Dynamical Gene-Environment Networks under Ellipsoidal Uncertainty – Set-Theoretic Regression Analysis Based on Ellipsoidal OR

Erik Kropat, Gerhard-Wilhelm Weber and Selma Belen

Abstract We consider dynamical gene-environment networks under ellipsoidal uncertainty and discuss the corresponding set-theoretic regression models. Clustering techniques are applied for an identification of functionally related groups of genes and environmental factors. Clusters can partially overlap as single genes possibly regulate multiple groups of data items. The uncertain states of cluster elements are represented in terms of ellipsoids referring to stochastic dependencies between the multivariate data variables. The time-dependent behaviour of the system variables and clusters is determined by a regulatory system with (affine-) linear coupling rules. Explicit representations of the uncertain multivariate future states of the system are calculated by ellipsoidal calculus. Various set-theoretic regression models are introduced in order to estimate the unknown system parameters. Hereby, we extend our *Ellipsoidal Operations Research* previously introduced for gene-environment networks of strictly disjoint clusters to possibly overlapping clusters. We analyze the corresponding optimization problems, in particular in view of their solvability by interior point methods and semidefinite programming and we conclude with a discussion of structural frontiers and future research challenges.

Erik Kropat
Universität der Bundeswehr München, Institute for Theoretical Computer Science, Mathematics and Operations Research, Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
e-mail: erik.kropat@unibw.de

Gerhard-Wilhelm Weber
Middle East Technical University, Institute of Applied Mathematics, 06531 Ankara, Turkey.
Honorary positions: Faculty of Economics, Business and Law, University of Siegen, Germany;
Center for Research on Optimization and Control, University of Aveiro, Portugal;
Faculty of Science, Universiti Teknologi Malaysia (UTM), Skudai, Malaysia
e-mail: gweber@metu.edu.tr

Selma Belen
CAG University, Yenice-Tarsus, 33800 Mersin, Turkey
e-mail: sbelen@cag.edu.tr

1 Introduction

The development of microarray technologies enabled researchers in genetics to monitor the expression values of thousands of genes simultaneously. The availability of such huge data sets challenged bioinformatics and mathematics and led to the development of new methods for knowledge discovery in functional genomic data sets. Many concepts from data mining and statistical analysis were applied in order to reveal the cellular processes involved. In particular, *clustering techniques* were used for an identification of functionally related groups of genes. Among them were, for example, techniques as *k-means* [47, 82], *hierarchical clustering* [17, 26, 47], *self-organizing maps* [44, 36, 52], *principle component analysis* [63, 79], *singular value decomposition* [6, 63] and *support vector machines* [15, 35]. However, these methods often resulted in *disjoint clusters* and in many applications such a *hard clustering* is too strict because of the quality of the data or the presence of outliers and errors. In addition, a single gene (or a group of genes) can have a regulating effect on various clusters of genes, as for example in context with the identification of synexpression groups and the analysis of synexpression control networks [37]. *Fuzzy-clustering* [57] or other methods resulting in a partially overlapping cluster decomposition can alleviate the effects of noise-prone data and can lead to a more flexible representation of interconnections between groups of data. Although these methods - exact or flexible - proved to be sufficient in identifying, e.g., damaged or cancerous genes and groups of regulating genetic items, they are nevertheless considered as *static methods* which do not shed any light on the time-dependent behaviour of the genetic network. *Time-series analysis* [27] or related approaches can be applied to forecast the time-dependent states of the expression values. In our studies [19, 67], we demonstrate the way how we develop a time-discrete dynamics whose parameters we identified and how we use the given data in order to test the goodness of the regression. Since the time-discrete dynamics can be gained by various kinds of discretization schemes, the aforementioned comparison also helps to test the quality of these schemes and rank them. It turned out that 3rd order Heun's method has a smoothing effect with respect to the prediction, which is regarded to be very good and 'natural' in the (natural) context of gene-environment networks, and that it leads to a faster convergence to equilibrium points of the dynamics.

In ways such as aforementioned, the gained time-discrete dynamics supports the prediction. As we explained in [59, 76, 71], we can also in further ways use these dynamics for a texting of our model, i.e., of the quality of data fitting. Actually, in various application contexts it is known or at least considered to be guaranteed, that the 'expression levels' of state variables are staying in bounded intervals. If, however, our discrete dynamics emerged in some direction in an unbounded kind, then we could conclude that this dynamics, e.g., it parameters identified by us, cannot be accepted. In such a case, the hypothesis of the model has to be rejected and, within our entire learning processes, improvements in the model structure be made and the parameter estimation restarted.

Here, we combine methods from clustering theory and dynamic systems under the presence of errors and uncertainty. For an analysis of the interconnections between clusters of genes and environmental factors and a prediction of their future states we study *gene-environment networks under ellipsoidal uncertainty*. In contrast to other models of *genetic systems*, these networks capture and assess the regulating effects of additional environmental factors. In general, gene-environment networks consist of two major groups of data items – the *genes* (or proteins and other molecules) and the so-called *environmental factors*, which stand for other cell components like toxins, transcription factors etc. that often play an important but nevertheless underestimated role in the regulatory system. We note that beside genetic systems many other examples from life sciences and systems biology refer to *gene-environment networks*. Among them are, e.g., *metabolic networks* [13, 45, 74], *immunological networks* [23], networks in *toxicogenomics* [38], but also *social-* and *ecological networks* [22]. We refer to [21, 28, 53, 54, 55, 80, 81] for applications, practical examples and numerical calculations. Recent studies on gene-environment networks focussed on errors and uncertainty. The potential deviation from measurement values and predictions of *each gene* was measured in terms of error intervals by imposing bounds on each variable. Various regression models have been developed and studied with the help of *generalized Chebychev approximation* and, equivalently to that approximation, *semi-infinite* and even *generalized semi-infinite optimization* [62, 64, 66, 68, 69, 71, 74, 75, 77]. However, error intervals referring to single variables do not reflect correlations of the multivariate data within specific clusters of genetic and environmental items. In our approach we apply clustering techniques for an identification of functionally related groups of genes (or environmental factors) commonly exerting influence on other groups of genes and/or environmental items. In particular, we focus on *possibly overlapping clusters* and by this we further extend the approach from [31], where a strict subdivision of data was assumed. Each cluster stands for a group of correlated data items. In order to measure data uncertainty, the multivariate state of genes (or environmental factors) in a cluster will be represented in terms of *ellipsoids*. These uncertainty sets refer to stochastic representations of errors and are directly related to Gaussian distributions and the corresponding covariance matrices. Error ellipsoids are considered as more flexible than error intervals where stochastic dependencies among any two of the errors made in the measurement of expression values and environmental levels are not taken into account explicitly. However, the two approaches are related, because any confidence ellipsoid can be inscribed into a sufficiently large and suitably oriented parallelpipe or, in reverse, it can be contained in such a paraxial set.

The dynamics of the uncertain (ellipsoidal) states of genes and environmental factors are represented by a time-dependent *regulatory model*. The coupling rules of this model are based on *ellipsoidal calculus* and they determine the interactions between the various clusters. With this model, predictions of the future states can be calculated explicitly. In addition, we introduce an *iterative procedure* for calculating the centers and shape matrices of the ellipsoidal states. The parameter constellation of the regulatory model refers to the topology and the degree of connectivity of the underlying gene-environment network. For an estimation of the unknown parameters, various set-theoretic regression models are introduced. These models are heavily effected by the cluster decomposition and the overlap of clusters. The associated objective functions of the regression models compare the ellipsoidal predictions of the regulatory model and the results from microarray experiments and environmental measurements, however, in a set-theoretic sense. They depend on the distance of cluster centers and on nonnegative criteria functions which measure, e.g., the sum of squares of semiaxes (which corresponds to the trace of the configuration matrix) or the length of the largest semiaxes (which corresponds to the eigenvalues of the shape matrix). We note that semi-definite programming and interior point methods can be applied for solution.

In general, gene-environment networks comprise thousands of genes and additional factors. For this reason, the underlying network has a high number of branches. Often the connections between the clusters are weak so that the related contribution to the system is negligible. In order to reduce complexity we delete weak connections. This could be achieved by introducing bounds on the number of incoming branches. By imposing such additional constraints we obtain mixed integer regression problems. Since these constraints are very strict, we turn to a further relaxation that could be achieved by replacing binary constraints with continuous constraints leading to regression models based on continuous optimization.

The Chapter is organized as follows: In Section 2, we review basic facts about ellipsoidal calculus required for a representation of the dynamic states of multivariate noise prone data. In Section 3, the time-dependent regulatory model for overlapping groups of genes and environmental factors under ellipsoidal uncertainty is introduced. In addition, we provide an algorithm that allows to calculate predictions of future states of groups of genes and environmental items in terms of centers and shape matrices of the corresponding ellipsoids. In Section 4, we turn to a set-theoretic regression analysis for parameter estimation of the dynamic (ellipsoidal) system. Various regression models are introduced and we discuss their solvability by means of semi-definite programming. Finally, we address a reduction of complexity by network rarefication in Section 5, where we further extend the dynamic model and discuss related mixed integer approximation and a relaxation based on continuous optimization.

2 Ellipsoidal Calculus

The time-dependent multivariate states of the gene-environment network under consideration will be represented in terms of ellipsoidal sets. Predictions of the future ellipsoidal states are calculated with a time-discrete model based on *ellipsoidal calculus*. Here, we shortly review the basic operations of ellipsoidal calculus such as *sums*, *intersections (fusions)* and *affine-linear transformations* of ellipsoids. The family of ellipsoids in $\mathbb{R}^p$ is closed with respect to affine-linear transformations but neither the sum nor the intersection is generally ellipsoidal, so both must be approximated by ellipsoidal sets.

2.1 Ellipsoidal Descriptions

An *ellipsoid* in $\mathbb{R}^p$ will be parameterized in terms of its center $c \in \mathbb{R}^p$ and a symmetric non-negative definite *configuration (or shape) matrix* $\Sigma \in \mathbb{R}^{p \times p}$ as

$$\mathscr{E}(c,\Sigma) = \{\Sigma^{1/2}u + c \,|\, \|u\| \leq 1\},$$

where $\Sigma^{1/2}$ is any matrix square root satisfying $\Sigma^{1/2}(\Sigma^{1/2})^T = \Sigma$. When Σ is of full rank, the non-degenerate ellipsoid $\mathscr{E}(c,\Sigma)$ may be expressed as

$$\mathscr{E}(c,\Sigma) = \{x \in \mathbb{R}^p \,|\, (x-c)^T \Sigma^{-1}(x-c) \leq 1\}.$$

The eigenvectors of Σ point in the directions of principal semiaxes of $\mathscr{E}$. The lengths of the semiaxes of the ellipsoid $\mathscr{E}(c,\Sigma)$ are given by $\sqrt{\lambda_i}$, where λ_i are the eigenvalues of Σ for $i = 1,\ldots,p$. The volume of the ellipsoid $\mathscr{E}(c,\Sigma)$ is given by $\mathrm{vol}\,\mathscr{E}(c,\Sigma) = V_p\sqrt{\det(\Sigma)}$, where V_p is the volume of the unit ball in $\mathbb{R}^p$, i.e.,

$$V_p = \begin{cases} \dfrac{\pi^{p/2}}{(p/2)!} & \text{, for even } p, \\[2ex] \dfrac{2^p \pi^{(p-1)/2}((p-1)/2)!}{p!} & \text{, for odd } p. \end{cases}$$

2.2 Affine Transformations

The family of ellipsoids is closed with respect to *affine transformations*. Given an ellipsoid $\mathscr{E}(c,\Sigma) \subset \mathbb{R}^p$, matrix $A \in \mathbb{R}^{m \times p}$ and vector $b \in \mathbb{R}^m$ we get $A\mathscr{E}(c,\Sigma) + b = \mathscr{E}(Ac+b, A\Sigma A^T)$. Thus, ellipsoids are preserved under affine transformation. If the rows of A are linearly independent (which implies $m \leq p$), and $b = 0$, the affine transformation is called *projection* [34].

