Reconstruction of switching thresholds in piecewise-affine models of genetic regulatory networks

Samuel Drulhe, Hidde De Jong, Giancarlo Ferrari-Trecate

ISSN 0249-0803

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

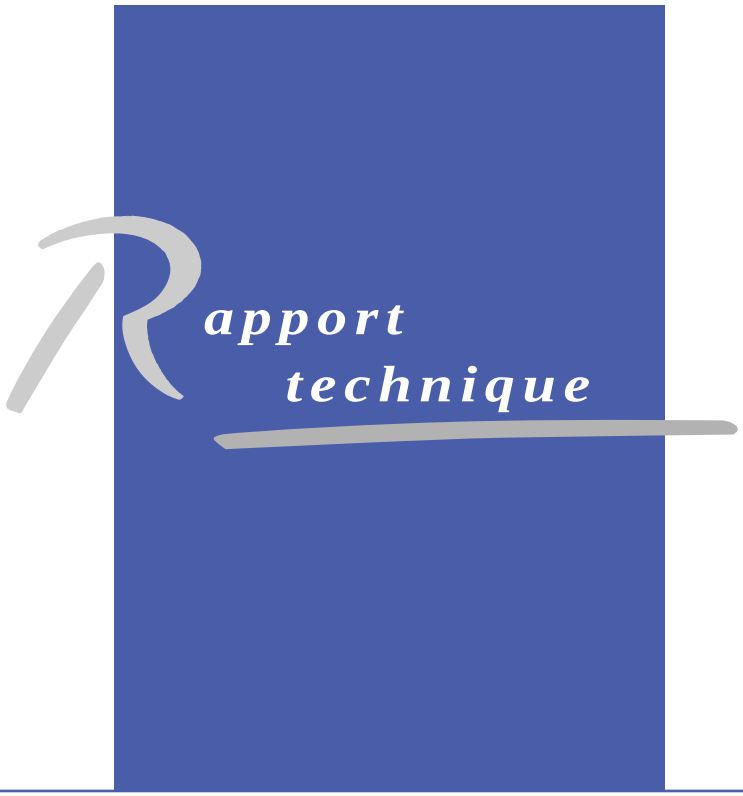
Reconstruction of Switching Thresholds in Piecewise-Affine Models of Genetic Regulatory Networks

Samuel Drulhe — Giancarlo Ferrari-Trecate — Hidde de Jong

N° 0322

February 2006

Thème BIO



INRIA
RHÔNE-ALPES

Reconstruction of Switching Thresholds in Piecewise-Affine Models of Genetic Regulatory Networks

Samuel Drulhe[*†] , Giancarlo Ferrari-Trecate[‡§] , Hidde de Jong[*]

Thème BIO — Systèmes biologiques
Projet Hélix

Rapport technique n° 0322 — February 2006 — 24 pages

Abstract: Recent advances of experimental techniques in biology have led to the production of enormous amounts of data on the dynamics of genetic regulatory networks. In this technical report, we present an approach for the identification of PieceWise-Affine (PWA) models of genetic regulatory networks from experimental data, focusing on the reconstruction of switching thresholds associated with regulatory interactions. In particular, our method takes into account geometric constraints specific to models of genetic regulatory networks. We show the feasibility of our approach by the reconstruction of switching thresholds in a PWA model of the carbon starvation response in the bacterium *Escherichia coli*.

Key-words: hybrid systems, piecewise-affine differential equations, differential equations with discontinuous righthand sides, quantitative analysis, identification, switching thresholds, genetic regulatory networks, nutritional stress response, Escherichia coli

[*] Institut National de Recherche en Informatique et en Automatique, Unité de recherche Rhône-Alpes, Grenoble, France
[†] Université Joseph Fourier, Grenoble, France
[‡] Institut National de Recherche en Informatique et en Automatique, Unité de recherche Rocquencourt, Paris, France
[§] Università degli studi di Pavia, Pavia, Italy

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier (France)
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Reconstruction des seuils de transition dans le cadre de modèles affines par morceaux de réseaux de régulation génique

Résumé : Les récentes améliorations de techniques expérimentales en biologie ont débouché sur la production de grandes quantités de données portant sur la dynamique des réseaux de régulation génique. Dans ce rapport, nous présentons une approche pour l'identification de modèles affines par morceaux de réseaux de régulation génique à partir des données expérimentales, en se concentrant sur la reconstruction des seuils de transition associés aux interactions de commande régulatoire. En particulier, notre méthode prend en considération des contraintes d'ordre géométrique qui sont spécifiques à ces modèles de régulation génique. Enfin, nous illustrons l'utilisation de notre approche par la reconstruction des seuils de transition d'un modèle affine par morceaux de réponse à un stress en carbone à l'intérieur de la bactérie *Escherichia coli*.

Mots-clés : systêmes hybrides, équations différentielles linéaires par morceaux, équations différentielles discontinues, simulation quantitative, identification, seuils de transition, réseaux de régulation génique, réponse à un stress nutritionel, Escherichia coli

1 Introduction

Recent advances of experimental techniques in biology have led to the production of enormous amounts of data on the dynamics of cellular processes. Prominent examples of such techniques are DNA microarrays [17] and gene reporter systems [23], which allow gene expression to be measured with varying degrees of precision and throughput. One of the major challenges in biology today consists in the analysis and interpretation of these data, with a view to identifying the networks of interactions between genes, proteins, and small molecules that regulate the observed processes. The mapping of these *genetic regulatory networks* is a key issue for understanding the functioning of a cell and for designing interventions of biotechnological or biomedical relevance.

The problem of identifying genetic regulatory networks from gene expression data has attracted much attention over the last ten years. Most approaches are based on the use of *linear* models (*e.g.*, [4, 7, 10, 24]), for which powerful identification algorithms exist. However, given that the underlying biological processes are usually strongly nonlinear, the models are valid only near an equilibrium point (see [16] for an exception). While there have been some approaches based on *nonlinear* models of genetic regulatory networks, the practical applicability of these models is often compromised by the intrinsic mathematical and computational difficulty of nonlinear system identification. Not surprisingly, most authors have therefore focused on specific classes of nonlinear models, with restrictions that reduce the number of parameters and simplify the mathematical form (*e.g.*, [13, 15, 21]).

Another class of models that seems to strike a good compromise between the advantages and disadvantages of linear and nonlinear models are the *PieceWise-Affine (PWA)* models of genetic regulatory networks introduced by Glass and Kauffman in the 1970s [12]. The study of these models and their generalizations has been an active research area in both mathematical biology and hybrid systems theory (*e.g.*, [1, 2, 6, 8, 11, 18]). Notwithstanding their simple mathematical form, PWA systems capture essential aspects of gene regulation, as demonstrated by several modeling studies of regulatory networks of biological interest [11, 22]. Moreover, powerful techniques for the identification of PWA systems have been developed in the field of hybrid systems (see [14] and the references therein), which might be profitably applied to the reconstruction of genetic regulatory networks from experimental data.

Although the available hybrid identification algorithms provide a good starting point, they are generic in nature and therefore not well-adapted to a number of constraints specific to PWA models of genetic regulatory networks. First of all, the state space regions associated with modes of the system are hyperrectangular, as they are defined by switching thresholds of the concentration variables. Second, there exist strong dependencies between the modes of the system, as a consequence of the coordinated control of gene expression. Third, the aim of the system identification process is not to generate a single model, but *all* models with a minimal number of regulatory interactions that are consistent with the experimental data.

The aim of our paper is to make a first step towards the adaptation of existing algorithms for the identification of PWA models so as to take into account the above constraints. In particular, we focus on a crucial stage of the identification process: the estimation of the switching thresholds that partition the state space into hyperrectangular regions. We introduce an algorithm that, given gene expression time-series data classified according to the regulatory modes, produces all minimal sets of switching thresholds. We thus assume here that the preliminary problem of detecting mode switches in time-series data has been solved [14], although we are of course well aware that the underlying classification algorithms will probably have to be tailored to gene expression data as well. In order to illustrate the feasibility of our approach, we apply the threshold identification algorithm to a PWA model of the carbon starvation response in *Escherichia coli* [1, 22]. The gene expression data have been obtained by simulation, while adjusting the noise level and the sampling frequency to the real data that will ultimately be available to us. The work presented in this paper is complementary to the approach of Perkins and colleagues [19], who focus on the reconstruction of the regulatory modes once the switching thresholds of the system are known.

In the next two sections, we will review PWA models of genetic regulatory networks and discuss the use of hybrid identification techniques for their reconstruction. In Sections 4 to 6 we introduce the notions of cut and multicut, formulate the switching threshold reconstructing problem in terms of these concepts, and introduce algorithms that, under suitable assumptions, reconstruct minimal sets of switching thresholds from gene expression data. Section 7 presents the results of the multicut algorithm in the context of the *E. coli* carbon starvation model. In the final section we summarize our contributions and indicate directions for further research.

2 Piecewise-affine models of genetic regulatory networks

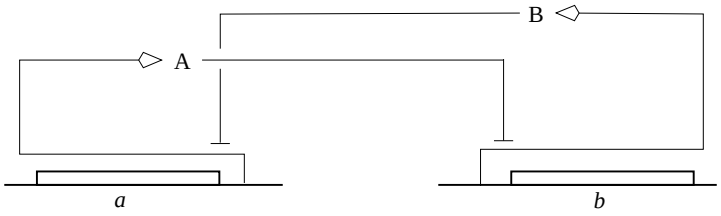
A variety of model formalisms have been proposed to describe the dynamics of genetic regulatory networks (see [5] for a review). A formalism particularly well-adapted to the currently available experimental data is the following class of *PWA differential equations* [12]:

$$\dot{x} = h(x) = f(x) - g(x)x, \tag{1}$$

where $x = [x_1, \ldots, x_n]' \in \Omega \subset \mathbb{R}^n_{\geq 0}$ is a vector of cellular protein concentrations, $f = [f_1, \ldots, f_n]'$, $g = \mathrm{diag}\,(g_1, \ldots, g_n)$, and Ω is a bounded, n-dimensional hyperrectangle. In equation (1), the rate of change of each protein concentration x_i is the difference of the rate of synthesis $f_i(x)$ and the rate of degradation $g_i(x)x_i$. The map f_i is defined as a sum of terms having the general form $\kappa_i^l\, b_i^l(x)$, where $\kappa_i^l > 0$ is a rate parameter and $b_i^l(x) : \Omega \to \{0, 1\}$ a piecewise-constant function defined in terms of the scalar step functions s^+ and s^- defined as

$$s^+(x_i, \theta_i) = \begin{cases} 1 & \text{if} \quad x_i > \theta_i \\ 0 & \text{if} \quad x_i < \theta_i \end{cases} \qquad \text{and} \qquad s^-(x_i, \theta_i) = 1 - s^+(x_i, \theta_i), \tag{2}$$

with $\theta_i > 0$ a constant denoting a threshold concentration for x_i. The step functions are reasonable approximations of sigmoid functions, which represent the switch-like character of the interactions found in gene regulation. The map g_i, which expresses regulation of protein degradation, is defined analogously, except that it is required to be strictly positive.



(a)

$$\dot{x}_a = \kappa_a\, s^-(x_b, \theta_b) - \gamma_a\, x_a,$$
$$\dot{x}_b = \kappa_b\, s^-(x_a, \theta_a) - \gamma_b\, x_b.$$

(b)

Figure 1: (a) Example of a simple genetic regulatory network, composed of two genes a and b, the proteins A and B, and their regulatory interactions. (b) PWA model of the network in (a).

Figure 1(a) shows an example of a simple *mutual-inhibition* genetic regulatory network. The genes a and b, transcribed from separate promoters, encode the proteins A and B, each of which controls the expression of the other gene. Repression of the genes is achieved by binding of the proteins to regulatory sites overlapping with the promoters. In Figure 1(b) the PWA differential equations of the example network are shown. Gene a is expressed at a rate κ_a, if the concentration of protein B is below its threshold θ_b, that is, if $s^-(x_b, \theta_b) = 1$. Analogously, gene b is expressed at a rate κ_b, if the concentration of protein A is below the threshold θ_a. The degradation of the proteins is not regulated in this case and therefore proportional to the concentration of the proteins (with degradation parameters γ_a or γ_b).

