REFINEMENTS, EXTENSIONS AND MODERN APPLICATIONS OF CONIC MULTIVARIATE ADAPTIVE REGRESSION SPLINES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF APPLIED MATHEMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

FATMA YERLİKAYA-ÖZKURT

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
SCIENTIFIC COMPUTING

DECEMBER 2013

Approval of the thesis:

REFINEMENTS, EXTENSIONS AND MODERN APPLICATIONS OF CONIC MULTIVARIATE ADAPTIVE REGRESSION SPLINES

submitted by **FATMA YERLİKAYA-ÖZKURT** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Department of Scientific Computing, Middle East Technical University** by,

Prof. Dr. Bülent Karasözen
Director, Graduate School of **Applied Mathematics** _____

Prof. Dr. Bülent Karasözen
Head of Department, **Scientific Computing** _____

Prof. Dr. Gerhard-Wilhelm Weber
Supervisor, **Scientific Computing, METU** _____

Examining Committee Members:

Prof. Dr. Bülent Karasözen
Scientific Computing, METU _____

Prof. Dr. Gerhard-Wilhelm Weber
Scientific Computing, METU _____

Prof. Dr. Mustafa Pınar
Department of Industrial Engineering, Bilkent University _____

Assoc. Prof. Dr. Sevtap Kestel
Actuarial Sciences, METU _____

Assoc. Prof. Dr. Ömür Uğur
Scientific Computing, METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: FATMA YERLİKAYA-ÖZKURT

Signature :

ABSTRACT

REFINEMENTS, EXTENSIONS AND MODERN APPLICATIONS OF CONIC MULTIVARIATE ADAPTIVE REGRESSION SPLINES

Yerlikaya-Özkurt, Fatma

Ph.D., Department of Scientific Computing

Supervisor : Prof. Dr. Gerhard-Wilhelm Weber

December 2013, 90 pages

Conic Multivariate Adaptive Regression Splines (CMARS) which has been developed at the Institute of Applied Mathematics, METU, as an alternative approach to the well-known data mining tool Multivariate Adaptive Regression Splines (MARS). CMARS is based on given data and a penalized residual sum of squares for MARS, interpreted as a Tikhonov Regularization problem. CMARS treats this problem by a continuous optimization technique called Conic Quadratic Programming (CQP).

This doctoral thesis adapts the CMARS model into a wide frame of advanced methods of statistics and applied mathematics. The first application is using CMARS in Generalized Partial Linear Models (GPLMs), a particular form of a semiparametric model, which extends the Generalized Linear Models (GLMs) in that the usual parametric terms are augmented by a single nonparametric component. We prefer GLMs because of their flexibility to the variety of statistical problems and the availability of software to fit the models. There are different kinds of estimation methods for GPLMs. One of the great advantages of semiparametric models consists of some grouping (linear and nonlinear or parametric and nonparametric) which could be done for the input dimensions (or features) in order to assign appropriate submodels to the groups specifically. In this thesis, for the estimation of the parametric model part, we apply the least-squares estimation. On the other hand, we consider CMARS for the nonparametric part to estimate the smooth function. This new algorithm, called CGPLM, has the advantage of higher speed and less complexity, as it accesses the use of interior point

methods.

The other extension is the use of CMARS method for the outlier identification problem. For this purposes, we provide a new solution by using regularization and CQP techniques to the mean-shift outlier model, which is considered as a parametric method. After that the proposed method is improved by using CMARS to represent the nonlinear structure in the data.

The second track of this doctorate study is the use of CMARS method for the parameter identification of Stochastic Differential Equations (SDEs) driven by Brownian motions and fractional Brownian motions (fBms). Both systems of SDEs with standard multi-dimensional Brownian motions and systems of SDEs having correlated Brownian motions are covered in this thesis. Moreover, we introduce the CMARS method to estimate both the spline coefficients and, especially, the Hurst parameter of the SDEs driven by fBms. The theoretical results of this study may lead new implementations and applications in science, technology and finance.

This PhD thesis ends with a conclusion and an outlook to future studies.

Keywords : Conic multivariate adaptive regression splines, stochastic differential equations, fractional Brownian motion, conic generalized partial linear model, outlier identification

ÖZ

KONİK ÇOK DEĞİŞKENLİ UYARLANABİLİR REGRESYON EĞRİLERİNİN GELİŞTİRİLMESİ, UZANTILARI VE MODERN UYGULAMALARI

Yerlikaya-Özkurt, Fatma

Doktora, Bilimsel Hesaplama Bölümü

Tez Yöneticisi : Prof. Dr. Gerhard-Wilhelm Weber

Aralık 2013, 90 sayfa

Konik Çok Değişkenli Uyarlanabilir Regresyon Eğrileri (CMARS) iyi bilinen veri madenciliği tekniklerinden biri olan Çok Değişkenli Uyarlanabilir Regresyon Eğrilerine (MARS) alternatif olarak ODTÜ Uygulamalı Matematik Enstitüsünde geliştirilmiştir. CMARS belirlenmiş veriye ve bir Tikhonov düzenleme problemi olarak yorumlanmış MARS için cezalandırılmış hata kareler toplamına dayanmaktadır. CMARS bu problemi konik karesel programlama (CQP) olarak adlandırılan sürekli optimizasyon tekniği ile ele almaktadır.

Bu doktora tezi CMARS modelini istatistik ve uygulamalı matematiğin ileri metodlarının geniş bir çerçevesi içerisinde uyarlamaktadır. İlk uygulama yarı parametrik bir modelin özel bir formu olan Genelleştirilmiş Kısmi Doğrusal Modellerde (GPLMs) CMARS'ın kullanılmasıdır. GPLMs genel parametrik terimlerin parametrik olmayan bir parça ile birleştirildiği Genelleştirilmiş Doğrusal Modellerin (GLMs) farklı bir şeklidir. Burada GLMs'in tercih edilmesinin nedeni çeşitli istatistiksel problemlere olan esnekliği ve model uyumu için gerekli olan hazır yazılımların varlığıdır. GPLMs için çeşitli tahmin yöntemleri bulunmaktadır. Yarı parametrik modellerin en büyük avantaj-larından biri bazı grupların (doğrusal ve doğrusal olmayan veya parametrik ve paramet-rik olmayan) girdi boyutlarına veya değişkenlerine göre uygun olan alt modeller belirleyebilmeyi içermesidir. Bu tezde, parametrik model kısmının tahmini için en küçük kareler tahmin yöntemini kullandık. Diğer taraftan, CMARS'ı parametrik olmayan kısımdaki pürüzsüz fonksiyon tahmini için düşündük. CGPLM olarak isim-

lendirilen bu yeni algoritma iç nokta metodunu kullandığı için yüksek hız ve düşük karmaşıklık avantajına sahiptir.

Diğer bir uygulama CMARS yönteminin aykırı gözlem belirlemesi probleminin çözümü için kullanılmasıdır. Bu amaç ile biz düzenleme ve CQP tekniklerini kullanarak parametrik bir yöntem olan ortalama-kaydırma aykırı gözlem modeli için yeni bir çözüm ürettik. Sonrasında önerilen yöntem veri içindeki doğrusal olmayan yapının gösterilmesi için CMARS'ın kullanılması ile daha da geliştirildi.

Bu doktora tezinin ikinci kolu Brown hareketleri ve/ya kademeli Brown hareketleri (fBms) ile elde edilmiş Stokastik Diferansiyel Denklemlerin (SDEs) parametrelerinin belirlenmesi için CMARS yönteminin kullanılmasıdır. Bu tez hem standart çok boyutlu Brown hareketleri ile SDEs sistemlerini hem de ilintili Brown haraketleri ile SDEs sistemlerini kapsamaktadır. Ayrıca, burada CMARS yönteminin eğri katsayısı ve özellikle fBms ile elde edilmiş SDEs'in Hurst parametresi tahmini için kullandık. Bu çalışmanın teorik sonuçlarının bilim, teknoloji ve finansta yeni yorumlama ve uygulamalara öncülük edebileceği düşünülmektedir.

Bu doktora tezi sonuç ve gelecekte yapılacak çalışmalara bir bakış ile sona ermektedir.

Anahtar Kelimeler: Konik çok değişkenli uyarlanabilir regresyon eğrileri, stokastik diferansiyel denklemler, kademeli Brown hareketi, konik genelleştirilmiş kısmi doğrusal model, aykırı gözlem belirleme

In memory of my parents, Dilber and Nadir Yerlikaya

ACKNOWLEDGMENTS

I would like to gratefully and sincerely thank my advisor, Prof. Dr. Gerhard Wilhelm Weber, for his guidance, understanding, and most importantly, his friendship during my graduate studies at Middle East Technical University (METU). He provides a well rounded experience consistent my long-term career goals. For everything you have done for me, Prof. Dr. Weber, I thank you.

I would like to thank the members of my doctoral committee for their input, valuable discussions and accessibility.

I would like to sincerely thank Assoc. Prof. Dr. Pakize Taylan for helpful discussions on my PhD studies.

I am very thankful to Assist. Prof. Dr. Yeliz Yolcu-Okur and Assist. Prof. Dr. Ceren Vardar-Acar for their friendship and discussions on fractional Brownian motion.

I would like to express my deepest appreciation and thanks to Prof. Dr. İnci Batmaz for her assistance and guidance in getting my graduate career and through our studies with her.

Advice and comments given by Assoc. Prof. Dr. Aysegül Askan has been great help in ending my PhD. I also thank her for introducing me to earthquake engineering as well for the support on the way.

I would like to thank all members of the Department of Industrial Engineering at METU for their friendship and the pleasant atmosphere, especially, Prof. Dr. Gülser Köksal, Prof. Dr. Nur Evin Özdemirel and Assist. Prof. Dr. Cem İyigün with whom I worked closely.

I deeply thank all members of the Institute of Applied Mathematics (IAM). Additionally, my special thank goes to the director of IAM, Prof. Dr. Bülent Karasözen, for his suggestions and supports to my academic life.

I also gratefully acknowledge the financial support of Turkish Scientific and Technical Research Council (TUBİTAK).

Finally, and most importantly, I would like to thank my husband, Metin Özkurt, for his support, encouragement, quiet patience and unwavering love. Furthermore, none of my studies would have been possible without the support of my big family. Thanks each of them for their continuous support, care, encouragement, endless love and patience. They always believed and encouraged me. Thanks for being there for me all the time.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Adj-R^2	Adjusted R^2
ANN	Artificial Neural Network
CART	Classification and Regression Trees
CGAM	Conic Generalized Additive Model
CGPLM	Conic Generalized Partial Linear Model
CMARS	Conic Multivariate Adaptive Regression Splines
CMARS-MSOM	Mean-Shift Outlier Model with Conic Multivariate Adaptive Regression Splines
CMSOM	Conic Mean-Shift Outlier Model
CQP	Conic Quadratic Programming
DDE	Delay Differential Equation
fBm	Fractional Brownian Motion
GAM	Generalized Additive Model
GAM & CQP	Generalized Additive Model with Conic Quadratic Programming
GCV	Generalized Cross-Validation
GLM	Generalized Linear Model
GPLM	Generalized Partial Linear Model
IDE	Impulsive Differential Equation
IPM	Interior Point Method
LM	Linear Model
LTS	Least Trimmed Square
MAE	Mean Absolute Error
MARS	Multivariate Adaptive Regression Splines
ML	Maximum-Likelihood
MLE	Maximum-Likelihood Estimation
MLR	Multiple Linear Regression
MSE	Mean Squared Error
MSOM	Mean-Shift Outlier Model
ODE	Ordinal Differential Equation
PDE	Partial Differential Equation

PRSS	Penalized Residual Sum of Squares
PWI	Proportion of Residuals within Three-Sigma
R^2	Multiple Coefficient of Determination
RSS	Residual Sums of Squares
SDE	Stochastic Differential Equation
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TR	Tikhonov Regularization
UK	United Kingdom
US	United States

CHAPTER 1

INTRODUCTION

Data mining is an important tool in many areas such as science, engineering, healthcare, medicine, business, economics and finance. The data sets, generated in big laboratories and institutions, and processed by powerful computing systems, and the increasingly complex applications in those areas, are large, complex, and generally "noisy." Extracting information from the massive amounts of data requires the use of sophisticated and high performance techniques and algorithms, based on statistics and optimization. These techniques have to be usable by scientists, engineers and researchers.

Data mining methods enable us to analyze the data stored in various data repositories gathered for different purposes such as quality improvement in industry [30] or developing early warning systems [5] in various fields such as finance, environment, energy, etc. There is a number of predictive data mining tools useful for establishing mathematical or statistical relationships between several factors of interest. Some widely used ones are *Classification and Regression Trees (CART)*, *Artificial Neural Network (ANN)*, *Multiple Linear Regression (MLR)*, and *Support Vector Machine (SVM)*. Among these tools, *Multivariate Adaptive Regression Splines (MARS)* is a well-known predictive data mining method capable of modeling high-dimensional data with nonlinear structure [14]. The flexible nature of MARS modeling leads to a successful implementation of the method in various application areas [14]. Reported success of the method attracted the attention of many researchers to tackle with the problems of MARS method or, alternatively, to extend it further and to improve its capability. In one of these studies, *Conic Multivariate Adaptive Regression Splines (CMARS)* has been developed as an alternative to the backward stepwise part of the MARS algorithm [70].

1.1 The Conic Multivariate Adaptive Regression Splines Method

CMARS does not make any specific assumption about the underlying functional relationship between the dependent and independent variables to estimate a general model function [59, 65, 70]. CMARS is introduced by linear combinations of the basis functions that are used in MARS. The selection of basis functions is data-based and specific to the problem at hand. CMARS uses one-dimensional basis functions of the

form $c^+(x,\tau) = [+(x-\tau)]_+$ and $c^-(x,\tau) = [-(x-\tau)]_+$, where $[q]_+ := \max\{0, q\}$ [65, 70]. Each function is piecewise linear, with a knot at the value τ, and the corresponding couple of functions is called a *reflected pair*. The visualization of these basis functions is demonstrated in Figure 1.1.

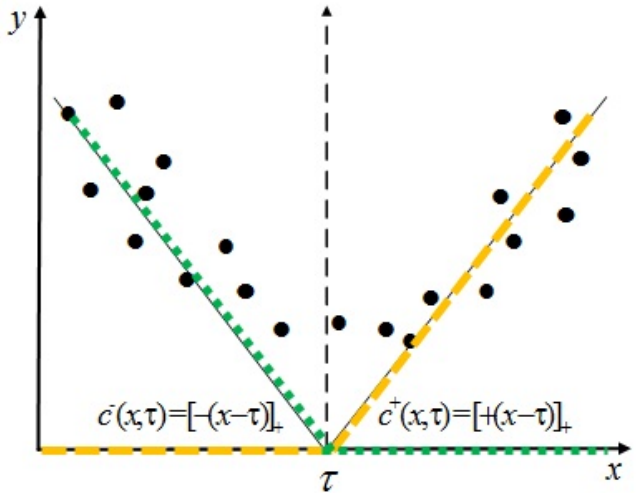

Figure 1.1: An illustration of the basis functions [70].

A set of basis functions is given as follows:

$$\wp := \{(x_j - \tau)_+, (\tau - x_j)_+ \mid \tau \in \{x_{1,j}, x_{2,j}, ..., x_{N,j}\},\ j \in \{1, 2, ..., p\}\}. \quad (1.1)$$

A CMARS model function f is represented by a linear combination of basis functions which is successively built up by the set $\wp$, as described below:

$$Y = f(\boldsymbol{x}) + \epsilon = \theta_0 + \sum_{m=1}^{M} \theta_m \psi_m(\boldsymbol{x}^m) + \epsilon. \quad (1.2)$$

Here, Y is a response variable, $\boldsymbol{x}^m = (x_1, x_2, ..., x_p)^T$ a vector of predictors for the corresponding mth multivariate basis function. Furthermore, θ_m are the unknown coefficients for the mth basis function $(m = 1, 2, ..., M)$ or for the constant 1 $(m = 0)$, and ϵ is an additive stochastic component, often called *noise*, which is assumed to have zero mean and finite variance. In Eqn. (1.2), ψ_m $(m = 1, 2, ..., M)$ are basis functions as products of two or more one-dimensional basis functions. The form of the mth basis function can be written as follows:

$$\psi_m(\boldsymbol{x}^m) := \prod_{j=1}^{K_m} [s_{\kappa_j^m} \cdot (x_{\kappa_j^m} - \tau_{\kappa_j^m})]_+; \quad (1.3)$$

these are nonsmooth *spline* functions. Here, $\boldsymbol{x}^m$ is the vector of variable contributed to the mth basis function, K_m is the number of *truncated linear* functions multiplied in

the mth basis function, $x_{\kappa_j^m}$ is the predictor variable corresponding to the jth truncated linear function in the mth basis function, $\tau_{\kappa_j^m}$ is the knot value corresponding to the variable $x_{\kappa_j^m}$, and $s_{\kappa_j^m}$ is the selected sign $+1$ or -1 [14, 59].

Determination of all elements in the set of basis functions given in Eqn. (1.1) is an important issue. For this purpose, the starting function is taken as $\psi_0(\boldsymbol{x}^0) = 1$. Then, the basis function that causes the most amount of reduction in the *Residual Sums of Squares (RSS)* is included in the model. The new basis functions can be restricted to interactions of a maximum order. An advantage of the form given in Eqn. (1.3) lies in its ability to estimate the contributions of the basis functions so that both the additive and the interactive effects of the predictors are allowed to determine the dependent variable. Interaction basis functions are created recursively by multiplying an existing basis function with a truncated linear function involving a new variable. In a CMARS approximation, both the existing basis function and the newly created interaction basis function are used. In order to demonstrate the interaction basis function form, the basis function in Figure 1.2 is defined as the multiplication of two basis functions $\psi_1 = \max\{0, M_w - 5.7\}$ and $\psi_2 = \max\{0, R_{jb} - 37.34\}$ from the application of CMARS for ground motion prediction [71].

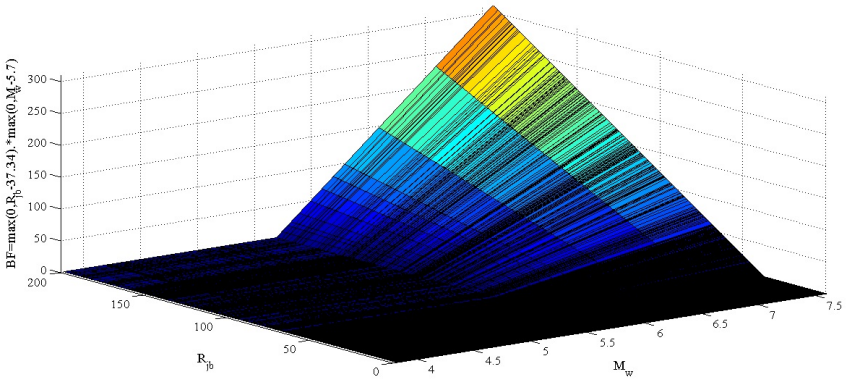

Figure 1.2: An illustration of the interaction basis function [71].

The MARS algorithm consists of two main parts [14]. In the first part (forward algorithm), the basis functions are selected to minimize the *lack-of-fit* until a user-specified maximum number of basis functions, M_{max}, is attained. This process results in a large model which overfits the data. In the second part (backward algorithm), to overcome this disadvantage, the basis functions contributing least to the residual squared errors are removed, thus causing less complex model. In both parts, MARS algorithm uses *Generalized Cross-Validation (GCV)* as a variable selection criterion. GCV is defined as follows:

$$GCV := \frac{1}{N} \frac{\sum_{i=1}^{N} (y_i - \hat{f}_\alpha(\boldsymbol{x}_i))^2}{(1 - \tilde{C}(\alpha)/N)^2}, \tag{1.4}$$

where $\tilde{C}(\alpha) := u + dK$ [14]. Here, N is the number of sample observations, u is the number of linearly independent basis functions, K is the number of knots selected

in the forward process, and d is a cost for basis function optimization as well as a smoothing parameter for the procedure [14, 20].

CMARS is developed as an alternative to the backward part of the MARS algorithm. CMARS uses up to all basis functions generated by the forward algorithm of MARS [65, 70]. CMARS is constructed by a *Penalized Residual Sum of Squares (PRSS)* parameter estimation problem, instead of an ordinary least-squares estimation problem as it occurs in MARS method. In the sense of a trade-off, the PRSS problem aims at a highest possible accuracy and a smallest possible complexity of the model. PRSS with penalty parameters λ_m and with M_{max} basis functions which are accumulated in the first part of the MARS algorithm, has the following form:

$$PRSS := \sum_{i=1}^{N}(y_i - f(\tilde{\boldsymbol{x}}_i))^2 + \sum_{m=1}^{M_{max}} \lambda_m \sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_m}} \int_{Q^m} \theta_m^2 \left[D_{r,s}^{\boldsymbol{\alpha}}\psi_m(\boldsymbol{t}^m)\right]^2 d\boldsymbol{t}^m, \tag{1.5}$$

where $V_m := \left\{\kappa_j^m | j = 1, 2, ..., K_m\right\}$ is the variable set associated with the mth basis function, $\boldsymbol{t}^m = \left(t_{m_1}, t_{m_2}, ..., t_{m_{K_m}}\right)^T$ represents the vector of variables which contribute to the mth basis function. Our integral for the mth basis function is defined over Q^m. The subdivision of Q^m is determined with the special coordinates of $\boldsymbol{t}^m$. Moreover, $D_{r,s}^{\boldsymbol{\alpha}}\psi_m(\boldsymbol{t}^m) := \frac{\partial^{|\boldsymbol{\alpha}|}\psi_m}{\partial^{\alpha_1} t_r^m \partial^{\alpha_2} t_s^m}(\boldsymbol{t}^m)$ for $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T$, $|\boldsymbol{\alpha}| := \alpha_1 + \alpha_2$, where $\alpha_1, \alpha_2 \in \{0, 1\}$ [65, 70]. As can be seen from Eqn. (1.5), PRSS is composed of two parts representing accuracy and complexity, respectively [20]. They are tried to be compromised by using the penalty parameters λ_m [65].

In Eqn. (1.5), the integral symbol, "$\int$", is used as dummy in the sense of $\int_{Q^m}$, where Q^m is some appropriately large K_m-dimensional parallel-pipe where the integration takes place. Since the multi-dimensional integrals in Eqn. (1.5) are difficult to evaluate, we discretize these integrals and rearrange PRSS in the following form [70]:

$$\begin{aligned} PRSS \;\approx\; & \sum_{i=1}^{N}\left(y_i - \boldsymbol{\psi}(\tilde{\boldsymbol{d}}_i)\boldsymbol{\theta}\right)^2 \\ & + \sum_{m=1}^{M_{max}} \lambda_m \theta_m^2 \sum_{i=1}^{(N+1)^{K_m}} \left(\sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_m}} \left[D_{r,s}^{\boldsymbol{\alpha}}\psi_m(\hat{\boldsymbol{x}}_i^m)\right]^2 \right) \Delta\hat{\boldsymbol{x}}_i^m, \end{aligned} \tag{1.6}$$

where $\boldsymbol{\psi}(\tilde{\boldsymbol{d}}_i) := \left(1, \psi_1(\tilde{\boldsymbol{x}}_i^1), ..., \psi_M(\tilde{\boldsymbol{x}}_i^M), \psi_{M+1}(\tilde{\boldsymbol{x}}_i^{M+1}), ..., \psi_{M_{max}}(\tilde{\boldsymbol{x}}_i^{M_{max}})\right)^T$, and $\boldsymbol{\theta} := (\theta_0, \theta_1, ..., \theta_{M_{max}})^T$ with the point $\tilde{\boldsymbol{d}}_i := \left(\tilde{\boldsymbol{x}}_i^1, \tilde{\boldsymbol{x}}_i^2, ..., \tilde{\boldsymbol{x}}_i^M, \tilde{\boldsymbol{x}}_i^{M+1}, ..., \tilde{\boldsymbol{x}}_i^{M_{max}}\right)^T$ in the argument, and with $(\sigma^{\kappa_j})_{j\in\{1,2,...,p\}} \in \{0, 1, 2, ..., N+1\}^{K_m}$ and

$$\hat{\boldsymbol{x}}_i^m = \left(\tilde{x}_{l_{\sigma^{\kappa_1^m}}^{\kappa_1^m}, \kappa_1^m}, ..., \tilde{x}_{l_{\sigma^{\kappa_{K_m}^m}}^{\kappa_{K_m}^m}, \kappa_{K_m}^m} \right), \quad \Delta\hat{\boldsymbol{x}}_i^m := \prod_{j=1}^{K_m} \left(\tilde{x}_{l_{\sigma^{\kappa_j}+1, \kappa_j^m}^{\kappa_j^m}} - \tilde{x}_{l_{\sigma^{\kappa_j}, \kappa_j^m}^{\kappa_j^m}} \right). \tag{1.7}$$

To put the PRSS in Eqn. (1.6) a simpler form in order to be able to handle the problem easily, we use a uniform penalization, $\lambda = \lambda_m$, for each derivative term. Then, PRSS turns into a *Tikhonov Regularization (TR) problem* as described below [3]:

$$PRSS \approx \left\| \boldsymbol{y} - \boldsymbol{\psi}(\tilde{\boldsymbol{d}})\boldsymbol{\theta} \right\|_2^2 + \lambda \left\| \boldsymbol{L}\boldsymbol{\theta} \right\|_2^2, \tag{1.8}$$

where $\boldsymbol{L}$ is constructed by the discretizations of the high-dimensional integrals given in Eqn. (1.5). The model approximations as presented in Eqn. (1.8) are carefully prepared [3, 70]. They play an important role in order to raise a final model approximation which is linear in the unknown spline parameters. After unifying some discretized complexity terms and including them into inequality constraints, a *Conic Quadratic Programming (CQP)* problem is obtained which uses the powerful *Interior Point Methods (IPMs)* [41, 49]. The formulation of a CQP is given as follows:

$$\begin{aligned} &\underset{\boldsymbol{\theta}}{\text{minimize}} \quad t, \\ &\text{subject to} \quad \left\| \boldsymbol{y} - \boldsymbol{\psi}(\tilde{\boldsymbol{d}})\boldsymbol{\theta} \right\|_2 \leq t, \quad \left\| \boldsymbol{L}\boldsymbol{\theta} \right\|_2 \leq \sqrt{\tilde{M}}, \end{aligned} \tag{1.9}$$

Here, the optimization problem given in Eqn. (1.9) is solved by IPMs via the optimization software MOSEK [37, 41]. There can be many solutions to this problem for some chosen complexity bound values, $\tilde{M}$, which are determined by trial and error [20]. As a representative solution, however, we have plotted on a log-log scale. The curve of the optimal values of $\left\| \boldsymbol{L}\boldsymbol{\theta} \right\|_2$ and $\left\| \boldsymbol{y} - \boldsymbol{\psi}(\tilde{\boldsymbol{d}})\boldsymbol{\theta} \right\|_2$ often take on an *efficiency* (or *L*) *curve* [65, 70]. An illustration of this L-curve is given in Figure 1.3. In this figure, the optimal value configuration point at the L-curve is the corner point, which is marked by a bold point. The selected value gives the best solution for both accuracy and complexity in terms of PRSS in Eqn. (1.8). The L-curve, which is an efficiency frontier, can be regarded as an implicit function of a parameter that is in the role of an upper bound; in case of several such parameters, we obtain an efficiency surface.

In addition to the discrepancy principle, another popular criterion for picking the value of λ is the L-curve criterion in which the value of λ is selected that gives the solution at or closest to the corner of the L-curve [70]. In fact, the use of these IPMs is one of the main advantages that is prepared by our model-based mathematical approach. For this reason, problem formulation in Eqn. (1.8) is our preferred realization of the Tikhonov regularization.

CMARS method is based on the CQP approach. This approach, supported by mathematical programming, is elegant and powerful. Just as linear programming, semidefinite programming, and robust optimization, CQP uses the theory of IPMs (also known to be a class of barrier methods) which guarantees feasibility throughout the entire iteration procedures. On the other hand, penalty methods and Tikhonov Regularization can be considered as exterior point methods with possible infeasibility [41].

Moreover, CMARS is one of the nonparametric approaches that make no specific assumptions to estimate model parameters based on given data sets. Real-world processes from the financial sector or from nature are often characterized by their huge

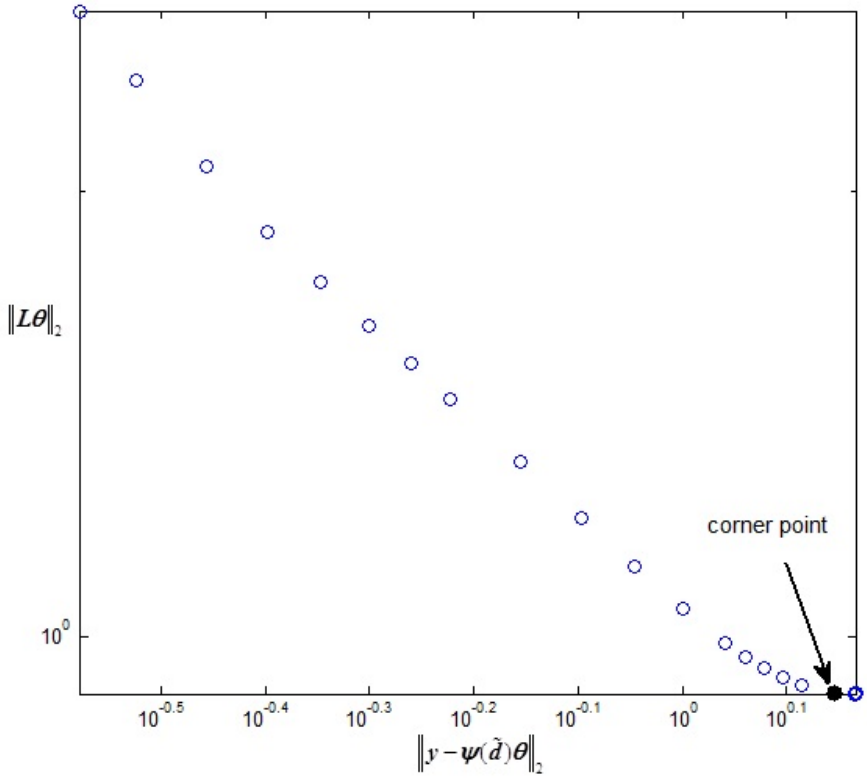

Figure 1.3: Demonstration of L-curve.

quantity and variation [13]. CMARS is able to model such data by using the advantages of recent developments in optimization theory, computer hardware and software. The flexible, *adaptive* nature of CMARS modeling has led to a successful implementation on various processes with nonlinear structure, e.g., in industrial engineering, image processing and earthquake engineering [32, 65, 71].

1.2 Applications of Conic Multivariate Adaptive Regression Splines and Comparison with Other Data Mining Methods

In order to assess the power of the CMARS method, several real-life data obtained from well-known data repositories [65] as well as simulation data applications are implemented. These data sets used in CMARS applications are gathered from various study areas such as life, finance, industry, social and business, etc. Also, CMARS is rigorously evaluated and compared with some other predictive data mining methods such as MARS, CART and *Generalized Additive Models with CQP (GAMs & CQP)* for classification and MLR and MARS for prediction [72].

In one of the CMARS related studies [53], CMARS and CART are applied to the area of finance to model sovereign default. In this application, since the dependent variable is type of categorical, particularly binary, classification trees are utilized to construct the CART model. In that study, the classification capabilities of the models developed

are determined by the scoring results of training and test samples. Results indicate that, in the training sample, CART has a better classification capability than CMARS with respect to the classification accuracy. In the test sample, however, CMARS gives more accurate results than CART with respect to the same measure. As a result of CART algorithm, modeling an additive structure with CART is not easy, and the estimation variance of a CART model increases with continuous explanatory variables. On the contrary, CMARS captures such a structure easily and, thus, it seems more suitable for future predictions than CART.

In the comparison study, the data sets used for CART are also applied to GAMs & CQP [53]; that approach can also be called *Conic Generalized Additive Models (CGAMs)*. Results indicate that the classification capabilities of CGAMs and CMARS in the test sample are accurate. CMARS performs better than CGAMs and it discovers the main structure of the data set better.

In the study where CMARS and MLR are compared, the models are developed for two simulation data sets. When the structure of data is linear, CMARS performs as good as MLR does; when there is nonlinearity, however, CMARS outperforms MLR. There are two main disadvantages of MLR over CMARS. First, human expertise is usually needed, and as a result, it may take a long time to develop MLR models. Secondly, the *white-noise* assumptions should be satisfied to obtain statistically valid inferences but these assumptions may not be realistic in most real-life applications [39, 73].

The prediction performances of CMARS and MARS are evaluated both on real-life and simulated data sets. In one of the simulation applications, both MARS and CMARS models are developed involving lower- as well as higher-order interactions [6]. According to the results of the real-life applications, for all training data sets, CMARS performs better and it is more robust than MARS with respect to all prediction performance measures considered. In addition, although CMARS is more stable than MARS, MARS is more robust than CMARS in stability with respect to most of the prediction performance measures. While stability is measured by the proportion of the performance measure values on training and test data sets, robustness is assessed by standard deviation of the performance measures under different data features. Furthermore, an effect of sample size on the performance of methods is also detected. For example, CMARS performs better than MARS mostly on medium to large training samples. If we consider the performance with respect to scale, almost for all measures, MARS and CMARS perform better (or the same) as the scale changes from small to large. To compare the efficiencies of methods, computational run times (in seconds) for each training sample is recorded on the same computer. Run times seem to be related to the sample size but not to the problem scale. MARS may provide solutions very fast compared to CMARS since its applications are run on professional software, Salford MARS. The run times of CMARS method increase almost up to 3 to 5 times as much to that of MARS as the sample size increases. Moreover, in the same study, a simulation study conducted shows that CMARS method has a better performance than MARS on the noisy data [65, 70].

