RCMARS: Robustification of CMARS with Different Scenarios under Polyhedral Uncertainty Set

Ayşe Özmen *, Gerhard Wilhelm Weber *
İnci Batmaz, Erik Kropat*****

**Institute of AppliedMathematics, Middle East Technical University, Ankara, Turkey (e-mail: ayseozmen@gmail.com).*
**Institute of AppliedMathematics, Middle East Technical University, Ankara, Turkey (e-mail: gweber@metu.edu.tr).*
*** Department of Statistics, Middle East Technical University, Ankara, Turkey (e-mail: ibatmaz@metu.edu.tr).*
**** Institute for Theoretical Computer Science, Mathematics and Operations Research, Universitat der Bundeswehr München, Neubiberg, Germany, (e-mail: erik.kropat@unibw.de).*

Abstract: Our recently developed CMARS is powerful in handling complex and heterogeneous data. We include into CMARS the existence of uncertainty about the scenarios. Indeed, data include noise in both output and input variables. Therefore, solutions of the optimization problem may reveal a remarkable sensitivity to perturbations in the parameters of the problem. The data uncertainty results in uncertain constraints and objective function. To overcome this difficulty, we refine our CMARS algorithm by a robust optimization technique proposed to cope with data uncertainty. In our previous study, we present the new Robust CMARS (RCMARS) in theory and method and illustrate it with a numerical example (Özmen et al., 2010). In this study, we present RCMARS results with different uncertainty scenarios for our numerical example.

Keywords: Regression, Optimization, Robustness, Regularization, Uncertainty.

1. INTRODUCTION

MARS has been applied successfully to many fields of science, economy and technology in recent years. It bases on a modern methodology from statistical learning, which is important in both regression and classification. MARS builds flexible high-dimensional nonparametric regression models, and presents a great promise for fitting nonlinear multivariate functions. It generates an additive model in two-stage process: the forward and backward stepwise algorithms. In CMARS method, the backward stepwise algorithm is not applied. Instead of this, a *Penalized Residual Sum of Squares* (PRSS) is employed for MARS as a *Tikhonov regularization* (TR) problem. This two-objective optimization problem is treated using the continuous optimization technique called *Conic Quadratic Programming* (*CQP*) (Weber et al., 2009).

CMARS is an alternative method to a well-known regression tool MARS from data mining and estimation theory. With this study, we further improve CMARS so that it can treat uncertainty in the data. In fact, generally, data may include noise in both input and output variable. This means that the data of the regression problem are not exactly known or may not be exactly measured, or the exact solution of the problem may not be carried out because of intrinsic inaccuracy of the devices (Boni, 2007). Moreover, the data can undergo small changes by variations in the optimal experimental design. These altogether leads to uncertainty in the objective function and in possible constraints. To handle this, we refine our CMARS algorithm by an important robust optimization developed by Ben-Tal and Nemirovski (2001, 2002), and El-Ghaoui and Lebret (1997).

Robust optimization is a modeling methodology to process optimization problems in which the data are uncertain and are only known to belong to some uncertainty set, except for outliers. The purpose of robust optimization is to find an optimal or near optimal solution which is feasible for every possible realization of the uncertain scenarios (Bertsimas et al., 2008, Werner, 2008).

Below, we firstly analyze how uncertainty incorporated into the CMARS model with complexity terms in the form of integrals of squared first- and second-order derivatives of the model functions and, then, the discretized TR and, finally, the CQP form of the problem. Then, we introduce a *robustification* of CMARS with robust optimization under polyhedral uncertainty and ellipsoidal uncertainty (Özmen et al., 2010, El-Ghaoui, 2003). Because of the computational effort which our robustification of CMARS will easily need, we also present the concept of a *weak robustification*.

This paper is organized as follows. In Section 2, RCMARS is introduced in theory and method. RCMARS results with different uncertainty scenarios for the numerical example studied in our previous study (Özmen et al., 2010) are presented in Section 3. Conclusion and further studies are stated in the last section.

