Level 2

Emergent Models of *Cellular Functions*

In the previous chapter, the data format was determined so that cell-specific information (or any network-originated information) could compactly be captured. The next task is to find the higher-level presentations, or the actual model structures, so that the underlying data can efficiently be exploited, and the essence of the cellular behaviors truly becomes manifested.

The key issue in this chapter are the *models,* or how to construct them in an appropriate way. It has to be recognized that *models are always false,* only showing a narrow projection of the complexity in real life systems. But good models can give intuitions.

Very simple mathematics only is employed here, and the model structures will be linear. There is nothing new in the mathematics — it is the interpretations that play the central role. Appropriate interpretations make it possible to escape from the reductionistic level to explicitly holistic models. These "emergent models" become practical when the components-oriented modeling view is exhausted. The new model structures can be seen as revealing the functions that take place in the complex system.

2.1 About "system semantics"

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 how they should be represented. Indeed, what there is, what one can know about them — these problem fields are called *ontology* and *epistemology,* respectively (these issues are studied again in chapters 7 and 10). 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.

For the purposes of concrete modeling, the notion of semantics has to be formalized in some way: This very abstract concept is given here very concrete contents, compromising 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 and compressing it. 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.

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, whereas finding representations for the low-level domain-area features is domain-area specific.

The former task — coding the domain-area information in concrete data structures — was studied in the previous chapter, whereas the latter task — connecting the atoms of information into relevant structures — is studied in this chapter.

The higher-level semantics determines how the information atoms are connected. In our numbers-based environments, a practical and robust approach towards capturing such *contextual semantics* is offered by correlations-based measures. If the data is defined appropriately so that it captures the dynamical balances in the system, the simple contextual dependency structures can also be seen to capture *cybernetic semantics* of the domain (see chapter 7). Assuming that information is conveyed in visible co-variations among data, structuring of lower-level data can be implemented by the mathematical machinery without need of outside expert guidance. Despite the trivial-sounding starting point, non-trivial results can be found when the mathematical structures cumulate. This makes it possible to reach "smart" models that adapt in unknown environments.

2.2 Constraints vs. degrees of freedom

The mathematical machinery has been traditionally used for solving engineering-like, reductionistic problems. However, the focus is changing: One should be capable of abstracting away the details and seeing the "big picture". In such cases one simply cannot go in the traditional bottom up direction — one has to go top-down, explicitly starting from the system level. And one cannot assume

there is some existing *a priori* model structure — the models have to be based on observations. There are many challenges when new ways of thinking are exploited.

2.2.1 System models and identification

It is assumed that in a system the data are somehow bound together, and it is this bond that captures the essence of the system. The model structure derived in the previous chapter was of the following form, explicitly characterizing the bond between variables

$$0 = \Gamma^T z, \tag{2.1}$$

this matrix expression consisting of n separate scalar equations determining connections among variables in z. Indeed, this formulation is the very traditional approach to presenting structures within systems. For example, assuming that the matrix Γ consisting of a single column, and the data vectors $z(k)$, for k indexing the discrete time axis, are defined as

$$\Gamma = \begin{pmatrix} -1 \\ a_1 \\ \vdots \\ a_d \\ b_0 \\ b_1 \\ \vdots \\ b_d \end{pmatrix}, \quad \text{and} \quad z(k) = \begin{pmatrix} y(k) \\ y(k-1) \\ \vdots \\ y(k-d) \\ u(k) \\ u(k-1) \\ \vdots \\ u(k-d) \end{pmatrix}, \tag{2.2}$$

the connection among variables can be rewritten also in the form

$$y(k) = \sum_{i=1}^{d} a_i y(k-i) + \sum_{j=0}^{d} b_j u(k-j). \tag{2.3}$$

As it turns out, this is the traditional way of representing dynamics of a d'th order SISO (single input, single output) system. A huge body of theory has been developed, for example, for identifying the system parameters a_i and b_j based on a set of observations of the variables $y(\kappa)$ and $u(\kappa)$ for $k_0 \leq \kappa \leq k$ (for example, see [2]).

The models of the form (2.1) assume that the linear combination of the variables should be exactly zero — however, as the measurement values always are inaccurate, this does not exactly hold, and one has to extend the original model:

$$e = \Gamma^T z. \tag{2.4}$$

Here, e is the model error vector — the goal of identification of parameters in Γ is transformed into an optimization problem, where one tries to minimize the overall error variance. Very much effort has been put on enhancing the

numerical properties of the identification algorithms, typically starting directly from the formulation (2.3), and for making them more reliable and robust — after all, the determination of the parameters is typically based on least-squares matching, and there are various reasons for problems [42].

First, a special challenge in traditional identification is caused by the nonideal noise properties. Different variables can be corrupted by the noise in different ways. And, in the case of colored noise, the uncorrelatedness assumption of the noise samples becomes compromised, and the parameter estimates become biased.

Second, if trying to capture all available information — by employing all available variables — in the models, as was proposed in the previous chapter, determination of the parameters sooner or later becomes an ill-defined task. As the large number of variables are more or less redundant, they are no more strictly linearly independent of each other, and the numerical properties of the algorithms can become very poor. For example, the variables $y(k-i)$ in (2.3) are in principle separate variables, but because of the smoothness in the signal behaviors, the variables are certainly not independent. The data covariance matrix (matrix that needs to be inverted in least-squares fitting) becomes badly conditioned.

In today's applications, these problems with high dimensionality severely plague the traditional modeling approaches. It is not only the high number of input data that causes problems, but the whole model structure is challenged. Applying the traditional model structure it is easy the implement SISO models, but one should also be capable of tackling with more complex systems consisting of various submodels — as in the metabolic system there exist various simultaneous balance reactions taking place at the same time. In principle, the data representation in (2.1) is naturally a MIMO structure, being a framework for presenting various simultaneous equations just as well. This structure only needs to be efficiently utilized. Are there alternatives to traditional ways of describing (locally linear) models?

