

# Robustification of Conic Generalized Partial Linear Models under Polyhedral Uncertainty

Ayşe Özmen<sup>a</sup>, Gerhard-Wilhelm Weber<sup>a1</sup>, Erik Kropat<sup>b</sup>

<sup>a</sup>*Institute of Applied Mathematics, Middle East Technical University, 06531, Ankara, Turkey*

<sup>b</sup>*Institute for Theoretical Computer Science, Mathematics and Operations Research, University of Bundeswehr München, 85557, Neubiberg, Germany*

## Abstract

In emerging areas like engineering, finance and control design, it is supposed that the input data are known exactly and equal to some nominal values to construct a model. In real life, however, we have noise in both output and input data. In inverse problems of modelling and data mining, solutions can represent a remarkable sensitivity with respect to perturbations in the parameters, and a computed solution can be highly infeasible, suboptimal, or both. Hence, new models have to be developed when optimization results are combined within real-life applications. Generalized Partial Linear Model (GPLM) combines of two different regression models each of which is employed on different parts of the data set and it is adequate to high dimensional, non-normal and nonlinear data sets that have the flexibility to effectively consider all anomalies. In our previous study, Conic GPLM (CGPLM) was introduced using CMARS and Logistic Regression. Moreover, we included the existence of uncertainty regarding future scenarios into CMARS and linear/logit regression part in CGPLM, and robustified it through robust optimization that is a method to address uncertainty in optimization problems. In this study, we apply RCGPLM on the financial market data taken from a real-world data set as a sample and represent results according to variance.

**Keywords:** RCMARS, CGPLM, RCGPLM, Uncertainty, Robust Optimization

## 1. Introduction

With an increased volatility and thus uncertainty factors, financial crises in recent years introduce a high “uncertainty” into the data taken from the financial sectors and overall to any data related to the financial markets [1]. So, it may be expected that the known statistical models do not give trustworthy results. Robust Optimization reaches great attention from both theoretical and practical point of view as a modelling framework for immunizing against parametric uncertainties. It is a modelling methodology for processing optimization problems in which the data are uncertain and only known to belong to some uncertainty set [7].

---

<sup>1</sup> Corresponding author. E-mail: gweber@metu.edu.tr

Honorary affiliations: Faculty of Economics, Business and Law University of Siegen, Germany School of Science, Information Technology and Engineering, University of Ballarat, Australia

Data are supposed to contain fixed input variables for Multivariate Adaptive Regression Spline (MARS) and its modified version Conic MARS (CMARS). But, in reality, data include noise in output and input variables. Therefore, in our earlier study, to be able to deal with not only fixed but also random type of input data, we included the existence of uncertainty in the future scenarios into CMARS [26, 27] and refined it by *robust optimization* developed by Ben-Tal and Nemirovski [3, 4, 5], El-Ghaoui et al [12, 13]. We call it as RCMARS and, through a robustification of it, diminish the estimation *variance*. Therefore, we arrive at the RCMARS methods [20, 21, 22].

In CMARS and its robustification RCMARS, we must solve an extra problem (by Software MARS [17]), etc.), which is the knot selection not required for the linear part. Therefore, in our earlier study, Conic Generalized Partial Linear Model (CGPLM) was presented as a semiparametric model by using the contribution of a continuous regression model CMARS and a discrete regression model Logistic Regression in [10, 28]. Also, *Robust Generalized Partial Linear Model (RCGPLM)* was received by RCMARS in [23, 24]. CGPLM and RCGPLM base on partial linear model which splits linear and nonlinear variables and model them individually.

In this study, we apply RCGPLM on the financial data by using the contribution of a continuous regression models RCMARS and Linear Regression. Moreover, we describe the concept of a *weak robustification*, because of the computational effort that RGPLM requires. We propose to reduce estimation variance through a robustification in CGPLM.

The works of some scholar's demonstrated that financial decision making for a rational agent is fundamentally a question of achieving an optimal trade-off between risk and return. In this way, robustification is beginning to attract more attention in finance. Therefore, in the financial sector, this study may contribute to some existing approaches of pre-processing financial decision making as done by, e.g., Resampling and Black-Litterman approaches in portfolio optimization [16, 29]. When in these projects, special goals are a sound "diversification" and portfolios which are perceived "natural", in our paper, we focus on the risk aspect and, additionally, provide many control variables for a fine tuning by the modeller and, eventually, the decision maker.

Our paper is structured as follows. In Section 2, RCGPLM is presented in theory and method. In Section 3, robust counterparts of RCGPLM with polyhedral uncertainty are given. Section 4 belongs to the application part, prepared to use a financial data set and to apply RCGPLM. A conclusion and outlook to further investigations are offered in Section 5.