2. RCMARS MODEL

2.1 CMARS Model with Uncertainty

For CMARS, the large model that has the maximum number of basis functions (BFs), $M_{\max}$, is created by Salford MARS (2009). The following general model is considered to represent the relation between the input variables and the response:

$$Y = f(\breve{\boldsymbol{X}}) + \varepsilon, \tag{1}$$

where Y is the response variable, $\breve{\boldsymbol{X}} = (\breve{X}_1, \breve{X}_2, ..., \breve{X}_p)^T$ is a vector of predictor variables, and ε is an additive stochastic component which is assumed to have zero mean and finite variance. The aim is to build reflected pairs for each input variable $\breve{X}_j$ $(j = 1, 2, ..., p)$ with p-dimensional knots $\tau_i = (\tau_{i,1}, \tau_{i,2}, ..., \tau_{i,p})^T$ $(i = 1, 2, \ldots, N)$ at or just nearby each input data vectors. Moreover, $\breve{X}_j$ are assumed to be normally distributed random variables. Here, the following general model is considered for each input $\breve{X}_j$:

$$\breve{X}_j = \overline{X} + \xi_j \quad (j = 1, 2, ..., p).$$

To robustify CMARS, we employ the robust optimization on the BFs provided by the MARS model, and we assume that the input and output variables of our model are all random variables. They lead us to the *uncertainty sets*, which are assumed to contain *confidence intervals (CIs)* (refer to (Özmen et al., 2010) for more details).

MARS method employs expansions of piecewise linear BFs based on the new dataset that have uncertainties. We prefer the following notation for the piecewise linear BFs (Friedman, 1991):

$$c^+(\breve{\breve{x}}, \tau) = (\breve{\breve{x}} - \tau)_+, \; c^-(\breve{\breve{x}}, \tau) = (\breve{\breve{x}} - \tau)_-,$$

where $(q)_+ := \max\{0, q\}$, $(q)_- := \max\{0, -q\}$, and τ is a univariate knot. Incorporating the uncertainty sets $U_1 \subseteq \mathbb{R}^{N \times M_{\max}}$, and $U_2 \subseteq \mathbb{R}^N$, defined in Section 2.3, into the data $(\breve{\breve{\boldsymbol{x}}}_i, \breve{\breve{y}}_i)$ $(i = 1, 2, ..., N)$, the multiplicative form of the mth BF can be written as

$$\psi_m(\breve{\breve{\boldsymbol{x}}}_i) := \prod_{j=1}^{K_m} (\breve{\breve{x}}_{i\kappa_j^m} - \tau_{\kappa_j^m})_{\pm} \quad \text{for } i = 1, 2, ..., N, \tag{2}$$

where K_m is the number of truncated linear functions multiplied in the mth basis function. Then, for the CMARS model with uncertainty, PRSS will have the following representation:

$$\begin{aligned} PRSS \; &:= \sum_{i=1}^{N} (\breve{\breve{y}}_i - f(\breve{\breve{\boldsymbol{x}}}_i))^2 \\ &+ \sum_{m=1}^{M_{\max}} \phi_m \sum_{|\boldsymbol{\theta}|=1}^{2} \sum_{\substack{r<s \\ r,s \in V(m)}} \int \alpha_m^2 [D_{r,s}^{\theta} \psi_m(\boldsymbol{t}^m)]^2 \, d\boldsymbol{t}^m, \end{aligned} \tag{3}$$

$$\boldsymbol{\theta}^T = (\theta_1, \theta_2)$$

where $V(m) := \{\kappa_j^m \mid j = 1, 2, ..., K_m\}$ is the variable set associated with the mth basis function. After using the discretization to approximate the multi-dimensional integrals $\int \alpha_m^2 \left[D_{r,s}^{\theta} \psi_m(t^m) \right]^2 dt^m$ (Weber et al., 2009), our PRSS with uncertainty will be as follows:

$$PRSS \approx \left\| \breve{\breve{\boldsymbol{y}}} - \boldsymbol{\psi}(\breve{\breve{\boldsymbol{b}}}) \boldsymbol{\alpha} \right\|_2^2 + \phi \left\| \boldsymbol{L} \boldsymbol{\alpha} \right\|_2^2. \tag{4}$$

Here, *PRSS* problem looks like a classical TR problem with $\phi \geq 0$, i.e., $\phi = \lambda^2$ for some $\lambda \in \mathbb{R}$. Then, it can be coped with the CQP (Weber et al., 2009). The second (complexity) part of our PRSS model will remain the same as it is in CMARS after we incorporate a *"perturbation"* into the real input data $\breve{\boldsymbol{x}}_i$ in each dimension and into the output data $\breve{y}_i$ because we do not make any changes for the function in the multi-dimensional integrals.

