Lesson 11

Conclusion:* About "Emergent Models"

Statistical methods seem to be efficient tools for data analysis. But will these methods always be inferior to first-principles models — are they only describing *surface-level* reflections of internal phenomena, can they ever capture the true *essence* of systems?

What is this "essence", then? Modeling is about hiding details and concentrating information, one has to abstract away irrelevant details. Again, when determining what *is* irrelevant, one is facing ontological assumptions. We already know how to model simple systems, but when studying *complex systems,* new ways of thinking are needed.

This final chapter tries to illustrate the possibilities that may someday come true. It is shown here how the multivariate statistical methods can perhaps offer new conceptual tools for mastering the complexity in systems. It is the *differences that make a difference:* If there exist phenomena that cannot be seen in observations, they can be ignored. And it is the multivariate methods that can capture such phenomena — if the way of looking at systems is adjusted in an approprite way. The traditional methods only capture narrow projections of the behavioral wealth, whereas the multivariate methods can give a more *holistic* view. This view is presented in closer detail in [**?**]; here, only excerpts from there are reviewed.

11.1 Capturing semantics in data

To make it possible to apply multivariate methods for capturing the system essence, the data needs to be defined so that the phenomena of relevance are represented there. The key question is: How to capture the essential information, or *domain-area semantics* in the data? To define data so that the important features are available there for further modeling, one needs a concrete application area. Here, the application area throughout this chapter is the *realm of chemical systems.*

11.1.1 What is "semantics"?

The model should be an interface between the system and outside world, providing best possible information transfer. The model structure should be a compromise between the properties of the system and the properties of the applications. What are the model structures like that support the new tools and new ways of thinking, simultaneously taking into account the system itself?

When searching for *good models,* philosophical questions cannot be avoided: It is such modeling issues that have been studied for millennia — what is the nature of systems, and and how they should be represented. Indeed, what there is, what one can we know about them, these problem fields are called *ontology* and *epistemology,* respectively. Earlier in this report, ontological questions have been discussed in simple terms — now these discussions need to be extended slightly. Here all these mutually related issues are collected under the common concept of *semantics:* What is the essence of a system, and how this essence should be interpreted?

Semantics conveys *meaning.* Traditionally, it is thought that semantics cannot exist outside human brain. However, to reach "smart models" that can adapt in new environments, one needs to make this meaning machine-readable and machine-understandable. Otherwise, no abstraction of relevant vs. irrelevant phenomena can be automatically carried out. Indeed, one is facing a huge challenge here, but something *can* be done.

Just as was done earlier when ontologies were studied, now this semantics is formalized: This very abstract concept is given here very concrete contents, compromizing between intuitions (what would be nice) and reality (what can be implemented in reality). It can even be said that a *good model formalizes the semantics of the domain field, making it visible.* Now there are two levels of semantics to be captured:

1. **Low-level semantics.** The formless complexity of the underlying system has to be captured in concrete homogeneous data. The "atoms" of semantics constitute the connection between the numeric representations and the physical realm, so that the properties of the system are appropriately coded and made visible to the higher-level machineries. In concrete terms, one has to define "probes" and put them in the system appropriately. The measurements delivered by the probes still need to be interpreted, or features need to be extracted from the measurements by applying appropriate data preprocessing.

2. **Higher-level semantics.** The high number of structureless low-level features have to be connected into *structures* of semantic atoms. Assuming that the semantic atoms are available, this higher-level task is *simpler,* being more generic. In our numbers-based environments, a practical approach towards such *contextual semantics,* where relevant lower-level structures are to be appropriately combined, is again offered by correlations-based measures. As has been shown before, assuming that information is conveyed in co-variations among data, structuring of lower-level data can be implemented by the mathematical machinery without need of outside expert guidance.

Indeed, analyses of this higher-level semantics processing have been carried out already a lot in this report, and they can be implemented implicitly by the presented multivariate statistical tools. But representation of the low-level domain-area features is domain-area specific, and needs to be studied separately in each case. To have a solid grounding, one somehow needs to limit the overwhelming diversity of available measurements by applying some assumptions concerning the nature of systems being studied.

11.1.2 Neocybernetic starting points

The traditional models need to be explicitly controlled by the domain area expert, and the structure needs to be determined before the machinery (identification algorithms, etc.) take over. When modeling complex systems, the structure is hidden, it is not known beforehand. The objective is *automatic abstraction,* letting the structures automatically emerge. And the statistical tools naturally carry out abstraction: Individual observations are not assumed to be significant, only phenomena that remain consistent over the long-term observation periods.

To use statistical methods in a plausible way, the observations need to have statistical relevance. To reach this, the observations need to be *stationary,* that is, there need to exist some consistent statistical structure in the data. To make this possible, to be able to collect stationary data from a complex process, there has to be *balance,* at least as seen in the wider scale.

To find general ways of modeling, something has to be assumed. It turns out that such a rigid enough structural modeling framework where there is possibility of individual structures to emerge is that of *neocybernetics:* One assumes dynamic balance in the system where the internal interactions and feedbacks implement tensions that maintain the system integrity. One can forget the underlying interaction structures if they are just capable of providing appropriate stabilizing internal controls.

In the neocybernetic framework, one does not study all mathematically possible systems — only the physically reasonable ones that are in balance with their environment. Natural systems typically fulfill this assumption, and one would like the industrial systems to fulfill this assumption. What is more, good controls, however they are implemented, keep the system near its setpoint, regardless of the environmental disturbances: This means that linearity of the models can reasonably be assumed.

So, to apply multivariate methods, one has to concentrate on such (thermo)dynamic balances. The data needs to be selected so that is reflects this framework to make it possible to later determine appropriate models. As it turns out, the domain of *chemical systems* offers a compact framework for such studies.

11.1.3 Modeling chemical systems

Study a hypothetical example reaction, where there are α reactants on the left hand side, being denoted as A_i, $1 \leq i \leq \alpha$, and the β products on the right

hand side are B_j, $1 \leq j \leq \beta$:

$$a_1\mathrm{A}_1 + \cdots + a_\alpha\mathrm{A}_\alpha \quad \underset{k_\mathrm{A}}{\overset{k_\mathrm{B}}{\Leftrightarrow}} \quad b_1\mathrm{B}_1 + \cdots + b_\beta\mathrm{B}_\beta, \quad \Delta H. \tag{11.1}$$

Processes are typically reversible, so that the reaction can take place in both directions (k_B being the reaction speed in forward and k_A in backward direction). Symbol ΔH denotes the change in enthalpy, or inner energy, when the reaction takes place.

One needs a mathematically more compact representation for chemical reactions. How to "cybernetize" chemical reaction models applying the neocybernetic principles?