2.3 Sums of K Ellipsoids

Given K bounded ellipsoids of $\mathbb{R}^p$, $\mathscr{E}_k = \mathscr{E}(c_k,\Sigma_k)$, $k = 1,\ldots,K$, their *geometric (Minkowksi) sum* $\mathscr{E}_1 + \mathscr{E}_1 = \{z_1 + z_2 \,|\, z_1 \in \mathscr{E}_1,\, z_2 \in \mathscr{E}_2\}$ is not generally an ellipsoid. However, it can be tightly approximated by parameterized families of external ellipsoids. We adapt the notion of the minimal trace ellipsoid from [20] and introduce the outer ellipsoidal approximation $\mathscr{E}(\sigma,P) = \oplus_{k=1}^K \mathscr{E}_k$ containing the sum $\mathscr{S} = \sum_{k=1}^K \mathscr{E}_k$ of ellipsoids which is defined by

$$\sigma = \sum_{k=1}^{K} c_k$$

and

$$P = \left(\sum_{k=1}^{K} \sqrt{\operatorname{Tr} \Sigma_k} \right) \left(\sum_{k=1}^{K} \frac{\Sigma_k}{\sqrt{\operatorname{Tr} \Sigma_k}} \right).$$

2.4 Intersection of Ellipsoids

The intersection of two ellipsoids is generally not an ellipsoid. For this reason we replace this set by the outer ellipsoidal approximation of minimal volume and adapt the notion of *fusion* of ellipsoids from [49]. Given two non-degenerate ellipsoids $\mathscr{E}(c_1, \Sigma_1)$ and $\mathscr{E}(c_2, \Sigma_2)$ in $\mathbb{R}^p$ with $\mathscr{E}(c_1, \Sigma_1) \cap \mathscr{E}(c_2, \Sigma_2) \neq \emptyset$ we define an ellipsoid

$$\begin{aligned} \mathscr{E}_\lambda(c_0, \Sigma_0) := \{ x \in \mathbb{R}^p \mid & \lambda (x - c_1)^T \Sigma_1^{-1} (x - c_1) \\ & + (1 - \lambda)(x - c_2)^T \Sigma_2^{-1} (x - c_2) \leq 1 \}, \end{aligned}$$

where $\lambda \in [0, 1]$. The ellipsoid $\mathscr{E}_\lambda(c_0, \Sigma_0)$ coincides with $\mathscr{E}(c_1, \Sigma_1)$ and $\mathscr{E}(c_2, \Sigma_2)$ for $\lambda = 1$ and $\lambda = 0$, respectively. In order to determine a tight external ellipsoidal approximation $\mathscr{E}_\lambda(c_0, \Sigma_0)$ of the intersection of $\mathscr{E}(c_1, \Sigma_1)$ and $\mathscr{E}(c_2, \Sigma_2)$, we introduce

$$\mathscr{X} := \lambda \Sigma_1^{-1} + (1 - \lambda) \Sigma_2^{-1}$$

and

$$\tau := 1 - \lambda (1 - \lambda)(c_2 - c_1)^T \Sigma_2^{-1} \mathscr{X}^{-1} \Sigma_1^{-1} (c_2 - c_1).$$

The ellipsoid $\mathscr{E}_\lambda(c_0, \Sigma_0)$ is given by the center

$$c_0 = \mathscr{X}^{-1} (\lambda \Sigma_1^{-1} c_1 + (1 - \lambda) \Sigma_2^{-1} c_2)$$

and shape matrix

$$\Sigma_0 = \tau \mathscr{X}^{-1}.$$

The *fusion* of $\mathscr{E}(c_1, \Sigma_1)$ and $\mathscr{E}(c_2, \Sigma_2)$, whose intersection is a nonempty bounded region, is defined as the ellipsoid $\mathscr{E}_\lambda(c_0, \Sigma_0)$ for the value $\lambda \in [0, 1]$ that minimizes its volume [49]. The fusion of $\mathscr{E}(c_1, \Sigma_1)$ and $\mathscr{E}(c_2, \Sigma_2)$ is $\mathscr{E}(c_1, \Sigma_1)$, if $\mathscr{E}(c_1, \Sigma_1) \subset \mathscr{E}(c_2, \Sigma_2)$; or $\mathscr{E}(c_2, \Sigma_2)$, if $\mathscr{E}(c_2, \Sigma_2) \subset \mathscr{E}(c_1, \Sigma_1)$; otherwise, it is $\mathscr{E}_\lambda(c_0, \Sigma_0)$ defined as above where λ is the only root in $(0, 1)$ of the following polynomial of degree $2p - 1$:

$$\begin{aligned} & \tau (\det \mathscr{X}) \operatorname{Tr} (\operatorname{co}(\mathscr{X})(\Sigma_1^{-1} - \Sigma_2^{-1})) - p (\det \mathscr{X})^2 \\ & \times (2 c_0^T \Sigma_1^{-1} c_1 - 2 c_0^T \Sigma_2^{-1} c_2 + c_0^T (\Sigma_2^{-1} - \Sigma_1^{-1}) c_0 - c_1^T \Sigma_1^{-1} c_1 + c_2^T \Sigma_2^{-1} c_2) = 0. \end{aligned}$$

Here, $\operatorname{co}(\mathscr{X})$ denotes the matrix of cofactors of $\mathscr{X}$. Since $\mathscr{X}^{-1} = \operatorname{co}(\mathscr{X}) / \det \mathscr{X}$, we represent this polynomial as

$$\tau(\det \mathscr{X})^2 \operatorname{Tr}(\mathscr{X}^{-1}(\Sigma_1^{-1} - \Sigma_2^{-1})) - p(\det \mathscr{X})^2$$
$$\times (2c_0^T \Sigma_1^{-1} c_1 - 2c_0^T \Sigma_2^{-1} c_2 + c_0^T (\Sigma_2^{-1} - \Sigma_1^{-1}) c_0 - c_1^T \Sigma_1^{-1} c_1 + c_2^T \Sigma_2^{-1} c_2) = 0.$$

We note that it is also possible to define an inner ellipsoidal approximation. The method of finding the internal ellipsoidal approximation of the intersection of two ellipsoids is described in [61].

3 Gene-Environment Systems under Ellipsoidal Uncertainty

3.1 Clusters of Gene-Environment Data

Various approaches from clustering and classification can be applied to analyze the structure of gene-environment networks. In this way, certain groups of genes and environmental factors can be identified which exert a more or less regulating influence on other groups of data items. Usually these groups cannot be divided unambiguously since a single gene can have a regulating effect on various clusters of genes and, thus, belongs to different clusters. In addition, the quality of the available data sets may not be sufficient for an identification of disjoint groups. For this reason, we assume that in the preprocessing step of clustering a number of *overlapping* clusters of genes and environmental factors can be identified. Such a partition can be achieved for example with one of the many variants of *fuzzy-c-means clustering* [57]. The specific gene-environment network under consideration consists of n genes and m environmental factors, where the vector $\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_n)^T$ denotes the expression values of the genes and the vector $\mathbb{E} = (\mathbb{E}_1, \ldots, \mathbb{E}_m)^T$ stands for the values of the environmental factors. The set of genes is divided in R overlapping clusters $C_r \subset \{1, \ldots, n\}$, $r = 1, \ldots, R$ and the set of all environmental items is divided in S overlapping clusters $D_s \subset \{1, \ldots, m\}$, $s = 1, \ldots, S$. We note that the paper [31] focussed on disjoint clusters assuming a strict sub-division of the variables where the relations $C_{r_1} \cap C_{r_2} = \emptyset$ for all $r_1 \neq r_2$ and $D_{s_1} \cap D_{s_2} = \emptyset$ for all $s_1 \neq s_2$ are fulfilled.

The *(crisp) states* of the elements of these clusters are given by subsets of the vectors $\mathbb{X}$ and $\mathbb{E}$. That means, we assign a $|C_r|$-subvector X_r of $\mathbb{X}$ to each cluster of genes which is given by the indices of C_r. Similarly, E_s is a $|D_s|$-subvector of $\mathbb{E}$ given by the indices of D_s.

For a representation of the *uncertain states* of the aforementioned clusters we identify the clusters with error ellipsoids. That means, the vectors X_r represent ellipsoidal states of the genes in cluster C_r given by the ellipsoid $\mathscr{E}(\mu_r, \Sigma_r) \subset \mathbb{R}^{|C_r|}$ and E_s represent the ellipsoidal states of the environmental items in cluster D_s given by the ellipsoids $\mathscr{E}(\rho_s, \Pi_s) \subset \mathbb{R}^{|D_s|}$. The ellipsoid $\mathscr{E}(\mu_r, \Sigma_r)$ is characterized by $|C_r| + |C_r|^2$ coefficients and the ellipsoid $\mathscr{E}(\rho_s, \Pi_s)$ is determined by $|D_s| + |D_s|^2$ variables. The number of coefficients can be reduced by assuming symmetric shape matrices what refers to specific correlation of the data variables. We note that ellipsoids can be identified with intervals if clusters are singletons. It is also possible that some of the variables are exactly known. In this situation, the ellipsoids $\mathscr{E}(\mu_r, \Sigma_r)$ and $\mathscr{E}(\rho_s, \Pi_s)$ are flat. However, as we are interested in approximations we can avoid this by imposing lower bounds on the semiaxes lengths or an artificial extension in the corresponding coordinate directions of length $\varepsilon > 0$. Similarly, degenerate or needle-shaped ellipsoids can be avoided by imposing upper bounds on the extension of the semiaxes.

3.2 The Linear Model

In this section, we introduce a dynamic model that allows to predict the time-dependent (ellipsoidal) states of the clusters in the gene-environment regulatory network. This model is based on four types of cluster interactions and regulating effects:

(GG) genetic cluster regulates genetic cluster
(EG) environmental cluster regulates genetic cluster
(GE) genetic cluster regulates environmental cluster
(EE) environmental cluster regulates environmental cluster.