When we estimate the BFs $(\breve{\breve{x}}_{i\kappa_j^m} - \tau_{\kappa_j^m})_{\pm}$ in (2), we can rewrite them as the following term:

$$(\breve{\breve{x}}_{i\kappa_j^m} - \tau_{\kappa_j^m})_{\pm} \leq (\breve{x}_{i\kappa_j^m} - \tau_{\kappa_j^m})_{\pm} + (\Delta_{i\kappa_j^m} + (\pm \mathrm{A}_{i\kappa_j^m}))_{\pm}. \tag{5}$$

Here, $\mathrm{A}_{i\kappa_j^m}$ is interpreted and employed as a *control variable*. Since the value of this control variable directly affect the size of our uncertainty set U_1, and our uncertainty sets are unknown but bounded, $\mathrm{A}_{i\kappa_j^m}$ is restricted by a value $\gamma_{i\kappa_j^m}$. When we consider the conservative (risk averse) case, "worst case" for the value of $\mathrm{A}_{i\kappa_j^m}$, it will be equal to $\gamma_{i\kappa_j^m}$. However, when the absolute value of our uncertainty set is very high, it may take too much time to find a solution or we may not find any meaningful solution for our problems. Therefore, we may consider the risk friendly case to select the value of $\mathrm{A}_{i\kappa_j^m}$ between zero and the absolute value of $\mathrm{A}_{i\kappa_j^m}$,i.e $\breve{\mathrm{A}}_{i\kappa_j^m} \in [0, |\mathrm{A}_{i\kappa_j^m}|]$. Here, to simplify notion, we still preserve the notion $\mathrm{A}_{i\kappa_j^m}$ for $\breve{\mathrm{A}}_{i\kappa_j^m}$. To estimate the values $\psi_m(\breve{\boldsymbol{x}}_i)$ and $\psi_m(\breve{\breve{\boldsymbol{x}}}_i)$, we can employ (5) in the following form where all the "+" and "-" signs belong to each other respectively (Özmen et al., 2010):

$$\begin{aligned} \underbrace{\prod_{j=1}^{K_m} (\breve{\breve{x}}_{i\kappa_j^m} - \tau_{\kappa_j^m})_{\pm}}_{=:\psi_m(\breve{\breve{\boldsymbol{x}}}_i)} &\leq \underbrace{\prod_{j=1}^{K_m} (\breve{x}_{i\kappa_j^m} - \tau_{\kappa_j^m})_{\pm}}_{=:\psi_m(\breve{\boldsymbol{x}}_i)} \\ &+ \sum_{\substack{A \subseteq \{1, ... K_m\} \\ \neq}} \prod_{a \in A} (\breve{x}_{ia} - \tau_a)_{\pm} \prod_{b \in \{1, ... K_m\}/A} ((\pm \mathrm{A}_{ib}) + \Delta_{ib})_{\pm} \end{aligned}$$

$$(i = 1, 2, ..., N).$$

Here, we can obtain the bounding form given below with symmetry:

$$\psi_m(\breve{\breve{x}}_i) - \psi_m(\breve{x}_i) \le \hat{u}_{im}$$
$$\underbrace{\psi_m(\breve{x}_i) - \psi_m(\breve{\breve{x}}_i) \le \hat{\hat{u}}_{im}}_{\Downarrow}$$
$$\left|\psi_m(\breve{\breve{x}}_i) - \psi_m(\breve{x}_i)\right| \le \max\{\hat{u}_{im}, \hat{\hat{u}}_{im}\}.$$

So our uncertainty value $|u_{im}|$ can be estimated in the following way for each BF (Özmen et al., 2010):

$$|u_{im}| \le \sum_{\substack{A \subseteq \{1,\dots,K_m\} \\ \neq}} \mathrm{B}_i^{|A|-1} \prod_{a \in A} \rho_{ia} \prod_{b \in \{1,\dots,K_m\}/A} (\gamma_{ib} + \rho_{ib}). \qquad (6)$$

Here, with $|A|$ we denote the cardinality size of the set *A*. B_i is also considered to be and applied as a *control variable*. The value of B_i is equal to 2 in cases without outliers, but for outliers, it will be greater than 2. For these cases, we will have to select different values for B_i. When we consider the conservative case, we do not want to ignore any outliers. Therefore the values of B_i may be very large for some variables in the input data, and the absolute values of our uncertainty set may be very high because of the values of this control variable. If the absolute value of our uncertainty set is very high, it may take too much time to find a solution or we may not find any meaningful solution for our problem. Consequently, instead of the conservative case, we may consider a more risk friendly case to select the values of B_i for the outlier case. For visualization, see Fig. 1:

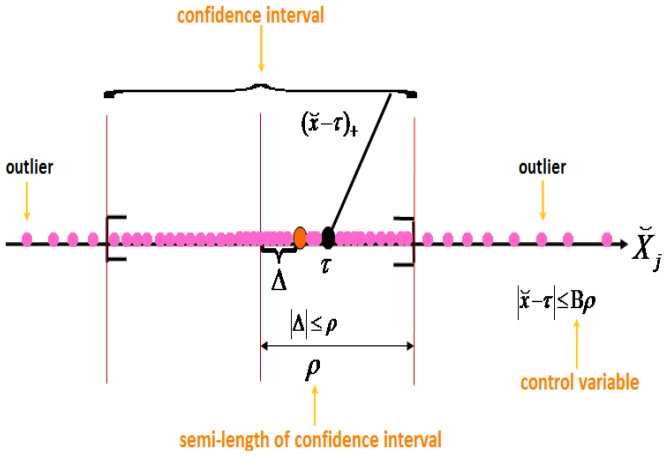

Fig. 1. The CIs of perturbation Δ and $|\breve{x}-\tau|$.