Two studies have been conducted for comparison of MARS and CMARS for classification. In one of them, both methods are applied to Pima Indians Diabete data to

determine if a person has diabetes or not [59, 62]. According to the results, both methods perform almost the same. However, in term of the true diagnose of the disease, CMARS performs better than MARS. In both training and test samples, MARS and CMARS scoring models are almost the same; on the other side, CMARS discovers the main structure of data better. Depending on these findings, the study in [59] concludes that CMARS gives a superior estimate compared to MARS.

The comparison of MARS and CMARS models for the data to predict the countries' credit default possibilities show variations in training and tests samples [53]. CMARS outperforms MARS according to the default country assignation. Also, its performance measures reveal higher stabilities than MARS.

1.3 Extensions and Advances on Conic Multivariate Adaptive Regression Splines

The mentioned success of the applications using CMARS attracted the attention of us to refine and extend the modern applications of CMARS by advanced methods of applied mathematics and statistics. The main purpose of this thesis is to benefit from the advances in the mathematics of data mining with optimization methods, and to address the importance of the application of special data mining method on different data sets, and to identify and explore the connections between data mining and other applied sciences.

As an extension and modern application of CMARS method, Generalized Partial Linear Models (GPLMs) have been combined with CMARS. Here, a GPLM utilizes semiparametric methods and augments the usual parametric terms by a single nonparametric component of a continuous covariate. In recent years, the class of Generalized Linear Models (GLMs) has gained popularity as a statistical modeling tool. This popularity is due in part to the flexibility of GLMs in addressing a variety of statistical problems and to the availability of software to fit the models. In this thesis, GPLMs - an extensions of GLMs - are motivated, introduced, and formally presented with CMARS model and CQP-based optimization techniques. We call this new technique as Conic Generalized Partial Linear Models (CGPLMs).

Another extension and a modern use of CMARS method have been made for the outlier identification problem. In statistical applications, regression models based on data may give misleading results when data contain outliers. We deal with the outliers problem in linear regression in two stages. Firstly, we use the Mean-Shift Outlier Model (MSOM) as a parametric method and providing a new solution by using regularization and convex optimization techniques, CQP. Then, the power of CMARS method is used and the proposed method is improved.

As an advance, we employ CMARS method for the parameter identification of Stochastic Differential Equations (SDEs) in a simplified manner. SDEs are widely used to represent noisy and dynamical real-world problems and they play an important role in many field of applied science, especially, in finance [43]. Since CMARS provides a functional form of the model by using basis functions which is a special form of splines, in our case, the parameters of SDEs are constructed by these "discretely" non-

smooth functions. Then, coefficients of basis functions are found by using CQP. Both systems of SDEs with standard multi-dimensional Brownian motions and systems of SDEs having correlated Brownian motions are considered and identified in this thesis. Moreover, various parameters of SDEs driven by fractional Brownian motions (fBms) are identified by the mentioned approach. Here, we introduce the CMARS method to estimate both the spline coefficients and, especially, the Hurst parameter of the SDEs driven by fBms, too.

This thesis is organized as follows: In Chapter 2, we begin with CGPLMs, including CMARS for the nonlinear model part. We introduce CMARS and CQP which is not just method but a wide field of optimization, for the outlier identification problem in Chapter 3. In Chapter 4, we give a new methodology based on CMARS for the parameter estimation of SDEs. We analyze and focus on a numerical approximation of multi-dimensional SDEs and correlated systems of SDEs in the same chapter. In Chapter 5, we introduce the method CMARS, especially, for the estimation of the Hurst parameter of SDEs driven by fBms. Finally, a conclusion of the thesis and an outlook to future research are presented in Chapter 6.

CHAPTER 2

CONIC GENERALIZED PARTIAL LINEAR MODELS WITH CONIC MULTIVARIATE ADAPTIVE REGRESSION SPLINES

2.1 Introduction

Analysis of large and complex data sets is very important to describe social, physical and biological systems, but is not easy. In order to comprise the best possible model for the data on hand, most applied researchers use regression models [42]. The following subsections provide information about extensions and improvements on well-known regression models.

2.1.1 Generalized Linear Models

Almost in every field of science such as engineering, applied science, social science or economics, scientists and engineers use regression models for analyzing and describing data sets to understand the underlying patterns and to disclose the relationship between the predictors and response variable [42]. The most popular one is the class of *Generalized Linear Models (GLMs)* which is an extension of traditional linear regression models. This class will be our preparatory set of models based on which we shall introduce *Generalized Partial Linear Models (GPLMs)* soon. This popularity of GLMs is due in part to the flexibility of GLMs in addressing a variety of statistical problems and to the availability of software to fit the models [40]. The class of GLMs allows the mean of a dependent variable to depend on a linear predictor through a nonlinear link function, and it permits the probability distribution of the response to be any member of an exponential family of distributions. Many widely used statistical models belong to GLMs, since both linear and nonlinear regression models are unified under the framework of GLMs. GLMs include traditional linear models with normal errors, logistic and probit models for binary data, log-linear models for multinomial data. Many other useful statistical models such as the Poisson, binomial, Gamma, and normal distribution can be formulated as GLMs by the selection of an appropriate link function and response probability distribution [40, 58].

A GLM has the following basic structure:

$$\eta = H(\mu) = \boldsymbol{x}^T \boldsymbol{\beta},$$

where H is a smooth monotonic *link function*, $\boldsymbol{x}$ is a vector of the observed values of explanatory variables and $\boldsymbol{\beta}$ is a vector of unknown parameters. In addition, a GLM usually makes the distributional assumptions that the response variables, Y_i, are independent and can have any distribution from exponential family density

$$Y \sim f_Y(y, \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\},$$

where a, b and c are arbitrary "scale" parameters, and θ is called a natural parameter. We can also obtain a general expression for the mean and variance of dependent variable Y using log likelihood of θ, $\mu = \mathrm{E}(Y) = b'(\theta)$ and $\mathrm{Var}(Y) = b''(\theta)\, a(\phi)$. Here, $(\cdot)'$ and $(\cdot)''$ represent the first-order and second-order derivative of b in terms of θ, respectively. Generally, $a(\phi)$ is defined as $a(\theta) := \phi/w$ and $\mathrm{Var}(Y) = \mathrm{V}(\mu)\,\phi$, where $\mathrm{V}(\mu) := b''(\theta)/w$ [58, 69]. Some commonly used exponential family density functions are listed in Table 2.1.

2.1.2 Generalized Partial Linear Models

A particular semiparametric model of interest is the *Generalized Partial Linear Model (GPLM)* which extends the GLMs in away that the usual parametric terms are augmented by a single nonparametric component. As a result, a GPLM separates the set of independent variables into two subsets and combines a linear model with a nonlinear model by addition. For example, in biological or financial sectors, many processes expressed by stochastic differential equations can be stated by a linear submodel for the deterministic drift term and by a nonlinear submodel for the stochastic diffusion term. Such combinations of submodels provide better accuracy and stability for complex and large data sets [38]. Moreover, GPLM approach demonstrates the capability of handling both classification and prediction problems, e.g., prediction of credit default [66], image processing and speech processing [60]. In this chapter, we consider GPLM as a prediction tool. The general GPLM model is given by the following function

$$f(\boldsymbol{X}, \boldsymbol{T}) = \boldsymbol{X}^T\boldsymbol{\beta} + \gamma(\boldsymbol{T}) + \epsilon \tag{2.1}$$

or, after modeling the response variable Y by Eqn. (2.1) and an additional noise component ϵ, and applying the expectation value, by

$$\mathrm{E}(Y|\boldsymbol{X}, \boldsymbol{T}) = G\left\{\boldsymbol{X}^T\boldsymbol{\beta} + \gamma(\boldsymbol{T})\right\}. \tag{2.2}$$

Here, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a finite-dimensional parameter and $\gamma(\cdot)$ is a smooth function which is introduced to estimate the nonlinear part of GPLM. In this chapter, to represent the smooth function, nonparametric models are preferred because they are more flexible than nonlinear models. Here, we also assume that vectors $\boldsymbol{X}$ and $\boldsymbol{T}$ come from a decomposition of the set of explanatory variables. While $\boldsymbol{X}$ denotes an m-dimensional random vector which typically represents discrete covariates, $\boldsymbol{T}$ is a q-dimensional random vector of continuous covariates which is to be modeled in a nonparametric way [38]. The variable $\boldsymbol{T}$ can be seen as an extraneous or nuisance variable.

Table 2.1: Some commonly used exponential family distributions [69].

Distribution Name	$f\left(y\right)$	Range	θ	ϕ
Normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$	$-\infty < y < \infty$	μ	σ^2
Poisson	$\frac{\mu^y \exp(-\mu)}{y!}$	$y = 0, 1, 2, \ldots$	$\log\left(\mu\right)$	1
Binomial	$\binom{n}{y}\binom{\mu}{n}\left(1-\frac{\mu}{n}\right)^{n-y}$	$y = 0, 1, 2, \ldots, n$	$\log\left(\frac{\mu}{n-\mu}\right)$	1
Gamma	$\frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}\exp\left(-\frac{\nu y}{\mu}\right)$	$y < 0$	$-\frac{1}{\mu}$	$\frac{1}{\nu}$
Inverse Gaussian	$\sqrt{\frac{\gamma}{2\pi y^3}}\exp\left(\frac{-\gamma\left(y-\mu^2\right)}{2\mu^2 y}\right)$	$y < 0$	$-\frac{1}{2\mu^2}$	$\frac{1}{\gamma}$

Distribution Name	$a\left(\phi\right)$	$b\left(\theta\right)$	$c\left(y,\phi\right)$	$\mathrm{V}\left(\mu\right)$
Normal	$\phi\left(=\sigma^2\right)$	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left(\frac{y^2}{\phi}+\log 2\pi\phi\right)$	1
Poisson	$\phi\left(=1\right)$	$\exp\left(\theta\right)$	$-\log(y!)$	μ
Binomial	$\phi\left(=1\right)$	$n\log\left(1+e^{\theta}\right)$	$\log\binom{n}{y}$	$\mu\left(1-\mu/n\right)$
Gamma	$\phi\left(=\frac{1}{\nu}\right)$	$-\log\left(-\theta\right)$	$\nu\log\left(\nu y\right)-\log\left(y\Gamma\left(\nu\right)\right)$	μ^2
Inverse Gaussian	$\phi\left(=\frac{1}{\gamma}\right)$	$-\sqrt{-2\theta}$	$-\frac{1}{2}\left(\log\left(2\pi y^3\phi\right)+\frac{1}{\phi y}\right)$	μ^3

There are different kinds of estimation methods for GPLM. Generally, the estimation methods for the model of Eqn. (2.2) are based on the idea that an estimate $\hat{\boldsymbol{\beta}}$ can be found for a known $\gamma(\cdot)$ and an estimate $\hat{\gamma}(\cdot)$ can be found for a known $\boldsymbol{\beta}$, especially, for $\hat{\boldsymbol{\beta}}$. Now, we will concentrate on special estimation types of $\gamma(\cdot)$ and $\boldsymbol{\beta}$ based on CMARS and with the help of CQP [60, 74].

The rest of this chapter is organized as follows. Section 2.2 includes Conic Generalized Partial Linear Models (CGPLMs) and our estimation approach with CMARS. Finally, Section 2.3 presents applications and results of the methodology.

2.2 Conic Generalized Partial Linear Models

Let us consider the model in Eqn. (2.2), where we assume that $G = H^{-1}$ is a known link function which links the mean of the dependent variable, $\mu = \mathrm{E}(Y|\boldsymbol{X}, \boldsymbol{T})$, to the predictors. In this case, Eqn. (2.2) can be considered as a semiparametric generalized linear model, because all terms are linear; i.e.,

$$H(\mu) = \eta(\boldsymbol{X}, \boldsymbol{T}) = \boldsymbol{X}^T\boldsymbol{\beta} + \gamma(\boldsymbol{T}) = \sum_{j=1}^{p} X_j\beta_j + \gamma(\boldsymbol{T}). \tag{2.3}$$

Now, to obtain the GPLM, we consider data values and data vectors: $\bar{y}_i$, $\bar{\boldsymbol{x}}_i$, $\bar{\boldsymbol{t}}_i$ for $(i = 1, 2, \ldots, N)$. Then, $\mu_i = G(\eta_i)$ and $\eta_i = H(\mu_i) = \bar{\boldsymbol{x}}_i^T\boldsymbol{\beta} + \gamma(\bar{\boldsymbol{t}}_i)$, and $\gamma(\cdot)$ is a smooth function [58].

In many studies, the theory does not put any restrictions on the form of $\gamma(\cdot)$, i.e., it does not say whether $\gamma(\cdot)$ is linear, quadratic, increasing in $\boldsymbol{T}$ or its components, etc. [38]. Hence, it is up to the empirical analysis uses data to find out more about $\gamma(\cdot)$. When we take $\boldsymbol{\beta} = \mathbf{0}$ in Eqn. (2.3), then the estimate of $\gamma(\cdot)$ is based on a nonparametric regression technique that contains a smoother, i.e., a "mollifying" operation, such as a regularization.

A *smoother* can be described as a tool for briefly expressing the form of response variable as a function of one or more predictor variables. The name smoother is coming from the form of the estimation that includes less variables than the response itself. Since a smoother does not suppose a rigid form for the relationship between response and predictors, it has a nonparametric nature. Because of this important property, a smoother is often preferred to be used for nonparametric regression [19].

One of the main uses of the smoother is to estimate the relationship between the response and the predictors and, hence, to provide an important structure for the estimation of additive models [16]. There are many useful smoothers such as polynomial, bin, running mean, running line, loess, Gaussian kernel, smoothing spline, regression spline and natural spline. If we have more than one predictor, then an estimation of a p-dimensional surface can be done easily by a generalization of the running mean, locally weighted running line, and kernel smoothers. On the other hand, multivariate tensor product splines are known as an another generalization [64]. These are functional to generalize univariate regression splines. The process is based on the construction of

multi-dimensional basis functions by multiplying one-dimensional basis function with two or more than two one-dimensional basis functions. MARS basis function is an example for this kind of a smoother [19, 20]. In this chapter, we prefer to present smoothers by CMARS model and its algorithm.

CMARS uses expansions in piecewise linear basis functions; thus, $\gamma(\boldsymbol{t})$ can be represented by a linear combination of these successively built basis functions and the intercept, θ_0 , such that Eqn. (2.3) becomes

$$\eta = H(\mu) = \boldsymbol{x}^T\boldsymbol{\beta} + \theta_0 + \sum_{m=1}^{M_{max}} \theta_m \psi_m(\boldsymbol{x}^m) + \epsilon. \tag{2.4}$$

Here, ψ_m $(m = 1, 2, ..., M_{max})$ are basis functions from $\wp$ or products of two or more such functions, ψ_m is taken from a set of M_{max} linearly independent basis elements, and θ_m are the unknown coefficients for the mth basis function $(m = 1, 2, ..., M_{max})$ or for the constant 1 $(m = 0)$. A set of eligible knots is assigned separately for each variable dimension and is chosen to approximately coincide with the input levels represented in the data.

Let us consider Eqn. (2.3). Here, we can write the equation as

$$\eta = H(\mu) = \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{\psi}(\boldsymbol{t})^T\boldsymbol{\theta}, \tag{2.5}$$

where $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_{M_{max}})^T$ and $\boldsymbol{\psi}^T(\boldsymbol{t}) = \left(1, \psi_1(\boldsymbol{t}^1), \ldots, \psi_{M_{max}}\left(\boldsymbol{t}^{M_{max}}\right)\right)^T$. Furthermore, let us refer to PRSS with M_{max} basis functions having been accumulated for the CMARS model. Now, PRSS has the following form [60]:

$$\begin{aligned} PRSS \;:=\; & \sum_{i=1}^{N}\left(\eta_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{\psi}(\boldsymbol{t}_i)^T\boldsymbol{\theta}\right)^2 \\ & + \sum_{m=1}^{M_{max}} \lambda_m \sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_m}} \int_{Q^m} \theta_m^2 \left[D_{r,s}^{\boldsymbol{\alpha}}\psi_m(\boldsymbol{t}^m)\right]^2 d\boldsymbol{t}^m, \end{aligned} \tag{2.6}$$

where $V_m := \left\{\kappa_j^m \mid j = 1, 2, ..., K_m\right\}$ is the variable set associated with the mth basis function and $\boldsymbol{t}^m = \left(t_{m_1}, t_{m_2}, ..., t_{m_{K_m}}\right)^T$ represents the vector of variables which contribute to the mth basis function. Moreover,

$$D_{r,s}^{\boldsymbol{\alpha}}\psi_m(\boldsymbol{t}^m) := \frac{\partial^{|\boldsymbol{\alpha}|}\psi_m}{\partial^{\alpha_1}t_r^m\,\partial^{\alpha_2}t_s^m}(\boldsymbol{t}^m) \tag{2.7}$$

for $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T$, $|\boldsymbol{\alpha}| := \alpha_1 + \alpha_2$, where $\alpha_1, \alpha_2 \in \{0, 1\}$. Our optimization problem bases on a trade-off between both accuracy, i.e., a small RSS, and not too high complexity. This trade-off is established through the penalty parameters. In this chapter,

we firstly represent that trade-off by penalty methods, such as regularization techniques and by CQP [59, 70].

If we take into account the Eqn. (2.4) in Eqn. (2.6), then the objective function in Eqn. (2.4) will be of the following form:

$$\begin{aligned} PRSS &= \sum_{i=1}^{N}\left(\eta_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \theta_0 - \sum_{m=1}^{M_{max}}\theta_m\psi_m(\boldsymbol{t}_i^m)\right)^2 \\ &\quad + \sum_{m=1}^{M_{max}}\lambda_m \sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_m}} \int_{Q^m} \theta_m^2 \left[D_{r,s}^{\boldsymbol{\alpha}}\psi_m(\boldsymbol{t}^m)\right]^2 d\boldsymbol{t}^m, \end{aligned} \tag{2.8}$$

where the vector $\bar{\boldsymbol{t}}_i^m = \left(\bar{t}_{i,1}^m, \bar{t}_{i,2}^m, \ldots, \bar{t}_{i,d_m}^m\right)^T$ denotes any of the input vectors and $\bar{\boldsymbol{t}}_i^m = \left(\bar{t}_{i,\kappa_1}^m, \bar{t}_{i,\kappa_2}^m, \ldots, \bar{t}_{i,\kappa_{d_m}}^m\right)^T$ $(\kappa = 1, 2, \ldots, p^m)$ stands for the corresponding projection vectors of $\bar{\boldsymbol{t}}_i^m$ onto those coordinates which contribute to the mth basis function, ψ_m, they are related with the ith link function η_i. We recall that those coordinates are collected in the set V_m. Let us note that the second-order derivatives of the piecewise linear functions ψ_m $(m = 1, 2, \ldots, M_{max})$ and, hence, the penalty terms related, are vanishing. Now, we can rearrange the representation of PRSS as follows:

$$\begin{aligned} PRSS &= \sum_{i=1}^{N}\left(\eta_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{\psi}\left(\bar{\boldsymbol{d}}_i\right)^T\boldsymbol{\theta}\right)^2 \\ &\quad + \sum_{m=1}^{M_{max}}\lambda_m \sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_m}} \int_{Q^m} \theta_m^2 \left[D_{r,s}^{\boldsymbol{\alpha}}\psi_m(\boldsymbol{t}^m)\right]^2 d\boldsymbol{t}^m, \end{aligned} \tag{2.9}$$

where $\boldsymbol{\psi}\left(\bar{\boldsymbol{d}}_i\right) = \left(1, \psi_1(\boldsymbol{t}_i^1), \ldots, \psi_{M_{max}}(\boldsymbol{t}_i^{M_{max}})\right)^T$, $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_{M_{max}})^T$ with the matrix $\bar{\boldsymbol{d}}_i = \left(\boldsymbol{t}_i^1, \boldsymbol{t}_i^2, \ldots, \boldsymbol{t}_i^{M_{max}}\right)^T$ in the argument. To approximate the multi-dimensional integrals

$$\int_{Q^m} \theta_m^2 \left[D_{r,s}^{\boldsymbol{\alpha}}\psi_m(\boldsymbol{t}^m)\right]^2 d\boldsymbol{t}^m, \tag{2.10}$$

a suitable discretization and model approximation are used. In fact, we approximate the discretized form of the integrals by Riemann sums as follows [70]:

$$\begin{aligned} &\int_{Q^m} \theta_m^2 \left[D_{r,s}^{\boldsymbol{\alpha}}\psi_m(\boldsymbol{t}^m)\right]^2 d\boldsymbol{t}^m \quad \approx \\ &\sum_{(\sigma^\kappa)_{\kappa\in\{1,2,\ldots,p^m\}}\in\{0,1,2,\ldots,N+1\}^{p^m}} \theta_m^2 \quad \left[D_{r,s}^{\boldsymbol{\alpha}}\psi_m(\bar{t}_{i_{\sigma^\kappa},\kappa}^m, \ldots, \bar{t}_{i_{\sigma^\kappa},\kappa}^m)\right]^2 \prod_{\kappa=1}^{p^m}\left(\bar{t}_{i_{\sigma^\kappa+1},\kappa}^m - \bar{t}_{i_{\sigma^\kappa},\kappa}^m\right). \end{aligned} \tag{2.11}$$

We can rearrange PRSS in this form:

$$PRSS \approx \sum_{i=1}^{N} \left(\eta_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{\psi}\left(\bar{\boldsymbol{d}}_i\right)^T\boldsymbol{\theta}\right)^2$$
$$+\sum_{m=1}^{M_{max}} \lambda_m \sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_m}} \sum_{(\sigma^\kappa)} \theta_m^2 \left[D_{r,s}^{\boldsymbol{\alpha}}\psi_m(\bar{t}^m_{i_{\sigma^\kappa},\kappa},...,\bar{t}^m_{i_{\sigma^\kappa},\kappa})\right]^2 \prod_{\kappa=1}^{p^m}\left(\bar{t}^m_{i_{\sigma^\kappa+1},\kappa} - \bar{t}^m_{i_{\sigma^\kappa},\kappa}\right),$$

where $(\sigma^\kappa)_{\kappa\in\{1,2,...,p^m\}} \in \{0,1,2,...,N+1\}^{p^m}$. Let us introduce some more notation related with the sequence (σ^κ):

$$\hat{\boldsymbol{t}}_i^m = \left(\bar{t}^m_{i_{\sigma^\kappa},\kappa},...,\bar{t}^m_{i_{\sigma^\kappa},\kappa}\right), \quad \Delta\hat{\boldsymbol{t}}_i^m = \prod_{\kappa=1}^{p^m}\left(\bar{t}^m_{i_{\sigma^\kappa+1},\kappa} - \bar{t}^m_{i_{\sigma^\kappa},\kappa}\right). \tag{2.12}$$

By Eqn. (2.12), we can approximate PRSS as

$$\begin{aligned} PRSS \;\approx\; & \sum_{i=1}^{N}\left(\eta_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{\psi}\left(\bar{\boldsymbol{d}}_i\right)^T\boldsymbol{\theta}\right)^2 \\ & +\sum_{m=1}^{M_{max}}\lambda_m\theta_m^2 \sum_{i=1}^{(N+1)^{p^m}}\left(\sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2}\sum_{\substack{r<s \\ r,s\in V_m}}\left[D_{r,s}^{\boldsymbol{\alpha}}\psi_m(\hat{\boldsymbol{t}}_i^m)\right]^2\right)\Delta\hat{\boldsymbol{t}}_i^m. \end{aligned} \tag{2.13}$$

For a short representation, we can rewrite the approximate relation in Eqn. (2.13) as

$$PRSS \approx \left\|\boldsymbol{\eta} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\psi}\left(\bar{\boldsymbol{d}}\right)\boldsymbol{\theta}\right\|_2^2 + \sum_{m=1}^{M_{max}}\lambda_m \sum_{i=1}^{(N+1)^{p^m}} L_{im}^2\theta_m^2, \tag{2.14}$$

where $\boldsymbol{\psi}\left(\bar{\boldsymbol{d}}\right) = \left(\mathbf{1}, \boldsymbol{\psi}\left(\bar{\boldsymbol{d}}_1\right), \boldsymbol{\psi}\left(\bar{\boldsymbol{d}}_2\right), \ldots, \boldsymbol{\psi}\left(\bar{\boldsymbol{d}}_N\right)\right)^T$ and $\boldsymbol{X}$ are $(N\times(M_{max}+1))$- and $(N\times p)$-matrices, respectively, $\|\cdot\|_2$ denotes the Euclidean norm and the numbers L_{im}^2 are defined by their roots:

$$L_{im} := \left[\left(\sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2}\sum_{\substack{r<s \\ r,s\in V_m}}\left[D_{r,s}^{\boldsymbol{\alpha}}\psi_m(\hat{\boldsymbol{t}}_i^m)\right]^2\right)\Delta\hat{\boldsymbol{t}}_i^m\right]^{1/2}. \tag{2.15}$$

If we consider Eqn. (2.14), we can write PRSS as

$$PRSS \approx \|\boldsymbol{\eta} - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2^2 + \sum_{m=1}^{M_{max}}\lambda_m \sum_{i=1}^{(N+1)^{p^m}} L_{im}^2\theta_m^2, \tag{2.16}$$

where $\boldsymbol{X}^* = \left(\boldsymbol{X} \mid \boldsymbol{\psi}\left(\bar{\boldsymbol{d}}\right)\right)$ is a block matrix constructed by $(N \times p)$-matrix $\boldsymbol{X}$ and $(N \times (M_{max}+1))$-matrix $\boldsymbol{\psi}\left(\bar{\boldsymbol{d}}\right)$, and $\boldsymbol{\beta}^* = \left(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T\right)^T$ is a vector composed of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ vectors. Then, we deal with the linear systems of equations, $\boldsymbol{X}^*\boldsymbol{\beta}^* = \boldsymbol{\eta}$, approximately. This problem may be ill-posed (irregular or unstable). For this reason, we firstly approach our problem, PRSS, as a TR problem [3, 9]; because this regularization belongs to the most commonly used methods of making ill-posed problems well-posed (regular or stable). A Tikhonov solution can be expressed quite easily in terms of *Singular Value Decomposition (SVD)* of the coefficient matrix $\boldsymbol{X}^*$ of a regarded linear system of equations, $\boldsymbol{X}^*\boldsymbol{\beta}^* = \boldsymbol{\eta}$.

For this purpose we consider the formula in Eqn. (2.16) again, arranging it as follows:

$$\begin{aligned} PRSS &\approx \|\boldsymbol{\eta} - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2^2 \\ &\quad + \sum_{m=1}^{M_{max}} \lambda_m \left[(L_{1m}\theta_m)^2 + (L_{2m}\theta_m)^2 + \ldots + \left(L_{(N+1)^{p^m}m}\theta_m\right)^2 \right], \\ PRSS &= \|\boldsymbol{\eta} - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2^2 + \sum_{m=1}^{M_{max}} \lambda_m \|\boldsymbol{L}_m\theta_m\|_2^2, \end{aligned} \tag{2.17}$$

where $\boldsymbol{L}_m = \left(L_{1m}, L_{2m}, \ldots, L_{(N+1)^{p^m}m}\right)^T$ $(m = 1, 2, \ldots, M_{max})$. However, rather than a singleton, there is a finite sequence of the trade-off or penalty parameters, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_{M_{max}})^T$, such that this equation is not yet a Tikhonov Regularization with a single such parameter. For this reason, let us make a uniform penalization by taking the same λ for each derivative term. Then, our approximation for the PRSS can be rearranged as

$$PRSS \approx \|\boldsymbol{\eta} - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2^2 + \lambda \|\boldsymbol{L}^*\boldsymbol{\beta}^*\|_2^2, \tag{2.18}$$

where $\boldsymbol{L}$ is a diagonal $((M_{max}+1) \times (M_{max}+1))$-matrix with first column $\boldsymbol{L}_0 = \boldsymbol{0}_{(N+1)^{p^m}}$ and the other columns being the vectors $\boldsymbol{L}_m$ introduced above. Moreover, $\boldsymbol{\theta}$ is an $((M_{max}+1) \times 1)$-parameter vector to be estimated through the data points. Furthermore, we consider the high-dimensional matrix $\boldsymbol{L}^* = (\boldsymbol{R} \mid \boldsymbol{L})$, where $\boldsymbol{R}$ is an $(p \times p)$-matrix formed by the zeroth-, first- or second-order derivatives of $\boldsymbol{\beta}$. These derivatives are given by zeroth-, first- or second-order difference quotients of $\boldsymbol{\beta}$, regarded as a function that is evaluated at the points i, or i and $i+1$, or $i-1$, i and $i+1$, respectively. Those later difference quotients approximate the first- and second-order derivatives, whereas the zeroth-order derivative of the vector m is m itself. Altogether, these terms are comprised by products $\boldsymbol{R}\boldsymbol{\beta}$ of $\boldsymbol{\beta}$ with matrices that represent the discrete differential operators of zeroth-, first- and second-order, respectively. These matrices are in the form of a band-structure with values 1 or -1, 1, or 1, -2, 1, on the band. To achieve $M_{max}+1$ row vectors, we can fill up the matrices $\boldsymbol{R}$ on 0th-, 1th- and 2nd-order discrete derivatives, which have less many rows if $M_{max}+1 > p$, by rows $\boldsymbol{0}^T$, respectively. Now, our PRRS problem looks like a classical Tikhonov Regularization problem with $\phi > 0$, i.e., $\lambda = \phi^2$ for some $\phi \in \mathbb{R}\setminus\{0\}$, as follows [59, 70]:

$$PRSS \approx \|\boldsymbol{\eta} - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2^2 + \lambda \|\boldsymbol{L}^*\boldsymbol{\beta}^*\|_2^2. \tag{2.19}$$

We easily note that our Tikhonov Regularization problem has multiple objective functions through a linear combination of $\|\boldsymbol{\eta} - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2^2$ and $\|\boldsymbol{L}^*\boldsymbol{\theta}^*\|_2^2$. In fact, we select the solution such that it minimizes both the first objective $\|\boldsymbol{\eta} - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2^2$ and the second objective $\|\boldsymbol{L}^*\boldsymbol{\theta}^*\|_2^2$ in the sense of a compromise (trade-off) solution. For a contribution to the dependence of locally linear embedding on regularization parameter(s), we refer to [65].

The penalty parameter λ in the PRSS and the parametrical upper bound $\tilde{M}$ in a constraint of the CQP can be chosen in a related way, determined via Tikhonov Regularization. This regularization method utilities an efficiency curve that comes from a plot of the optimal solutions to the problem in Eqn. (2.13) according to a large (finite) number of parameter values, as points in a coordinate scheme with two axes. At one axis, the complexity is denoted, whereas the other axis stands for the length of the residual vector (or goodness-of-fit). In this regularization method, logarithmical scales are employed such that some "kink" (corner) kind of a point on the efficiency boundary, called the L-curve according to its more pronounced shape, is obtained. This point is regarded to be the closest to the origin and therefore it is often chosen, together with the corresponding penalty parameter [3]. For our approach with CQP, in [70], a lot of numerical experience is presented, related with varying upper bounds $\sqrt{\tilde{M}}$ by using the software package of MOSEK. The solutions obtained at the upper bounds are expected to be the same non-dominated solutions as obtained in Eqn. (2.19), to the multi-objective problem of minimizing complexity and maximizing goodness-of-fit [70].