## **2. Robust Conic Generalized Partial Linear Model (CGPLM)**

Since a data set may contain linear and nonlinear variables, different models obtain models for the linear and nonlinear parts separately applying *Generalized Partial Linear Models (GPLMs)*. A particular semiparametric model of interest is the GPLM that extends the GLM in which the usual parametric terms are enlarged by a nonparametric component [19].

Let us represent the existence of uncertainty in the future scenarios within CGPLM [11, 15] in the following form [23]:

$$E(Y|\tilde{\mathbf{X}},\tilde{\mathbf{T}}) = G(\underbrace{\tilde{\mathbf{X}}^T}_{\text{noisy variable}} \boldsymbol{\beta} + \underbrace{\zeta(\tilde{\mathbf{T}})}_{\text{noisy variable}}), \quad (1)$$

where  $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_n)^T$  is a finite dimensional parameter and  $\zeta(\cdot)$  is a smooth function which is tried to predict by  $B$ -splines. Here,  $\tilde{\mathbf{X}} = (\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n)^T$  and  $\tilde{\mathbf{T}} = (\tilde{T}_1, \tilde{T}_2, \dots, \tilde{T}_p)^T$  stand for a decomposition of variables,  $\tilde{\mathbf{X}}$  denoting an  $n$ -variate vector of variables with a linear pattern,  $\tilde{\mathbf{T}}$  denoting a  $q$ -variate vector of variables with a nonlinear pattern to be estimated through a nonlinear model. In this study, we shall focus on special types of estimation  $\zeta(\cdot)$  by RCMARS.

To obtain the GPLM, we encounter observation values  $\tilde{y}_i, \tilde{\mathbf{x}}_i, \tilde{\mathbf{t}}_i$  ( $i=1, 2, \dots, N$ ), giving  $\tilde{\mu}_i = G(\tilde{\gamma}_i)$  and  $\tilde{\gamma}_i = H(\tilde{\mu}_i) = \tilde{\mathbf{x}}_i^T \boldsymbol{\beta} + \zeta(\tilde{\mathbf{t}}_i)$  with a smooth function  $\zeta(\cdot)$  [28]. Here,  $\tilde{X}_j$  ( $j=1, 2, \dots, n$ ) and  $\tilde{T}_j$  ( $j=1, 2, \dots, q$ ) are assumed to be normally distributed random variables and the following configuration is considered for each one of the input variable  $\tilde{X}_j$  and  $\tilde{T}_j$  [23]:

$$\begin{aligned} \tilde{X}_j &= \bar{X} + \zeta_j \quad (j=1, 2, \dots, n). \\ \tilde{T}_j &= \bar{T} + \xi_j \quad (j=1, 2, \dots, q). \end{aligned} \quad (2)$$

To perform a robustification of CGPLM, we employ robust optimization on the linear and the nonlinear parts of CGPLM, and, in the equation (2), we suppose that the input and output variables of CGPLM are given by random variables. They lead us to *uncertainty sets*, which are refined to contain *confidence intervals (CIs)* (for more details, we refer to [21, 22]). In each dimension, we incorporate a *perturbation* or *uncertainty* into the real input data  $\tilde{\mathbf{x}}_i, \tilde{\mathbf{t}}_i$ , and into the output data  $\tilde{y}_i$ . Then, our model under uncertainty can be stated as an additive semiparametric model [23, 24]:

$$H(\tilde{\mu}) = \gamma(\tilde{\mathbf{X}}, \tilde{\mathbf{T}}) = \tilde{\mathbf{X}}^T \boldsymbol{\beta} + \zeta(\tilde{\mathbf{T}}) = \sum_{j=1}^n \tilde{X}_j \beta_j + \zeta(\tilde{\mathbf{T}}). \quad (3)$$

To construct a RCGPLM, we consider observation values  $\tilde{\tilde{y}}_i, \tilde{\tilde{\mathbf{x}}}_i, \tilde{\tilde{\mathbf{t}}}_i$  ( $i=1, 2, \dots, N$ ) after implying the uncertainty. By this,  $\tilde{\tilde{\mu}}_i = G(\tilde{\tilde{\gamma}}_i)$  and  $\tilde{\tilde{\gamma}}_i = H(\tilde{\tilde{\mu}}_i) = \tilde{\tilde{\mathbf{x}}}_i^T \boldsymbol{\beta} + \zeta(\tilde{\tilde{\mathbf{t}}}_i)$  with a smooth function  $\zeta(\cdot)$ .