2.2 Robustification of the CMARS Model

The CMARS model depends on parameters. Small perturbations in data can result in different model parameters. This may cause unstable solutions. In CMARS, the purpose is to decrease the estimation error while keeping efficiency as high as possible. To achieve this purpose, we apply some approaches such as usage of more robust estimators, scenario optimization and robust counterpart. Using robustification in CMARS, we aim to reduce the estimation variance.

To make reduction in the complexity of the regression method MARS, which especially means sensitivity with respect to noise in the data, we do a penalization in the form of TR and study it as a CQP problem in CMARS model. Regularization from CMARS is already some kind of robustification, however, in our study, we additionally robustify CMARS with the help of the robust optimization approach, which is some kind of regularization in the input and output domain. Therefore, we have some changes in the part of $\left\|\boldsymbol{\psi}(\breve{\boldsymbol{b}})\boldsymbol{\alpha} - \breve{\boldsymbol{y}}\right\|_2^2$, when we do our robustification of CMARS for both the input and output variables by including uncertainty with the help of robust optimization. We, however, need not any change in the integration function of complexity part of PRSS model in the equation (3). Therefore, the part of $\|\boldsymbol{L\alpha}\|_2^2$ is the same as in CMARS.

2.3 Selecting the Shape of Uncertainty Sets

The robust optimization approach is based on making the optimization models robust regarding constraint violations by solving *robust counterparts* of these problems in prespecified *uncertainty sets* for the uncertain parameters (Fabozzi et al., 2007). These uncertainty sets base on statistical estimates and probabilistic guarantees on the solution. The robust optimization problem can be solved efficiently when the uncertainty set has a special shape (Fabozzi et al., 2007). These special shapes for uncertainty sets can be either *ellipsoidal* or *polyhedral*.

If ellipsoidal uncertainty sets are employed, robustification is more successful than employing of polyhedral uncertainty sets (Schöttle et al., 2006). Nevertheless, using ellipsoidal uncertainty sets increase the complexity of optimization problems. In this paper, we study our robust CQP (second order optimization problem (SCOP)) and we shall find out that it remains CQP. Therefore, we will guarantee polyhedral uncertainty sets by an interval concept for input and output data in our model; our robust CQP (SCOP) will be traced back directly as a CQP. Therefore, in this paper, we only focus on *polyhedral uncertainty* with different uncertain scenarios.

2.4 Polyhedral Uncertainty and Robust Counterpart for the CMARS Model

To study the robustness problem, we assume that the given model uncertainty is represented by a family of matrices $\boldsymbol{\psi}(\breve{\breve{\boldsymbol{b}}}) = \boldsymbol{\psi}(\breve{\boldsymbol{b}}) + \boldsymbol{U}$ and vectors $\breve{\breve{\boldsymbol{y}}} = \breve{\boldsymbol{y}} + \boldsymbol{v}$, where $\boldsymbol{U} \in U_1$ and $\boldsymbol{v} \in U_2$ are unknown but bounded sets. Here, the uncertainty matrix $\boldsymbol{U} \in U_1$ and uncertainty vector $\boldsymbol{v} \in U_2$ defined by

$$\boldsymbol{U} = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1M_{\max}} \\ u_{21} & u_{22} & \dots & u_{2M_{\max}} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N1} & u_{N2} & \dots & u_{NM_{\max}} \end{bmatrix}, \boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}. \quad (7)$$

Since we do not want to increase the complexity of our optimization problems, we select the uncertainty sets U_1 and U_2 as of type *polyhedral* for both input and output data in our model to study our robustness problem. Based on these sets, the robust counterpart is defined as follows:

$$\min_{\boldsymbol{\alpha}} \max_{\substack{\boldsymbol{W}\in U_1 \\ \boldsymbol{z}\in U_2}} \left\|\boldsymbol{W}\boldsymbol{\alpha}-\boldsymbol{z}\right\|_2^2 + \phi\left\|\boldsymbol{L}\boldsymbol{\alpha}\right\|_2^2.$$