Let us tackle the Tikhonov Regularization problem in Eqn. (2.19) with the CQP, which is a continuous optimization technique. Indeed, based on an appropriate choice of a bound $\tilde{M}$, we can state the following optimization problem:

$$\begin{aligned} \underset{\boldsymbol{\beta}^*}{\text{minimize}} \quad & \|\boldsymbol{\eta} - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2^2 \\ \text{subject to} \quad & \|\boldsymbol{L}^*\boldsymbol{\beta}^*\|_2^2 \leq \tilde{M}. \end{aligned} \tag{2.20}$$

Let us underline that this choice of $\tilde{M}$ is the outcome of the careful learning process as explained in Chapter 1 [20]. In Eqn. (2.19), we have the least-squares objective function $\|\boldsymbol{\eta} - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2^2$ and the inequality constraint function $-\|\boldsymbol{L}^*\boldsymbol{\beta}^*\|_2^2 + \tilde{M}$, which is requested to be nonnegative for feasibility. Now, we equivalently write our optimization problem as follows:

$$\begin{aligned} \underset{t,\boldsymbol{\beta}^*}{\text{minimize}} \quad & t, \\ \text{subject to} \quad & \|\boldsymbol{\eta} - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2 \leq t, \\ & \|\boldsymbol{L}^*\boldsymbol{\beta}^*\|_2 \leq \sqrt{\tilde{M}}. \end{aligned} \tag{2.21}$$

Let us use modern methods of continuous optimization techniques, especially, from CQP where we employ the subsequent basic notation [41]:

$$\underset{\boldsymbol{x}}{\text{minimize}} \ \boldsymbol{c}^T\boldsymbol{x}, \quad \text{subject to} \ \|\boldsymbol{D}_i\boldsymbol{x} - \boldsymbol{d}_i\|_2 \leq \boldsymbol{p}_i^T\boldsymbol{x} - q_i \ \ (i = 1, 2, ..., k). \tag{2.22}$$

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

$$\boldsymbol{c} = (1, \mathbf{0}_{M_{max}+p+1}^T)^T,\ \boldsymbol{x} = (t, \boldsymbol{\beta}^{*T})^T,\ \boldsymbol{D}_1 = (\mathbf{0}_N, \boldsymbol{X}^*),\ \boldsymbol{d}_1 = \boldsymbol{\eta},\ \boldsymbol{p}_1 = (1, 0, ..., 0)^T,$$
$$q_1 = 0,\ \boldsymbol{D}_2 = (\mathbf{0}_{M_{max}+p+1}, \boldsymbol{L}^*),\ \boldsymbol{d}_2 = \mathbf{0}_{M_{max}+p+1,}\ \boldsymbol{p}_2 = \mathbf{0}_{M_{max}+p+2} \text{ and } q_2 = -\sqrt{\tilde{M}}.$$

Now, having written the Tikhonov Regularization task for GPLMs by evaluating the CMARS with CQP problem, we call it *Conic Generalized Partial Linear Models (CGPLM)*. In fact, CGPLM provides a solution by applying the developed CQP techniques. These kinds of well-structured convex optimization problems have also been studied by Weber et al. for new approaches to regression and classification. In this respect, CGPLM has the advantage of higher speed and less complexity [65], and it permits the use of IPMs [41]. In order to write the optimality condition for this problem, we firstly reformulate the problem in Eqn. (2.21) as follows:

$$\begin{aligned}
\underset{t,\boldsymbol{\beta}^*}{\text{minimize}} \quad & t, \\
\text{such that} \quad & \boldsymbol{\chi} := \begin{pmatrix} \mathbf{0}_N & \boldsymbol{X}^* \\ 1 & \mathbf{0}_{M_{max}+p+1}^T \end{pmatrix} + \begin{pmatrix} t \\ \boldsymbol{\beta}^* \end{pmatrix} + \begin{pmatrix} -\boldsymbol{\eta} \\ 0 \end{pmatrix}, \\
& \boldsymbol{\eta} := \begin{pmatrix} \mathbf{0}_{M_{max}+p+1} & \boldsymbol{L}^* \\ 0 & \mathbf{0}_{M_{max}+p+1}^T \end{pmatrix} \begin{pmatrix} t \\ \boldsymbol{\beta}^* \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{M_{max}+p+1} \\ \sqrt{\tilde{M}} \end{pmatrix}, \\
& \boldsymbol{\chi} \in \boldsymbol{L}^{N+1},\ \boldsymbol{\eta} \in \boldsymbol{L}^{M_{max}+p+2}, \qquad (2.23)
\end{aligned}$$

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

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

The *dual problem* to the latter primal one is given by

$$\begin{aligned}
& \text{maximize} \quad (\boldsymbol{\eta}^T, 0)\omega_1 + \left(\mathbf{0}_{M_{max}+p+1}^T, -\sqrt{\tilde{M}} \right) \omega_2 \\
& \text{such that} \quad \begin{pmatrix} \mathbf{0}_N^T & 1 \\ \boldsymbol{X}^* & \mathbf{0}_{M_{max}+p+1}^T \end{pmatrix} \boldsymbol{\omega}_1 + \begin{pmatrix} \mathbf{0}_{M_{max}+p+1}^T & 0 \\ (\boldsymbol{L}^*)^T & \mathbf{0}_{M_{max}+p+1} \end{pmatrix} \boldsymbol{\omega}_2 = \begin{pmatrix} 1 \\ \mathbf{0}_{M_{max}+p+1} \end{pmatrix}, \\
& \boldsymbol{\omega}_1 \in L^{N+1},\ \boldsymbol{\omega}_2 \in L^{M_{max}+p+2}. \qquad (2.24)
\end{aligned}$$

Moreover, $(t, \boldsymbol{\theta}, \boldsymbol{\chi}, \boldsymbol{\eta}, \boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ is a *primal dual optimal solution* if and only if

$$
\begin{aligned}
&\boldsymbol{\chi} := \begin{pmatrix} \mathbf{0}_N & \boldsymbol{X}^* \\ 1 & \mathbf{0}_{M_{max}+p+1}^T \end{pmatrix} + \begin{pmatrix} t \\ \boldsymbol{\beta}^* \end{pmatrix} + \begin{pmatrix} -\boldsymbol{\eta} \\ 0 \end{pmatrix}, \\
&\boldsymbol{\eta} := \begin{pmatrix} \mathbf{0}_{M_{max}+p+1} & \boldsymbol{L}^* \\ 0 & \mathbf{0}_{M_{max}+p+1}^T \end{pmatrix} \begin{pmatrix} t \\ \boldsymbol{\beta}^* \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{M_{max}+p+1} \\ \sqrt{\tilde{M}} \end{pmatrix}, \\
&\begin{pmatrix} \mathbf{0}_N^T & 1 \\ \boldsymbol{X}^* & \mathbf{0}_{M_{max}+p+1}^T \end{pmatrix} \boldsymbol{\omega}_1 + \begin{pmatrix} \mathbf{0}_{M_{max}+p+1}^T & 0 \\ (\boldsymbol{L}^*)^T & \mathbf{0}_{M_{max}+p+1} \end{pmatrix} \boldsymbol{\omega}_2 = \begin{pmatrix} 1 \\ \mathbf{0}_{M_{max}+p+1} \end{pmatrix}, \\
&\boldsymbol{\omega}_1^T \boldsymbol{\chi} = 0, \;\; \boldsymbol{\omega}_2^T \boldsymbol{\eta} = 0, \\
&\boldsymbol{\omega}_1 \in L^{N+1}, \;\; \boldsymbol{\omega}_2 \in L^{M_{max}+p+2}, \\
&\boldsymbol{\chi} \in L^{N+1}, \;\; \boldsymbol{\eta} \in L^{M_{max}+p+2}.
\end{aligned} \tag{2.25}
$$

2.2.1 Estimation Procedure for Conic Generalized Partial Linear Models

After the unified theoretical treatment of our CGPLM in Section 2.2, we now come to a practical treatment. Here, the basic idea will be that we first represent the given data by the linear model. Then, we will subtract that model, along the different observations, from the data, in order, finally to approximate the residual data vector by the nonlinear model. In this bi-level approach, we call the first step a *preprocessing*, namely, for our CMARS technique of the second step. Here, CMARS can be considered as a refinement and fine-tuning step with respect to the linear model [60].

For the estimation of the linear parametric part of the CGPLM, a regularized least-squares aproximation is applied [66]. As a result, the following form is obtained:

$$
Y^{preproc} = \boldsymbol{X}^T \boldsymbol{\beta} + \epsilon = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon. \tag{2.26}
$$

Now, we get the linear model coefficients' vector $\boldsymbol{\beta}^{preproc}$ after solving the Eqn. (2.26). We apply our regularization technique with CQP for this purpose. Then, for each data point, a value $\bar{\boldsymbol{X}}_i^T \bar{\boldsymbol{\beta}}^{preproc}$ is subtracted (without intercept, β_0) from the corresponding response, and $\hat{\boldsymbol{y}}$ is obtained by

$$
(\boldsymbol{\eta} =) \hat{\boldsymbol{y}} := \boldsymbol{y} - \bar{\boldsymbol{X}} \bar{\boldsymbol{\beta}}^{preproc}.
$$

In fact, $\boldsymbol{y}$ is the given response data vector, $\bar{\boldsymbol{X}}$ is the given input data matrix except its first column (of entries 1) which is multiplied by the intercept β_0, and $\bar{\boldsymbol{\beta}}^{preproc}$ is the vector of the further unknown parameters. Thus, the residual vector, $\hat{\boldsymbol{y}}$, is used as the new response vector to construct our nonlinear CMARS model, based on the given input data also, for an estimation of nonparametric part of CGPLM. Then, the entire CGPLM is identified, constituted of two parts [60].

2.3 Application and Results

In this section, the implementation of CGPLM algorithm (with CMARS) and the comparison study of the Linear Model (LM), CMARS and CGPLM (with CMARS) will be presented. To solve the CQP problem, MOSEK software is preferred by us. Two different data sets with different sample sizes and numbers of independent variables are selected from regression test problems from the *UCI Machine Learning Repository* [62]. The first data set is the *Concrete Compressive Strength* test data [62]. Based on its sample size, this data set is the largest data set used in this chapter. It includes 1030 observations and 8 input variables, which are cement, blast furnace slag, fly ash, water, super plasticizer, coarse aggregate, age, and fine aggregate. The second data set for the application part of this chapter is the *Concrete Slump* test data [62]. This data set includes 103 observations. There are seven input variables, which are cement, slag, fly ash, water, SP, coarse aggregate, as well as fine aggregate, and an output variable known as 28-day compressive strength (MPa). All input variables for each data sets are quantitative.

For our approaches with CQP, many different upper bounds are tested by using the software package of MOSEK to find non-dominated solutions as explained in Chapter 1. The $\tilde{M}$ values used in our CQP problem between 1 and 200 are: 1, 1.5, 2, 2.5, 3, 3.5, ..., 199, 199.5, 200. For Concrete Compressive Strength test data, the selected $\tilde{M}$ values for CMARS and CGPLM (with CMARS) are 102 and 100, respectively. On the other hand, for Concrete Slump test data, 87 is the $\tilde{M}$ value for CMARS and 99 the $\tilde{M}$ value for CGPLM (with CMARS).

The formulation and evaluation of the well-known performance measures used in this thesis are given in Table 2.2 [72].

Table 2.2: Prediction Performance Measures.

Name of Measure	Evaluation	Formula
Coefficient of Determination (R^2)	Higher values are the better.	$R^2 := 1 - \frac{\sum_{i=1}^{N}(\bar{y}_i - \hat{\bar{y}}_i)^2}{\sum_{i=1}^{N}(\bar{y}_i - \bar{\bar{y}})^2}$
Adjusted R^2 (Adj-R^2)	Higher values are the better.	$Adj - R^2 := 1 - (1 - R^2)\frac{N-1}{N-p-1}$
Mean Absolute Error (MAE)	Smaller values are the better.	$MAE := \frac{1}{N}\sum_{i=1}^{N}\left\|\bar{y}_i - \hat{\bar{y}}_i\right\|$
Mean Squared Error (MSE)	Smaller values are the better	$MSE := \frac{1}{N-p}\sum_{i=1}^{N}(\bar{y}_i - \hat{\bar{y}}_i)^2$
Proportion of Residuals within Three Sigma (PWI)	Higer values are the better.	$PWI := \frac{a}{N}$

Here, $\bar{y}_i$ is ith observed response value; $\hat{\bar{y}}_i$ is ith fitted response; $\bar{\bar{y}}$ is the mean response; N is the number of observations; p is the number of terms in the model; $e_i = \bar{y}_i - \hat{\bar{y}}_i$ is

ith ordinary residual; $\bar{e}$ is the mean of ordinary residual; $\sigma = \sqrt{\left(\sum_{i=1}^{N}(e_i - \bar{e})^2\right)/N}$ is the standard deviation of ordinary residual; $T = 3 \cdot \sigma$ denotes 3σ as a threshold; a is the sum of the indicator variables where the condition $|e_i| < T$ is satisfied [72].

For the application of CGPLM, the independent variables for each data set are separated according to their linear and nonlinear patterns by the help of statistical software MINITAB [35]. After the construction of LM as a parametric approach, CMARS as a nonparametric approach and CGPLM as a semiparametric approach for both data sets, the performance measures that are described in Appendix A are computed. The results are presented in Table 2.3.

As it can be seen from Table 2.3, CMARS performs better than LM with respect to all measures for both data sets. On the other hand, CGPLM performs better than both LM and CMARS according to all performance measures. As a result, we can say that nonparametric models, as represented here by CMARS give more accurate results than parametric models (namely, LM), since nonparametric models are more flexible and efficient to express the nonlinearities. On the other hand, semiparametric models CGPLM give more reliable results than both parametric and nonparametric models because they reveal not only the linear structures but also nonlinearities and nonconvexities in the high-dimensional and complex data sets.

Table 2.3: Performance results of LM, CMARS and CGPLM for the given data sets.

Data Sets	**Performance Measures**	**Parametric** *LM*	**Nonparametric** *CMARS*	**Semiparametric** *CGPLM*
Concrete Compressive Strength Data Set	MAE	8.2133	3.9924	*3.6797
	RMSE	10.3991	5.4516	*4.9373
	R^2	0.6155	0.8992	*0.9252
	Adj-R^2	0.6125	0.8935	*0.9202
	PWI	*0.9980	0.9932	0.9902
Concrete Slump Test Data Set	MAE	10.2559	7.6841	*7.2680
	RMSE	12.8583	11.4831	*10.4763
	R^2	0.5010	0.6607	*0.7168
	Adj-R^2	0.4643	0.5444	*0.6519
	PWI	1.0000	1.0000	1.0000

* indicates better performance

CHAPTER 3

OUTLIER IDENTIFICATION WITH CONIC MULTIVARIATE ADAPTIVE REGRESSION SPLINES

3.1 Introduction

Data sets are very important for statistical analysis and data mining. Some of the observations in a data set deviate from other observations to a great extent, arousing suspicion that they were generated by some different mechanism. These observations are generally considered as an error but they might carry important information. Hawkins (1980) called such an observation an *outlier*. However, Barnet and Lewis (1994) similarly defined an outlier as an observation that appears to deviate noticeably from other observations in the data set [4]. A geometrical representation of some outliers is shown in Figure 3.1.

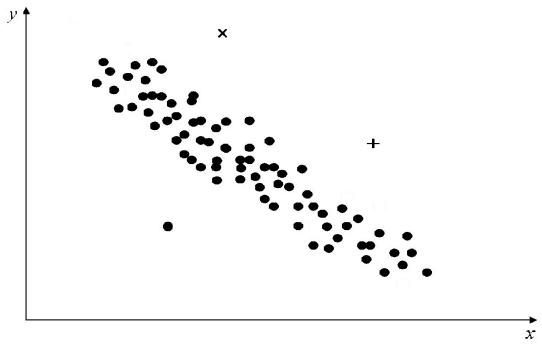

Figure 3.1: An illustration of some outliers in a two-dimensional space; the three outliers are marked by **.**, **+**, **x**.

If the data set contains a single or few outliers, the identification of such observations is not difficult. However, in most cases, data sets contain many outlier observations; therefore, identifying such observations becomes rather difficult because of the effects of masking and swamping [18]. *Masking* occurs when one outlier is not detected because of the presence of others, while *swamping* occurs whenever a non-outlier is wrongly identified, caused by the effect of some hidden outliers [44].

In such cases, these observations will cause biased parameter estimations; then, a sta-

tistical study can yield a misleading interpretation of a model via the estimates of the individual model coefficients. Therefore, it is very important to identify these observations and bound their influence through methods that are less sensitive against outliers in the data, i.e., robust methods, or via the elimination of the outliers from the data set. In this chapter, a mathematical alternative to robust statistical methods is provided by constructing a CQP with its very efficient IPMs and CMARS [61].

In Section 3.2, outlier identification method for linear models are reviewed. In Section 3.3, an alternative approach with Tikhonov Regularization and CQP is presented for mean-shift outlier model which is introduced in Section 3.2. In Section 3.4, by using the power of CMARS, the proposed method is improved. Finally, in Section 3.5, applications and comparisons of the presented methods are provided by using four data sets with different characteristics.

3.2 Outlier Identification Methods for Linear Models

In this chapter, outlier identification methods are considered for parametric and non-parametric methods. Parametric methods are not appropriate for high-dimensional and nonlinear data sets since they require certain distributional assumptions to be validated as well as a fixed structure for the parametric form.

Shortly recalling Chapter 2, as a parametric modeling, the LM [48], with p independent variables, is given by

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon,$$

where Y is the response variable and X_j $(j = 1, 2, ..., p)$ are the random input variables and $\boldsymbol{X} = (X_1, X_2, ..., X_p)^T$ represents the vector of predictors. The coefficient (or unknown parameter) β_0 is the intercept, the parameters β_j are the regression coefficients related with the independent variables X_j $(j = 1, 2, ..., p)$, and ϵ is the random error term, called noise. In other words, $Y = f(\boldsymbol{X}) + \epsilon$ with $f(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$ $(\boldsymbol{x} \in \mathbb{R}^p)$. The data response values and data input vectors are $y_i, \boldsymbol{x}_i$ $(i = 1, 2, \ldots, N)$, repectively. When these data values and points are inserted into the model, then the LM turns into the following linear system:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{3.1}$$

Here, $\boldsymbol{y}$ is an $(N \times 1)$-vector of the response data, y_i, $\boldsymbol{X}$ is an $(N \times (p+1))$-matrix of independent variables with the $(1 \times (p+1))$-vectors $\boldsymbol{x}_i^T$ $(i = 1, 2, ..., N)$ of input data as the rows, $\boldsymbol{\beta}$ is a $((p+1) \times 1)$-vector of unknown parameters. Sometimes, we call $\boldsymbol{X}$ as the *design matrix*. Furthermore, $\boldsymbol{\epsilon}$ is an $(N \times 1)$-vector of residuals, identically distributed random errors whose conditional mean and variance are given by $E(\boldsymbol{\epsilon} \mid \boldsymbol{x}) = 0$ and $\text{Var}(\boldsymbol{\epsilon} \mid \boldsymbol{x}) = \sigma^2 \boldsymbol{I}_N$, respectively. Here, σ^2 is an unknown parameter, and $\boldsymbol{I}_N$ is the identity matrix of order N. Assuming that we have at least as many observations as unknowns, i.e., $N \geq p+1$, and that the matrix $\boldsymbol{X}$ has full rank, the least-squares estimates

of $\boldsymbol{\beta}$ and σ^2 are given by $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$ and $\boldsymbol{y}^T\left(\boldsymbol{I}-\boldsymbol{H}\right)\boldsymbol{y}/\left(N-p-1\right)$, where $\boldsymbol{H} := \boldsymbol{X}^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ is the so-called *hat operator matrix*, yielding the estimated response $\hat{\boldsymbol{y}} = \boldsymbol{H}\hat{\boldsymbol{\beta}}$ [48].

For LM, there are different approaches for outlier identification. These approaches are separated into two categories: (i) *direct* approaches and (ii) *indirect* approaches. Indirect approaches are using residuals from a robust fit.

Direct approaches: Hadi and Simonoff [18] presented two methods for direct approaches. In their first method, they initially separate the data set into a clean subset without outliers and a complementary subset that contains all potential outliers. They assume that the data set contains K outliers. The most likely outliers are defined as the subset of K observations, which produce the largest reduction in the RSS when deleted [18]. Then, they test the outlyingness of the remaining points relative to the clean subset. Let U be the set of indexes of the observations in the clean subset with initial size $h = \lfloor\left(N+k-1\right)/2\rfloor$, which means the integer part of $\left(N+k-1\right)/2$. Here, k is the number of possible outliers. Then, they compute the scale d_i based on the clean subset U as follows:

$$d_i := \begin{cases} \frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_U}{\hat{\sigma}_U\sqrt{1-\boldsymbol{x}_i^T\left(\boldsymbol{X}_U^T\boldsymbol{X}_U\right)^{-1}}}, & \text{if} \quad i \in U, \\ \\ \frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_U}{\hat{\sigma}_U\sqrt{1+\boldsymbol{x}_i^T\left(\boldsymbol{X}_U^T\boldsymbol{X}_U\right)^{-1}}}, & \text{if} \quad i \notin U, \end{cases}$$

where d_i is the scale Studentized residual if $i \in U$ or the prediction error if $i \notin U$. Here, y_U and $\boldsymbol{X}_U$ together are the subset of observations, and $\hat{\boldsymbol{\beta}}_U$ and $\hat{\sigma}_U$ are the estimated regression coefficients and the MSE, respectively, indexed by U. As a third step, Hadi and Simonoff arrange the observations in ascending order according to the $|d_i|$ and they compare d_{s+1} by $t_{(\alpha/2(s+1),s-k)}$. Here, d_{s+1} is the $(s+1)$th-order statistic of $|d_i|$, and $t_{(\alpha/2(s+1),s-k)}$ is the upper critical value of the t distribution with $s-k$ degrees of freedom. If $d_{s+1} \geq t_{(\alpha/2(s+1),s-k)}$, then Hadi and Simonoff declare all observations satisfying $|d_i| \geq t_{(\alpha/2(s+1),s-k)}$ as outliers, and they stop the computation. Otherwise, they construct a new subset, U, taking the first $s+1$ ordered observations. If $N = s+1$, then they declare "no outliers" in the data and stop computation; otherwise, they continue to the second step. Their second approach is similar to the first one; however, it just differs in step 1. For more details, one may refer to [18]. According to Pena and Yohai [44], the success of the procedure is based on the initial clean subset of the data. This procedure works well for low-leverage outliers, i.e., for values that are outliers with respect to an independent variable. Deletion of this type of outlier shows a low influence on the regression fit. However, this procedure may fail when the sample contains a set of several high-leverage outliers. This type of deletion technique reveals a strong leverage and influence on the regression coefficients.

Indirect approaches: An indirect approach to outlier identification can be achieved through a robust regression estimate. The aim of robust regression is to provide resistant (stable) results in the presence of outliers. To achieve this stability, robust regres-

sion limits the influence of outliers on the parameter estimates. Three classes of problems have been addressed with robust regression techniques: problems with outliers in the y-direction (response direction), problems with multivariate outliers in the covariate space (i.e., outliers in the $\boldsymbol{x}$-space, which are also referred to as leverage points) and problems with outliers in both the y-direction and the $\boldsymbol{x}$-space. Many methods have been developed for these problems. In statistical applications of outlier detection and robust regression, however, the methods that are most commonly used today are *M estimation* (this class of estimators can be regarded as a generalization of *Maximum-Likelihood Estimation (MLE)*) [23] and *Least Trimmed Square (LTS) estimation* [51]. However, in this chapter, an outlier identification method called *Mean-Shift Outlier Model (MSOM)* [11, 28] (see Subsection 3.2.1) is considered, which is different from both M estimation and LTS estimation [61].

3.2.1 Mean-Shift Outlier Model

In evaluating the given outliers detection methods, the MSOM has to be mentioned; it is an indirect method. In fact, since MSOM gives the same RSS as the model fitted after omitting the relevant observations, it is convenient for studying the regression model in the presence of outliers. This model is given by

$$Y = \boldsymbol{X}^T\boldsymbol{\beta} + \Delta\delta + \epsilon, \tag{3.2}$$

where $\Delta \in \{0, 1\}$ is a constant "switching" or selection term, and δ is the unknown parameter for outlier observation. In the presence of an outlier, $\Delta = 1$, and the importance of an outlier are represented by the value $1 \cdot \delta$. The system after all data inserted into the model is as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}_i\delta + \boldsymbol{\epsilon}, \tag{3.3}$$

where $\boldsymbol{e}_i$ is the ith standard unit vector, i.e., $\boldsymbol{e}_i = (0, ..., 1, 0, ..., 0)^T$ $(i = 1, 2, ..., N)$. In this model, it is assumed that either y_i or $\boldsymbol{x}_i^T\boldsymbol{\beta}$ deviates systematically by some value δ from the equation $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$. Then, according to the ith observation, $\left(y_i, \boldsymbol{x}_i^T\boldsymbol{\beta}\right)$ would have a different intercept than the remaining observations, and $\left(y_i, \boldsymbol{x}_i^T\boldsymbol{\beta}\right)$ would hence represent an outlier. To check this fact, we test the hypothesis

$$H_0 : \delta = 0 \quad \text{(i.e., } E(Y) = \boldsymbol{X}\boldsymbol{\beta}\text{)}$$

against the alternative

$$H_1 : \delta \neq 0 \quad \text{(i.e., } E(Y) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}_i\delta\text{)},$$

using the likelihood-ratio test statistic [46]

$$F_i = \frac{\left(RSS(H_0) - RSS(H_1)\right)/1}{RSS(H_1)/\left(N - p - 1\right)}. \tag{3.4}$$

Here, $RSS(H_0)$ is the residual sum of the squares in the model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, containing all N of the observations, and $RSS(H_1)$ is the residual sum of squares in the model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}_i\delta + \boldsymbol{\epsilon}$, i.e.,

$$RSS(H_0) = \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2 = \boldsymbol{y}^T\left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{y}\left(N - p\right)\hat{\sigma}^2,$$

$$RSS(H_1) = \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \delta_i\right)^2,$$

where δ_i will be δ or 0 according to the outlier variable. Additionally, the relationship between $RSS(H_0)$ and $RSS(H_1)$ can be written as

$$RSS(H_1) = RSS(H_0) - \frac{\hat{\epsilon}_i^2}{1 - h_{ii}}, \tag{3.5}$$

where $\hat{\boldsymbol{\epsilon}} = \left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{y}$, $\hat{\epsilon}_i = \boldsymbol{e}_i^T\hat{\boldsymbol{\epsilon}}$, and $\boldsymbol{e}_i^T\boldsymbol{H}\boldsymbol{e}_i = h_{ii}$. When the ith observation $(y_i, \boldsymbol{x}_i)$ is omitted, then the estimator of σ_i^2 is defined by

$$s_{-i}^2 = \frac{y_i^T\left(\boldsymbol{I} - \boldsymbol{H}_i\right)\boldsymbol{y}_i}{N - p - 1}, \tag{3.6}$$

where $\boldsymbol{H}_i$ and $\boldsymbol{y}_i$ represent the hat matrix and the response vector after omission of the ith observation, respectively; σ_i is the standard deviation of the ith residual. If σ_i is taken as its estimation, $\hat{\sigma}_i = s_{-i}\sqrt{1 - h_{ii}}$, then the test statistic in Eqn. (3.4) may be written as

$$F_i = \frac{\hat{\epsilon}_i^2}{s_{-i}^2\left(1 - h_{ii}\right)} = \left(r_i^*\right)^2 \quad \left(i = 1, 2, ..., N\right), \tag{3.7}$$

where r_i^* is called the ith *externally Studentized residuals* [46]. Furthermore, the non-centrality parameter approaches zero as h_{ii} increases. Therefore, the detection of outliers becomes difficult if the ith leverage h_{ii} is large [46].

3.3 Alternative Approach for Mean-Shift Outlier Model with Tikhonov Regularization and Conic Quadratic Programming

As we mentioned, the goal of robust regression is to provide resistant (stable) results in the presence of outliers, limiting the influence of outliers on the parameter estimates. For this reason, the Tikhonov Regularization [3] method is considered and applied by us for MSOM. Tikhonov Regularization belongs to the most commonly used methods for making ill-posed problems well-posed (regular or stable) as we mentioned before in Section 2.2 [3, 9]. There are different types of basic formulation for the Tikhonov Regularization problems, all of which are represented as minimization problems. In

this section, a Tikhonov Regularization problem is constructed for the MSOM, and it is solved by using CQP [7]; this problem will be called *Conic Mean-Shift Outlier Model (CMSOM)*.

For CMSOM, we try to construct minimization problems mentioned by the Tikhonov Regularization for the MSOM. Let us assume that m observations y_i, where $m < N$, deviate systematically from the corresponding model $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$ by δ_i, as detected by one of the direct methods such as the Hadi and Simonoff [18] test statistic F_i given by Eqn. (3.4), Cook's distance [11] or Studentized residuals [11]. That is, m observations are outliers. Then, MSOM takes the form of the fallowing linear system

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E}\boldsymbol{\delta} + \boldsymbol{\epsilon}, \tag{3.8}$$

where $\boldsymbol{X}$ is a full rank $(N \times (p+1))$-matrix of independent (or explanatory) variables, $\boldsymbol{E}$ is an $(N \times m)$-matrix with m indicator vectors, and $\boldsymbol{\delta}$ is an $(m \times 1)$-vector of the regression coefficients of the indicator vector. In fact, MSOM can be stated as

$$\boldsymbol{y} = \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{3.9}$$

where $\boldsymbol{X}^* = (\boldsymbol{X} \mid \boldsymbol{E})$ is a block-structured $(N \times (p+1+m))$-matrix constructed by the matrices $\boldsymbol{X}$ and $\boldsymbol{E}$, and $\boldsymbol{\beta}^* = \left(\boldsymbol{\beta}^T, \boldsymbol{\delta}^T\right)^T$ is an $((p+1+m) \times 1)$-vector, constructed by the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$. Under the discrepancy principle [9], the Tikhonov Regularization problems for the regression model (3.8) can be written as

$$\underset{\boldsymbol{\beta}^*}{\text{minimize}} \quad \|\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2^2 + \lambda \|\boldsymbol{L}\boldsymbol{\beta}^*\|_2^2. \tag{3.10}$$

Here, λ can be considered as a penalty parameter. This parameter establishes the trade-off between both accuracy, i.e., a small RSS, and not too high complexity. The regularization matrix, $\boldsymbol{L}$, is a diagonal $((p+m+1) \times (p+m+1))$-matrix to represent the discrete differential operators of zeroth-, first- and second- order derivatives, filling up with rows $\boldsymbol{0}^T$ if needed. In this section, $\boldsymbol{L}$ is specified as an identity matrix.

The Tikhonov Regularization problem in Eqn. (3.10) can be addressed through CQP. Based on an appropriate choice of a bound $\tilde{M}$, the following optimization problem can be stated:

$$\begin{aligned} \underset{t,\boldsymbol{\beta}^*}{\text{minimize}} \quad & t, \\ \text{subject to} \quad & \|\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2 \leq t, \\ & \|\boldsymbol{L}\boldsymbol{\beta}^*\|_2 \leq \sqrt{\tilde{M}}. \end{aligned} \tag{3.11}$$

Note that through the thesis, $\tilde{M}$, the bound of the our optimization problem, is selected by a trial and error approach.

3.4 Improvements on Conic Mean-Shift Outlier Model with Conic Multivariate Adaptive Regression Splines

There are several well-known advantages of CMARS over traditional LM. One major drawback of LM is that in spite of all corrective measures applied, in some cases the constant error variance assumption may not be validated. In such a case, the theory underlying LM fails, and it does not provide the best linear and unbiased estimators for the model parameters anymore. For such data sets, CMARS is known to provide better predictions powerfully and clearly, according to the different comparison criteria [65].

If there are linear relationships between the variables of the given data set, CMARS method performs as good as LM does [72]. However, for high-dimensional data sets with a large number of predictors which exhibit nonlinear (complex) relationships, the complex structure of the data may prevent LM from developing valid and adequate statistical models. For these kind of prediction problems, there is numerical evidence that CMARS method provides even a better fit than traditional LM [72].

Finally, traditional regression models require certain distributional assumptions to be validated as well as a fixed structure for the parametric form. Thus these models need human expertise in their use and it may take a longer time to construct parametric models. In certain cases, we may even fail to develop LM models. On the other hand, CMARS models are developed "automatically," especially, adaptively requiring less human intervention. In addition, for complex data sets, the prediction algorithm should not adapt a parametric form prior to modeling but rather should explore the inherent structure of the data set to propose a nonparametric form. Thus, for the outlier detection problem, CMARS is employed as a novel and an effective tool.

The following model is constructed to remove the deficiency of CMSOM:

$$Y = \theta_0 + \sum_{m=1}^{M_{max}} \theta_m \psi_m(\boldsymbol{X}^m) + \Delta\delta + \epsilon. \tag{3.12}$$

Here, ψ_m $(m = 1, 2, ..., M_{max})$ are basis functions taken from a set of M_{max} linearly independent basis elements, θ_m are the unknown coefficients for the mth basis function $(m = 1, 2, ..., M_{max})$ or for the constant 1 $(m = 0)$. A special advantage of the basis function is lying in its ability to estimate the contributions of the basis functions so that both the additive and the interactive effects of the predictors are allowed to determine the dependent response variable. It should be noted that from now on we use the bar, $\bar{(\cdot)}$, on top of the data in order to distinguish them from both estimations and generic variables. Thus, the model with data inserted is leading to the following linear system:

$$\boldsymbol{y} = \boldsymbol{\psi}\left(\bar{\boldsymbol{d}}_i\right)\boldsymbol{\theta} + \boldsymbol{E}\boldsymbol{\delta} + \boldsymbol{\epsilon}, \tag{3.13}$$

where $\boldsymbol{\psi}\left(\bar{\boldsymbol{d}}_i\right) := \left(1, \psi_1(\bar{\boldsymbol{x}}_i^1), ..., \psi_M(\bar{\boldsymbol{x}}_i^M), \psi_{M+1}(\bar{\boldsymbol{x}}_i^{M+1}), ..., \psi_{M_{max}}(\bar{\boldsymbol{x}}_i^{M_{max}})\right)^T$, $\boldsymbol{\theta} := (\theta_0, \theta_1, ..., \theta_{M_{max}})^T$ with the point $\bar{\boldsymbol{d}}_i := \left(\bar{\boldsymbol{x}}_i^1, \bar{\boldsymbol{x}}_i^2, ..., \bar{\boldsymbol{x}}_i^M, \bar{\boldsymbol{x}}_i^{M+1}, ..., \bar{\boldsymbol{x}}_i^{M_{max}}\right)^T$ in the argument.

