the set of variables will be used and 2) a preliminary test is run in order to verify if the hypothesis of linear relationship between the set of variables is true. The counterpart of this approach is the relevant computational effort needed. At present no sufficiently powerful (fast and reliable) software has been prepared and most of the available algorithms (alike the ours) have been programmed for research reasons or may be operatively used when the dimension of the dataset is not too large [4].

Consider an asymmetric relationship between two sets of multivariate variables, Y, X. For the sake of graphical representability we will focus on the case $p = q$. In the following we will give some details on the extension to the case $p \neq q$. Suppose that $\{\Omega_Y, \mathcal{B}_Y, G_{Y|X,\theta}, \boldsymbol{\Theta}\}$ is the parametric probabilistic space of Y, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a parameter vector (or matrix of parameters). Suppose that a transfer function $g : \Omega_X \to \Omega_Y$ exists and that measurement errors mask the true g. We will approximate g by a function $f(X, \boldsymbol{\theta}, E)$ where E is a matrix of r.v.s.. As most of the calibrating models are supposed to be linear (see §2), then it turns out to be relevant to study the 'degree of linearity' of f or more extensively its (at least locally) 'degree of invertibility'.

Let us first define what functional f we consider and then how to test if f is linear. Consider the following symbolic transformation

$$\mathcal{A}_X = \{A_{X_1}, ..., A_{X_w}\} \underbrace{\to \{A^*_{Y_1}, ..., A^*_{Y_w}\} \overset{co(A^*_Y)}{\to}}_{f} \{A_{Y_1}, ..., A_{Y_w}\} = \mathcal{A}_Y \qquad (8)$$

where for each convex polytope $A_{X_i} \in \mathcal{A}_X, i = 1, 2, ..., w$, we take the set $A^*_{Y_i}$ with elements in Y matching the vertices of A_{X_i} and then we consider the polytope $co(A^*_{Y_i}) = A_{Y_i}$. If $co^{-1}(A_Y)$ exists then $\mathcal{A}_X \overset{f}{\to} \mathcal{A}_X$ is a 1:1 application. If the points left after the $(w-1)$th polytope are less then $q+1$, they will be considered as a unique set simply transferred through f to the corresponding data in Y.

If f is linear and in the not statistical case where the r.v. E does not exists, we will obtain a result similar to the one given in Fig.1b where $\mathbf{B}$ is a matrix of known coefficients. If we introduce the disturbance $E \sim N_2(\mathbf{0}, \sigma^2\mathbf{I})$,

[4] We ourself have implemented a software to draw the $co(\cdot)$ function and to simulate the overall proposal (see. Zappa and Salini (2003)).

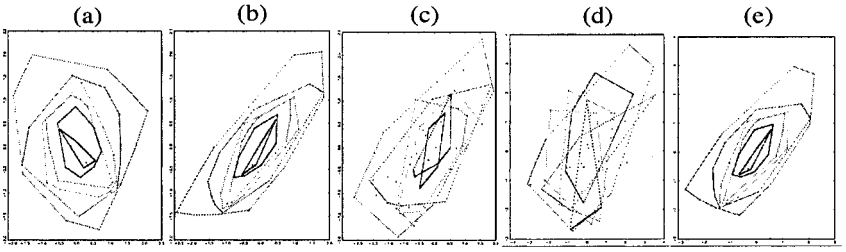

Fig. 1. Convex hulls of *(a)* $\mathbf{X}$; *(b)* $\mathbf{Y} = \mathbf{XB}$; *(c)* $\mathbf{Y} = \mathbf{XB} + \mathbf{E}_1$ where $\mathbf{E}_1$ comes from $E_1 \sim N(\mathbf{0}, \frac{1}{16}\mathbf{I})$; *(d)* $\mathbf{Y} = \mathbf{XB} + \mathbf{E}_2$ where $\mathbf{E}_2$ comes from $E_1 \sim N(\mathbf{0}, \frac{1}{16}\mathbf{I})$; *(e)* $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$ using case (b) and the least squares estimate for $\mathbf{B}$

from Fig 1c,1d it emerges that (8) will reproduce with a good approximation the true $\mathcal{A}_Y$ only when the contribution of the r.v.s is small that is when the explanatory variables are well identified and only errors likely due to measurement errors give low contribution to distortion. This is the typical calibration problem where the calibrating sample is generally accurately chosen.