Here, U_1 is a polytope with $2^{N\cdot M_{\max}}$ vertices $\boldsymbol{W}^1, \boldsymbol{W}^2, \dots, \boldsymbol{W}^{2^{N\cdot M_{\max}}}$. It is not exactly known, but belongs to a convex bounded uncertain domain U_1 given by

$$U_1 = \left\{ \sum_{j=1}^{2^{N\cdot M_{\max}}} \delta_j \boldsymbol{W}^j \,\middle|\, \delta_j \ge 0 \ (j \in \{1,2,\dots,2^{N\cdot M_{\max}}\}), \sum_{j=1}^{2^{N\cdot M_{\max}}} \delta_j = 1 \right\}, \quad (8)$$

where $U_1 = \text{conv}\{\boldsymbol{W}^1, \boldsymbol{W}^2, \dots, \boldsymbol{W}^{2^{N\cdot M_{\max}}}\}$ is the convex hull. Furthermore, U_2 is a polytope with 2^N vertices $\boldsymbol{z}^1, \boldsymbol{z}^2, \dots, \boldsymbol{z}^{2^N}$. It is not exactly known, but belongs to a bounded uncertain domain U_2 given by

$$U_2 = \left\{ \sum_{i=1}^{2^N} \varphi_i \boldsymbol{z}^i \;\middle|\; \varphi_i \ge 0 \ (i \in \{1,2,\dots,2^N\}), \sum_{i=1}^{2^N} \varphi_i = 1 \right\}, \quad (9)$$

where $U_2 = \text{conv}\{\boldsymbol{z}^1, \boldsymbol{z}^2, \dots, \boldsymbol{z}^{2^N}\}$ is the convex hull.

Any uncertainty sets U_1 and U_2 can be represented as a convex combination of vertices $\boldsymbol{W}^j$ (j=1,2,…, $2^{N\times M_{\max}}$) and $\boldsymbol{z}^i$ (i=1,…, 2^N) of the polytope. The entries are found to have become intervals. Therefore, our matrix $\boldsymbol{W}$ and vector $\boldsymbol{z}$ with uncertainty are lying in the Cartesian product of intervals that are parallelpipes. To give an easy illustration, the Cartesian product of intervals in general and, especially, for three entries can be represented by the Fig. 2.

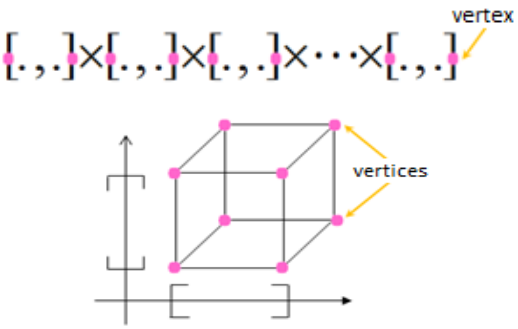

Fig. 2. Cartesian product of intervals for three entries.

Here, the matrix $\boldsymbol{W}$ is represented as a vector with uncertainty which generates a parallelpipe. We have a $(N \times M_{\max})$-matrix $\boldsymbol{W} = (w_{ij})_{\substack{i=1,2,\dots,N \\ j=1,2,\dots,M_{\max}}}$ and we can write it as a vector $\boldsymbol{t} = (t_k)_{k=1,2,\dots,N\times M_{\max}}$, where $t_k := u_{ij}$ with $k = i+(j-1)N$. So, our $\boldsymbol{W}$ matrix can be canonically represented as a vector $\boldsymbol{t}_k = (t_1, t_2, \dots, t_{N\times M_{\max}})^T$ by successively aligning the columns of $\boldsymbol{W}$.

2.4 Robust CQP with the Polyhedral Uncertainty

For our CMARS model, the optimization problem is written as follows:

$$\begin{aligned} &\min_{t,\boldsymbol{\alpha}} t, \\ &\text{subject to } \left\|\boldsymbol{\psi}(\breve{\boldsymbol{b}})\boldsymbol{\alpha} - \breve{\boldsymbol{y}}\right\|_2 \le t, \\ &\qquad \left\|\boldsymbol{L}\boldsymbol{\alpha}\right\|_2 \le \sqrt{\breve{M}}. \end{aligned}$$

When *polyhedral* uncertainty is used for the CMARS model based on the uncertainty sets U_1 and U_2, the robust counterpart is defined by

$$\min_{\boldsymbol{\alpha}} \max_{\substack{\boldsymbol{W}\in U_1 \\ \boldsymbol{z}\in U_2}} \left\|\boldsymbol{W}\boldsymbol{\alpha}-\boldsymbol{z}\right\|_2^2 + \phi\left\|\boldsymbol{L}\boldsymbol{\alpha}\right\|_2^2.$$