Now, our *MSOM with CMARS (CMARS-MSOM)* can be stated as

$$\boldsymbol{y} = \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{3.14}$$

where $\boldsymbol{X}^* = \left(\boldsymbol{\psi}\left(\bar{\boldsymbol{d}}_i\right)|\boldsymbol{E}\right)$ is a block-structured $(N \times (M_{max} + m + 1))$-matrix formed by the matrices $\boldsymbol{X}$ and $\boldsymbol{E}$, and $\boldsymbol{\beta}^* = \left(\boldsymbol{\theta}^T, \boldsymbol{\delta}^T\right)^T$ is an $((M_{max} + m + 1) \times 1)$-vector constructed by the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$.

Moreover, $\boldsymbol{\theta}$ is an $((M_{max} + 1) \times 1)$-parameter vector to be estimated through the data points. One of the components of the high-dimensional matrix, $\boldsymbol{L}^* = (\boldsymbol{R} \mid \boldsymbol{L})$, is $\boldsymbol{R}$ which is an $((M_{max} + 1) \times (M_{max} + 1))$-matrix formed by integrals of squared first- or second-order derivatives of the basis functions of CMARS. Now, our PRRS problem turns to a classical Tikhonov Regularization problem with penalty parameter, λ, as follows [59, 70]:

$$\underset{\boldsymbol{\beta}^*}{\text{minimize}} \quad \|\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2 + \lambda \|\boldsymbol{L}^*\boldsymbol{\beta}^*\|_2 . \tag{3.15}$$

Now, we equivalently write our optimization problem as follows:

$$\begin{aligned} \underset{t,\boldsymbol{\beta}^*}{\text{minimize}} \quad & t, \\ \text{subject to} \quad & \|\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2 \leq t, \\ & \|\boldsymbol{L}^*\boldsymbol{\beta}^*\|_2 \leq \sqrt{\tilde{M}}. \end{aligned} \tag{3.16}$$

To write the optimality conditions for this problem, we first reformulate the problem of Eqn. (3.16) as follows:

$$\begin{aligned} &\underset{t,\boldsymbol{\beta}^*}{\text{minimize}} \quad t, \\ &\text{such that} \\ &\boldsymbol{\chi} := \begin{pmatrix} \boldsymbol{0}_{N\times 1} & \boldsymbol{X}^* \\ 1 & \boldsymbol{0}^T_{(M_{max}+m+1)\times 1} \end{pmatrix} + \begin{pmatrix} t \\ \boldsymbol{\beta}^* \end{pmatrix} + \begin{pmatrix} -\boldsymbol{y} \\ 0 \end{pmatrix}, \\ &\boldsymbol{\eta} := \begin{pmatrix} \boldsymbol{0}_{(M_{max}+m+1)\times 1} & \boldsymbol{L}^* \\ 0 & \boldsymbol{0}^T_{(M_{max}+m+1)\times 1} \end{pmatrix} \begin{pmatrix} t \\ \boldsymbol{\beta}^* \end{pmatrix} + \begin{pmatrix} \boldsymbol{0}_{(M_{max}+m+1)\times 1} \\ \sqrt{\tilde{M}} \end{pmatrix}, \\ &\boldsymbol{\chi} \in \boldsymbol{L}^{N+1},\ \boldsymbol{\eta} \in \boldsymbol{L}^{M_{max}+m+2}, \end{aligned}$$

where L^{N+1}, $L^{M_{max}+m+2}$ are the $(N+1)$- and $(M_{max}+m+2)$-dimensional Lorentz cones. A primal-dual optimal solution is $(t, \boldsymbol{\theta}, \boldsymbol{\chi}, \boldsymbol{\eta}, \boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$.

In this chapter, we are addressing a need for *regularization* and perform it in two ways and different respects:

1. We take possible outliers into account at all by mathematically modeling and then determining to what an amount we let any such data having an impact on our model, herewith also "identifying" them. This is a first way to make the

model "regular" with respect to the existence of outliers. We may say that by the basic settings of our model, we "reach out" to possible outliers and represent their possible character and contribution to the model by a parameter δ, which has to be assessed numerically.

2. As always in this thesis, whether we are in a standard LM or in our nonlinear CMARS model, we do a regularization with the help of CQP programs which represent a Tikhonov Regularization problem that, besides of model accuracy, addresses its complexity, too. Besides through the linear system, which bases on model and data, the parameter vector of the outlier detection and assessment of item 1 is included. Now, that item of our efforts has become subordinate to the regularization of item 2 also. One may say that we even assess the basic problematic of outliers and to what an extent to take it into account via this second respect of item 2.

3.5 Numerical Performance Comparisons of LM, MSOM, CMSOM and CMARS-MSOM

In this section, the performances of four approaches (LM, MSOM, CMSOM and CMARS-MSOM) are compared according to some general measures. Firstly, the data sets and the performance measures are described. Then, the outcomes of the comparison studies are presented.

3.5.1 Data Sets

To compare the performance of the LM, MSOM, CMSOM and CMARS-CMSOM, four data sets, with different sample sizes (N) and numbers of independent variables (p), are selected from regression test problems. These test problems are application taken from the industry, chemistry and computer science areas.

Data Set 1: The first data set is the delivery time data taken from Rousseeuw and Leroy [50], page 155, Table 23. In this data set, which contains 25 observations collected from the service of a vending machine, y is the delivery time, X_1 is the number of cases that the product stocked, and X_2 is the distance walked by the route driver.

Data Set 2: As the second data set for comparison, the stack data set is considered. This data set is a well-known stack-loss data set presented by Brownlee [52] and was taken from SAS Customer Support. The data describe the operation of a plant for the oxidation of ammonia to nitric acid and consist of 21 4-dimensional observations. The explanatory variables for the response stack-loss (y) are the rate of operation (X_1), the cooling water inlet temperature (X_2), and the acid concentration (X_3).

Data Set 3: The third data set, a simulation data generated by using MATLAB [34], has 100 observations. To determine which methods perform better, a test problem is taken from [22]. In this problem, data are artificially created from the following model

function:

$$f(\boldsymbol{X}) = 24.55X_1 + 26.75X_2 + 39X_3 + 40.50X_4,$$

where the regressors X_1, X_2, X_3, and X_4 are assumed to have a uniform (-10, 10) distribution. To obtain the response, normally distributed noise with 0 mean is added to the model function $f(\boldsymbol{X})$.

Data Set 4: The fourth data set of this chapter is the Concrete slump test data [62]. The data set includes 103 observations. There are seven input variables, which are cement, slag, fly ash, water, SP, coarse aggregate, and fine aggregate, and an output variable known as 28-day compressive strength (MPa).

3.5.2 Applications and Results

For these applications, a special code has been written using MATLAB for CMSOM and CMARS-MSOM. Besides, we use MATLAB while implementing LM, MSOM, CMSOM and CMARS-MSOM.

To find the potential outliers for each data set, we apply the following ***outlier detection procedure***:

1. *LM is constructed to fit the data.*
2. *The fit values and ordinary residuals are computed from Step 1.*
3. *Studentized residual and Cook's distance values are calculated.*
4. *The potential outlier specified after Step 3 is removed from the data set, and Steps 1 and 2 are repeated to check for a better fit. (An observation is a potential outlier if it has larger Cook's distance and Studentized residual values.)*
5. *To find and remove the other potential outlier observations, Step 1 is repeated until all of the outlier observations are eliminated.*

After applying above procedure to all data sets, The number of potential outliers is presented in Table 3.1.

In our applications, to solve the CQP problem, MOSEK software is preferred [37]. As an upper bound of the CQP problem for CMSOM and CMARS-MSOM methods, the selected $\tilde{M}$ values are listed in Table 3.2.

The performance of the methods LM, MSOM, CMSOM and CMARS-MSOM are calculated for all data sets, and the results are given in Table 3.3. The values of some measures such as MAE and RMSE vary with the data set studied, i.e., there are no well-defined bounds on these values of the performance measures. For others, the values close to one are the better. For Data Sets 1 and 3, CMSOM gives a slightly better performance than LM and MSOM, according to all the chosen measures. For

Table 3.1: The number of potential outliers for the given data sets.

Data Sets		Data Features		# Potential Outliers
Number	**Name**	***N***	***p***	***k***
1	Delivery time data	25	2	2
2	Stack data	21	3	5
3	Artificial data	100	4	3
4	Concrete slump test data	103	7	5

Table 3.2: The selected $\tilde{M}$ values for the applications of CMSOM and CMARS-MSOM methods on the given data sets.

Data Sets	**1**	**2**	**3**	**4**
$\tilde{M}$ for CMSOM	35	69	101	98
$\tilde{M}$ for CMARS-MSOM	84	76	112	100

the other data sets, CMSOM and MSOM produce the same performance with respect to all the performance measures. When considering outliers, the results for LM show a weak or inferior performance compared to CMSOM and MSOM. Besides, with the advantages of CMARS method, CMARS-MSOM performs better than the other models. Since CMARS works effectively for high-dimensional and nonlinear data sets, CMARS-MSOM is successfully applied to more complex data sets containing outlier observations.

3.6 Concluding Remarks

This chapter provides a new contribution to the challenges of the mean-shift outlier regression problem by enabling the accessibility and usability of modern methods of continuous optimization and data mining. Herewith, a bridge has been offered between statistical learning, inverse problems and the strong tools prepared for well-structured convex optimization problems [61].

This chapter on the detection and following treatment of outliers benefits from prior data and, hence, model regularization, which we achieve with the help of CQP. Furthermore, the handling of outliers itself can be regarded as an extension of this process of regularization. Both are measurements oriented in the same direction of "smoothening" the data via a model that represents core information and they are orchestrated in this chapter.

In the field of the stochastic process of financial mathematics, there is a particular case of an outlier, which is called a *spike* [25]. As time proceeds, it can be regarded as a double jump, i.e., a jump up or down followed by a jump down or up, respectively.

Table 3.3: Performance results of the methods LM, MSOM, CMSOM and CMARS-MSOM for the data sets studied.

Data Set	**Models**	**MAE**	**RMSE**	**R^2**	**Adj-R^2**	**PWI**
1	LM	2.2867	3.3367	0.9596	0.9538	1.0000
	MSOM	1.6654	2.5473	0.9787	0.9743	1.0000
	CMSOM	1.6589	2.4826	0.9787	0.9755	1.0000
	CMARS-MSOM	*1.2634	*2.2048	*0.9890	*0.9798	1.0000
2	LM	2.3974	3.2846	0.9114	0.8957	1.0000
	MSOM	0.7383	1.2735	0.9898	0.9843	1.0000
	CMSOM	0.7383	1.2735	0.9898	0.9843	1.0000
	CMARS-MSOM	*0.5603	*1.2722	*0.9929	*0.9843	1.0000
3	LM	30.8827	39.6762	0.9901	0.9897	1.0000
	MSOM	29.8022	37.7668	0.9914	0.9907	1.0000
	CMSOM	27.6567	*35.5002	0.9924	0.9920	1.0000
	CMARS-MSOM	*27.3291	35.6751	*0.9926	*0.9916	1.0000
4	LM	1.8938	2.6087	0.8969	0.8892	0.9903
	MSOM	1.5020	2.1310	0.9348	0.9261	1.0000
	CMSOM	1.5020	2.1310	0.9348	0.9261	1.0000
	CMARS-MSOM	*0.6906	*0.91818	*0.9882	*0.9868	1.0000

* indicates better performance

Here, the second jump can be regarded as a type of (maybe partial) "withdrawal" of the first jump, which explains the outlier property of the spike. On the other hand, in different areas of technology, such as electrical engineering and mechanical engineering, an outlier can have an infinite response value as a *Dirac delta* (*impulse*) *function*. Then, it is an "infinite outlier" or "infinite (double) jump," and through infiniteness, it becomes very essential and can be employed systematically [3].

CHAPTER 4

PARAMETER IDENTIFICATION OF STOCHASTIC DIFFERENTIAL EQUATIONS BY CONIC MULTIVARIATE ADAPTIVE REGRESSION SPLINES

4.1 Introduction

Financial processes as well as processes in nature and technology are subject to random effects and noise. Generally, these time-dependent processes are characterized by their large number and by a high frequency. Describing high-frequency financial observations is a challenging issue because of the discretely discontinuous piecewise constant structures of their large data sets [26]. Thus, any related mathematical model for such given observations has to be constructed with a careful learning process since it faces a high sensitivity with respect to the slightest perturbations, as well as non-smoothness of the data. Such a mathematical model serves as the basis of prediction for the future [57].

Stochastic Differential Equations (SDEs) turn out to be an increasingly common tool for the representation of natural processes, and they play an important role to model the physical phenomena whose dynamics are affected by random noise [31]. In fact, these equations find application in many fields such as physics, biology, engineering, medicine, telecommunication and, in particular, finance. SDEs, however, are often hard to represent and to resolve because they involve unknown parameters or model functions nonlinearly, which need to be estimated from observations of the process [43].

Several approaches have been developed for parameter estimation of SDEs. Generally, these are difficult and sometimes impossible to apply because of the assumptions that have to be provided [8]. Therefore, we deal with this challenge by expressing SDEs with their parameters in a simplified manner of approximation with our CMARS method.

As already stated in Chapter 1, CMARS is one of the nonparametric approaches that make no specific assumptions to estimate the components of SDEs on some given stochastic process. Real-world processes from the financial sector or from nature are often characterized by their huge quantity and variations [13]. CMARS is able to model such data by using the advantages of recent developments in optimization the-

ory, numerical mathematics, computer software and hardware. The fundamental objective of our study is to propose CMARS as an identification tool for SDEs and to test this method on different kinds of stochastic processes without any assumptions [65, 70]. In this chapter, the stochastic processes could be price, logarithmic price, volatility, interest rate, wealth processes, etc. For studies in financial mathematics where our work could contribute as a tool or module, the references can be given as examples [13, 26, 31].

This chapter is organized as follows. In Section 4.2, a brief introduction to one-dimensional SDEs is given. Section 4.3 represents the proposed methodology for the parameter estimation of one-dimensional SDEs. In Section 4.4, some information on the problem of parameter estimation of multi-dimensional SDEs is presented. An introduction to multi-dimensional SDEs and the proposed methodology for parameter estimation of multi-dimensional SDEs are given in Section 4.5. In Section 4.6, correlated systems of SDEs are considered and analyzed.

4.2 One-dimensional Stochastic Differential Equations

Dynamical processes in nature, technology and economy are usually modeled by means of a deterministic differential equation such as ordinal differential equation (ODE), partial differential equation (PDE), impulsive differential equation (IDE), or delay differential equation (DDE). Consider an ODE with initial vector $x_0 \in \mathbb{R}$:

$$\dot{x}(t) = \left(:= \frac{d}{dt} x(t) \right) = a(x,t), \qquad x(t_0) = x_0.$$

where $x(\cdot) : [0, \infty) \to \mathbb{R}$ is a solution satisfying the initial condition $x(t_0) = x_0$ and $a : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$ is a smooth function. In general, $x(t)$ is not known explicitly, but it can be approximated by $a(x,t)$. We note that, rather than the initial value, we may also refer to a terminal value $x\left(t_T\right) = x_T$ at some "maturity time" $T > 0$.

On the other hand, in many applications, the process contains random effects and dependencies, and it cannot be modeled by the ODE. If the coefficients of the differential equation contain random effects, we cannot obtain a realistic mathematical model. Therefore, in order to describe the behaviour of the noise in differential equations, the SDEs are used. We call this real-valued special case among the system of SDEs studied in Section 4.5-4.6, a *one-dimensional SDE*. In this section, we briefly recall some concepts of SDEs and, especially, of a Wiener process, which is also called a standard Brownian motion.

Suppose we have a stochastic process $X = (X_t : t \geq 0)$, also written as $(X_t)_{t \geq 0}$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and presented by the following mathematical model [43]:

$$dX_t = a(t, X_t)dt + b(t, X_t)dW_t, \tag{4.1}$$

where $a(t, X_t)$ and $b(t, X_t)$ characterize the *drift* and *diffusion* terms, respectively. These terms satisfy the conditions to guarantee existence and uniqueness of unknown

process $(X_t)_{t\geq 0}$ on a certain interval, given values X_0 and X_T, respectively. Finally, $(W_t)_{t\geq 0}$ denotes a Wiener process at time t.

A *Wiener process* is a continuous-time stochastic process that fulfills the following three properties [2]:

- $W_0 = 0$ and $E[W_t] = 0$ for all $t \geq 0$.
- $W_t - W_s$ is $N(0, t-s)$ for all $0 \leq s \leq t$.
- Given any finitely many times $0 < t_1 < t_2 < ... < t_N$, the random variables (increments) W_{t_1}, $W_{t_2} - W_{t_1}$, ..., $W_{t_N} - W_{t_{N-1}}$ are independent.

From the Central Limit Theorem we learn that a Wiener process is a continuous and centered Gaussian process with the following conditions [43]:

$$E(W_t) = 0, \quad E(W_t^2) = \text{Var}(W_t) = t \quad \text{for any time} \quad t \geq 0.$$

Even over a small time interval, a Wiener process has an oscillating behavior of so-called *infinite variation*. Therefore, it is not differentiable for any time $t \geq 0$. To obtain our approximate and, then, smoothened model, we symbolically treat the real-valued stochastic process $(W_t)_{t\geq 0}$ as if it was differentiable; this is the first approach and a widespread notation in literature. We emphasize that this notation and SDEs, in general, can be exactly defined by *integral equations* which are based on *Itô calculus* [2, 29].

4.3 Parameter Identification of One-dimensional Stochastic Differential Equations with Conic Multivariate Adaptive Regression Splines

Estimation of the parameters of SDEs is very important in practice because in many applications, such as modeling the behavior of stock prices, the deterministic components in the dynamics is hard to represent [8]. The method which we adopt to estimate the drift and diffusion terms of SDEs is based on both discretization of the SDEs and our CMARS method. Since the realizations of the unknown $x_t = X_t(\omega)$ $(\omega \in \Omega)$ are not known exactly, we want to numerically approximate the time-discretized version of it and, hence, the time-continuous version, by this approach. In the following, we may suppress ω often for the sake of convenience.

4.3.1 Discretization of One-dimensional Stochastic Differential Equations

Simulation methods for SDEs are usually based on discrete approximations of the continuous-time solution curves. There are many discretization schemes for SDEs

such as *Euler* and *Runge-Kutta* discretization. Some of the discretization methods have a strong-order convergence. Although one of the most widely applied schemes of approximation is the Euler method which is also used to generate solutions to ordinary differential equations, we choose the *Milstein* scheme which employs Itô's Lemma to increase the accuracy of the approximation by adding a second-order term [29].

In this section, we consider one-dimensional real valued stochastic process, X_t. In the one-dimensional case, we represent the Milstein approximation of the unknown values $x_i := X_{t_i}$ at any increasing times $t_i \geq 0$, in short: $\hat{x}_i$ $(i = 1, 2, ..., N)$, as follows [77]:

$$\begin{aligned} \hat{x}_{i+1} &= \hat{x}_i + a(\hat{x}_i, t_i)(t_{i+1} - t_i) + b(\hat{x}_i, t_i)(\hat{w}_{i+1} - \hat{w}_i) \\ &\quad + \frac{1}{2} b(\hat{x}_i, t_i) b_x(\hat{x}_i, t_i) \left((\hat{w}_{i+1} - \hat{w}_i)^2 - (t_{i+1} - t_i) \right). \end{aligned}$$

Here, b_x is the partial derivative of $b(\hat{x}_i, t_i)$ with respect to x and $\hat{w}_i$ stands for an estimation of W_{t_i}. As soon as we refer to sometimes t_i, we may denote W_{t_i} by W_i. We write $\bar{W}_i$ in order to denote real-valued specifications in the context of random numbers drawn and data referred to. If we particularly address the finitely many given data points $(\bar{x}_i, \bar{t}_i)$, where $0 \leq \bar{t}_1 < \bar{t}_2 < ... < \bar{t}_i < \bar{t}_{i+1} < ... < \bar{t}_N$ are sampling times, we obtain a more symbolic form:

$$\begin{aligned} \bar{y}_i &= a(\bar{x}_i, \bar{t}_i) + b(\bar{x}_i, \bar{t}_i) \frac{\Delta \bar{w}_i}{\bar{h}_i} \\ &\quad + \frac{1}{2} b(\bar{x}_i, \bar{t}_i) b_x(\bar{x}_i, \bar{t}_i) \left(\frac{(\Delta \bar{w}_i)^2}{\bar{h}_i} - 1 \right) \quad (i = 1, 2, ..., N), \end{aligned} \tag{4.2}$$

where the value $\bar{y}_i$ represents the difference quotients raised on the ith data values $\bar{x}_i$ $(i = 1, 2, ..., N)$ and on step lengths $\Delta \bar{t}_i := \bar{h}_i = \bar{t}_{i+1} - \bar{t}_i$ between neighbouring sampling times:

$$\bar{y}_i = \begin{cases} (1/\bar{h}_i) \cdot (\bar{x}_{i+1} - \bar{x}_i), & \text{if} \quad i = 1, 2, ..., N-1, \\ (1/\bar{h}_N) \cdot (\bar{x}_N - \bar{x}_{N-1}), & \text{if} \quad i = N. \end{cases}$$

However, Eqn. (4.2) cannot be expected to hold in an exact sense, since it includes real data, affected by uncertainity; therefore, we satisfy them best in an approximate sense that refers to a minimal value of RSS, which is also called *Least Squares of Errors*. For the ease of exposition, the symbol, "=," is used instead of the approximation symbol "≈."

Since $W_{t_i} \sim N(0, t)$, the increments ΔW_{t_i} are independent on non-overlapping intervals and, moreover, $\text{Var}(\Delta W_{t_i}) = \Delta \bar{t}_i$. Therefore, the increments having normal distribution can be simulated with the help of standard normal distributed random numbers Z_i. Herewith, we obtain a discrete model for a Wiener process [29]:

$$\Delta W_i = Z_i \sqrt{\Delta \bar{t}_i}, \quad Z_i \sim N(0, 1). \tag{4.3}$$

If we insert Eqn. (4.3) into our Milstein approximation and draw random numbers $\bar{z}_i$ of Z_i (by a pseudo random number generator) [56, 77], we receive the following form:

$$\bar{y}_i = a(\bar{x}_i, \bar{t}_i) + b(\bar{x}_i, \bar{t}_i)\frac{\bar{z}_i}{\sqrt{\bar{h}_i}} + \frac{1}{2}b(\bar{x}_i, \bar{t}_i)b_x(\bar{x}_i, \bar{t}_i)\left(\bar{z}_i^2 - 1\right), \tag{4.4}$$

which we abbreviate by

$$\bar{y}_i = \bar{G}_i + \bar{H}_i c_i + \left(\bar{H}_i^{'}\bar{H}_i\right) d_i. \tag{4.5}$$

Here, $c_i := \bar{z}_i/\sqrt{\bar{h}_i}$, $d_i := (1/2)\cdot(\bar{z}_i^2 - 1)$, $\bar{G}_i := a(\bar{x}_i, \bar{t}_i)$ and $\bar{H}_i := b(\bar{x}_i, \bar{t}_i)$ $(i = 1, 2, ..., N)$. In view of the large number of functions, parameters and variables implied by us, as a small abuse of notation, here we exceptionally use capital letters for real-valued functions and for real numbers. For simplicity, we use prime "$(\cdot)^{'}$" to denote the partial derivative with respect to the variable x [56].

4.3.2 Construction of the Minimization Problem for the Parameter Estimation of One-dimensional Stochastic Differential Equations

To determine the unknown values of $\bar{G}_i$ and $\bar{H}_i$, we consider the following minimization problem:

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \sum_{i=1}^{N}\left(\bar{y}_i - \left(\bar{G}_i + \bar{H}_i c_i + \left(\bar{H}_i^{'}\bar{H}_i\right) d_i\right)\right)^2,$$

where the vector $\boldsymbol{\theta}$ comprises all the parameters in the Milstein approximation. We point out that also vector-valued processes can be studied, referring to sums of terms in the squared Euclidean norm $\|\cdot\|_2^2$.

This optimization problem could be solved with classical Gaussian least-squares estimation. However, the data underlying our stochastic process, especially, in the financial sector, have a high variation and oscillation. Therefore, a parameter estimation method has to be used which will take into account this high variation and will give a smoother approximation of the data [56].

4.3.3 Parameter Estimation of One-dimensional Stochastic Differential Equations using Conic Multivariate Adaptive Regression Splines

Spline approximation is very flexible, because it is *adaptive*, as it is explained below. It can be regularized in order to avoid too large an oscillation of the solutions; such an oscillation might be permitted by a high-degree polynomial approximation and be based on strongly varying data, by outliers or "spikes" existing. Splines can be described as linear combinations of basis splines which are usually low-dimensional. We

will approximate the data $(\bar{x}_i, \bar{t}_i)$ "smoothly," but we will compromise smoothness by the simplicity of certain discrete nondifferentiabilities [20]. In this chapter, we use CMARS' basis functions as our splines. The nondifferentiability at the knot points is a compromise for having "easy," in fact, piecewise linear one-dimensional basis functions.

CMARS is a powerful estimation method and it approximates the data in a "smoothening" way. In this section, a special CMARS application is done in the dimension $p = 2$. For this reason we approximate each function underlying the numbers $\bar{G}_i := a(\bar{x}_i, \bar{t}_i)$, $\bar{H}_i := b(\bar{x}_i, \bar{t}_i)$ and $\bar{F}_i := \bar{H}_i' \bar{H}_i$ by using the basis functions of CMARS and then introducing a regularization of the basic model. This treatment is very useful for the stability of our model in the presence of many and highly varying data [77].

The basis functions can be established in an additive way with respect to some subsets variables (coordinates), i.e., of subvectors, rather than separated variables known from *Generalized Additive Models (GAMs)* [57]. Referring to each such a subvector, the basis functions are multiplicative, as we shall state below. Here, it is also possible to construct $\bar{G}_i$, $\bar{H}_i c_i$ and $\bar{F}_i d_i$ as products of two basis functions, respectively. These are taken from a set of linearly independent one-dimensional basis elements. The construction of $\bar{G}_i$, $\bar{H}_i c_i$ and $\bar{F}_i d_i$ with CMARS' basis functions are represented as follows:

$$\begin{aligned} \bar{G}_i &= \beta_0 + \sum_{l=1}^{d^g} \beta_l \mathrm{b}_l(\bar{\boldsymbol{u}}_{i,\mathrm{b}}^l), \\ \bar{H}_i c_i &= \gamma_0 + \sum_{m=1}^{d^h} \gamma_m \mathrm{c}_m(\bar{\boldsymbol{u}}_{i,\mathrm{c}}^m), \\ \bar{F}_i d_i &= \delta_0 + \sum_{n=1}^{d^f} \delta_n \mathrm{d}_n(\bar{\boldsymbol{u}}_{i,\mathrm{d}}^n), \end{aligned} \tag{4.6}$$

where we use the unifying notation $\bar{\boldsymbol{u}}_{i,\mathrm{b}}^l,\ \bar{\boldsymbol{u}}_{i,\mathrm{c}}^m,\ \bar{\boldsymbol{u}}_{i,\mathrm{d}}^n = (\bar{x}_i, \bar{t}_i)\ (i = 1, 2, ..., N)$, and $d^g,\ d^h,\ d^f \in \mathbb{N}_0$ are suitable integers. The forms of b_l, c_m and d_n in Eqn. (4.6) given by CMARS basis functions are constructed subsequently:

$$\begin{aligned} \mathrm{b}_l(\bar{\boldsymbol{u}}_{\mathrm{b}}^l) &= \prod_{k=1}^{2} [s_{\kappa_k^l} \cdot (\bar{u}_{\kappa_k^l} - \tau_{\kappa_k^l})]_+, \\ \mathrm{c}_m(\bar{\boldsymbol{u}}_{\mathrm{c}}^m) &= \prod_{k=1}^{2} [s_{\kappa_k^m} \cdot (\bar{u}_{\kappa_k^m} - \tau_{\kappa_k^m})]_+, \\ \mathrm{d}_n(\bar{\boldsymbol{u}}_{\mathrm{d}}^n) &= \prod_{k=1}^{2} [s_{\kappa_k^n} \cdot (\bar{u}_{\kappa_k^n} - \tau_{\kappa_k^n})]_+. \end{aligned} \tag{4.7}$$

Here, $[q]_+ = \max\{0, q\}$ $(q \in \mathbb{R})$, $\bar{u}_{\kappa_k^l}$, $\bar{u}_{\kappa_k^m}$ and $\bar{u}_{\kappa_k^n}$ are the input variables corresponding to the kth truncated linear function in the lth, mth and nth basis functions, $\tau_{\kappa_k^l}$,

$\tau_{\kappa_k^m}$ and $\tau_{\kappa_k^n}$ are the knots value corresponding to the variables $\bar{u}_{\kappa_k^l}$, $\bar{u}_{\kappa_k^m}$ and $\bar{u}_{\kappa_k^n}$, and $s_{\kappa_k^l}$, $s_{\kappa_k^m}$ and $s_{\kappa_k^n}$ are the selected signs $+1$ or -1, respectively. For this case, permitted interaction is only up to two factors [14, 65, 70].

4.3.3.1 Construction of the Penalized Residual Sum of Squares Problem

We construct the PRSS as a regularization procedure for Eqn. (4.5) in the following form [20]:

$$\begin{aligned} PRSS \;:=\; & \sum_{i=1}^{N}(\bar{y}_i - \big(\bar{G}_i + \bar{H}_i c_i + \bar{F}_i d_i\big))^2 \\ & + \sum_{l=1}^{d^g}\lambda_l \sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_l}} \int_{Q^l} \beta_l^2 \left[D^{\boldsymbol{\alpha}}_{r,s}\mathrm{b}_l(\bar{\boldsymbol{u}}^l_{\mathrm{b}})\right]^2 d\boldsymbol{u}^l_{\mathrm{b}} \\ & + \sum_{m=1}^{d^h}\mu_m \sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_m}} \int_{Q^m} \gamma_m^2 \left[D^{\boldsymbol{\alpha}}_{r,s}\mathrm{c}_m(\bar{\boldsymbol{u}}^m_{\mathrm{c}})\right]^2 d\boldsymbol{u}^m_{\mathrm{c}} \\ & + \sum_{n=1}^{d^f}\nu_n \sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_n}} \int_{Q^n} \delta_n^2 \left[D^{\boldsymbol{\alpha}}_{r,s}\mathrm{d}_n(\bar{\boldsymbol{u}}^n_{\mathrm{d}})\right]^2 d\boldsymbol{u}^n_{\mathrm{d}}, \end{aligned} \tag{4.8}$$

where $V_l := \left\{\kappa_k^l |\; k=1,2\right\}$, $V_m := \{\kappa_k^m |\; k=1,2\}$ and $V_n := \{\kappa_k^n |\; k=1,2\}$ are the variable sets associated with the lth, mth and nth basis functions b_l, c_m and d_n, and $\bar{\boldsymbol{u}}^l_{\mathrm{b}}$, $\bar{\boldsymbol{u}}^m_{\mathrm{c}}$ and $\bar{\boldsymbol{u}}^n_{\mathrm{d}}$ represent the vectors of variables which contribute to the lth, mth and nth basis functions, respectively. Furthermore, we refer to the derivative terms

$$\begin{aligned} D^{\boldsymbol{\alpha}}_{r,s}\mathrm{b}_l(\bar{\boldsymbol{u}}^l_{\mathrm{b}}) &:= \frac{\partial^{|\boldsymbol{\alpha}|}\mathrm{b}_l}{\partial^{\alpha_1}u_r^l\;\partial^{\alpha_2}u_s^l}(\bar{\boldsymbol{u}}^l_{\mathrm{b}}), \\ D^{\boldsymbol{\alpha}}_{r,s}\mathrm{c}_m(\bar{\boldsymbol{u}}^m_{\mathrm{c}}) &:= \frac{\partial^{|\boldsymbol{\alpha}|}\mathrm{c}_m}{\partial^{\alpha_1}u_r^m\;\partial^{\alpha_2}u_s^m}(\bar{\boldsymbol{u}}^m_{\mathrm{c}}), \\ D^{\boldsymbol{\alpha}}_{r,s}\mathrm{d}_n(\bar{\boldsymbol{u}}^n_{\mathrm{d}}) &:= \frac{\partial^{|\boldsymbol{\alpha}|}\mathrm{d}_n}{\partial^{\alpha_1}u_r^n\;\partial^{\alpha_2}u_s^n}(\bar{\boldsymbol{u}}^n_{\mathrm{d}}), \end{aligned} \tag{4.9}$$

where $\boldsymbol{\alpha} = (\alpha_1,\alpha_2)^T$, $|\boldsymbol{\alpha}| := \alpha_1+\alpha_2$, and $\alpha_1,\alpha_2 \in \{0,1\}$. Indeed, we note that in any case where $\alpha_i = 2$, the derivatives $D^{\boldsymbol{\alpha}}_{r,s}\mathrm{b}_l(\bar{\boldsymbol{u}}^l_{\mathrm{b}})$, $D^{\boldsymbol{\alpha}}_{r,s}\mathrm{c}_m(\bar{\boldsymbol{u}}^m_{\mathrm{c}})$ and $D^{\boldsymbol{\alpha}}_{r,s}\mathrm{d}_n(\bar{\boldsymbol{u}}^n_{\mathrm{d}})$ vanish, and by referring to $r<s$, we have applied Schwarz's Theorem [70].