In the linear part of our estimation, we introduce a new variable  $\tilde{\tilde{Y}}^{preproc}$  with the help of  $\beta_0, \beta_j$  and  $\tilde{\tilde{\mathbf{x}}}_j$  ( $j=1, 2, \dots, n$ ):

$$\tilde{\tilde{Y}}^{preproc} = \tilde{\tilde{\mathbf{X}}}^T \boldsymbol{\beta}^{preproc} + \varepsilon = \beta_0 + \sum_{j=1}^n \tilde{\tilde{\mathbf{x}}}_j \beta_j + \varepsilon. \quad (4)$$

In order to define the knots of MARS with the remaining  $q$  nonlinear variables for residual part, after getting the regression coefficients with the optimal vector  $\boldsymbol{\beta}^{preproc}$  in (4), the linear least-squares model  $\boldsymbol{X}\boldsymbol{\beta}^{preproc}$  is subtracted from  $\check{\check{y}}$ . Here,  $\boldsymbol{X}$  is the design matrix based on the input data. So, for the nonlinear part, the response data vector  $\check{\check{\boldsymbol{y}}}$  is given by [23, 24]

$$\check{\check{\boldsymbol{y}}} = \check{\check{\boldsymbol{y}}} - \check{\check{\boldsymbol{X}}}\boldsymbol{\beta}^{preproc}. \quad (5)$$

In our model (3), the smooth function  $\zeta(\cdot)$  of RCGPLM can be estimated by RCMARS and  $\gamma$  can be represented as a linear combination of *basis functions (BFs)*  $\psi_m$ . Consequently, model (5) obtains the form

$$\check{\check{\boldsymbol{y}}} = H(\check{\check{\boldsymbol{\mu}}}) = \alpha_0 + \sum_{m=1}^M \alpha_m \psi_m(\check{\check{\boldsymbol{t}}}^m), \quad (6)$$

where  $\alpha_m$  is the unknown coefficient of the  $m$ th BF ( $m=1,2,\dots,m$ ) and  $\alpha_0$  is the intercept term. Here,  $\psi_m$  ( $m=1,2,\dots,m$ ) is a basis function, being a product of two or more one-dimensional structured linear functions and taken from a set of  $M$  linearly independent basis elements. Then, by MARS, a set of eligible knots values is chosen and assigned separately for input variables. Multiplying an existing basis function with a truncated linear function including a new variable, interaction BFs are obtained. Piecewise linear BFs in MARS method are expanded based on the new data set that has uncertainties. So, the piecewise linear BFs have the following notation [14]:

$$c^+(\check{\check{x}}, \tau) = (\check{\check{x}} - \tau)_+, \quad c^-(\check{\check{x}}, \tau) = (\check{\check{x}} - \tau)_-, \quad (7)$$

where  $(q)_+ := \max\{0, q\}$ ,  $(q)_- := \max\{0, -q\}$ , and  $\tau$  is a univariate knot. Therefore, the existing BF and the newly created interaction BF as well are employed in the approximation by MARS. Given the observations presented by the data  $(\check{\check{t}}_i, \check{\check{y}}_i)$  ( $i = 1, 2, \dots, N$ ), the  $m$ th BF is following form [14]:

$$\psi_m(\check{\check{\boldsymbol{t}}}) := \prod_{j=1}^{K_m} [s_{\kappa_j^m} \cdot (\check{\check{t}}_{\kappa_j^m} - \tau_{\kappa_j^m})]_+. \quad (8)$$

Here,  $K_m$  means the number of truncated linear functions multiplied in the  $m$ th BF,  $t_{\kappa_j^m}$  means the input variable corresponding to the  $k$ th truncated linear function in the  $m$ th BF,  $\tau_{\kappa_j^m}$  means the knot value corresponding to the variable  $t_{\kappa_j^m}$  and  $s_{\kappa_j^m}$  means the selected sign +1 or -1. To predict our function  $\gamma(\check{\check{\boldsymbol{t}}})$  in model (6) instead of the backward step-wise algorithm of MARS [17], penalty terms are applied in addition to the least-squares estimation to control the lack-of-fit from the complexity viewpoint of the estimation. For the GPLM with RCMARS, *Penalized Residual*

*Sum of Squares (PRSS)*, with BFs according to (8) accumulated in the forward stepwise algorithm of MARS, are constructed by the equation [22]:

$$PRSS := \sum_{i=1}^N \left( \tilde{\gamma}_i - \boldsymbol{\alpha}^T \boldsymbol{\psi}(\tilde{\mathbf{t}}_i) \right)^2 + \sum_{m=1}^{M_{\max}} \phi_m \sum_{|\boldsymbol{\theta}|=1}^2 \sum_{r < s} \int_{\boldsymbol{\theta}^T = (\theta_1, \theta_2)} \alpha_m^2 \left[ D_{r,s}^{\boldsymbol{\theta}} \psi_m(\mathbf{t}^m) \right]^2 d\mathbf{t}^m. \quad (9)$$