To test if f is at least locally linear, making use of Liu's Q statistics and recalling that it has the properties to be invariant if f is affine, proceed as follows. For each element $A_{X_i} \in \mathcal{A}_X$, using (8) compute the sets $A_{Y_i}, A^*_{Y_i}$. Then compute

$$Q(G^*_{X_n}, A_{X_i}) = \frac{1}{m_{X_i}} \sum_{j=1}^{m_{X_i}} r_{G^*_{X n_i}}(x_j) \qquad Q(G^*_{Y_n}, A_{Y_i}) = \frac{1}{m_{Y_i}} \sum_{j=1}^{m_{Y_i}} r_{G^*_{Y n_i}}(y_j)$$

where $m_{X_i}(m_{Y_i})$ is the number of elements in $A_{X_i}(A_{Y_i})$, n_i is the number of samples in $\mathbf{X}(\mathbf{Y})$ after having peeled off the set $A_{X_i}(A^*_{Y_i})$, and $G^*_{X_n}(G^*_{Y_n})$ is the empirical distribution of $X(Y)$ without the set $A_{X_i}(A^*_{Y_i})$. Suppose that the paired samples in $\{\mathbf{X}, \mathbf{Y}\}$ are independent. We wish to compare $Q(G^*_{Y_n}, A_{Y_i})$ with $Q(G^*_{X_n}, A_{X_i})$ (which has the role of conditioning value). Using theorem 6.2 of Liu and Singh (1993) we may state the following result:

$$\sum_{i=1}^{w} \left[Q(G^*_{Y_n}, A_{Y_i}) - Q(G^*_{X_n}, A_{X_i})\right] \overset{n_i \to \infty}{\sim} N\left(c, \frac{1}{12} \sum_{i=1}^{w} \frac{1}{m_{Y_i}}\right) \tag{9}$$

The proof of (9) is resides in the sum of independent normal variables. For small samples the distribution in (9) should be premultiplied by the ratio (m_{Y_i}/m_{X_i}): this is needed because the set A_{Y_i} may not have the same cardinality of the corresponding A_{X_i}. As the dataset increases, under H_0, this ratio is almost 1.

Then the test on linearity may be formulated as:

$$\begin{cases} H_0 : c = 0 & \text{then } f \text{ is a linear application} \\ H_1 : c \neq 0 & \text{then } f \text{ is not a linear application} \end{cases}$$

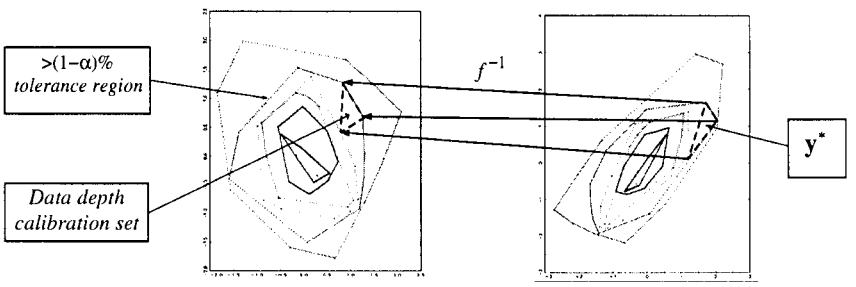

Fig. 2. Data depth calibration set. (At the left: $co(\mathbf{X})$; at the right $co(\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}})$)

Note that the above procedure may be implemented for any combination of p, q. Some problems may exist only for $p < q$. A_{X_i} must have at least $q + 1$ vertices otherwise we cannot build a convex set in Ω_Y. A conservative solution is to search for $q - p$ additional points, possibly internal to the region defined by the convex hull of A_{X_i} , such that volume of the corresponding hull A_{Y_i} is the largest. The results of a simulation based on 500 replications of (9) under the hypothesis of existence of linearity is reported in Zappa and Salini (2003). It has been noticed that the convergence of (9) to a normal distribution is matched even when n is as small as 10.

