Robust continuum regression

Sven Serneels[1] Peter Filzmoser[2] Christophe Croux[3]

Pierre J. Van Espen[1*]

[1]Department of Chemistry, University of Antwerp, Belgium

[2] Department of Statistics and Probability Theory, Vienna University of Technology, Austria

[3] Department of Applied Economics, K.U. Leuven, Belgium

November 17, 2004

*Correspondence to P. Van Espen, Departement Scheikunde, Universiteit Antwerpen, Universiteitsplein 1, 2610 Antwerpen (Belgium) E-mail: piet.vanespen@ua.ac.be. . Tel.: +32/3/8202358; Fax.: +32/3/8202376

Robust continuum regression

Abstract

Several applications of continuum regression to non-contaminated data have shown that a significant improvement in predictive power can be obtained compared to the three standard techniques which it encompasses (Ordinary least Squares, Principal Component Regression and Partial Least Squares). For contaminated data continuum regression may yield aberrant estimates due to its non-robustness with respect to outliers. Also for data originating from a distribution which significantly differs from the normal distribution, continuum regression may yield very inefficient estimates. In the current paper, robust continuum regression (RCR) is proposed. To construct the estimator, an algorithm based on projection pursuit is proposed. The robustness and good efficiency properties of RCR are shown by means of a simulation study. An application to an X-ray fluorescence analysis of hydrometallurgical samples illustrates the method's applicability in practice.

Keywords: Continuum regression (CR), Projection Pursuit, Robust continuum regression (RCR), Robust multivariate calibration.

1 Introduction

Parametric statistics has been developed as a science which endeavors to proffer applied scientists the ability to draw conclusive inference from data. The methodology

is based upon the random nature of the samples one has at disposition, combined with some assumptions made beforehand. These assumptions nearly always encompass the assertion that the data be drawn from a specified type of distributions, often taken to be the class of normal distributions. This order of proceeding has over the last century been successful in sundry practical applications, albeit the probability that the data have exactly been originated by the distribution assumed, is close to zero. From this insight a new branch of the statistical sciences has emerged, robust statistics. In robust statistics, one develops methods that take into account that the true data generating distribution is not necessarily equal to the imposed model distribution. In particular, robust methods can cope with the presence of outliers, being observations that are not all generated by the model. As in many practical applications the statistical model assumptions are violated, such that the ensuing inference becomes unreliable, robust statistics has become a mainstay in any field of applied sciences where one expects not to have enough control over the process of data generation. Chemometrics is no exception when regarding its vulnerability to possible erroneous or outlying observations.

A problem frequently addressed in applied sciences is the prediction of a dependent variable based on a linear model. Ever since Gauß [1] first touched this subject, adaptations and new estimation procedures have been designed. A special case frequently occurring in chemometrics consists of an ill-conditioned problem where the number of samples at hand is vastly exceeded by the number of explica-

tory variables, some of which may be correlated as well, so that the least squares regression estimates become unstable or do not even exist. In order to remediate these problems, various techniques have been proposed, all of which try to reduce the number of variables by compressing the data into a smaller set of uncorrelated, so-called *latent*, variables. A major question arising whilst applying this methodology, is how these latent variables should be defined, such that an optimal prediction of the dependent variable from these latent variables is obtained. For example, in principal component regression (PCR) the latent variables are linear combinations of the predictor variables having maximal variance. Another possibility is to perform partial least squares (PLS) [2], which constructs latent variables maximizing their covariance with the predictand, and one can expect this method to be better fit for prediction than PCR which constructs latent variables regardless of the predictand. Envisaging the necessity of a more general objective function, Stone and Brooks [3] proposed a joint maximization criterion called *continuum regression* (CR), which encompassed the before mentioned latent variables regression techniques as well as ordinary least squares regression. In continuum regression, a parameter δ (belonging to the interval $[0, 1]$) needs to be chosen or selected by cross-validation enabling one to decide which value of δ is best for the data at hand. Most values of δ do not correspond to existing methods, and justify the existence of continuum regression in its own right. Continuum regression only reduces to Ordinary Least Squares (OLS), PLS and PCR if δ equals 0, 0.5 or 1, respectively. Sundry practical applications in varying fields of science have shown that the application of continuum regres-

sion indeed improves prediction compared to the methods that existed before (see e.g. [4, 5, 6]).

It is thus the main purpose of this paper, to provide a robust version of the continuum regression framework. We will directly robustify the criterion which defines continuum regression by using robust estimators of variance and covariance. This framework provides the possibility to define a plethora of different robust continuum regression estimators, depending on which robust estimators of variance and covariance have been plugged into the criterion. In the current paper, we propose an algorithm to compute robust continuum regression estimators. As robust estimator of variance, we focus on the trimmed variance, being simple to compute and combining good robustness and efficiency properties. For similar reasons, trimmed sample covariances will estimate the covariance. We will show both by a simulation study and a practical application, that the robust continuum regression estimator we propose, is a valuable alternative for the existing robust estimation methods.

2 Definition of classical and robust continuum regression

Continuum regression was proposed by [3] as a unified regression technique embracing ordinary least squares, partial least squares and principal component regression.