So, the robust CQP for our optimization problem is represented in the following form:

$$\begin{aligned} &\min_{t,\boldsymbol{\alpha}} t, \\ &\text{subject to } \left\|\boldsymbol{W}\boldsymbol{\alpha}-\boldsymbol{z}\right\|_2 \le t \quad \forall \ \underbrace{\boldsymbol{W}}_{=\sum_{j=1}^{2^{N\cdot M_{\max}}} \delta_j \boldsymbol{W}^j} \in U_1, \ \underbrace{\boldsymbol{z}}_{=\sum_{i=1}^{2^N} \varphi_i \boldsymbol{z}^i} \in U_2, \\ &\qquad \left\|\boldsymbol{L}\boldsymbol{\alpha}\right\|_2 \le \sqrt{\breve{M}}. \end{aligned}$$

If U_1 and U_2 are polytopes which are described by their vertices as

$$U_1 = \text{conv}\{\boldsymbol{W}^1, \boldsymbol{W}^2, \dots, \boldsymbol{W}^{2^{N\cdot M_{\max}}}\}, \ U_2 = \text{conv}\{\boldsymbol{z}^1, \boldsymbol{z}^2, \dots, \boldsymbol{z}^{2^N}\},$$

then our robust CQP can be equivalently stated as a standard CQP (El-Ghaoui, 2003) as the following:

$$\begin{aligned} &\min_{t,\boldsymbol{\alpha}} t, \\ &\text{subject to } \left\|\boldsymbol{W}^j\boldsymbol{\alpha} - \boldsymbol{z}^i\right\|_2 \le t \ \ (i = 1,2,\dots,2^N \ ; j = 1,2,\dots,2^{N\cdot M_{\max}}), \\ &\qquad \left\|\boldsymbol{L}\boldsymbol{\alpha}\right\|_2 \le \sqrt{\breve{M}}. \end{aligned} \quad (10)$$

Let us use modern methods of *continuous optimization techniques*, especially, from CQP where the basic notation is employed (Ben-Tal at al., 2001):

$$\min_{x} \ \mathbf{c}^T \boldsymbol{x},$$

subject to $\left\| \boldsymbol{D}_i \boldsymbol{x} - \boldsymbol{d}_i \right\| \le \boldsymbol{p}_i^T \boldsymbol{x} - q_i \qquad (i=1,2,...,k).$

In fact, we see that our optimization problem is a CQP with

$$\boldsymbol{c} = (1, \mathbf{0}_{M_{\max}+1}^T)^T, \ x = (t, \boldsymbol{\alpha}^T)^T, \ \boldsymbol{D}_1 = (\mathbf{0}_N, \boldsymbol{W}^j),$$

$$\boldsymbol{d}_1 = \boldsymbol{z}^i, \ \boldsymbol{p}_1 = (1, 0, ..., 0)^T, \ q_1 = 0,$$

$$\boldsymbol{D}_2 = (\mathbf{0}_{M_{\max}+1}, \boldsymbol{L}), \boldsymbol{d}_2 = \mathbf{0}_{M_{\max}+1}, \ \boldsymbol{p}_2 = \mathbf{0}_{M_{\max}+2} \text{ and } q_2 = -\sqrt{\breve{M}}.$$

In order to write the optimality condition for this problem, we reformulate the problem (10) as follows:

$$\min_{t,\boldsymbol{\alpha}} \ t,$$

such that

$$\boldsymbol{\chi}^{i,j} := \begin{pmatrix} \mathbf{0}_N & \boldsymbol{W}^j \\ 1 & \mathbf{0}_{M_{\max}+1}^T \end{pmatrix} \begin{pmatrix} t \\ \boldsymbol{\alpha} \end{pmatrix} + \begin{pmatrix} -\boldsymbol{z}^i \\ 0 \end{pmatrix},$$

$$\boldsymbol{\eta} := \begin{pmatrix} \mathbf{0}_{M_{\max}+1} & \boldsymbol{L} \\ 0 & \mathbf{0}_{M_{\max}+1}^T \end{pmatrix} \begin{pmatrix} t \\ \boldsymbol{\alpha} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{M_{\max}+1} \\ \sqrt{\breve{M}} \end{pmatrix},$$

$$\boldsymbol{\chi}^{i,j} \in L^{N+1}, \ \boldsymbol{\eta} \in L^{M_{\max}+2},$$

where L^{N+1}, $L^{M_{\max}+2}$ are the $(N+1)$- and $(M_{\max}+2)$-dimensional ice-cream (*or second-order, or Lorentz*) cones, defined by:

$$L^{N+1} := \left\{ \boldsymbol{x} = (x_1, ..., x_{N+1})^T \in \mathbb{R}^{N+1} \,\middle|\, x_{N+1} \ge \sqrt{x_1^2 + x_2^2 + ... + x_N^2} \right\}$$

$$(N \ge 1).$$

3. A NUMERICAL STUDY

The implementation of RCMARS algorithm is illustrated by a numerical example in our previous study (Özmen et al., 2010). In that implementation, first, the largest model is constructed by using Salford MARS Version 3 (2009). Then, to apply the robust optimization technique on the CMARS model, we incorporate a perturbation (uncertainty) into the real input data, $\bar{\boldsymbol{x}}$, in each dimension and into the output data, $\boldsymbol{y}$. For this aim, the uncertainty matrices and vectors based on *polyhedral uncertainty sets* are obtained by using (8) and (9). Then, using the equation (6), uncertainty is evaluated for all input and output values which are represented by CIs. The boundaries of CIs are assumed to be (-3, 3) after the variables are transformed into the standard normal distribution. However, the uncertainty matrix for input data has a huge size, and the computer does not have an enough capacity to solve our problem for this uncertainty matrix. In fact, we have a *tradeoff* between tractability and robustification. To handle this problem, we obtain different *weak RCMARS* (*WRCMARS*) models for each observation, and solve them by using MOSEK program (2008). After we obtain the MOSEK models and find the *t* values for all auxiliary problems, using the *worst-case* approach, we select the solution which has the *maximum t* value. Then, we continue our calculations by using the parameter estimates $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4$ and α_5 obtained from the auxiliary problem which has the highest *t* value (See Özmen et al. (2010) for the details).

In this study, we obtain uncertainty matrices, ***U***, for the input data and uncertainty vectors, ***v,*** for the output data as the form of (7) by using four different intervals which are ±3, ±3e-6, ±3e-7, and as a special case, mid-point value of our interval (i.e. zero length interval). We calculate our parameters with 16 different uncertainty scenarios using these values under polyhedral uncertainty sets. All of the parameter estimates for different uncertainty scenarios are shown in Table 1, 2, 3, and 4. Note here that we defined the values $\sqrt{\breve{M}}$ by a model-free method. When we apply the $\sqrt{\breve{M}}$ values in our RCMARS code and solve by using MOSEK, RCMARS provides us several solutions, but, here, we use the $\sqrt{\breve{M}}$ value which has minimum value of PRSS in the equation (4).

Table 1. Parameter estimates and the model performances I

v	**±3**			
U	**±3**	**±3e-6**	**±3e-7**	**zero**
$\boldsymbol{\alpha_0}$	0.1230	-0.0634	-0.0773	-0.3732
$\boldsymbol{\alpha_1}$	-0.3131	-0.0526	-0.0577	0.0274
$\boldsymbol{\alpha_2}$	0.0000	0.2596	0.3141	0.1136
$\boldsymbol{\alpha_3}$	0.0109	-0.0029	-0.0044	-0.0700
$\boldsymbol{\alpha_4}$	0.0000	-0.0206	-0.0315	-0.0657
$\boldsymbol{\alpha_5}$	0.0000	-0.0021	-0.0016	0.5238
AAE	0.7822	0.7241	0.7109	0.4885
RMSE	1.1814	1.1063	1.0862	0.7888
r	0.2124	0.6516	0.6617	0.7648

Table 2. Parameter estimates and the model performances II

v	**±3e-6**			
U	**±3**	**±3e-6**	**±3e-7**	**zero**
$\boldsymbol{\alpha_0}$	0.1230	-0.0654	-0.0815	-0.3733
$\boldsymbol{\alpha_1}$	-0.3133	-0.0528	-0.0592	0.0274
$\boldsymbol{\alpha_2}$	0.0000	0.2592	0.3297	0.1136
$\boldsymbol{\alpha_3}$	0.0110	-0.0033	-0.0046	-0.0700
$\boldsymbol{\alpha_4}$	0.0000	-0.0179	-0.0337	-0.0656
$\boldsymbol{\alpha_5}$	0.0000	0.0001	-0.0018	0.5238
AAE	0.7822	0.7232	0.7080	0.4885
RMSE	1.1814	1.1043	1.0809	0.7888
r	0.2124	0.6536	0.6631	0.7648