As shown in Section 3.1, each cluster corresponds to a functionally related group of genes or environmental factors and the uncertain states of these clusters are represented in terms of parameterized ellipsoids

$$X_r = \mathscr{E}(\mu_r, \Sigma_r) \subset \mathbb{R}^{|C_r|}, \quad E_s = \mathscr{E}(\rho_s, \Pi_s) \subset \mathbb{R}^{|D_s|}.$$

With the ellipsoidal calculus introduced in Section 2, the dynamics and interactions between the various clusters of genetic and environmental items is given by the linear model

$$\left.\begin{aligned} X_j^{(\kappa+1)} &= \xi_{j0} + \left(\bigoplus_{r=1}^{R} \mathscr{A}_{jr}^{GG} X_r^{(\kappa)}\right) + \left(\bigoplus_{s=1}^{S} \mathscr{A}_{js}^{EG} E_s^{(\kappa)}\right) \\ E_i^{(\kappa+1)} &= \zeta_{i0} + \left(\bigoplus_{r=1}^{R} \mathscr{A}_{ir}^{GE} X_r^{(\kappa)}\right) + \left(\bigoplus_{s=1}^{S} \mathscr{A}_{is}^{EE} E_s^{(\kappa)}\right) \end{aligned}\right\} \; (EC)$$

with $\kappa \geq 0$ and $j = 1, 2, \ldots, R$, $i = 1, 2, \ldots, S$. The system (EC) is defined by (affine) linear coupling rules, what implies that all future states of genetic and environmental clusters are ellipsoids themselves. In particular, the sums $\bigoplus_{r=1}^{R} \mathscr{A}_{jr}^{GG} X_r^{(\kappa)}$ and $\bigoplus_{s=1}^{S} \mathscr{A}_{js}^{EG} E_s^{(\kappa)}$ describe the *cumulative effects* of all genetic and environmental clusters exerted on the elements of cluster C_j in a set theoretic or ellipsoidal sense. In the same way, the (ellipsoidal) sums $\bigoplus_{r=1}^{R} \mathscr{A}_{ir}^{GE} X_r^{(\kappa)}$ and $\bigoplus_{s=1}^{S} \mathscr{A}_{is}^{EE} E_s^{(\kappa)}$ refer to the additive genetic and environmental effects on cluster D_i. The *degree of connectivity* between the individual clusters is given by the (unknown) interactions matrices $\mathscr{A}_{jr}^{GG} \in \mathbb{R}^{|C_j|\times|C_r|}$, $\mathscr{A}_{js}^{EG} \in \mathbb{R}^{|C_j|\times|D_s|}$, $\mathscr{A}_{ir}^{GE} \in \mathbb{R}^{|D_i|\times|C_r|}$, and $\mathscr{A}_{is}^{EE} \in \mathbb{R}^{|D_i|\times|D_s|}$. These matrices are in turn sub-matrices of the general interaction matrices $\mathscr{A}^{GG} \in \mathbb{R}^{n\times n}$, $\mathscr{A}^{EG} \in \mathbb{R}^{n\times m}$, $\mathscr{A}^{GE} \in \mathbb{R}^{m\times n}$, $\mathscr{A}^{EE} \in \mathbb{R}^{m\times m}$. In case of disjoint clusters, the aforementioned sub-matrices constitute distinct building blocks of the interaction matrices [31, 32]. For overlapping clusters, the structure of the interaction matrices is much more complicated; it reflects the cluster structure and the sub-matrices are partly composed of the same elements. This also holds for the intercepts $\xi_{j0} \in \mathbb{R}^{|C_j|}$ and $\zeta_{i0} \in \mathbb{R}^{|D_i|}$ which are partly overlapping sub-vectors of the vectors $\xi_0 = (\xi_{10},\ldots,\xi_{n0})^T \in \mathbb{R}^n$ and $\zeta_0 = (\zeta_{10},\ldots,\zeta_{m0})^T \in \mathbb{R}^m$, respectively. We note that the initial values of the linear system (EC) can be defined by the first genetic and environmental measurements, i.e., $X_j^{(0)} = \overline{X}_j^{(0)}$ and $E_i^{(0)} = \overline{E}_i^{(0)}$.

The unknown parameters of the linear model (EC) have to be determined by a regression analysis based on ellipsoidal data sets. Since all matrices and vectors of (EC) are parts of the general interaction matrices and intercepts, $n^2 + 2nm + m^2 + n + m = (n+m)^2 + n + m$ unknown parameters have to be determined. We will provide more details on regression analysis in Section 4.

REMARK *(Gene-environment networks).* The clusters of genes and environmental factors can be considered as the nodes of a so-called *gene-environment network*. Such a network usually consists of a high number of branches what refers to the inherent connections between the clusters. The branches between the nodes (or clusters) are weighted by the matrices and intercept vectors of the linear coupling rules of model (EC) and the nodes are weighted by the time-dependent ellipsoidal states of the clusters. Hereby, network analysis and concepts from discrete mathematics become applicable and features like connectedness, cycles and shortest paths can be investigated [29].

3.3 Algorithm

The regulatory system (EC) allows to predict the ellipsoidal states of genes and environmental factors with the set-theoretic calculus introduced in Section 2. In order to avoid set-valued calculations we propose to determine the centers and shape matrices of the predictions $X_j^{(\kappa+1)}$ and $E_s^{(\kappa+1)}$ of (ellipsoidal) genetic and environmental cluster states by an iterative procedure. Throughout this section we assume $\kappa \geq 0$. The states of the genetic clusters C_j, $j = 1,2,\ldots,R$, are given by the ellipsoids

$$X_j^{(\kappa+1)} = \mathscr{E}\left(\mu_j^{(\kappa+1)}, \Sigma_j^{(\kappa+1)}\right)$$

with center

$$\mu_j^{(\kappa+1)} = \xi_{j0} + \sum_{r=1}^{R} A_{jr}^{GG} \mu_r^{(\kappa)} + \sum_{s=1}^{S} A_{js}^{EG} \rho_s^{(\kappa)}$$

and shape matrix

$$\Sigma_j^{(\kappa+1)} = \left(\sqrt{\operatorname{Tr} \mathscr{G}_j^{(\kappa)}} + \sqrt{\operatorname{Tr} \mathscr{H}_j^{(\kappa)}} \right) \cdot \left(\frac{\mathscr{G}_j^{(\kappa)}}{\sqrt{\operatorname{Tr} \mathscr{G}_j^{(\kappa)}}} + \frac{\mathscr{H}_j^{(\kappa)}}{\sqrt{\operatorname{Tr} \mathscr{H}_j^{(\kappa)}}} \right),$$

where

$$\mathscr{G}_j^{(\kappa)} = \left(\sum_{r=1}^{R} \sqrt{\operatorname{Tr} G_{jr}^{GG}} \right) \cdot \left(\sum_{r=1}^{R} \frac{G_{jr}^{GG}}{\sqrt{\operatorname{Tr} G_{jr}^{GG}}} \right),$$

$$\mathscr{H}_j^{(\kappa)} = \left(\sum_{s=1}^{S} \sqrt{\operatorname{Tr} H_{js}^{EG}} \right) \cdot \left(\sum_{s=1}^{S} \frac{H_{js}^{EG}}{\sqrt{\operatorname{Tr} H_{js}^{EG}}} \right)$$

and

$$G_{jr}^{GG} = A_{jr}^{GG} \Sigma_r^{(\kappa)} (A_{jr}^{GG})^T, \quad H_{js}^{EG} = A_{js}^{EG} \Pi_s^{(\kappa)} (A_{js}^{EG})^T.$$

Similarly, the states of the environmental cluster D_i, $i = 1, 2, \ldots, S$, can be represented in terms of ellipsoids

$$E_i^{(\kappa+1)} = \mathscr{E}\left(\rho_i^{(\kappa+1)}, \Pi_i^{(\kappa+1)} \right)$$

with center

$$\rho_i^{(\kappa+1)} = \zeta_{i0} + \sum_{r=1}^{R} A_{ir}^{GE} \mu_r^{(\kappa)} + \sum_{s=1}^{S} A_{is}^{EE} \rho_s^{(\kappa)}$$

and shape matrix

$$\Pi_i^{(\kappa+1)} = \left(\sqrt{\operatorname{Tr} \mathscr{M}_i^{(\kappa)}} + \sqrt{\operatorname{Tr} \mathscr{N}_i^{(\kappa)}} \right) \cdot \left(\frac{\mathscr{M}_i^{(\kappa)}}{\sqrt{\operatorname{Tr} \mathscr{M}_i^{(\kappa)}}} + \frac{\mathscr{N}_i^{(\kappa)}}{\sqrt{\operatorname{Tr} \mathscr{N}_i^{(\kappa)}}} \right),$$

where

$$\mathscr{M}_i^{(\kappa)} = \left(\sum_{r=1}^{R} \sqrt{\operatorname{Tr} M_{ir}^{GE}} \right) \cdot \left(\sum_{r=1}^{R} \frac{M_{ir}^{GE}}{\sqrt{\operatorname{Tr} M_{ir}^{GE}}} \right),$$

$$\mathscr{N}_i^{(\kappa)} = \left(\sum_{s=1}^{S} \sqrt{\operatorname{Tr} N_{is}^{EE}} \right) \cdot \left(\sum_{s=1}^{S} \frac{N_{is}^{EE}}{\sqrt{\operatorname{Tr} N_{is}^{EE}}} \right)$$

and

$$M_{ir}^{GE} = A_{ir}^{GE} \Sigma_r^{(\kappa)} (A_{ir}^{GE})^T, \quad N_{is}^{EE} = A_{is}^{EE} \Pi_s^{(\kappa)} (A_{is}^{EE})^T.$$

4 Regression Analysis Under Ellipsoidal Uncertainty

4.1 The Regression Problem

The linear model (EC) depends on $(n+m)^2+n+m$ unknown parameters which define the system dynamics but also the strength of the interconnections between the genetic and environmental clusters. In this section, we introduce our main regression model for an estimation of these parameters and, thus, of the entries of the interaction matrices $\mathscr{A}_{jr}^{GG}$, $\mathscr{A}_{js}^{EG}$, $\mathscr{A}_{ir}^{GE}$, $\mathscr{A}_{is}^{EE}$ as well as the intercepts ξ_{j0} and ζ_{i0} for overlapping clusters. Since the model (EC) is based on ellipsoidal sets, a set-theoretic regression analysis has to be established. The input data is given in terms of ellipsoidal genetic and environmental observations

$$\overline{X}_r^{(\kappa)} = \mathscr{E}\big(\overline{\mu}_r^{(\kappa)}, \overline{\Sigma}_r^{(\kappa)}\big) \subset \mathbb{R}^{|C_r|}, \quad \overline{E}_s^{(\kappa)} = \mathscr{E}\big(\overline{\rho}_s^{(\kappa)}, \overline{\Pi}_s^{(\kappa)}\big) \subset \mathbb{R}^{|D_s|},$$

with $r=1,2,\ldots,R$, $s=1,2,\ldots,S$ and $\kappa=0,1,\ldots,T$ which are taken at sampling times $t_0<t_1<\ldots<t_T$. These measurements have to be compared with the first T predictions of the model (EC) given by

$$\widehat{X}_j^{(\kappa+1)} = \mathscr{E}\big(\widehat{\mu}_j^{(\kappa+1)}, \widehat{\Sigma}_j^{(\kappa+1)}\big) := \xi_{j0} + \left(\bigoplus_{r=1}^{R} \mathscr{A}_{jr}^{GG}\overline{X}_r^{(\kappa)}\right) + \left(\bigoplus_{s=1}^{S} \mathscr{A}_{js}^{EG}\overline{E}_s^{(\kappa)}\right),$$

$$\widehat{E}_i^{(\kappa+1)} = \mathscr{E}\big(\widehat{\rho}_i^{(\kappa+1)}, \widehat{\Pi}_i^{(\kappa+1)}\big) := \zeta_{i0} + \left(\bigoplus_{r=1}^{R} \mathscr{A}_{ir}^{GE}\overline{X}_r^{(\kappa)}\right) + \left(\bigoplus_{s=1}^{S} \mathscr{A}_{is}^{EE}\overline{E}_s^{(\kappa)}\right),$$

with $j=1,2,\ldots,R$, $i=1,2,\ldots,S$ and $\kappa=0,1,\ldots,T-1$.