Let $\boldsymbol{X}$ be a centred data matrix with n rows, containing the observations, and p columns, containing the predictor variables. Let $\boldsymbol{y}$ be a column vector containing

the n observations of the response variable. Continuum regression is basically a technique to estimate the vector of regression coefficients $\boldsymbol{\beta}$ in the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

with an error term $\boldsymbol{\varepsilon}$. As mentioned in the introduction, instead of directly solving (1), a latent variable model

$$\boldsymbol{y} = \boldsymbol{T}_h\boldsymbol{\xi} + \boldsymbol{\varepsilon} \tag{2}$$

is considered, with the so-called score matrix $\boldsymbol{T}_h = \boldsymbol{X}\boldsymbol{W}_h$ and $\boldsymbol{W}_h = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_h)$ being a $p \times h$ matrix of weights. The score matrix $\boldsymbol{T}_h$ contains the values of the h latent variables in its columns. Since h will typically be much smaller than p, the dimensionality of the regression problem is greatly reduced. The continuum regression weight vectors $\boldsymbol{w}_i$ $(i = 1, \ldots, h)$ are defined as proposed by [3], according to the criterion

$$\boldsymbol{w}_i = \underset{\boldsymbol{a}}{\operatorname{argmax}} \left\{ \operatorname{Cov}(\boldsymbol{X}\boldsymbol{a}, \boldsymbol{y})^2 \operatorname{Var}(\boldsymbol{X}\boldsymbol{a})^{\frac{\delta}{1-\delta}-1} \right\} \tag{3a}$$

under the constraints that

$$\| \boldsymbol{w}_i \| = 1 \quad \text{and} \quad \operatorname{Cov}(\boldsymbol{X}\boldsymbol{w}_i, \boldsymbol{X}\boldsymbol{w}_j) = 0 \text{ for } j < i. \tag{3b}$$

The parameter δ takes values between 0 and 1, and it adjusts the amount of information of the x-part to be considered for predicting the y-part. It is now easy to see from (3a) that we recover for $\delta = 0, 0.5, 1$ the well-known methods OLS, PLS and PCR, respectively.

In the criterion (3a), the abbreviations "Cov" and "Var" stand for the estimated covariance and variance, respectively. In classical continuum regression the usual sample covariance and variance estimators are used. But for robust continuum regression a robust estimator of covariance and variance needs to be used. Note that the covariance is only computed between two univariate variables, allowing for the use of simple robust covariance estimators, like a trimmed sample covariance.

The goal of continuum regression is to estimate the regression coefficients $\boldsymbol{\beta}$ in (1). In classical continuum regression the maximization problem (3) is either solved analytically which leads to a complex and inefficient algorithm, or approximated by a method called *continuum power regression* (CPR) [7], where a specific choice of δ and the dimension h has to be made. Then the parameter $\boldsymbol{\xi}$ in the model (2) is determined by ordinary least squares estimation, yielding

$$\hat{\boldsymbol{\xi}}_{\delta,h}^{CPR} = (\boldsymbol{T}_h^T \boldsymbol{T}_h)^{-1} \boldsymbol{T}_h^T \boldsymbol{y}.$$

Since $\hat{\boldsymbol{y}}_{\delta,h}^{CPR} = \boldsymbol{X}\boldsymbol{W}_h \hat{\boldsymbol{\xi}}_{\delta,h}^{CPR}$, estimates for the regression coefficients $\boldsymbol{\beta}$ are given by

$$\hat{\boldsymbol{\beta}}_{\delta,h}^{CPR} = \boldsymbol{W}_h (\boldsymbol{T}_h^T \boldsymbol{T}_h)^{-1} \boldsymbol{T}_h^T \boldsymbol{y} \tag{4}$$

for given values of δ and h. In the robust case, the estimation of the regression coefficients has to be done in a robust manner, and will be outlined in the next Section. A remaining important question to address is how the optimal values for δ and h can be determined. We will touch this important aspect of continuum regression modeling in Section 4.

3 Algorithm

3.1 Continuum regression by projection pursuit (CR-PP)

In the current section we will focus on how to compute robust continuum regression for given values of δ and h. In the case of classical continuum regression, an analytical solution to the maximization problem can be obtained. In the case of robust continuum regression, the latter is impossible. We decided to adopt the approach of projection pursuit. Projection pursuit (PP) as such has been initially proposed in 1974 ([8]) in order to reveal some relevant directions in an arbitrary data set; it has thenceforth been applied to a myriad of statistical problems (see e.g. [9]). PP has been particularly successful in the context of the construction of multivariate robust estimators such as a robust PP estimate of the scatter matrix [9], principal component analysis [10, 11, 12] and canonical correlation analysis [13].