Table 3. Parameter estimates and the model performances III

v	±3e-7			
U	±3	±3e-6	±3e-7	zero
α_0	0.1230	-0.0597	-0.0838	-0.3733
α_1	-0.3133	-0.0513	-0.0600	0.0274
α_2	0.0000	0.2441	0.3375	0.1136
α_3	0.0110	-0.0023	-0.0045	-0.0700
α_4	0.0000	-0.0150	-0.0347	-0.0656
α_5	0.0000	-0.0031	-0.0017	0.5238
AAE	0.7822	0.7285	0.7065	0.4885
RMSE	1.1814	1.1130	1.0781	0.7888
r	0.2124	0.6443	0.6638	0.7648

Table 4. Parameter estimates and the model performances IV

v	zero			
U	±3	±3e-6	±3e-7	zero
α_0	0.1230	-0.0017	-0.0676	-0.3733
α_1	-0.3133	-0.0021	-0.0543	0.0274
α_2	0.0000	0.0074	0.2751	0.1136
α_3	0.0110	0.0000	-0.0029	-0.0700
α_4	0.0000	-0.0001	-0.0239	-0.0656
α_5	0.0000	-0.0001	-0.0016	0.5238
AAE	0.7822	0.7842	0.7200	0.4885
RMSE	1.1814	1.2057	1.1001	0.7888
r	0.2124	0.6191	0.6553	0.7648

Above results indicate that solutions obtained are sensitive to the limits of CIs. We obtain better performance results when the lengths of CIs are narrow. Moreover, when we use the mid-point of our interval values for both input and output data, which is certain data case, we obtain the same parameter estimates, and thus, the same model performances with the CMARS. This reveals that CMARS is a special case of RCMARS. In addition, according to the results, solutions are more sensitive to the changes in the CI limits of the input data than the output data.

4. CONCLUSION AND FURTHER STUDIES

In this paper, we first briefly review the theory and methods of RCMARS, a newly developed method for modeling uncertain data. Then the results of the sensitivity analysis on the parameter estimates, and thus, the model performances are presented. As expected, CMARS produces more accurate results than RCMARS. As the CIs on the variables become narrower, the performance results approaches to that of the CMARS.

As a future research, we are going to run the code for the data that include uncertainties, and then, evaluate the results with respect to the efficiency as well. In this respect, we will discuss stability of our RCMARS model. We will also use robust estimators to construct CIs for our data. We will develop the method further by considering other distributional assumptions than normal for the data.

REFERENCES

Ben-Tal, A., and Nemirovski, A., (2001). *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPR-SIAM Series on Optimization, SIAM, Philadelphia.

Ben-Tal, A., and Nemirovski, A., (2002). *Robust optimization - methodology and applications*, Mathematical Programming, 92, 3, 453-480.

Bertsimas, D., Brown, D.B., and Caramanis, C., (2008). *Theory and Applications of Robust Optimization*, Working paper, Sloan School of Management and Operations Research Center, MIT.

Boni, O., (2007). *Robust Solutions of Conic Quadratic Problems*, PhD Thesis, Technion, Israeli Institute of Technology, IE&M Faculty.

El-Ghaoui, L. and Lebret, H., (1997). *Robust solutions to least-square problems to uncertain data matrices*, SIAM J. Matrix Anal. Appl. 18, 1035–1064.

El-Ghaoui, L., (2003). *Robust Optimization and Applications*, IMA Tutorial.

Fabozzi, F.J., Kolm, P.N., Pachamanova, D.A., and Focardi, S.M., (2007). *Robust Portfolio Optimization and Management*, Wiley Finance.

Friedman, J.H., (1991). *Multivariate adaptive regression splines*, The Annals of Statistics, 19(1), 1-141.

MARS from Salford Systems, 2009. http://www.salfordsystems.com/mars/phb (accessed 25 Aug.).

MOSEK, (2008). A *very powerful commercial software for CQP*, http://www.mosek.com (accessed 05 Sep.).

Özmen, A., Weber,G-W., Batmaz, I., (2010). *The new robust CMARS (RCMARS) method*, preprint at Institute of Applied Mathematics, METU, ISI Proceedings of 24th MEC-EurOPT 2010 –Continuous Optimization and Information-Based Technologies In the Financial Sector, Izmir, Turkey, June 23-26, 2010, 362-368; ISBN 978-9955-28-598-4.

Schöttle, K. and Werner, R., (2006). *Consistency of robust portfolio estimators*, OR and Management Sciences.

Weber, G.-W., Batmaz, I., Köksal G., Taylan P., and Yerlikaya F., (2009). *CMARS: A New Contribution to Nonparametric Regression with Multivariate Adaptive Regression Splines Supported by Continuous Optimisation*, preprint at IAM, METU, submitted for publication.

Werner, R., (2008). *Cascading: an adjusted exchange method for robust conic programming*, CEJOR 16, 179–189.