2.2.2 Emergent models

The structure of the model (2.1) needs to be elaborated on: This can be accomplished as the model is interpreted in terms of linear algebra. Mathematically speaking, if there are μ separate variables, there are μ degrees of freedom in the data space, but each (linear) constraint decreases the number of degrees of freedom by one — specially, if there are ν linearly independent constraints, the number of remaining degrees of freedom is only $\mu - \nu$. The linear constraints constitute a *null space* within the data space: This means that in these directions there is no variability. The remaining $\mu - \nu$ 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

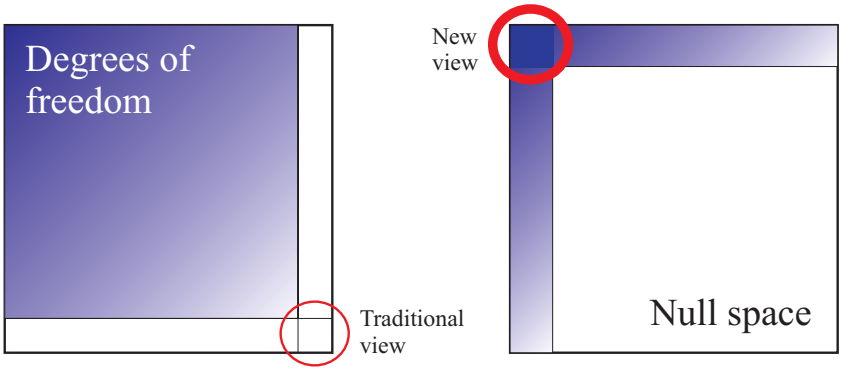

Figure 2.1: Schematic illustration of the covariance structure among data when there are few constraints (on the left), and when there are many constraints (on the right). The simplest presentation for the system properties changes as the number of constraints increases, or when the remaining degrees of freedom accordingly decrease

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. In the metabolic system there are dozens of individual underlying reactions controlling the cellular metabolics, the chemical levels being balanced accordingly.

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 not 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. 2.1, the covariance structure of the data space is depicted: When the null space of constraints is dead and dull, all interesting behaviors are concentrated in the directions of remaining freedoms. It can be assumed that relevant phenomena in the cell are revealed by the "metabolic degrees of freedom"; it turns out that when applying very compact and behaviors-oriented models, the system starts looking more or less "clever" — indeed, speaking in such terms has to be postponed to next section.

Whereas the visible constraints are emergent patterns resulting from underlying dynamic attractors, the degrees of freedom make it possible to model this process of emergence and the structure of such patterns.

How is this dichotomy between constraints and freedoms manifested in concrete terms? For example, study an infinite-dimensional distributed parameter system that is governed by partial differential equations — a very natural way to characterize natural complex systems. As these PDE's are spatially discretized, there is a large number of ordinary differential equations connecting the local variables. Remember that only together with the boundary conditions the PDE's can uniquely determine system behaviors, thus giving rise to a very complicated system of hybrid equations that can seldom be solved explicitly. The

constraints are now there in the form of dynamic and algebraic equations — the PDE's and the boundary conditions, respectively. The emergent behavior is typically manifested in terms of a low number of possible *modes.* For example, in the case of a vibrating plate, typically there only exist a few vibration modes; these "modes of freedom" can easier be modeled than the original constraints. Such freedoms-oriented approach is also quite natural, as then one directly concentrates on the time domain solutions of the equations that are immediately measurable in system behaviors.

Laws of nature are traditionally written in terms of constraints: The visible dependencies among observed phenomena are recorded. But, again, these surface patterns just emerge from underlying, more fundamental interactions. Perhaps one should rather start thinking in terms of "freedoms of nature".

It is difficult to escape the traditional ways of thinking: Traditional methods for analysis (modeling) and design (synthesis) are based on models that are based on constraints. And, indeed, constraints are the very basis of Wittgensteinian thinking: Languages are the means of structuring the world in terms of connections between concepts. This also holds what comes to formal languages like programming formalisms that the contemporary software tools are based on, and also traditional mathematics is based on finding ingenious proofs, or paths from a fact to another. Traditional mathematics exercises make nice pastime activity as the solutions typically are unique and hard to discover; however, the heavy mathematical machinery that is based on relevance is more general.

It is just as it is with detective stories: they make nice reading, but they are not plausible. Sherlock Holmes once said that "When you eliminate the impossible, whatever remains, however improbable, must be the truth". But in real life there are no clear-cut truths — modern detectives construct the "big picture" out from the mosaic of more or less contradictory evidence: The plausible explanation maximally fits the observations. This is today's world — as there is no unambiguous truth, it is *relevance* that is preferred; closer studies are needed here.

2.2.3 Towards inverse thinking

One needs to find appropriate mathematical formulations for the above intuitions. The leap is mainly conceptual — one has to go to the other end of the continuum, from structure orientation to data orientation. It is data originating from freedom structures that is more relevant than parameters originating from constraint structures. As it turns out, this approach makes it possible to avoid the age-old problem concerning symbolic and numeric representations: The structures are not fixed beforehand — or, actually, they are ignored altogether.

First, study the structure-oriented end of the continuum. For simplicity, assume that one wants to capture the nominal state when observations are available. Variations around the nominal state are interpreted — in the traditional spirit — as noise that should be eliminated from the model.