If the non parametric procedure is tested to be appropriate, then a standard parametric calibrating model may be implemented. In Fig.2 how to find a "data depth calibration set" is illustrated.

The procedure is:

1) Compute $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$ and apply (8).
2) Consider a new observation $\mathbf{y}^*$: find the smallest simplex (with $q + 1$ vertices) in Ω_Y that contains $\mathbf{y}^*$.
3) Through f^{-1} find the corresponding simplex in Ω_X: this will be called the *data depth calibration set.*
4) Find in Ω_X the convex hull that contains a pre-chosen $(1 - \alpha)\%$ of points. It will be interpreted as the fiducial region for the depth calibration set. Translating this region so that the depth calibration set is at the centre will result in the fiducial region for ξ: alike the fiducial approach on the construction of confidence region, we are $(1 - \alpha)\%$ sure of being right in this particular case, regarding the observations as fixed and setting up regions based on some belief in the values of the parameters that have generated those observations. This is the typical calibration context where a calibration experiment is run under well controlled conditions.

To find the calibrating depth region when $q < p$, use the vertices of the calibration depth region plus additional, possibly internal, $p - q$ points such that the simplex in Ω_X is the smallest.

4 Conclusions

A semiparametric procedure to build calibration confidence regions has been proposed. It may be used to test the very common hypothesis of linear relationship among the set of variables and it has the property to use all the information available to build the 'calibrating confidence region'. It may be used for any combination of p and q and the resulting region is limited and not empty (unlike what happens sometimes using the classical parametric approach). Further research is needed to solve some problems. First of all the computational effort needed: the algorithm is very time consuming and faster procedure must be implemented. A general form must be defined for the H_1 (the aim is to measure the power of the test) and it must be shown if the family of f to be tested in H_0 include only the linear model or other locally linear models. If the convex hull are each inside the others, it means that the link f is invertible: to what family does the link function belong to? Finally it must be measured up to what degree the random error disturbs the identification of the (supposed true) linear link and some simulations must be run to compare the classical and our new approach. A prospective can be the use of non parametric approach also in the estimation problem: either classical smoothing techniques, artificial neural networks or Kalman filter theory can be useful in presence of complexity and non normal error distributions.

References

BELLIO, R. (2002): Likelihood Methods for Controlled Calibration. *In printing in Scandinavian Journal of Statistics..*

BROWN, P. J. (1993): *Measurement, Regression and Calibration*, Oxford University Press, Oxford.

LIU, R., SINGH K. (1993): A Quality index based on Data Depth and Multivariate Rank Tests. *Journal of the American Statistical Association, 88, 252-260.*

MATHEW T., SHARMA M. K. (2002): Joint confidence regions in the multivariate calibration problem. *Journal of Statistical planning and Inference, 100, 427-441.*

SALINI S. (2003): *Taratura Statistica Multivariata*, Doctoral Thesis in Statistics - XV ciclo, Universit degli Studi di Milano - Bicocca.

SUNDBERG, R. (1999): Multivariate calibration - direct and indirect regression methodology (with discussion). *Scandinavian Journal of Statistics, 26, 161-207.*

ZAPPA D. (2002): A Nonparametric Capability Index for Multivariate Processes. *Studi in Onore di Angelo Zanella, Vita e Pensiero, 709-726.*

ZAPPA D., SALINI S. (2003): Some notes on Confidence Regions in Multivariate Calibration. *Istituto di Statistica, Universita' Cattolica del Sacro Cuore, SEP. 118*

ZUO Y., SERFLING R. (2000): General notions of statistical depth function. *The Annals of Statistics, 28, 2, 461-482.*

Part III

Applied Multivariate Statistics