Projection pursuit can be applied whenever the estimator to be constructed is defined by maximizing over all possible directions in the p dimensional space of a criterion computed solely from the data projected onto each direction, the projected data being one dimensional. Hence, as can be seen from (3a), PP can be applied to compute the weighting vectors in continuum regression. Note that a direction is characterized by a unit vector $\boldsymbol{a}$, and the data projected on it are given by $\boldsymbol{Xa}$. To sumarize, projection pursuit comes down to scanning all possible directions and computing an estimate of the criterion to be maximized for each direction. The direction which yields the maximal value for the criterion is the solution to the

maximization problem. When all possible directions are thus scanned, the solution obtained is exact. However, in practice, only a limited number of directions can be considered, so that the final solution obtained is only an approximation. Not only the accuracy of the solution obtained, but also the computation time required strongly depends on the number k of directions scanned. As it is very unlikely that the maximum should be found in a direction of the p dimensional space where no data are present, we propose to construct the k directions ($k \geq n$) to be considered as k arbitrary linear combinations of the data points at hand (the first n directions being the directions given by the n observations available).

As an illustration, the regression coefficients estimated by our projection pursuit algorithm (using the classical sample covariance and variance and for $\delta = 0.5$) for the first 24 observations of the (mean-centered) "Fearn" data [14] have been computed. Here and elsewhere in the article, computations were carried out in the Octave programming environment (University of Wisconsin, USA)[1]. The estimates obtained by our projection pursuit algorithm are compared in Table 1 to the SIMPLS [15] regression coefficients, corresponding to the exact solution when using the standard sample variance and covariance in the criterion. The numerical values obtained for both estimates are very similar indicating that the approximation is satisfactory. One might object to this statement that a relative difference of about ten percent can be observed, but in fact the main goal of any regression technique is prediction, which does not deteriorate when approximation errors of this order of magnitude

[1]The Octave m-files can be obtained upon request from the corresponding author.

affect the regression coefficients. An interesting question to address is in which case

Table 1: Regression coefficients of CR-PP ($\delta = 0.5$) compared to the SIMPLS regression for the Fearn data, 10^4 generated directions.

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
CR-PP	0.0039	0.0137	-0.0489	-0.0252	0.0021	0.0281
	0.0039	0.0417	0.1397	0.1287	0.0082	0.0017
	0.0041	0.0386	0.1317	0.1472	0.2459	0.2349
	0.0029	-0.0190	-0.1856	-0.1888	-0.2213	-0.2404
	0.0040	-0.0375	0.0098	0.0135	0.0104	0.0118
	0.0017	0.0036	-0.0239	-0.0830	-0.0460	-0.0356
SIMPLS	0.0039	0.0132	-0.0430	-0.0173	0.0007	0.0281
	0.0039	0.0432	0.1383	0.1291	0.0108	0.0017
	0.0041	0.0401	0.1284	0.1414	0.2443	0.2349
	0.0029	-0.0210	-0.1903	-0.1945	-0.2201	-0.2404
	0.0040	-0.0377	0.0100	0.0140	0.0104	0.0118
	0.0019	0.0035	-0.0193	-0.0800	-0.0472	-0.0356

the projection pursuit approximation might produce erroneous estimates. The convergence properties of projection pursuit have been described for related estimators (see e.g. [9, 10]) and hence we can assume similar convergence behaviour for the projection pursuit estimator presented here, implying that erroneously approximate

estimates will only occur in situtations where it is not viable to apply continuum regression *in se*.

Although in times of increasing computational power, computation times are not a convincing argument to opt for a certain method, robust estimators still frequently suffer from the drawback of a high computational cost making the methods less attractive for routine use. Hence, it is a necessary step to give the reader an idea of the computational performance of CR-PP. For the dataset used in the previous example, we computed the computation times of the estimator on a PC with a 2.2 GHz processor for a different number of directions constructed. The goal of Figure 1 is to show the dependency of the computation time of the method on the number k of constructed directions. From Table 1 we concluded that $k = 10^4$ directions results in a very good approximation of the exact solution. From Figure 1 it can be seen that the computation time for this choice of k remains sufficiently low.

[Figure 1 about here]

Another important factor affecting the computational cost is the dimension of the data matrix $\mathbf{X}$. As heeded in the introduction, a case occurring frequently in practice is the case where $p \gg n$. When regressors are very high-dimensional, it is standard to carry out a data compression before the PP algorithm itself. This is accomplished as follows: we carry out a full singular value decomposition on $\mathbf{X}^T$ such that:

$$\mathbf{X}^T = \mathbf{V}\boldsymbol{D}\mathbf{U}^T \tag{5}$$

In the case where $p \gg n$, it is well known that the matrices $\mathbf{V}$ and $\boldsymbol{D}$ take on the partitioned form:

$$\mathbf{V} = \begin{pmatrix} \tilde{\mathbf{V}} & \mathbf{0}_{p-n} \end{pmatrix} \tag{6a}$$

and

$$\boldsymbol{D} = \begin{pmatrix} \tilde{\boldsymbol{D}} \\ \\ \mathbf{0}_{p-n} \end{pmatrix} \tag{6b}$$

Now the projection pursuit algorithm is run with the modified data matrix $\tilde{\mathbf{X}} = \boldsymbol{U}\tilde{\boldsymbol{D}}^T$, reducing the dimension of the directions to be constructed to n instead of p, saving computational effort. The regression coefficients, identical to those computed without prior data compression, are given by

$$\hat{\boldsymbol{\beta}}_{\delta,h} = \tilde{\mathbf{V}}\tilde{\hat{\boldsymbol{\beta}}}_{\delta,h} \tag{7}$$

where $\tilde{\hat{\boldsymbol{\beta}}}_{\delta,h}$ are the regression coefficients relating $\tilde{\mathbf{X}}$ and $\boldsymbol{y}$. Including a data compression makes the method virtually independent of the dimension p of the data matrix. It can be concluded that only at a very high number of directions k considered the method becomes computationally intensive. Thus, CR-PP is fit for quotidian routine applications.