Information representation

The first problem is to represent such a chemical reaction formula in a practical numeric form. It seems that a practical way to code the reactions in a mathematically applicable form is to employ the vector formulation: Define a vector C containing all chemical concentrations so that all A_i and B_j are represented there among the elements. The "chemical state" can assumedly be captured in this vector, and individual reactions determine equations in that chemical space: If the coefficients $-a_i$ and b_j from (11.1) corresponding to the chemicals are collected in the vector G, one can express the total concentration changes in the system as

$$\Delta C = G\,\zeta. \tag{11.2}$$

Here, ζ is a scalar that reveals "how much" (and in which direction) that reaction has proceeded. When there are many simultaneous reactions taking place, there are various vectors G_i; the weighted sum of reaction vectors $\zeta_i G_i$ reveals the total changes in chemical contents, the weighting factors being collected in the vector ζ.

Using the above framework, metabolic systems can in principle be modeled: If one knows the rates of reactions, or the scalars ζ_i, the changes in the chemical contents can be determined. This idea of *invariances* within a chemical system have been widely applied for metabolic modeling; the key term here is *flux balance analysis (FBA)* (for example, see [**?**]). However, the rates x are not known beforehand, and, what is more, the reactions are typically not exactly known.

In many ways, the model structure (11.2) is not yet what one is looking for. The main problem there is that the flux balances only capture the *stoichiometric,* more or less *formal balance* among chemicals. It does not capture the *dynamic balance,* whether or not the reactions actually take place or not. Luckily, there exist also other ways to represent the chemical realm.

Thermodynamic balance

There is a big difference between what is *possible* and what is *probable,* that is, even though something may happen in principle, it will not actually happen. To

understand the dynamic balance, the reaction mechanisms need to be studied closer.

Assume that it takes a_1 molecules of A_1, a_2 molecules of A_2, etc., according to (11.1), for one unit reaction to take place. This means that all these molecules have to be located sufficiently near to each other at some time instant for the forward reaction to take place. The probability for one molecule to be within the required range is proportional to the number of such molecules in a volume unit; this molecular density is revealed by consentration (when the unit is mole/liter; by definition one mole always contains $6.022 \cdot 10^{23}$ particles). Assuming that the locations of the molecules are independent of each other, the probability for several of them being found within the range is proportional to the product of their concentrations. On the other hand, the reverse reaction probability is proportional to the concentrations of the right-hand-side molecules. Collected together, the rate of change for the concentration of the chemical A_1, for example, can be expressed as a difference between the backward reaction and forward reaction rates:

$$\frac{d\,C_{\mathrm{A}_1}}{d\,t} = -k_{\mathrm{B}} C_{\mathrm{A}_1}^{a_1} \cdots C_{\mathrm{A}_\alpha}^{a_\alpha} + k_{\mathrm{A}} C_{\mathrm{B}_1}^{b_1} \cdots C_{\mathrm{B}_\beta}^{b_\beta}. \tag{11.3}$$

In equilibrium state there holds $\frac{d\,C_{\mathrm{A}_1}}{d\,t} = 0$, etc., and one can define the constant characterizing the thermodynamic equilibrium (for example, see [**?**]):

$$K = \frac{k_{\mathrm{B}}}{k_{\mathrm{A}}} = \frac{C_{\mathrm{B}_1}^{b_1} \cdots C_{\mathrm{B}_\beta}^{b_\beta}}{C_{\mathrm{A}_1}^{a_1} \cdots C_{\mathrm{A}_\alpha}^{a_\alpha}}. \tag{11.4}$$

Linearity objective

One of the neocybernetic objectives is that of linearity. Clearly, the expression (11.4) is far from being linear — indeed, it is purely multiplicative. It turns out that applying a purely syntactic trick, linearity of the structures can be reached: Taking logarithms on both sides there holds

$$\log K' = b_1 \log C_{\mathrm{B}_1} + \cdots + b_\beta \log C_{\mathrm{B}_\beta} - a_1 \log C_{\mathrm{A}_1} + \cdots - a_\alpha \log C_{\mathrm{A}_\alpha}. \tag{11.5}$$

To get rid of constants and logarithms, it is also possible to differentiate the expression:

$$0 = b_1 \frac{\Delta C_{\mathrm{B}_1}}{\bar{C}_{\mathrm{B}_1}} + \cdots + b_\beta \frac{\Delta C_{\mathrm{B}_\beta}}{\bar{C}_{\mathrm{B}_\beta}} - a_1 \frac{\Delta C_{\mathrm{A}_1}}{\bar{C}_{\mathrm{A}_1}} + \cdots - a_\alpha \frac{\Delta C_{\mathrm{A}_\alpha}}{\bar{C}_{\mathrm{A}_\alpha}}, \tag{11.6}$$

where the variables $\Delta C_i / \bar{C}_i$ are deviations from the nominal values, divided by those nominal values, meaning that it is *relative changes* that are of interest. The differentiated model is only locally applicable, valid in the vicinity of the nominal value.

Multivariate representation

A single reaction formula can also be expressed in a linear form when the variables are appropriately selected. However, to model complex systems consisting

of various reactions, the data representation needs to be extended: The differing data vectors containing different sets of variables (the reactions employing different chemicals) have to be embedded in the same vector space to make them compatible.

Assume that the vector v is a vector containing all relevant variables capturing the state of the environment and the system itself, including, for example, relative changes in all chemical concentrations. This means that the vector Γ_i representing a single reaction can contain various zeros, assuming that the corresponding chemicals are not contributing in the reaction i. If the vectors Γ_i are collected as columns in the matrix Γ, one can write the individual expressions in (11.6) in the matrix form where one row is allocated to each of the reactions:

$$0 = \Gamma^T \, v. \tag{11.7}$$

This expression needs to be compared to flux balance analysis: Now one only needs to study levels of concentrations, not changes in them. This is indeed essential in complex chemical systems, where the energy and matter flows cannot be exactly managed. The key point to observe here is that analysis of complicated reaction networks can be avoided: No matter what has caused the observed chemical levels, only the prevailing tensions in the system are of essence. The underlying assumption is that the system is robust and redundant: Individual pathways are of no special importance as there exist various alternative routes in the network.

It turns out that reactions can in principle be characterized applying linear algebra in the space of chemical concentrations, being compatible with the multivariate methods. However, the results still need to be interpreted appropriately. Nothing mathematically very special is being done — as there seldom is in the field of linear theory! — but when seen from the appropriate point of view, new conceptual tools for modeling of complex systems can be available.

11.2 From constraints to degrees of freedom

As shown above, the domain-area information can be captured in data. However, this representation feels somewhat hollow, and it is difficult to believe that domain-area *knowledge* could ever be captured this way. However, it can be claimed that *freedoms-oriented* way of modeling is just as natural as the *constraints-oriented* approach is. To understand the meaning of this claim, closer analyses are needed.