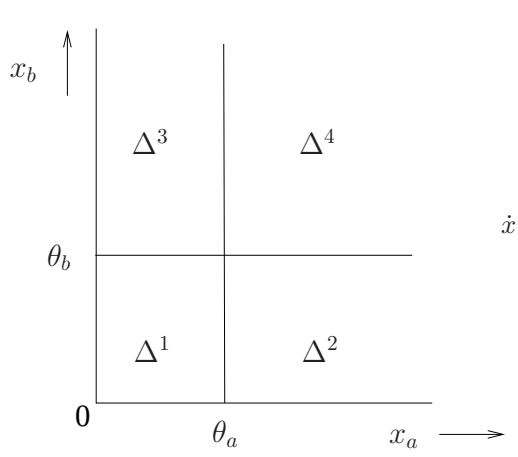
We now show how model (1) can be recast into a standard PWA system. Consider the union of threshold hyperplanes

$$\Theta = \cup_{i \in \{1,\ldots,n\}, l_i \in \{1,\ldots,p_i\}} \{x \in \Omega : x_i = \theta_i^{l_i}\}, \tag{3}$$

where p_i denotes the number of thresholds for x_i. Θ splits Ω in open hyperrectangular regions Δ^j, $j = 1, \ldots, s$, $s = \prod_{i=1}^n (p_i+1)$, called *regulatory domains*. One can show that if $x \in \Delta^j$, then model (1) reduces to $\dot{x} = \mu^j - \nu^j x$, where $\mu^j = f(x)$ is a constant vector and $\nu^j = g(x)$ is a constant diagonal matrix. In summary, when $x \in \Omega \backslash \Theta$, model (1) is equivalent to the PWA system

$$\dot{x} = h(x) = \mu^j - \nu^j x, \quad \text{if} \quad \lambda(x) = j, \quad j = 1, \ldots, s, \tag{4}$$

where the switching function λ is defined as: $\lambda(x) = j$, if and only if $x \in \Delta^j$. Note that in every domain Δ^j, the map $h(x)$ is affine and in each mode of operation the state variables evolve independently of each other (Figure 2).



$$\dot{x} = \begin{bmatrix} \dot{x}_a \\ \dot{x}_b \end{bmatrix} = \begin{cases} \begin{bmatrix} \kappa_a \\ \kappa_b \end{bmatrix} - \begin{bmatrix} \gamma_a & 0 \\ 0 & \gamma_b \end{bmatrix} \begin{bmatrix} x_a \\ x_b \end{bmatrix}, & \text{if } \lambda(x) = 1 \\ \begin{bmatrix} \kappa_a \\ 0 \end{bmatrix} - \begin{bmatrix} \gamma_a & 0 \\ 0 & \gamma_b \end{bmatrix} \begin{bmatrix} x_a \\ x_b \end{bmatrix}, & \text{if } \lambda(x) = 2 \\ \begin{bmatrix} 0 \\ \kappa_b \end{bmatrix} - \begin{bmatrix} \gamma_a & 0 \\ 0 & \gamma_b \end{bmatrix} \begin{bmatrix} x_a \\ x_b \end{bmatrix}, & \text{if } \lambda(x) = 3 \\ - \begin{bmatrix} \gamma_a & 0 \\ 0 & \gamma_b \end{bmatrix} \begin{bmatrix} x_a \\ x_b \end{bmatrix}, & \text{if } \lambda(x) = 4 \end{cases}$$

Figure 2: PWA system representing the model in Figure 1, corresponding to the subdivision of Ω into four regulatory domains: $\Delta_1, \ldots, \Delta_4$.

3 Hybrid system identification of genetic regulatory networks

Experimental techniques in biology, like DNA microarrays and gene reporter systems, allow gene expression to be measured at discrete time instants. In what follows, we assume that data are obtained with a uniform sampling period $T > 0$, where T is small with respect to the time constants of gene expression. We denote by $\hat{x}(k)$, $k = 1, \ldots, N+1$, the measured vectors of concentrations $\hat{x}(kT)$. By approximating derivatives through first-order differences, from (4) one obtains the following data model:

$$\hat{x}(k+1) = (I - T\nu^j)\,\hat{x}(k) + T\mu^j + \epsilon(k), \quad \text{if} \quad \lambda(\hat{x}(k)) = j, \tag{5}$$

where $\epsilon(k)$ is an additive noise corrupting the measurements. By focusing on the dynamics of a single protein concentration, say $\hat{x}_i$, model (5) becomes

$$\hat{x}_i(k+1) = \begin{bmatrix} \hat{x}_i(k) & 1 \end{bmatrix} \phi^j + \epsilon(k), \quad \text{if} \quad \lambda(\hat{x}(k)) = j, \tag{6}$$

where $\phi^j = \begin{bmatrix} 1 - T(\nu^j)_{ii} & T(\mu^j)_i \end{bmatrix}'$. [1]

Over the last few years, several hybrid system identification algorithms have been proposed for the reconstruction of so-called PieceWise AutoRegressive eXogenous (PWARX) models (see [14] for a review). Without going into details (which can be found in [9]), we just highlight that (6) is a PWARX system with input $u(k) = [\hat{x}_1(k), \ldots, \hat{x}_{l \neq i}(k), \ldots, \hat{x}_n(k)]'$ and output $y(k) = \hat{x}_i(k)$. The identification of model (6) amounts to three tasks: (i) the estimation of the number of domains s (when s is not known *a priori*); (ii) the estimation of the hyperrectangular domains Δ^j; (iii) the estimation of the parameters vectors ϕ^j. All hybrid identification techniques available in the literature perform tasks (ii) and (iii), while some have also the built-in capability to estimate the number of modes.

In the sequel, we focus on task (ii), which usually requires an intermediate result produced by all of the above algorithms: the reconstruction of the *switching sequence* $\lambda(\hat{x}(k))$, $k = 1, \ldots, N$. More specifically, as illustrated in [9], a domain Δ^j is found by looking for the $s-1$ hyperplanes separating the set $\mathcal{F}_j = \{\hat{x}(k) : \lambda(\hat{x}(k)) = j\}$ from all sets $\mathcal{F}_l = \{\hat{x}(k) : \lambda(\hat{x}(k)) = l\}$, $l \neq j$. These hyperplanes can be obtained through pattern-recognition techniques such as Multicategory Robust Linear Programming (MRLP) [3] or Support Vector Classifiers (SVC) [25].

A problem with this approach is that both MRLP and SVC do not impose any constraint on the hyperplanes to be estimated. As a consequence, even if the switching sequence is perfectly known, there is no guarantee that the estimated domains Δ^j will be hyperrectangular. This may result in hybrid models that are meaningless from a biological point of view, since they do not preserve the concept of a switching threshold associated with a concentration variable. Another problem with existing techniques is that they produce a single model. This is not realistic in our case, because only a fraction of the modes are encountered in the experiments. As a

[1] $(\nu^j)_{ii}$ is the element at position (i,i) of ν^j, $(\mu^j)_i$ is the ith element of μ^j.

consequence, several hybrid models of the network, each characterized by a different combination of thresholds for the variables, may be consistent with the data and need to be considered.

For all of these reasons, we propose a pattern recognition algorithm tailored to the features of PWARX models of genetic regulatory networks in the next three sections.

4 Switching thresholds and multicuts

Let $\mathcal{F}_1, \ldots, \mathcal{F}_s$ be disjoint sets collecting finitely-many points in $\mathbb{R}^n$ and $\mathcal{F}^* = \{\mathcal{F}_1, \ldots, \mathcal{F}_s\}$. Hereafter, we focus on the problem of separating the sets in $\mathcal{F}^*$ with hyperplanes parallel to the linear combination of $n-1$ axes. In order to illustrate the main concepts, we will use the collection $\mathcal{F}^*$ depicted in Figure 3(a). Pairs of distinct sets in $\mathcal{F}^*$ will often be indexed by means of pairs in $U = \{(p,q) \in \{1,\ldots,s\}^2 : p < q\}$. The cardinality of any finite set $\mathcal{A}$ will be denoted by $|\mathcal{A}|$.

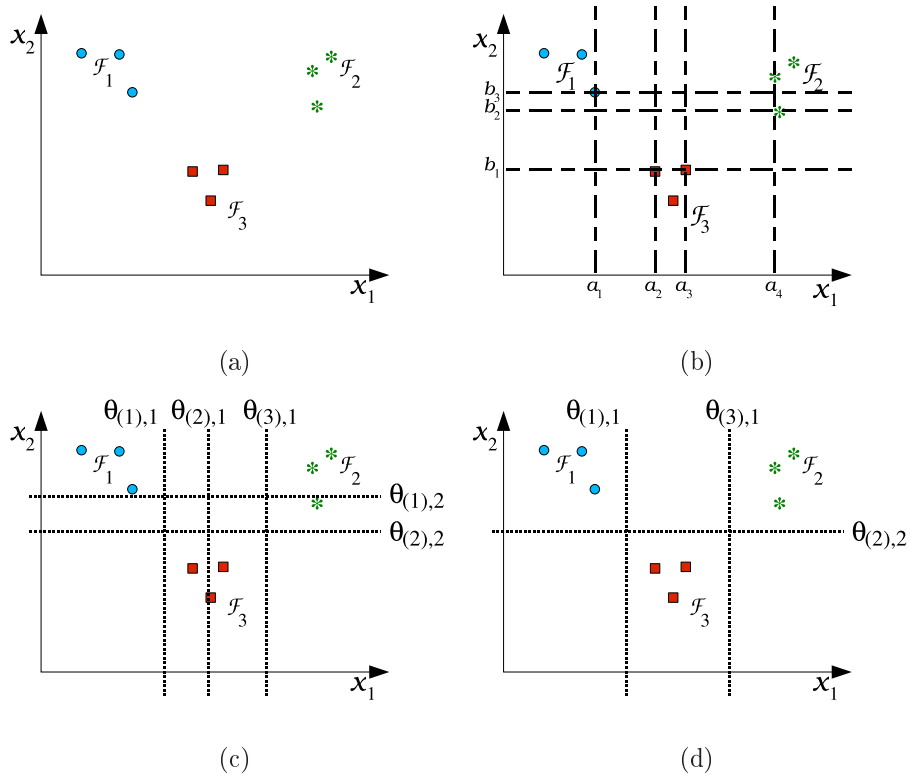


Figure 3: Simple example of multicuts. (a) Data sets $\mathcal{F}^*$. (b) Boundaries of the classes of equivalence. (c) Multicut $\mathcal{C}^*$. (d) Multicut $Max_{\preceq}\, \mathcal{C}^*$.

4.1 Separating axis-parallel hyperplanes

Definition 1 (Ap-hyperplane) *An* axis-parallel (ap-) hyperplane *in* $\mathbb{R}^n$ *with direction* $i \in \{1,\ldots,n\}$ *is a hyperplane of equation* $x_i = \alpha$, $\alpha \in \mathbb{R}$, *or equivalently, the zero level set of the function* $\theta(x) = x_i - \alpha$.

By abuse of notation, θ will denote both an ap-hyperplane and its associated function. The function $dir(\theta)$ gives the direction i of the ap-hyperplane θ, while the function $\mathrm{Z}(\theta)$ gives the zero-level α. Together, $dir(\theta)$ and $\mathrm{Z}(\theta)$ unambiguously define θ.

Definition 2 (Separability) *Let* $\mathcal{F}_p$ *and* $\mathcal{F}_q$ *be disjoint sets collecting finitely many points in* $\mathbb{R}^n$. *An ap-hyperplane* θ *in* $\mathbb{R}^n$ separates $\mathcal{F}_p$ *and* $\mathcal{F}_q$ *if there exists* $\delta \in \{+1,-1\}$ *such that for all* $x \in \mathcal{F}_p \cup \mathcal{F}_q$ *one has*

$$\begin{cases} \delta\,\theta(x) > 0, & \text{if } x \in \mathcal{F}_p, \\ \delta\,\theta(x) < 0, & \text{if } x \in \mathcal{F}_q. \end{cases} \tag{7}$$

In this case, we write $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$. *$\mathcal{F}_p$ and $\mathcal{F}_q$ are* separable *if there exists an ap-hyperplane separating the sets.*

In Figure 3(c), $\mathcal{F}_1$ and $\mathcal{F}_2$ are separable since there exist ap-hyperplanes in the x_1-direction (*e.g.*, $\theta_{(1),1}$ and $\theta_{(2),1}$), such that all points in $\mathcal{F}_1$ lie on one side of each hyperplane and all points of $\mathcal{F}_2$ on the other side. Notice that the sets $\mathcal{F}_1$ and $\mathcal{F}_2$ are not separable in the x_2-direction. As it can be verified in Figure 3, the ap-hyperplane $\theta_{(1),1}$ separates more sets than the ap-hyperplane $\theta_{(2),1}$.

The example of $\theta_{(2),1}$ illustrates that the data sets may lie on either side of the ap-hyperplane, or straddle the ap-hyerplane. To indicate the relative positions of the sets of $\mathcal{F}^*$, we introduce the set-valued functions:

$$\begin{aligned} \mathcal{I}_0(\theta) &= \{j : (\exists x \in \mathcal{F}_j, \theta(x) \leqslant 0) \text{ and } (\exists x \in \mathcal{F}_j, \theta(x) \geqslant 0)\}, \\ \mathcal{I}_-(\theta) &= \{j : \forall x \in \mathcal{F}_j, \theta(x) < 0\}, \\ \mathcal{I}_+(\theta) &= \{j : \forall x \in \mathcal{F}_j, \theta(x) > 0\}. \end{aligned} \tag{8}$$

It is straightforward to see that $\mathcal{I}_0(\theta)$, $\mathcal{I}_-(\theta)$ and $\mathcal{I}_+(\theta)$ constitute a partition of $\{1, \ldots, s\}$ ($s = |\mathcal{F}^*|$), i.e.

$$\mathcal{I}_-(\theta) \cup \mathcal{I}_+(\theta) \cup \mathcal{I}_0(\theta) = \{1, \ldots, s\} \text{ and the sets are disjoint.} \tag{9}$$

From Definition 1 and (8) it also follows that:

$$\begin{aligned} &\forall p \in \mathcal{I}_-(\theta), \forall x \in \mathcal{F}_p, x_{dir(\theta)} < \mathrm{Z}(\theta), \\ &\forall q \in \mathcal{I}_+(\theta), \forall x \in \mathcal{F}_q, x_{dir(\theta)} > \mathrm{Z}(\theta). \end{aligned} \tag{10}$$

Then, $(p, q) \in \mathcal{I}_-(\theta) \times \mathcal{I}_+(\theta)$ if and only if $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$.

In Figure 3, $\theta_{(1),1}$ separates both $\mathcal{F}_1$ from $\mathcal{F}_2$ and $\mathcal{F}_1$ from $\mathcal{F}_3$. Hence, one has $\mathcal{I}_-(\theta_{(1),1}) = \{1\}$ and $\mathcal{I}_+(\theta_{(1),1}) = \{2, 3\}$. Note also that the ap-hyperplane $\theta_{(2),1}$ separates only $\mathcal{F}_1$ from $\mathcal{F}_2$. The difference in separation power of ap-hyperplanes can be formally defined as follows.

Definition 3 (Separation power) *The* separation power *of an ap-hyperplane θ is the set-valued function*

$$S(\theta) = \{(p, q) \in U : \mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q\}. \tag{11}$$

In the remainder of this section, we focus on ap-hyperplanes in the set $\Theta = \{\theta : S(\theta) \neq \emptyset\}$. We also assume that Θ is non-empty.

The comparison of the separation power of ap-hyperplanes in Θ with the same direction motivates the introduction of equivalence classes of ap-hyperplanes.