In our set-theoretic regression, we try to maximize the overlap of the predictions and measurement values (both ellipsoids). For this reason, we introduce the ellipsoidal approximation of the intersection given by

$$\Delta X_r^{(\kappa)} := \widehat{X}_r^{(\kappa)} \cap \overline{X}_r^{(\kappa)} \quad \text{and} \quad \Delta E_s^{(\kappa)} := \widehat{E}_s^{(\kappa)} \cap \overline{E}_s^{(\kappa)},$$

with $r=1,2,\ldots,R$, $s=1,2,\ldots,S$ and $\kappa=1,\ldots,T$, where $\cap$ denotes the fusion of ellipsoids introduced in Subsection 2.4. In addition, the centers of the ellipsoids are adjusted, so that their squared distance

$$\left\| \widehat{\mu}_r^{(\kappa)} - \overline{\mu}_r^{(\kappa)} \right\|_2^2 \quad \text{and} \quad \left\| \widehat{\rho}_s^{(\kappa)} - \overline{\rho}_s^{(\kappa)} \right\|_2^2$$

becomes minimized (cf. Figure 1). This leads us to the following regression problem:

$$(R) \quad \text{Maximize} \sum_{\kappa=1}^{T} \left\{ \sum_{r=1}^{R} \left[\left\| \Delta X_r^{(\kappa)} \right\|_* - \left\| \widehat{\mu}_r^{(\kappa)} - \overline{\mu}_r^{(\kappa)} \right\|_2^2 \right] + \sum_{s=1}^{S} \left[\left\| \Delta E_s^{(\kappa)} \right\|_* - \left\| \widehat{\rho}_s^{(\kappa)} - \overline{\rho}_s^{(\kappa)} \right\|_2^2 \right] \right\}.$$

Here, $\|\cdot\|_*$ denotes a measure that reflects the geometrical size of the intersections (fusions) and we assume that $\|\Delta X_r^{(\kappa)}\|_* = 0$, if $\Delta X_r^{(\kappa)} = \emptyset$ and $\|\Delta E_s^{(\kappa)}\|_* = 0$, if $\Delta E_s^{(\kappa)} = \emptyset$. There exist various measures related to the shape of the intersections, e.g., the *volume* (which corresponds to the ellipsoid matrix determinant), the *sum of squares of semiaxes* (which corresponds to the trace of the configuration matrix) or the *length of the largest semiaxes* (which corresponds to the eigenvalues of the shape matrix). For further details on geometrical (ellipsoidal) measures and the related regression problems we refer to [32].

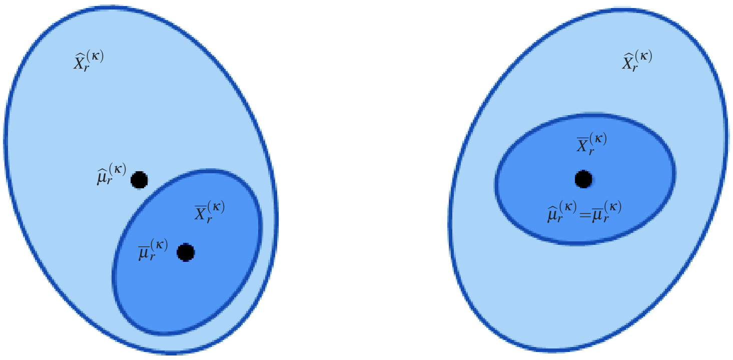
$\widehat{X}_r^{(\kappa)}$
$\widehat{\mu}_r^{(\kappa)}$
$\overline{X}_r^{(\kappa)}$
$\overline{\mu}_r^{(\kappa)}$
$\widehat{X}_r^{(\kappa)}$
$\overline{X}_r^{(\kappa)}$
$\widehat{\mu}_r^{(\kappa)}=\overline{\mu}_r^{(\kappa)}$

Fig. 1 Overlap of ellipsoids: The intersections of the two ellipsoids $\widehat{X}_r^{(\kappa)}$ and $\overline{X}_r^{(\kappa)}$ have the same geometrical size with the same measure of fusions on the left and the right side. On the right side, the centers $\widehat{\mu}_r^{(\kappa)}$ and $\overline{\mu}_r^{(\kappa)}$ are adjusted in order to minimize the difference between the centers of ellipsoids.

4.2 Variants of the Regression Problem

In this section, we introduce specific formulations of the regression model (R). As mentioned above, the objective function of this model depends on a measure of the geometrical size of the intersections (fusions) $\Delta X_r^{(\kappa)}$ and $\Delta E_s^{(\kappa)}$ which is related to the corresponding shape matrices. In general, nonnegative-valued criteria functions $\psi(\mathscr{E}(0,Q))$ defined on the set of all nondegenerate ellipsoids can be applied to measure the size of a p-dimensional ellipsoid $\mathscr{E}(0,Q)$. These functions are monotonous by increasing with respect to inclusion, i.e., $\psi(\mathscr{E}_1) \leq \psi(\mathscr{E}_2)$ if $\mathscr{E}_1 \subseteq \mathscr{E}_2$. Such measures are, e.g.,

(a) *the trace of Q*,

$$\psi_T(\mathscr{E}(0,Q)) := \operatorname{Tr} Q = \lambda_1 + \ldots + \lambda_p,$$

where λ_i are the eigenvalues of Q (i.e., $\operatorname{Tr} Q$ is equal to the sum of the squares of the semiaxes),

(b) *the trace of square of Q*,

$$\psi_{TS}(\mathscr{E}(0,Q)) := \operatorname{Tr} Q^2,$$

(c) *the diameter*,

$$\psi_{Dia}(\mathscr{E}(0,Q)) := \operatorname{diam}(\mathscr{E}(0,Q)) := d,$$

where

$$\max\{\lambda_i \in \mathbb{R} \,|\, i = 1,\ldots,p\} = \left(\frac{d}{2}\right)^2,$$

so that $d/2$ is the radius of the smallest p-dimensional ball that includes $\mathscr{E}(0,Q)$.

For further details on criteria functions we refer to [33], p. 101. The measures stated above lead to different representations of the regression problem (R) and in the following sections we study them in more detail.

Remark *(Representation of fusions).* For numerical calculations and an estimation of parameters of the regression problem (R), explicit representations of the fusions $\Delta X_r^{(\kappa)}$ and $\Delta E_s^{(\kappa)}$ are required. In the following we explain how these representations can be calculated with the ellipsoidal calculus of Section 2:

The fusion $\Delta X_r^{(\kappa)} = \widehat{X}_r^{(\kappa)} \cap \overline{X}_{C_r}^{(\kappa)}$ is an ellipsoid $\mathscr{E}\big(\Delta\mu_r^{(\kappa)}, \Delta\Sigma_r^{(\kappa)}\big)$ with center

$$\Delta\mu_r^{(\kappa)} = \big[\mathscr{X}_r^{(\kappa)}\big]^{-1}\big(\lambda\,\big[\widehat{\Sigma}_r^{(\kappa)}\big]^{-1}\widehat{\mu}_r^{(\kappa)} + (1-\lambda)\,\big[\overline{\Sigma}_r^{(\kappa)}\big]^{-1}\overline{\mu}_r^{(\kappa)}\big)$$

and shape matrix

$$\Delta\Sigma_r^{(\kappa)} = \xi_r^{(\kappa)}\,\big[\mathscr{X}_r^{(\kappa)}\big]^{-1},$$

where

$$\mathscr{X}_r^{(\kappa)} := \lambda\,\big[\widehat{\Sigma}_r^{(\kappa)}\big]^{-1} + (1-\lambda)\,\big[\overline{\Sigma}_r^{(\kappa)}\big]^{-1}$$

and

$$\xi_r^{(\kappa)} := 1 - \lambda(1-\lambda)\big(\overline{\mu}_r^{(\kappa)} - \widehat{\mu}_r^{(\kappa)}\big)^T\,\big[\overline{\Sigma}_r^{(\kappa)}\big]^{-1}\big[\mathscr{X}_r^{(\kappa)}\big]^{-1}\big[\widehat{\Sigma}_r^{(\kappa)}\big]^{-1}\big(\overline{\mu}_r^{(\kappa)} - \widehat{\mu}_r^{(\kappa)}\big).$$

The parameter λ is the only root in $(0,1)$ of the following polynomial of degree $2|C_r|-1$:

$$\begin{aligned}
&\xi_r^{(\kappa)}\big(\det \mathscr{X}_r^{(\kappa)}\big)^2 \operatorname{Tr}\Big(\big[\mathscr{X}_r^{(\kappa)}\big]^{-1}\Big(\big[\widehat{\Sigma}_r^{(\kappa)}\big]^{-1}-\big[\overline{\Sigma}_r^{(\kappa)}\big]^{-1}\Big)\Big)-|C_r|\big(\det \mathscr{X}_r^{(\kappa)}\big)^2\\
&\times\Big(2\big[\Delta\mu_r^{(\kappa)}\big]^T\big[\widehat{\Sigma}_r^{(\kappa)}\big]^{-1}\widehat{\mu}_r^{(\kappa)}-2\big[\Delta\mu_r^{(\kappa)}\big]^T\big[\overline{\Sigma}_r^{(\kappa)}\big]^{-1}\overline{\mu}_r^{(\kappa)}\\
&+\big[\Delta\mu_r^{(\kappa)}\big]^T\Big(\big[\overline{\Sigma}_r^{(\kappa)}\big]^{-1}-\big[\widehat{\Sigma}_r^{(\kappa)}\big]^{-1}\Big)\Delta\mu_r^{(\kappa)}-\big[\widehat{\mu}_r^{(\kappa)}\big]^T\big[\widehat{\Sigma}_r^{(\kappa)}\big]^{-1}\widehat{\mu}_r^{(\kappa)}\\
&+\big[\overline{\mu}_r^{(\kappa)}\big]^T\big[\overline{\Sigma}_r^{(\kappa)}\big]^{-1}\overline{\mu}_r^{(\kappa)}\Big)=0.
\end{aligned}$$

Similarly, the fusion $\Delta E_s^{(\kappa)}=\widehat{E}_s^{(\kappa)}\cap\overline{E}_s^{(\kappa)}$ is an ellipsoid $\mathscr{E}\big(\Delta\rho_s^{(\kappa)},\Delta\Pi_s^{(\kappa)}\big)$ with center

$$\Delta\rho_s^{(\kappa)}=\big[\mathscr{Y}_s^{(\kappa)}\big]^{-1}\big(\lambda\big[\widehat{\Pi}_s^{(\kappa)}\big]^{-1}\widehat{\rho}_s^{(\kappa)}+(1-\lambda)\big[\overline{\Pi}_s^{(\kappa)}\big]^{-1}\overline{\rho}_s^{(\kappa)}\big)$$

and shape matrix

$$\Delta\Pi_s^{(\kappa)}=\eta_s^{(\kappa)}\big[\mathscr{Y}_s^{(\kappa)}\big]^{-1},$$

where

$$\mathscr{Y}_s^{(\kappa)}:=\lambda\big[\widehat{\Pi}_s^{(\kappa)}\big]^{-1}+(1-\lambda)\big[\overline{\Pi}_s^{(\kappa)}\big]^{-1}$$

and

$$\eta_s^{(\kappa)}:=1-\lambda(1-\lambda)\big(\overline{\rho}_s^{(\kappa)}-\widehat{\rho}_s^{(\kappa)}\big)^T\big[\overline{\Pi}_s^{(\kappa)}\big]^{-1}\big[\mathscr{Y}_s^{(\kappa)}\big]^{-1}\big[\widehat{\Pi}_s^{(\kappa)}\big]^{-1}\big(\overline{\rho}_s^{(\kappa)}-\widehat{\rho}_s^{(\kappa)}\big).$$

The parameter λ is the only root in $(0,1)$ of the following polynomial of degree $2|D_s|-1$:

$$\begin{aligned}
&\eta_s^{(\kappa)}\big(\det \mathscr{Y}_s^{(\kappa)}\big)^2 \operatorname{Tr}\Big(\big[\mathscr{Y}_s^{(\kappa)}\big]^{-1}\Big(\big[\widehat{\Pi}_s^{(\kappa)}\big]^{-1}-\big[\overline{\Pi}_s^{(\kappa)}\big]^{-1}\Big)\Big)-|D_s|\big(\det \mathscr{Y}_s^{(\kappa)}\big)^2\\
&\times\Big(2\big[\Delta\rho_s^{(\kappa)}\big]^T\big[\widehat{\Pi}_s^{(\kappa)}\big]^{-1}\widehat{\rho}_s^{(\kappa)}-2\big[\Delta\rho_s^{(\kappa)}\big]^T\big[\overline{\Pi}_s^{(\kappa)}\big]^{-1}\overline{\rho}_s^{(\kappa)}\\
&+\big[\Delta\rho_s^{(\kappa)}\big]^T\Big(\big[\overline{\Pi}_s^{(\kappa)}\big]^{-1}-\big[\widehat{\Pi}_s^{(\kappa)}\big]^{-1}\Big)\Delta\rho_s^{(\kappa)}-\big[\widehat{\rho}_s^{(\kappa)}\big]^T\big[\widehat{\Pi}_s^{(\kappa)}\big]^{-1}\widehat{\rho}_s^{(\kappa)}\\
&+\big[\overline{\rho}_s^{(\kappa)}\big]^T\big[\overline{\Pi}_s^{(\kappa)}\big]^{-1}\overline{\rho}_s^{(\kappa)}\Big)=0.
\end{aligned}$$