Here,  $V(m) := \{ \kappa_j^m \mid j=1,2,\dots,K_m \}$  is the variable set used in the  $m$ th BF  $\psi_m$ ,  $\tilde{\mathbf{t}}_i = (\tilde{t}_{i,1}, \tilde{t}_{i,2}, \dots, \tilde{t}_{i,q})^T$  show the inputs and  $\mathbf{t}^m = (t_{m_1}, t_{m_2}, \dots, t_{m_{\kappa_m}})^T$  represents the vector of variables that contribute to the  $m$ th BF  $\psi_m$ , which are related with the  $i$ th link function  $\zeta_i$  ( $i=1,2,\dots,N$ ). The terms  $\phi_m \geq 0$  are in the role of *penalty parameters* ( $m=1,2,\dots,M_{\max}$ ). Additionally,  $D_{r,s}^{\boldsymbol{\theta}} \psi_m(\mathbf{t}^m) = \left( \frac{\partial^{\boldsymbol{\theta}} \psi_m}{\partial^{\theta_r} t_r^m \partial^{\theta_s} t_s^m} \right) (\mathbf{t}^m)$  shows the first- or second-order derivatives. The integrals of the first-order derivatives measure the *flatness* of the model functions whereas the integrals of the second-order derivatives measure the *instability* and *complexity* in the model. After the discretized form is applied to approximate the multi-dimensional integrals, our PRSS of CGPLM in equation (9) may be presented as (see [21, 27] for more details):

$$PRSS \approx \left\| \tilde{\boldsymbol{\gamma}} - \boldsymbol{\psi}(\tilde{\mathbf{t}}) \boldsymbol{\alpha} \right\|_2^2 + \phi \left\| \mathbf{L} \boldsymbol{\alpha} \right\|_2^2. \quad (10)$$

Herewith, our PRSS problem looks like a classical *Tikhonov Regularization (TR) problem* [2] with some  $\phi > 0, \phi = \lambda^2$  for some  $\lambda \in \mathbb{R}$ . Let us express the TR problem (10) through Conic Quadratic Programming (CQP) which is a convex optimization methodology. Therefore, PRSS can be easily formulated as a CQP problem and, indeed, referring to an approximate selection of a bound  $\tilde{M}$ , we state our problem as follows [20]:

$$\begin{aligned} & \underset{t, \boldsymbol{\alpha}}{\text{minimize}} \quad t, \\ & \text{subject to} \quad \left\| \tilde{\boldsymbol{\gamma}} - \boldsymbol{\psi}(\tilde{\mathbf{t}}) \boldsymbol{\alpha} \right\|_2 \leq t, \\ & \quad \quad \quad \left\| \mathbf{L} \boldsymbol{\alpha} \right\|_2 \leq \sqrt{\tilde{M}}. \end{aligned} \quad (11)$$

### 3. Robust Counterparts of RCGPLM with Polyhedral Uncertainty

In our study, we employ the form of *polyhedral sets* as our uncertainty sets as we continue to elaborate our optimization problem as a CQP problem. Whenever *polyhedral* uncertainty is employed for the *linear part* of CGPLM, addressing uncertainty sets  $U_1^1$  and  $U_2^1$  the robust counterpart is represented by [23]

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \max_{\substack{\mathbf{w}_1 \in U_1^1 \\ \mathbf{z}_1 \in U_2^1}} \left\| \mathbf{z}_1 - \mathbf{W}_1 \boldsymbol{\beta} \right\|_2^2 + \phi \left\| \mathbf{K} \boldsymbol{\beta} \right\|_2^2, \quad (12)$$

where  $U_1^1$  is a polytope with  $2^{N \cdot n}$  vertices  $\mathbf{W}_1^1, \mathbf{W}_1^2, \dots, \mathbf{W}_1^{2^n}$ . Actually, it is not exactly known, but belongs to a convex bounded uncertain domain

$$U_1^1 = \left\{ \sum_{j=1}^{2^{N \cdot n}} \delta_j \mathbf{W}_1^j \mid \delta_j \geq 0 (j \in \{1, 2, \dots, 2^{N \cdot n}\}), \sum_{j=1}^{2^{N \cdot n}} \delta_j = 1 \right\}, \quad (13)$$

i.e.,  $U_1^1 = \text{conv}\{\mathbf{W}_1^1, \mathbf{W}_1^2, \dots, \mathbf{W}_1^{2^n}\}$  being the convex hull. Furthermore,  $U_2^1$  is a polytope with  $2^N$  vertices  $\mathbf{z}_1^1, \mathbf{z}_1^2, \dots, \mathbf{z}_1^{2^N}$ . It is not exactly known, but considered as the convex bounded uncertain domain

$$U_2^1 = \left\{ \sum_{i=1}^{2^N} \varphi_i \mathbf{z}_1^i \mid \varphi_i \geq 0 (i \in \{1, 2, \dots, 2^N\}), \sum_{i=1}^{2^N} \varphi_i = 1 \right\}, \quad (14)$$

i.e.,  $U_2^1 = \text{conv}\{\mathbf{z}_1^1, \mathbf{z}_1^2, \dots, \mathbf{z}_1^{2^N}\}$ . Here, the matrix  $\mathbf{W}_1$  and the vector  $\mathbf{z}_1$  with uncertainty are lying in Cartesian products of intervals which are parallelepipeds [21]. Since  $U_1^1$  and  $U_2^1$  are polytopes they can be described via their vertices:

$$\begin{aligned} U_1^1 &= \text{conv}\{\mathbf{W}_1^1, \mathbf{W}_1^2, \dots, \mathbf{W}_1^n\}, \\ U_2^1 &= \text{conv}\{\mathbf{z}_1^1, \mathbf{z}_1^2, \dots, \mathbf{z}_1^N\}. \end{aligned} \quad (15)$$