3.2 Robust continuum regression by projection pursuit (RCR-PP)

In order to obtain a robust estimate of the continuum regression vector of regression coefficients, the only adaptation to the PP algorithm that in principle has to be

done, is to alter the maximization criterion. This means that we evaluate Criterion (3) with robust measures of covariance and variance at k constructed directions and conclude that the point maximizing the criterion is the robust continuum regression weighting vector. Note that when using robust variances and covariances it is not possible to solve the optimization problem by a closed formula, and one needs to resort to approximations like the PP algorithm described before.

As robust counterpart one could consider a robust estimate of the joint covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{y}$ and decompose this matrix into the parts needed in (3). However, these estimates often require either $n > 2p$ or a high computational cost. Fortunately, we can take advantage here of the PP formulation of the problem: one only needs to compute robust variances of univariate variables or covariances between a pair of univariate variables. Simple robust estimators are given by an α-trimmed covariance between $\mathbf{X}\boldsymbol{a}$ and $\boldsymbol{y}$ and an α-trimmed variance of $\mathbf{X}\boldsymbol{a}$ in Equation (3a). Here α ($0 < \alpha < 0.5$) determines the trimming proportion. The α-trimmed covariance between two data vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with n univariate observations is defined as follows. First the trimmed means $\bar{x}_\alpha$ and $\bar{y}_\alpha$ of both data vectors are computed by dropping the smallest and largest l observations and computing the average of the remaining $n - 2l$ observations, where $l = [n\alpha] + 1$. (Here $[k]$ gives the smallest integer larger than k. The α-trimmed covariance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is then computed as

$$\mathrm{Cov}_\alpha(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n - 2l} \sum_{i=l+1}^{n-l} z_{(i)} \quad \text{with} \quad z_i = (x_i - \bar{x}_\alpha)(y_i - \bar{y}_\alpha) \tag{8}$$

and $z_{(1)} \leq \ldots \leq z_{(n)}$ are the cross-products z_i sorted from smallest to largest. The α-

trimmed variance is obtained by setting $\boldsymbol{x} = \boldsymbol{y}$ in the above formula. The parameter α determines the robustness of the procedure: a high value for α makes the method more robust to outliers. On the other hand, a high value of α implies that one deviates more from the usual definition of sample variance and covariance, yielding a loss in efficiency in the statistical sense, i.e. one may expect the estimates to be prone to a higher variance (at least at normal models). Unless otherwise stated, throughout the paper we took $\alpha = 0.1$ as a good compromise between robustness and efficiency.

Once the weight matrix $\boldsymbol{W}_h$ is computed, we can proceed with the regression model (2) since $\boldsymbol{T}_h = \boldsymbol{X}\boldsymbol{W}_h$. Of course, we will not use the least squares estimator explained in Section 2, but we perform robust multiple linear regression of $\boldsymbol{y}$ on $\boldsymbol{T}_h$. We denote the estimated parameters by $\hat{\boldsymbol{\xi}}_{\delta,h}^{RCR}$. In analogy to Equation (4), the robust estimator of the regression coefficients $\boldsymbol{\beta}$ is obtained by $\hat{\boldsymbol{\beta}}_{\delta,h}^{RCR} = \boldsymbol{W}_h \hat{\boldsymbol{\xi}}_{\delta,h}^{RCR}$.

Note that the robust regression to be performed is a robust regression of an n vector on an $n \times h$ matrix. Since h will in practical applications be of modest size, virtually *any* robust regression method can be used here. In our implementation, we opted to use a Huber M-regression [16] estimator, but this method can be replaced by any other robust regression method (see e.g. [17]).

A final aspect to discuss is the construction of the weighting vectors in case $h > 1$. In order to comply with the second side condition of the maximization criterion (in Equation (3b)), a deflation of the original data matrix is carried out such that the

estimated score vectors are uncorrelated. This is done in the usual way [2]:

$$\boldsymbol{E}_i = \left(\boldsymbol{I}_n - \sum_{j=1}^{i-1} \frac{\hat{\boldsymbol{t}}_j \hat{\boldsymbol{t}}_j^T}{\hat{\boldsymbol{t}}_j^T \hat{\boldsymbol{t}}_j} \right) \boldsymbol{X} \tag{9}$$

for $i > 1$. In order to obtain the weighting vector $\hat{\boldsymbol{w}}_i$, the algorithm is run with as inputs the deflated data matrix $\boldsymbol{E}_i$ and the response vector $\boldsymbol{y}$.