Definition 4 (Equivalence) *Two ap-hyperplanes $\theta, \theta' \in \Theta$ are* equivalent *if $dir(\theta) = dir(\theta')$ and $S(\theta) = S(\theta')$. Equivalent ap-hyperplanes will be denoted by $\theta \sim \theta'$ and the equivalence class of θ by $[\theta] = \{\theta' : \theta' \sim \theta\}$. The* quotient set *will be denoted $\mathcal{E}^* = \Theta / \sim$.*

As it can be verified by means of the Definition 4, the ap-hyperplanes $\theta_{(1),1}$ and $\theta_{(2),1}$ in Figure 3(c) are not equivalent.

It is helpful to generalize the functions defined for an ap-hyperplane in Θ to its equivalence class in $\mathcal{E}^*$ using the notion of invariance. We recall that, given an equivalence relation $\sim$ on a set X and a function $f : X \to Y$, f is *invariant* under $\sim$ if $x \sim y$ implies $f(x) = f(y)$. It is obvious that the functions *dir* and S are invariant under the equivalence relation $\sim$ introduced in Definition 4.

Proposition 1 *The functions $\mathcal{I}_0$, $\mathcal{I}_-$ and $\mathcal{I}_+$ are invariant under the equivalence relation $\sim$.*

Proof. Let $\mathcal{E} \in \mathcal{E}^*$ and (θ, θ') be a pair of ap-hyperplanes belonging to $\mathcal{E}$. The case such that $\mathrm{Z}(\theta) = \mathrm{Z}(\theta')$ is trivial. If $\mathrm{Z}(\theta) \neq \mathrm{Z}(\theta')$, it follows from $S(\theta) = S(\theta')$ that $\mathcal{I}_-(\theta) = \mathcal{I}_-(\theta')$ and $\mathcal{I}_+(\theta) = \mathcal{I}_+(\theta')$. Then, in view of (9), $\mathcal{I}_0(\theta) = \{1, \ldots, s\} \setminus (\mathcal{I}_-(\theta) \cup \mathcal{I}_+(\theta)) = \{1, \ldots, s\} \setminus (\mathcal{I}_-(\theta') \cup \mathcal{I}_+(\theta')) = \mathcal{I}_0(\theta')$. □

Next, we characterize the zero-levels of ap-hyperplanes belonging to the same equivalence class $\mathcal{E}$. Consider the following functions:

- $g_-(\mathcal{E}) = \max_{p \in \mathcal{I}_-(\mathcal{E})} \max_{x \in \mathcal{F}_p} x_{dir(\mathcal{E})}$,

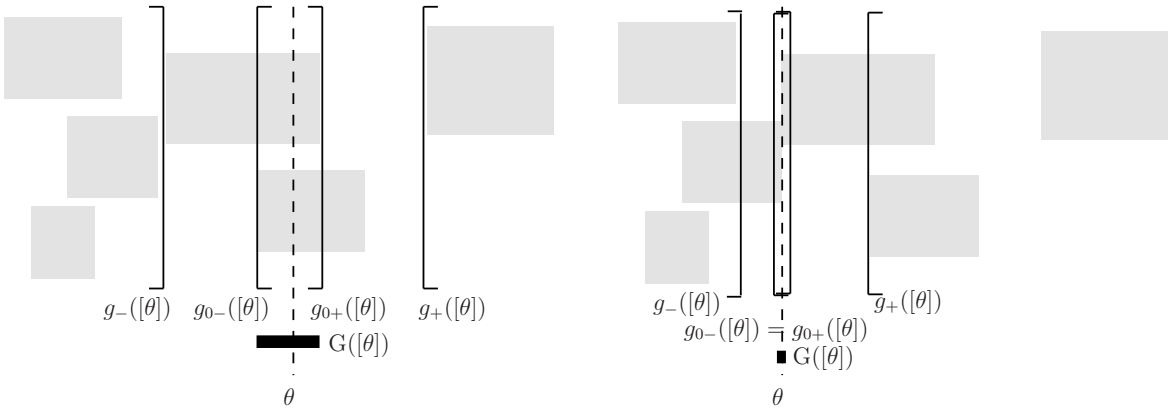


Figure 4: Intervals $\mathrm{G}([\theta])$ and the functions $g_-([\theta])$, $g_+([\theta])$, $g_{0-}([\theta])$ and $g_{0+}([\theta])$ for arbitrary chosen sets of points and a given ap-hyperplane θ (dashed line). Gray rectangles represent the smallest rectangles including all points in a set (i.e. there is at least a point lying on each boundary of each rectangle). The interval $\mathrm{G}([\theta])$ spanned by the equivalence class $[\theta]$ is depicted in bold. (a) $\mathrm{G}([\theta])$ is a closed interval. (b) $\mathrm{G}([\theta])$ is a point.

- $g_+(\mathcal{E}) = \min_{q \in \mathcal{I}_+(\mathcal{E})} \min_{x \in \mathcal{F}_q} x_{dir(\mathcal{E})}$,
- $g_{0-}(\mathcal{E}) = \max_{j \in \mathcal{I}_0(\mathcal{E})} \min_{x \in \mathcal{F}_j} x_{dir(\mathcal{E})}$, if $\mathcal{I}_0(\mathcal{E}) \neq \emptyset$,
- $g_{0+}(\mathcal{E}) = \min_{j \in \mathcal{I}_0(\mathcal{E})} \max_{x \in \mathcal{F}_j} x_{dir(\mathcal{E})}$, if $\mathcal{I}_0(\mathcal{E}) \neq \emptyset$.

It should be noticed that since we assume that $S(\mathcal{E}) \neq \emptyset$, the maps $g_-(\cdot)$ and $g_+(\cdot)$ are always defined on $\mathcal{E}^*$. However, $g_{0-}(\mathcal{E})$ and $g_{0+}(\mathcal{E})$ make sense only if $\mathcal{I}_0(\mathcal{E}) \neq \emptyset$. In view of (10), the functions $g_-(\mathcal{E})$ and $g_+(\mathcal{E})$ verify, by construction:

$$\begin{array}{l} \forall \theta \in \mathcal{E}, \mathrm{Z}(\theta) > g_-(\mathcal{E}), \\ \forall \theta \in \mathcal{E}, \mathrm{Z}(\theta) < g_+(\mathcal{E}). \end{array} \tag{12}$$

Then, from (12) one has that $\forall \theta \in \mathcal{E}$, $g_-(\mathcal{E}) < \mathrm{Z}(\theta) < g_+(\mathcal{E})$ and hence $g_-(\mathcal{E}) < g_+(\mathcal{E})$. When $\mathcal{I}_0(\mathcal{E}) \neq \emptyset$, one also has that $\forall \theta \in \mathcal{E}$, $\forall j \in \mathcal{I}_0(\mathcal{E})$, $\min_{x \in \mathcal{F}_j} x_{dir(\mathcal{E})} \leqslant \mathrm{Z}(\theta) \leqslant \max_{x \in \mathcal{F}_j} x_{dir(\mathcal{E})}$, and hence $\forall (j, j') \in \mathcal{I}_0(\mathcal{E})^2$, $\min_{x \in \mathcal{F}_j} x_{dir(\mathcal{E})} \leqslant \mathrm{Z}(\theta) \leqslant \max_{x \in \mathcal{F}_{j'}} x_{dir(\mathcal{E})}$. As a consequence, one obtains the inequality $g_{0-}(\mathcal{E}) \leqslant g_{0+}(\mathcal{E})$.

Consider the following interval associated to $\mathcal{E} \in \mathcal{E}^*$:

$$\mathrm{G}(\mathcal{E}) = \left\{ \begin{array}{l}]g_-(\mathcal{E}), g_+(\mathcal{E})[\text{ , if } \mathcal{I}_0(\mathcal{E}) = \emptyset, \\ ]g_-(\mathcal{E}), g_+(\mathcal{E})[\cap [g_{0-}(\mathcal{E}), g_{0+}(\mathcal{E})], \text{ otherwise.} \end{array} \right. \tag{13}$$

Figure 4 gives a graphic representation of $\mathrm{G}(\mathcal{E})$. From (13) it is easy to realize that $\mathrm{G}(\mathcal{E})$ can be opened, half openened, closed (as in Figure 4(a)), or even a single point (as in Figure 4(b)). The next proposition characterizes a given equivalence class through G.

Proposition 2 *For every $\mathcal{E} \in \mathcal{E}^*$:*

$$\mathcal{E} = \{\theta : dir(\theta) = dir(\mathcal{E}) \text{ and } \mathrm{Z}(\theta) \in \mathrm{G}(\mathcal{E})\} \tag{14}$$

Proof. For $\mathcal{E} \in \mathcal{E}^*$ and $\theta \in \Theta$ such that $dir(\theta) = dir(\mathcal{E})$, we need to prove that $\theta \in \mathcal{E}$ if and only if $\mathrm{Z}(\theta) \in \mathrm{G}(\mathcal{E})$.

($\Rightarrow$) Let $\theta \in \mathcal{E}$, then $\mathrm{Z}(\theta) \in]g_-(\mathcal{E}), g_+(\mathcal{E})[$ (see Equation 12) and, if defined, $\mathrm{Z}(\theta) \in [g_{0-}(\mathcal{E}), g_{0+}(\mathcal{E})]$. As a consequence, $\mathrm{Z}(\theta) \in \mathrm{G}(\mathcal{E})$, also proving that $\mathrm{G}(\mathcal{E}) \neq \emptyset$ whenever $\mathcal{E} \in \mathcal{E}^*$.

($\Leftarrow$) Let $\theta \in \Theta$ such that $dir(\theta) = dir(\mathcal{E})$ and $\mathrm{Z}(\theta) \in \mathrm{G}(\mathcal{E})$. Because $\mathrm{Z}(\theta) \in]g_-(\mathcal{E}), g_+(\mathcal{E})[$, we have $\forall p \in \mathcal{I}_-(\mathcal{E}), \forall x \in \mathcal{F}_p, x_{dir(\mathcal{E})} < \mathrm{Z}(\theta)$ and $\forall q \in \mathcal{I}_+(\mathcal{E}), \forall x \in \mathcal{F}_q, x_{dir(\mathcal{E})} > \mathrm{Z}(\theta)$, so that $\mathcal{I}_-(\mathcal{E}) \subseteq \mathcal{I}_-(\theta)$ and $\mathcal{I}_+(\mathcal{E}) \subseteq \mathcal{I}_+(\theta)$. This implies that $S(\mathcal{E}) \subseteq S(\theta)$. We treat now separately the cases $\mathcal{I}_0(\mathcal{E}) = \emptyset$ and $\mathcal{I}_0(\mathcal{E}) \neq \emptyset$.

If $\mathcal{I}_0(\mathcal{E}) = \emptyset$, in view of (9) we have $\mathcal{I}_-(\mathcal{E}) \cup \mathcal{I}_+(\mathcal{E}) = \{1, \ldots, s\} \subseteq \mathcal{I}_-(\theta) \cup \mathcal{I}_+(\theta)$. From the inclusions $\{1, \ldots, s\} \subseteq \mathcal{I}_-(\theta) \cup \mathcal{I}_+(\theta) \subseteq \{1, \ldots, s\}$ one deduces that $\mathcal{I}_-(\theta) \cup \mathcal{I}_+(\theta) = \{1, \ldots, s\}$ which implies that $\mathcal{I}_0(\theta) = \emptyset = \mathcal{I}_0(\mathcal{E})$. Hence $\mathcal{I}_-(\mathcal{E}) = \mathcal{I}_-(\theta)$, $\mathcal{I}_+(\mathcal{E}) = \mathcal{I}_+(\theta)$, and then $S(\mathcal{E}) = S(\theta)$.

(a) (b)

Figure 5: (a) Poset diagram for the set of cuts $\mathcal{C}^*$ in Figure 3. The diagram shows, *e.g.*, $\theta_{(2),1} \preceq \theta_{(1),1}$. (b) Poset diagram for the down-set of $\mathcal{M} = \{\theta_{(1),1}, \theta_{(3),1}, \theta_{(2),2}\}$, which is a multicut for Figure 3. In fact, $\mathcal{M}$ equals $\mathrm{Max}_{\preceq} \mathcal{C}^*$.

If $\mathcal{I}_0(\mathcal{E}) \neq \emptyset$, one has $\mathrm{Z}(\theta) \in [g_{0-}(\mathcal{E}), g_{0+}(\mathcal{E})]$ which implies, by construction, that any element of $\mathcal{I}_0(\mathcal{E})$ is also element of $\mathcal{I}_0(\theta)$, i.e. $\mathcal{I}_0(\mathcal{E}) \subseteq \mathcal{I}_0(\theta)$. Now, in view of (9), one has $\mathcal{I}_-(\mathcal{E}) \cup \mathcal{I}_+(\mathcal{E}) \subseteq \mathcal{I}_-(\theta) \cup \mathcal{I}_+(\theta)$ and then $\mathcal{I}_0(\mathcal{E}) \supseteq \mathcal{I}_0(\theta)$. Since then $\mathcal{I}_0(\mathcal{E}) = \mathcal{I}_0(\theta)$, it follows again that $S(\theta) = S(\mathcal{E})$. Finally, if $\mathrm{Z}(\theta) \in \mathrm{G}(\mathcal{E})$, then $S(\theta) = S(\mathcal{E})$, that is equivalent to $\mathcal{E} = [\theta]$. □

Example 1 *In Figure 3(b), the dashed lines correspond to the extrema of the intervals associated with the different equivalence classes. For instance,* $g_-([\theta_{(1),1}]) = a_1$, $g_+([\theta_{(1),1}]) = a_2$, *and* $\mathcal{I}_0(\theta_{(1),1}) = \emptyset$, *so that* $\mathrm{G}([\theta_{(1),1}]) =]a_1, a_2[$. *Similarly,* $\mathrm{G}([\theta_{(3),1}]) =]a_3, a_4[$. *Another example is* $\theta_{(2),1}$*:* $g_-([\theta_{(2),1}]) = a_1$, $g_+([\theta_{(2),1}]) = a_4$, *and, because* $\mathcal{I}_0(\theta_{(2),1}) = \{3\}$, *one has* $g_{0-}([\theta_{(2),1}]) = a_2$ *and* $g_{0+}([\theta_{(2),1}]) = a_3$. *It follows that* $\mathrm{G}([\theta_{(2),1}]) = [a_2, a_3]$. *These are all the equivalence classes in the first direction.*

The next Proposition shows that number of distinct equivalence classes is finite.