Now, since we employ CQP for the linear part of our RCGPLM, the optimization problem (12) can be stated as a *standard* CQP [6] in the following form [22]:

$$\begin{aligned} &\underset{\varpi_1, \boldsymbol{\beta}}{\text{minimize}} \quad \varpi_1, \\ &\text{subject to} \quad \|\mathbf{z}_1^i - \mathbf{W}_1^j \boldsymbol{\beta}\|_2 \leq \varpi_1 \quad (i = 1, 2, \dots, N ; j = 1, 2, \dots, n), \\ &\quad \quad \quad \|\mathbf{K} \boldsymbol{\beta}\|_2 \leq \sqrt{\widetilde{M}_1}, \end{aligned} \quad (16)$$

with some selected parameter values of  $\widetilde{M}_1$ . We can solve this robust CQP problem (16) by the help of MOSEK [18].

If *polyhedral* uncertainty is implied for the *nonlinear part* of CGPLM, addressing on uncertainty sets  $U_1^2$  and  $U_2^2$ , the robust counterpart is described by [23]

$$\underset{\boldsymbol{\alpha}}{\text{minimize}} \quad \max_{\substack{\mathbf{W}_2 \in U_1^2 \\ \mathbf{z}_2 \in U_2}} \|\mathbf{z}_2 - \mathbf{W}_2 \boldsymbol{\alpha}\|_2^2 + \phi \|\mathbf{L} \boldsymbol{\alpha}\|_2^2, \quad (17)$$

where  $U_1^2$  is a polytope with  $2^{N \cdot M_{\max}}$  vertices  $\mathbf{W}_2^1, \mathbf{W}_2^2, \dots, \mathbf{W}_2^{2^{N \cdot M_{\max}}}$ . Similar to  $U_1^1$ ,  $U_1^2$  is set by

$$U_1^2 = \left\{ \sum_{j=1}^{2^{N \cdot M_{\max}}} \delta_j \mathbf{W}_2^j \mid \delta_j \geq 0 (j \in \{1, 2, \dots, 2^{N \cdot M_{\max}}\}), \sum_{j=1}^{2^{N \cdot M_{\max}}} \delta_j = 1 \right\}, \quad (18)$$

i.e.,  $U_1^2 = \text{conv}\{\mathbf{W}_2^1, \mathbf{W}_2^2, \dots, \mathbf{W}_2^{2^{N \cdot M_{\max}}}\}$ . Furthermore,  $U_2^2$  is a polytope with  $2^N$  vertices  $\mathbf{z}_2^1, \mathbf{z}_2^2, \dots, \mathbf{z}_2^{2^N}$ :

$$U_2^2 = \left\{ \sum_{i=1}^{2^N} \varphi_i \mathbf{z}_2^i \mid \varphi_i \geq 0 (i \in \{1, 2, \dots, 2^N\}), \sum_{i=1}^{2^N} \varphi_i = 1 \right\}, \quad (19)$$

i.e.,  $U_2^2 = \text{conv}\{\mathbf{z}_2^1, \mathbf{z}_2^2, \dots, \mathbf{z}_2^{2^N}\}$  [21]. Whenever  $U_1^2$  and  $U_2^2$  are polytopes given by

$$\begin{aligned} U_1^2 &= \text{conv}\{\mathbf{W}_2^1, \mathbf{W}_2^2, \dots, \mathbf{W}_2^{M_{\max}}\}, \\ U_2^2 &= \text{conv}\{\mathbf{z}_2^1, \mathbf{z}_2^2, \dots, \mathbf{z}_2^N\}, \end{aligned} \quad (20)$$

and we employ CQP again, then our robust problem (17) for nonlinear part can be equivalently stated as a *standard* CQP [6]:

$$\begin{aligned} &\underset{\varpi_2, \boldsymbol{\alpha}}{\text{minimize}} \quad \varpi_2 \\ &\text{subject to} \quad \|\mathbf{z}_2^i - \mathbf{W}_2^j \boldsymbol{\alpha}\|_2 \leq \varpi_2 \quad (i = 1, 2, \dots, N ; j = 1, 2, \dots, M_{\max}), \\ &\quad \|\mathbf{L}\boldsymbol{\alpha}\|_2 \leq \sqrt{\tilde{M}_2}. \end{aligned} \quad (21)$$

We can solve this robust CQP problem (21) by the help of MOSEK [18] with some parameter values of  $\tilde{M}_1$  chosen.

