13 Simplicity in a behavioural, non-parametric context

Dirk Tempelaar

Motivation

The concept of simplicity is inextricably related to models and the process of modelling (dynamic) phenomena. For that reason, the definition of simplicity critically depends upon the modelling paradigm chosen. In this contribution, we will describe a non-parametric approach to modelling dynamic phenomena, and the role the concept of simplicity plays within that approach. But before turning to the descriptive part, we will first mention the most important sources of inspiration which led us to choose this specific modelling paradigm and the behavioural definition of simplicity that is congruent to it.

The first source of inspiration is to be found in the apparent paradox built into parametric versions of the concept of simplicity, like Schwartz’s information criterion, SIC, Akaike’s information criterion, AIC, or the minimal description length, MDL (see Keuzenkamp and McAleer, 1995). Concepts such as simplicity, or complexity as its mirror image, are particularly useful in modelling exercises in which economic theory only plays a rather modest role. This type of modelling application is characterized by a lack of economic theory or, just the opposite, by a rich affluence of contradicting economic hypotheses, and puts the data in the forefront when deciding upon the model structure. In so doing it creates a need for principles like Occam’s razor. But at the same time, in this type of application with weak theory and strong data, deep structural parameters are scarce, or even absent. And when model parameters don’t have a clear physical (economic) interpretation, one can wonder if parametric implementations of Occam’s razor aren’t a contradiction in themselves.

The lack of interpretability of parameter values in economic models is just one of the aspects of Kalman’s critique of econometrics. In his theory of modelling, Kalman (1980, 1982a, 1983) makes a distinction between natural sciences, in which laws of nature live, and system-determined sciences, such as economics, computer science or engineering, in which system-independent laws do not exist. In the absence of laws and abso-
lutenly definable parameters, the system-theoretic paradigm of modelling becomes useful. According to this paradigm, modelling is a data-driven procedure, so as to prevent the use of any form of 'prejudice': assumptions unrelated to the data but that are imposed on the data without any check against the data (Kalman, 1982b). Prejudice can manifest itself in different ways, such as not treating all variables symmetrically, but instead assuming a priori classification into noiseless variables (the exogenous ones) and noisy variables (the endogenous ones); economic restrictions on parameter values to achieve over-identification and probabilistic assumptions. It is our ambition to opt for a modelling framework and an interpretation of the concept of simplicity that meets Kalman’s critique and avoids untestable assumptions as much as possible.

Together with the issue of internal consistency, we should consider the external consistency of informational criteria for simplicity. In defining a concept that has so many connotations in daily life, and is used in so many different domains, one at least hopes to be able to find a definition that is congruent to its meaning in everyday language and can be transferred from one domain to another. The first problem with defining simplicity in terms of the number of adjustable parameters of the equations describing the object or concept under consideration is that this definition severely restricts its range of applicability. We more often than not use the concept in situations in which parameters and equations are alien. Moreover, information-criterion-based definitions of simplicity possess in some cases rather counter-intuitive properties. To clarify, just compare two different pieces of canvas, the first being tightly stretched within a picture frame, the second not being tightened at all and, for that reason, folded and crumpled. By all possible standards, the second canvas is more complex and less simple than the first one, being more curved, less aesthetic, and necessitating a higher dimensional space to describe it (see next section for an elaboration on these issues). However, when counting parameters, we quite surprisingly arrive at the opposite conclusion: the first canvas, being constrained in its manifestation (and realizing we need parameters to define such a constraint) is more complex than the second, being not constrained at all! The fact that simplicity is not so much defined in terms of the manifestation itself, but rather in terms of the mechanisms that constrain that manifestation, is the cause of this counter-intuitive outcome. The more constrained the manifestation, the more complex the mechanism of constraining but, at least at an intuitive level, the simpler the (class of) shapes that are left over after constrainning. This property of counter-intuitive outcomes is not specific to information-criterion-based implementations of simplicity only, but can be gen-
eralized to other syntactic implementations of the concept of simplicity (as opposed to semantic construals of simplicity, for which we will opt in this contribution: see Sober, 2001).