Assuming that there are many sources of noise, one can abstract away the properties of individual noise sources. According to the Central Limit Theorem,

one can assume that the net effect of all noise sources is such that the error distribution is Gaussian, that is the observations are distributed along a high-dimensional bell-shaped curve around the mean value; it is this mean value vector ζ that is being searched for. For the data distribution one can write the density function

$$p(e) = \frac{1}{\sqrt{(2\pi)^{\dim\{z\}}\,|\mathrm{E}\{zz^T\}|}}\,\mathrm{e}^{-\frac{1}{2}\,(z-\zeta)^T\mathrm{E}\{zz^T\}^{-1}(z-\zeta)}. \tag{2.5}$$

In the spirit of maximum likelihood identification, one selects the best estimate for ζ by maximizing the overall probability of the measurements

$$\hat{\zeta} = \arg\max_{\zeta}\left\{\mathrm{E}\left\{p(e)\right\}\right\} \tag{2.6}$$

by adjusting the center of the distribution appropriately. Because logarithm is a monotonous function, maximization of (2.6) equals minimization of

$$-\ln p(e) = c + \frac{1}{2}\cdot(z-\zeta)^T\,\mathrm{E}\{zz^T\}^{-1}\,(z-\zeta)\,. \tag{2.7}$$

When looking at this goodness criterion, it is evident that the "natural" scaling of variables is reached if the measurements are preprocessed as

$$z' = \mathrm{E}\{zz^T\}^{-1/2}\,z. \tag{2.8}$$

In the space of these new variables z', it is simply the Euclidean distances (or their squares) that reflect the differences between vectors. This scaling explicitly emphasizes the null space directions where there exists no variation in the data space, thus boosting the constraints-oriented thinking. For the original data, the weighting matrix when evaluating distances is $W = \mathrm{E}\{zz^T\}^{-1}$; it is revealing to note that for Gaussian data this expression is called *Fisher information matrix.* Information is assumed to be in the inverses of covariances.

This is the today's realm. The problem with the scaling (2.8) here is that if the dimension of z is excessive, the scaling matrix becomes badly conditioned: If there are linearly dependent variables, the inverse matrix cannot be found. In cybernetic systems the variables typically are highly redundant due to the high number of underlying constraints.

To proceed, one needs to look at (2.1): Even though the roles of Γ and z are intuitively clear, this can be incorrect intuition. Mathematically, if Γ is a vector, the roles of these two vectors are identical. There is duality among structure and variables: The visible manifestations of structure are numbers in vectors, just as the data is. It can be assumed that the information delivered by observations is distributed among the structure part and the data part. Normally, it is assumed that observations represent data — however, in this case when the constraints dominate, it can be assumed that *observations represent structure.* The situation needs to be turned upside-down: The information that is normally used for modeling is now regarded as noise, and only the "leftovers" not exploited by the traditional modeling approaches are concentrated on.

This kind of problems of traditional thinking can be concretized: For example, inverse covariance weighting results in excessive emphasis on linearly dependent

variables, the identification procedures trying to distinguish between identical variables — what comes to representing the real properties of the data, such emphasis is counterproductive. Another example: When identification is carried out in the parameter space rather than in the data space, iterative adaptation steps trying to pull the parameters towards better locations, pathological effects can take place, specially, if the parameterization represents a dynamic model. The reason for this is that dynamic behaviors are related to the poles and zeros of the parameter polynomials rather than to the parameters themselves; convex combinations of parameter vectors do not necessarily reflect the properties of those vectors at all.

Now the model is constructed to capture the properties of the data directly, not the properties of some man-made parameterization.

What this intuition means in practice, what are the consequences? Traditionally, when searching for the structure, it is thought that variation outside the assumed structure is noise — now it is assumed that this remaining variation is interesting, reflecting those behaviors that have not been paralyzed by the constraints. Somewhat intuitively, one could employ the idea of *symmetry pursuit,* defining the data-oriented portion of the measurements as the inverse of the weighting in (2.8):

$$z'' = \mathrm{E}\{zz^T\}^{1/2}\, z. \tag{2.9}$$

This can be expressed also in another way: The symmetric weighting matrix among measurements becomes (see next section)

$$W = \mathrm{E}\{zz^T\}, \tag{2.10}$$

rather than being $\mathrm{E}\{zz^T\}^{-1}$, as in the (2.8) case. This means that directions of variation in the data are explicitly emphasized. What is nice is that no matrix inversions are needed, and such operations remain well-behaving even for high-dimensional data.

The motivation for the data weighting was here rather intuitive — however, in the next chapter this issue will be concentrated on from another perspective. It can be claimed that *such weighting mathematically corresponds to the view of data that locally controlled systems actually see* in their environments.

2.3 Technical exploitation

For the rest of this chapter, assume that the presented view of data were appropriate, and study the conceptual tools that are in place when this view is being functionalized. The approach to modeling here is synthetic rather than analytic: The approach is "technical", not trying to capture the actual underlying processes but only trying to imitate the results. The key point here is to present the best possible tools — multivariate statistical mathematics — and in the next chapter it is shown that there truly can exist some connection to real life.

2.3.1 Subspaces and mappings

It is beneficial to see the more general setting, or what the presented framework looks like when see from the point of view of mathematics and mathematical tools. When functionalizing the freedoms-based model structure, one faces a pattern matching problem where linear algebra is needed.

Data preprocessing

Forget about all connotations that the variables in z may have, and apply conditioning to this data so that the technical assumptions become optimally fulfilled.

The first assumption is that of model linearity. Typically, problems are caused by the fact that data from linearized models are *affine,* that is, additive constants are needed in formulas. To get rid of the affine terms, the data can be transformed to follow a strictly linear model, for example, applying mean-centering — this is the standard approach when doing strictly data-based modeling where the nominal values of the variables are not known.