#### 4. Application of RCGPLM in Financial Market

For the implementation of RCGPLM algorithm developed, we use a data set as a sample from the real-world financial market data which are chosen as the time series data for the empirical part from the website of Central Bank of the Republic of Turkey [9]. The data that cover the time horizon between January 1999 and December 2000 include the four economic indicators which are the most usually used ones for the interpretation of an economic situation.

For the financial markets, an index is an imaginary portfolio of securities presenting a particular market or a portion of it. In our data set, *ISE 100 stock index* is the dependent variable because of a statistical measure of change in an economy or a securities market. It has its own calculation methodology and is usually employed in terms of a change from a base value. Therefore, the percentage change is more important than the actual numerical value. The independent variables are *ISE Trading Volume*, *Capacity Usage Ratio*, and *Credit Volume*. Additionally, one indicator

from the USA is taken for analyze: *Fed Funds Interest Rate* because of the strong effect of the USA on the economy of Turkey and the world. So, we have 4 predictor variables ( [25] for more detail ):

$x_1$  : ISE Trading Volume,  
 $x_2$  : Capacity Usage Ratio,  
 $x_3$  : Credit Volume,  
 $x_4$  : Federal Funds Interest Rate,  
with 24 observations.

For RCGPLM application, the input variables divide two parts. ISE Trading Volume and Federal Funds Interest Rate are selected for linear part and Capacity Usage Ratio and Credit Volume are chosen for nonlinear part. Firstly, we validate our assumption that the input variables and the output variable are distributed normally and we transform the variables into the standard normal distribution, the CI is obtained to be the interval (-3, 3). In fact, we have a *tradeoff* between tractability and robustification. Since we do not have enough computer capacity to solve our problem for uncertainty matrices, we formulate linear and nonlinear part of our RCGPLM as a CQP problem for each observation using the *combinatorial approach*, which we call as *weak robustification*. Consequently, we generate different *weak RCGPLM (WRCGPLM)* models for both the linear and the nonlinear part.

To apply the robust optimization technique on the linear part of CGPLM model, the uncertainty matrices and vectors based on *polyhedral uncertainty sets* are obtained by using (13) and (14) and, uncertainty is calculated for all input and output values which are represented by CIs. Then, perturbation (uncertainty) is included into the real input data  $\tilde{x}_i$ , in each dimension, and into the output data  $y_i$  ( $i = 1, 2, \dots, 24$ ).

To construct our WRCGPLM model for linear part, we obtain the 24 different models. We solve them separately by using MOSEK program [18] and find the  $\varpi_1$  values for all auxiliary problems. Then, using the *worst-case* approach, we choose the solution that has the *maximum*  $\varpi_1$  value from the equation (16) and we continue our calculations using the parameter value  $\alpha_0, \alpha_1, \alpha_2$  which we find from the auxiliary problem that has the highest  $\varpi_2$  value.

After getting the regression coefficients with the optimal vector  $\beta^{preproc}$ , we subtract the linear least-square model  $\tilde{\tilde{X}}\beta^{preproc}$  from the  $y$  and we evaluate the output vector  $\mathcal{Y}$  for the nonlinear part. Then, for nonlinear part we obtain the largest model with  $\mathcal{Y}$  by using the Salford MARS Version 3 [17]. This largest model has the following BFs:



$$\begin{aligned}
\psi_1(x) &= \max\{0, t_2 - 0.2855\}, & \psi_2(x) &= \max\{0, 0.2855 - t_2\}, \\
\psi_3(x) &= \max\{0, t_2 + 0.2430\}, & \psi_4(x) &= \max\{0, -0.2430 - t_2\}, \\
\psi_5(x) &= \max\{0, t_1 + 1.8979\} \cdot \max\{0, t_2 + 0.2430\} \\
\psi_6(x) &= \max\{0, t_2 + 0.7508\}, & \psi_7(x) &= \max\{0, -0.7508 - t_2\}, \\
\psi_8(x) &= \max\{0, t_1 + 1.8979\} \cdot \max\{0, t_2 + 0.7508\}.
\end{aligned}$$

Thus, the large model is presented as follows:

$$\begin{aligned}
y &= \alpha_0 + \sum_{m=1}^M \alpha_m \psi_m(\mathbf{t}) + \varepsilon \\
&= \alpha_0 + \alpha_1 \max\{0, t_2 - 0.2855\} + \alpha_2 \max\{0, 0.2855 - t_2\} + \alpha_3 \max\{0, t_2 + 0.2430\} \\
&\quad + \alpha_4 \max\{0, -0.2430 - t_2\} + \alpha_5 \max\{0, t_1 + 1.8979\} \cdot \max\{0, t_2 + 0.2430\} \\
&\quad + \alpha_6 \max\{0, t_2 + 0.7508\} + \alpha_7 \max\{0, -0.7508 - t_2\} \\
&\quad + \alpha_8 \max\{0, t_1 + 1.8979\} \cdot \max\{0, t_2 + 0.7508\} + \varepsilon.
\end{aligned}$$