In the modelling of dynamic phenomena from observed data, the generic case is that no (non-trivial) models exist that are able to explain the data exactly. The standard way out of the problem that all models, from a Popperian point of view, are falsified by the data is to adopt a probabilistic modelling approach. This comes down to postulating a model as a set of equations containing as yet unspecified parameters, and postulating random elements, for example by assuming that the observed variables, or observed time series, are a realization of a stochastic process. Such assumptions guarantee that every finite observed time series can occur with positive probability density and in this sense the data will no longer falsify the model. In specific applications, and thus as a general philosophy, this approach has drawbacks. For example, in many applications the lack of fit between data and any model is not in the first place due to randomness or measurement noise, but due to the fact that one consciously uses a model that is too simple: that cannot capture the complexity of the phenomenon one has observed. In such a case, it is appealing to formulate the modelling problem as a deterministic approximation problem instead (Willems, 1986b, p. 675). This raises the question of what is meant by an approximate model and how one should judge such a model against data that, in a strict sense, falsify the model. The paradigm on which approximate deterministic modelling is based are the objectives of low complexity, or high simplicity, and high accuracy, and not unbiasedness, consistency and efficiency as would be the case in the probabilistic framework.

A further source of inspiration is drawn from the apparent contradiction in the views of Popper and Jeffreys on simplicity, as chronicled by Keuzenkamp and McAleer (1995, pp. 10, 12). Keuzenkamp and McAleer demonstrate that Popper's view, implying that 'theories that are simple and easy to falsify, i.e. theories with high empirical content' are linked with improbability, is flawed, judged on its Bayesian, probabilistic consistency. As an alternative, Keuzenkamp and McAleer opt for Jeffreys' view on simplicity: 'variation is random until the contrary is shown', implying that probability is increasing with simplicity. The judgment on these two views seems to go hand in hand with the methodological framework used to assess both views. In this contribution, we will break a larse for Popper's view, this time reasoning within a behavioural framework of modelling.
Behavioural approach

In our common parlance, simplicity is related to manifestation: ‘For both objects and concepts the most obvious meanings of “simple” relate to homogeneity in structure and the absence, or at least small number, of discernible parts or partitions ... A simple dress, painting, symphony, theory, or society has fewer lines, curves, colors, classes, or other internal differentiations than a more complicated one’ (Slobodkin, 1992, p. 2). This description is the very first attempt by Slobodkin to capture the idea of simplicity and the starting point of what can be viewed as a kind of Bildungsroman of the concept simplicity within the domains of art, religion and science. The pinnacle of this development is to be found in the last chapter, ‘Masters of Reality’, where simplicity is defined as dimensionality. Plato’s myth of the cave, populated by prisoners who observe only the two-dimensional shadows of the three-dimensional objects outside their cave, serves as an illuminating example (Slobodkin, 1992, p. 206). In principle, by noting the changes in shapes of these shadows, the prisoners might eventually have developed a three-dimensional theory of objects; but they lacked the ability to do that. However, Slobodkin continues:

We have now the general techniques for considering how many dimensions are relevant to analyse some sets of data. In a general sense, more complex things require more terms in their description than more simple things. The number of dimensions required to specify a situation depends on how much complexity we are willing to consider, and this may depend on what we want to do or say... It is ... often useful to ... consider how many dimensions might be relevant and how they should be arranged in ‘space’ in order to suggest ways of thinking about data. There is no necessary numerical limit to the number of dimensions involved. The more kinds of things we know about anything, the more likely we are able to deal with it in a realistic way. (Slobodkin, 1992, pp. 206, 207)

Opting for a semantic construal of simplicity, and also defining the complexity of a system – being the opposite of simplicity – as the dimension of the system, characterizes the modelling approach called ‘behavioural modelling’. Without giving a full exposition on this approach – for that purpose we refer to the numerous publications of Willems and Heij, on which this contribution is largely based – it is our intention to provide the reader with a sufficient notion of behavioural modelling as to sense the perspective of an alternative definition of simplicity and its role in modelling. In this behavioural approach, all concepts are semantic by nature: simplicity, complexity, but also accuracy and the concept model itself. This semantic nature implies a non-parametric approach: a model is not the same thing as a struc-
Simplicity in a behavioural, non-parametric context

atural model, consisting of mathematical equations with unknown parameters; a model is defined in terms of its behaviour, that is, the set of all outcomes that are allowed by the system. In the dynamic case, and we will concentrate on that, these are trajectories in time. Behavioural model selection, like traditional estimation, involves a trade-off between the complexity and the accuracy of the model. Accuracy is measured as the error or misfit of the system with respect to the observed behaviour (time series). Contrary to other modelling approaches, fit isn’t measured in terms of local restrictions, as in equation error methods, but in terms of the global distance from the observed time series to the nearest time series within the model. This implies a non-parametric implementation of fit: it is defined purely in terms of external behaviour and involves no parameters. As we have indicated before, complexity, the second criterion, is also defined in a non-parametric way. In the static case, the complexity is the dimension of the system, which equals the number of driving forces (inputs or exogenous variables). In the dynamic case, depending upon the representation chosen, complexity is determined by the number of states and the number of driving forces (in a state space representation), or by the number of equations and the number of lags (in a polynomial representation).