11.2.1 Constraint-based models

Traditional models are typically based on *constraints.* This means that system properties are captured by formulas of the general form

$$0 = f(v), \tag{11.8}$$

where f is some scalar or vector-valued function of the variable vector v. For example, the chemical model in (11.7) is a special (linear) case consisting of

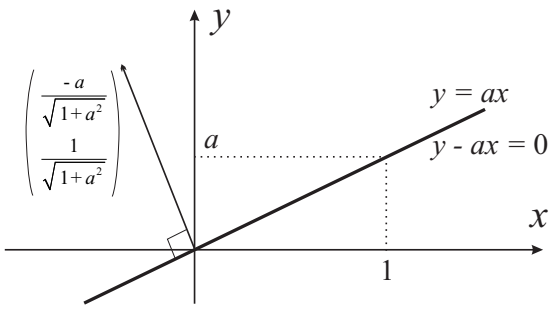

Figure 11.1: How the parameter vector represents the normal vector

various independent equations or constraints in a matrix form. What is more, the linear multivariate models that have been studied in previous chapters, or the models of the form $y = F^T x$, can be written as $0 = F^T x - y$, so that when one defines

$$\Gamma = \begin{pmatrix} F \\ \hline -I \end{pmatrix}, \qquad \text{and} \qquad v = \begin{pmatrix} x \\ \hline y \end{pmatrix}, \tag{11.9}$$

this is again of the form (11.7), and simultaneously a special case of (**??**). Note that such models are not unique — the vectors Γ_i can be freely scaled without affecting the validity of the equations. So, to make such a presentation less ambiguous, from now on assume that the vectors in Γ are normalized to unit length, so that $\Gamma_i^T \Gamma_i = 1$.

To better understand the structure of models that are presented in such constraints-oriented form, study a single-output case, so that y_i is scalar, and Γ_i is a vector. Whereas $y_i = F_i^T x$ defines a *one-dimensional null-space* in the high-dimensional variable space of v, and because the inner product $\Gamma_i^T v$ between the data and the vector Γ_i is zero, this vector defines a unit vector that is *orthogonal to this subspace.*

Further, to illustrate the above fact, for a moment study a case where the input data also is scalar, so that there holds $y = ax$ for some scalar a. This case is shown in Fig. 11.1: As the variable x varies, the variable y follows it following the linear dependency. When the x–y pairs are projected onto the normal vector, the projection length for variable pairs that fulfill the constraint is always zero. However, because of noise, this seldom exactly holds, and one has $e = \Gamma^T v$ for some non-vanishing e. Because of the orthonormal nature of Γ_i, the dot product $\Gamma_i^T v$ directly tells the distance between the data point v and the model. This gives an explicit solution to the error-in-variables problem presented in chapter 4: All variables have similar roles, all containing noise. Indeed, cleverly minimizing this model error gives yet another regression strategy, and this will be briefly studied in what follows.

11.2.2 Total Least Squares

One approach to implementing the EIV model (see Sec. 4.2.1) is the *Total Least Squares (TLS)* algorithm [11]. Following the idea presented above, search for such a regression hyperplane that when data points are orthogonally projected onto this plane, the (squared) distances reach minimum.

Here we continue with the single-output study for output y_i, so that

$$y_i = F_i^T x = F_{i,1}x_1 + \cdots + F_{i,n}x_n, \tag{11.10}$$

equalling

$$0 = \Gamma_i^T v \tag{11.11}$$

for

$$\Gamma_i = \begin{pmatrix} F_{i,1} \\ \vdots \\ F_{i_n} \\ -1 \end{pmatrix}, \qquad \text{and} \qquad v = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ y_i \end{pmatrix}. \tag{11.12}$$

The dimension of the "augmented" data space of v, and the length of the vector Γ_i, is $n+1$. As was observed above, Γ_i is orthogonal to the subspace that is "allowed" by the model. Further assuming that Γ_i is normalized, so that $\|\Gamma_i\| = 1$, the dot product $e = \Gamma_i^T v$ directly tells the shortest distance (positive or negative) from the point v to the regression hyperplane (for points lying exactly on the plane this measure, of course, giving 0, according to the model). the average of squared distances for a set of points $v(1)$ to $v(k)$ can be expressed as

$$\frac{1}{k} \cdot \sum_{\kappa=1}^{k} e^2(\kappa) = \frac{1}{k} \cdot \sum_{\kappa=1}^{k} \left(\Gamma_i^T v(\kappa) \cdot v^T(\kappa)\Gamma_i\right) = \frac{1}{k} \cdot \Gamma_i^T \cdot V^T V \cdot \Gamma_i, \tag{11.13}$$

where

$$\underset{k\times n+1}{V} = \left(\; X \; \middle| \; Y_i \; \right). \tag{11.14}$$

To minimize this with the requirement that the normal vector must be normalized,

$$\begin{array}{ll} \text{Minimize} & \frac{1}{k} \cdot \Gamma_i^T \cdot V^T V \cdot \Gamma_i \\ \text{when} & \Gamma_i^T \Gamma_i = 1, \end{array} \tag{11.15}$$

leads to the Lagrangian formulation (see page 20) where one has

$$\left\{ \begin{array}{lcll} f(\Gamma_i) & = & \frac{1}{k} \cdot \Gamma_i^T \cdot V^T V \cdot \Gamma_i, & \text{when} \\ g(\Gamma_i) & = & 1 - \Gamma_i^T \Gamma_i. & \end{array} \right. \tag{11.16}$$

The cost criterion becomes

$$J(\Gamma_i) = \frac{1}{k} \cdot \Gamma_i^T V^T V \Gamma_i + \lambda_i(1 - \Gamma_i^T \Gamma_i). \tag{11.17}$$

This results in

$$\frac{d}{d\,\Gamma_i} \left(\frac{1}{k} \cdot \Gamma_i^T V^T V \Gamma_i + \lambda_i(1 - \Gamma_i^T \Gamma_i) \right) = 0, \tag{11.18}$$

giving

$$\frac{1}{k} \cdot 2V^T V \cdot \Gamma_i - 2\lambda_i \cdot \Gamma_i = 0, \tag{11.19}$$

or

$$\frac{1}{k} \cdot V^T V \cdot \Gamma_i = \lambda_i \cdot \Gamma_i. \tag{11.20}$$

The distance minimization has become an *eigenvalue problem* with the searched normal vector Γ_i being an eigenvector of the data covariance matrix $R = \frac{1}{k} \cdot V^T V$. However, as compared to principal component analysis, the searched normal vector is given by the principal component corresponding to the *least significant* eigenvalue — zero eigenvalue meaning exact match with the assumed model structure: In such a case, there must exist an exact linear dependency between the variables, and this dependency can be extracted as the model. Remembering the definition of the vector Γ_i, the final regression formula solved as

$$y_i = \frac{\Gamma_{i,1}}{\Gamma_{i,n+1}} \cdot x_1 + \cdots + \frac{\Gamma_{i,n}}{\Gamma_{i,n+1}} \cdot x_n. \tag{11.21}$$

For a multivariate system, the same analysis can be repeated for all outputs y_i separately; note that the eigenproblem is generally different for all outputs. However, one needs to be careful: In the derivation y_i was interpreted as any of the other input variables, meaning that it is not the output that was explicitly being explained (as is the case with MLR). This means that the TLS model not necessarily gives a good regression model for estimating the output.