To employ the robust optimization technique on the nonlinear part of CGPLM model, similar to linear part, the uncertainty matrices and vectors based on *polyhedral uncertainty sets* are constructed by using (18) and (19) and, uncertainty is evaluated for all input and output values which are represented by CIs. Then perturbation (uncertainty) is incorporated into the real input data  $\tilde{t}_i$ , in each dimension, and into the output data  $\gamma_i$  ( $i=1,2,\dots,24$ ).

Afterward, similar to linear part, we obtain 24 different *WRCGPLM* models for nonlinear part, and solve them by using MOSEK program [18]. After getting the  $\varpi_2$  values for all auxiliary problems, using the *worst-case* approach, we choose the solution which has the *maximum*  $\varpi_2$  value from the equation (21). Then we continue our calculations using the parameter value  $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7$  and  $\alpha_8$  which we find from the auxiliary problem that has the highest  $\varpi_2$  value by using the worst-case approach. Therefore, for our optimal problem, we evaluate the regression coefficients and variances of linear and nonlinear parts of RCGPLM in Table 1.

Table1. Parameter values and variances for linear and nonlinear parts of RCGPLM

	$\alpha_0$	$\alpha_1$	$\alpha_2$	$\alpha_3$	$\alpha_4$	$\alpha_5$	$\alpha_6$	$\alpha_7$	$\alpha_8$	<b>variance</b>
<b>linear part</b>	0,026	0,595	0,233							0,577
<b>nonlinear part</b>	0,586	-0,708	-0,929	0,153	-0,249	0,088	0,063	-0,189	0,000	0,501

## 5. Conclusion and Further Studies

In this investigation, we solve our data analysis and robustify CGPLM by both a robust approach and a combinatorial approach, called the *weakly robust case*, to tackle uncertainties that exist in real-world data and to make our refined approach *feasible*. Herewith, we purpose to diminish the estimation variance. For this aim, we have developed a theory and method, and also programmed it to be able to implement the method. We used both the efficient Interior Point Methods of MOSEK [18], together with MATLAB and some developed parts of the statistical software of MARS, called Salford MARS [17]. This has proved to be an excellent symbiosis of codes, and we want to develop and refine that interface in the future.

The “goodness” of RCGPLM expresses itself through a tradeoff between exactness and stability; the first goal which is the classical one in model identification, is compromised (via some parameters) with robustness concerning the response variables. As our new contribution, robustification enters into the input variables and, herewith, in the model design. Via further “control” parameters, we provide a tuning of how much risk averse the modeller wishes to be.

Aiming at a precise choice of any of these parameters, we will apply comparison criteria from statistics, as we did before [21, 22, 25]. An additional performance measure might also come from geometry and clustering theory: by our separation between “linearly” and “nonlinearly” involved groups of variables  $X$  and  $T$  from equation (2), we benefit from the “shape” and structure of the data. This is a main advantage of partially linear models, *PLMs*; we go one generalizing step further and permit two main working fields of supervised learning to be addressed at once now: regression and classification, modelled by *GPLMs*. Together with our robustification and the excellent numerical and complexity properties of Interior Point Methods which we use [8, 30], a tremendous improvement of our earlier works which appear as special cases now, and that it will mean a great achievement in future studies.

Then, we will work on our new introduced *process version* of GPLM [24] for modelling, optimization and robustification of dynamical networks. We shall study a real-world application of this approach to validate and to investigate the performance of our GPLM, CGPLM, RGPLM and RCGPLM for target environment networks and we shall develop our method under other distributional assumptions for the data than being normal.

## References

- [1] E. Andreou, E. Ghysels, A. Kourtellos, *Should Macroeconomic Forecasters Use Daily Financial Data and How?*, Available at SSRN: <http://ssrn.com/abstract=1711899>, 2010.
- [2] R.C. Aster, B. Borchers and C. Thurber, *Parameter Estimation and Inverse Problems*, Academic Press, Boston, 2004.
- [3] A. Ben-Tal and A. Nemirovski, *Robust convex optimization*, Math. Oper. Res. 23 (1998), pp. 769–805.
- [4] A. Ben-Tal, and A. Nemirovski, *Robust solutions to uncertain Linear Programs*, Operations Research Letters, 25, 1:1-13, 1999.
- [5] A. Ben-Tal, and A. Nemirovski, *Robust solutions of Linear Programming problems contaminated with uncertain data*, Math. Progr. 88:411–424, 2000.