Before we elaborate further on the behavioural approach, we will descend one last time into Plato’s cave, to discover if and to what degree the above definition of complexity, or simplicity, does indeed correspond to its usual meaning. The prisoners in the cave couldn’t observe the three-dimensional world outside, but only a two-dimensional reduction of it, in the form of shades projected on the wall. Both in Slobodkin’s language and in behavioural modelling, this reduction is a simplification: the shades have a simpler form than their three-dimensional originals. One can pursue this example even one step further by assuming that Plato’s cave resembles a cavern with an extremely narrow entrance. In that case, all shadows are one-dimensional, implying a further reduction of the behaviour compatible with the projection operation. However, if we describe the behaviour by means of a parametric model, each reduction step implies the introduction of another equation containing several parameters, thus decreasing parametric simplicity instead of increasing it.

**Behavioural models and equations**

Assume we have a *phenomenon* we want to model. In defining a proper language, we assume that the phenomenon produces elements in a set $U$
that we call the universe. Elements of $U$ will be called the outcomes of the phenomenon. Now, a mathematical model for the phenomenon claims that certain outcomes are possible, whilst others are not. Hence a model recognizes a certain subset $B$ of $U$. This subset will be called the behaviour (of the model). Formally: a behavioural model is a pair $(U, B)$ with $U$ the universe – its elements are called outcomes – and $B$ the behaviour.

**Example:** the price of a good being non-negative may be viewed as modelling the phenomenon price. Thus $U = \mathbb{R}$ and $B = [0, \infty)$. And a second example: the production possibilities of an economy, facing a Cobb-Douglas production function, is described by the model: $U = \mathbb{R}^3_+$ and $B = \{(Y, K, L) \in \mathbb{R}^3_+: Y = \alpha K^\beta L^{1-\beta}; \alpha, \beta \in \mathbb{R}_+; 0 \leq \alpha; 0 \leq \beta \leq 1\}$.

In applications, models are more often than not described by equations. We will make use of them in the behavioural approach as follows: let $U$ be a universe, $E$ an abstract set, called the equating space, and $f: U \to E$. The behavioural model $(U, B)$ with $B = \{u \in U | f(u) = 0\}$ is said to be described by behavioural equation(s).

Whereas equations uniquely specify the behaviour, the converse is obviously not true. Since we have a tendency to think of mathematical models in terms of equations, most models being presented in that form, it is important to emphasize their ancillary role: it is the behaviour, the solution set of the behavioural equations, and not the behavioural equations themselves, which is the essential result of a modelling procedure. The equations are just a (non-unique) representation of that behaviour.

**Dynamic systems**

The next step is the application of this view of mathematical models to set up a language for dynamic systems. In many disciplines, there is a tendency to view systems as processors, producing output signals from input signals. There are without doubt applications where it is eminently clear what the inputs and the outputs are. However, there are also many applications where this input–output structure is not at all evident. So we will view a dynamic system in the logical context of static systems defined before as symmetric models, in which all variables are treated on the same footing. The only distinguishing feature is that now the phenomenon produces outcomes that are functions of time: the universe $U$ is a function space. More formally, this can be described as follows: a dynamic system $\Sigma$ is a triple, $\Sigma = (T, W, B)$, with $T \subset \mathbb{R}$ the time axis, $W$ the signal space, and $B \subset W^T$ the behaviour of the system.
Simplicity in a behavioural, non-parametric context

This description does not refer to input-output maps, relations or behavioural equations: a dynamic system is defined as a family of trajectories. In the triptych that constitutes a dynamic system, \( T \) is the time set of interest. Usually \( T \) is an (infinite) interval in \( \mathbb{R} \) or \( \mathbb{Z} \), and in the discrete time series case \( T = \mathbb{Z} \) or \( T = \mathbb{Z}_+ \). \( W \) is the space in which the time signals that the system produces take on their values, and \( B \) is a family of \( W \)-valued time trajectories. Whereas the sets \( T \) and \( W \) define the setting, \( B \) formalizes the laws that govern the system. According to the dynamic model \( \Sigma \), time signals in \( B \) can in principle occur and are compatible with the laws of the system, while those outside \( B \) cannot occur and are prohibited.