4 Selection of the optimal δ and h

The optimal values for δ and h are usually determined by dint of cross-validation [3]. An adapted criterion has been reported which allows to determine the optimal δ analytically [6, 18]. However, as this is only applicable to classical continuum regression, it will not be usable for the robust version of continuum regression we present here and will henceforth be disregarded. For robust continuum regression, we propose also to use cross-validation, although in a slightly modified way. Different types of cross-validation exist. In the context of classical PLS regression, it has been reported that the correct number of factors (optimal h) is only found by means of a full cross-validation (i.e. leave-multiple-out cross-validation with random repeats); simpler approaches such as leave-one-out cross-validation have been shown to over-estimate the optimal number of factors [19]. Hence, it is obvious that also in the case of robust continuum regression, a variant of full cross-validation should be implemented. However, full cross-validation may lead to an erroneous estimate in case outliers are present in the data. As in cross-validation random subsets are selected from the data, it is highly probable that in the selected subset to be

predicted, outliers will be present. Prediction errors for these outliers will be large, since they are not coming from the same model being estimated by the calibration sample. Hence, a robust cross-validation must be performed, essentially in the same way as the classical cross-validation, except that we propose to compute a trimmed mean squared error (MSE) as a measure to evaluate the predictive performance. For example, a 20%-trimmed mean squared error does not take into account the largest 20% of the squared errors when computing the MSE. Note that the choice for a trimmed variance and covariance in the criterion of RCR is in principle independent of the choice for the trimmed mean squared error in the cross-validation procedure. In Section 6 we will illustrate the selection procedure for δ and h on a real data set.

5 Simulation study

In the current section, we will show the robust continuum regression estimator to be resistant against outlying observations by means of a simulation study. It would also be possible to prove the method's robustness properties by theoretical arguments such as the influence function. However, even the influence function of classical continuum regression itself has not yet been established, except in the special cases of $\delta = 0$ [20] and $\delta = 0.5$ [21]. In the current paper, we will limit ourselves to show the robustness properties of the method by simulations.

We generated a data matrix $\mathbf{X}$ of size $n \times p$ according to a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{C}$. Without loss of generality, $\boldsymbol{C}$ is

taken as a diagonal matrix and we selected diagonal elements $\{1, 1/2, \ldots, 1/p\}$. The matrix $\boldsymbol{W}_h$ is constructed in such a way that for $\boldsymbol{T}_h = \boldsymbol{X}\boldsymbol{W}_h$ the model constraints (3b) are fulfilled, the values for $\boldsymbol{\xi}$ in the latent regression model (2) were generated from a uniform random distribution in $[-1, 1]$. All these generated matrices are fixed for a particular simulation setup, so we work with a fixed-designed regression. We simulate from the regression model (2) by generating m different error terms $\boldsymbol{\varepsilon}$. The distribution of the error terms is chosen to be

a) a standard normal distribution $N(0, 1)$,

b) a Student's t distribution t_2 with 2 degrees of freedom,

c) the outlier generating model $0.8 \cdot N(0, 1) + 0.2 \cdot N(15, 1)$.

The latter distribution will be denoted by "O", and is a typical model for extreme shift outliers. The student t_2 distribution has heavier tails than a normal distribution and can be considered as generating moderate size outliers. The normal distribution is the uncontaminated model distribution.

To keep the influence of $\boldsymbol{\varepsilon}$ small, we multiply the generated values by 0.1. Since

$$\boldsymbol{y} = \boldsymbol{T}_h\boldsymbol{\xi} + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{W}_h\boldsymbol{\xi} + \boldsymbol{\varepsilon}, \tag{10}$$

we know the true regression parameter $\boldsymbol{\beta} = \boldsymbol{W}_h\boldsymbol{\xi}$ in the original regression model (1), and can make a comparison with the estimated regression parameters by the mean squared error (MSE) defined as

$$MSE(\hat{\boldsymbol{\beta}}_{\delta,h}) = \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\delta,h}^{(i)}\right)^T\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\delta,h}^{(i)}\right). \tag{11}$$

Here $\hat{\boldsymbol{\beta}}_{\delta,h}^{(i)}$ is the estimated vector of regression coefficients in the i-th simulation for classical or robust CR. The number of simulation replications was $m = 300$.

Table 2 shows the simulation results for a situation with more variables than observations ($n = 30$, $p = 300$). We used a model with 3 latent variables, and thus the results were computed for $h = 3$. In all simulations we used a trimming constant $\alpha = 0.1$ for computing the robust covariances and variances in the objective function of robust CR based on projection pursuit. The algorithm RCR-PP was applied by considering $k = 1000$ projection directions. In general, the resulting MSEs are

Table 2: Comparison of MSE for Continuum Power Regression (CPR) and Robust Continuum Regression with a Projection Pursuit algorithm (RCR-PP) for simulated data of dimensions 30×300 with a true latent structure of $h = 3$. The error term was simulated from a $N(0, 1)$, a t_2 distribution, and an extreme outlier generating distribution O.

δ		0.1	0.25	0.5	0.75	0.9
$N(0,1)$	CPR	0.084	0.083	0.061	0.045	0.047
	RCR-PP	0.128	0.123	0.118	0.133	0.152
t_2	CPR	0.781	0.777	0.573	0.093	0.077
	RCR-PP	0.139	0.147	0.121	0.074	0.056
O	CPR	55.544	55.304	46.134	11.223	2.890
	RCR-PP	1.784	1.7620	1.633	1.284	1.470

smaller for higher values of δ. The optimal choice of δ depends on the data, and in this simulation scheme a higher value of δ is preferable. For normally distributed errors, the loss in MSE of RCR-PP with respect to CPR is rather limited, except for higher values of δ where there is a price to pay for the robustness of RCR-PP.