4.2.1 The Trace Criterion

We now turn to the specific formulations of the regression problem (R). The first criterion is based on the *traces of the shape matrices* of the fusions $\Delta X_r^{(\kappa)}$ and $\Delta E_s^{(\kappa)}$. The geometrical size of these ellipsoids is measured in terms of their (squared) lengths of semiaxes and, thus, the traces of the shape matrices $\Delta\Sigma_r^{(\kappa)}$ and $\Delta\Pi_s^{(\kappa)}$:

$$(R_{Tr}) \qquad \text{Maximize} \qquad \sum_{\kappa=1}^{T}\left\{\sum_{r=1}^{R}\left[\mathrm{Tr}\left(\Delta\Sigma_r^{(\kappa)}\right)-\sum_{j=1}^{|C_r|}\left(\widehat{\mu}_{r,j}^{(\kappa)}-\overline{\mu}_{r,j}^{(\kappa)}\right)^2\right]\right.$$
$$\left.+\sum_{s=1}^{S}\left[\mathrm{Tr}\left(\Delta\Pi_s^{(\kappa)}\right)-\sum_{i=1}^{|D_s|}\left(\widehat{\rho}_{s,i}^{(\kappa)}-\overline{\rho}_{s,i}^{(\kappa)}\right)^2\right]\right\}.$$

The trace of the shape matrix of an ellipsoid is equal to the sum of the squares of the semiaxes. For this reason, the regression problem takes the form

$$(R'_{Tr}) \qquad \text{Maximize} \qquad \sum_{\kappa=1}^{T}\left\{\sum_{r=1}^{R}\sum_{j=1}^{|C_r|}\left[\lambda_{r,j}^{(\kappa)}-\left(\widehat{\mu}_{r,j}^{(\kappa)}-\overline{\mu}_{r,j}^{(\kappa)}\right)^2\right]\right.$$
$$\left.+\sum_{s=1}^{S}\sum_{i=1}^{|D_s|}\left[\Lambda_{s,i}^{(\kappa)}-\left(\widehat{\rho}_{s,i}^{(\kappa)}-\overline{\rho}_{s,i}^{(\kappa)}\right)^2\right]\right\},$$

where $\lambda_{r,j}^{(\kappa)}$ and $\Lambda_{s,i}^{(\kappa)}$ are the eigenvalues of $\Delta\Sigma_r^{(\kappa)}$ and $\Delta\Pi_s^{(\kappa)}$, respectively.

4.2.2 The Trace of the Square Criterion

When we measure the size of an ellipsoid by the *traces of the square* of its shape matrix, we obtain the following regression problem:

$$(R_{TS}) \qquad \text{Maximize} \qquad \sum_{\kappa=0}^{T}\left\{\sum_{r=1}^{R}\left[\mathrm{Tr}\left(\Delta\Sigma_r^{(\kappa)}\right)^2-\sum_{j=1}^{|C_r|}\left(\widehat{\mu}_{r,j}^{(\kappa)}-\overline{\mu}_{r,j}^{(\kappa)}\right)^2\right]\right.$$
$$\left.+\sum_{s=1}^{S}\left[\mathrm{Tr}\left(\Delta\Pi_s^{(\kappa)}\right)^2-\sum_{i=1}^{|D_s|}\left(\widehat{\rho}_{s,i}^{(\kappa)}-\overline{\rho}_{s,i}^{(\kappa)}\right)^2\right]\right\}.$$

4.2.3 The Diameter Criterion

The maximal extension of the fusions can be used to define a further regression model. Here, the *diameter* of the ellipsoids $\Delta X_r^{(\kappa)}$ and $\Delta E_s^{(\kappa)}$ (or the size of the smallest balls which include the fusions) is used in the objective function :

$$(R_{Dia}) \qquad \text{Maximize} \qquad \sum_{\kappa=0}^{T}\left\{\sum_{r=1}^{R}\left[\mathrm{diam}\left(\mathscr{E}\left(0,\Delta\Sigma_r^{(\kappa)}\right)\right)-\sum_{j=1}^{|C_r|}\left(\widehat{\mu}_{r,j}^{(\kappa)}-\overline{\mu}_{r,j}^{(\kappa)}\right)^2\right]\right.$$
$$\left.+\sum_{s=1}^{S}\left[\mathrm{diam}\left(\mathscr{E}\left(0,\Delta\Pi_s^{(\kappa)}\right)\right)-\sum_{i=1}^{|D_s|}\left(\widehat{\rho}_{s,i}^{(\kappa)}-\overline{\rho}_{s,i}^{(\kappa)}\right)^2\right]\right\}.$$

An equivalent formulation of (R_{Dia}) can be given in terms of the eigenvalues of $\Delta\Sigma_r^{(\kappa)}$ and $\Delta\Pi_s^{(\kappa)}$:

$$(R'_{Dia}) \qquad \text{Maximize} \qquad \begin{aligned} &\sum_{\kappa=1}^{T} \Bigg\{ \sum_{r=1}^{R} \Bigg[2 \cdot \sqrt{\lambda_r^{(\kappa)}} - \sum_{j=1}^{|C_r|} \big(\widehat{\mu}_{r,j}^{(\kappa)} - \overline{\mu}_{r,j}^{(\kappa)}\big)^2 \Bigg] \\ &\qquad + \sum_{s=1}^{S} \Bigg[2 \cdot \sqrt{\Lambda_s^{(\kappa)}} - \sum_{i=1}^{|D_s|} \big(\widehat{\rho}_{s,i}^{(\kappa)} - \overline{\rho}_{s,i}^{(\kappa)}\big)^2 \Bigg\} \end{aligned}$$

with $\lambda_r^{(\kappa)} := \max\{\lambda_{r,j}^{(\kappa)} \mid j = 1, \ldots, |C_r|\}$ and $\Lambda_s^{(\kappa)} := \max\{\Lambda_{s,i}^{(\kappa)} \mid i = 1, \ldots, |D_s|\}$. As the objective function of (R'_{Dia}) is nonsmooth with well-understood max-type functions [64, 65, 66] but not Lipschitz-continuous, we also introduce the additional regression problem

$$(R''_{Dia}) \qquad \text{Maximize} \qquad \begin{aligned} &\sum_{\kappa=0}^{T} \Bigg\{ \sum_{r=1}^{R} \Bigg[\lambda_r^{(\kappa)} - \sum_{j=1}^{|C_r|} \big(\widehat{\mu}_{r,j}^{(\kappa)} - \overline{\mu}_{r,j}^{(\kappa)}\big)^2 \Bigg] \\ &\qquad + \sum_{s=1}^{S} \Bigg[\Lambda_s^{(\kappa)} - \sum_{i=1}^{|D_s|} \big(\widehat{\rho}_{s,i}^{(\kappa)} - \overline{\rho}_{s,i}^{(\kappa)}\big)^2 \Bigg] \Bigg] \Bigg\} \end{aligned}$$

as an alternative proposal.

4.3 Optimization Methods

In this section, we summarize solution methods for the regression models of the previous subsections. The objective functions of these volume-related programming problems depend on, e.g., the eigenvalues of symmetric positive semidefinite shape matrices $\Delta\Sigma_r^{(\kappa)}$ and $\Delta\Pi_s^{(\kappa)}$ as well as the distance of the centers $\Delta\mu_r^{(\kappa)}$ and $\Delta\rho_s^{(\kappa)}$ of the fusions $\Delta X_r^{(\kappa)}$ and $\Delta E_s^{(\kappa)}$. For this reason, methods from semidefinite programming [14] can be applied. In particular, the regression model (R'_{Tr}) refers to sums of all eigenvalues of the shape matrices $\Delta\Sigma_r^{(\kappa)}$ and $\Delta\Pi_s^{(\kappa)}$. These objective functions can also be considered as positive semidefinite representable functions ([11], p. 80) and *interior point methods* can applied [39, 40, 41, 43]. Alternatively, asscociated *bilevel problems* can be introduced which could be solved by *gradient methods*. In fact, in [42] structural frontiers of conic programming are discussed with other optimization methods compared, and future applications in machine learning and data mining prepared. However, we would like to underline that in the areas regression and classification of statistical learning (cf. e.g., [25, 50]), our optimization based methods provided and further promise very good and competitive results [16, 28, 56, 70, 78].

5 Network Rarefication Based on Mixed Integer Programming and Continuous Programming

The gene-environment network of clusters defined by the interaction matrices of the linear model (EC) is usually highly-interconnected. The regression analysis of the previous sections allows an identification of the corresponding degrees of connectivity of all branches between the clusters, although its actual influence is small and negligible. In addition, a particular cluster is generally influenced by a limited (not too high) number of regulating genetic and environmental clusters and usually it regulates only a small number of clusters. For this reason, we introduce a method for diminishing the number of branches with the aim of network rarefication during the regression process. This goal could be achieved by introducing upper bounds on the indegrees and outdegrees of nodes of the gene-environment network. In other words, the number of clusters regulating a specific genetic or environmental cluster in our network as well as the number of clusters regulated by a particular cluster has to be bounded. We impose *binary constraints* in order to decide whether or not there is a connection between two clusters and by this we obtain a *mixed-integer optimization problem*. After this model is provided, we will pass to *continuous optimization* and introduce a model with *continuous constraints*. This is because of the exclusive nature of binary constraints that could even destroy the connectivity of the gene-environment network.

5.1 Mixed Integer Regression Problem

Given two clusters A, B we use the notation $A \leftarrow B$ if cluster A is regulated by cluster B and $A \not\leftarrow B$ if cluster A is not regulated by cluster B. Now, we define the Boolean matrices

$$\chi_{jr}^{GG} = \begin{cases} 1 & \text{, if } C_j \leftarrow C_r \\ 0 & \text{, if } C_j \not\leftarrow C_r, \end{cases} \qquad \chi_{js}^{EG} = \begin{cases} 1 & \text{, if } C_j \leftarrow D_s \\ 0 & \text{, if } C_j \not\leftarrow D_s, \end{cases}$$

$$\chi_{ir}^{GE} = \begin{cases} 1 & \text{, if } D_i \leftarrow C_r \\ 0 & \text{, if } D_i \not\leftarrow C_r, \end{cases} \qquad \chi_{is}^{EE} = \begin{cases} 1 & \text{, if } D_i \leftarrow D_s \\ 0 & \text{, if } D_i \not\leftarrow D_s, \end{cases}$$

indicating whether or not pairs of clusters in our regulatory network are directly related. If two clusters are not related, the corresponding parts of the matrices A^{GG}, A^{GE}, A^{EG}, A^{EE} have zero entries in case of a disjoint cluster decomposition.