This TLS method can also be called "last principal component analysis", as compared to PCA, where the solution (to the problem of maximizing variance rather than minimizing variation) is given in terms of the *most significant* principal components. This is an indication of the need for new thinking, indeed, *inverse thinking:* Rather than concentrating on the null space, or the *constraints,* one concentrates on *freedoms,* what is left outside, where there still exists non-nullified information.

TLS is an example of experiments when trying to rehabilitate the old way of thinking. However, the problems of very high dimensions are not solved. If there is a high number of redundant variables, many of the eigenvalues are practically zero. Which of the minor eigenvectors to select, then? This selection becomes very sensitive: With another data with another noise realization the ordering can become very different — giving a completely different model. This means that the noise sensitivity of the TLS model is increased unreasonably. And, as observed before, it is this noise sensitivitity that is a crucial matter when constructing good regression models.

11.2.3 Emergent models

Mathematically speaking, if there are n separate variables, there are n degrees of freedom in the data space, but each (linear) constraint decreases the number

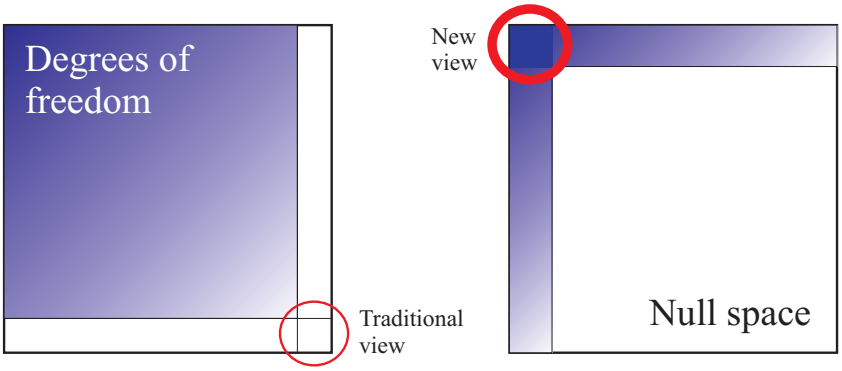

Figure 11.2: Schematic illustration of the covariance structure among data when there are few constraints (on the left), and when there are many constraints (on the right). The simplest presentation for the system properties changes as the number of constraints increases, or when the remaining degrees of freedom accordingly decrease

of degrees of freedom by one — specially, if there are ν linearly independent constraints, the number of remaining degrees of freedom is only $N = n - \nu$. Summarizing: The linear constraints constitute a null space within the data space: This means that in these directions there is no variability. The remaining N directions in the data space constitute a linear subspace where all variation among variables is concentrated.

What do these degrees of freedom mean in practice? Originally, if there were completely separate unconnected variables (subsystems), there would be the maximum number of freedoms. When subsystems become connected, when interactions between them are established, the variables become coupled, thus reducing the number of free variables. Further, when feedbacks are introduced, the remaining inputs and outputs of the subsystems can still be connected. It is specially typical in cybernetic systems where this scenario holds: Ability to recover after disturbances is a manifestation of tightly interconnected system. In such systems it is only a few degrees of freedom that remain more or less loosely controlled.

The key point here is that essentially the same dependencies among variables can be captured in terms of degrees of freedom as with constraints. At some point, when the number of constraints increases, the *most economical* representation changes: The simplest model with the least parameters is no more the constraints-oriented model but the freedoms-oriented model (whatever it will be). According to the *Ockham's razor,* one needs to switch to *emergent models* when the system is cybernetic enough. In Fig. 11.2, the covariance structure of the data space is schematically depicted: When the null space of constraints is dead and dull, all interesting behaviors are concentrated in the directions of remaining freedoms.

It is difficult to escape the traditional ways of thinking: Traditional methods for analysis (modeling) and design (synthesis) are always based on models that are based on constraints.

it is the multivariate statistical methods that directly attack the degrees of freedom, abstracting away the structural details, that help to escape the constraints. Even though this opposite view of modeling sounds unintuitive, it turns out that the freedoms-oriented models are *more intuitive* than the constraints-oriented models, being based on the explicit time-domain features, as visualized below.

11.2.4 Examples

To visualize the freedoms-oriented model structures, exploit dynamic intuitions: Assume that the available variables are successive measurements of some signal y, so that samples are indexed as $y(\kappa)$, $y(\kappa-1)$, etc. Originally, it is assumed that these samples are independent of each other — it is the task of the (dynamic) model to connect the variables together. Assuming that the constraint-oriented model is

$$y(\kappa) = ay(\kappa - 1), \tag{11.22}$$

there is a direct connection to Fig. 11.1. Constructing the augmented data space as

$$v(\kappa) = \begin{pmatrix} y(\kappa - 1) \\ y(\kappa) \end{pmatrix}, \tag{11.23}$$

the whole data space $\mathcal{S}$ is spanned by the constraint vector and the freedom vector together:

$$\mathcal{S} = \left(\begin{array}{c|c} \Gamma & \theta \end{array} \right) = \left(\begin{array}{c|c} \frac{a}{\sqrt{1+a^2}} & \frac{1}{\sqrt{1+a^2}} \\ \frac{-1}{\sqrt{1+a^2}} & \frac{a}{\sqrt{1+a^2}} \end{array} \right). \tag{11.24}$$

The freedom-oriented way of describing the model is also

$$\theta = \begin{pmatrix} a \\ -1 \end{pmatrix} / \sqrt{1 + a^2}. \tag{11.25}$$

It is difficult to see here anything that would outperform the original model. However, now assume that there are three variables that are connected together by a model:

$$\begin{cases} y(\kappa) = ay(\kappa - 1) \\ y(\kappa + 1) = ay(\kappa). \end{cases} \tag{11.26}$$

This exactly corresponds to the model (11.22) where there are redundant variables. The key point here is that one does not know beforehand whether some of the variables are redundant — when modeling complex systems, this is typically the case. The data vectors are now

$$v(\kappa) = \begin{pmatrix} y(\kappa - 1) \\ y(\kappa) \\ y(\kappa + 1) \end{pmatrix}. \tag{11.27}$$

In this case, the constraint vectors without normalization are

$$\Gamma = \begin{pmatrix} a & 0 \\ -1 & a \\ 0 & -1 \end{pmatrix}. \tag{11.28}$$