However, these problems are only faced when doing constraints-oriented modeling: When concentrating on the freedoms, no mean-centering is necessary.

The second assumption is quadratic nature of cost criteria. The reason for this is that easily manipulated and explicit formulas can be reached. The quadratic criteria mean that variations in the data are emphasized, and to reach reasonable models, appropriate scaling of data needs to be carried out. Assuming that all variables are equally informative, different variables can be equally "visible" by normalizing them to have unit variance, because units are arbitrary. This means that one uses either correlation matrices (if data is mean-centered) or cosine matrices (if data is not centered) as association matrices (see [92]).

Whatever are the data preprocessing steps, the original data z will hereafter be denoted ζ.

The data scaling is very crucial, affecting the results very much – the normalization should be motivated better. Indeed, as shown in Sec. 3.3, if the data is coming from a truly cybernetic system, it turns out that normalization is the *natural* way of seeing inter-system signals.

Pattern matching

In concrete terms, the freedoms-based model characterizes the location of an observation in the data space in terms of the degrees of freedom. The degrees of freedom are manifested as n linear *feature vectors* φ_i being collected in the matrix φ. Because of linearity, features can be freely scaled and added together. The observed patterns, combinations of the variables, are assumed to be weighed sums of such features, so that one can write

$$\hat{\zeta} = \sum_{i=1}^{n} \xi_i \varphi_i = \varphi \xi, \tag{2.11}$$

where ξ is the vector of weighting factors. If vectors φ_i are seen as coordinate axes, ξ_i are the coordinate values. Use of the feature model becomes an associative pattern matching process against data.

Assuming that $n < m$, arbitrary variable combinations ζ cannot be exactly represented by the features, and when searching for the best possible match, or estimate $\hat{\zeta}$, one is facing an optimization problem. When the representation error $\zeta - \varphi\xi$, weighted appropriately, is minimized, one can write the quadratic criterion

$$J(\xi, \zeta) = \frac{1}{2} (\zeta - \varphi\xi)^T W (\zeta - \varphi\xi) . \tag{2.12}$$

The unique minimum is found when the gradient vector is set to zero:

$$\frac{d J(\xi, \zeta)}{d\xi} = \varphi^T W \varphi\xi - \varphi^T W \zeta = 0, \tag{2.13}$$

giving the unique solution

$$\xi = \left(\varphi^T W \varphi\right)^{-1} \varphi^T W \zeta. \tag{2.14}$$

In practice, this implements a mapping from an m dimensional space of ζ onto the n dimensional subspace of ξ spanned by the feature axes. Variables in ξ are called *latent* or *hidden variables.* Because of the data compression, exact match is not found, and one can only hope that the ignored variation is *noise,* not actual *information.*

How to distinguish between noise and information, then? Formally, there is no difference in the manifestation of variations in the data, and one has to apply *ontological assumptions* concerning the nature of relevant properties in the data. First, following the above discussions, one should select the weighting matrix as

$$W = \mathrm{E}\{\zeta\zeta^T\}. \tag{2.15}$$

Selection of the feature vectors so that they would represent the most important degrees of freedom can also be explicitly solved, and the solution is given by PCA presented in Sec. 2.3.2. This means that one should choose φ so that the *subspace of the n most significant principal components of data* is spanned by columns φ. Indeed, it is not necessary that the features are exactly the covariance matrix eigenvalues, $\varphi_i = \theta_i$, but it suffices that there holds

$$\varphi = \theta D \tag{2.16}$$

for some orthogonal transformation matrix D, so that $D^T = D^{-1}$. Correspondingly, the latent variables are modified as $\xi' = D^{-1}\xi$. The optimal selection of features is also non-unique — regardless of how the (non-singular) basis is constructed out from the matrix θ, the same variation can be captured. This means that after PCA, different kinds of *factor analysis* techniques, rotations, etc., can be applied to find a physically better motivated basis. For example, if the variables ζ have some constant bias, so that they are not zero-mean, it is possible to determine variables ξ that also have non-zero-mean — they can even always remain positive. When the variables represent some physical quantities, such non-negative coding is more plausible.

Regression based on latent variables

When the latent variables ξ are available, they can be exploited, for example, for *regression,* mapping data from the latent basis ξ onto some output space of y, so that $y = f^T\xi$. If the mapping is implemented through the low-dimensional latent basis rather than directly from the variables ζ, noise gets filtered out, and more robust estimates for the output can be found.

Similarly as above, the criterion for a good mapping model is minimization of quadratic criterion. When written for a single output variable y_j at a time, the mapping error becomes $\epsilon_j = y_j - f_j^T\xi$, and a reasonable criterion is found when the variance of this error, or $\mathrm{E}\{\epsilon_j^2\}$, is minimized. However, in some badly conditioned cases a generalization is in place: Robust regression models are found when the *regularized* criterion is applied where the parameter sizes are also emphasized:

$$J_j(f_j) = \mathrm{E}\{\left(y_j - f_j^T\xi\right)\left(y_j - f_j^T\xi\right)^T\} + \frac{1}{q}f_j^T f_j. \tag{2.17}$$

When the gradient is set to zero,

$$\frac{d\,J_j(f_j)}{df_j} = 2\left(\mathrm{E}\{\xi\xi^T\} + \frac{1}{q}I_n\right)f_j - 2\mathrm{E}\{y_j\xi^T\}^T = 0, \tag{2.18}$$

one can find the unique solution:

$$f_j = \left(\mathrm{E}\{\xi\xi^T\} + \frac{1}{q}I_n\right)^{-1}\mathrm{E}\{y_j\xi^T\}^T. \tag{2.19}$$