Proposition 3 *The cardinality of* $\mathcal{E}^*$ *is finite.*

Proof. Since S is invariant under $\sim$, it can be taught as a function $S : \mathcal{E}^* \to \mathcal{P}(U)$ where $\mathcal{P}(U)$ denotes the power set of U. For two equivalence classes $(\mathcal{E}_1, \mathcal{E}_2) \in \mathcal{E}^{*2}$ having the same direction one has that $\mathcal{E}_1 \neq \mathcal{E}_2$ implies $S(\mathcal{E}_1) \neq S(\mathcal{E}_2)$ (see Definition 4). Therefore, the number of different equivalence classes with the same direction cannot be larger than $|\mathcal{P}(U)|$, that is finite. Since the number of possible directions n is also finite, the bound $|\mathcal{E}^*| \leq n|\mathcal{P}(U)|$ follows. □

4.2 Cuts

Although all ap-hyperplanes in an equivalence class $\mathcal{E} \in \mathcal{E}^*$ have the same separation power, only one is optimal in a statistical sense [25]. This ap-hyperplane will be called a *cut*.

Definition 5 (Cut) *Let* $\mathcal{E} \in \mathcal{E}^*$ *and* $i = dir(\mathcal{E})$. *The* cut *associated to* $\mathcal{E}$ *is the ap-hyperplane* $\theta \in \mathcal{E}$ *such that* $\mathrm{Z}(\theta)$ *is the midpoint of the interval* $\mathrm{G}(\mathcal{E})$.

Since $\mathrm{G}(\mathcal{E}) \neq \emptyset$ whenever $\mathcal{E} \in \mathcal{E}^*$ there exists an isomorphism between the cuts and the equivalence classes. We denote the set of all cuts by $\mathcal{C}^*$. Since $\mathcal{E}^*$ and $\mathcal{C}^*$ are isomorphic, the cardinality of $\mathcal{C}^*$ is also finite.

For the three data sets in Figure 3(a), $\mathcal{C}^*$ is composed of the five cuts $\theta_{(1),1}$, $\theta_{(2),1}$, $\theta_{(3),1}$, $\theta_{(1),2}$, and $\theta_{(2),2}$ represented in Figure 3(c) by means of dotted lines. Intuitively, we would be inclined to say that the cut $\theta_{(1),1}$ is more powerful than $\theta_{(2),1}$, in the sense that the former separates $\mathcal{F}_1$ and $\mathcal{F}_2$ as well as $\mathcal{F}_1$ and $\mathcal{F}_3$, whereas the latter separates only $\mathcal{F}_1$ and $\mathcal{F}_2$ (that is, $S(\theta_{(1),1}) = \{(1,2),(1,3)\}$ and $S(\theta_{(2),1}) = \{(1,2)\}$). This motivates the introduction of an order relation on $\mathcal{C}^*$, denoted by $\preceq$.

Definition 6 (Order relation on cuts) *Let* $\theta, \theta' \in \mathcal{C}^*$. *We define the binary relation* $\preceq$ *over* $\mathcal{C}^*$*:*

$$\theta \preceq \theta' \text{ if } S(\theta) \subseteq S(\theta') \text{ and } dir(\theta) = dir(\theta'). \tag{15}$$

It is straightforward to show that $\preceq$ is reflexive, antisymmetric, and transitive, and hence $\preceq$ is a partial order on $\mathcal{C}^*$. That is, $\mathcal{C}^*$ is a poset (partially ordered set). We refer the reader to [20] for a general introduction to posets.

The poset diagram corresponding to the example in Figure 3 is shown in Figure 5(a). As for any other poset, $\mathcal{C}^*$ admits maximal and minimal elements. The sets of maximal and minimal elements of $\mathcal{C}^*$ are denoted by

$\text{Max}_{\preceq}\,\mathcal{C}^*$ and $\text{Min}_{\preceq}\,\mathcal{C}^*$, respectively. For instance, in Figure 5(a), $\text{Max}_{\preceq}\,\mathcal{C}^* = \{\theta_{(1),1}, \theta_{(3),1}, \theta_{(2),2}\}$. As suggested by the figure, for many purposes it is practical to restrict $\mathcal{C}^*$ to a particular direction, that is, for $i \in \{1, \ldots, n\}$,

$$\mathcal{C}^{i*} = \{\theta \in \mathcal{C}^* : dir(\theta) = i\}. \tag{16}$$

$\mathcal{C}^{i*}$ is a subset of $\mathcal{C}^*$ inheriting the partial order from $\mathcal{C}^*$. In Figure 5(a), $\text{Max}_{\preceq}\,\mathcal{C}^{1*} = \{\theta_{(1),1}, \theta_{(3),1}\}$ and $\text{Max}_{\preceq}\,\mathcal{C}^{2*} = \{\theta_{(2),2}\}$.

4.3 Multicuts

In general, several cuts will be required to separate all sets in $\mathcal{F}^*$. This motivates the introduction of multicuts.

Definition 7 (Multicut) *A* multicut $\mathcal{M}$ *of* $\mathcal{F}^*$ *is a finite set of cuts such that for all* $(p, q) \in U$ *there exists a* $\theta \in \mathcal{M}$, *such that* $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$. *A collection* $\mathcal{F}^*$ *is said to be* m-separable *if there exists a multicut of* $\mathcal{F}^*$.

We call $\mathcal{M}^*$ the set of multicuts. Due to the fact that $\mathcal{C}^*$ is finite, $\mathcal{M}^*$ is finite as well. Notice that $\mathcal{M}^*$ may be empty, that is, $\mathcal{F}^*$ may not be m-separable. In the example of Figure 3, $\mathcal{M} = \{\theta_{(3),1}, \theta_{(1),2}\}$ is a multicut since we have $S(\theta_{(3),1}) = \{(1,2),(2,3)\}$ and $S(\theta_{(1),2}) = \{(1,3)\}$.

The following two propositions state relevant properties of multicuts.

Proposition 4 $\mathcal{M}$ *is a multicut if and only if* $U = \cup_{\theta \in \mathcal{M}} S(\theta)$.

Proof. ($\Rightarrow$) By contradiction, let $(p, q) \in U$ and $(p, q) \notin \cup_{\theta \in \mathcal{M}} S(\theta)$. Since $\mathcal{M}$ is a multicut, there exists θ verifying $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$, that is, $(p, q) \in S(\theta)$.
($\Leftarrow$) $\mathcal{M}$ is a finite set of cuts with $\cup_{\theta \in \mathcal{M}} S(\theta) = U$. Then, by Definition 7, $\mathcal{M}$ is a multicut. □

Proposition 5 $\mathcal{F}^*$ *is m-separable if and only if* $\mathcal{C}^*$ *is a multicut.*

Proof. ($\Rightarrow$) Let $(p, q) \in U$. By assumption, $\mathcal{F}_p$ and $\mathcal{F}_q$ are separable and there exists $\bar{\theta}$ that separates them. Let θ be the cut in $[\bar{\theta}]$. Then $\theta \in \mathcal{C}^*$ and $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$. Since the previous argument can be repeated for all $(p, q) \in U$, it follows that $\mathcal{C}^*$ is a multicut. ($\Leftarrow$) Obvious. □

We define an obvious partial order relation on the set of multicuts $\mathcal{M}^*$, the set inclusion $\subseteq$. The poset $\mathcal{M}^*$ for the example in Figure 3 consists of 20 multicuts (see Figure 6 for the representation of its diagram).

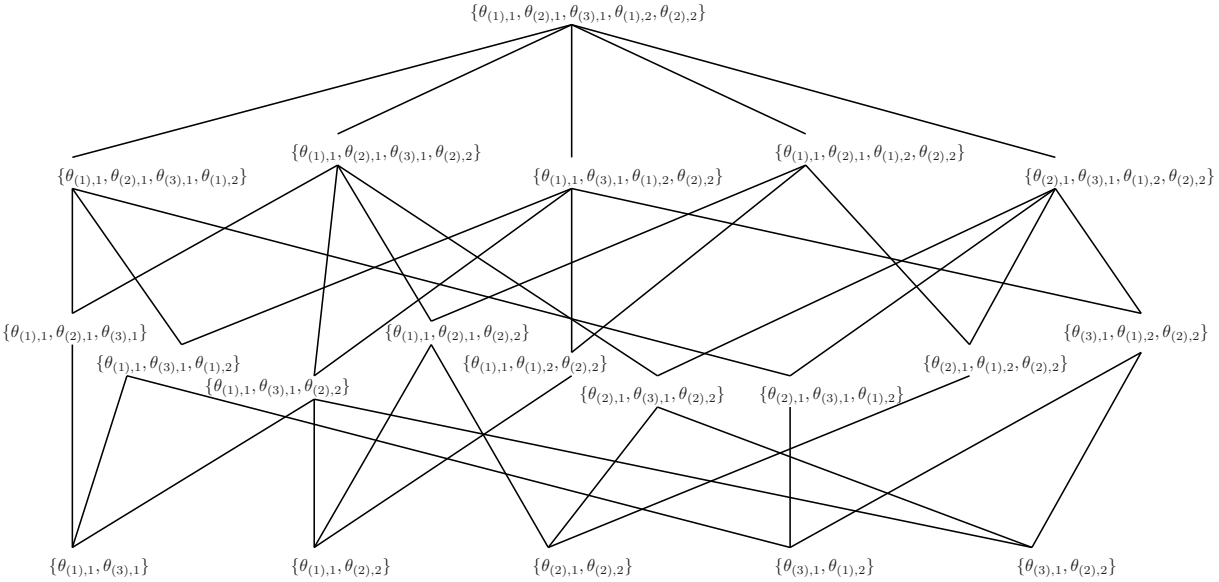


Figure 6: Poset diagram for the set of multicuts $\mathcal{M}^*$ corresponding to the example in Figure 3. The set of all cuts $\{\theta_{(1),1}, \theta_{(2),1}, \theta_{(3),1}, \theta_{(1),2}, \theta_{(2),2}\}$ is obviously the maximal element.

To every subset $\mathcal{B}$ of $\mathcal{M}^*$ we can associate a down-set, which consists of the multicuts in $\mathcal{M}^*$ upper bounded (according to $\subseteq$) by some multicut in $\mathcal{B}$. For reasons that will become clear below, we focus here on the down-set of singletons $\mathcal{B} = \{\mathcal{M}\}$, for some $\mathcal{M} \in \mathcal{M}^*$.

Definition 8 (Down-set of multicut set) *The* down-set *of* $\{\mathcal{M}\}$, $\mathcal{M} \in \mathcal{M}^*$, *denoted by* $\downarrow\{\mathcal{M}\}$, *is defined by*

$$\downarrow\{\mathcal{M}\} = \{\mathcal{M}' \in \mathcal{M}^* : \mathcal{M}' \subseteq \mathcal{M}\}. \tag{17}$$

Consider the multicut $\mathrm{Max}_{\preceq}\mathcal{C}^*$ in the example (Figure 5(a)). The down-set of $\{\mathrm{Max}_{\preceq}\mathcal{C}^*\}$ equals

$$\{\{\theta_{(1),1}, \theta_{(3),1}, \theta_{(2),2}\}, \{\theta_{(3),1}, \theta_{(2),2}\}, \{\theta_{(1),1}, \theta_{(2),2}\}, \{\theta_{(1),1}, \theta_{(3),1}\}\}$$

and is represented in Figure 5(b). We note that $\downarrow\{\mathcal{M}\}$ is also a poset with respect to set inclusion.

5 Formulation of switching threshold reconstruction problem

The introduction of the concepts of cut and multicut, and the partial orders defined on them, allows us to formulate the problem of reconstructing switching thresholds in a more precise way. Each cut θ corresponds to a switching threshold $\mathrm{Z}(\theta)$ for the concentration variable with index $dir(\theta)$. When the system crosses this threshold, the dynamics of the PWA system may switch from one regulatory mode to another. By extension, a multicut corresponds to a set of switching thresholds. These thresholds have the property to allow all classified data sets, associated with different regulatory modes of the system, to be separated.

In general, the available data are consistent with a large number of multicuts, and thus PWA models of the genetic regulatory network. *A priori* there is no reason to prefer one of these models above the others. However, in practice we are most interested in the minimal models that account for the available data, that is, in those models that contain a minimal number of thresholds and separate all pairs of sets in $\mathcal{F}^*$. Assuming that the set of data points is m-separable, so that $\mathcal{C}^*$ is a multicut, it seems reasonable to accept as solutions all multicuts in $\mathrm{Min}\downarrow\{\mathcal{C}^*\}$. That is, we want to find the multicuts that are minimal elements of the down-set of $\{\mathcal{C}^*\}$.

Notice though that $\mathcal{C}^*$ may contain many cuts with a weak separation power that could be eliminated beforehand if we are only interested in finding minimal multicuts. That is, we can remove cuts $\theta \in \mathcal{C}^*$ if there exists another $\theta' \in \mathcal{C}^*$, $\theta' \not\sim \theta$, such that $\theta \preceq \theta'$. Eliminating these cuts does not affect the m-separability of the sets of data points, as indicated by the following proposition which should be compared with Proposition 5.