The *indegree* of the genetic cluster C_j in our regulatory network is defined with respect to all genetic and environmental clusters by

$$\deg(C_j)_{in}^{GG} := \sum_{r=1}^{R} \chi_{jr}^{GG} \quad \text{and} \quad \deg(C_j)_{in}^{EG} := \sum_{s=1}^{S} \chi_{js}^{EG},$$

where $j \in \{1,\dots,R\}$. By this, the indegrees $\deg(C_j)^{GG}$ and $\deg(C_j)^{EG}$ count the number of genetic and environmental clusters which regulate cluster C_j. The *overall indegree* of the genetic cluster C_j is defined by

$$\deg(C_j)_{in} := \deg(C_j)_{in}^{GG} + \deg(C_j)_{in}^{EG}.$$

Similarly, for $i \in \{1,\dots,S\}$ the *indegree* of cluster D_i with respect to the environmental clusters and the genetic clusters is given by

$$\deg(D_i)_{in}^{GE} := \sum_{r=1}^{R} \chi_{ir}^{GE} \quad \text{and} \quad \deg(D_i)_{in}^{EE} := \sum_{s=1}^{S} \chi_{is}^{EE}.$$

In this way, the indegrees $\deg(D_i)^{GE}$ and $\deg(D_i)^{EE}$ count the number of genetic and environmental clusters which regulate cluster D_i. The *overall indegree* of cluster D_i is defined by

$$\deg(D_i)_{in} := \deg(D_i)_{in}^{GE} + \deg(D_i)_{in}^{EE}.$$

In the same way, bounds on the outdegree, i.e., the number of outgoing branches can be introduced. Firstly, binary values are defined to determine whether or not there is an outgoing connection:

$$\zeta_{jr}^{GG} = \begin{cases} 1 & \text{, if } C_r \leftarrow C_j \\ 0 & \text{, if } C_r \not\leftarrow C_j, \end{cases} \qquad \zeta_{js}^{EG} = \begin{cases} 1 & \text{, if } D_s \leftarrow C_j \\ 0 & \text{, if } D_s \not\leftarrow C_j, \end{cases}$$

$$\zeta_{ir}^{GE} = \begin{cases} 1 & \text{, if } C_r \leftarrow D_i \\ 0 & \text{, if } C_r \not\leftarrow D_i, \end{cases} \qquad \zeta_{is}^{EE} = \begin{cases} 1 & \text{, if } D_s \leftarrow D_i \\ 0 & \text{, if } D_s \not\leftarrow D_i. \end{cases}$$

Now, the *outdegree* of genetic cluster C_j with respect to all genetic and environmental clusters can be expressed as

$$\deg(C_j)_{out}^{GG} := \sum_{r=1}^{R} \zeta_{jr}^{GG} \quad \text{and} \quad \deg(C_j)_{out}^{GE} := \sum_{s=1}^{S} \zeta_{js}^{GE},$$

where $j \in \{1,\dots,R\}$. The outdegrees $\deg(C_j)_{out}^{GG}$ and $\deg(C_j)_{out}^{EG}$ count the number of genetic and environmental clusters *regulated by* cluster C_j. The *overall outdegree* of genetic cluster C_j is given by

$$\deg(C_j)_{out} := \deg(C_j)_{out}^{GG} + \deg(C_j)_{out}^{EG}.$$

The *outegree* of environmental cluster D_i with respect to the environmental clusters and the genetic clusters is defined by

$$\deg(D_i)_{out}^{GE} := \sum_{r=1}^{R} \zeta_{ir}^{GE} \quad \text{and} \quad \deg(D_i)_{out}^{EE} := \sum_{s=1}^{S} \zeta_{is}^{EE},$$

where $i \in \{1,\ldots,S\}$. The outdegrees $\deg(D_i)^{GE}_{out}$ and $\deg(D_i)^{EE}_{out}$ count the number of genetic and environmental clusters *regulated by* cluster D_i. The *overall outdegree* of cluster D_i is

$$\deg(D_i)_{out} := \deg(D_i)^{GE}_{out} + \deg(D_i)^{EE}_{out}.$$

As mentioned above, we introduce upper bounds on the indegrees and the outdegrees of the nodes (clusters) with the aim of network rarefication. These values have to be given by the practitioner and they can depend on any a priori information. Including these additional constraints, we obtain the following *mixed integer optimization problem*:

$$\textit{(MI1)} \quad \begin{cases} \text{Maximize} \displaystyle\sum_{\kappa=1}^{T} \left\{ \sum_{r=1}^{R} \left\| \Delta X_r^{(\kappa)} \right\|_* - \left\| \widehat{\mu}_r^{(\kappa)} - \overline{\mu}_r^{(\kappa)} \right\|_2^2 \right. \\ \qquad\qquad\qquad \left. + \displaystyle\sum_{s=1}^{S} \left\| \Delta E_s^{(\kappa)} \right\|_* - \left\| \widehat{\rho}_s^{(\kappa)} - \overline{\rho}_s^{(\kappa)} \right\|_2^2 \right\} \\ \\ \text{subject to } \deg(C_j)^{GG}_{in} \le \alpha_j^{GG},\ j = 1,\ldots,R \\ \qquad\qquad\quad \deg(C_j)^{EG}_{in} \le \alpha_j^{EG},\ j = 1,\ldots,R \\ \qquad\qquad\quad \deg(D_i)^{EG}_{in} \le \alpha_i^{EG},\ i\ = 1,\ldots,S \\ \qquad\qquad\quad \deg(D_i)^{EE}_{in} \le \alpha_i^{EE},\ i\ = 1,\ldots,S \\ \\ \qquad\qquad\quad \deg(C_j)^{GG}_{out} \le \beta_j^{GG},\ j = 1,\ldots,R \\ \qquad\qquad\quad \deg(C_j)^{EG}_{out} \le \beta_j^{EG},\ j = 1,\ldots,R \\ \qquad\qquad\quad \deg(D_i)^{GE}_{out} \le \beta_i^{GE},\ i\ = 1,\ldots,S \\ \qquad\qquad\quad \deg(D_i)^{EE}_{out} \le \beta_i^{EE},\ i\ = 1,\ldots,S. \end{cases}$$

In model *(MI1)*, individual bounds on the indegrees and outdegrees of each genetic and environmental cluster are imposed what allows us to control the connectivity of the gene-environment network. This approach is an extension of the regression problems with bounds on the indegrees in [31, 32]. Similar mixed-integer problems for an analysis of gene-environment networks based on interval arithmetics were presented in [68, 69, 71].

In model *(MI1)* the number of connections of each cluster with genetic and environmental clusters are considered separately. In a further step, we can combine these bounds and impose restrictions on the total number of all ingoing or outgoing branches of each cluster:

$$
\textit{(MI2)} \quad \begin{cases} \text{Maximize } \sum_{\kappa=1}^{T} \left\{ \sum_{r=1}^{R} \left\| \Delta X_r^{(\kappa)} \right\|_* - \left\| \widehat{\mu}_r^{(\kappa)} - \overline{\mu}_r^{(\kappa)} \right\|_2^2 \right. \\ \qquad\qquad \left. + \sum_{s=1}^{S} \left\| \Delta E_s^{(\kappa)} \right\|_* - \left\| \widehat{\rho}_s^{(\kappa)} - \overline{\rho}_s^{(\kappa)} \right\|_2^2 \right\} \\ \text{subject to } \deg(C_j)_{in} \le \gamma_j, \ j = 1, \dots, R \\ \qquad\qquad \deg(D_i)_{in} \le \delta_i, \ i = 1, \dots, S \\ \\ \qquad\qquad \deg(C_j)_{out} \le \varepsilon_j, \ j = 1, \dots, R \\ \qquad\qquad \deg(D_i)_{out} \le \varphi_i, \ i = 1, \dots, S. \end{cases}
$$

5.2 Continuous Programming

As mentioned above, the binary constraints in *(MI1)* and *(MI2)* can lead to restrictions of the network connectivity. For this reason, continuous optimization is applied for a relaxation of *(MI1)* by replacing the binary variables χ_{jr}^{GG}, χ_{js}^{EG}, χ_{ir}^{GE} and χ_{is}^{EE} with real variables $P_{jr}^{GG}, P_{js}^{EG}, P_{ir}^{GE}, P_{is}^{EE} \in [0,1]$, which is also interpretable as probabilities (we refer to [51] for optimization models with probabilistic constraints). These variables should linearly depend on the corresponding elements of $\mathscr{A}_{jr}^{GG}, \mathscr{A}_{js}^{EG}, \mathscr{A}_{ir}^{GE}, \mathscr{A}_{is}^{EE}$.

The real-valued *indegree* of cluster C_j in our regulatory network with respect to the genetic and environmental clusters are now defined by

$$
\deg(C_j)_{in}^{GG} := \sum_{r=1}^{R} P_{jr}^{GG}\big(\mathscr{A}_{jr}^{GG}\big) \quad \text{and} \quad \deg(C_j)_{in}^{EG} := \sum_{s=1}^{S} P_{js}^{EG}\big(\mathscr{A}_{js}^{EG}\big),
$$

respectively. Similarly, the real-valued *indegree* of cluster D_i is given by

$$
\deg(D_i)_{in}^{GE} := \sum_{r=1}^{R} P_{ir}^{GE}\big(\mathscr{A}_{ir}^{GE}\big) \quad \text{and} \quad \deg(D_i)_{in}^{EE} := \sum_{s=1}^{S} P_{is}^{EE}\big(\mathscr{A}_{is}^{EE}\big).
$$

In the same way, we can adapt the outdegrees of clusters by replacing the binary variables ζ_{jr}^{GG}, ζ_{js}^{EG}, ζ_{ir}^{GE} and ζ_{is}^{EE} with real variables $Q_{jr}^{GG}, Q_{js}^{EG}, Q_{ir}^{GE}, Q_{is}^{EE} \in [0,1]$ linearly depending on the corresponding elements of $\mathscr{A}_{jr}^{GG}, \mathscr{A}_{js}^{EG}, \mathscr{A}_{ir}^{GE}, \mathscr{A}_{is}^{EE}$. Now, the real-valued *outdegrees* of cluster C_j with respect to the genetic and environmental clusters are defined by

$$
\deg(C_j)_{out}^{GG} := \sum_{r=1}^{R} Q_{jr}^{GG}\big(\mathscr{A}_{jr}^{GG}\big) \quad \text{and} \quad \deg(C_j)_{out}^{EG} := \sum_{s=1}^{S} Q_{js}^{EG}\big(\mathscr{A}_{js}^{EG}\big),
$$

respectively. Similarly, the real-valued *outdegree* of cluster D_i is given by

$$\deg(D_i)^{GE}_{out} := \sum_{r=1}^{R} P^{GE}_{ir}\left(\mathscr{A}^{GE}_{ir}\right) \quad \text{and} \quad \deg(D_i)^{EE}_{out} := \sum_{s=1}^{S} Q^{EE}_{is}\left(\mathscr{A}^{EE}_{is}\right).$$