When this procedure is carried out for all outputs y_j separately, one can see that essentially the same formula is found in each case, and one can write a combined expression for all individual mappings as

$$f = \left(\mathrm{E}\{\xi\xi^T\} + \frac{1}{q}I_n\right)^{-1}\mathrm{E}\{y\xi^T\}^T. \tag{2.20}$$

If there is no need for regularization, that is, if the covariance $\mathrm{E}\{\xi\xi^T\}$ is invertible, one can use the standard formulation

$$f = \mathrm{E}\{\xi\xi^T\}^{-1}\mathrm{E}\{y\xi^T\}^T. \tag{2.21}$$

A special regression case is where the output is chosen to be the original data, $y = \zeta$, so that *reconstruction* of the data is being carried out, noise hopefully being filtered out during the compression process:

$$\hat{\zeta} = \mathrm{E}\{\zeta\xi^T\}\mathrm{E}\{\xi\xi^T\}^{-1}\,\xi. \tag{2.22}$$

Now there are technical tools for implementing mappings from data onto the feature subspace and back. The remaining problem is the determination of that feature subspace.

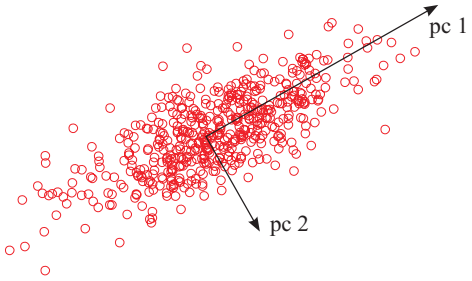

Figure 2.2: Principal component analysis reveals the variation structure in data

2.3.2 Multivariate tools

The distinction between constraints and freedoms can be elaborated yet in another way. Remember that traditionally one wants to minimize the sum of squared errors over the set of measurement data:

$$\Gamma = \arg\min_{\Gamma} \left\{ \mathrm{E}\{e^T e\} \right\}, \quad \text{when} \quad |\Gamma_i| = 1 \quad \text{for all} \quad i. \tag{2.23}$$

In this vector formulation, to have a well-conditioned optimization task, one has to fix the model vector size to avoid trivial solutions $\Gamma_i = 0$ (this is reached by introducing the additional restriction $|\Gamma_i| = 1$). This constrained optimization problem results in search for constraints in the traditional sense — indeed, the solution here is the method called *Total Least Squares*. When searching for the freedoms instead, the objective is exactly opposite:

$$\varphi = \arg\max_{\varphi} \left\{ \mathrm{E}\{\xi^T \xi\} \right\}, \quad \text{when} \quad |\varphi_i| = 1 \quad \text{for all} \quad i. \tag{2.24}$$

Note that even though it is freedoms that are searched, the mathematical machinery again is based on constrained optimization — constraints simply are the kernel of today's models! Here, vectors φ and ξ have been used to emphasize their different roles as compared to Γ and e: Defining $\xi = \varphi^T \zeta$, it is now the "error" ξ that is to be maximized, and φ is the axis along which this maximum variation in data is reached. If the vectors Γ_i and φ_i are interpreted as directions in the data space, mathematically speaking they reveal maximum orthogonality and maximum parallelity among these vectors and data, respectively. Applying the objective (2.24), it is assumed that variation in data is interpreted as information, whereas traditionally variation is seen as noise. And, specially, it is *covariation* among variables that carries information: Covariations can reveal the underlying "common causes" that are reflected in the measurements.

The solution to the problem (2.24) is given by *principal component analysis* or *PCA* (for example, see [6]). Without going into details (for example, see [42]), the basic results can be summarized as follows.

> The degrees of freedom can be analyzed using the data covariance matrix $\mathrm{E}\{\zeta\zeta^T\}$. The variability is distributed in the data space along the eigenvector directions of this matrix, variance in the eigenvector direction θ_i being given by the eigenvalue λ_i:
>
> $$\mathrm{E}\{\zeta\zeta^T\}\theta_i = \lambda_i \theta_i. \tag{2.25}$$

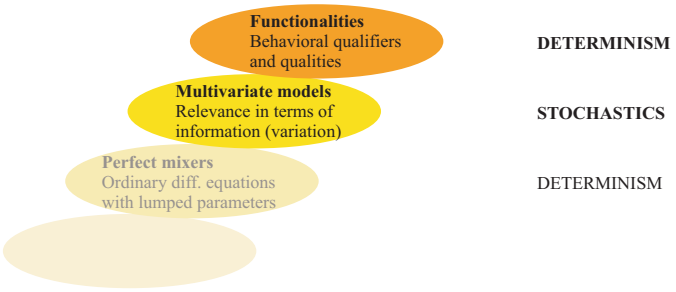

Figure 2.3: The remaining levels in the hierarchy of models in Fig. 3

Principal component analysis gives a structured view of freedoms in the data: The axes θ_i corresponding to most significant eigenvalues span a subspace where most of the variation in data is found (see Fig. 2.2). If $n < m$, meaning that the high-dimensional data is projected onto a lower-dimensional principal subspace, data compression takes place where the data variation is maximally preserved: If an n dimensional PCA basis is exploited, the model captures $\sum_{i=1}^{n} \lambda_i$ of the total variation in data — assuming data normalization, this total variation is $\sum_{j=1}^{m} \lambda_j = m$.

It turns out (because of symmetricity of the matrix $\mathrm{E}\{\zeta\zeta^T\}$) that the eigenvectors are orthogonal (indeed, orthonormal), so that the principal component directions can be used as a well-conditioned subspace basis vectors in a mathematically efficient way.