Proposition 6 *$\mathcal{F}^*$ is m-separable if and only if $Max_{\preceq}\mathcal{C}^*$ is a multicut.*

Proof. ($\Rightarrow$) Since $\mathcal{F}^*$ is m-separable, $\mathcal{C}^*$ is a multicut (Proposition 5). For any cut $\theta \in \mathcal{C}^*$, there exists a cut $\tilde{\theta} \in \mathrm{Max}_{\preceq}\mathcal{C}^*$ such that $S(\theta) \subseteq S(\tilde{\theta})$. As a consequence any pair of sets $\mathcal{F}_p, \mathcal{F}_q$ (with $(p,q) \in U$) is separated by at least one cut in $\mathrm{Max}_{\preceq}\mathcal{C}^*$, which shows that $\mathrm{Max}_{\preceq}\mathcal{C}^*$ is a multicut. ($\Leftarrow$) Obvious. □

Once $\mathcal{C}^*$ has been reduced to $\mathrm{Max}_{\preceq}\mathcal{C}^*$, our switching threshold reconstruction problem can be recast into that of computing the set

$$\mathrm{Min}_{\subseteq}\downarrow\{\mathrm{Max}_{\preceq}\mathcal{C}^*\}. \tag{18}$$

That is, we are interested in the minimal elements of the down-set of the set of maximal cuts. Notice that $\mathrm{Max}_{\subseteq}\downarrow\{\mathrm{Max}_{\preceq}\mathcal{C}^*\}$ is $\{\mathrm{Max}_{\preceq}\mathcal{C}^*\}$ itself, so that we will call $\mathrm{Max}_{\preceq}\mathcal{C}^*$ the *maximal multicut*. In the example of Figure 3, $\mathrm{Max}_{\preceq}\mathcal{C}^*$ consists of three cuts, as shown in Figure 5(a). That is, two cuts with obvious weaker separation power have been eliminated ($\theta_{(2),1}$ and $\theta_{(1),2}$). The down-set of $\{\mathrm{Max}_{\preceq}\mathcal{C}^*\}$ is shown in Figure 5(b). It has three minimal multicuts: $\{\theta_{(1),1}, \theta_{(3),1}\}$, $\{\theta_{(1),1}, \theta_{(2),2}\}$, and $\{\theta_{(3),1}, \theta_{(2),2}\}$.

As illustrated by the example, there will generally be several minimal multicuts. We can distinguish between *locally* and *globally* minimal multicuts.

Definition 9 *Let $\mathcal{M}$ be a multicut of $\mathcal{F}^*$. $\mathcal{M}$ is* locally minimal *if for all $\theta \in \mathcal{M}$, the set $\mathcal{M}\backslash\{\theta\}$ is not a multicut of $\mathcal{F}^*$. $\mathcal{M}$ is* globally minimal *if*

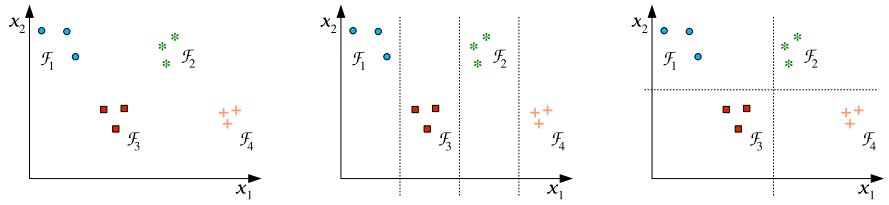
$$|\mathcal{M}| = \min_{\tilde{\mathcal{M}} \in \mathcal{M}^*_{min}} |\tilde{\mathcal{M}}|, \tag{19}$$

where $\mathcal{M}^*_{min}$ is the set of all locally minimal multicuts of $\mathcal{F}^*$.

Proposition 7 *The elements of $Min_{\subseteq}\downarrow\{Max_{\preceq}\mathcal{C}^*\}$ are locally minimal multicuts.*

Proof. $\text{Min}_{\subseteq} \downarrow \{\text{Max}_{\preceq} \mathcal{C}^*\}$ is the set of the minimal multicuts (by inclusion) of the down-set of the multicut $\text{Max}_{\preceq} \mathcal{C}^*$. By contradiction, assume that there exists $\mathcal{M} \in \text{Min}_{\subseteq} \downarrow \{\text{Max}_{\preceq} \mathcal{C}^*\}$ and $\theta \in \mathcal{M}$ such that $\mathcal{M} \setminus \{\theta\}$ is a multicut. Then $\mathcal{M} \setminus \{\theta\}$ belongs to $\downarrow \{\text{Max}_{\preceq} \mathcal{C}^*\}$ and hence $\mathcal{M}$ is not minimal by inclusion. It contradicts our assumption: as a consequence all multicuts of $\text{Min}_{\subseteq} \downarrow \{\text{Max}_{\preceq} \mathcal{C}^*\}$ are locally minimal. □

The elements of $\text{Min}_{\subseteq} \downarrow \{\text{Max}_{\preceq} \mathcal{C}^*\}$ are locally minimal multicuts, but they are not necessarily globally minimal. We show this fact using the following illustration. In Figure 7, both multicuts $\mathcal{M}_1$ and $\mathcal{M}_2$ (shown in Figures 7(b) and 7(c), respectively) are locally minimal. However, since $|\mathcal{M}_1| = 3$, $|\mathcal{M}_2| = 2$ and there is no singleton multicut of $\mathcal{F}^*$, $\mathcal{M}_2$ is globally minimal.



(a) (b) (c)

Figure 7: Locally and globally minimal multicuts of $\mathcal{F}^*$. (a) Data sets $\mathcal{F}^*$. (b) Locally minimal multicut $\mathcal{M}_1$ of $\mathcal{F}^*$. (c) Globally minimal multicut $\mathcal{M}_2$ of $\mathcal{F}^*$.

The above considerations lead us to a final refinement of the problem statement:

$$\text{find } \textit{all} \text{ globally minimal multicuts in } \text{Min}_{\subseteq} \downarrow \{\text{Max}_{\preceq} \mathcal{C}^*\}. \tag{20}$$

6 Algorithms for computing switching thresholds

In this section we present an approach for computing the multicuts satisfying criterion (20), and thus inferring the minimal set of switching thresholds for a PWA model of a genetic regulatory network from a classified data set $\mathcal{F}^*$. In particular, we introduce algorithms for computing the set of all cuts ($\mathcal{C}^*$), the set of maximal cuts ($\text{Max}_{\preceq} \mathcal{C}^*$), and the globally minimal multicuts (i.e. the multicuts in $\text{Min}_{\subseteq} \downarrow \{\text{Max}_{\preceq} \mathcal{C}^*\}$ that are globally minimal).

6.1 Computing cuts and maximal cuts

The computation of the set of all cuts is based on Definition 5 and requires the enumeration of all equivalence classes. Before giving an algorithm for accomplishing this task, we provide an example based on Figure 3(b). Consider the first direction and define $a_0 = \max_{x \in \mathcal{F}_1} x_1$ and $a_5 = \min_{x \in \mathcal{F}_2} x_1$. The ordered list $I_1 : a_0 \leq a_1 \leq \ldots \leq a_5$ can be readely obtained by computing the quantities $\max_{x \in \mathcal{F}_i} x_1$ and $\min_{x \in \mathcal{F}_i} x_1$ for all sets $\mathcal{F}_i$ belonging to $[a_0, a_5]$. Using the results in Example 1, the intervals $\text{G}(\mathcal{E})$ associated to equivalence classes can be obtained discarding the extreme points a_0 and a_5 from I_1 and building intervals defined by consecutive elements in I_1 (i.e. $]a_1, a_2[$, $[a_2, a_3]$ and $]a_3, a_4[$). The cuts associated to each equivalence class are found computing the midpoints of these intervals and are represented in Figure 3(c). We highlight that this procedure will produce *all* cuts in the first direction. We also stress that as far as one is interested in computing cuts it is unnecessary to assess if the intervals $\text{G}(\mathcal{E})$ include or not their extreme points.

This method can be generalized in order to consider also the case of equivalence classes made by a single point, as in the example of Figure 4(b). The overall procedure is summarized in Algorithm 1.

Algorithm 1 *Find the set of all cuts $\mathcal{C}^*$ of $\mathcal{F}^*$*

Set $\mathcal{C}^* = \emptyset$. Initialize the lists $m_i = \emptyset$ and $M_i = \emptyset$, for all $i = 1, \ldots, n$,
for** $i = 1, \ldots, n$ **do
  for** $j = 1, \ldots, s$ **do
    $m_i = m_i \cup \{\min_{x \in \mathcal{F}_j} x_i\}$,
    $M_i = M_i \cup \{\max_{x \in \mathcal{F}_j} x_i\}$.

6: **end for**
7: **for** $j = 1, \ldots, s$ **do**
8: **if** $m_i(j) < \min(M_i)$ **then**
9: $m_i = m_i \setminus \{m_i(j)\}$.
10: **end if**
11: **if** $M_i(j) > \max(m_i)$ **then**
12: $M_i = M_i \setminus \{M_i(j)\}$.
13: **end if**
14: **end for**
15: $I_i = m_i \cup M_i$.
16: *Sort the elements in* I_i *in an ascending order.*
17: **for** $k = 1, \ldots, |I_i| - 1$ **do**
18: *Let* θ *be the ap-hyperplane such that* $dir(\theta) = i$ *and* $\mathrm{Z}(\theta) = I_i(k) + \frac{I_i(k+1) - I_i(k)}{2}$, *add the cut* θ *to* $\mathcal{C}^*$.
19: **if** $k > 1$ *and* $I_i(k) \in m_i$ *and* $I_i(k) \in M_i$ **then**
20: *Let* θ *be the ap-hyperplane such that* $dir(\theta) = i$ *and* $\mathrm{Z}(\theta) = I_i(k)$, *add the cut* θ *to* $\mathcal{C}^*$.
21: **end if**
22: **end for**
23: **end for**

The next proposition shows the correctness of Algorithm 1.

Proposition 8 *The set* $\mathcal{C}^*$ *computed by Algorithm 1 contains all cuts associated to* $\mathcal{F}^*$.

Proof. It is enough to prove that the lines 3-22 of the Algorithm compute all cuts in a given direction i. Indeed, if this is true, the outer loop starting on line 2 will find all cuts in all directions. In order to simplify the notations, for $j \in \{1, \ldots, s\}$, we define :

$$\begin{aligned} m_{ij} &= \min_{x \in \mathcal{F}_j} x_i, \\ M_{ij} &= \max_{x \in \mathcal{F}_j} x_i, \\ \tilde{m}_i &= \{m_{ij} : m_{ij} < \min\left(\{M_{ij}\}_{j=1}^s\right)\}, \\ \tilde{M}_i &= \{M_{ij} : M_{ij} > \max\left(\{m_{ij}\}_{j=1}^s\right)\}. \end{aligned}$$

For sake of clarity, we split the proof in different statements. The first one concerns a key property of the set I_i computed in lines 3-15 of the algorithm.
Statement 1: If $\exists \mathcal{E} \in \mathcal{E}^*$, $\mathrm{dir}(\mathcal{E}) = i$, *then the boundaries of* $\mathrm{G}(\mathcal{E})$ *are points in* $I_i = m_i \cup M_i$, *where* $m_i = \{m_{ij}\}_{j=1}^s \setminus \tilde{m}_i$ *and* $M_i = \{M_{ij}\}_{j=1}^s \setminus \tilde{M}_i$.
Proof of statement 1. Since $\mathcal{E} \in \mathcal{E}^*$ one has $S(\mathcal{E}) \neq \emptyset$ and from the definitions of g_- and g_+, it follows that $\exists (p, q) \in U : g_-(\mathcal{E}) = M_{ip}$ and $g_+(\mathcal{E}) = m_{iq}$. Moreover, from (12) one has $M_{ip} < m_{iq}$ that implies $M_{ip} \notin \tilde{M}_i$ and $m_{iq} \notin \tilde{m}_i$. If $\mathcal{I}_0(\mathcal{E}) = \emptyset$, from Proposition 2 and (13) one has that $\mathrm{G}(\mathcal{E}) =]M_{ip}, m_{iq}[$ and the statement is proved. If $\mathcal{I}_0(\mathcal{E}) \neq \emptyset$, from the definitions of g_{0-} and g_{0+} it follows that $\exists (p', q') \in \mathcal{I}_0(\mathcal{E})^2$ such that $g_{0-}(\mathcal{E}) = m_{ip'}$ and $g_{0+}(\mathcal{E}) = M_{iq'}$. Then, $\mathrm{G}(\mathcal{E}) =]M_{ip}, m_{iq}[\cap [m_{ip'}, M_{iq'}]$. Let $\alpha_1,\ \alpha_2,\ \alpha_1 \leq \alpha_2$ denote the extremes of $\mathrm{G}(\mathcal{E})$. If $m_{ip'} \leq M_{ip}$, then $\alpha_1 = M_{ip} \notin \tilde{M}_i$. Moreover, if $M_{ip} < m_{ip'}$, one has $m_{ip'} \notin \tilde{m}_i$ and also $\alpha_1 = m_{ip'}$. Similarly, if $m_{iq} \leq M_{iq'}$, then $\alpha_2 = m_{iq} \notin \tilde{m}_i$ and if $m_{iq} > M_{iq'}$, one has $M_{iq'} \notin \tilde{m}_i$ and also $\alpha_2 = M_{iq'}$.□

Next, we motivate the fact that the algorithm skips the computation of the cuts (that is done in lines 18-21) when I_i is empty.
Statement 2: The set I_i *is empty if and only if there is no equivalence class in* $\mathcal{E}^*$ *with direction* i.
Proof of statement 2. The emptiness of I_i means that $\nexists (p, q) \in \{1, \ldots s\}^2$ such that $M_{ip} < m_{iq}$ or, equivalently, that no pair of sets $\mathcal{F}_p$ and $\mathcal{F}_q$ are separable along the i-th direction. If $\nexists \mathcal{E} \in \mathcal{E}^*$ with $dir(\mathcal{E}) = i$ one has that $\nexists (p, q) \in \{1, \ldots s\}^2$ such that $\mathcal{F}_p$ and $\mathcal{F}_q$ are separable along the i-th direction. Then, $\forall (p, q) \in \{1, \ldots s\}^2$ it holds $M_{ip} > m_{iq}$ implying that $M_{ip} \in \tilde{M}_i$ and $m_{iq} \in \tilde{m}_i$. It follows that $m_i = M_i = \emptyset$ and then $I_i = M_i \cup m_i = \emptyset$. □

The last step, is to show the correctness of lines 18-21 that compute the cuts.
Statement 3: Let $I_i \neq \emptyset$ *be sorted in ascending order. An ap-hyperplane* θ *is a cut with direction* i *if and only if one of the following conditions is verified:*

a. $\mathrm{Z}(\theta) = I_i(k) + \frac{I_i(k+1) - I_i(k)}{2}$

b. $k > 1$, $k < |I_i|$, $I_i(k) \in m_i$, $I_i(k) \in M_i$ *and* $\mathrm{Z}(\theta) = I_i(k)$.