Our optimization problem in Eqn. (4.8) is based on the trade-off between accuracy, i.e., a small RSS, and a reduced complexity. This trade-off is established via penalty parameters λ_l, μ_m, $\nu_n \geq 0$. Large values of λ_l, μ_m and ν_n yield smoother curves and smaller ones result in more fluctuation. In this section, we approach that trade-off by a penalty approach, such as regularization techniques, and we conduct it through CQP.

When we use our multiplicative-additive form based on our multiplicative-basis splines for each model function, then the first part of the PRSS becomes

$$\sum_{i=1}^{N}(\bar{y}_i - \left(\bar{G}_i + \bar{H}_i c_i + \bar{F}_i d_i\right))^2 =$$
$$\sum_{i=1}^{N}\left(\bar{y}_i - \left(\beta_0 + \sum_{l=1}^{d^g}\beta_l \mathrm{b}_l(\bar{\boldsymbol{u}}_{i,\mathrm{b}}^l) + \gamma_0 + \sum_{m=1}^{d^h}\gamma_m \mathrm{c}_m(\bar{\boldsymbol{u}}_{i,\mathrm{c}}^m) + \delta_0 + \sum_{n=1}^{d^f}\delta_n \mathrm{d}_n(\bar{\boldsymbol{u}}_{i,\mathrm{d}}^n)\right)\right)^2 .$$

For the sake of convenience, we introduce the following matrix notation:

$$\beta_0 + \sum_{l=1}^{d^g}\beta_l \mathrm{b}_l(\bar{\boldsymbol{u}}_{i,\mathrm{b}}^l) + \gamma_0 + \sum_{m=1}^{d^h}\gamma_m \mathrm{c}_m(\bar{\boldsymbol{u}}_{i,\mathrm{c}}^m) + \delta_0 + \sum_{n=1}^{d^f}\delta_n \mathrm{d}_n(\bar{\boldsymbol{u}}_{i,\mathrm{d}}^n) = \bar{\boldsymbol{A}}_i\boldsymbol{\theta},$$

with $\bar{\boldsymbol{A}}_i := (\boldsymbol{b}_i,\ \boldsymbol{c}_i,\ \boldsymbol{d}_i)$, where $\boldsymbol{b}_i := \left(1, \mathrm{b}_1\left(\bar{\boldsymbol{u}}_{i,\mathrm{b}}^l\right), \mathrm{b}_2\left(\bar{\boldsymbol{u}}_{i,\mathrm{b}}^l\right), ..., \mathrm{b}_{d^g}\left(\bar{\boldsymbol{u}}_{i,\mathrm{b}}^l\right)\right)$,
$\boldsymbol{c}_i := \left(1, \mathrm{c}_1\left(\bar{\boldsymbol{u}}_{i,\mathrm{c}}^m\right), \mathrm{c}_2\left(\bar{\boldsymbol{u}}_{i,\mathrm{c}}^m\right), ..., \mathrm{c}_{d^h}\left(\bar{\boldsymbol{u}}_{i,\mathrm{c}}^m\right)\right)$,
$\boldsymbol{d}_i := \left(1, \mathrm{d}_1\left(\bar{\boldsymbol{u}}_{i,\mathrm{d}}^n\right), \mathrm{d}_2\left(\bar{\boldsymbol{u}}_{i,\mathrm{d}}^n\right), ..., \mathrm{d}_{d^f}\left(\bar{\boldsymbol{u}}_{i,\mathrm{d}}^n\right)\right)$, and $\boldsymbol{\theta} := \left(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\delta}^T\right)^T$, where
$\boldsymbol{\beta} := (\beta_0, \beta_1, \beta_2, ..., \beta_{d^h})^T$, $\boldsymbol{\gamma} := (\gamma_0, \gamma_1, \gamma_2, ..., \gamma_{d^f})^T$, $\boldsymbol{\delta} := (\delta_0, \delta_1, \delta_2, ..., \delta_{d^g})^T$.

Then, firstly, we obtain the residual sum of squares as the squared length of the difference vector between $\bar{\mathbf{y}}$ and $\bar{\boldsymbol{A}}\boldsymbol{\theta}$, where the matrix $\bar{\boldsymbol{A}} = \left(\bar{\boldsymbol{A}}_1^T, \bar{\boldsymbol{A}}_2^T, ..., \bar{\boldsymbol{A}}_N^T\right)^T$ contains the row vectors $\bar{\boldsymbol{A}}_i$, and the vector of difference quotients, $\bar{\mathbf{y}} = (\bar{y}_1, \bar{y}_2, ..., \bar{y}_N)^T$, represents the change rates of the given data [65, 77]:

$$\sum_{i=1}^{N}(\bar{y}_i - \bar{\boldsymbol{A}}_i\boldsymbol{\theta})^2 = \left\|\bar{\mathbf{y}} - \bar{\boldsymbol{A}}\boldsymbol{\theta}\right\|_2^2 .$$

Secondly, we obtain a discretized form of each integration term in Eqn. (4.8) by using Riemann sums. To approximate the multi-dimensional integrals, subsequently, the discretizations and model approximations are carefully prepared and refined. In a canonical way, the input data generate a subdivision of any sufficiently large parallelpipe $Q^l,\ Q^m,\ Q^n$ which contain all data points as elements. Let Q^l, Q^m and Q^n be a parallepipes which encompass all our input data; we represent them by

$$Q^l = \prod_{k=1}^{2} Q_k^l, \quad Q^m = \prod_{k=1}^{2} Q_k^m, \quad Q^n = \prod_{k=1}^{2} Q_k^n,$$

where $Q_k^\chi = [a_k^\chi, b_k^\chi]$ and $a_k^\chi \leq \bar{u}_{i,k} \leq b_k^\chi$ $(k = 1, 2;\ i = 1, 2, ..., N;\ \chi = l, m, n)$. Without loss of generality, for all k we may reorder the coordinates of the input data points: $\bar{u}_{l_1^k,k} \leq \bar{u}_{l_2^k,k} \leq ... \leq \bar{u}_{l_N^k,k}$, where $l_\sigma^j = 1, 2, ..., N$ $(k = 1, 2;\ \sigma = 1, 2, ..., N)$, and $\bar{u}_{l_\sigma^k,k}$, the kth component of $\bar{\boldsymbol{u}}_{l_\sigma^k}$ is the l_σ^kth input vector after reordering. Without

loss of generality we can assume $\bar{u}_{l^k_\sigma,k} \neq \bar{u}_{l^k_\varphi,k}$ for all $\sigma, \varphi = 1, 2, ..., N$, with $\sigma \neq \varphi$; i.e., $\bar{u}_{l^k_1,k} < \bar{u}_{l^k_2,k} < ... < \bar{u}_{l^k_N,k}$. Indeed, whenever "=" is attained for some coordinate, we obtained subparallelpipes of a lower dimension and its approximation, i.e., zero sets within the following integration process. Now, let us denote

$$\bar{u}_{l^k_0,k} = a^\chi_k,\ l^k_0 = 0, \quad \text{and} \quad \bar{u}_{l^k_{N+1},k} = b^\chi_k,\ l^k_{N+1} = N+1.$$

In order to increase the readability of the following formulas, we make a small simplification without loss of generality. Therefore, we assume $l^k_\sigma = l_\sigma$ $(k = 1, 2)$. Otherwise, we could just replace the index (or discrete parameter) l_σ by a k-depending, i.e., feature-depending, one called l^k_σ. Now, we can represent Q^χ_k as

$$Q^\chi_k = \bigcup_{\sigma^k=0}^{N} \prod_{k=1}^{2} \left[\bar{u}_{l_{\sigma^k},k}, \bar{u}_{l_{\sigma^{k+1}},k}\right] \quad (k = 1, 2;\ \chi = l, m, n).$$

The notation, the subdivision and the approximation presented above are done for all functions $\left[D^{\boldsymbol{\alpha}}_{r,s}\mathrm{b}_l(\bar{\boldsymbol{u}}^l_{\mathrm{b}})\right]^2$, $\left[D^{\boldsymbol{\alpha}}_{r,s}\mathrm{c}_m(\bar{\boldsymbol{u}}^m_{\mathrm{c}})\right]^2$ and $\left[D^{\boldsymbol{\alpha}}_{r,s}\mathrm{d}_n(\bar{\boldsymbol{u}}^n_{\mathrm{d}})\right]^2$, with the corresponding variables and lower dimensions of $\bar{\boldsymbol{u}}^l_{\mathrm{b}}$, $\bar{\boldsymbol{u}}^m_{\mathrm{c}}$ and $\bar{\boldsymbol{u}}^n_{\mathrm{d}}$ also. Then, if we apply this idea to our case, we discretize and represent each integration by the squared length of a vector after *Riemann sum* discretization as follows [65, 70]:

$$\begin{aligned}\int_{Q^l} \beta^2_l \left[D^{\boldsymbol{\alpha}}_{r,s}\mathrm{b}_l(\bar{\boldsymbol{u}}^l_{\mathrm{b}})\right]^2 d\boldsymbol{u}^l_{\mathrm{b}} \ &\approx \ \sum_{i=1}^{(N+1)^2} \left(\sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_m}} \beta^2_l \left[D^{\boldsymbol{\alpha}}_{r,s}\mathrm{b}_l(\hat{\boldsymbol{u}}^l_{i,\mathrm{b}})\right]^2 \right) \Delta\hat{\boldsymbol{u}}^l_{i,\mathrm{b}} \\ &= \ \sum_{i=1}^{(N+1)^2} \left(\bar{L}^{\mathrm{b}}_{il}\right)^2 \beta^2_l = \left\| \bar{\boldsymbol{L}}^{\mathrm{b}}_l \beta_l \right\|^2_2, \quad \text{where}\end{aligned}$$

$$\bar{L}^{\mathrm{b}}_{il} \ = \ \left[\left(\sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_m}} \beta^2_l \left[D^{\boldsymbol{\alpha}}_{r,s}\mathrm{b}_l(\hat{\boldsymbol{u}}^l_{i,\mathrm{b}})\right]^2 \right) \Delta\hat{\boldsymbol{u}}^l_{i,\mathrm{b}} \right]^{1/2}$$

with $\hat{\boldsymbol{u}}^l_{i,\mathrm{b}} = \left(\bar{u}_{l_{\sigma^{\kappa^l_k}},\kappa^l_k}, ..., \bar{u}_{l_{\sigma^{\kappa^l_k}},\kappa^l_k} \right)$, $\Delta\hat{\boldsymbol{u}}^l_{i,\mathrm{b}} = \prod_{k=1}^{2} \left[\bar{u}_{l_{\sigma^{\kappa^l_{k+1}}},\kappa^l_k} - \bar{u}_{l_{\sigma^{\kappa^l_k}},\kappa^l_k} \right]$ and with the sequence $\left(\sigma^{\kappa_j}\right)_{j\in\{1,2\}} \in \{0, 1, 2, ..., N+1\}^2$.

The same process is applied for the following multi-dimensional integrals:

$$\int_{Q^m} \gamma^2_l \left[D^{\boldsymbol{\alpha}}_{r,s}\mathrm{c}_m(\bar{\boldsymbol{u}}^m_{\mathrm{c}})\right]^2 d\boldsymbol{u}^m_{\mathrm{c}} \approx \sum_{i=1}^{(N+1)^2} \left(\bar{L}^{\mathrm{c}}_{im}\right)^2 \gamma^2_m = \left\| \bar{\boldsymbol{L}}^{\mathrm{c}}_m \gamma_m \right\|^2_2,$$

$$\int_{Q^n} \delta^2_n \left[D^{\boldsymbol{\alpha}}_{r,s}\mathrm{d}_n(\bar{\boldsymbol{u}}^n_{\mathrm{d}})\right]^2 d\boldsymbol{u}^n_{\mathrm{d}} \approx \sum_{i=1}^{(N+1)^2} \left(\bar{L}^{\mathrm{d}}_{in}\right)^2 \delta^2_n = \left\| \bar{\boldsymbol{L}}^{\mathrm{d}}_n \varphi_n \right\|^2_2.$$

Here, we refer to

$$\bar{L}^{\mathrm{c}}_{im} = \left[\left(\sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_m}} \gamma_m^2 \left[D^{\boldsymbol{\alpha}}_{r,s} \mathrm{c}_m(\hat{\boldsymbol{u}}^m_{i,\mathrm{c}}) \right]^2 \right) \Delta\hat{\boldsymbol{u}}^m_{i,\mathrm{c}} \right]^{1/2},$$

$$\bar{L}^{\mathrm{d}}_{in} = \left[\left(\sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_m}} \delta_n^2 \left[D^{\boldsymbol{\alpha}}_{r,s} \mathrm{d}_n(\hat{\boldsymbol{u}}^n_{i,\mathrm{d}}) \right]^2 \right) \Delta\hat{\boldsymbol{u}}^n_{i,\mathrm{d}} \right]^{1/2},$$

where $\hat{\boldsymbol{u}}^m_{i,\mathrm{c}} = \left(\bar{u}_{l_{\sigma^{\kappa^m_k},\kappa^m_k}}, ..., \bar{u}_{l_{\sigma^{\kappa^m_k},\kappa^m_k}} \right)$ and $\Delta\hat{\boldsymbol{u}}^m_{i,\mathrm{c}} = \prod_{k=1}^{2} \left[\bar{u}_{l_{\sigma^{\kappa^m_{k+1}},\kappa^m_k}} - \bar{u}_{l_{\sigma^{\kappa^m_k},\kappa^m_k}} \right]$; $\hat{\boldsymbol{u}}^n_{i,\mathrm{d}} = \left(\bar{u}_{l_{\sigma^{\kappa^n_k},\kappa^n_k}}, ..., \bar{u}_{l_{\sigma^{\kappa^n_k},\kappa^n_k}} \right)$ and $\Delta\hat{\boldsymbol{u}}^n_{i,\mathrm{d}} = \prod_{k=1}^{2} \left[\bar{u}_{l_{\sigma^{\kappa^n_{k+1}},\kappa^n_k}} - \bar{u}_{l_{\sigma^{\kappa^n_k},\kappa^n_k}} \right]$ [65, 70].

For a short representation, we can rewrite the approximate relation in Eqn. (4.8) as

$$PRSS \approx \left\| \bar{\mathbf{y}} - \bar{\boldsymbol{A}}\boldsymbol{\theta} \right\|_2^2 + \sum_{l=1}^{d^g} \lambda_l \left\| \bar{\boldsymbol{L}}^{\mathrm{b}}_l \beta_l \right\|_2^2 + \sum_{m=1}^{d^h} \mu_m \left\| \bar{\boldsymbol{L}}^{\mathrm{c}}_m \gamma_m \right\|_2^2 + \sum_{n=1}^{d^f} \nu_n \left\| \bar{\boldsymbol{L}}^{\mathrm{d}}_n \delta_n \right\|_2^2, \quad (4.10)$$

for $\bar{\boldsymbol{L}}^{\mathrm{b}}_l := \left(\bar{L}^{\mathrm{b}}_{1l}, \bar{L}^{\mathrm{b}}_{2l}, ..., \bar{L}^{\mathrm{b}}_{(N+1)^2 l} \right)^T (l = 1, 2, ..., d^g)$, $\bar{\boldsymbol{L}}^{\mathrm{c}}_m := \left(\bar{L}^{\mathrm{c}}_{1m}, \bar{L}^{\mathrm{c}}_{2m}, ..., \bar{L}^{\mathrm{c}}_{(N+1)^2 m} \right)^T \left(l = 1, 2, ..., d^h \right)$, and $\bar{\boldsymbol{L}}^{\mathrm{d}}_n := \left(\bar{L}^{\mathrm{d}}_{1n}, \bar{L}^{\mathrm{d}}_{2n}, ..., \bar{L}^{\mathrm{d}}_{(N+1)^2 n} \right)^T \left(l = 1, 2, ..., d^f \right)$. However, there is not a singleton, but an $M_{max} = \left(d^g + d^h + d^f \right)$-tuple of penalty (or trade-off, or smoothing) parameters: $(\lambda_1, \lambda_2, ..., \lambda_{d^g}, \mu_1, \mu_2, ..., \mu_{d^h}, \tau_1, \tau_2, ..., \tau_{d^f})$. Therefore, the minimization of PRSS is not a Tikhonov Regularization problem already as that need to have a single such multiplier. Therefore, let us make a *uniform* penalization by taking the same λ for each derivative term. Then, our approximation of PRSS can be rearranged as

$$PRSS \approx \left\| \bar{\mathbf{y}} - \bar{\boldsymbol{A}}\boldsymbol{\theta} \right\|_2^2 + \lambda \left\| \bar{\boldsymbol{L}}\boldsymbol{\theta} \right\|_2^2,$$

where $\bar{\boldsymbol{L}}$ is a diagonal $(M_{max}+1)\times(M_{max}+1)$-matrix with first column $\bar{\boldsymbol{L}}_0 = \mathbf{0}_{(N+1)^2}$ and the further columns $\bar{\boldsymbol{L}}^{\mathrm{b}}_l$, $\bar{\boldsymbol{L}}^{\mathrm{c}}_m$ and $\bar{\boldsymbol{L}}^{\mathrm{d}}_n$, which we presented above. Moreover, $\boldsymbol{\theta}$ is an $((M_{max}+1)\times 1)$-parameter vector to be estimated through the data points. Now, finally, our PRSS problem looks as a standard Tikhonov Regularization problem [3], however with a particular *regularization matrix* $\bar{\boldsymbol{L}}$.

4.3.3.2 Construction of the Conic Quadratic Programming Problem

Not solving the Tikhonov Regularization problem, but explaining how to treat this problem by using convex optimization techniques which we suppose to become a complementary and alternative technology to the concept of Tikhonov Regularization on

inverse problems we apply the elegant framework of CQP. With the statistical learning-based choice of the bound $\tilde{M}$ as we explained in Chapter 1, we represent the Tikhonov Regularization problem, as the following optimization program [65, 70]:

$$\begin{aligned}
&\underset{t,\boldsymbol{\theta}}{\text{minimize}}\ \ t,\\
&\text{such that}\ \ \boldsymbol{\chi} := \begin{bmatrix} \mathbf{0}_N & \bar{\boldsymbol{A}} \\ 1 & \mathbf{0}_{M_{max}+1}^T \end{bmatrix} \begin{bmatrix} t \\ \boldsymbol{\theta} \end{bmatrix} + \begin{bmatrix} -\bar{\mathbf{y}} \\ 0 \end{bmatrix},\\
&\qquad\qquad \boldsymbol{\eta} := \begin{bmatrix} \mathbf{0}_{M_{max}+1} & \bar{\boldsymbol{L}} \\ 0 & \mathbf{0}_{M_{max}+1}^T \end{bmatrix} \begin{bmatrix} t \\ \boldsymbol{\theta} \end{bmatrix} + \begin{bmatrix} \mathbf{0}_{M_{max}+1} \\ \sqrt{\tilde{M}} \end{bmatrix},\\
&\qquad\qquad \boldsymbol{\chi} \in L^{N+1},\ \boldsymbol{\eta} \in L^{M_{max}+2}.
\end{aligned} \tag{4.11}$$

In order to state the optimality conditions, we firstly write the dual problem to the latter problem as

$$\begin{aligned}
&\text{maximize}\quad (\bar{\mathbf{y}}^T, 0)\boldsymbol{\omega}_1 + \left(\mathbf{0}_{M_{max}+1}^T, -\sqrt{\tilde{M}}\right)\boldsymbol{\omega}_2,\\
&\text{such that}\ \begin{bmatrix} \mathbf{0}_N^T & 1 \\ \bar{\boldsymbol{A}}^T & \mathbf{0}_{M_{max}+1}^T \end{bmatrix}\boldsymbol{\omega}_1 + \begin{bmatrix} \mathbf{0}_{M_{max}+1}^T & 0 \\ \bar{\boldsymbol{L}}^T & \mathbf{0}_{M_{max}+1} \end{bmatrix}\boldsymbol{\omega}_2 = \begin{bmatrix} 1 \\ \mathbf{0}_{M_{max}+1} \end{bmatrix},\\
&\qquad \boldsymbol{\omega}_1 \in L^{N+1},\ \boldsymbol{\omega}_2 \in L^{M_{max}+2}.
\end{aligned}$$

Furthermore, $(t, \boldsymbol{\theta}, \boldsymbol{\chi}, \boldsymbol{\eta}, \boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ is a *primal-dual optimal solution* if and only if

$$\begin{aligned}
&\boldsymbol{\chi} = \begin{bmatrix} \mathbf{0}_N & \bar{\boldsymbol{A}} \\ 1 & \mathbf{0}_{M_{max}+1}^T \end{bmatrix} \begin{bmatrix} t \\ \boldsymbol{\theta} \end{bmatrix} + \begin{bmatrix} -\bar{\mathbf{y}} \\ 0 \end{bmatrix},\\
&\boldsymbol{\eta} = \begin{bmatrix} \mathbf{0}_{M_{max}+1} & \bar{\boldsymbol{L}} \\ 0 & \mathbf{0}_{M_{max}+1}^T \end{bmatrix} \begin{bmatrix} t \\ \boldsymbol{\theta} \end{bmatrix} + \begin{bmatrix} \mathbf{0}_{M_{max}+1} \\ \sqrt{\tilde{M}} \end{bmatrix},\\
&\begin{bmatrix} \mathbf{0}_N^T & 1 \\ \bar{\boldsymbol{A}}^T & \mathbf{0}_{M_{max}+1}^T \end{bmatrix}\boldsymbol{\omega}_1 + \begin{bmatrix} \mathbf{0}_{M_{max}+1}^T & 0 \\ \bar{\boldsymbol{L}}^T & \mathbf{0}_{M_{max}+1} \end{bmatrix}\boldsymbol{\omega}_2 = \begin{bmatrix} 1 \\ \mathbf{0}_{M_{max}+1} \end{bmatrix},\\
&\boldsymbol{\omega}_1^T\boldsymbol{\chi} = 0,\ \ \boldsymbol{\omega}_2^T\boldsymbol{\eta} = 0,\\
&\boldsymbol{\omega}_1 \in L^{N+1},\ \ \boldsymbol{\omega}_2 \in L^{M_{max}+2},\ \ \boldsymbol{\chi} \in L^{N+1},\ \ \boldsymbol{\eta} \in L^{M_{max}+2}.
\end{aligned}$$

For the application parts of Chapters 4 and 5, we do not give $\tilde{M}$ values. Let us refer to Subsection 1.1 where we explain our choice of the parameter value $\tilde{M}$, given a concrete problem and particular data set. In the future, we aim at including $\tilde{M}$ into the optimization problem as a further decision variable. Thus, our optimal solution will be model-free, data set dependent and heuristic parts of our methodology turn to become integrated, calculus supported approaches and techniques.

4.3.4 Applications and Results of One-dimensional Stochastic Differential Equations

In order to test our approach and its theory presented in Section 4.3, two different stochastic processes are selected from the field of finance. In applications, while CMARS' basis functions are built by using Salford MARS, the CMARS model is constructed by running a MATLAB code [34]. To solve the CQP problem, MOSEK [37] software is preferred. At the beginning, we describe the data sets and the performance measures. Then, we present the outcomes of the proposed approach.

Data Sets: As a numerical example, we choose a data set of simulating equity markets of different countries. These data sets are obtained through *default simulate* method in the Econometrics Toolbox in MATLAB. By using this method, a single path of correlated equity index prices is simulated for Canada, France, Germany, Japan, United Kingdom (UK) and United States (US) over one calendar year, respectively. In our numerical example, we select two of these processes: the US and the UK prices. These stochastic processes include prices of 250 trading days with the initial value of 100.

Firstly, we construct our CMARS model for the whole data set of the two countries (US and UK) to show and analyze the approximation of the model. The representative set of basis functions for the drift and the diffusion terms of the stochastic processes are listed below:

$$\wp_{US} = \{\psi_1,\ \psi_2,\ \psi_3,\ \psi_4,\ \psi_5,\ \psi_6\},$$

where

$$\begin{aligned}
\psi_1 &= \max\{0, x_{US} - 109.67\}, \quad \psi_2 = \max\{0, 109.67 - x_{US}\}, \\
\psi_3 &= \max\{0, x_{US} - 115.99\}, \quad \psi_4 = \max\{0, x_{US} - 107.59\}, \\
\psi_5 &= \max\{0, x_{US} - 118.26\}, \quad \psi_6 = \max\{0, x_{US} - 119.56\}.
\end{aligned}$$

And

$$\wp_{US} = \{\psi_1,\ \psi_2,\ \psi_3,\ \psi_4,\ \psi_5,\ \psi_6,\ \psi_7,\ \psi_8,\ \psi_9\},$$

where

$$\begin{aligned}
\psi_1 &= \max\{0, x_{UK} - 96.89\}, \quad \psi_2 = \max\{0, 96.89 - x_{UK}\}, \\
\psi_3 &= \max\{0, x_{UK} - 100.97\}, \quad \psi_4 = \max\{0, x_{UK} - 102.83\}, \\
\psi_5 &= \max\{0, x_{UK} - 114.31\}, \quad \psi_6 = \max\{0, x_{UK} - 104.88\}, \\
\psi_7 &= \max\{0, x_{UK} - 108.30\}, \quad \psi_8 = \max\{0, x_{UK} - 112.86\}, \\
\psi_9 &= \max\{0, x_{UK} - 111.58\}.
\end{aligned}$$

The CMARS algorithm constructs 6 basis functions for the US price; on the other hand, for the UK price it produces 9 basis functions. For both processes, each basis function is piecewise linear. Therefore, it provides convenience for interpretation and demonstration.

Figure 4.1 shows the path of the US price and the proposed approximation to this stochastic process with CMARS. The same is done for the UK price also. Figure 4.2 shows the path of the UK price and the proposed approximation to it. The approximation gives a promising result and captures the structures of both US and UK prices.

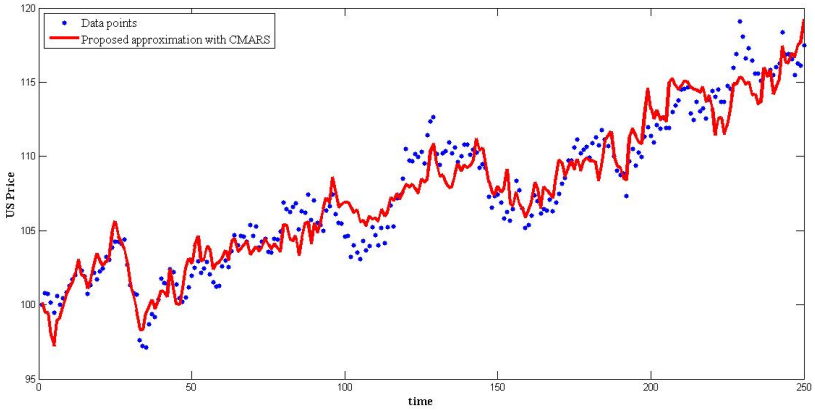

Figure 4.1: Demonstration of proposed approximation with CMARS for US price.

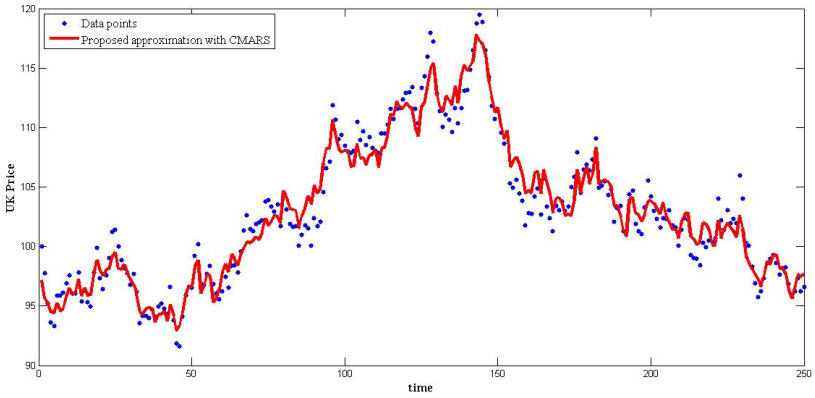

Figure 4.2: Demonstration of proposed approximation with CMARS for UK price.

After the models are built, comparison measures listed in Table 2.2 are calculated to show the estimation performance of CMARS method for each data set. For the stochastic processes used in this study, our approximation reveals a good performance according to these measures, as shown in Table 4.1. The smaller values of MAE and MSE indicate a strong estimation of the process parameters. On the other hand, the further measures (R^2, Adj-R^2, PWI) indicate a better performance if their values are closer to 1.0.

Table 4.1: Evaluation of proposed approximation with well-known statistical measures.

Performance Measures	**US Price**	**UK Price**
R^2	0.9950	0.9623
Adj-R^2	0.9946	0.9573
MAE	0.3012	0.9018
MSE	0.1836	2.0669
PWI	0.9800	0.9720

4.3.5 Conclusion and Outlook for of One-dimensional Stochastic Differential Equations

Identification of SDEs plays an important role in many application areas, including biology, chemistry, epidemiology, mechanics, microelectronics, economics and, especially, finance. A fundamental problem in estimating parameters in SDEs is the structure of the stochastic processes that can be of any distributional characteristics and any nonlinear and nonconvex form of the data set.

CMARS is a data mining methodology applied by us on a problem with dynamical contents and properties: the evolvement of financial and economical characteristics in time and under stochastic uncertainty. By this, the number of data grows at each point in time (a discrete time, after the approximation that we do), where it could be responded by another "constraint" or model compartment. By this, we could close the gap in terms of degree of freedom. We really gain from the regarded data set in terms of geometry and topology, do really "get into," or adapt to, the data and their spatial-temporal patterns, benefitting from structural characteristics of the data. In this sense, CMARS is a very good choice as in each dimension of the two input variables, we obtain a zig-zagging (piecewise linear) function, where the linear pieces mimic the data along subintervals. This adaptive approach is learning-based and "smart," and we use modern optimization theory for this purpose.

SDEs, incorporating random effects, are computationally expensive, especially when the processes are very complex, hence, when the number of parameters to be estimated is high. Therefore, advanced regularization and optimization techniques are required. In this chapter, using CQP programs gives us the opportunity to handle such problems. In addition to benefiting from well-structured optimization, CMARS basis functions yield a higher flexibility for the estimation of the stochastic process' parameters.

Differently from the approach by GAMs, which we used for the identification of SDEs in [67, 68], our approach by CMARS is a multiplicative one, herewith taking into account "interactions," e.g., any kind of dependence patterns between the input dimensions of time and space, as much as it can be "learned" through our adaptive method of CMARS. In fact, while the additivity of GAMs is understood in terms of separated variables where each input dimension receives its own spline functions, in our CMARS method, the additivity is understood in the multivariate terms of subvectors,

where multiplication of one-dimensional splines goes along the indices of the subvectors' components.

Finally, our adaptive method with elementary basis splines is much more flexible than classical models, from finance and other fields, with their few parameters at their fixed places [32, 65, 71]. Applications of our proposed approach to simulate equity markets are successfully demonstrated. We believe that the presented approach provides an alternative tool for the modeling of SDEs.

The next section will concentrate an extension to the general case of parameter estimation in the multi-dimensional systems' case of SDEs based on our approximation with CMARS, in theory, methods and applications.

4.4 Parameter Identification of Multi-dimensional Stochastic Differential Equations with Conic Multivariate Adaptive Regression Splines

While we have addressed one-dimensional SDEs in previous sections, from now on we focus on a numerical approximation of multi-dimensional SDEs. If a finite number of SDEs are in the model, the multi-dimensional case should be considered. In finance, there are some examples which include this situation, such as modeling the price evolution of multiple stocks. This arises if a decision maker wants to know the properties of some portfolio strategy with multiple assets. Other examples are the modeling of the evolution of interest rates, volatilities or variances. Often, these models occur in short-rate models and in complied systems, respectively. Let us add consumption and portfolio processes, together with wealth processes that imply both, they are addressed in modern portfolio optimization as control and state variables, respectively [21, 47].

Although numerical approximations of multi-dimensional stochastic processes and SDEs are computationally costly problems, they are an important methodology of coping with several sources of randomness and correlation. Both systems of SDEs with standard m-dimensional Brownian motions and systems of SDEs having correlated Brownian motions are considered in the following section.

4.5 Multi-dimensional Stochastic Differential Equations

A system of SDEs is the formulation of a d-dimensional stochastic process $\boldsymbol{X} = (\boldsymbol{X}_t : t \in [0, \infty))$, also denoted by $(\boldsymbol{X}_t)_{t \geq 0}$, in terms of d differential equations comprised:

$$d\boldsymbol{X}_t = \boldsymbol{a}(\boldsymbol{X}_t, t)dt + \boldsymbol{b}(\boldsymbol{X}_t, t)d\boldsymbol{W}_t, \tag{4.12}$$

where $\boldsymbol{a}(\boldsymbol{X}_t, t)$ represents a deterministic part, $\boldsymbol{b}(\boldsymbol{X}_t, t)d\boldsymbol{W}_t$ is a stochastic part, and $\boldsymbol{W}_t$ denotes an m-dimensional Brownian motion (or a Wiener process) at time t. Here, $\boldsymbol{a} : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^d$ and $\boldsymbol{b} : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^{d \times m}$ are measurable vector and matrix

functions, respectively. For the underlying probability space $(\Omega, \mathbb{P}, \mathcal{F})$, e.g., equipped with a filtration $(\mathcal{F}_t)_{t\geq 0}$, we refer to [43, 47].