The constraints span a two-dimensional subspace in the three-dimensional variable space – the remaining degree of freedom can be solved by orthogonalization, for example applying the *Gramm-Schmidt procedure.* To start with, one can take any linearly independent vector:

$$\begin{array}{c} \left(\begin{array}{cc|c} a & 0 & 1 \\ -1 & a & 0 \\ 0 & -1 & 0 \end{array}\right) \rightarrow \left(\begin{array}{cc|c} a & \frac{a^2}{1+a^2} & \frac{1}{1+a^2} \\ -1 & \frac{a^3}{1+a^2} & \frac{a}{1+a^2} \\ 0 & -1 & 0 \end{array}\right) \\ \rightarrow \left(\begin{array}{cc|c} a & \frac{a^2}{1+a^2} & \frac{1}{1+a^2+a^4} \\ -1 & \frac{a^3}{1+a^2} & \frac{a}{1+a^2+a^4} \\ 0 & -1 & \frac{a^2}{1+a^2+a^4} \end{array}\right). \end{array} \tag{11.29}$$

This means that the model becomes

$$\theta = \begin{pmatrix} 1 \\ a \\ a^2 \end{pmatrix} \Big/ \sqrt{1+a^2+a^4}. \tag{11.30}$$

The "axis of freedom" clearly has an *exponential outlook* in the data space. This is in exact correspondence with the actual time-domain behavior of a system that is characterized by a model of the form (11.22). Indeed, the degrees of freedom determine "behavioral fragments", so that the actual observations can be constructed as combinations of them. The patterns can be scaled arbitrarily to optimize the match — these scaling factors are the latent variables in z.

When working on simple cases, the approach is not crucial. But when new variables are introduced, each of them typically comes with an accompanying constraint, and it is only the degrees of freedom that truly reflect the essential dependency structures in the system. When modeling complex systems, it is assumed that the number of variables should not be limited artificially: Each of the new variables *can* contain some fresh information — the "accdompanying constraint" does not necessarily reduce the degrees of freedom in the augmented space exactly by one. Whereas the constraints-oriented modeling approach becomes a unmanageable mess, the freedoms-oriented models become clearer and clearer as the data dimension increases. The higher the number of variables is, the more appropriate is the pattern-based representation seems to become.

How about the interpretations when there is a higher number of remaining degrees of freedom? Study the model

$$y(\kappa) = a_1 y(\kappa-1) + a_2 y(\kappa-2), \tag{11.31}$$

or

$$0 = a_0 y(\kappa) - a_1 y(\kappa-1) - a_2 y(\kappa-2). \tag{11.32}$$

Now there is one constraint in the three-dimensional space, and two remaining degrees of freedom:

$$\Gamma = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \quad \text{and} \quad v(\kappa) = \begin{pmatrix} y(\kappa) \\ y(\kappa-1) \\ y(\kappa-2) \end{pmatrix}. \tag{11.33}$$

The degrees of freedom for such a dynamic system have always the same interpretation: Typically, if there is considerable inertia in the system, the most significant principal component stands for a filter for finding the average momentary value of y, as being revealed by the latent variable $z_1(\kappa)$, and the second principal component stands for the trend prototype: The latent variable $z_2(\kappa)$ reveals the rate of change in the signal (see exercises). In this sense, there again exist very natural interpretations for the model structures.

In this kind of rather simple cases, there is a trade-off between approaches. The constraints-based model is stronger when it comes to analysis of dynamic phenomena (as the roots of the coefficient polynomial reveal the dynamic modes beyond the signal), whereas for the freedoms-oriented model such time-domain analyses need to be separately carried out (as presented in chapter 9), meaning that a heavier machinery needs to be employed.

The freedoms-oriented model is based on features taht constitute patterns that together explain the observations in the data space, assuming that there are some dependencies and redundancies in the behaviors. Determination of the system state becomes a *pattern recognition task*. Specially, when in the case of chemical systems, it is "chemical pattern matching" that is being carried out — and this is carried out automatically by the underlying thermodynamic processes.

11.3 Case studies

To illustrate the above approaches, two practical application examples are presented, where the "chemical semantics" is appropriate. Both of these complex processes are being currently studied at HUT Control Engineering Laboratory.

11.3.1 Characterizing the state in practical processes

To apply the ideas, the theoretical derivations still need to be extended towards practice. The data vector v needs to be further studied to make it possible to capture all *internal tensions* in complex chemical systems. As it turns out, the following extensions can, for example, be implemented without ruining the linear structure among the variables:

- **Temperature.** According to the Arrhenius formula, the reaction coefficients are functions of the temperature, reactions becoming faster as the temperature rises, so that $k \propto \exp(c/T)$. This means that when this is substituted in the formulas, and when logarithms and differentiations are carried out, the model remains linear if the new variable is defined as $v_T = \Delta T/\bar{T}^2$.

- **Acidity.** The pH value of a solution is defined in terms of a nonlinear formula: $\mathrm{pH} = -\lg C_{\mathrm{H}^+}$. Because it is essentially logarithm taken of a concentration variable, one can directly include the changes in the pH value among the variables, $v_{\mathrm{pH}} = \Delta\mathrm{pH}$.

- **Voltage.** In electrochemical reactions, one should characterize the the "concentration of electrons". However, it turns out that acording to the Butler-Volmer theory [**?**], the amount of free electrons is exponentially proportional to the voltage. This means that, after taking the logarithms, the "electron pressure" can be characterized by the variable $v_{\mathrm{e}^-} = \Delta U$.

- **Dissipation.** It has been assumed that the systems being studied are in thermodynamic balance. This homeostasis can be extended, however: The steady state can be determined not only in terms of the variables, but also in terms of their derivatives. This means that one can study *dissipative systems,* where the rate of change remains constant, a constant flow of chemical flowing into or out from the system. Looking at the formula (11.3), it is clear that model linearity is not lost if one has variables like $v_{\dot{C}} = \Delta\dot{C}/\bar{\dot{C}}$.

- **Mass flows.** The concentration-oriented variables can be transformed into masses (molarities) when multiplied by volumes, meaning that after taking logarithms, the structure is linear. Similarly, the volumetric dissipation rates change into mass flows; further, surface phenomena (coating, etc.) are related to the surface area, so that if the volumes or areas change, one can include variables of the form $v_A = \Delta A/\bar{A}$ and $v_V = \Delta V/\bar{V}$.

- **Physical phenomena.** It is evident that structures that are originally linear, like phenomena that represent diffusion between compartments, etc., can directly be integrated in the model, assuming that appropriate variables (deviations from the nominal state) are included among the variables. What is more, smooth nonlinearities become affine when they are locally linearized, and, further, they become linear when developed around the nominal state.

In strong liquids one cannot always apply concentrations, but one has to employ *activities* instead, or actual activation probabilities. If it is assumed that these activities are some power functions of the concentration so that $\mathcal{A} = a_1 C^{a_2}$, after taking logarithms the model still remains linear in terms of the original concentrations. This means that — even though linearity is not compromized — the variables may become multiplied by some unknown factors, so that there is some scaling effect.

The vector v selected here is the measurement vector, containing *all* possible quantities that can affect the system behavior — internal system variables and external environmental variables alike. This data presentation can capture the chemical domain semantics, and in different environments the models have different interpretations.