- [6] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPR-SIAM Series on Optimization, SIAM, Philadelphia, 2001.
- [7] A. Ben-Tal and A. Nemirovski, *Robust Optimization– methodology and applications*, Math. Program, 92(3) (2002), pp. 453-480.
- [8] D. Bertsimas, D.B. Brown, and C. Caramanis, *Theory and applications of robust Optimization*, Tech. Rep., University of Texas, Austin, TX, USA, 2007.
- [9] Central Bank of the Republic of Turkey: <http://www.tcmb.gov.tr>.
- [10] Z. Çavuşoğlu, Predicting Debt Crises in Emerging Markets Using Generalized Partial Linear Models, Term Project, Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey, 2010.
- [11] G. Çelik, Parameter Estimation in Generalized Partial Linear Models with Conic Quadratic Programming, M.S. Thesis, METU, 2010.
- [12] L. El-Ghaoui and H. Lebret, *Robust solutions to least-square problems to uncertain data matrices*, SIAM J. Matrix Anal. Appl. 18:1035–1064, 1997.
- [13] L. El-Ghaoui, F. Oustry, and H. Lebret, *Robust solutions to uncertain semidefinite programs*, SIAM J. Optim. 9:33–52, 1998.
- [14] J.H. Friedman, *Multivariate adaptive regression splines*, The Ann. Statist., 19 (1) (1991), pp. 1-141.
- [15] B. Kayhan, Parameter Estimation in Generalized Partial Linear Models with Tikhonov Regularization Method, 2010, M.S. Thesis, METU, Ankara.
- [16] B. Litterman, F. Litterman and R. Litterman, *Global Portfolio Optimization*, Financial Analysts Journal, September/October 1992, pp. 28-43.
- [17] MARS Salford Systems; software available at <http://www.salfordsystems.com>.
- [19] MOSEK; A very powerful commercial software for CQP, available at <http://www.mosek.com>.
- [18] M. Müller, *Estimation and Testing in Generalized Partial Linear Models – A Comparative Study*, Statistics and Computing 11, 299-309, 2001.
- [20] A. Özmen, G.-W. Weber, and I. Batmaz, *The new robust CMARS (RCMARS) method*, in *ISI Proceedings of 24th MEC-EuroOPT 2010–Continuous Optimization and Information-Based Technologies in the Financial Sector*, Izmir, Turkey, 2010, pp. 362-368.
- [21] A., Özmen, *Robust Conic Quadratic Programming Applied to Quality Improvement- A Robustification of CMARS*, Master Thesis, METU, Ankara, Turkey, 2010.
- [22] A. Özmen, G.-W. Weber, I. Batmaz, and E. Kropat, *RCMARS: Robustification of CMARS with Different Scenarios under Polyhedral Uncertainty Set*, in *Communications in Nonlinear Science and Numerical Simulation (CNSNS)*, Special Issue Nonlinear, Fractional and Complex Systems with Discontinuity and Chaos, D. Baleanu and J.A. Tenreiro Machado (guest editors), 2011, DOI:10.1016/j.cnsns.2011.04.001.
- [23] A. Özmen, G.-W. Weber, *Robust Conic Generalized Partial Linear Models Using RCMARS Method – A Robustification of CGPLM*. In *Proceedings of Fifth Global Conference on Power Control and Optimization PCO June 1 – 3, 2011, Dubai*, ISBN: 983-44483-49.
- [24] A. Özmen, G.-W. Weber, Z. Çavuşoğlu and Ö. Defterli, *A new robust conic GPLM method with an Application to Finance and regulatory systems: prediction of credit default and a process version*, preprint at IAM, METU, Ankara, 2011.

- [25] A. Özmen, G.-W. Weber, A. Karimov, *A New Robust Optimization Tool Applied on Financial Data*, submitted to Pacific Journal of Optimization.
- [26] P. Taylan, G.-W. Weber and A. Beck, *New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and technology*, in the special issue in honour of Prof. Dr. Alexander Rubinov, B. Burachik and X. Yang (guest eds.) of Optimization 56, 5-6, 675–698, 2007.
- [27] G.-W. Weber, I. Batmaz, G. Köksal, P. Taylan, and F. Yerlikaya, *CMARS: A new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization*, Inverse Problems in Science and Engineering, 2011, DOI:10.1080/17415977.2011.624770.
- [28] G.-W. Weber, Z. Çavuşoğlu and A. Özmen, *Predicting Default Probabilities in Emerging Markets by New Conic Generalized Partial Linear Models and Their Optimization*, to appear Advances in Continuous Optimization with Applications in Finance, special issue of Optimization.
- [29] R. Werner, *Consistency of robust portfolio estimates*, Optimization in Finance, Coimbra, 2007.
- [30] R. Werner, *Cascading: an adjusted exchange method for robust conic programming*, CEJOR Cent. Eur. J. Oper. Res., 16, 2008, pp.179-189.