Let us write the system of SDEs with its components (namely, d processes of state and m-dimensional Brownian motions):

$$\begin{aligned} dX_t^1 &= a^1(\boldsymbol{X}_t, t)dt + b^{11}(\boldsymbol{X}_t, t)dW_t^1 + \ldots + b^{1m}(\boldsymbol{X}_t, t)dW_t^m, \\ dX_t^2 &= a^2(\boldsymbol{X}_t, t)dt + b^{21}(\boldsymbol{X}_t, t)dW_t^1 + \ldots + b^{2m}(\boldsymbol{X}_t, t)dW_t^m, \\ &\vdots \\ dX_t^d &= a^d(\boldsymbol{X}_t, t)dt + b^{d1}(\boldsymbol{X}_t, t)dW_t^1 + \ldots + b^{dm}(\boldsymbol{X}_t, t)dW_t^m. \end{aligned} \tag{4.13}$$

In Eqn. (4.13), we refer to

$$\boldsymbol{X}_t = \begin{bmatrix} X_t^1 \\ X_t^2 \\ \vdots \\ X_t^d(t) \end{bmatrix}, \quad \boldsymbol{a}(\boldsymbol{X}_t, t) = \begin{bmatrix} a^1(\boldsymbol{X}_t, t) \\ a^2(\boldsymbol{X}_t, t) \\ \vdots \\ a^d(\boldsymbol{X}_t, t) \end{bmatrix}, \quad d\boldsymbol{W}_t = \begin{bmatrix} dW_t^1 \\ dW_t^2 \\ \vdots \\ dW_t^m \end{bmatrix},$$

$$\boldsymbol{b}(\boldsymbol{X}_t, t) = \begin{bmatrix} b^{11}(\boldsymbol{X}_t, t) & b^{12}(\boldsymbol{X}_t, t) & \ldots & b^{1m}(\boldsymbol{X}_t, t) \\ b^{21}(\boldsymbol{X}_t, t) & b^{22}(\boldsymbol{X}_t, t) & \ldots & b^{2m}(\boldsymbol{X}_t, t) \\ \vdots & \vdots & \ddots & \vdots \\ b^{d1}(\boldsymbol{X}_t, t) & b^{d2}(\boldsymbol{X}_t, t) & \ldots & b^{dm}(\boldsymbol{X}_t, t) \end{bmatrix}.$$

For further details on multi-dimensional SDEs, we refer the reader to [21, 47].

One of the simplest ways to discretize a multi-dimensional process of SDE is the *Euler method*, which approximates the integrals by using the left-point rule. The *Milstein scheme*, which has the order 1.0 of strong convergence, is stronger than Euler method. The Milstein approximation is identical to the Euler approximation if there is no function of $\boldsymbol{x}$ in the diffusion term of the SDEs. Generally, Milstein method converges to the correct stochastic solution process more quickly than Euler method when the step size goes to zero [29].

Since the distribution of the process is not known, the simulated discretized version of the SDE should be simulated [76]. For multi-dimensional case, the kth component of Milstein Scheme, where $k = 1, 2, ..., d$, is given by

$$\begin{aligned} \hat{x}_{i+1}^k &= \hat{x}_i^k + a^k(\hat{\boldsymbol{x}}_i, t_i)(t_{i+1} - t_i) \\ &\quad + \sum_{j=1}^{m} b^{k,j}(\hat{\boldsymbol{x}}_i, t_i)(\hat{w}_{i+1}^j - \hat{w}_i^j) + \sum_{j_1, j_2 = 1}^{m} L^{k,j_1} b^{k,j_2}(\hat{\boldsymbol{x}}_i, t_i) I_{j_1, j_2}. \end{aligned} \tag{4.14}$$

Here, for $j_1 = j_2$:

$$I_{j_1, j_1} := \frac{1}{2}\left((\hat{w}_{i+1}^{j_1} - \hat{w}_i^{j_1})^2 - (t_{i+1} - t_i)\right),$$

and, for $j_1 \neq j_2$:

$$I_{j_1,j_2} := \frac{1}{2}\left((\hat{w}_{i+1}^{j_1} - \hat{w}_i^{j_1}) - (\hat{w}_{i+1}^{j_2} - \hat{w}_i^{j_2})\right),$$

and we refer to the operator

$$L^{k,j_1} := \sum_{l=1}^{d} b^{k,j_1}(\hat{\boldsymbol{x}}_i, t_i)\frac{\partial}{\partial x^l}.$$

If we rearrange Eqn. (4.14) and insert the data values and also simulated values of the Brownian motions for the equation, the approximation looks as follows [29]:

$$\begin{aligned}\bar{y}_i^k &= a^k(\bar{\boldsymbol{x}}_i, \bar{t}_i) + \frac{1}{\bar{h}_i}\sum_{j=1}^{m} b^{k,j}(\bar{\boldsymbol{x}}_i, \bar{t}_i)(\bar{w}_{i+1}^j - \bar{w}_i^j) \\ &\quad + \frac{1}{\bar{h}_i}\sum_{j_1,j_2=1}^{m} L^{k,j_1} b^{k,j_2}(\bar{\boldsymbol{x}}_i, \bar{t}_i) I_{j_1,j_2}. \end{aligned} \tag{4.15}$$

Here, the value $\bar{y}_i^k$ represents the difference quotient raised on the ith and $(i+1)$st data values $\bar{x}_i^k$ and $\bar{x}_{i+1}^k$, respectively, and on step lengths $\Delta\bar{t}_i := \bar{h}_i = \bar{t}_{i+1} - \bar{t}_i$ between neighbouring sampling times:

$$\bar{y}_i^k = \begin{cases} \frac{1}{\bar{h}_i} \cdot \left(\bar{x}_{i+1}^k - \bar{x}_i^k\right), & \text{if } i = 1, 2, ..., N-1, \\ \frac{1}{\bar{h}_N} \cdot \left(\bar{x}_N^k - \bar{x}_{N-1}^k\right), & \text{if } i = N. \end{cases}$$

Since $W_{t_i}^j \sim N(0,t)$ $(j = 1, 2, ..., m)$, the increments $\Delta W_{t_i}^j$ are independent on non-overlapping intervals which means that $W_{t_2}^j - W_{t_1}^j$, ..., $W_{t_N}^j - W_{t_{N-1}}^j$ are independent with any times $0 < t_1 < t_2 < ... < t_N$. Moreover, $\mathrm{Var}(\Delta W_{t_i}^j) = \Delta t_i$, where we refer to increments $\Delta W_{t_i}^j := W_{t_{i+1}}^j - W_{t_i}^j$ and $\Delta t_i := t_{i+1} - t_i$. With respect to this setting, the increments having normal distribution can be simulated with the help of standard normal distributed random numbers Z_{t_i}. In order to simplify the rest of the equations, we prefer to write W_i^j and Z_i^j instead of $W_{t_i}^j$ and $Z_{t_i}^j$, respectively. Herewith, we obtain a discrete model for a Wiener process [29]:

$$\Delta W_i^j = Z_i^j\sqrt{\Delta t_i}, \quad Z_i^j \sim N(0,1); \tag{4.16}$$

the multivariate distributions and correlations will be referred to in next section. Based on these preparations,

$$\bar{y}_i^k = a^k(\bar{\boldsymbol{x}}_i, \bar{t}_i) + \sum_{j=1}^{m} b^{k,j}(\bar{\boldsymbol{x}}_i, \bar{t}_i)\frac{\bar{Z}_i^j}{\sqrt{\bar{h}_i}} + \frac{1}{\bar{h}_i}\sum_{j_1,j_2=1}^{m} L^{k,j_1} b^{k,j_2}(\bar{\boldsymbol{x}}_i, \bar{t}_i) I_{j_1,j_2}. \tag{4.17}$$

For simplicity, if we choose $m = d = 2$, we obtain:

$$\begin{aligned}
\bar{y}_i^k &= a^k(\bar{\boldsymbol{x}}_i, \bar{t}_i) + b^{k,1}(\bar{\boldsymbol{x}}_i, \bar{t}_i)\frac{\Delta\bar{w}_i^1}{\bar{h}_i} + b^{k,2}(\bar{\boldsymbol{x}}_i, \bar{t}_i)\frac{\Delta\bar{w}_i^2}{\bar{h}_i} \\
&+\frac{1}{2}\left(\frac{\left(\Delta\bar{w}_i^1\right)^2}{\bar{h}_i} - 1\right) b^{k,1}(\bar{\boldsymbol{x}}_i, \bar{t}_i)\left(\sum_{l=1}^{2} b^{k,1}(\bar{\boldsymbol{x}}_i, \bar{t}_i)\frac{\partial}{\partial x^l}\right) \\
&+\frac{1}{2\bar{h}_i}\left(\Delta\bar{w}_i^1 \cdot \Delta\bar{w}_i^2\right) b^{k,2}(\bar{\boldsymbol{x}}_i, \bar{t}_i)\left(\sum_{l=1}^{2} b^{k,1}(\bar{\boldsymbol{x}}_i, \bar{t}_i)\frac{\partial}{\partial x^l}\right) \\
&+\frac{1}{2\bar{h}_i}\left(\Delta\bar{w}_i^2 \cdot \Delta\bar{w}_i^1\right) b^{k,1}(\bar{\boldsymbol{x}}_i, \bar{t}_i)\left(\sum_{l=1}^{2} b^{k,2}(\bar{\boldsymbol{x}}_i, \bar{t}_i)\frac{\partial}{\partial x^l}\right) \\
&+\frac{1}{2}\left(\frac{\left(\Delta\bar{w}_i^2\right)^2}{\bar{h}_i} - 1\right) b^{k,2}(\bar{\boldsymbol{x}}_i, \bar{t}_i)\left(\sum_{l=1}^{2} b^{k,2}(\bar{\boldsymbol{x}}_i, \bar{t}_i)\frac{\partial}{\partial x^l}\right). \qquad (4.18)
\end{aligned}$$

In a more compact form, this can be written as follows:

$$\bar{y}_i^k = \bar{G}_i^k + \bar{H}_i^k + \bar{F}_i^k. \qquad (4.19)$$

Here, $\bar{G}_i^k := a^k(\bar{\boldsymbol{x}}_i, \bar{t}_i)$, $\bar{H}_i^k := \left(1/\bar{h}_i\right) \cdot \sum_{j=1}^{m} b^{k,j}(\bar{\boldsymbol{x}}_i, \bar{t}_i)(\bar{w}_{i+1}^j - \bar{w}_i^j)$ and $\bar{F}_i^k := \left(1/\bar{h}_i\right) \cdot \sum_{j_1,j_2=1}^{m} L^{k,j_1} b^{k,j_2}(\bar{\boldsymbol{x}}_i, \bar{t}_i) I_{j_1,j_2}$.

In order to find the minimum of the difference of the right- and the left-hand side of Eqn. (4.19), the following minimization problem is introduced [77]:

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \sum_{i=1}^{N}\left(\bar{y}_i^k - \left(\bar{G}_i^k + \bar{H}_i^k + \bar{F}_i^k\right)\right)^2, \qquad (4.20)$$

where the vector $\boldsymbol{\theta}$ comprises all the parameters in the Milstein approximation.

In the previous sections, our computationally efficient estimation method called CMARS, was proposed and used for the parameter identification of a one-dimensional SDE. In this section, we aim to estimate multi-dimensional SDEs in a simplified manner by using CMARS method. Therefore, for each function, we use additive form of the CMARS' basis function:

$$\begin{aligned}
\bar{G}_i^k &= a^k(\bar{\boldsymbol{x}}_i, \bar{t}_i) = \beta_0 + \sum_{l=1}^{d^g} \beta_l \mathrm{b}_l(\bar{\boldsymbol{u}}_{i,\mathrm{b}}^l), \\
\bar{H}_i^k &= \frac{1}{\bar{h}_i} \cdot \sum_{j=1}^{m} b^{k,j}(\bar{\boldsymbol{x}}_i, t_i)(\bar{w}_{i+1}^j - \bar{w}_i^j) = \gamma_0 + \sum_{m=1}^{d^h} \gamma_m \mathrm{c}_m(\bar{\boldsymbol{u}}_{i,\mathrm{c}}^m), \\
\bar{F}_i^k &= \frac{1}{\bar{h}_i} \cdot \sum_{j_1,j_2=1}^{m} L^{j_1} b^{k,j_2}(\bar{\boldsymbol{x}}_i, t_i) I_{j_1,j_2} = \delta_0 + \sum_{n=1}^{d^f} \delta_n \mathrm{d}_n(\bar{\boldsymbol{u}}_{i,\mathrm{d}}^n). \qquad (4.21)
\end{aligned}$$

The summation of these three terms in Eqn. (4.21) gives the following expression:

$$\begin{aligned}\bar{G}_i^k + \bar{H}_i^k + \bar{F}_i^k &= \beta_0 + \sum_{l=1}^{d^g} \beta_l \mathrm{b}_l(\bar{\boldsymbol{u}}_{i,\mathrm{b}}^l) + \gamma_0 + \sum_{m=1}^{d^h} \gamma_m \mathrm{c}_m(\bar{\boldsymbol{u}}_{i,\mathrm{c}}^m) + \delta_0 + \sum_{n=1}^{d^f} \delta_n \mathrm{d}_n(\bar{\boldsymbol{u}}_{i,\mathrm{d}}^n) \\ &= \bar{\boldsymbol{A}}_i^k \boldsymbol{\theta}, \end{aligned} \tag{4.22}$$

with $\bar{\boldsymbol{A}}_i^k := \left(\boldsymbol{b}_i^k,\ \boldsymbol{c}_i^k,\ \boldsymbol{d}_i^k\right)$, where $\boldsymbol{b}_i^k := \left(1, \mathrm{b}_1\left(\bar{\boldsymbol{u}}_{i,\mathrm{b}}^l\right), \mathrm{b}_2\left(\bar{\boldsymbol{u}}_{i,\mathrm{b}}^l\right), ..., \mathrm{b}_{d^g}\left(\bar{\boldsymbol{u}}_{i,\mathrm{b}}^l\right)\right)$,
$\boldsymbol{c}_i^k := \left(1, \mathrm{c}_1\left(\bar{\boldsymbol{u}}_{i,\mathrm{c}}^m\right), \mathrm{c}_2\left(\bar{\boldsymbol{u}}_{i,\mathrm{c}}^m\right), ..., \mathrm{c}_{d^h}\left(\bar{\boldsymbol{u}}_{i,\mathrm{c}}^m\right)\right)$,
$\boldsymbol{d}_i^k := \left(1, \mathrm{d}_1\left(\bar{\boldsymbol{u}}_{i,\mathrm{d}}^n\right), \mathrm{d}_2\left(\bar{\boldsymbol{u}}_{i,\mathrm{d}}^n\right), ..., \mathrm{d}_{d^f}\left(\bar{\boldsymbol{u}}_{i,\mathrm{d}}^n\right)\right)$, and $\boldsymbol{\theta}^k := \left(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\delta}^T\right)^T$, where
$\boldsymbol{\beta}^k := (\beta_0, \beta_1, \beta_2, ..., \beta_{d^h})^T$, $\boldsymbol{\gamma}^k := (\gamma_0, \gamma_1, \gamma_2, ..., \gamma_{d^f})^T$, $\boldsymbol{\delta}^k := (\delta_0, \delta_1, \delta_2, ..., \delta_{d^g})^T$.

Then, we can represent the summation term in Eqn. (4.20) as follows:

$$\sum_{i=1}^{N} (\bar{y}_i^k - \bar{\boldsymbol{A}}_i^k \boldsymbol{\theta}^k)^2 = \left\| \bar{\mathbf{y}}^k - \bar{\boldsymbol{A}}^k \boldsymbol{\theta}^k \right\|_2^2.$$

The matrix notation of two-dimensional case is given below:

$$\begin{bmatrix} \bar{y}_1^1 \\ \bar{y}_2^1 \\ \vdots \\ \bar{y}_N^1 \\ \bar{y}_{N+1}^2 \\ \vdots \\ \bar{y}_{2N}^2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{N\times((1+3m)\cdot M_{max})}^1 & \mathbf{0}_{N\times((1+3m)\cdot M_{max})}^2 \\ \mathbf{0}_{N\times((1+3m)\cdot M_{max})}^1 & \boldsymbol{A}_{N\times((1+3m)\cdot M_{max})}^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_{((1+3m)\cdot M_{max})\times 1}^1 \\ \boldsymbol{\theta}_{((1+3m)\cdot M_{max})\times 1}^2 \end{bmatrix},$$

where the dimensions of $\bar{\mathbf{y}}$ is $2N \times 1$, of $\bar{\boldsymbol{A}}$ is $2N \times ((2+6m)\cdot 2M_{max})$, and of $\boldsymbol{\theta}$ is $((2+6m)\cdot 2M_{max}) \times 1$. To approximate the multi-dimensional integral, the same procedure as described in Section 4.3 is followed for each dimensions of multi-dimensional SDEs. Firstly, PRSS is constructed for a regularization as by Eqn. (4.10):

$$PRSS \approx \left\| \bar{\mathbf{y}}^k - \bar{\boldsymbol{A}}^k \boldsymbol{\theta}^k \right\|_2^2 + \sum_{l=1}^{d^g} \lambda_l \left\| \bar{\boldsymbol{L}}_l^{\mathrm{b},k} \beta_l \right\|_2^2 + \sum_{m=1}^{d^h} \mu_m \left\| \bar{\boldsymbol{L}}_m^{\mathrm{c},k} \gamma_m \right\|_2^2 + \sum_{n=1}^{d^f} \nu_n \left\| \bar{\boldsymbol{L}}_n^{\mathrm{d},k} \delta_n \right\|_2^2,$$

where $(l = 1, 2, ..., d^g)$ $\bar{\boldsymbol{L}}_l^{\mathrm{b},k} := \left(\bar{L}_{1l}^{\mathrm{b},k}, \bar{L}_{2l}^{\mathrm{b},k}, ..., \bar{L}_{(N+1)^2 l}^{\mathrm{b},k}\right)^T$, $\left(l = 1, 2, ..., d^h\right)$ $\bar{\boldsymbol{L}}_m^{\mathrm{c},k} := \left(\bar{L}_{1m}^{\mathrm{c},k}, \bar{L}_{2m}^{\mathrm{c},k}, ..., \bar{L}_{(N+1)^2 m}^{\mathrm{c},k}\right)^T$ and $\left(l = 1, 2, ..., d^f\right)$ $\bar{\boldsymbol{L}}_n^{\mathrm{d},k} := \left(\bar{L}_{1n}^{\mathrm{d},k}, \bar{L}_{2n}^{\mathrm{d},k}, ..., \bar{L}_{(N+1)^2 n}^{\mathrm{d},k}\right)^T$.
Here, $M_{max} = d^g + d^h + d^f$ is the number of penalty parameters. After turning to a uniform penalization via a single factor λ for each derivative term, the approximation of PRSS becomes a task of Tikhonov Regularization. Now, we refer to a minimization of the following right-hand side:

$$PRSS \approx \left\| \bar{\mathbf{y}}^k - \bar{\boldsymbol{A}}^k \boldsymbol{\theta}^k \right\|_2^2 + \lambda \left\| \bar{\boldsymbol{L}}^k \boldsymbol{\theta}^k \right\|_2^2,$$

where $\bar{\boldsymbol{L}}^k$ is a diagonal $((M_{max}+1)\times(M_{max}+1))$-matrix with first column $\bar{\boldsymbol{L}}_0^k = \mathbf{0}_{(N+1)^2}$. Then, the Tikhonov Regularization problem is formulated as a CQP problem as in Eqn. (4.11):

$$\begin{aligned}
&\underset{t,\boldsymbol{\theta}}{\text{minimize}}\ \ t,\\
&\text{such that}\quad \boldsymbol{\chi} := \begin{bmatrix} \mathbf{0}_N & \bar{\boldsymbol{A}}^k \\ 1 & \mathbf{0}_{M_{max}+1}^T \end{bmatrix}\begin{bmatrix} t \\ \boldsymbol{\theta} \end{bmatrix} + \begin{bmatrix} -\bar{\mathbf{y}}^k \\ 0 \end{bmatrix},\\
&\qquad\qquad \boldsymbol{\eta} := \begin{bmatrix} \mathbf{0}_{M_{max}+1} & \bar{\boldsymbol{L}}^k \\ 0 & \mathbf{0}_{M_{max}+1}^T \end{bmatrix}\begin{bmatrix} t \\ \boldsymbol{\theta} \end{bmatrix} + \begin{bmatrix} \mathbf{0}_{M_{max}+1} \\ \sqrt{\tilde{M}} \end{bmatrix},\\
&\qquad\qquad \boldsymbol{\chi} \in L^{N+1},\ \boldsymbol{\eta} \in L^{M_{max}+2}.
\end{aligned} \tag{4.23}$$

After these steps, the unknown parameters of the SDEs' components are obtained by solving the CQP problem.

4.5.1 Numerical Example of Multi-dimensional Stochastic Differential Equations

As a numerical example of the given procedure, the simulating equity markets described in Section 4.3 are used. These data are preferred in order to examine the differences between one- and multi-dimensional SDEs.

The data sets are demonstrated in Figure 4.3.

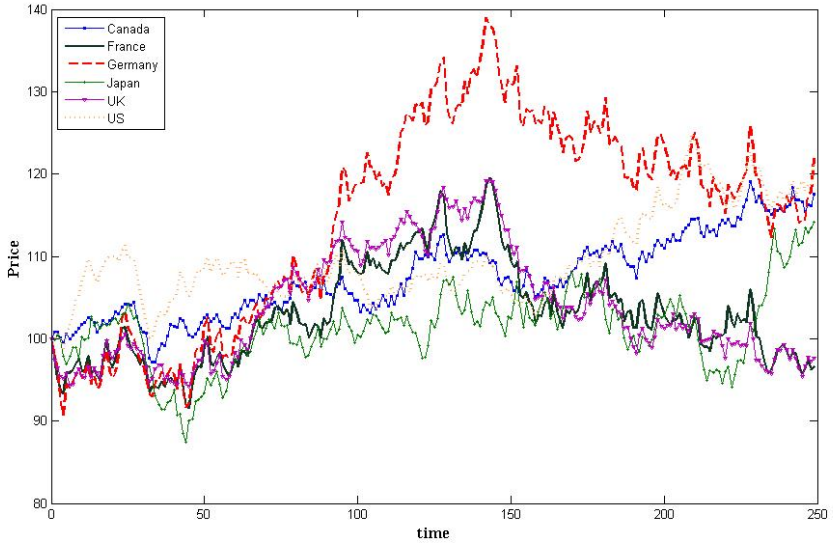

Figure 4.3: Demonstration of the multi-dimensional market model data sets.

For the numerical example, UK and US prices are selected again, as in Section 4.4. The discretization and approximation approaches are built for each process. The comparison measures listed in Table 4.1 are calculated. For the stochastic processes used in this study, our approximation reveals a good performance according to these measures, as shown in Table 4.2. The smaller values for MAE and MSE indicate a strong estimation of the process parameters. On the other hand, the further measures (R^2, Adj-R^2, PWI) indicate a better performance if their values are closer to 1.0. If we compare these results with the previous results given in Table 4.1, the approximation is improved for both processes.

Table 4.2: Evaluation of proposed approximation with well-known statistical measures.

Performance Measures	**US Price**	**UK Price**
R^2	0.9921	0.9854
Adj-R^2	0.9921	0.9853
MAE	0.3462	0.4692
MSE	0.2498	0.5341
PWI	0.9840	0.9800

Considering that there is not a correlation between these stochastic processes, the proposed approximation gives a promising result and captures the structures of both US and UK prices. In order to see the two processes in a single figure synoptically, Figure 4.4 is provided.

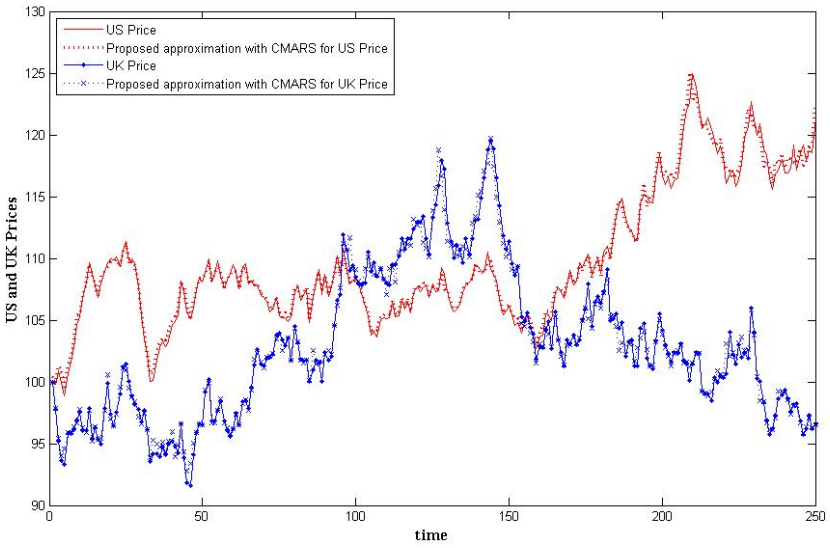

Figure 4.4: Demonstration of proposed approximation with CMARS for US price versus UK price.

4.6 Multi-dimensional Stochastic Differential Equations for Correlated Brownian Motions

One of the main concerns of this section is to transform the referred system of SDEs with correlated Brownian motions to a one having uncorrelated Wiener processes and, then, to apply our proposed approach given in Section 4.3 to the transformed systems.

Many practical situations and scientific configurations that depend on a variety of factors can be modeled as a stochastic process that is driven by a multi-dimensional Wiener processes. These random factors may be independent, but in reality, there is often correlation between them. Usually, the statistical correlation ρ of two random variables X_1 and X_2 is defined as

$$\rho\left(X_1, X_2\right) = \frac{\operatorname{cov}\left(X_1, X_2\right)}{\sqrt{\operatorname{Var}\left(X_1\right)}\sqrt{\operatorname{Var}\left(X_2\right)}}.$$

Note that if $\rho\left(X_1, X_2\right) = 0$, X_1 and X_2 are uncorrelated; otherwise, there is a correlation between X_1 and X_2.

The correlation between W_t^i and W_t^j is represented by the (constant in time) *correlation coefficients* ρ_{ij} $(i, j = 1, 2, ..., m)$, at any regarded time t [63]. For correlated multi-dimensional Wiener processes, the following correlation matrix is introduced:

$$\boldsymbol{R} := \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1d} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{d1} & \rho_{d2} & \cdots & \rho_{dd} \end{bmatrix},$$

where $-1 < \rho_{ij} < 1$ for all i, j with $i \neq j$ and $\rho_{ii} = 1$. The matrix $\boldsymbol{R}$ is a *symmetric* and *positive definite* $d \times d$-matrix, e.g., $\rho_{ij} = \rho_{ji}$ $(i, j = 1, 2, ..., d)$ and $\boldsymbol{x}^T\boldsymbol{R}\boldsymbol{x} > 0$ for all $\boldsymbol{x} \in \mathbb{R}^m \setminus \{\boldsymbol{0}\}$. This implies $\det\left(\boldsymbol{R}\right) \neq 0$. To create noise processes with a specified correlation, we can use *Cholesky decomposition* as a fast algorithm through an upper (or lower) triangular matrix $\boldsymbol{\Gamma} = \left(\Gamma_{ij}\right)_{1 \leq i,j \leq m}$ [63]. The Cholesky decomposition of $\boldsymbol{R}$ is:

$$\boldsymbol{R} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T,$$

where $\boldsymbol{\Gamma}^T$ is the transposed matrix of $\boldsymbol{\Gamma}$. We take $\boldsymbol{Z} = \left(\boldsymbol{Z}_t\right)_{t \geq 0}$, $\boldsymbol{Z}_t = \left(Z_t^1, Z_t^2, \ldots, Z_t^m\right)^T$, as a standard m-dimensional Brownian motion whose components are independent at all times t and we define a correlated vector-valued process $\boldsymbol{W} = \left(\boldsymbol{W}_t\right)_{t \geq 0}$, $\boldsymbol{W}_t = \left(W_t^1, W_t^2, \ldots, W_t^m\right)^T$, as follows:

$$\boldsymbol{W}_t := \boldsymbol{\Gamma}\boldsymbol{Z}_t,$$

or, in terms of the components:

$$W_t^i = \sum_{j=1}^{m} \Gamma_{ij} Z_t^j \quad (i = 1, 2, ..., m).$$

The *multi-dimensional correlated Brownian motion*, $(\boldsymbol{W}_t)_{t\geq 0}$ of dimension m, with its correlation matrix $\boldsymbol{R}$ has the following properties:

- $\boldsymbol{W}_0 = \mathbf{0}$, the zero vector in $\mathbb{R}^m$.
- If $s < t$, then the increment $\boldsymbol{W}_t - \boldsymbol{W}_s$ is multivariate normal distributed, with mean $\mathbf{0}$ and variance-covariance matrix $(t-s)\cdot\boldsymbol{R}$:

$$\boldsymbol{W}_t - \boldsymbol{W}_s \sim N\left(\mathbf{0}, (t-s)\cdot\boldsymbol{R}\right).$$

- If $0 \leq r < s < t$, then the random variables $\boldsymbol{W}_t - \boldsymbol{W}_s$ and $\boldsymbol{W}_s - \boldsymbol{W}_r$ are independent. This means that each component of the former is independent of each component of later.
- The paths $t \longmapsto \boldsymbol{W}_t$ are continuous with probability 1.

If $\boldsymbol{R} = \boldsymbol{I}$, the correlation matrix is an identity matrix, then $(\boldsymbol{W}_t)_{t\geq 0}$ is a standard Brownian motion. Thus, for such a case, independence means having zero correlation.