11.3.2 Case 1: Modeling an industrial nickel plating process[1]

In printed wiring boards, one needs a layer of nickel as an oxidation barrier between the copper electric circuitry and gold finishing (see Fig. 11.3). This nickel-phosphor layer can be created, for example, using electrochemical processes. The properties of the nickel layer can be affected by changing its phosphor content. It is clear that one should be capable of monitoring and controlling the layer thickness, and also its phosphor content so that the set values would be reached.

The chemical reactions taking place in the plating process are very complex, and not completely known. Four contradictory sets of reactions have been proposed to characterize the process, but none of them seems to satisfactorily explain observed behaviors. Not only is the exact process structure unknown — not all chemicals are either known, as the compositions of the commercial reagents are business secrets. However, the processes are slow, and it is evident that the appropriately operated coating process remains well in balance. All these observations are well in line with the assumptions beyond the freedoms-oriented modeling.

The process state can be characterized in terms of its acidity or pH (controlled using ammonia to be between 4.7 and 5.0), temperature (to be around 80 degrees centigrade), nickel concentration (controlled by adding nickel sulphate), and electrical potentials. The dynamics is also affected by the loading, or the total area to be plated simultaneously in the bath. In addition to these, additional chemicals are present, some of them are known, like the *reducers* (sodium hypophosphite), and some are not (different kinds of activators and inhibitors); the contribution of the residues of reaction chemicals is also estimated: The variable MTO (or "metal turn-over") descibes the aging of the process liquids, being supposedly proportional to the concentrations of the unspecified chemicals. All these state variables can be recorded or calculated in a practically continuous manner.

It is the properties of the final nickel surface that cannot be measured on-line: The layer thickness should be around 4 μm, and it should contain some 7 – 10 weight percent phosphor. Information of these is available only after laboratory analyses, once or twice a day, and a model is needed to estimate these quantities in a reliable way. To implement such soft sensors, the multivariate regression models were constructed.

As it is typically the case, the model (or data preprocessing) needs to be tailored to match the problem domain. The state variables were mean-centered and normalized in the traditional way — but, in addition to these variables, new ones could also be employed. This nicely illustrates the benefits of the simple linear model structure.

Because the relative changes in the momentary layer growth rate assumedly are linear functions of changes in the other state variables, the overall relative change is reached when one integrates the momentary rate over the bath time. And because of the linearity of this mapping model F, the integration can be

[1]The simulations were carried out by Mr. Hans-Christian Pfisterer

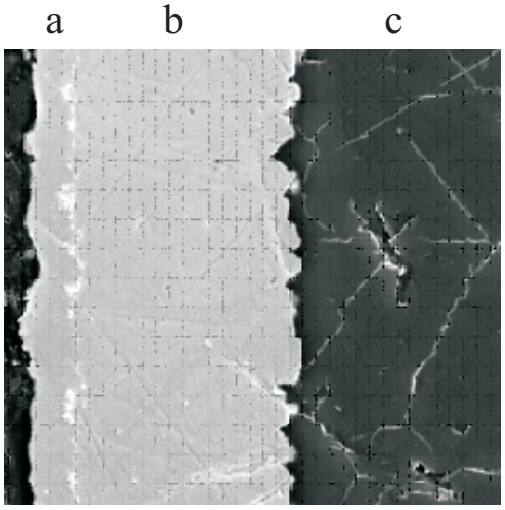

Figure 11.3: Cross-section of a test plate: **a** - the Ni-P layer (about 5 μm); **b** - copper layer; **c** - base (epoxy laminate)

moved "through" the model:

$$\Delta l(t) = \int_{t_0}^{t} \Delta \dot{l}(\tau)/\bar{\dot{l}} d\tau = \int_{t_0}^{t} F^T v(\tau)\, d\tau = F^T \int_{t_0}^{t} v(\tau)\, d\tau. \tag{11.34}$$

This means that if one includes the integrals of relative changes among the x variables, a linear model should be capable of capturing the layer changes around the nominal cumulation rates. These nominal absolute values need to be separately modeled, or if the bath time of the board is also included among the input variables, it is the same model that suffices.

It is always difficult to evaluate the performance of the models in an unbiased way — however, in this case we are lucky: There is an explicit model derived specially for this process, starting from physico-chemical first principles, the free parameters being optimally tuned to match the observations. It can be assumed that this model is the best model one can construct for the process, as that modeling effort gained the the Best Diploma Thesis Prize of 2004 in Finland (as granted by TEK, the Finnish Association of Graduate Engineers). The results are shown in Figs. 11.4 and 11.5: Even though not all phenomena can be estimated by the model of four PCA-based latent variables (see Fig. 11.5), it seems that the same problems are faced by all models regardless of their construction. The data-oriented model where no process-specific knowledge is exploited is well comparable with the expert-tuned physical model that is based on a set of highly nonlinear differential equations: The validation errors for fresh data have the same orders of magnitude (results for two set of validation data shown in the figures).