For the t_2 distribution we clearly see the advantage of the robust method over the classical. This becomes even more visible for the outlier contamination scheme. Note that for all considered values of δ the robust procedure performs better, and the difference in MSE is very pronounced for the smaller values of δ. Continuum Power Regression becomes even completely unreliable for values of δ up to 0.5.

In a next simulation we were interested in a configuration where $n > p$. We chose $n = 60$ observations and $p = 30$ variables of the $\boldsymbol{X}$ matrix. The resulting MSEs are presented in Table 3. Also for these simulated data, the value of δ should be chosen to be larger than 0.5. In general, the results support the same conclusions as before. This suggests that the algorithms are suitable for both situations $n > p$ and $n < p$.

6 Example

In order to illustrate the methodology proposed in the current article, we show the results of robust continuum regression applied to an X-Ray analysis of hydrometallurgical solutions. The data have previously been described in [22]. In order to obtain quantitative results within a reasonable time span, PLS calibration and quantification were successfully applied. PLS is not frequently applied to X-Ray

Table 3: As Table 2, but now for simulated data of dimensions 60×30 .

δ		0.1	0.25	0.5	0.75	0.9
$N(0,1)$	CPR	0.145	0.130	0.039	0.034	0.034
	RCR-PP	0.048	0.048	0.044	0.043	0.044
t_2	CPR	2.214	1.934	0.472	0.076	0.067
	RCR-PP	0.113	0.098	0.077	0.070	0.069
O	CPR	147.380	138.955	74.6269	4.060	2.414
	RCR-PP	4.723	3.591	1.953	1.522	2.121

spectrometry because classical data handling, which consists of a spectral analysis (i.e. net peak area estimation) and subsequent application of a calibration model based on the physical properties of the (X-ray) method, yields more precise results. However, in this case classical analysis was considered to be overly time-consuming.

We use the data matrix consisting of 22 samples as proposed by Lemberge et al [22]. Concentrations of copper, nickel and arsenic had to be predicted. In the current paper, we take as an example the calibration for arsenic, as the process for the two remaining elements is analogous.

The data matrix has not been analyzed with respect to the presence of possible outliers. However, it can be expected that two "outliers" in the statistical sense will be present in the data, as the last two samples had on purpose been chosen to lie slightly outside the calibration range. As outliers do not have a pernicous

effect on calibration by RCR, it is not necessary to run an entirely robust outlier detection technique before doing the RCR calibration. Howbeit, a computationally fast outlier detection technique for classical Partial Least Squares (corresponding to the central value $\delta = 0.5$) gives an idea which value for the trimming constant α should be chosen in calibration. As detection technique we use the Squared Influence Diagnostic plot [21], which is based on the influence function of PLS for each observation. In Figure 2 the SID is shown for all observations in the data set.

[Figure 2 about here]

It is observed that indeed the last two samples can be considered outlying. Moreover, one "true" outlier is also present in the data (observation 9). To give enough safeguard against these outliers, the parameter α for computing the trimmed variances and covariances in the objective function was set equal to 0.1.

For selecting the optimal values of δ and h we proceeded as outlined in Section 4. The data will be split up in half, the first half being taken as the calibration set, whereas the second half will be taken as the validation set. In the classical PLS calibration, the optimal model dimensionality was estimated by means of full cross validation and the optimal number of components was found at 4. In the robust case, a 20% trimmed cross-validation was carried out for h ranging from 1 to 6 and δ ranging from 0.1 up to 0.9. The (trimmed) root mean squared errors of cross validation are shown in Figure 3.

[Figure 3 about here]

Based on this cross-validation, we conclude that the optimal model complexity is found at 5 latent variables and the optimal δ in this case equals 0.11.

In order to compare both approaches, we computed the PLS and RCR vectors of regression coefficients at their respective optimal model complexities and hence computed the predicted values for observations in the validation set (for the data considered here, computation of the RCR regression coefficients took about 7 s). The obtained root trimmed mean squared errors of prediction equalled 0.514 for PLS and 0.415 for RCR (computed with the PP-algorithm with $k = 10^4$), respectively, which amounts to a relative gain of about 25% when using the robust method. The difference between the squared prediction errors turns out to be significant (using a sign-test, being more robust than a standard t-test). Note that the two pseudo outliers will not be well predicted by the robust method, and are trimmed away when computing the trimmed MSE. The robust method is meant to fit the majority of the data well, and not the outliers. The standard PLS predictions tries to predict all observations, and will give better predictions for the two pseudo outliers, but not for the main part of the samples.

7 Conclusions

In the current paper we proposed a framework for robust continuum regression. Application of robust continuum regression to contaminated data sets should combine

the benefits of a robust calibration technique to the versatility of continuum regression. Continuum regressions allows for a better fit and more predictive power by finding the optimal point in a continuum range of models from OLS over PLS to PCR. Robustness and efficiency of the method have been corroborated both by a simulation study and by an example.