Principal components offer a very powerful mathematical framework — but is it physically meaningful? Complexity intuition says that self-organization of structures necessitates some kind of nonlinearity and instability: To reach emergence of differences, one needs positive Lyapunov exponents in functions, and to stabilize such divergent processes, nonlinearity is needed. However, as analysis of PCA reveals, there exist structures in data that can be motivated also in linear terms and using stable dynamic characterizations. Indeed, as will turn out later in chapter 3, the PCA intuitions will be of crucial importance when studying the properties of cybernetic systems. This means that the emergent patterns are very different as compared to the traditional chaotic images; the PCA patterns are based on global rather than local properties of functions.

2.3.3 New levels in emergence hierarchies

In Fig. 3, it was shown how deterministic and stochastic approaches can be seen to alternate in the hierarchy of emergent levels. Now the multivariate statistical models determine yet another stochastic level above the highest deterministic one: Information from the lower levels is extracted in the form of variations, and among that data, statistical dependency structures are determined in terms of covariations (see Fig. 2.3). Because such covariation structures can be found applying convergent algorithms, one is escaping the (mental) deadlock: Structures *can* emerge even in balance systems, one does not always need chaos and positive feedback to shake the underlying structures apart.

But a further transition from this stochastic level to a yet higher deterministic

level is more or less straightforward. If there are statistical structures that can be employed to compress the statistical data, such abstracted phenomena can be named, thus introducing new distinct concepts. As the statistical structures represent dynamical balances, the essence of such concepts is that they are *attractors* in the data space dictated by the properties of the environment. The domain-oriented "concepts" are manifested as *emergent functionalities*. These functionalities are non-programmed, they are not explicitly designed; they do not reflect the intentions of humans but the properties of the interplay between the system and the environment.

Using such higher-level concepts, the functioning of a complex system can be appropriately structured: there are names for behaviors that are assumedly relevant, being manifested in observations. A "natural language" based on such concepts would be beneficial when trying to characterize and understand the functioning of a complex system. Based on the low-level semantics, interpretations of the emergent concepts are self-explanatory.

However, complex systems differ from each other, and the "axes of relevance" cannot always be defined in such a straightforward way. For example, in industrial plants it is the *quality measures* that are the most important quantities when characterizing the plant operation. The industrial plants do not simply reflect their environment; they are constructed for some special purpose, and the qualities cannot be dictated only by the environment, but the intentions of the system designers have to be taken into account. Generally in technical systems the operation (and "evolution"; see chapter 3) is goal-directed — rather than reflecting the environment directly, the emergent structures should reflect the coupling between the input space (environment) and the output space (qualities). Rather than employing PCA, the model structures should implement the cross-correlations among the two spaces. For engineering-like development of the processes, or "artificial evolution", there are other regression techniques available, for example PLS and CCR (see [92]).

Natural systems are simpler than technical — assumedly they just want to survive, trying to match with their environments (see next chapter), so that one can employ the PCA-based models directly. One can assume that it can only be the visible variables that determine the observable behaviors; if the variables are selected and scaled appropriately, there is no reason why a mathematical machinery could not capture the same phenomena that are followed by the biological machinery. A more detailed example of emergent-level modeling is given below. The emergent functionalities reflect match with environment; as seen from above, such behavior seems *clever* in its environment.

2.4 Towards system biology

Finally, study how the presented approaches can be exploited for modeling cellular systems in practice — and how they perhaps could be exploited.

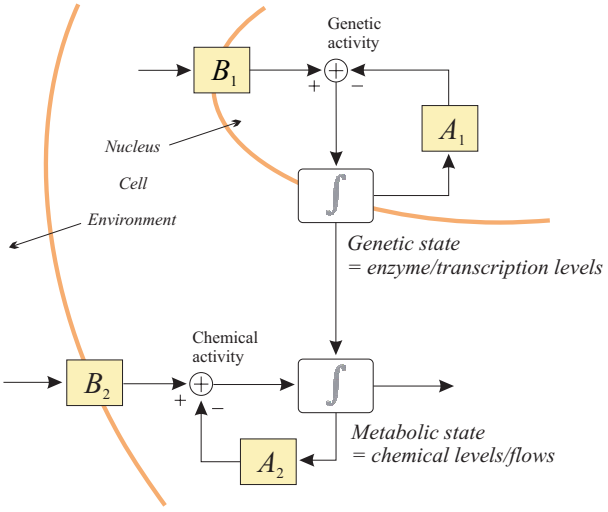

Figure 2.4: Two time scales in the cellular system

2.4.1 Facing real systems

Traditionally, modeling of biological (cellular) systems has been carried out in the rather traditional spirit: The goal has been to determine the constraints, individual dependency structures, exploiting more or less straightforward, SISO-type model structures. Data collection of cellular metabolics is typically carried out applying one-variable-at-a-time experiments. Similarly at the genetic level: Single "knock-out genes" can be explicitly deactivated to study their functions, resulting in non-natural behaviors. One reason for these simplified approaches are the practical problems, as the cellular state is difficult to measure, but new solutions are being introduced (for example, the microarray techniques for measuring the whole array of genetic activities simultaneously). Indeed, today there exists plenty of data, but this data is not necessarily well-conditioned or of good quality: The level of measurement noise is high.

Reaching reliable measurements is challenging, because the responses vary in different circumstances when the environment changes. What is more, because of buffering effects in the cells, huge dosages of reagents are needed (single input) to have noticeable responses (single output). On the other hand, these effects cannot be focused, being reflected to the whole set of variables. The experiments do not really characterize typical behaviors — the cell may become crippled altogether. Another traditional problem in metabolic systems is that they seem to be highly redundant (this also applies to gene expression). It seems that there typically is not just a single reaction mechanism explaining the processes, making it difficult to uniquely identify causal structures and model parameters. What makes this still more difficult is that not all chemicals can be recorded, and not all reactions are even known.