Inherent to modelling is the use of a restricted model class. This restriction is prior in nature: independent of the data, we impose a certain structure on the models we are willing to accept. In the model selection phase, we subsequently infer additional structure from the data. The restrictions we introduce are the following: we focus on discrete time systems, so the time set equals the set of integers: \( T = \mathbb{Z} \). The system we study contains \( q \) variables that take on real values and evolve in discrete time: so the signal set is \( \mathbb{R}^q \) for some integer \( q \geq 1 \). A system then is a set of \( q \)-dimensional time series: \( B \subset (\mathbb{R}^q)^\mathbb{Z} \). As model class we take the subclass of the systems in \((\mathbb{R}^q)^\mathbb{Z}\) that are linear, time invariant and complete.

- A system \( B \subset (\mathbb{R}^q)^\mathbb{Z} \) is called **linear** if it is a linear subspace of \((\mathbb{R}^q)^\mathbb{Z}\).
- A system \( B \) is called **time invariant** if its time lag \( LB \) also satisfies the laws of the system: \( LB = B \).
- A system \( B \) is called **complete** if, in order to check whether a time series \( w \) belongs to \( B \) or not, it is sufficient to consider it does so for all finite time intervals. Completeness is only an interesting property when \( T \) is an unbounded subset of \( \mathbb{R} \), for example in the discrete time series case \( T = \mathbb{Z} \) or \( T = \mathbb{Z}_+ \). Then completeness simply says that the behaviour at plus or minus infinity is of no consequence for deciding whether or not \( w \) belongs to \( B \) (the behaviour is closed in the topology of pointwise convergence).

The class of linear, time-invariant and complete systems is of special interest since each member of this class corresponds to a linear, time-invariant, finite-dimensional system. And such systems can be represented with behavioural equations having alternative formats: each element of the model class \( B \) of linear, time-invariant and complete systems allows a representation according to each of the following parametrizations (see Willems, 1986a, Heij, 1989, or Heij et al., 1997):
• polynomial (autoregressive) representation: \( R(L)w = 0 \)
• input/output representation: \( P(L)y = Q(L)u, \ w = \text{col}(u, y) \)
• state space representation: \( L^{-1}x = Ax + Bu, \ w = Cx + Du, \ w = \text{col}(u, y) \)
• transfer function representation: \( w = G(L)u. \)

These representations exhibit different degrees of structure. For example, in the input/output representation a causality structure is shown, and external variables are subdivided into two classes: the inputs, the causes: free variables that are left unexplained; and the outputs, the effects: bound variables that are explained by the model. However, in the AR-representation such a causality structure is absent. State representations display the state as an auxiliary variable: the state is a memory function which shows what part of the past is relevant for the future behaviour. The input/state/output representations display both their causality and their memory structure.

The fact that each linear, time-invariant and complete system can be represented in each of the above formats implies for this class of systems that causality (and memory) is a matter of representation and not an \textit{a priori} axiom which needs to be imposed. In other words, in this model class one can always obtain causality by properly interpreting the variables.

**Modelling objectives**

The selection of an approximate deterministic model involves three ingredients: the objectives of modelling, denoted by \( \pi \), a set of conceivable data, denoted by \( D \), and a model class, denoted by \( M \). For model selection we need a procedure which for a given data set assigns a model in the prespecified class \( M \). This \textit{data modelling procedure} can be expressed as a map: \( P: D \rightarrow 2^M \). The aim of any procedure is to find models that are optimal with respect to the objectives \( \pi \).

In constructing models from data, two general objectives are placed in the forefront:

• we want to infer from the data as much \textit{structure} as possible, and
• the model is an \textit{accurate} description of the data.

Structure is generated by laws: the more laws a model claims, the more structure it implies. Each law imposes restrictions on the variables under consideration, and thus implies a reduction of the dimension of the solution space. For that reason, greater simplicity is a synonym for more
structure, in the behavioural view. Also, a simpler model is more easily falsified, since it counts more restrictions each of which can be falsified separately. As a consequence, the simplicity principle, the falsifiability principle and the striving for structure are synonyms, once we accept the behavioural view on systems and modelling.