Let us mention that it is possible to write the system of SDEs for correlated Brownian motion [43]:

$$\begin{aligned}
dX_t^1 &= a^1(\boldsymbol{X}_t,t)dt + b^{11}(\boldsymbol{X}_t,t)\Gamma_{11}dZ_t^1 + \ldots + b^{1m}(\boldsymbol{X}_t,t)\Gamma_{1m}dZ_t^m,\\
dX_t^2 &= a^2(\boldsymbol{X}_t,t)dt + b^{21}(\boldsymbol{X}_t,t)\Gamma_{21}dZ_t^1 + \ldots + b^{2m}(\boldsymbol{X}_t,t)\Gamma_{2m}dZ_t^m,\\
&\vdots\\
dX_t^d &= a^d(\boldsymbol{X}_t,t)dt + b^{d1}(\boldsymbol{X}_t,t)\Gamma_{d1}dZ_t^1 + \ldots + b^{dm}(\boldsymbol{X}_t,t)\Gamma_{dm}dZ_t^m. \quad (4.24)
\end{aligned}$$

For simplicity, we continue with the two-dimensional case. Then the system turns into the following form:

$$\begin{aligned}
dX_t^1 &= a^1(\boldsymbol{X}_t,t)dt + b^{11}(\boldsymbol{X}_t,t)\Gamma_{11}dZ_t^1 + b^{12}(\boldsymbol{X}_t,t)\Gamma_{12}dZ_t^2,\\
dX_t^2 &= a^2(\boldsymbol{X}_t,t)dt + b^{21}(\boldsymbol{X}_t,t)\Gamma_{21}dZ_t^1 + b^{22}(\boldsymbol{X}_t,t)\Gamma_{22}dZ_t^2. \quad (4.25)
\end{aligned}$$

Here, the correlation matrix $\boldsymbol{R}$ can be written as:

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

since $\boldsymbol{R}$ is a symetric matrix, we take $\rho_{12} = \rho_{21} = \rho$. By Cholesky decomposition, the matrix $\boldsymbol{R}$ can be written as follows:

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} 1 & \rho \\ 0 & \sqrt{1-\rho^2} \end{bmatrix}.$$

Since

$$\boldsymbol{W} = \boldsymbol{\Gamma Z} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} Z^1 \\ Z^2 \end{bmatrix},$$

Eqn. (4.25) becomes

$$\begin{aligned} dX_t^1 &= a^1(\boldsymbol{X}_t,t)dt + b^{11}(\boldsymbol{X}_t,t)dZ_t^1, \\ dX_t^2 &= a^2(\boldsymbol{X}_t,t)dt + b^{21}(\boldsymbol{X}_t,t)\rho dZ_t^1 + b^{22}(\boldsymbol{X}_t,t)\sqrt{1-\rho^2}dZ_t^2. \end{aligned} \tag{4.26}$$

For the two-dimensional system of SDEs given in Eqn. (4.26), the first component of Milstein Scheme is as follows:

$$\begin{aligned} \hat{x}_{i+1}^1 &= \hat{x}_i^1 + a^1(\hat{\boldsymbol{x}}_i,t_i)(t_{i+1}-t_i) \\ &\quad + \sum_{j=1}^{2} b^{1,j}(\hat{\boldsymbol{x}}_i,t_i)(\hat{z}_{i+1}^j - \hat{z}_i^j) + \sum_{j_1,j_2=1}^{2} L^{1,j_1}b^{1,j_2}(\hat{\boldsymbol{x}}_i,t_i)I_{j_1,j_2}. \end{aligned}$$

The first component of the system of SDEs has just the term $b^{11}(\hat{\boldsymbol{x}}_i,t_i)$, Milstein approximation gets the following form:

$$\begin{aligned} \hat{x}_{i+1}^1 &= \hat{x}_i^1 + a^1(\hat{\boldsymbol{x}}_i,t_i)(t_{i+1}-t_i) \\ &\quad + b^{1,1}(\hat{\boldsymbol{x}}_i,t_i)(\hat{z}_{i+1}^1 - \hat{z}_i^1) + L^{1,1}b^{1,1}(\hat{\boldsymbol{x}}_i,t_i)I_{1,1}. \end{aligned} \tag{4.27}$$

After rearranging Eqn. (4.27), the approximation becomes

$$\begin{aligned} \bar{y}_i^1 &= a^1(\bar{\boldsymbol{x}}_i,\bar{t}_i) \\ &\quad + \frac{1}{\bar{h}_i} b^{1,1}(\bar{\boldsymbol{x}}_i,\bar{t}_i)(\bar{z}_{i+1}^1 - \bar{z}_i^1) + \frac{1}{\bar{h}_i} L^{1,1}b^{1,1}(\bar{\boldsymbol{x}}_i,\bar{t}_i)I_{1,1}. \end{aligned}$$

If we replace $L^{1,1}$ and $I_{1,1}$ with their open forms, we get subsequent equation:

$$\bar{y}_i^1 = a^1(\bar{\boldsymbol{x}}_i,\bar{t}_i) + b^{1,1}(\bar{\boldsymbol{x}}_i,\bar{t}_i)\frac{\Delta\bar{z}_i^1}{\bar{h}_i} + \frac{1}{2}\left(\frac{(\Delta\bar{z}_i^1)^2}{\bar{h}_i} - 1\right) b^{1,1}(\bar{\boldsymbol{x}}_i,\bar{t}_i)\left(\sum_{l=1}^{2} b^{1,1}(\bar{\boldsymbol{x}}_i,\bar{t}_i)\frac{\partial}{\partial x^1}\right).$$

On the other hand, the second component of Milstein Scheme for the given SDEs is as follows:

$$\begin{aligned} \hat{x}_{i+1}^2 &= \hat{x}_i^2 + a^2(\hat{\boldsymbol{x}}_i,t_i)(t_{i+1}-t_i) \\ &\quad + \sum_{j=1}^{2} b^{2,j}(\hat{\boldsymbol{x}}_i,t_i)(\hat{z}_{i+1}^j - \hat{z}_i^j) + \sum_{j_1,j_2=1}^{2} L^{2,j_1}b^{2,j_2}(\hat{\boldsymbol{x}}_i,t_i)I_{j_1,j_2}. \end{aligned} \tag{4.28}$$

Equivalently to Eqn. (4.28),

$$\begin{aligned} \bar{y}_i^2 &= a^2(\bar{\boldsymbol{x}}_i,\bar{t}_i) + \rho\left(1/\bar{h}_i\right) b^{2,1}(\bar{\boldsymbol{x}}_i,\bar{t}_i)(\bar{z}_{i+1}^1 - \bar{z}_i^1) \\ &\quad + \left(\sqrt{1-\rho^2}\right)\left(1/\bar{h}_i\right) b^{2,2}(\bar{\boldsymbol{x}}_i,\bar{t}_i)(\bar{z}_{i+1}^2 - \bar{z}_i^2) \\ &\quad + \rho\left(1/\bar{h}_i\right) L^{2,1}b^{2,1}(\bar{\boldsymbol{x}}_i,\bar{t}_i)I_{1,1} + \left(\sqrt{1-\rho^2}\right)\left(1/\bar{h}_i\right) L^{2,1}b^{2,2}(\bar{\boldsymbol{x}}_i,\bar{t}_i)I_{1,2} \\ &\quad + \rho\left(1/\bar{h}_i\right) L^{2,2}b^{2,1}(\bar{\boldsymbol{x}}_i,\bar{t}_i)I_{2,1} + \left(\sqrt{1-\rho^2}\right)\left(1/\bar{h}_i\right) L^{2,2}b^{2,2}(\bar{\boldsymbol{x}}_i,\bar{t}_i)I_{2,2}. \end{aligned}$$

If we replace $L^{2,1}$, $L^{2,2}$ and $I_{1,1}$, $I_{1,2}$, $I_{2,1}$, $I_{2,2}$ with their open forms, we get the equation below:

$$\begin{aligned}
\bar{y}_i^2 \;=\; & a^2(\bar{\boldsymbol{x}}_i,\bar{t}_i)+\rho b^{2,1}(\bar{\boldsymbol{x}}_i,\bar{t}_i)\frac{\Delta\bar{z}_i^1}{\bar{h}_i}+\left(\sqrt{1-\rho^2}\right)b^{2,2}(\bar{\boldsymbol{x}}_i,\bar{t}_i)\frac{\Delta\bar{z}_i^2}{\bar{h}_i}\\
&+\rho^2\left(\frac{1}{2}\left(\frac{\left(\Delta\bar{z}_i^1\right)^2}{\bar{h}_i}-1\right)b^{2,1}(\bar{\boldsymbol{x}}_i,\bar{t}_i)\left(\sum_{l=1}^{2}b^{2,1}(\bar{\boldsymbol{x}}_i,\bar{t}_i)\frac{\partial}{\partial x^l}\right)\right)\\
&+\rho\left(\sqrt{1-\rho^2}\right)\left(\frac{1}{2\bar{h}_i}\left(\Delta\bar{z}_i^1\cdot\Delta\bar{z}_i^2\right)b^{2,2}(\bar{\boldsymbol{x}}_i,\bar{t}_i)\left(\sum_{l=1}^{2}b^{2,1}(\bar{\boldsymbol{x}}_i,\bar{t}_i)\frac{\partial}{\partial x^l}\right)\right)\\
&+\rho\left(\sqrt{1-\rho^2}\right)\left(\frac{1}{2\bar{h}_i}\left(\Delta\bar{z}_i^2\cdot\Delta\bar{z}_i^1\right)b^{2,1}(\bar{\boldsymbol{x}}_i,\bar{t}_i)\left(\sum_{l=1}^{2}b^{2,2}(\bar{\boldsymbol{x}}_i,\bar{t}_i)\frac{\partial}{\partial x^l}\right)\right)\\
&+\left(\sqrt{1-\rho^2}\right)^2\left(\frac{1}{2}\left(\frac{\left(\Delta\bar{z}_i^2\right)^2}{\bar{h}_i}-1\right)b^{2,2}(\bar{\boldsymbol{x}}_i,\bar{t}_i)\left(\sum_{l=1}^{2}b^{2,2}(\bar{\boldsymbol{x}}_i,\bar{t}_i)\frac{\partial}{\partial x^l}\right)\right).
\end{aligned}$$

Then, we apply the same procedure as presented in Section 4.3. In order to avoid repetition, we continue with the numerical example of the multi-dimensional SDEs with correlated Brownian motions.

4.6.1 Numerical Example of Multi-dimensional Stochastic Differential Equations for Correlated Brownian Motions

This subsection includes the implementation of the proposed methodology to multi-dimensional stochastic processes inducing dependence and correlation. In order to simulate the data set, the nested functions for simulating equity markets in MATLAB are used. The simulated data set induces 1358 observations, i.e., roughly more than four years of daily data.

The daily prices of six countries (Canada, France, Germany, Japan, UK and US) are converted to returns. For simplicity, two simulated processes, France and UK, are selected for the application of this part. The reason to select France and UK is having a stronger relationship than the other equity index prices. Thus, we can predict one of the process with the other process more accurately. The estimated correlation matrix for France and UK is as follows:

$$\boldsymbol{R}_{France,\,UK}=\begin{bmatrix}1 & 0.8575\\ 0.8575 & 1\end{bmatrix}.$$

The value 0.8575 shows that there is a strong correlation between those two processes. The lower and upper Cholesky factors of the given correlation matrix is:

$$\boldsymbol{R}_{France,\,UK}=\begin{bmatrix}1 & 0\\ 0.8575 & 0.5145\end{bmatrix}\begin{bmatrix}1 & 0.8575\\ 0 & 0.5145\end{bmatrix}.$$

After this we construct CMARS basis functions for each generated process. For France, the following basis functions are obtained:

$$\begin{aligned}
\psi_1 &= \max\{0, X_{France} - 0.0329\}, & \psi_2 &= \max\{0, 0.0329 - X_{France}\}, \\
\psi_3 &= \max\{0, X_{UK} - 0.0355\}, & \psi_4 &= \max\{0, 0.0355 - X_{UK}\}, \\
\psi_5 &= \max\{0, X_{UK} - 0.0190\}, & \psi_6 &= \max\{0, 0.0190 - X_{UK}\}, \\
\psi_7 &= \max\{0, X_{France} - 0.0345\}, & \psi_8 &= \max\{0, 0.0345 - X_{France}\}, \\
\psi_9 &= \max\{0, X_{UK} - 0.0250\}, & \psi_{10} &= \max\{0, 0.0250 - X_{UK}\}.
\end{aligned}$$

For UK, the basis functions are as follows:

$$\begin{aligned}
\psi_1 &= \max\{0, X_{UK} - 0.0280\}, & \psi_2 &= \max\{0, 0.0280 - X_{UK}\}, \\
\psi_3 &= \max\{0, X_{France} + 0.0768\}, & \psi_4 &= \max\{0, X_{UK} - 0.0109\}, \\
\psi_5 &= \max\{0, 0.0109 - X_{UK}\}, & \psi_6 &= \max\{0, X_{UK} - 0.0318\}, \\
\psi_7 &= \max\{0, 0.0318 - X_{UK}\}, & \psi_8 &= \max\{0, X_{France} - 0.0049\}, \\
\psi_9 &= \max\{0, 0.0049 - X_{France}\}, & \psi_{10} &= \max\{0, X_{UK} + 0.0559\}.
\end{aligned}$$

There are 10 basis functions with main effects for each one of our processes. For this two-dimensional SDE system, we generate parameters of the SDEs to find the best fit.

Finally, the performances of the proposed approximation with CMARS are calculated according to well-known performance measures described in Appendix A. The results of this application are summarized in Table 4.3.

Table 4.3: Evaluation of proposed approximation with well-known statistical measures.

Performance Measures	**France Return**	**UK Return**
R^2	0.5324	0.5659
Adj-R^2	0.5203	0.5398
MAE	0.0101	0.0077
MSE	0.0002	0.0001
PWI	0.9853	0.9801

By this numerical example, we are able to identify the parameters of the given correlated stochastic processes with the help of our data mining method. This advance gives knowledge and insight about stochastic processes.

The main concern of modeling stochastic processes is how to determine data values and to forecast *future* values of the process based on available information *in-between*. The special feature of stochastic processes such as stock prices or exchange rates are changing over time. Furthermore, there may be several missing data or not available data due to many reasons such as regular holidays for stock prices.

If already *forecasted*, *anticipated*, "*ideal*" or "*desired*" future data are presumed [1], then our methodology can be applied over such a future horizon. However, in this

regard, future research will need to contribute to stochastic foundations of it in terms of martingales and related concepts [55].

Nonparametric methods are very promising to estimate the parameters of stochastic processes, because the functional forms of the processes are obtained without any assumptions. For this reason, CMARS model is suggested to explain volatility or variance in stochastic process, and it has been successfully applied to finance data. By CMARS technique and its functional form, not only the structure of the stochastic process is specified fully, but also future values of the process can be estimated from them, in the aforementioned sense and in future research.

CHAPTER 5

PARAMETER IDENTIFICATION AND ESTIMATION OF HURST PARAMETER OF STOCHASTIC DIFFERENTIAL EQUATIONS DRIVEN BY FRACTIONAL BROWNIAN MOTIONS USING CONIC MULTIVARIATE ADAPTIVE REGRESSION SPLINES

5.1 Introduction

Fractional Brownian motions (fBm) is a widely used concept for modeling various situations such as the level of water in a river, the temperature at a specific place, empirical volatility of a stock, the price dynamics of electricity. That utilization appears naturally in those phenomena because of its capability of explaining the *dependence* structure in real-life observations. A main purpose of introducing the concept of an fBm, $\left(W_t^H\right)_{t\geq 0}$, lies in a notion of random fluctuation of a time-continuous stochastic processes, $(X_t)_{t\geq 0}$, which is wider than that given by a Brownian motion $(W_t)_{t\geq 0}$. In fact, the connection between $(X_t)_{t\geq 0}$ and $(W_t)_{t\geq 0}$ or, in our case, $\left(W_t^H\right)_{t\geq 0}$, is implied by our model which is a SDE, where the (fractional) Brownian motion is a key component in the second, actually, random or diffusion term. Since the (fractional) Brownian motion fulfills certain axioms, it can be regarded as a formatted or normalized random fluctuation; moreover, fBm is a continuous zero-mean Gaussian process with stationary increments. Therefore, in the SDE a factor occurs in front of the differential (fractional) Brownian term; that factor plays the role of volatility. The fBm is characterized by a parameter, the so-called Hurst parameter H. A fBm with Hurst parameter $H > 1/2$ is called a *persistent* process, i.e., the increments of this process are positively correlated. On the other hand, the increments of fBm with $H < 1/2$ are called an *anti-persistent* process with increments being negatively correlated. For $H = 1/2$, fBm corresponds to a Brownian motion which has independent increments. The following references are for more information on fBm and its applications [36, 43, 45]. We will estimate the Hurst parameter together with the model coefficients that will be linearly involved in the representation of the entire SDE, to be more precise: in the time-discrete approximation which we will study. This estimation will be based on given data and supported by modern optimization techniques [75]. In this chapter, the aforementioned items will become represented more closely.

It is highly important to identify the value of a Hurst parameter in order to understand the structure of the process and its applications since the calculations dramatically differ according to the value of H. Therefore, some techniques have been developed to estimate Hurst parameter which can be categorized into three groups; heuristics, maximum-likelihood and wavelet-based estimators. In the group of heuristics estimators, there is the so-called R/S estimator which was firstly proposed by Hurst [24], followed by the methods of correlogram, variogram, variance plot, and partial correlations plot. Due to lack of accuracy of heuristics estimators, *Maximum-Likelihood (ML) estimators* were developed. Being weakly consistent is the main disadvantage of ML estimators. In parallel to ML estimators, wavelet-based estimators were suggested because of the popularity of wavelet decomposition of fBm [10, 54].

In search of faster and efficient ways to estimate the Hurst parameter H, we suggest a new numerical and computational method [75].

This chapter is organized as follows: In Section 5.2, we start with explaining the properties of our madel given as SDEs driven by fBm. In Section 5.3, we introduce the method CMARS relating it to the Hurst parameter estimation of our model. In Section 5.4, we give an application of our study, in order to test the theory which we have developed. Finally, we present a brief conclusion and a general outlook based on our study.

5.2 Stochastic Differential Equations with Fractional Brownian Motions

SDEs generated by fBm are widely used to represent noisy and real-world problems. They play an important role in many fields of science such as finance, physics, biotechnology and engineering. In this section, we briefly recall some concepts of fBm and stochastic differential equations driven by fBm.

5.2.1 Fractional Brownian Motions

Let H be a constant in the interval $(0, 1)$. An fBm with Hurst parameter H, $W^H := (W_t^H)_{t\geq 0}$, is a continuous and centered Gaussian process with covariance function

$$\mathrm{E}[W_t^H W_s^H] = \frac{1}{2}(t^{2H} + s^{2H} - |t-s|^{2H}).$$

We note that, for $H = 1/2$, fBm corresponds to a standard Brownian motion which has independent increments. For a *standard fBm*, W^H:

- $W_0^H = 0$ and $E[W_t^H] = 0$ for all $t \geq 0$.
- W^H has homogenous increments, i.e., $W_{t+s}^H - W_s^H$ has the same law as W_t^H, for all $s, t \geq 0$.
- W^H is a Gaussian process and $E[(W_t^H)^2] = t^{2H}$ $(t \geq 0)$, for all $H \in (0, 1)$.

- W^H has continuous trajectories.

The Hurst parameter H of fBm explains the dependency of data [12, 43, 45]. Indeed, the correlation between increments for $s, t \geq 0$ can be obtained by

$$\mathrm{E}[(W^H_{t+h}) - W^H_t)(W^H_{s+h} - W^H_s)] = \tfrac{h^{2H}}{2}[(n+1)^{2H} + (n-1)^{2H} - 2n^{2H}].$$

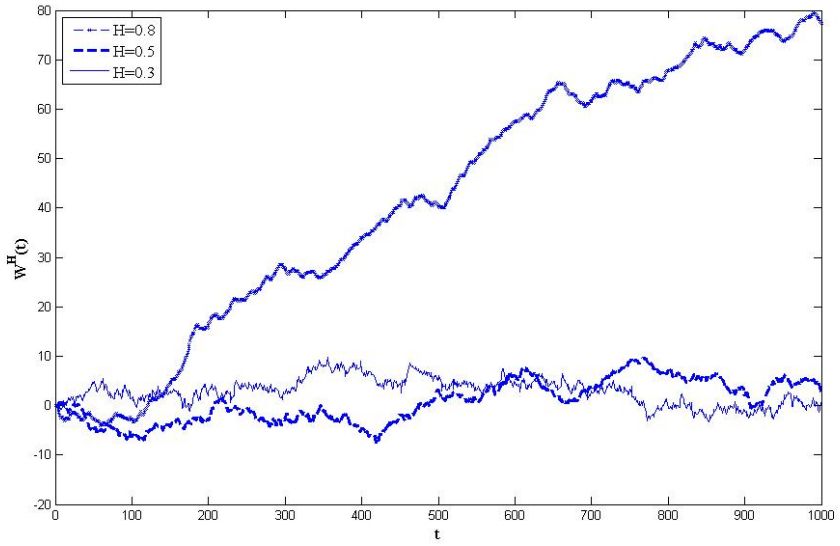

Figure 5.1: Sample paths of fBm with Hurst parameter values, H=0.3, 0.5, 0.8, simulated by Cholesky method.

Observations with $H > 1/2$ have positively correlated increments and display long-range dependence, while the observations with $H < 1/2$ have a negatively correlated increments and display a short-range dependence structure. Figure 5.1 shows the paths of fBm with different Hurst parameter values. While the Brownian motion case (i.e., $H = 1/2$) represents a "formatted" or normalized rondom fluctuation, with a conic corridor of variance, given by time t itself, the cases $H > 1/2$ ($H < 1/2$) mean some kind of self-supporting and self-sustaining (or self-degrading, respectively) of the fBm process in time t. Therefore, it is crucial to find the Hurst parameter of a stochastic process for understanding various phenomena in diverse fields from engineering to finance. For example, it is observed that the prices of electricity in a liberated electricity market have spikes which can be regarded as negatively correlated increments. This phenomenon can be modeled by a SDE driven by a fBm with $H < 1/2$. On the other hand, in financial markets, the prices of stocks usually display a long-range dependence which can be explained by an SDE driven by a fBm with $H > 1/2$. In this chapter, we concentrate on finding H for the stochastic processes which are the strong solutions of SDEs with fBm. Hence, we first recall some fundamental properties of them.

5.2.2 Stochastic Differential Equations Driven by Fractional Brownian Motions

Suppose we have a stochastic process $X = (X_t : t \geq 0)$, or $(X_t)_{t\geq 0}$, defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ which is the strong solution of the following SDE:

$$dX_t = a(t, X_t)dt + b(t, X_t)dW_t^H. \tag{5.1}$$

Here, $a(t, X_t)$ and $b(t, X_t)$ represent the drift and diffusion terms satisfying the conditions of existence and uniqueness theorem for $t \geq 0$. Furthermore, the distribution of the variable X is unknown because it depends on the functional form of the drift and diffusion terms. Later, in Section 5.3.1, this variable is estimated by discretization of SDE. For some classes of such SDEs, where the functions $a(t, X_t)$ and $b(t, X_t)$ obey a certain, mostly easy analytical or model representation, the distribution of $(X_t)_{t\geq 0}$ is known, especially, by explicit solution formulas existing. In the possibly most famous cases of so-called geometric or arithmetic Brownian motion with $H = 1/2$, standing for price or logarithmic price processes, solution formulas of initial value (or terminal-value) problems of Eqn. (5.1) are well-know, together with the probability distribution of the variable X_t. That distribution is usually expressed through its mean and variance; in other cases, one knows moments (of some order) of the distribution or data-based approximations of the moments [43]. We note that it is necessary to have the integrator as a semi-martingale in the theory of stochastic integration. However, since fBm is not a semi-martingale, one should extend the usual settings as in the definition of the Itô integral and define the integration with respect to fBm in a new pathwise integration technique. Alos et al. [2] construct the theory of integration with respect to general Gaussian proceses to overcome this obstacle [36, 45].

There have been comprehensive studies on statistical inferences for processes satisfying SDEs driven by Brownian motion. However, the recent interest is on SDEs driven by fBm since there have not been adequate studies on this topic. The purpose of this chapter is to estimate the Hurst parameter of the following SDE with initial value $x_0 \in \mathbb{R}$:

$$dX_t = a(t, X_t)dt + b\, dW_t^H, \qquad X(0) = x_0, \tag{5.2}$$

by CMARS methodology, together with the spline coefficients. Let us note that the $b(t, X_t) \equiv b$ term in Eqn. (5.1) is taken as a *constant*.

5.3 Estimation of Hurst Parameter Using Conic Multivariate Adaptive Regression Splines Method

In this section, as an alternative to the existing methods for estimation of the Hurst parameter, CMARS and the proposed methodology on SDEs will be introduced [65, 77]. For that purpose, firstly, we present the methodology and show how to apply this technique for finding the Hurst parameter of SDE defined in Eqn. (5.2).

5.3.1 Discretization of Stochastic Differential Equations with Fractional Brownian Motion

In general, the distribution of the stochastic process $(X_t)_{t\geq 0}$ is not known. Therefore, the discretized version of the SDEs, $\hat{X}_i$, should be simulated [29]. There are many discretization schemes for the SDEs generated by fBm such as Euler and Milstein scheme [15]. In this chapter, *Euler approximation* is used since Milstein approximation contains the derivatives of the diffusion term in Eqn. (5.2) which is equal to 0. The Euler approximation of the Eqn. (5.2) is:

$$\hat{x}_{i+1} = \hat{x}_i + a(\hat{x}_i, t_i)(t_{i+1} - t_i) + b(\hat{x}_i, t_i)(w_{i+1}^H - w_i^H). \tag{5.3}$$

For finitely many given data points $(\bar{x}_i, \bar{t}_i)$ $(i = 1, 2, \ldots, N)$, the symbolic form of the approximation can be given as follows:

$$\bar{y}_i = a(\bar{x}_i, \bar{t}_i) + b(\bar{x}_i, \bar{t}_i)\frac{\Delta \bar{w}_i^H}{\bar{h}_i}, \tag{5.4}$$

where $\Delta \bar{w}_i^H = \bar{w}_{i+1}^H - \bar{w}_i^H$ is a centered Gaussian random variable, $\bar{h}_i = \bar{t}_{i+1} - \bar{t}_i := \Delta \bar{t}_i$ represents step lengths and

$$\bar{y}_i = \begin{cases} \frac{\bar{x}_{i+1} - \bar{x}_i}{\bar{h}_i} & \text{if} \quad i = 1, 2, ..., N-1, \\ \\ \frac{\bar{x}_N - \bar{x}_{N-1}}{\bar{h}_N} & \text{if} \quad i = N, \end{cases}$$

are the difference quotients raised on the ith data values $(i = 1, 2, ..., N)$. A more compact form of the Eqn. (5.4) is defined by

$$\bar{y}_i = \bar{G}_i + \bar{F}_i c_i, \tag{5.5}$$

where $\bar{G}_i := a(\bar{x}_i, \bar{t}_i)$, $\bar{F}_i := b(\bar{x}_i, \bar{t}_i)$, and $c_i := \Delta \bar{w}_i^H / \bar{h}_i$. We note that Eqn. (5.5) can be considered as an approximation of the problem. The expressions stated subsequently until the end of Section 6.3 are described parametrically with respect to the Hurst parameter H. In Section 6.4, we shall specify it by numeric values.

5.3.2 Parameter Estimation of Coefficients and Hurst Parameter

To determine the unknown values in Eqn. (5.5), the following minimization problem is constructed using some abbreviated notation of the approximation [56]:

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \sum_{i=1}^{N} \left\| \bar{y}_i - \left(\bar{G}_i + \bar{F}_i c_i \right) \right\|_2^2 .$$

Here, $\boldsymbol{\theta}$ comprises all unknown parameters in the Euler approximation. To solve this optimization problem and to give a smoother, regularized approximation to the data,

we employ CMARS method which controls any high "variation" in the data. CMARS' basis functions are gradually constructed for the approximation of $\bar{G}_i$ and $\bar{F}_i$ with data $\bar{\boldsymbol{U}}^l_{i,B},\ \bar{\boldsymbol{U}}^m_{i,C} = (\bar{x}_i, \bar{t}_i)$ according to the following approaches [77]:

$$\bar{G}_i = \alpha_0 + \sum_{l=1}^{d^B} \alpha_l B_l(\bar{\boldsymbol{U}}^l_{i,B}), \quad \text{and} \quad \bar{F}_i c_i = \beta_0 + \sum_{m=1}^{d^C} \beta_m C_m(\bar{\boldsymbol{U}}^m_{i,C}).$$

Here, the forms of the basis functions are $B_l(\bar{\boldsymbol{U}}^l_B) = \prod_{k=1}^{2}[s^B_{\kappa^l_k} \cdot (x^B_{\kappa^l_k} - \tau^B_{\kappa^l_k})]_+$ and $C_m(\bar{\boldsymbol{U}}^m_C) = \prod_{k=1}^{2}[s^C_{\kappa^m_k} \cdot (x^C_{\kappa^m_k} - \tau^C_{\kappa^m_k})]_+$. Here, we choose the numbers K^B_l and K^C_m as maximal, namely, as 2.

We construct the PRSS for our minimization problem in the following form:

$$\begin{aligned} PRSS \ := \ & \sum_{i=1}^{N} (\dot{\bar{X}}_i - \left(\bar{G}_i + \bar{F}_i c_i\right))^2 + \\ & \sum_{l=1}^{d^B} \lambda_l \sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_l}} \int_{Q^B_l} \alpha_l^2 \left[D^{\boldsymbol{\alpha}}_{r,s} B_l(\bar{\boldsymbol{U}}^l_B)\right]^2 d\boldsymbol{U}^l_B + \\ & \sum_{m=1}^{d^C} \mu_m \sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_m}} \int_{Q^C_m} \beta_m^2 \left[D^{\boldsymbol{\alpha}}_{r,s} C_m(\bar{\boldsymbol{U}}^m_C)\right]^2 d\boldsymbol{U}^m_C. \end{aligned} \tag{5.6}$$

Here, the multipliers $\lambda_l,\ \mu_m \geq 0$ are smoothing parameters and they provide a trade-off between both accuracy and complexity. To approximate two multi-dimensional integrals in Eqn. (5.6), parallepipes $Q^B_l = \left[a^l_{1,B}, b^l_{1,B}\right] \times \left[a^l_{2,B}, b^l_{2,B}\right] = \prod_{k=1}^{2} Q^l_{k,B}$ and $Q^C_m = \left[a^m_{1,C}, b^m_{1,C}\right] \times \left[a^m_{2,C}, b^m_{2,C}\right] = \prod_{k=1}^{2} Q^m_{k,C}$ which encompass all our input data are constructed. Then, the following discretization is applied for the first multi-dimensional integral:

$$\begin{aligned} \int_{Q^B_l} \alpha_l^2 \left[D^{\boldsymbol{\alpha}}_{r,s} B_l(\bar{\boldsymbol{U}}^l_B)\right]^2 d\boldsymbol{U}^l_B \ &\approx \ \sum_{i=1}^{(N+1)^2} \left(\sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_m}} \alpha_l^2 \left[D^{\boldsymbol{\alpha}}_{r,s} B_l(\hat{\boldsymbol{U}}^l_{i,B})\right]^2 \right) \Delta\hat{\boldsymbol{U}}^l_{i,B} \\ &= \ \sum_{i=1}^{(N+1)^2} \left(\bar{L}^B_{il}\right)^2 \alpha_l^2 = \left\| \bar{\boldsymbol{L}}^B_l \alpha_l \right\|_2^2. \end{aligned} \tag{5.7}$$

The same discretization is also applied for the second multi-dimensional integral in Eqn. (5.6). For simplicity, we introduce PRSS in the following matrix-norm notation:

$$PRSS \approx \left\| \dot{\bar{\mathbf{X}}} - \bar{\boldsymbol{A}}\boldsymbol{\theta} \right\|_2^2 + \sum_{l=1}^{d^B} \lambda_l \left\| \bar{\boldsymbol{L}}^B_l \alpha_l \right\|_2^2 + \sum_{m=1}^{d^C} \mu_m \left\| \bar{\boldsymbol{L}}^C_m \beta_m \right\|_2^2, \tag{5.8}$$

where $\dot{\bar{\mathbf{X}}} = \left(\dot{\bar{X}}_1, \dot{\bar{X}}_2, ..., \dot{\bar{X}}_N\right)^T$, $\bar{\boldsymbol{A}} = \left(\bar{\boldsymbol{A}}_1^T, \bar{\boldsymbol{A}}_2^T, ..., \bar{\boldsymbol{A}}_N^T\right)^T$, $\boldsymbol{\theta} = \left(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T\right)^T$, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, ..., \alpha_{d^B})^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_{d^C})^T$. Moreover, $\bar{G}_i + \bar{F}_i c_i = \bar{\boldsymbol{A}}_i \boldsymbol{\theta}$, for $l = 1, 2, ..., d^B$ $\bar{\boldsymbol{L}}_l^B = \left(L_{1l}^B, L_{2l}^B, ..., L_{(N+1)^2 l}^B\right)^T$ and for $m = 1, 2, ..., d^C$ $\bar{\boldsymbol{L}}_m^C = \left(L_{1m}^C, L_{2m}^C, ..., L_{(N+1)^2 m}^C\right)^T$ [77].

Using uniform penalization by taking the same λ for each derivative term, the regularized approximation problem of PRSS turns into a Tikhonov Regularization problem:

$$PRSS \approx \left\| \dot{\bar{\mathbf{X}}} - \bar{\boldsymbol{A}}\boldsymbol{\theta} \right\|_2^2 + \lambda \left\| \bar{\boldsymbol{L}}\boldsymbol{\theta} \right\|_2^2. \tag{5.9}$$

Here, $\lambda = \lambda_1 = ... = \lambda_{d^B} = \mu_1 = ... = \mu_{d^C}$, and $\bar{\boldsymbol{L}}$ is an $((M_{max}+1) \times (M_{max}+1))$-diagonal matrix with first column $\boldsymbol{L}_0 = \mathbf{0}_{(N+1)^2}$ and the other columns being the vectors $\boldsymbol{L}_l^B$, $\boldsymbol{L}_m^C$, introduced above, where $M_{max} = d^B + d^C$.