We provided an algorithm based on projection pursuit, which has been shown to be efficient in the computational sense. Even for a large number of directions to be scanned, e.g. 10^4, the estimator can still be computed within a reasonable time span. In any practical application, however, the time consuming step is the robust cross-validation phase. Computation times for cross-validation depend of the number of iterations that is considered sufficient, as well as of the intervals of the continuum parameter at which one wants to evaluate the estimator and of course also of the size of the data. For spectrometric data, even moderate settings of both tunable parameters may require computation times of about an hour.

Simulations have shown that for normally distributed data, the MSE for the robust methods are fairly close to those obtained with the classical estimator, indicating that the RCR proposed here has a reasonable statistical efficiency. Analogous simulations for non-normal data, including data containing outliers, yielded a vast decrease in Mean Squared Error for the robust approach compared to its classical techniques, leading to the conclusion that the methodology proposed here is indeed robust.

RCR is proposed as a continuum regression framework of which the estimator

corresponding to $\delta = 0.5$ is a new robust partial least squares estimator. Howbeit, the goal of the current paper was not to design a new robust PLS estimator, albeit the latter is an implicit consequence. One can expect that a technique which has specifically been designed as a robust PLS technique, should yield better results than RCR. This disadvantage is compensated for by the fact that RCR is more versatile and allows to vary δ. In any practical analysis where the optimal δ does not equal 0.5, such as in the practical example, a lower RMSEP can be obtained by using RCR at the optimal delta value than by applying any PLS estimator.

Hitherto, we proposed a robust continuum regression estimator for univariate $\boldsymbol{y}$. In some applications it may be interesting to have at hand a multivariate version of the estimator. This is beyond the scope of this paper, but we believe that the main ideas of this paper can be generalized to this multivariate setting.

Acknowledgements

Research financed (in part) by a PhD Grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen), and (in part) by the "Fonds voor Wetenschappelijk Onderzoek" and the Research Fund K.U.Leuven.

References

[1] C.F. Gauß, Werke, 4 (1826), 1-93.

[2] H. Wold, in: P.R. Krishnaiah (ed.), Multivariate Analysis III, Academic Press, New York (1973), 383-407.

[3] M. Stone, R.J. Brooks, J. R. Statist. Assoc. B, 52 (1990), 237-269.

[4] O.K. Chung, J.B. Ohm, M.S. Caley, B.W. Seabourn, Cereal Chem., 78 (2001), 493-497.

[5] M.C. Ortiz, J. Arcos, L. Sarabia, Chemom. Intell. Lab. Syst., 34 (1996), 245-262.

[6] J. Malpass, D. Salt, M. Ford, Pestic. Sc., 46 (1996), 282-284.

[7] S. de Jong, B.M. Wise, N.L. Ricker, J. Chemometr., 15 (2001), 85-100.

[8] J.H. Friedman, J.W. Tukey, IEEE Trans. Comp. Ser. C, 23 (1974),881-889.

[9] P.J. Huber, Ann. Statist., 13 (1985), 435-475.

[10] G. Li, Z. Chen, J. Am. Statist. Assoc., 80 (1985), 759-766.

[11] J. Xie, J. Wang, Y. Liang, L. Sun, R. Yu, J. Chemometr., 7 (1993), 527-541.

[12] C. Croux, A. Ruiz-Gazen, High-Breakdown Estimators for Principal Components, The Projection Pursuit Approach Revisited, to appear in J. of Multivariate Analysis.

[13] J.A. Branco, C. Croux, P. Filzmoser, M.R. Oliviera (2005), " Robust Canonical Correlations: A Comparative Study", Computational Statistics, to appear.

[14] T. Fearn, Appl. Stat., 32 (1983), 73.

[15] S. de Jong, Chemometr. Intell. Lab. Syst., 18 (1993), 251-263.

[16] P.J. Huber, Robust Statistics, Wiley, NY (1981), 153-198.

[17] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley and Sons, New York, 1987.

[18] J. Malpass, The implementation of continuum regression as a SAS procedure, internal report, University of Portmouth (1996).

[19] K. Baumann, H. Albert and M. von Korff, J. Chemometr. 16 (2002), 339-350.

[20] R.D. Cook and S. Weisberg, Residuals and influence in regression, Chapman and Hall, New York, USA, 1982, pp. 101-148.

[21] S. Serneels, C. Croux, P.J. Van Espen, Chemometr. Intell. Lab. Syst., 71 (2004), 13-20.

[22] P. Lemberge, P.J. Van Espen, X-Ray Spectrom., 28 (1999), 77-85.