When we replace the strict binary constraints of the mixed-integer problem *(MI1)* with the aforementioned ‘soft constraints’, we obtain the following *continuous programming problem*:

$$\textit{(C1)} \quad \begin{cases} \text{Maximize } \displaystyle\sum_{\kappa=1}^{T} \left\{ \sum_{r=1}^{R} \left\| \Delta X_r^{(\kappa)} \right\|_* - \left\| \widehat{\mu}_r^{(\kappa)} - \overline{\mu}_r^{(\kappa)} \right\|_2^2 \right. \\ \qquad\qquad \left. + \displaystyle\sum_{s=1}^{S} \left\| \Delta E_s^{(\kappa)} \right\|_* - \left\| \widehat{\rho}_s^{(\kappa)} - \overline{\rho}_s^{(\kappa)} \right\|_2^2 \right\} \\ \\ \text{subject to } \displaystyle\sum_{r=1}^{R} P^{GG}_{jr}\left(\mathscr{A}^{GG}_{jr}\right) \le \alpha^{GG}_j, \; j = 1, \ldots, R \\ \qquad\qquad \displaystyle\sum_{s=1}^{S} P^{EG}_{js}\left(\mathscr{A}^{EG}_{js}\right) \le \alpha^{EG}_j, \; j = 1, \ldots, R \\ \qquad\qquad \displaystyle\sum_{r=1}^{R} P^{GE}_{ir}\left(\mathscr{A}^{GE}_{ir}\right) \le \alpha^{GE}_i, \; i \; = 1, \ldots, S \\ \qquad\qquad \displaystyle\sum_{s=1}^{S} P^{EE}_{is}\left(\mathscr{A}^{EE}_{is}\right) \le \alpha^{EE}_i, \; i \; = 1, \ldots, S \\ \\ \qquad\qquad \displaystyle\sum_{r=1}^{R} Q^{GG}_{jr}\left(\mathscr{A}^{GG}_{jr}\right) \le \beta^{GG}_j, \; j = 1, \ldots, R \\ \qquad\qquad \displaystyle\sum_{s=1}^{S} Q^{EG}_{js}\left(\mathscr{A}^{EG}_{js}\right) \le \beta^{EG}_j, \; j = 1, \ldots, R \\ \qquad\qquad \displaystyle\sum_{r=1}^{R} Q^{GE}_{ir}\left(\mathscr{A}^{GE}_{ir}\right) \le \beta^{GE}_i, \; i \; = 1, \ldots, S \\ \qquad\qquad \displaystyle\sum_{s=1}^{S} Q^{EE}_{is}\left(\mathscr{A}^{EE}_{is}\right) \le \beta^{EE}_i, \; i \; = 1, \ldots, S. \end{cases}$$

We can also combine these real-valued restrictions on the total number of all ingoing or outgoing branches of each cluster and so we obtain a relaxation of model *(M2)* in terms of the following *continuous programming problem*:

$$(C2)\quad \begin{cases} \text{Maximize} \displaystyle\sum_{\kappa=1}^{T} \left\{ \sum_{r=1}^{R} \left\| \Delta X_r^{(\kappa)} \right\|_* - \left\| \widehat{\mu}_r^{(\kappa)} - \overline{\mu}_r^{(\kappa)} \right\|_2^2 \right. \\ \qquad\qquad \left. + \displaystyle\sum_{s=1}^{S} \left\| \Delta E_s^{(\kappa)} \right\|_* - \left\| \widehat{\rho}_s^{(\kappa)} - \overline{\rho}_s^{(\kappa)} \right\|_2^2 \right\} \\ \\ \text{subject to } \displaystyle\sum_{r=1}^{R} P_{jr}^{GG}\left(\mathscr{A}_{jr}^{GG}\right) + \sum_{s=1}^{S} P_{js}^{EG}\left(\mathscr{A}_{js}^{EG}\right) \leq \gamma_j,\ j = 1,\ldots,R \\ \qquad\qquad \displaystyle\sum_{r=1}^{R} P_{ir}^{GE}\left(\mathscr{A}_{ir}^{GE}\right) + \sum_{s=1}^{S} P_{is}^{EE}\left(\mathscr{A}_{is}^{EE}\right) \leq \beta_i,\ i\ = 1,\ldots,S \\ \\ \qquad\qquad \displaystyle\sum_{r=1}^{R} Q_{jr}^{GG}\left(\mathscr{A}_{jr}^{GG}\right) + \sum_{s=1}^{S} Q_{js}^{EG}\left(\mathscr{A}_{js}^{EG}\right) \leq \gamma_j,\ j = 1,\ldots,R \\ \qquad\qquad \displaystyle\sum_{r=1}^{R} Q_{ir}^{GE}\left(\mathscr{A}_{ir}^{GE}\right) + \sum_{s=1}^{S} Q_{is}^{EE}\left(\mathscr{A}_{is}^{EE}\right) \leq \delta_i,\ i\ = 1,\ldots,S. \end{cases}$$

6 Conclusion

We considered dynamical gene-environment networks under ellipsoidal uncertainty. The hidden interconnections and regulating effects of possibly overlapping clusters of genetic and environmental items are revealed by a new set-theoretic regression methodology. By this, we further extend our *Ellipsoidal Operations Research* introduced in [31] that was based on our and our colleagues studies on regulatory systems in computational biology and life sciences with data uncertainty. In these studies, interval arithmetics was applied to express various kinds of errors and uncertainty. Here, we further extend this approach by considering stochastic dependencies between groups of genes and environmental factors. Furthermore, the clusters of data items are not strictly divided as in [31] and they can overlap what is motivated by the fact that a single gene or environmental factor can influence more than one other gene. The representation of data uncertainty in terms of ellipsoids is more flexible than the error intervals of single variables. In particular, ellipsoids are directly related to covariance matrices. For this reason, the often used Gaussian random noise refers to our ellipsoidal approach. However, Gaussian random distributions are often considered as simplifications. In future works, we will extend our regression models based on ellipsoidal uncertainty and we will focus on set-theoretic approaches with semi-algebraic sets and approximations of convex or non-convex error sets. This new perception will be combined with refined optimization methods that offer a new perspective for the analysis of regulatory systems under uncertainty.

Acknowledgements The authors express their cordial gratitude to Professor Alberto Pinto for inviting us to contribute with our chapter to this distinguished book in honour of Prof. Dr. Mauricio Peixoto and Prof. Dr. David Rand.

References

1. S.Z. Alparslan Gök: Cooperative Interval Games. PhD Thesis, Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey (2009).
2. S.Z. Alparslan Gök, R. Branzei, S. Tijs: Convex interval games. Preprint at IAM, Middle East Technical University, Ankara, Turkey, and Center for Economic Research, Tilburg University, The Netherlands, 2008.
3. S.Z. Alparslan Gök, R. Branzei, S. Tijs: Airport interval games and their Shapley value. To appear in *Operations Research and Decisions*, **2**, 9–18 (2009).
4. S.Z. Alparslan Gök, S. Miquel, S. Tijs: Cooperation under interval uncertainty. Math. Methods Oper. Res., **69**, 99–109 (2009).
5. S.Z. Alparslan Gök, G.-W. Weber: Cooperative games under ellipsoidal uncertainty. To appear in the Proceedings of PCO 2010, 3rd Global Conference on Power Control and Optimization, February 2-4, 2010, Gold Coast, Queensland, Australia.
6. O. Alter, P.O. Brown, D. Botstein: Singular value decomposition for genome-wide expression data processing and modeling. PNAS, **97**, 18, 10101–10106 (2000).
7. A. Aster, B. Borchers, C. Thurber: Parameter estimation and inverse problems. Academic Press (2004).
8. A.M. Bagirov, J. Yearwood: A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problems. European J. Oper. Res., **170**, 2, 578–596 (2006).
9. Z. Barzily, Z.V. Volkovich, B. Akteke-Öztürk, G.-W. Weber: Cluster stability using minimal spanning tree. ISI Proceedings of 20th Mini-EURO Conference, *Continuous Optimization and Knowledge-Based Technologies*, Neringa, Lithuania, May 20-23, (2008), pp. 248–252.
10. R. Benedetti: Real algebraic and semi-algebraic sets. Hermann, Ed. des Sciences et des Arts, Paris (1990).
11. A. Ben-Tal: Conic and robust optimization. Lecture notes (2002)
Available at http://iew3.technion.ac.il/Home/Users/morbt.phtml.
12. J. Bochnak, M. Coste, M.-F. Roy: Real algebraic geometry. Springer 1998.
13. E. Borenstein, M.W. Feldman: Topological signatures of species interactions in metabolic networks. J. Comput. Biol., **16**, 2, 191–200 (2009), DOI: 10.1089/cmb.2008.06TT.
14. S. Boyd, L. Vandenberghe: Convex Optimization. Cambridge University Press (2004).
15. M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares, D. Haussler: Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS **97**, 1, 262-267 (2000).
16. E. Büyükbebeci: Comparison of MARS, CMARS and CART in Predicting Default Probabilities for Emerging Markets. MSc. Term Project Report/Thesis in Financial Mathematics, at IAM, METU, Ankara, August 2009.
17. H. Chipman, R. Tibshirani: Hybrid hierarchical clustering with applications to microarray data. Biostatistics, **7**, 2, 286–301 (2006).
18. C. De Mol, S. Mosci, M. Traskine, A. Verri: A Regularized Method for Selecting Nested Groups of Relevant Genes from Microarray Data. Journal of Computational Biology, **16**, 5, 677-690 (2009).
19. Ö. Defterli, A. Fügenschuh, G.-W. Weber: New discretization and optimization techniques with results in the dynamics of gene-environment networks. In the proceedings of PCO 2010, 3rd Global Conference on Power Control and Optimization, February 2-4, 2010, Gold Coast, Queensland, Australia (ISBN: 978-983-44483-1-8).
20. P. Durieu, É. Walter, B. Polyak: Multi-input multi-output ellipsoidal state bounding. J. Optim. Theory Appl., **111**, 2, 273–303 (2001).
21. J. Gebert, M. Lätsch, E.M.P. Quek, G.-W. Weber: Analyzing and optimizing genetic network structure via path-finding. Journal of Computational Technologies, **9**, 3, 3–12 (2004).
22. A. Gökmen, S. Kayalgil, G.-W. Weber, I. Gökmen, M. Ecevit, A. Sürmeli, T. Bali, Y. Ecevit, H. Gökmen, D.J. DeTombe: Balaban Valley Project: Improving the Quality of Life in Rural Area in Turkey. International Scientific Journal of Methods and Models of Complexity, **7**, 1 (2004).

23. J.R. Harris, W. Nystad, P. Magnus: Using genes and environments to define asthma and related phenotypes: applications to multivariate data. Clinical and Experimental Allergy, **28**, 1, 43–45 (1998).
24. T. Hastie, R. Tibshirani: Discriminant adaptive nearest neighbor classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, **18**, 6, 607–616 (1996).
25. T. Hastie, R. Tibshirani, J. Friedman: The elements of statistical learning. Springer (2001).
26. J. Herrero, A. Valencia, J. Dopazo: A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics, **17**, 2, 126–136 (2001).
27. S.D. Hooper, S. Boué, R. Krause, L.J. Jensen, C.E. Mason, M. Ghanim, K.P. White, E.E.M. Furlong, P. Bork: Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis. Molecular Systems Biology, **3**, 72, 2007.
28. A. Işcanoğlu, G.-W. Weber, P. Taylan: Predicting default probabilities with generalized additive models for emerging markets. Invited lecture, Graduate Summer School on New Advances in Statistics, METU (2007).
29. E. Kropat, S. Pickl, A. Rössler, and G.-W. Weber, On theoretical and practical relations between discrete optimization and nonlinear optimization, in special issue Colloquy Optimization Structure and Stability of Dynamical Systems (at the occasion of the colloquy with the same name, Cologne, October 2000) of *Journal of Computational Technologies*, **7**, 2002), pp. 27–62.
30. E. Kropat, G.-W. Weber, B. Akteke-Öztürk: Eco-finance networks under uncertainty. In J. Herskovits, A. Canelas, H. Cortes, and M. Aroztegui, eds.: Proceedings of the International Conference on Engineering Optimization (ISBN 978857650156-5, CD), EngOpt 2008, Rio de Janeiro, Brazil (2008).
31. E. Kropat, G.-W. Weber, J.-J. Rückmann: Regression analysis for clusters in gene-environment networks based on ellipsoidal calculus and optimization. Preprint 157 at IAM, METU, Ankara, Turkey (2009). Submitted to Dynamics of Continuous, Discrete and Impulsive Systems.
32. E. Kropat, G.-W. Weber, C.S. Pedamallu: Regulatory networks under ellipsoidal uncertainty - optimization theory and dynamical systems. Preprint at IAM, METU, Ankara, Turkey, 2009. Submitted to SIAM Journal on Optimization.
33. A.B. Kurzhanski, I. Vályi: Ellipsoidal Calculus for Estimation and Control. Birkhäuser (1997).
34. A.A. Kurzhanski, P. Varaiya: Ellipsoidal Toolbox Manual. EECS Department, University of California, Berkeley (2008).
35. Y. Lee, Y. Lin, G. Wahba: Multicategory Support Vector Machines: Theory and Application to the Classification of Microarray Data and Satellite Radiance Data. Journal of the American Statistical Association, **99**, 67–81 (2004).
36. S. Mahony, J. O McInerney, T.J. Smith, A. Golden: Gene prediction using the Self-Organizing Map: automatic generation of multiple gene models. BMC Bioinformatics **5**, 23 (2004). doi:10.1186/1471-2105-5-23.
37. M. Marvanova, P. Toronen, M. Storvik, M. Lakso, E. Castren, G. Wong: Synexpression analysis of ESTs in the rat brain reveals distinct patterns and potential drug targets. Molecular Brain Research, **104**, 2, 176-183 (2002).
38. W.B. Mattes, S.D. Pettit, S.A. Sansone, P.R. Bushel, M.D. Waters: Database development in toxicogenomics: issues and efforts. Environmental Health Perspectives, **112**, (4), 495–505. (2004).
39. A. Nemirovski: Five Lectures on Modern Convex Optimization. C.O.R.E. Summer School on Modern Convex Optimization (2002). Available at http://iew3.technion.ac.il/Labs/Opt/opt/LN/Final.pdf.
40. A. Nemirovski: Lectures on Modern Convex Optimization, Israel Institute of Technology (2002). Available at http://iew3.technion.ac.il/Labs/Opt/opt/LN/Final.pdf.
41. A. Nemirovski: Interior Point Polynomial Time Algorithms in Convex Programming, Lecture Notes (2004). Available at https://itweb.isye.gatech.edu.