Proof of statement 3. Associated to an ap-hyperplane, we introduce the following sets that will be useful in the sequel:

$$M_>(\theta) = \{M_{ij} \in I_i : M_{ij} > \mathrm{Z}(\theta)\}, \quad M_<(\theta) = \{M_{ij} \in I_i : M_{ij} < \mathrm{Z}(\theta)\},$$
$$m_>(\theta) = \{m_{ij} \in I_i : m_{ij} > \mathrm{Z}(\theta)\}, \quad m_<(\theta) = \{m_{ij} \in I_i : m_{ij} < \mathrm{Z}(\theta)\}.$$

($\Rightarrow$) Assume that θ is a cut and let $\mathcal{E} = [\theta]$. As shown in the proof of statement 1, $g_-(\mathcal{E})$ and $g_+(\mathcal{E})$ are elements of I_i and so are $g_{0-}(\mathcal{E})$ and $g_{0+}(\mathcal{E})$, when they are defined. We need to show that the boundaries of $\mathrm{G}(\mathcal{E})$ are consecutive or equal elements of I_i. Let $(\alpha_1, \alpha_2) \in {I_i}^2$ such that $[\alpha_1, \alpha_2] = cl(\mathrm{G}(\mathcal{E}))$ (where cl is the set closure). Then, $\forall (p,q) \in \mathcal{I}_-(\mathcal{E}) \times \mathcal{I}_+(\mathcal{E})$, $m_{ip} \leqslant M_{ip} \leqslant \alpha_1 \leqslant \alpha_2 \leqslant m_{iq} \leqslant M_{iq}$. If $\mathcal{I}_0(\mathcal{E}) = \emptyset$, then $\mathcal{I}_-(\mathcal{E}) \cup \mathcal{I}_+(\mathcal{E}) = \{1, \ldots, s\}$, so that there is no element of I_i between α_1 and α_2. If $\mathcal{I}_0(\mathcal{E}) \neq \emptyset$, then we deduce from the intersection in (13) that the closure of $\mathrm{G}(\mathcal{E})$ can also be written as $[\alpha_1, \alpha_2] = [\max(g_-(\mathcal{E}), g_{0-}(\mathcal{E})), \min(g_+(\mathcal{E}), g_{0+}(\mathcal{E}))]$. Consequently, we also have that $\forall (j,j') \in \mathcal{I}_0(\mathcal{E})^2$, $m_{ij} \leqslant \alpha_1 \leqslant \alpha_2 \leqslant M_{ij'}$. Again, there is no element of I_i between α_1 and α_2.

In order to conclude the proof, we treat separately the case when $\mathcal{E}$ is a singleton and the case when $\mathcal{E}$ is not a singleton. If $\mathcal{E}$ is a singleton, from Proposition 2 one has that $\mathrm{G}(\mathcal{E})$ is a point and this is possible when $g_{0-}(\mathcal{E}) = g_{0+}(\mathcal{E})$. Since, as shown in the proof of statement 1, $\exists (p,q) \in \{1, \ldots s\}^2$ such that $g_{0-}(\mathcal{E}) = m_{ip}$, and $g_{0+}(\mathcal{E}) = M_{iq}$, since the unique element of $\mathcal{E}$ is such that $\mathrm{Z}(\theta) = g_{0-}(\mathcal{E}) = g_{0+}(\mathcal{E})$, the conditions $I_i(k) \in m_i$, $I_i(k) \in M_i$ and $\mathrm{Z}(\theta) = I_i(k)$ in point (b) are fulfilled for some k. Assume by contradiction that they are verified for $k = 1$. Then, $M_<(\theta) = \emptyset$ and hence $\mathcal{I}_-(\theta) = \emptyset$ leading to the conclusion that θ is not a cut. Analogously, if $k = |I_i|$, then, $m_>(\theta) = \emptyset$ and hence $\mathcal{I}_+(\theta) = \emptyset$ implying again that θ is not a cut. If $\mathcal{E}$ is not a singleton, then $\alpha_1 < \alpha_2$ and hence $(\alpha_1, \alpha_2) = (I_i(k), I_i(k+1))$ for some k. In this case, the fomula of $\mathrm{Z}(\theta)$ in point (a) follows from Definition 5.

($\Leftarrow$) We first prove that if $I_i = \emptyset$, then $|I_i| \geq 2$. $I_i \neq \emptyset$ implies that there exists at least an equivalence class $\mathcal{E} \in \mathcal{E}^*$ with direction i (see Statement 2). Moreover, both $g_-(\mathcal{E})$ and $g_+(\mathcal{E})$ must be points in I_i and since $g_-(\mathcal{E}) < g_+(\mathcal{E})$, we have shown that $|I_i| \geq 2$.

We assume now that point (a) holds. In this case we set $\alpha_1 = I_i(k) < I_i(k+1) = \alpha_2$ and we show that α_1, α_2 are the boundaries of the interval $\mathrm{G}(\mathcal{E})$ for some $\mathcal{E} \in \mathcal{E}^*$ (indeed, this implies that the ap-hyperplane θ with $\mathrm{Z}(\theta)$ computed as in point (a) is a cut). We consider a pair of ap-hyperplanes (θ, θ') with direction i and assume that $\mathrm{Z}(\theta) \in]\alpha_1, \alpha_2[$. Note that the sets $M_<(\theta)$ and $m_>(\theta)$ are nonempty and then $\exists (p,q) \in U$ such that $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$. If $\mathrm{Z}(\theta') \in]\alpha_1, \alpha_2[$, by using the fact that α_1 and α_2 are consecutive elements of I_i, one has

$$M_<(\theta') = M_<(\theta), \quad m_>(\theta') = m_>(\theta), \quad M_>(\theta') = M_>(\theta), \quad m_<(\theta') = m_<(\theta) \tag{21}$$

The first equality in (21) implies that $\mathcal{I}_-(\theta') = \mathcal{I}_-(\theta)$, the second equality implies that $\mathcal{I}_+(\theta') = \mathcal{I}_+(\theta)$ and the last two equalities yield $\mathcal{I}_0(\theta') = \mathcal{I}_0(\theta)$. Hence $\theta \sim \theta'$.

On the other hand, if $\mathrm{Z}(\theta') \notin [\alpha_1, \alpha_2]$, from the fact that α_1 and α_2 are consecutive elements of I_i we deduce that at least one equality in (21) is not verified and hence $\theta \not\sim \theta'$. This argument shows that α_1, α_2 are the boundaries of $\mathrm{G}(\mathcal{E})$.

In the last part of the proof, we assume that point (b) holds. First, since $k > 1$ and $k < |I_i|$, one has that $I(k) \in (m_i \cap M_i) \setminus \{\min I_i, \max I_i\}$. This shows that the sets $M_<(\theta)$ and $m_>(\theta)$ are nonempty and then $\exists (p,q) \in U$ such that $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$. We show now that $[\theta] = \{\theta\}$. From point (b), there exists $(p,q) \in \{1, \ldots, s\}^2$ such that $I_i(k) = m_{ip} = M_{iq}$ that is $p \in \mathcal{I}_0([\theta])$ and $q \in \mathcal{I}_0([\theta])$. We show now that no other ap-hyperplane is equivalent to θ. If θ' is an ap-hyperplane with $dir(\theta) = i$ and $\min I_i < \mathrm{Z}(\theta') < \mathrm{Z}(\theta)$, then it belongs to Θ (because $\mathrm{Z}(\theta) > \min I_i$) and verifies $q \in \mathcal{I}_+(\theta')$. This means that $\theta' \not\sim \theta$. On the other hand, if θ' is an ap-hyperplane with $dir(\theta') = i$ and $\max I_i > \mathrm{Z}(\theta') > \mathrm{Z}(\theta)$, then it belongs to Θ (because $\mathrm{Z}(\theta) < \max I_i$) and verifies $p \in \mathcal{I}_-(\theta')$. As in the previous case, one obtains $\theta' \not\sim \theta$.

□

The computation of the set of maximal cuts considers $\mathcal{C}^*$ separately in each of the n directions, using the restriction of the poset introduced in Section 4.2. The procedure is summarized in Algorithm 2.

Algorithm 2 *Compute* $Max_{\preceq} \mathcal{C}^*$

1: *Set* $\bar{\mathcal{C}}^{i*} = \emptyset$, $i = 1, \ldots, n$.
2: ***for*** $i = 1, \ldots, n$ ***do***
3: ***for*** $k = 1, \ldots, |\mathcal{C}^{i*}|$ ***do***
4: *Set* $max\_flag = true$.
5: ***for*** $l = \{1, \ldots, |\mathcal{C}^{i*}|\} \setminus \{k\}$ ***do***

Let θ be the k-th cut of $\mathcal{C}^{i*}$ and θ' the l-th one:
  if $S(\theta) \subset S(\theta')$ **then**
    Set $max\_flag = false$.
  end if
end for
if $max\_flag = true$ **then**
  Add θ to $\bar{\mathcal{C}}^{i*}$.
end if
end for
end for
Set $Max_{\preceq}\mathcal{C}^* = \cup_{i=1}^{n}\bar{\mathcal{C}}^{i*}$.

6.2 Computing globally minimal multicuts

In order to compute the globally minimal multicuts, we could in principle enumerate all subsets of $Max_{\preceq}\mathcal{C}^*$ and verify minimality by means of Definitions 7 and 9. However, this procedure is computationally prohibitive even for simple examples. Therefore, in the sequel, we present an additional result on multicuts that will allow us to propose more efficient methods for finding globally minimal multicuts, based on branch-and-bound techniques.

Definition 10 (Redundancy) *Let $\mathcal{M}$ be a multicut of $\mathcal{F}^*$. A cut $\theta \in \mathcal{M}$ is* redundant *in $\mathcal{M}$, if $S(\theta) \subseteq \cup_{\theta' \in \mathcal{M}\setminus\{\theta\}} S(\theta')$.*

In the example of Figure 3, each of the three cuts in the multicut $\{\theta_{(1),1}, \theta_{(3),1}, \theta_{(2),2}\}$ is redundant. The following proposition shows that redundant cuts can be safely ignored.

Proposition 9 *A multicut $\mathcal{M}$ of $\mathcal{F}^*$ is locally minimal if and only if no $\theta \in \mathcal{M}$ is redundant in $\mathcal{M}$.*

Proof. ($\Rightarrow$) Assume, by contradiction, that there exists a $\theta \in \mathcal{M}$ that is redundant. Then, from Definition 10, one has

$$U = \cup_{\tilde{\theta} \in \mathcal{M}} S(\tilde{\theta}) = \cup_{\tilde{\theta} \in \mathcal{M}\setminus\{\theta\}} S(\tilde{\theta}),$$

thus showing that $\mathcal{M}\setminus\{\theta\}$ is a multicut of $\mathcal{F}^*$ (because of Proposition 4). Then $\mathcal{M}$ is not locally minimal.

($\Leftarrow$) Assume, by contradiction, that the multicut $\mathcal{M}$ is not locally minimal. Then, there exists a cut $\theta \in \mathcal{M}$ such that $\tilde{\mathcal{M}} = \mathcal{M}\setminus\{\theta\}$ is a multicut of $\mathcal{F}^*$. But since θ is not redundant, the set $\tilde{S} = S(\theta)\setminus \cup_{\tilde{\theta} \in \tilde{\mathcal{M}}} S(\tilde{\theta})$ is not empty. This means that there is a pair of indexes $(p, q) \in U$ such that $(p, q) \notin \cup_{\theta \in \tilde{\mathcal{M}}} S(\theta)$. By virtue of Proposition 4, this implies that $\tilde{\mathcal{M}}$ is not a multicut. □

Definition 11 (Kernel) *Let $\mathcal{M}$ be a multicut of $\mathcal{F}^*$. The* kernel *of $\mathcal{M}$ is defined as*

$$\ker(\mathcal{M}) = \{\theta \in \mathcal{M} : \exists u \in S(\theta), \nexists \theta' \in \mathcal{M} \setminus \{\theta\}, u \in S(\theta')\}. \tag{22}$$

Algorithm 3 *Create the set $\mathcal{M}^*_{min}$ of all globally minimal multicuts*
Initialize the global variables $\mathcal{M}^*_{min} = \emptyset$ and $best = |Max_{\preceq}\mathcal{C}^*|$. Initialize $\mathcal{M}_{in} = \ker(Max_{\preceq}\mathcal{C}^*)$
if $U = \cup_{\theta \in \mathcal{M}_{in}} S(\theta)$ **then**
  Append $\ker(Max_{\preceq}\mathcal{C}^*)$ *to $\mathcal{M}^*_{min}$ and exit*
else
  Branch($\mathcal{M}_{in}$)
end if

function Branch($\mathcal{M}_{in}$)
for all $\theta \in Max_{\preceq}\mathcal{C}^*\setminus\mathcal{M}_{in}$ **do**
  if $S(\theta) \not\subseteq \cup_{\theta' \in \mathcal{M}_{in}} S(\theta')$ **then** *//θ is not redundant in $\mathcal{M}_{in} \cup \{\theta\}$.*
    Set $\mathcal{M}_{out} = \mathcal{M}_{in} \cup \{\theta\}$
    if $U = \cup_{\theta' \in \mathcal{M}_{out}} S(\theta')$ **then** *//$\mathcal{M}_{out}$ is a multicut.*
      if $|\mathcal{M}_{out}| = best$ *and* $\mathcal{M}_{out} \notin \mathcal{M}^*_{min}$ **then**
        Append $\mathcal{M}_{out}$ to $\mathcal{M}^*_{min}$
      else if $|\mathcal{M}_{out}| < best$ **then**

```
8:          Set M*_min = {M_out} and best = |M_out|      //Reset M*_min and update best.
9:        end if
10:     else if |M_out| < best then
11:       Branch(M_out)
12:     end if
13:   end if
14: end for
```

From Definition 11, it is apparent that $\ker(Max_{\preceq}\mathcal{C}^*)$ collects the cuts in $\mathcal{M}$ that must belong to every minimal multicut, otherwise at least one pair of sets in $\mathcal{F}^*$ will not be separated. For $\mathcal{M} = \{\theta_{(1),1}, \theta_{(3),1}, \theta_{(2),2}\}$ in the example of Figure 3, $\ker(\mathcal{M})$ is empty: none of the cuts is indispensable.