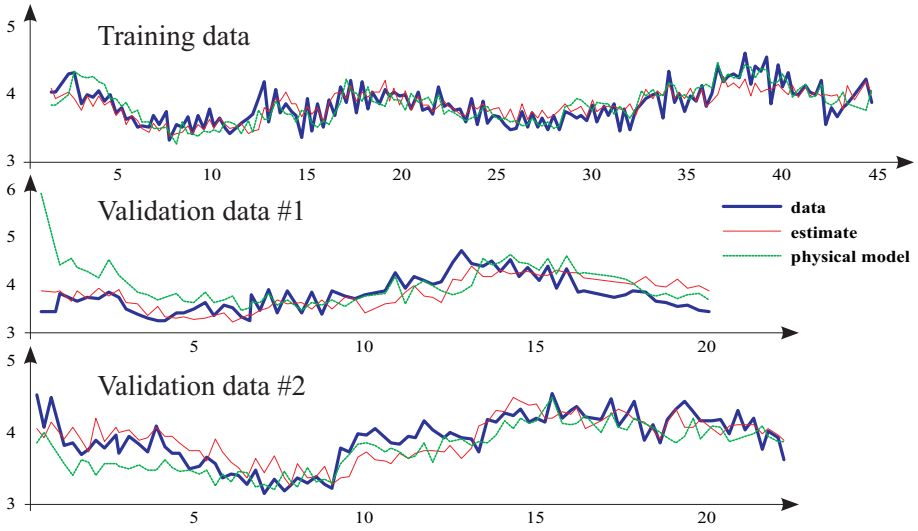

Figure 11.4: Estimates for nickel layer thickness

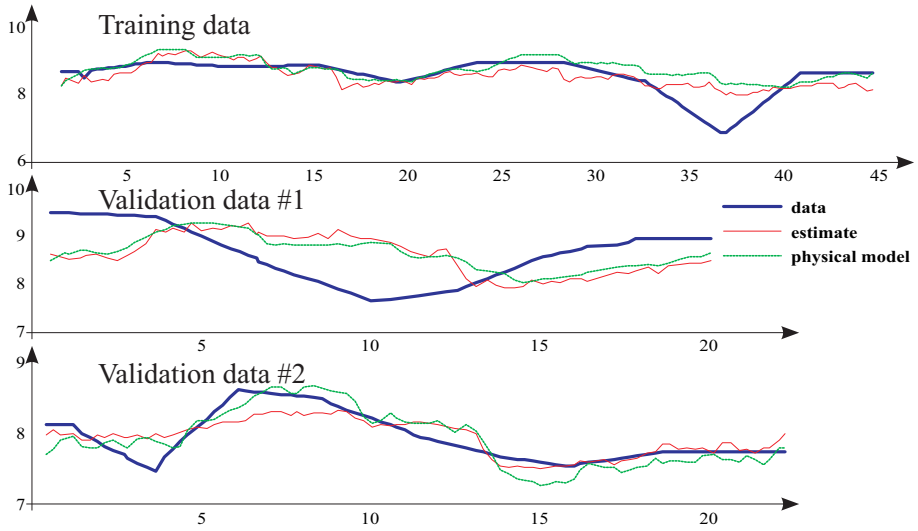

Figure 11.5: Estimates for phosphor content

11.3.3 Case 2: Modeling genetic networks and metabolic systems[2]

The previous example was a man-made system based on more or less designed chemical reactions, the reaction mechanisms being predetermined to explicitly implement intended behaviors, and a (more or less accurate) first-principles model could also be constructed. Now, study a natural system that is still much more complex, so that finding the explicit reaction mechanisms is even more complicated — perhaps the same principles of freedoms-based modeling apply?

When studying metabolic reactions, it is complex chains of reactions based on organic chemistry that should be mastered. What is more, these reactions are dictated by the genetic processes, where enzymes are produced. On the other hand, the chemical state affects the gene activities — this means that there are interacting genetic and metabolic networks that should be mastered. The closed control loops cannot be distinguished from each other, and the only realistic approach is to assume "pancausality", where the interactions and feedbacks constitute the tensions keeping the system in balance. As studied in chapter 2, genetic networks can be modeled applying the same model structures as the chemical processes — the metabolic processes are fast, whereas the genetic ones are slow (see Fig 11.6). Both of the levels can be combined in one model structure, making it perhaps possible to reach *systemic biology*. In the figure, the linear pattern recognition processes are expressed in terms of dynamic state-space models.

In the project SyMbolic (Systemic Models for Metabolic Dynamics and Gene Expression), funded by TEKES during 2004 – 2006, new kinds of models were derived for representing the cellular dynamics, and one of the approaches was the exploitation of the idea of emergent models [**?**].

There is plenty of data: The modern ChIP techniques, etc., provide huge amounts of measurements, as all gene activities can be simultaneously measured (for example, see [**?**]). Indeed, measuring gene activities (in terms of active messenger-RNA) is more straightforward than measuring the metabolites. Even though there is plenty of data, it is not optimally conditioned for dynamic identification purposes: The dimension of data (in thousands) is higher than what is the number of samples (in hundreds), and the excitation sequences are not persistently exciting (being step experiments). What is more, the data is very noisy — partly because of the uncertainties in the measurement process, and partly because measurements carried out in different laboratories seem not to be quite compatible. This means that the statistical multivariate methods, and specially the latent variable approaches, are well motivated also from the pragmatic point of view.

Implicitly, the latent variable methods assume that there is redundancy among genetic and cellular functionalities — and, indeed, it has been shown that there are typically groups of genes rather than individual genes that are responsible for the functionalities. And also on the metabolic level: Processes in the cytoplasm are well buffered, and typically there are negligible responses if one only considers a single input and a single output. The multivariate methods make it

[2]The simulations were carried out by Mr. Olli Haavisto, M.Sc.

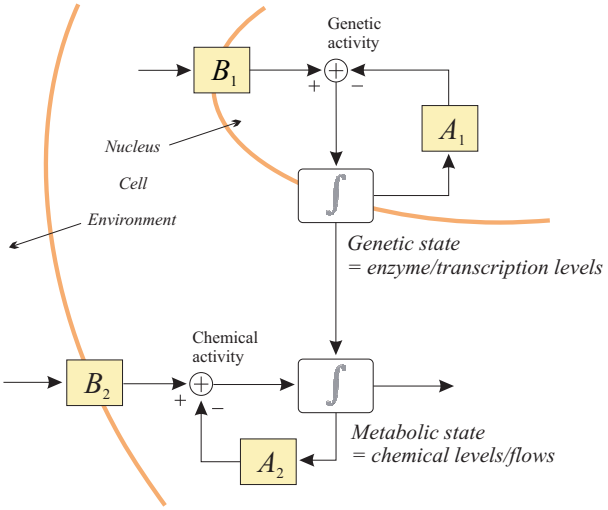

Figure 11.6: Two time scales in the cellular system

possible to study the whole grid of proteomic/metabolomic phenomena simultaneously — this means that one does not need to employ excessive excitation signals, or huge dosages, resulting in considerable disturbances in the cell behavior, or even death. The gentle approaches are necessary when one wants to study living cells rather than pathological, more or less irrelevant cases.

As an application example, modeling of data from yeast cell cultivations were used (see [**?**]). There were a few dozen experiments, where different kinds of step changes in the environment were executed, and the resulting gene activity transients were recorded. Modeling this data was quite a challenge, as there was not enough data. Even though the applied model structure was robust, no conclusive conclusions can be drawn.

As was observed above, metabolite concentrations and gene activities could be represented in the linear model structure, variaqbles being collected in a single vector. However, now the model was restructured so that dynamics was captured: The environmental variables (substrate properties, temperature, etc.) were collected in the input vector u, and the gene expression levels were collected in the output vector y. Mean-centering and normalization of data was carried out. The dimensions of the vectors were such that n_u was about ten, and m was about 4000; the number of latent variables N was selected as 4, and stochastic-deterministic subspace identification was applied.

The assumption beyond the adopted modeling approach is that balances are more characteristic to cellular systems than the transients are. And, indeed, it seems that the steady states are nicely modeled, whereas the transient behaviors are not reproduced by the model (see Fig. 11.7). Still, it seems that the extreme compression of the variable space does not ruin the steady-state correspondence. There seem to exist only few degrees of freedom left in the behavioral data.