All of this suggests that multivariate statistical methods are needed. When applying the multivariate methods, buffering is just a manifestation of the internal feedbacks, and observations of the new balance deliver valuable information concerning the underlying metabolic processes and functions. No one-input/one-output studies are needed. Also the problems with unclear causal dependencies

are avoided because of the pancausality assumption: First, the actual reactions are not searched for, but the "residual" variations; second, PCA is just the right tool to model redundant and noisy phenomena, because it transforms from the visible variables to new latent variables, where noise and redundancies among variables has been ripped off. Putting it freely: "If they are there, but if they cannot be seen, just ignore them". All relevant variables and dependencies cannot be detected, but they can be ignored as long as they do their job in maintaining the system balance.

As studied in chapter 3, the PCA-based model structures are motivated not only from the data analysis point of view. It can be claimed that in evolution it is the principal subspace that is naturally being pursued by surviving systems that are capable of most efficiently exploiting the environmental resources.

The objective here is to study living cell rather than pathological cases. Balances are more characteristic than transients, and it is steady states that are modeled. Because metabolic processes are well buffered, remaining near the nominal state, linear models are locally applicable. Rather than carrying out tests in a SISO manner, the whole grid of chemicals are studied simultaneously. This applies also to the transcription factors on the genetic level: 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 2.4). In the figure, the linear pattern recognition processes are expressed in terms of dynamic state-space models, implementing two overlapping processes levels of "generalized diffusion".

Both of the levels can be combined in one model structure, and all information can be included in the data vector. The modeling procedure goes as follows: The sets of metabolites, transcription factors, and relevant environmental conditions (temperature, pH, ...) are defined, and experiments are carried out in different conditions, collecting data during the transients and in steady state. The degrees of freedom are found, determining the metabolic and genetic functions. Data orientation is necessary, multivariate tools are needed as the signal details are abstracted away, whereas emergent long-term phenomena become visible. Stationarity and validity of statistical measures is assumed — however, this assumption does not strictly hold. When the system becomes more and more complex, and as the number of constraints increases, the situation becomes blurred: some of the constraints are more acute than the others, and the thermodynamic balances are not necessarily all reached instantly.

2.4.2 Case example: Modeling genetic networks and metabolic systems[1]

In the project SyMbolic (Systemic Models for Metabolic Dynamics and Gene Expression), funded by National Technology Agency of Finland (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.

As an application example, modeling of data from yeast cell cultivations were

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

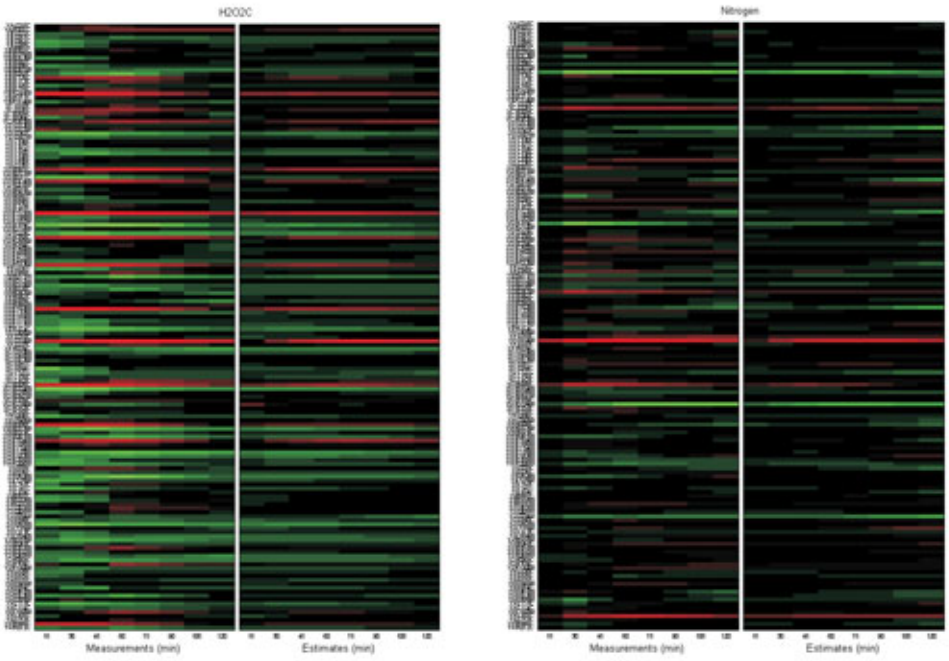

Figure 2.5: Two open-loop experiments with the model, showing 256 "stress genes" (red color meaning activity increase, green meaning activity decrease). Horizontal axis is time, and the rows represent individual genes. In the leftmost figures, hydrogen peroxide step is being simulated for two hours, 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 [35])

used (see [35]). There were a few dozen experiments available (from [15] and [32]), where different kinds of step changes in the environment had been executed, and the resulting gene activity transients had been recorded. The step experiments were interpreted to present "stress responses" of the yeast cells. Modeling this data was quite a challenge, as there was not enough data, and not all data was quite reliable. Indeed, there do not exist many reports of dynamic modeling of the cellular behaviors (one attempt that is also based on latent variables can be found, for example, in [39]).