The notions of redundancy and kernel are used to speed up the branch-and-bound algorithm summarized in Algorithm 3 computing the set $\mathcal{M}^*_{min} \subseteq \mathcal{M}^*$ of globally minimal multicuts. The basic idea is to start with a small subset of $Max_{\preceq}\mathcal{C}^*$, given by $\ker(Max_{\preceq}\mathcal{C}^*)$, and add new cuts iteratively.

During the execution of Algorithm 3, the global variable *best* stores the size of the smaller multicut found so far. If $\ker(Max_{\preceq}\mathcal{C}^*)$ is a multicut, it is also the only globally minimal multicut in $Max_{\preceq}\mathcal{C}^*$ and the algorithm terminates (lines 2 and 3 of the main procedure). Otherwise, the function *Branch* is called in order to add suitable cuts to $\ker(Max_{\preceq}\mathcal{C}^*)$.

At line 2 of the function *Branch*, the addition of a new cut θ to $\mathcal{M}_{in}$ is considered only if θ is not redundant in $\mathcal{M}_{out} = \mathcal{M}_{in} \cup \{\theta\}$ (see Proposition 9). Lines 4-8 process sets $\mathcal{M}_{out}$ that are multicuts and modify the set $\mathcal{M}^*_{min}$ accordingly. More specifically, a multicut of size *best* is added to $\mathcal{M}^*_{min}$ (line 6), while a multicut of size less than *best* causes the reset of the set $\mathcal{M}^*_{min}$ (line 8) and the update of *best*. These operations guarantee that only globally minimal multicuts will be stored in $\mathcal{M}^*_{min}$.

Lines 10 and 11 handle the set $\mathcal{M}_{out}$ when it is not a multicut. The possibility of adding another cut to $\mathcal{M}_{out}$ (through the call to *Branch* at line 11) is explored only if $|\mathcal{M}_{out}| < best$, thus avoiding consideration of multicuts that are larger than the smallest ones already found.

7 Reconstruction of switching thresholds in PWA model of carbon starvation response of *E. coli*

In order to test the applicability of the multicut approach, we have used it for the reconstruction of switching thresholds in a PWA model of the carbon starvation response in the bacterium *Escherichia coli*. In the absence of essential carbon sources, an *E. coli* population abandons exponential growth and enters a non-growth state called *stationary phase*. This growth-phase transition is accompanied by numerous physiological changes in the bacteria, concerning among other things the morphology and the metabolism of the cell, as well as gene expression. On the molecular level, the transition from exponential phase to stationary phase in response to a carbon stress is controlled by a complex genetic regulatory network.

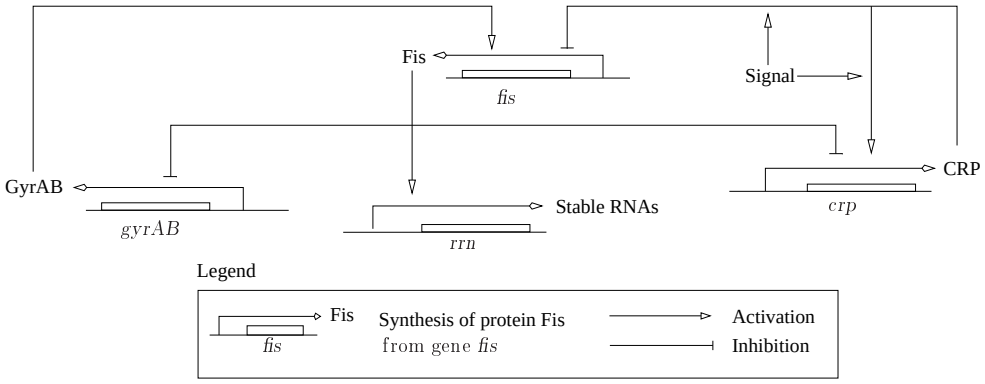
In previous work, we have developed a PWA model of the carbon starvation response in *E. coli* [22]. The model describes how a carbon stress signal is propagated through a network of interactions between global transcriptional regulators of the bacterium, so as to influence the synthesis of stable RNAs and thereby adapt the growth of the cell. For this study, we have used a simplified version of this model (Figure 8), which preserves essential properties of the qualitative dynamics predicted by the original model, as verified by means of the approach described in [1]. In response to a carbon starvation signal, the system switches from an equilibrium point characteristic for exponential growth to another equilibrium point, corresponding to stationary phase. Reentry into exponential phase after a carbon upshift gives rise to a damped oscillation towards the exponential-phase equilibrium point.

In collaboration with the laboratory of Johannes Geiselmann (Université Joseph Fourier, Grenoble), we are in the process of measuring gene expression over time in the absence and presence of carbon sources. The use of reporter genes encoding fluorescent and luminescent proteins makes it possible to obtain precise and densely-spaced measurements of the expression of the genes in the carbon starvation response network. This kind of data is well-suited for system identification purposes, as shown previously in [10, 21]. In this technical report, we use simulated data to test the multicut approach, staying close to the expected noise and sample density of the real measurements.

The numerical simulations have been carried out in Matlab with values for the parameters and initial conditions satisfying the inequality constraints inferred from the experimental literature [22]. The simulations describe the variation of the concentration of the proteins during the growth-phase transitions. Figure 9 gives

$$
\begin{aligned}
\dot{x}_{CRP} &= \kappa^0_{CRP} + \kappa^1_{CRP}\, s^-(x_{Fis}, \theta^1_{Fis})\, s^+(x_{CRP}, \theta^1_{CRP})\, s^+(x_S, \theta_S) - \gamma_{CRP}\, x_{CRP} \\
\dot{x}_{Fis} &= \kappa^1_{Fis}\, (1 - s^+(x_{CRP}, \theta^1_{CRP})\, s^+(x_S, \theta_S)) \\
&\quad + \kappa^2_{Fis}\, s^+(x_{GyrAB}, \theta_{GyrAB})\, (1 - s^+(x_{CRP}, \theta^1_{CRP})\, s^+(x_S, \theta_S)) - \gamma_{Fis}\, x_{Fis} \\
\dot{x}_{GyrAB} &= \kappa_{GyrAB}\, s^-(x_{Fis}, \theta^3_{Fis}) - \gamma_{GyrAB}\, x_{GyrAB} \\
\dot{x}_{rrn} &= \kappa_{rrn}\, s^+(x_{Fis}, \theta^2_{Fis}) - \gamma_{rrn}\, x_{rrn} \\
\dot{x}_S &= 0
\end{aligned}
$$

(a)



(b)

Figure 8: (a) Simplified PWA model of the carbon starvation network in *E. coli* [22]. The variables x_{CRP}, x_{Fis}, x_{GyrAB}, and x_{rrn} denote the concentrations of CRP, Fis, GyrAB, and stable RNAs, while x_S represents the carbon starvation signal ($s^+(x_S, \theta_S) = 1$ means that the carbon starvation signal is present). The variables have been rescaled to the interval $[0, 1]$, and the following parameter values have been used for the simulations: $\theta^1_{CRP} = 0.33$, $\theta^2_{CRP} = 0.67$, $\theta^1_{Fis} = 0.1$, $\theta^2_{Fis} = 0.5$, $\theta^3_{Fis} = 0.75$, $\theta_{GyrAB} = 0.5$, $\theta_{rrn} = 0.5$, $\theta_S = 0.5$, $\gamma_{CRP} = 0.5$, $\gamma_{Fis} = 2$, $\gamma_{GyrAB} = 1$, $\gamma_{rrn} = 1.5$, $\gamma_S = 0.5$, $\kappa^0_{CRP} = 0.25$, $\kappa^1_{CRP} = 0.4$, $\kappa^1_{Fis} = 0.6$, $\kappa^2_{Fis} = 1.15$, $\kappa_{GyrAB} = 0.75$, $\kappa_{rrn} = 1.12$, (b) Graphical representation of the PWA model, indicating genes and their regulatory interactions.

an indication of the data obtained from simulating the reentry into exponential phase after a carbon upshift. In order to separate the threshold reconstruction problem from the classification problem for the purpose of this report, we have generated the correct classification by detecting mode switches during simulation.

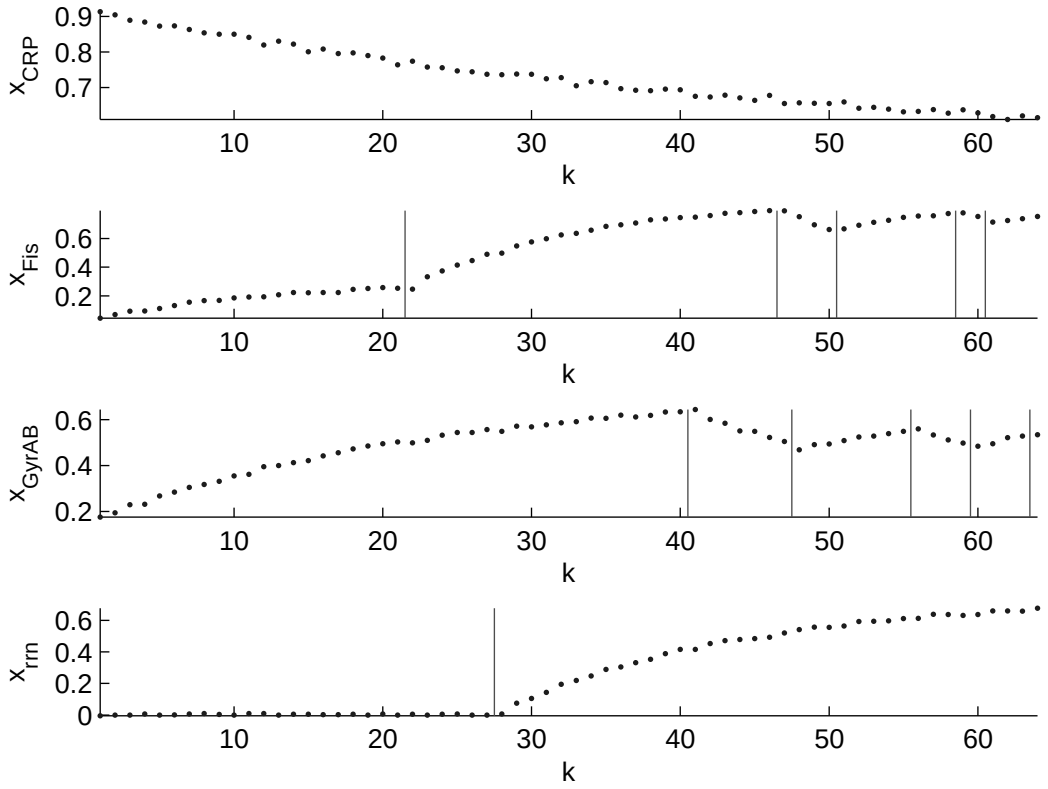


Figure 9: Simulation of the reentry into exponential phase following a carbon upshift, using the PWA model in Figure 8(a). In order to mimic the absence of a carbon stress, $x_S(0)$ has been set to 0. For each protein concentration variable, the mode switches are indicated by means of vertical bars.

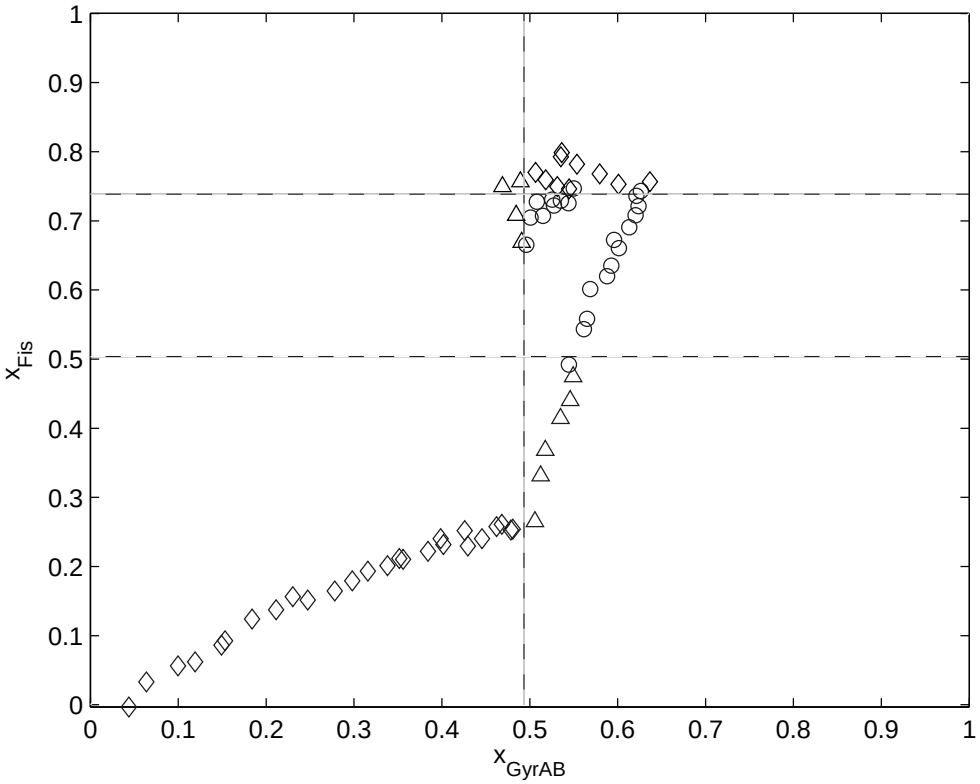
The resulting datasets have been analyzed by means of a Matlab implementation of the algorithms presented in Section 6. The results for the transition from stationary to exponential phase after a carbon upshift are summarized in Figures 10 and 11. The algorithm finds $\mathcal{C}^*$ consisting of six cuts, $\theta_1, \ldots, \theta_6$. In order to get an idea of the separation power of the cuts, Figure 10(b) pictures the projection of the data points on the (x_{Fis}, x_{GyrAB})-subspace. As can be seen, the cuts θ_2, θ_5, and θ_6 nicely separate the classes generated from the damped oscillation (Figure 9).