42. A. Nemirovski: Modern Convex Optimization. Lecture at PASCAL Workshop, Thurnau, Germany, March 16-18 (2005).
43. Y.E. Nesterov, A.S. Nemirovskii: Interior Point Polynomial Algorithms in Convex Programming. SIAM (1994).
44. J. Nikkila, P. Törönen, S. Kaski, J. Venna, E. Castrén, G. Wong: Analysis and visualization of gene expression data using self-organizing maps. Neural Networks, **15**, 8-9, 953–966 (2002).
45. M. Partner, N. Kashtan, U. Alon: Environmental variability and modularity of bacterial metabolic network. BMC Evolutionary Biology, **7**, (2007). 169doi:10.1186/1471-2148-7-169.
46. S. Pickl: Der τ-value als Kontrollparameter - Modellierung und Analyse eines Joint-Implementation Programmes mithilfe der dynamischen kooperativen Spieltheorie und der diskreten Optimierung. Thesis, Darmstadt University of Technology, Department of Mathematics (1998).
47. J. Quackenbush: Computational analysis of microarray data. Nature Reviews Genetics, **2**, 418–427 (2001).
48. M.N. Rivolta, A. Halsall, C.M. Johnson, M.A. Tones, M.C. Holley: Transcript Profiling of Functionally Related Groups of Genes During Conditional Differentiation of a Mammalian Cochlear Hair Cell Line. Genome Research, **12**, 1091-1099 (2002).
49. L. Ros, A. Sabater, F. Thomas: An ellipsoidal calculus based on propagation and fusion. IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, **32**, 4, 430–442 (2002).
50. Y. She: Sparse regression with exact clustering. PhD Thesis, Department of Statistics, Stanford University, USA (2008).
51. A. Shapiro, D. Dentcheva, A. Ruszczyński: Lectures on stochastic programming: modeling and theory. To be published by SIAM, Philadelphia (2009).
52. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T.R. Golub: Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. PNAS, **96**, 6, 2907–2912 (1999).
53. M. Taştan: Analysis and prediction of gene expression patterns by dynamical systems, and by a combinatorial algorithm. MSc Thesis, Institute of Applied Mathematics, METU, Turkey (2005).
54. M. Taştan, T. Ergenç, S.W. Pickl, G.-W. Weber: Stability analysis of gene expression patterns by dynamical systems and a combinatorial algorithm. In: HIBIT – Proceedings of International Symposium on Health Informatics and Bioinformatics, Turkey '05, Antalya, Turkey, 2005, pp. 67–75.
55. M. Taştan, S.W. Pickl, G.-W. Weber: Mathematical modeling and stability analysis of gene-expression patterns in an extended space and with Runge-Kutta discretization. In: Proceedings of Operations Research 2005, Bremen, September 2005, Springer, pp. 443–450.
56. P. Taylan, G.-W. Weber, A. Beck: New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and techology. In the special issue in honour of Prof. Dr. Alexander Rubinov, Burachik, B., Yang, X. (guest eds.) *Optimization*, **56**, 5-6, 1–24 (2007).
57. B. Thomas, G. Raju, W. Sonam: A modified fuzzy c-means algorithm for natural data exploration. World Academy of Science, Engineering and Technology, **49**, (2009).
58. Ö. Uğur, S.W. Pickl, G.-W. Weber, R. Wünschiers: Operational research meets biology: An algorithmic approach to analyze genetic networks and biological energy production. Preprint no. 50, Institute of Applied Mathematics, METU, 2006. Submitted for the special issue of *Optimization* at the occasion of the 5th Ballarat Workshop on Global and Non-Smooth Optimization: Theory, Methods, and Applications (2006).
59. Ö. Uğur, S.W. Pickl, G.-W. Weber, R. Wünschiers: An algorithmic approach to analyze genetic networks and biological energy production: an introduction and contribution where OR meets biology. Optimization, **58**, 1, 1–22 (2009).
60. Ö. Uğur, G.-W. Weber: Optimization and dynamics of gene-environment networks with intervals. In the special issue at the occasion of the 5th Ballarat Workshop on Global and

Non-Smooth Optimization: Theory, Methods and Applications, November 28-30, 2006, of *J. Ind. Manag. Optim.*, **3**, 2, 357–379 (2007).

61. A.Y. Vazhentsev: On internal ellipsoidal approximations for problems of control synthesis with bounded coordinates. J. Comput. System Sci. Int., **39**, 3, (2000).
62. F.G. Vázques, J.-J. Rückmann, O. Stein, G. Still: Generalized semi-infinite programming: A tutorial. Journal of computational and applied mathematics, **217**, 2, 394–419 (2008).
63. M. Wall, A. Rechtsteiner, L. Rocha: Singular Value Decomposition and Principal Component Analysis. In: A Practical Approach to Microarray Data Analysis (Berrar DP, Dubitzky W, Granzow M, eds.), 91–109, Kluwer: Norwell, MA (2003).
64. G.-W. Weber: Charakterisierung struktureller Stabilität in der nichtlinearen Optimierung. In Aachener Beiträge zur Mathematik 5, H.H. Bock, H.T. Jongen, and W. Plesken, eds., Augustinus publishing house (now: Mainz publishing house) Aachen (1992).
65. G.-W. Weber: Minimization of a max-type function: Characterization of structural stability. In: Parametric Optimization and Related Topics III, J. Guddat, J., H. Th. Jongen, and B. Kummer, and F. Nožička, eds., Peter Lang publishing house, Frankfurt a.M., Bern, New York (1993), pp. 519–538.
66. G.-W. Weber: Generalized semi-infinite optimization and related topics. Heldermann Publishing House, Research and Exposition in Mathematics 29, Lemgo, K.H. Hofmannn and R. Wille, eds. (2003).
67. G.-W. Weber, S.-Z. Alparslan-Gök, O. Defterli, E. Kropat: Modeling, Inference and Optimization of Regulatory Networks Based on Time Series Data. Preprint at Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey, submitted to *European Journal of Operational Research* (EJOR).
68. G.-W. Weber, S.Z. Alparslan-Gök, N. Dikmen: Environmental and life sciences: gene-environment networks - optimization, games and control - a survey on recent achievements. Invited paper, in the special issue of *Journal of Organisational Transformation and Social Change*, **5**, 3, (2008), pp. 197–233, guest editor: D. DeTombe.
69. G.-W. Weber, S.Z. Alparslan-Gök, B. Söyler: A new mathematical approach in environmental and life sciences: gene-environment networks and their dynamics. Environmental Modeling & Assessment, **14**, 2, 267-288 (2009).
70. G.-W. Weber, I. Batmaz, G. Köksal, P. Taylan F. Yerlikaya-Özkur: CMARS: A new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimisation. Preprint at IAM, METU, Ankara.
71. G.-W. Weber, E. Kropat, B. Akteke-Öztürk, Z.-K. Görgülü: A survey on OR and mathematical methods applied on gene-environment networks. Special Issue on *Innovative Approaches for Decision Analysis in Energy, Health, and Life Sciences* of *Central European Journal of Operations Research (CEJOR)* at the occasion of EURO XXII 2007 (Prague, Czech Republic, July 8-11, 2007), **17**, 3, 315–341 (2009).
72. G.-W. Weber, E. Kropat, A. Tezel, S. Belen: Optimization applied on regulatory and eco-finance networks - survey and new developments. To appear in *Pac. J. Optim.*, **6**, 3, Special Issue in memory of Professor Alexander Rubinov, M. Fukushima, M., et al., guest eds. (2010).
73. G.-W. Weber, S. Özögür-Akyüz, E. Kropat: A review on data mining and continuous optimization applications in computational biology and medicine. Embryo Today, Birth Defects Research (Part C), **87**, 165–181 (2009).
74. G.-W. Weber, P. Taylan, S.-Z. Alparslan-Gök, S. Özöğür, B. Akteke-Öztürk: Optimization of gene-environment networks in the presence of errors and uncertainty with Chebychev approximation. *TOP*, the Operational Research journal of SEIO (Spanish Statistics and Operations Research Society) **16**, 2, 284-318 (2008).
75. G.-W. Weber, A. Tezel: On generalized semi-infinite optimization of genetic networks. TOP, **15**, 1, 65–77 (2007).
76. G.-W. Weber, A. Tezel, P. Taylan, A. Soyler, M. Çetin: Mathematical contributions to dynamics and optimization of gene-environment networks. In Special Issue: In Celebration of Prof. Dr. Dr. Hubertus Th. Jongen's 60th Birthday, D. Pallaschke, O. Stein, guest eds., of *Optimization*, **57**, 2, 353–377 (2008).

77. G.-W. Weber, Ö. Uğur, P. Taylan, A. Tezel: On optimization, dynamics and uncertainty: a tutorial for gene-environment networks. In the Special Issue *Networks in Computational Biology* of *Discrete Appl. Math.*, **157**, 10, 2494–2513 (2009).
78. F. Yerlikaya: A new contribution to nonlinear robust regression and classification with MARS and its applications to data mining for quality control in manufacturing. Thesis, Middle East Technical University, Ankara, Turkey (2008).
79. K.Y. Yeung, W.L. Ruzzo: Principal component analysis for clustering gene expression data. Bioinformatics, **17**, 9, 763–774 (2001).
80. F.B. Yılmaz: A mathematical modeling and approximation of gene expression patterns by linear and quadratic regulatory relations and analysis of gene networks. MSc Thesis, Institute of Applied Mathematics, METU, Ankara, Turkey (2004).
81. F.B. Yılmaz, H. Öktem, G.-W. Weber: Mathematical modeling and approximation of gene expression patterns and gene networks. In *Operations Research Proceedings*, F. Fleuren, D. den Hertog, and P. Kort, eds., 280–287 (2005).
82. A. Zhang: Advanced Analysis of Gene Expression Microarray Data. World Scientific Pub. Co. Ltd. (2006).