It can be claimed that the degrees of freedom in a cellular system characterize *metabolic behaviors* or *functions.* When the environment changes, the new balance is found along these axes in the chemical space whan “chemical pattern matching” is carried out. For example, assuming that available glucose

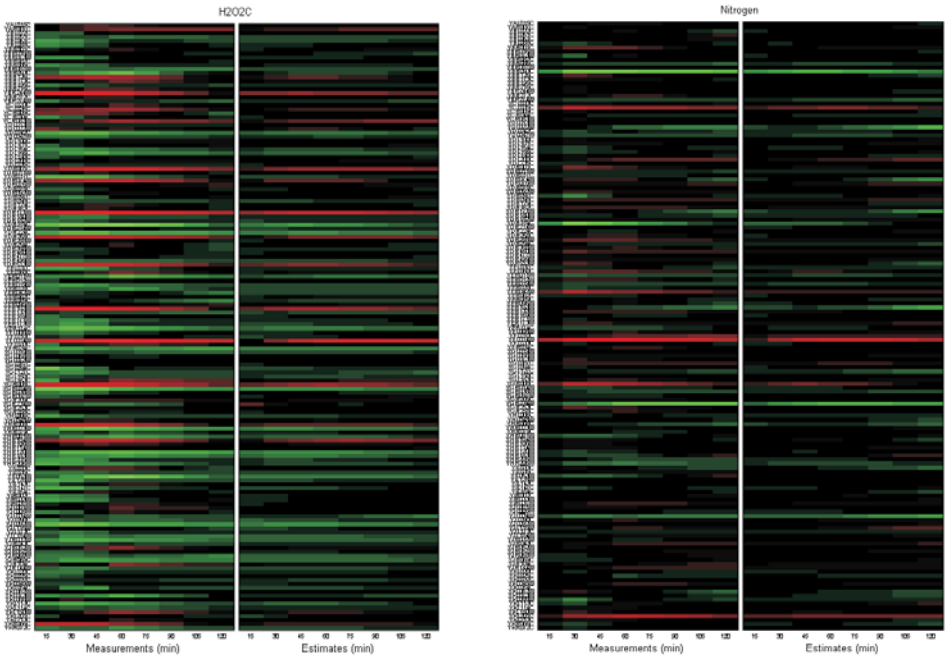

Figure 11.7: Two open-loop experiments with the model, showing 256 "stress genes" (red color meaning activity increase, green meaning activity decrease). In the leftmost figures, hydrogen peroxide step is being simulated for two hours, and in the rightmost ones, nitrogen step is simulated. In both cases, the actual behaviors in the genetic state are shown on the left, and the estimates given by the four-state model are shown on the right. Despite the transients, there is a good correspondence between the observations and the very low-dimensional model (see [**?**])

goes up, it is also mannose production that goes up, or some other processes that exploit glucose. There is only balance pursuit taking place: But after "anthropocentric", finalistically-loaded interpretations are employed, when some chemicals are interpreted as nutrients, some others as metabolic products, and the rest as waste, one reaches "emergent interpretations". When complexity cumulates, the balance reactions start looking goal-oriented, pre-planned, and "clever". Scarcity of some chemicals changes the balance appropriately, trying to compensate for the shortage.

11.4 Towards "artificial cells"

New conceptual tools become available as further interpretations are employed. In complex chemical systems, there seem to exist reserve mechanisms for compensating for the disturbances. This kind of buffering is characteristic not only to metabolic systems, but it seems to apply also in more general terms: *Le Chatelier principle* states that changes in environment are compensated by changes in the balance, so that the system tries to "escape" the changes. In

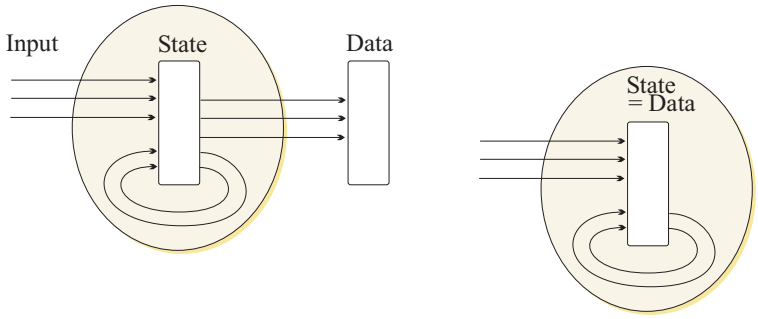

Figure 11.8: From *data modeling* (on the left) towards *system modeling* (on the right). The variables being measured are system variables: Because of pancausality, changing them also changes the system state

[**?**], the idea of "elastic systems" is proposed to characterize the reactions of cybernetic systems in general.

When the variables are selected appropriately, so that system semantics is captured, and if the pancausality assumption holds, the constructed modes are not only *data models* — they are *system models.* They can capture the fundamental essence of systems. They can be used not only for monitoring, but also for design and control construction: Changing variables appropriately also changes the resulting balance (see Fig. 11.8). The remaining degrees of freedom in the system reveal the possibilities of further controls to make the system still more balanced; in this sense, *process data mining* becomes possible, where information can be gathered directly from the behaviors, not from model-based assumptions. New kinds of models make it possible to implement new kinds of controls — higher-level controls. However, new challenges are faced: When new feedbacks are introduced, the set of freedoms changes. Control design becomes an iterative task, and new kinds of design tools are needed.

The ideas of biological cybernetic systems can be extended to technical (bio)processes: The still unbounded degrees of freedom can be regulated, new feedbacks can be constructed. Still better balanced "superorganisms" are constructed. The industrial systems are becoming like *artificial cells* themselves: Industrial plants also have *metabolism,* raw materials being exhausted and others being produced. Originally, the production can be far from optimum, but as soon as dependencies among variables are recognized, they can be used for constructing new feedback structures to implement more efficient and robust — better balanced — production. In both cases, in natural and man-made cells alike, it turns out that the goal of "evolution" is overall efficiency of production, no matter whether it is humans that are acting as agents for development or not. This can be reached by implementing mechanisms for reaching best possible production conditions; and this system integrity needs to be maintained without collapses. To maintain such balance, the system has to respond appropriately to the spectrum of disturbances coming from the environment.

It seems that the new approaches offer new possibilities for attacking the mysteries of evolutionary processes from a fresh point of view — such visions are studied closer in [**?**]).

Computer exercises

1. Assume that data of an oscillating system is collected and its time-series is analyzed, that is, dynamics is being captured in data, and study the covariance structure:

   ```
   y = sin([1:100]/2)';
   V = [y(1:98),y(2:99),y(3:100)];
   theta = regrPCA(V)
   ```

   Interpret the distribution of the eigenvalues. Also interpret the first and second eigenvector as patterns characterizing the signal.

2. Applying the same data, study the eigenvector with the vanishing eigenvalue (carrying out the Total Least Squares regression analysis):

   ```
   G = regrPCA(V,-1)              % Also "regrTLS" available
   abs(roots(G))
   ```

   Interpret the result. What happens with the above analyses (freedoms vs. constraints) if the data is extended so that

   ```
   V = [y(1:97),y(2:98),y(3:99),y(4:100)];
   theta = regrPCA(V)
   ```

   Try to interpret the eigenvectors and eigenvalues now. What can you say about the extensibility and robustness of the two approaches?