To each of the cuts corresponds a switching threshold, associated with a regulatory interaction in the network. For instance, one can verify in Figure 9 that when x_{Fis} crosses the threshold value 0.5 from below, the concentration x_{rrn} of stable RNAs starts to increase as well. This motivates the conclusion that the threshold where x_{Fis} equals 0.5 corresponds to the activation of the *rrn* operon by Fis, an interaction that is correctly inferred from the simulation data (Figure 9). Four of the cuts in the maximal multicut correspond to real switching thresholds of the system.

Applying Algorithm 3 to the maximal multicut yields three globally minimal multicuts, shown in Figure 11. Each of the multicuts consists of three cuts, two of which occur in every solution. The cut θ_6 corresponds to the switching threshold above which Fis starts to inhibit the expression of the gene *gyrAB*, while θ_2 represents the switching threshold associated with the activation of *fis* by GyrAB. Notice that the globally minimal multicuts $\mathcal{M}_2$ and $\mathcal{M}_3$ contain only cuts corresponding to correct switching thresholds, whereas for $\mathcal{M}_1$ two out of three thresholds are correct.

Cut	Variable	Threshold value	Interaction	Correct? (Y/N)
θ_1	x_{Fis}	0.26	Fis activates *fis*	N
θ_2	x_{GyrAB}	0.49	GyrAB activates *fis*	Y
θ_3	x_{rrn}	0.03	Stable RNAs activate *rrn*	N
θ_4	x_{CRP}	0.65	CRP inhibits *fis*	Y
θ_5	x_{Fis}	0.5	Fis activates *rrn*	Y
θ_6	x_{Fis}	0.74	Fis inhibits *gyrAB*	Y

(a)



(b)

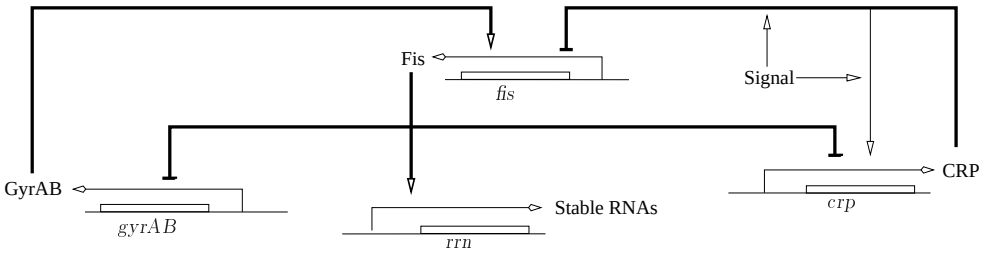
Figure 10: (a) Maximal multicut generated by Algorithm 1 for the data in Figure 9. (b) Illustration of the separation power of the cuts θ_2, θ_5, and θ_6, included in the globally minimal multicut $\mathcal{M}_3$ (Figure 11(a)). The data have been projected on the (x_{Fis}, x_{GyrAB})-subspace.

Repeating the above procedure for the second set of simulation data, corresponding to the entry into stationary phase, yields a maximal multicut consisting of four cuts, two of which correspond to a real switching threshold of the system and the remaining of them are compatible with the available data (results shown in Figure 12). From this information, Algorithm 3 generates four globally minimal multicuts, each composed of two cuts. One of the globally minimal multicuts entirely consists of cuts corresponding to correct switching thresholds, whereas in two cases one of the cuts corresponds to a non-existing threshold, and the last case consists of the spurious cuts.

Summarizing the results of the switching threshold reconstruction process, Figure 11 projects the best globally minimal multicut for the first and second data series on the graphical representation of the carbon starvation network. As can be seen, the multicut approach has inferred five out of six interactions from the data (only the autoactivation of CRP is missed). As for the worst globally minimal multicuts found by the algorithm, they nevertheless achieve the correct identification of three of the switching thresholds in the model. These results confirm the in-principle applicability of our approach.

Multicut	Cuts in multicut	Correct? (Y/N)
$\mathcal{M}_1$	$\{\theta_2, \theta_3, \theta_6\}$	$\{Y, N, Y\}$
$\mathcal{M}_2$	$\{\theta_2, \theta_4, \theta_6\}$	$\{Y, Y, Y\}$
$\mathcal{M}_3$	$\{\theta_2, \theta_5, \theta_6\}$	$\{Y, Y, Y\}$

(a)



(b)

Figure 11: (a) Globally minimal multicuts generated by Algorithm 3 from the maximal multicut in Figure 10. (b) Interactions (in bold) correctly identified by the best globally minimal multicuts obtained from the data for the transition to exponential phase after a carbon upshift ($\mathcal{M}_2$ in part (a) of the figure) and the entry into stationary phase ($\mathcal{M}_1$ in Figure 12(b)).

Cut	Variable	Threshold value	Interaction	Correct? (Y/N)
θ_1	x_{Fis}	0.52	Fis activates *rrn*	Y
θ_2	x_{GyrAB}	0.55	GyrAB inhibits *rrn*	N
θ_3	x_{Fis}	0.10	Fis inhibits *crp*	Y
θ_4	x_{rrn}	0.17	Stable RNAs inhibit *crp*	N

(a)

Multicut	Cuts in multicut	Correct? (Y/N)
$\mathcal{M}_1$	$\{\theta_1, \theta_3\}$	$\{Y, Y\}$
$\mathcal{M}_2$	$\{\theta_1, \theta_4\}$	$\{Y, N\}$
$\mathcal{M}_3$	$\{\theta_2, \theta_3\}$	$\{N, Y\}$
$\mathcal{M}_4$	$\{\theta_2, \theta_4\}$	$\{N, N\}$

(b)

Figure 12: (a) Maximal multicut for the data of the second set of simulation (b) Globally minimal multicuts generated by Algorithm 3 from the maximal multicut in Figure 12(a).

8 Conclusions

In this technical report we have proposed a pattern recognition technique for reconstructing all combinations of switching thresholds that are consistent with measured data in PWA models of genetic regulatory networks. We have shown how to recast this problem into finding all globally minimal multicuts of maximal cuts that separate different sets of points within a given collection. This algorithm is intended to be used jointly with hybrid identification procedures for classifying the data (*i.e.*, partitioning temporal gene expression data into subsets associated with different regulatory modes) and for reconstructing the values of synthesis/degradation parameters characterizing the dynamics of the network in different regulatory domains. Indeed, data classification is a preliminary requirement for estimating the switching thresholds, and complete models of genetic regulatory networks can be inferred only from the joint estimation of thresholds and synthesis/degradation parameters.

A potential pitfall of the multicut approach is that Algorithms 1, 2, and 3 have been derived under the assumption that the sets of points considered are m-separable. Although this assumption is satisfied in the example of Section 7, it may be violated in other situations for two main reasons. The first one is that noisy data may affect the quality of the results obtained through hybrid systems identification, and lead to a misclassification of some data points [14]. The second reason is that genetic regulatory networks may exhibit the same dynamics on different regulatory domains, a fact that may result in a structural loss of m-separability. However, we stress that even if some pairs of sets are not separable, this does not prevent the multicut algorithm from finding *some* of the thresholds. Most importantly, the m-separability assumption can be verified right after the execution of Algorithm 1. We also believe that even if the mathematical framework for multicuts developed in Sections 4 to 6 is tailored to an idealized case, it provides a sound background for developing new methods capable of dealing with m-inseparable collections of sets.

Acknowledgments: This research has been supported by the European Commission under project HYGEIA (NEST-4995).

References

[1] G. Batt, D. Ropers, H. de Jong, J. Geiselmann, M. Page, and D. Schneider. Qualitative analysis and verification of hybrid models of genetic regulatory networks: Nutritional stress response in *Escherichia coli*. In M. Morari and L. Thiele, editors, *Proc. Hybrid Systems: Computation and Control (HSCC 2005)*, volume 3414 of *LNCS*, pages 134–150. Springer-Verlag, Berlin, 2005.

[2] C. Belta, P. Finin, L.C.G.J.M. Habets, A.M. Halász, M. Imiliǹski, R.V. Kumar, and H. Rubin. Understanding the bacterial stringent response using reachability analysis of hybrid systems. In R. Alur and G.J. Pappas, editors, *Proc. Hybrid Systems: Computation and Control (HSCC 2004)*, volume 2993 of *LNCS*, pages 111–125. Springer-Verlag, Berlin, 2004.

[3] K.P. Bennett and O.L. Mangasarian. Multicategory discrimination via linear programming. *Optimization Methods and Software*, 3:27–39, 1993.

[4] T. Chen, H.L. He, and G.M. Church. Modeling gene expression with differential equations. In R.B. Altman, K. Lauderdale, A.K. Dunker, L. Hunter, and T.E. Klein, editors, *Pac. Symp. Biocomput., PSB'99*, volume 4, pages 29–40, 1999.

[5] H. de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *J. Comput. Biol.*, 9(1):67–103, 2002.

[6] H. de Jong, J.-L. Gouzé, C. Hernandez, M. Page, T. Sari, and J. Geiselmann. Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bull. Math. Biol.*, 66(2):301–340, 2004.

[7] P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: From co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, 2000.

[8] R. Edwards, H.T. Siegelmann, K. Aziza, and L. Glass. Symbolic dynamics and computation in model gene networks. *Chaos*, 11(1):160–169, 2001.

[9] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari. A clustering technique for the identification of piecewise affine and hybrid systems. *Automatica*, 39(2):205–217, Feb 2003.

[10] T.S. Gardner, D. di Bernardo, D. Lorenz, and J.J. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301(5629):102–105, 2003.

[11] R. Ghosh and C.J. Tomlin. Symbolic reachable set computation of piecewise affine hybrid automata and its application to biological modelling: Delta-Notch protein signalling. *Syst. Biol.*, 1(1):170–183, 2004.

[12] L. Glass and S.A. Kauffman. The logical analysis of continuous non-linear biochemical control networks. *J. Theor. Biol.*, 39(1):103–129, 1973.

[13] J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K.N. Kozlov, Manu, E. Myasnikova, C.E. Vanario-Alonso, M. Samsonova, D.H. Sharp, and J. Reinitz. Dynamic control of positional information in the early *Drosophila* embryo. *Nature*, 430(6997):368–371, 2004.

[14] A.Lj Juloski, W.P.M.H. Heemels, G. Ferrari-Trecate, R. Vidal, S. Paoletti, and J.H.G. Niessen. Comparison of four procedures for the identification of hybrid systems. In M. Morari and L. Thiele, editors, *Proc. Hybrid Systems: Computation and Control (HSCC-05)*, volume 3414 of *LNCS*, pages 354–369. Springer-Verlag, Berlin, 2005.

[15] S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi, and M. Tomita. Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics*, 19(5):643–650, 2003.

[16] S. Lemeille, A. Latifi, and J. Geiselmann. Inferring the connectivity of a regulatory network from mRNA quantification in *Synechocystis* PCC6803. *Nucleic Acids Res.*, 33(10):3381–3389, 2005.

[17] D.J. Lockhart and E.A. Winzeler. Genomics, gene expression and DNA arrays. *Nature*, 405(6788):827–836, 2000.

[18] T. Mestl, E. Plahte, and S.W. Omholt. A mathematical framework for describing and analysing gene regulatory networks. *J. Theor. Biol.*, 176(2):291–300, 1995.

[19] T.J. Perkins, M. Hallett, and L. Glass. Inferring models of gene expression dynamics. *J. Theor. Biol.*, 230(3):289–299, 2004.

[20] J. Roitman. *Introduction to Modern Set Theory.* Wiley-IEEE, 1990.

[21] M Ronen, R Rosenberg, B.I. Shraiman, and U. Alon. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. USA*, 99(16):10555–10560, 2002.

[22] D. Ropers, H. de Jong, M. Page, D. Schneider, and J. Geiselmann. Qualitative simulation of the carbon starvation response in *Escherichia coli*. *BioSystems*, 2006. In press.

[23] Y. Shav-Tal, R.H. Singer, and X. Darzacq. Imaging gene expression in single living cells. *Nat. Rev. Mol. Cell Biol.*, 5(10):855–861, 2004.

[24] E.P. van Someren, L.F.A. Wessels, and M.J.T. Reinders. Linear modeling of genetic networks from experimental data. In R. Altman and et *al.*, editors, *Proc. Eight Int. Conf. Intell. Syst. Mol. Biol., ISMB 2000*, pages 355–366, Menlo Park, CA, 2000. AAAI Press.

[25] V. Vapnik. *Statistical Learning Theory.* John Wiley, NY, 1998.

Contents

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-0803