Figures

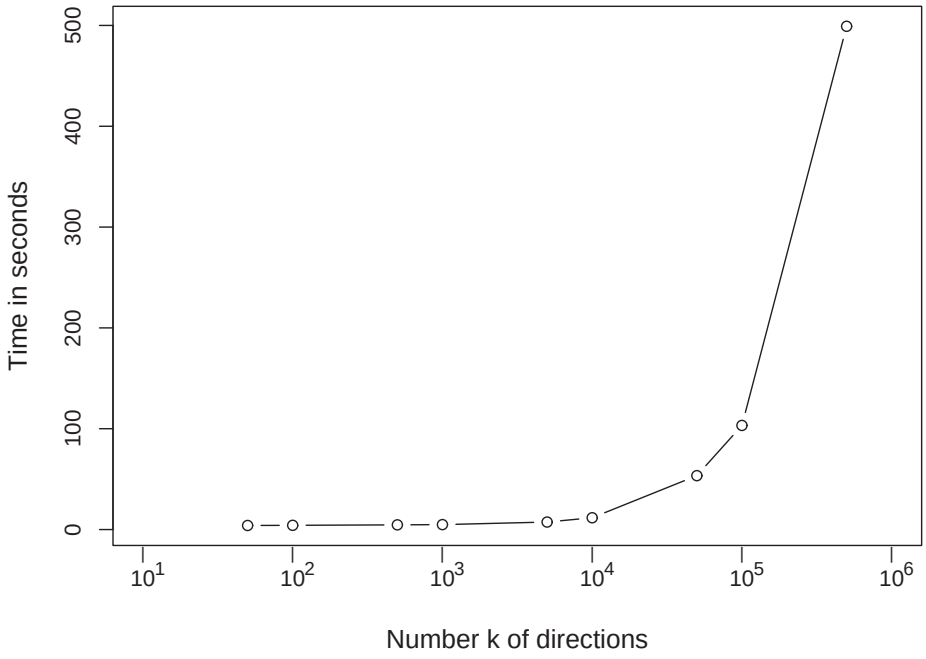

Figure 1: CR-PP computation times vs. increasing values of k

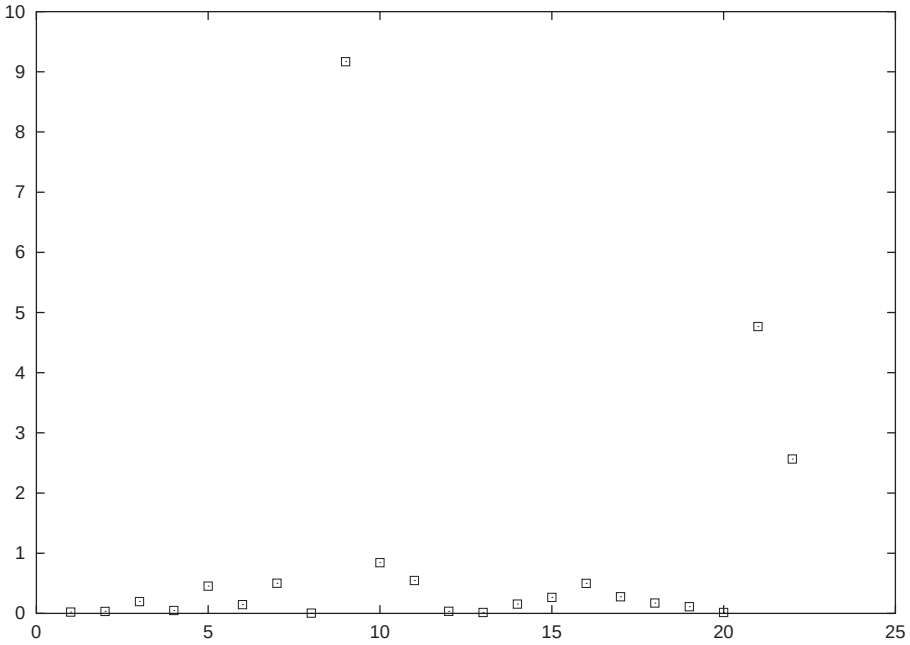

Figure 2: Squared Influence Diagnostic (SID) for all 22 samples of the data set

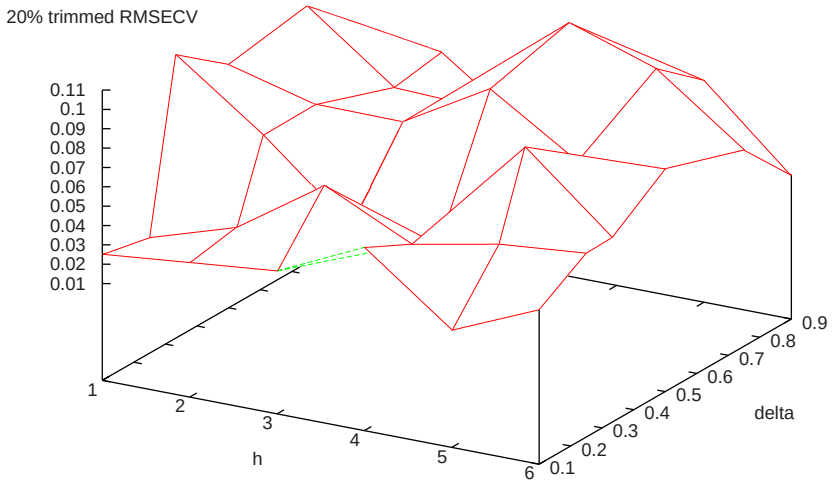

Figure 3: 20% trimmed root mean squared errors of cross-validation for different values of h and δ for RCR with $\alpha = 0.1$