Because the available data was in the form of step experiments, the model was restructured so that the experimental setting was captured: The causal structure from manipulated variables to observations was simulated in the model. 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. Rather than constructing a traditional static PCA model between these data sets, a dynamic model was constructed applying so called *stochastic-deterministic subspace identification* (see [60]). This means that also the time sequence among data is taken into account and exploited when the

latent variables x are constructed, the subspace identification algorithm automatically constructing a discrete-time state-space model (see [4]) for "generalized diffusion". Such a model can be efficiently exploited for implementing, for example, *Kalman filters* for optimally estimating the system state (see [24]). It can also be claimed that there is a connection to *Hidden Markov Models* here: The state sequences are reconstructed optimally, even though the probability interpretations are violated (this interpretation becomes more appropriate when the state variables are kept strictly non-negative; see chapter 6).

The dimensions of the vectors were selected so that m was about ten, and there were about 4000 output variables; the number of latent variables n was chosen to be 4.

When there are explicit transients in the data, the underlying assumptions about system stationarity are violated. This gives raise to model errors: There are slower and faster reactions taking place, some reaching their balance faster than the others. Indeed, a "Pandora's box" is opened when the balance assumption is abandoned — "extra" behaviors become visible in stress (transient) situations. What is more, complex transient reactions can take place in parts, where subprocesses follow each other; each of such intermediate products spans a new dimension in the variable space, and each chemical reaction introduces a new constraint, compensating for the increased dimensionality only after the balance is reached. The net effect is that the *invisible* dimensions in the variable space become visible during changes.

The assumption beyond the adopted modeling approach is, however, that balances are more characteristic to cellular systems than the transients are. And, indeed, it seems that at least the steady states are nicely modeled, whereas the transient behaviors are not reproduced as well by the model (see Fig. 2.5). Still, it seems that the extreme compression of the variable space does not ruin the steady-state correspondence. Truly, there seem to exist only few degrees of freedom left in the behavioral data.

2.4.3 "Artificial cells"?

When the presented model structure is seen in a perspective, it seems to open up new horizons. Using some imagination, it is easy to draw interesting interpretations.

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 when "chemical pattern matching" is carried out. For example, assuming that available glucose goes up, it is also mannose production that goes up, or some other processes that exploit glucose. In fact, 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, and abundance results in the opposite outcome, as being visible in the "activity vector" ξ (or x).

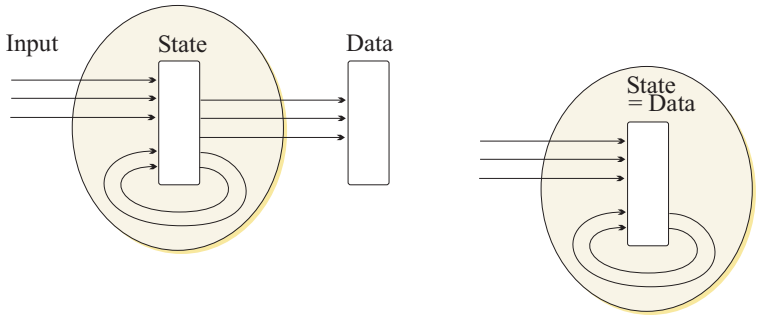

Figure 2.6: From *data modeling* (on the left) towards *system modeling* (on the right). The variables being measured are system variables (because of pancausality, changing them also changes the system state), and model structures being exploited are those of the system itself (both being based on the principal components of the measurement data; see next chapter)

A system model can be applied also for design and control. 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 and system-specific variables. 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. 2.6). 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* or real *knowledge 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, so that still better balanced higher-level "superorganisms" are constructed. On the other hand, the "broken" control loops can be fixed in the same way: For example, if the glucose level varies in the body more than it should, this can be compensated by insulin injections — along these lines, diabetes is treated manually today; but a simple automatic control loop could be implemented also as a step towards better lives of the "cybernetized patients".

Today, there are problems when trying to implement such integrated systems. For example, the glucose sensors need regeneration after a short time; after this problem is solved, new ones are sure to emerge. The key challenge is not how a single functionality — like sensitivity to a certain chemical — could be implemented, but how to keep the new system in a sustainable balance with its environment. This goal sounds very cybernetic. Indeed, it is the whole engineering-like thinking that has to be abandoned: Whereas one today concentrates only on a single functionality, it is the whole entity that has to survive

in the complex environment. The same challenges are faced in all applications of tomorrows medicine: If the new integrated systems are not in balance, the body rejects the transplants. Finding such balanced solutions is a holistic problem that cannot be solved reductionistically. One needs to change the whole way of thinking from invasive to humble: One has to admit that nature's own structures offer the most useful adapted solutions to the key problem, that of finding a sustainable equilibrium in the metabolic system. Indeed, there exist ready-to-exploit cell structures to be used as platforms for new functionalities; one only has to take the next step and tame and cultivate the bacteria, domesticating them. Rather than constructing completely new artificial cells, one has to obey those ways of thinking that nature has followed: New structures are constructed on existing ones, just redirecting and boosting the evolution.

This all does not only apply to medical engineering: The key challenge in future industrial systems is their life-long maintenance. It would be reasonable to implement some level of cybernetic self-repair or adaptability in those systems, too, rather than only fixing the broken parts. Tomorrow's industrial systems also need to be in balance with their surroundings, not fight against it.

The presented emergent models were just models, and models should not be mixed with reality. For example, how could one motivate the "chemical pattern matching" as a fundamental cellular principle? How could a system with no central control accomplish it, even if it would like to do it? And, to reach *real system biology,* it is not only the internal behaviors within the cell that need to be captured — the next level is the coordination among the cells, and, generally, among populations. The challenge is to find out how such orchestration can be explained in terms of local actions only.

When studying natural systems, it is difficult to get farther only studying available data and existing systems — one needs stronger modeling principles. One should not only try to explain phenomena: One should proactively try to find the underlying principles. This kind of ideas crystallize in the question: *What are the goals of systems?*