As we mentioned in previous chapters, Tikhonov Regularization problem can be solved by a CQP program. In order to write the optimality condition for this problem, we firstly reformulate our program as the subsequent primal problem [75, 77]:

$$\begin{aligned}
&\underset{t,\boldsymbol{\theta}}{\text{minimize}} \;\; t, \\
&\text{such that} \;\; \boldsymbol{\chi} := \begin{bmatrix} \mathbf{0}_N & \bar{\boldsymbol{A}} \\ 1 & \mathbf{0}_{M_{max}+1}^T \end{bmatrix} \begin{bmatrix} t \\ \boldsymbol{\theta} \end{bmatrix} + \begin{bmatrix} -\dot{\bar{\mathbf{X}}} \\ 0 \end{bmatrix}, \\
&\boldsymbol{\eta} := \begin{bmatrix} \mathbf{0}_{M_{max}+1} & \bar{\boldsymbol{L}} \\ 0 & \mathbf{0}_{M_{max}+1}^T \end{bmatrix} \begin{bmatrix} t \\ \boldsymbol{\theta} \end{bmatrix} + \begin{bmatrix} \mathbf{0}_{M_{max}+1} \\ \sqrt{\tilde{M}} \end{bmatrix}, \\
&\boldsymbol{\chi} \in L^{N+1}, \; \boldsymbol{\eta} \in L^{M_{max}+2}.
\end{aligned} \tag{5.10}$$

The dual problem to the latter problem is given by

$$\begin{aligned}
&\text{maximize} \quad (\dot{\bar{\mathbf{X}}}^T, 0)\boldsymbol{\omega}_1 + \left(\mathbf{0}_{M_{max}+1}^T, -\sqrt{\tilde{M}}\right)\boldsymbol{\omega}_2, \\
&\begin{bmatrix} \mathbf{0}_N^T & 1 \\ \bar{\boldsymbol{A}}^T & \mathbf{0}_{M_{max}+1}^T \end{bmatrix} \boldsymbol{\omega}_1 + \begin{bmatrix} \mathbf{0}_{M_{max}+1}^T & 0 \\ \bar{\boldsymbol{L}}^T & \mathbf{0}_{M_{max}+1} \end{bmatrix} \boldsymbol{\omega}_2 = \begin{bmatrix} 1 \\ \mathbf{0}_{M_{max}+1} \end{bmatrix}, \\
&\boldsymbol{\omega}_1 \in L^{N+1}, \; \boldsymbol{\omega}_2 \in L^{M_{max}+2}.
\end{aligned} \tag{5.11}$$

A primal-dual optimal solution $(t, \boldsymbol{\theta}, \boldsymbol{\chi}, \boldsymbol{\eta}, \boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ is obtained when the optimality

conditions given in Eqn. (5.12) are satisfied:

$$
\begin{aligned}
&\boldsymbol{\chi} := \begin{bmatrix} \mathbf{0}_N & \bar{\boldsymbol{A}} \\ 1 & \mathbf{0}_{M_{max}+1}^T \end{bmatrix} \begin{bmatrix} t \\ \boldsymbol{\theta} \end{bmatrix} + \begin{bmatrix} -\dot{\tilde{\mathbf{X}}} \\ 0 \end{bmatrix}, \\
&\boldsymbol{\eta} := \begin{bmatrix} \mathbf{0}_{M_{max}+1} & \bar{\boldsymbol{L}} \\ 0 & \mathbf{0}_{M_{max}+1}^T \end{bmatrix} \begin{bmatrix} t \\ \boldsymbol{\theta} \end{bmatrix} + \begin{bmatrix} \mathbf{0}_{M_{max}+1} \\ \sqrt{\tilde{M}} \end{bmatrix}, \\
&\begin{bmatrix} \mathbf{0}_N^T & 1 \\ \bar{\boldsymbol{A}}^T & \mathbf{0}_{M_{max}+1}^T \end{bmatrix} \boldsymbol{\omega}_1 + \begin{bmatrix} \mathbf{0}_{M_{max}+1}^T & 0 \\ \bar{\boldsymbol{L}}^T & \mathbf{0}_{M_{max}+1} \end{bmatrix} \boldsymbol{\omega}_2 = \begin{bmatrix} 1 \\ \mathbf{0}_{M_{max}+1} \end{bmatrix}, \\
&\boldsymbol{\omega}_1^T \boldsymbol{\chi} = 0, \;\; \boldsymbol{\omega}_2^T \boldsymbol{\eta} = 0, \\
&\boldsymbol{\omega}_1 \in L^{N+1}, \;\; \boldsymbol{\omega}_2 \in L^{M_{max}+2}, \;\; \boldsymbol{\chi} \in L^{N+1}, \;\; \boldsymbol{\eta} \in L^{M_{max}+2}.
\end{aligned} \tag{5.12}
$$

5.4 A Numerical Application

In order to test the theory developed in the previous section, we start with simulating the stochastic process for a fixed Hurst parameter H using Cholesky method [12]. Now, our aim is to estimate the exact value of this Hurst parameter of the simulated data. We generate various stochastic processes which are the strong solution of SDEs driven by fBm with different Hurst parameters. Next, we construct CMARS model for each generated process to find the best fit. For the implementation of CMARS algorithm, basis functions are built using Salford MARS software program [33] as in [34, 65]. The optimization problem given in Eqn. (5.10) is solved by using IPMs via the optimization software MOSEK [37, 41]. Finally, we examine the performances of CMARS fits according to well-known performance measures. The steps described above are applied for H=0.2, H=0.3, H=0.7 and H=0.8. The results of the applications are summarized in Table 5.1 [75].

In the case of anti-persistent processes, namely, H=0.2, H=0.3, the values of MAE and MSE are lower and the values of R^2, Adj-R^2 and PWI are higher than the values for the other Hurst parameter values. Similar results are also obtained for the case of persistent processes. Hence, this shows that according to performance measure criteria, the best CMARS fit gives us the correct Hurst parameter value.

Table 5.1: CMARS performances for fBm generated by H=0.2, H=0.3, H=0.7, and H=0.8.

	Performance Measures				
Hurst index	**MAE**	**MSE**	**R^2**	**Adj-R^2**	**PWI**
$H = 0.1$	0.8766	1.4827	0.0370	-0.1651	1.0000
$\boldsymbol{H = 0.2}$	*0.7480	*0.9868	*0.9739	*0.9684	1.0000
$H = 0.3$	0.8861	1.5266	0.0098	-0.1979	1.0000
$H = 0.4$	0.8770	1.4733	0.0430	-0.1577	1.0000
$H = 0.5$	0.8839	1.5201	0.0164	-0.1900	1.0000
$H = 0.1$	0.7138	0.9776	0.0679	-0.1276	1.0000
$H = 0.2$	0.7162	0.9901	0.0576	-0.1401	1.0000
$\boldsymbol{H = 0.3}$	*0.2516	*0.984	*0.9699	*0.9636	1.0000
$H = 0.4$	0.6926	0.9387	0.1055	-0.0821	1.0000
$H = 0.5$	0.7053	0.9763	0.0697	-0.1254	1.0000
$H = 0.5$	0.7031	0.9719	0.2250	0.0623	1.0000
$H = 0.6$	0.7048	0.9784	0.2198	0.0560	0.9898
$\boldsymbol{H = 0.7}$	*0.2602	*0.9582	*0.9182	*0.9010	1.0000
$H = 0.8$	0.7041	0.9506	0.2419	0.0828	1.0000
$H = 0.9$	0.7081	0.9781	0.2200	0.0563	1.0000
$H = 0.5$	0.6068	0.7841	0.3498	0.2133	1.0000
$H = 0.6$	0.6359	0.8015	0.3345	0.1948	1.0000
$H = 0.7$	0.6053	0.7389	0.3815	0.2517	1.0000
$\boldsymbol{H = 0.8}$	*0.0822	*0.9883	*0.9768	*0.9720	1.0000
$H = 0.9$	0.6006	0.7294	0.3894	0.2613	1.0000

* indicates better performance

CHAPTER 6

CONCLUSION AND OUTLOOK

Recent developments in computer science provide environments in order to collect numerous data from various sources. Data mining methods enable us to analyze data for different purposes in many fields, such as science, economics, finance, environment, engineering and energy. One of the modern methods of data mining, CMARS, has been developed as an alternative to the backward stepwise part of the MARS algorithm [65].

CMARS is a parameter estimation method for regression and classification, which simply involves regularization and employes continuous optimization. It constructs a functional form of the model by using basis functions, which are a special form of splines. In fact, our splines will be "discretely" nonsmooth, herewith reducing complexity also. Then, it estimates coefficients of basis functions using CQP. The flexible, *adaptive* nature of the CMARS modeling has led to a successful implementation on various processes with nonlinear structure, e.g., in industrial engineering, image processing and earthquake engineering [32, 65, 71].

The success of the CMARS applications led to adapting the CMARS model into a wide frame of advanced methods of statistics and applied mathematics. In the following paragraphs, we summarize the refinements and extensions of the CMARS method to identify and explore the connections between data mining and other applied sciences.

CGPLM, the first extension of CMARS, is become a very popular tool and shown to be effective in many areas such as finance, actuarial science, medicine, etc. Embedding our data mining methodology and optimization approach to GPLM is more recent and its popularity in the finance community is increasing. The main reason for using CMARS method here is that it is computationally fast and also overcomes some well-known deficiencies of traditional methods. In this thesis, we have aimed to present how the advantages and strengths of GPLM can be effectively combined with the computational power of data mining methods by incorporating CMARS and estimating all the related unknown coefficients of GPLM by the CQP. As a future work, CGPLM (with CMARS) can be applied to large data sets, e.g., related with the dynamics in the financial sector, with many potential predictors and various other emerging application areas.

In this thesis, the mean-shift outlier regression problem is newly represented as a CQP and it achieved excellent numerical results with CMARS, compared with a new regu-

larized, CQP supported approach by a linear model. These new methods can be applied to a wide range of application areas, such as medical signal processing, quality analysis and control in manufacturing, early warning systems in meteorology, ecology and the financial sector, especially for a new view at the analysis and control of financial "bubbles."

SDEs play an important role in the many field of science, especially, in finance. These equations, however, are usually hard to represent and resolve by a computer. In order to identify them in a simplified manner, we aim to use CMARS method. CMARS provides the opportunity for statistical inference of stochastic process parameters with the help of basis functions. Especially for large sample size, our estimation procedure based on special discretization and optimization techniques captures the behavior of the stochastic process very well.

In this thesis, a parameter estimation procedure for SDEs is presented and implemented. This procedure is based on a discretization of the SDEs and on the CMARS algorithm. CMARS holds for a reasonable approach to nonparametrically estimate the complete functional form of the deterministic drift and the diffusion term of a SDE by identifying unknown parameters of these functions based on discretization of Milstein approximation.

Differently from ODEs, SDEs incorporate random effects and they are computationally expensive. In the case of multi-dimensional stochastic processes, strong solvers are necessary to explore characteristics of systems. As introduced before, a main problem in estimating parameters of SDEs is the structure of the stochastic processes that can be of any distributional characteristics and any nonlinear and nonconvex form of the underlying data set.

In this thesis, a parameter estimation procedure for multi-dimensional SDEs has been presented and implemented. This procedure is based on a discretization of the SDEs by using Milstein approximation and CMARS. In addition to them, using CQP programs gives us the opportunity to benefit from well-structured convex optimization.

In this thesis, we particularly deal with the system case of SDEs driven by correlated Brownian motions which are experienced in real-life problems of finance such as in stochastic volatility and credit risk with stochastic interest. Moreover, implementations of the given approaches to financial data are successfully demonstrated.

Our future work will proceed on the generalization of parameter estimation for the multi-dimensional systems' case with vectors of fractional Brownian motions, in theory, methods and applications.

This thesis gives a new contribution to Hurst parameter estimation theory for the strong solution of SDEs driven by fBms using CMARS technique. The main superiority of our approach to the others is that it does not only estimate the Hurst parameter but also finds spline parameters of the stochastic process. Furthermore, our representation of financial and other processes is empowered by all the modeling and numerical advantages of the CMARS. By this, a bridge has been offered between convex optimization and Hurst parameter estimation theory.

In present thesis, we follow a *two-level* approach with the determination of the parameters at the *lower level*, except for the Hurst-parameter which was chosen at the following *upper level*. This approach can be regarded as a *parametric optimization* [17, 27]. In future research, we will deepen and extend this approach by both more *model-free* strategies (e.g., from statistics and data mining), especially, more *model-based* ones, and with a comparison of them. The model-based approaches will be of a more *integrated* mathematical nature and in the analytical line that we initiated in this thesis.

As a future investigation, the results and achievements on SDEs, presented in Chapter 4 and 5, can be applied to new research and application areas for them such as biology and medicine. By using the advantages of our new methodology based on CMARS and optimization theory, real-world processes from the nature are able to model. Thus, this extension in view of interest and application will be useful. Furthermore, we want to compare the performance of our methodologies to other parameter estimation methods of SDEs based on different stochastic processes.

The choice of the upper bound, $\tilde{M}$, in our CQP problems is based on an L-curve. In future, the optimal value of $\tilde{M}$ can be determined by an appropriate optimization-based approach where $\tilde{M}$ is included as a state or decision variable. Thus, our proposed approaches based on a CQP problem can select a better model for us. For this purpose, multi-objective approaches can be used.

One drawback of proposed methodologies is their computational speed. In our applications, preparation of the data sets, construction of the basis functions and discretizations of the given equations were made separately. This takes a lot of time. In this respect, we are going to improve our coding and make a user-friendly tool for its prospective users.

REFERENCES

[1] M. U. Akhmet, S. W. Öktem, H. Pickl, and G. W. Weber, An anticipatory extension of malthusian model, Computing Anticipatory Systems, 839, pp. 260–264, 2006, seventh International Conference on Computing Anticipatory Systems.

[2] B. E. Alos, O. Mazet, and D. Nualart, Stochastic calculus with respect to gaussian process, The Annals of Probability, 29, pp. 766–801, 2001.

[3] R. C. Aster, B. Borchers, and C. Thurber, *Parameter Estimation and Inverse Problems*, Academic Press, Burlington, 2012.

[4] V. Barnett and T. Lewis, *Outliers in Statistical Data*, Wiley, Great Britain, 1994.

[5] İ. Batmaz and G. Köksal, *Surveillance Technologies and Early Warning Systems: Data Mining Applications for Risk Detection*, chapter Overview of Knowledge Discovery in Databases Process and Data Mining for Surveillance Technologies and EWS, PA: IGI Global Publisher (Idea Group Publisher), Hershey, 2011.

[6] İ. Batmaz, F. Yerlikaya-Özkurt, E. Kartal-Koç, G. Köksal, and G. W. Weber, Evaluating the CMARS performance for modeling nonlinearities, Proceedings of the 3rd Global Conference on Power Control and Optimization, 1239, pp. 351–357, 2010.

[7] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms and Engineering Applications*, SIAM, Philadelphia, 2001.

[8] J. P. N. Bishwal, *Parameter Estimation in Stochastic Differential Equations*, Springer-Verlag, Berlin, 2008.

[9] T. Bonesky, Morozov's discrepancy principle and tikhonov-type functionals, Inverse Problems, 25(1), pp. 1–11, 2009.

[10] A. Chronopoulou and F. G. Viens, *Recent Advances in Stochastic Dynamics and Stochastic Analysis*, chapter Hurst Index Estimation for Self-similar Processes with Long-memory, pp. 1–28, World Scientific Publishing Co Pte Ltd, Hershey, 2009.

[11] R. D. Cook, *Residuals and Influence in Regression*, Chapman and Hall, New York, 1982.

[12] T. Dieker, *Simulation of Fractional Brownian Motion*, Thesis (MSc), University of Twente, Amsterdam, 2004.

[13] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer, New York, 2006.

[14] J. H. Friedman, Multivariate adaptive regression splines, The Annals of Statistics, 19(1), pp. 1–141, 1991.

[15] M. Gradinaru and I. Nourdin, Milstein's type schemes for fractional SDEs, Annales de l' Institut Henri Poincare-Probabilites et Statistiques, 45(4), pp. 1085–1098, 2009.

[16] C. Gu, *Smoothing Spline ANOVA Models*, Springer, New York, 2002.

[17] J. Guddat, F. Guerra, and H. T. Jongen, *Parametric Optimization: Singularities, Pathfollowing and Jumps*, John Wiley & Sons, BG Teubner, Stuttgart, Chichester, 1990.

[18] A. S. Hadi and J. S. Simonoff, Procedures for the identification of multiple outliers in linear models, Journal of the American Statistical Association, 88(424), pp. 1264–1272, 1993.

[19] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*, Chapman and Hall, New York, 1990.

[20] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*, Springer Verlag, New York, 2001.

[21] D. J. Higham, An algorithmic introduction to numerical simulation of stochastic differential equations, SIAM Journal Numerical Analysis, 43(3), pp. 525–546, 2001.

[22] W. Hock and K. Schittkowski, *Test Examples for Nonlinear Programming Codes*, Springer-Verlag, New York, 1981.

[23] P. J. Huber, *Robust Statistics*, John Wiley and Sons, Hoboken, NJ, 2009.

[24] H. Hurst, Long term storage capacity of reservoirs, Transactions of the American Society of Civil Engineers, 116, pp. 770–799, 1951.

[25] S. Jaimungal and V. Surkov, Levy-based cross-commodity models and derivative valuation, SIAM Journal on Financial Mathematics, 2(1), pp. 464–487, 2011.

[26] M. Jeanblanc, M. Yor, and M. Chesney, *Mathematical Methods for Financial Markets*, Springer-Verlag London Ltd., London, 2009.

[27] H. T. Jongen and G. W. Weber, On parametric nonlinear programming, Annals of Operations Research, 27, pp. 253–284, 1990.

[28] S. S. Kim, S. H. Park, and W. J. Krzanowski, Simultaneous variable selection and outlier identification in linear regression using the mean-shift outlier model, Journal of Applied Statistics, 35(3), pp. 283–291, 2008.

[29] P. E. Kloeden, E. Platen, and H. Schurz, *Numerical Solution of SDE Through Computer Experiments*, Springer-Verlag, New York, 1994.

[30] G. Köksal, İ. Batmaz, and M. C. Testik, *Expert Systems with Applications*, chapter A Review of Data Mining Applications for Quality Improvement in Manufacturing Industry, 2011.

[31] R. Korn and E. Korn, *Options Pricing and Portfolio Optimization: Modern Methods of Financial Mathematics*, Oxford University Press, 2001.

[32] S. Kuter, G. W. Weber, A. Özmen, and Z. Akyürek, *Springer Volume Modeling, Optimization, Dynamics and Bioeconomy*, chapter Modern Applied Mathematics for Alternative Modeling of the Atmospheric Effects on Satellite Images, Springer Proceedings in Mathematics, 2011.

[33] MARS, Salford Systems, available at http://www.salfordsystems.com/mars/phb.

[34] MATLAB, Version 7.5 (r2007b).

[35] MINITAB 15, Statistical Software, available at www.minitab.com.

[36] Y. Mishura, *Stochastic Calculus for Fractional Brownian Motion and Related Topics*, Springer-Verlag, Berlin, Heidelberg, 2008.

[37] MOSEK, A very powerful commercial software for CQP, available at http://www.mosek.com.

[38] M. Müller, Estimation and testing in generalized partial linear models: A comparative study, Statistics and Computing, 11, pp. 299–309, 2001.

[39] R. H. Myers and D. C. Montgomery, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, Wiley, New York, 2002.

[40] J. A. Nelder and R. W. M. Wedderburn, Generalized linear models, Journal of the Royal Statistical Society: Series A, 145, pp. 470–484, 1972.

[41] Y. E. Nesterov and A. S. Nemirovski, *Interior-point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.

[42] J. Neter, M. Kutner, W. Wasserman, and C. Nachtsheim, *Applied Linear Statistical Models*, MA: WCB/McGrawHill, Boston, 1996.

[43] B. K. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, Springer, Berlin, 2007.

[44] D. Pena and V. J. Yohai, The detection of influential subsets in linear regression by using an influential matrix, Journal of the Royal Statisical Association Series B, 57(1), pp. 145–156, 1995.

[45] B. L. S. Rao, *Statistical Inference for Fractional Diffusion Processes*, Wiley, UK, first edition, 2010.

[46] C. R. Rao, H. Toutenburg, and A. Fieger, *Linear Models: Least Squares and Alternatives*, Springer, 1999.

[47] M. Reiß, Lecture notes on stochastic differential equations for courses given at Humboldt University Berlin and University of Heidelberg, 2007, available at http://math.uni-heidelberg.de/studinfo/reiss/sode-lecture.pdf.

[48] A. C. Rencher, *Linear Models in Statistics*, John Wiley and Sons, New York, 2000.

[49] J. Renegar, *A Mathematical View of Interior-Point Methods in Convex Programming*, SIAM, Philadelphia, 2001.

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

[51] P. J. Rousseeuw and K. Van Driessen, Computing lts regression for large data sets, Data Mining and Knowledge Discovery, 12, pp. 29–45, 2006.

[52] SAS, Customer support, available at http://support.sas.com/.

[53] Ö. Sezgin-Alp, E. Büyükbebeci, A. Işcanoğlu-Çekiç, F. Yerlikaya-Özkurt, P. Taylan, and G. W. Weber, -CMARS and GAM and CQP- modern optimization methods applied to international credit default prediction, Journal of Computational and Applied Mathematics, 235, pp. 4639–4651, 2011.

[54] A. Sharkasi, M. Crane, H. J. Ruskin, and J. A. Matos, The reaction of stock markets to crashes and events: A comparison study between emerging and mature markets using wavelet transforms, Physica A, 368, pp. 511–521, 2006.

[55] S. E. Shreve, *Stochastic calculus for finance. II*, Springer Finance, Springer-Verlag, New York, 2004.

[56] P. Taylan and G. W. Weber, Organization in finance prepared by stochastic differential equations with additive and nonlinear models and continuous optimization, Organizacija (Organization - Journal of Management, Information Systems and Human Resources), 41, pp. 185–193, 2008.

[57] 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, Optimization, 56, pp. 1–24, 2007.

[58] P. Taylan, G. W. Weber, L. Liu, and F. Yerlikaya-Özkurt, On foundations of parameter estimation for generalized partial linear models with B-splines and continuous optimization, Computers and Mathematics with Applications, 60, pp. 134–143, 2010.

[59] P. Taylan, G. W. Weber, and F. Yerlikaya-Özkurt, A new approach to multivariate adaptive regression spline by using tikhonov regularization and continuous optimization, TOP (the Operational Research journal of SEIO (Spanish Statistics and Operations Research Society)), 18, pp. 377–395, 2010.

[60] P. Taylan, F. Yerlikaya-Özkurt, and G. W. Weber, Parameter estimation for semiparametric models with CMARS and its applications, Working Paper, IAM, METU, Turkey, 2010.

[61] P. Taylan, F. Yerlikaya-Özkurt, and G. W. Weber, An approach to the mean shift outlier model by tikhonov regularization and conic programming, Intelligent Data Analysis, 18, pp. 79–94, 2014, the speacial issue on Business Analytics and Intelligent Optimization.

[62] UCI, Machine Learning Repository, available at http://archive.ics.uci.edu/ml/.

[63] Ö. Uğur, *An Introduction to Computational Finance*, Imperial College Press, London, 2009.

[64] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, 1990.

[65] G. W. Weber, İ. Batmaz, G. Köksal, P. Taylan, and F. Yerlikaya-Özkurt, CMARS: A new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization, Inverse Problems in Science and Engineering, 20(3), pp. 371–400, 2012.

[66] G. W. Weber, Z. Cavusoglu, and A. Özmen, Predicting default probabilities in emerging markets by new conic generalized partial linear models and their optimization, Optimization: A Journal of Mathematical Programming and Operations Research, 61, pp. 443–457, 2012.

[67] G. W. Weber, P. Taylan, Z. K. Görgülü, H. Abd. Rahman, and A. Bahar, *Dynamics, Games and Science*, chapter Parameter Estimation in Stochastic Differential Equations, pp. 703–733, Springer-Verlag, 2011, iSBN:978-3-642-14787-6.

[68] G. W. Weber, P. Taylan, K. Yı ldırak, and Z. K. Görgülü, Financial regression and organization, Special Issue on Optimization in Finance of DCDIS-B (Dynamics of Continuous, Discrete and Impulsive Systems (Series B)), 17, pp. 149–174, 2010.

[69] S. N. Wood, *Generalized Additive Models, An Introduction with R*, Chapman and Hall, New York, 2006.

[70] F. Yerlikaya, *A New Contribution to Nonlinear Robust Regression and Classification with MARS and Its Application to Data Mining for Quality Control in Manufacturing*, Thesis (MSc), Middle East Technical University, Turkey, 2008.

[71] F. Yerlikaya-Özkurt, A. Askan, and G. W. Weber, An alternative approach to the ground motion prediction problem by a non-parametric adaptive regression method, Engineering Optimization, 2013, DOI: 10.1080/0305215X.2013.858141.

[72] F. Yerlikaya-Özkurt, İ. Batmaz, and G. W. Weber, *Springer Volume Modeling, Optimization, Dynamics and Bioeconomy, Series Springer Proceedings in Mathematics*, chapter A Review and New Contribution on Conic Multivariate Adaptive Regression Splines (CMARS): A Powerful Tool for Predictive Data Mining, 2013.

[73] F. Yerlikaya-Özkurt, P. Taylan, İ. Batmaz, G. Köksal, and G. W. Weber, A modification of MARS by Tikhonov regularization and conic quadratic programming for modeling quality data, 2009, EURO XXIII 2009, Bonn, Germany, July 5-8.

[74] F. Yerlikaya-Özkurt, P. Taylan, and G. W. Weber, Parameter estimation for semiparametric models with CMARS and its applications, 2010, EURO XXIV 2010, Lisbon, Portugal, July 11-14.

[75] F. Yerlikaya-Özkurt, C. Vardar-Acar, Y. Yolcu-Okur, and G. W. Weber, Estimation of the Hurst parameter for fractional Brownian motion using the CMARS method, Journal of Computational and Applied Mathematics, 259, pp. 843–850, 2014.

[76] F. Yerlikaya-Özkurt and G. W. Weber, Estimation of multi-dimensional stochastic differential equations with CMARS method, 2012, EURO XXV 2012, Vilnius, Lithuania, July 8-11.

[77] F. Yerlikaya-Özkurt and G. W. Weber, Identification of stochastic differential equations by conic optimization of multivariate adaptive regression splines, Working Paper, IAM, METU, Turkey, 2013.

APPENDIX A

Performance Measures Used in Comparisons

The performance measures with their general notations are presented as follows:

Mean Absolute Error (***MAE***): MAE is the average magnitude of the error; it is defined by

$$MAE := \frac{1}{N}\sum_{i=1}^{N}\left|\bar{y}_i - \hat{\bar{y}}_i\right|,$$

where $\bar{y}_i$ is ith observed response value and $\hat{\bar{y}}_i$ is ith fitted response value and N is the number of observations. The smaller value for this measure gives the better model.

Mean Square Error (***MSE***): MSE is defined as

$$MSE := \frac{1}{N-p}\sum_{i=1}^{N}(\bar{y}_i - \hat{\bar{y}}_i)^2,$$

where p is the number of terms in the model. The MSE represents an unbiased estimate of the error variance, and it depends on both the residual and number of predictive variables. Similar to the MAE, smaller values for the RMSE indicate a better model.

Multiple Coefficient of Determination (***R***^{2}): R^2 is a coefficient of determination. As the R^2 values become higher, a better fit is obtained for the model. The formulation of R^2 is as follows:

$$R^2 := 1 - \frac{\sum_{i=1}^{N}(\bar{y}_i - \hat{\bar{y}}_i)^2}{\sum_{i=1}^{N}(\bar{y}_i - \bar{\bar{y}})^2}.$$

Adjusted R^2 (***Adj-R***^{2}): This value is used to compute the number of predictors in the model. It is useful for comparing models with different numbers of independent variables. The higher the Adj-R^2 value is, the better the model fits the data. The formula is:

$$\text{Adj} - R^2 := 1 - (1 - R^2)\frac{N-1}{N-p-1},$$

where $N - p - 1 \neq 0$.

Percentage of Residuals within Three Standard Deviation of Mean (***PWI***): PWI is obtained by the sum of indicator variables over all observations divided by the total number of observations. The indicator variables take the value of 1 if the absolute value of the difference between the actual and predicted response is within some user-specified thresholds. In this thesis, 3 standard deviations of mean are used as a threshold.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Yerlikaya-Özkurt, Fatma
Nationality: Turkish
Date and Place of Birth: 1983, Iğdır
E-mail: fatmayerlikaya@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
M.S.	Scientific Computing, IAM, METU	2008
B.S.	Department of Mathematics, Ankara University	2005

AWARDS AND SCHOLARSHIPS

METU Best MSc. Thesis Award Winner, A New Contribution to Nonlinear Robust Regression and Classification with MARS and Its Application to Data Mining for Quality Control in Manufacturing (May 27, 2009).

TUBİTAK higher education scholarship for MSc. Degree (2006-2008).

TUBİTAK higher education scholarship for PhD. Degree (2008-2013).

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2013-	Aerospace Industries	Project consultant
2007-2009	Department of Industrial Engineering	Project assistance

PUBLICATIONS

Books/Chapters in a book

F. Yerlikaya-Özkurt, İ. Batmaz and G. W. Weber, A review of conic multivariate adaptive regression splines (CMARS): a powerful tool for predictive data mining, chapter in book, Springer volume Modeling, Optimization, Dynamics and Bioeconomy, series Springer Proceedings in Mathematics, D. Zilberman and A. Pinto, eds., 2013.

Papers/Conference Proceedings

P. Taylan, G. W. Weber and F. Yerlikaya-Özkurt, *Continuous optimization applied in MARS for modern applications in finance, science and technology*, ISI Proceedings of 20th Mini-EURO Conference Continuous Optimization and Knowledge-Based Technologies (Neringa, Lithuania, May 20-23, 2008) 317-322.

P. Taylan, G. W. Weber and F. Yerlikaya-Özkurt, *A new approach to multivariate adaptive regression splines by using Tikhonov regularization and continuous optimization*, TOP (the Operational Research journal of SEIO (Spanish Statistics and Operations Research Society)) 18, 2010, pp. 377-395.

İ. Batmaz, F. Yerlikaya-Özkurt, E. Kartal-Koç, G. Köksal and G. W. Weber, *Evaluating the CMARS performance for modeling nonlinearities*, in the Proceedings of the 3rd Global Conference on Power Control and Optimization, Gold Coast (Australia), 1239, 2010, pp. 351-357.

P. Taylan, G. W. Weber, L. Liu and F. Yerlikaya-Özkurt, *On foundations of parameter estimation for generalized partial linear models with B-splines and continuous optimization*, Computers and Mathematics with Applications (CAMWA) 60, 1, 2010, pp.134-143.

Ö. Sezgin-Alp, E. Büyükbebeci, A. Işcanoğlu-Çekiç, F. Yerlikaya-Özkurt, P. Taylan and G. W. Weber, *-CMARS and GAM and CQP- modern optimization methods applied to international credit default prediction*, Journal of Computational and Applied Mathematics (JCAM) 235, 2011, pp. 4639-4651.

G. W. Weber, İ. Batmaz, G. Köksal, P. Taylan and F. Yerlikaya-Özkurt, *CMARS: A new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization*, Inverse Problems in Science and Engineering, 20, 3, 2012, pp. 371-400.

P. Taylan, F. Yerlikaya-Özkurt and G. W. Weber, *An approach to mean shift outlier model (MSOM) by Tikhonov regularization and conic programming*, Intelligent Data Analysis, 18, 2014, pp. 79-94 .

F. Yerlikaya-Özkurt, C. Vardar-Acar, Y. Yolcu-Okur and G. W. Weber, *Estimation of Hurst parameter of fractional Brownian motion using CMARS method*, Journal of

Computational and Applied Mathematics, 259, 2014, pp. 843-850.

F. Yerlikaya-Özkurt, A. Askan and G. W. Weber, *An alternative approach to ground motion prediction problem by a nonparametric adaptive regression method*, Engineering Optimization, 2013. (DOI: 10.1080/0305215X.2013.858141)

TECHNICAL PRESENTATIONS

F. Yerlikaya-Özkurt, A. Askan and G. W. Weber, A Novel Nonparametric Adaptive Regression Methodology for Ground Motion Prediction, EURO XXVI 2013 EURO-INFORMS Joint International Conference, Rome, Italy, July 1-4, 2013.

F. Yerlikaya-Özkurt, C. Vardar-Acar, Y. Yolcu-Okur, and G. W. Weber, Estimation of Hurst Parameter of Fractional Brownian Motion Using CMARS Method, International Conference on Applied and Computational Mathematics Ankara, Turkey, October 3-6, 2012.

F. Yerlikaya-Özkurt and G. W. Weber, Estimation of Multi-dimensional Stochastic Differential Equations with CMARS Method, EURO XXV 2012, Vilnius, Lithuania, July 8-11, 2012.

F. Yerlikaya-Özkurt, P. Taylan and G. W. Weber, Mean Shift Outlier Model with MARS and Continuous Optimization, IFORS 2011, Melbourne, Australia, July 10-15, 2011.

F. Yerlikaya-Özkurt, P. Taylan and G. W. Weber, CMARS Method for Stochastic Differential Equations, OR 2011 Zurich, August 30 - September 2, 2011, Zurich, Switzerland.

F. Yerlikaya-Özkurt, P. Taylan and G. W. Weber, Parameter Estimation for Semiparametric Models with CMARS and Its Applications, EURO XXIV 2010, Lisbon, Portugal, July 11-14, 2010.

P. Taylan, F. Yerlikaya-Özkurt and G. W. Weber, A New Contribution to Mean Shift Outlier Model with Continuous Optimization, EURO XXIV 2010, Lisbon, Portugal, July 11-14, 2010.

F. Yerlikaya-Özkurt, G. W. Weber and P. Taylan, Parameter Estimation for Semiparametric Models with CMARS and Its Applications, 5th International Summer School Achievements and Applications of Contemporary Informatics, Mathematics and Physics, National University of Technology of the Ukraine, Kiev, Ukraine, August 3-15, 2010.

F. Yerlikaya-Özkurt, P. Taylan, İ. Batmaz, G. Köksal and G. W. Weber, A Modification of MARS by Tikhonov Regularization and Conic Quadratic Programming for Modeling Quality Data, EURO XXIII 2009, Bonn, Germany, July 5-8, 2009.

F. Yerlikaya-Özkurt, G. W. Weber and A. Özmen, Robustification of CMARS, 14th International Congress on Computational and Applied Mathematics (ICCAM), Antalya,

Turkey, September 29 - October 2, 2009.

F. Yerlikaya-Özkurt, G. W. Weber, P. Taylan, İ. Batmaz and G. Köksal, CMARS ile Doğrusal Olmayan Veri Yapılarının Modellenmesi, YA/EM'09: Yöneylem Araştırması ve Endüstri Mühendisliği 29. Ulusal Kongresi. Ankara, Türkiye, 22-24 Haziran, 2009. 94. CDROM.

F. Yerlikaya, G. W. Weber, P. Taylan, İ. Batmaz and G. Köksal, MARS Algoritmasında Tikhonov Düzenlemesi ve Çok Amaçlı Optimizasyon Kullanımı, YA/EM'08: Yöneylem Araştırması ve Endüstri Mühendisliği 28. Ulusal Kongresi. İstanbul, Türkiye, 30 Haziran-2 Temmuz, 2008. CDROM.

MEMBERSHIP IN PROFESSIONAL SOCIETIES

EURO - Association of European Operational Research Societies

EUROPT - The Continuous Optimization Working Group of EURO

SIAM - Society of Industrial and Applied Mathematics

YA/EM - Yöneylem Araştırması ve Endüstri Mühendisliği