The second principle is the corroboration principle: although we want to infer as many laws from the data as possible, we want to accept laws only if there is sufficient evidence for them from the data. Such evidence for a law is measured applying a criterion of fit of the law with respect to the data.

As a next step, we assume that the objectives \( \pi \) can be specified by a complexity map \( c: M \rightarrow C \) and a misfit map \( e: D \times M \rightarrow E \). Both spaces \( C \) and \( E \) are assumed to be partially ordered. It is desirable to design procedures that find models with both low complexity and small misfit: simple models that are corroborated by the data. However, in general these goals are conflicting. We therefore assume that \( \pi \) can be expressed as a utility function, that is a map \( u: C \times E \rightarrow U \), where \( U \) is a partially ordered set. In this case, our selection procedure \( P_\pi \) picks that model for which the complexity and misfit are such that the corresponding utility is maximal.

In case the complexity space \( C \) and the misfit space \( E \) are both totally ordered, two specific modelling procedures can be classified:

- \( P_{c,\text{tol}} \) or modelling under a complexity constraint: this procedure determines the most accurate model amongst those models that obey a prespecified maximal tolerated complexity level \( c_{\text{tol}} \).
- \( P_{e,\text{tol}} \) or modelling under a misfit constraint: this procedure determines the most simple model amongst those models that obey a prespecified maximal tolerated misfit level \( e_{\text{tol}} \).

These two different modelling strategies show some similarities to the 'simple to general' approach and the 'general to specific' or reductionist approach to modelling, respectively.

**Example** (see Heij, 1989, pp. 25, 28): the maximum likelihood estimation of a univariate time-series model. For instance, let \( M \) consist of the class of ARMA-models and \( D \) be the set of finite times series. For \( M \in M \) we can define the complexity \( c(M) \) as e.g. \( \max \{ d_1, d_2 \} \) with \( d_1 \) the degree of the AR-part and \( d_2 \) the degree of the MA-part. For \( M \in M \) and \( w \in D \) we define the misfit \( e(w, M) \) as the inverse of the likelihood of \( M \) for \( w \). For given \( e_{\text{tol}} \), the procedure \( P_{c,\text{tol}} \) models the time series by means of an ARMA-model of maximum likelihood, under the restriction
\[ \max \{d_1, d_2\} < c_{\text{tol}}. \] For ‘non-degenerate data’, the optimal solution satisfies \( d_1 = d_2 = c_{\text{tol}}. \) For given minimal tolerated likelihood \( 1/\varepsilon_{\text{tol}} \), the procedure \( P_{\varepsilon_{\text{tol}}} \) minimizes \( \max \{d_1, d_2\} \) under the likelihood constraint, which gives a solution of the model order selection problem.

**Behavioural modelling**

The restriction of the model class \( M \) to the aforementioned class of linear, time-invariant and finite-dimensional systems results in the following modelling problem: find a decomposition \( w = \hat{w} + \tilde{w} \) of the observed \( q \)-dimensional time series \( w \) where the unobserved, latent trajectory \( \hat{w} \) is the behaviour of a linear system. Such a behaviour is a subset \( B \subset (\mathbb{R}^n)^T \) of the set of all time series over \( Z \). For ‘non-degenerate data’ only high-dimensional linear systems with long lag structures will model the observed data exactly. In that case \( w = \hat{w} \). Linear models corresponding to simpler behaviours induce a non-zero approximation error \( \tilde{w} \). As stated before, the modelling objective is to trade off the accuracy of the model, inversely related to the size of this approximation error with the simplicity of the model. We measure the approximation error by the squared distance between the data and the linear system, that is:

\[
\|w - \hat{w}\|^2 = \sum_{t=1}^{T} \sum_{j=1}^{q} [w_j(t) - \hat{w}_j(t)]^2
\]

with \( T = [1..T] \) the time-interval of the observations.

The next step is to define the simplicity/complextiy of the behavioural model. Since each behavioural model allows for several representations, we can interpret this definition in different ways. First of all, \( \hat{w} \) can be represented as the solution set of the polynomial (autoregressive) equations:

\[ R(L) \cdot \hat{w} = 0 \]

with \( R \) a full row rank polynomial matrix in the lag operator \( L \). We interpret this representation as an input–output system in polynomial form, where \( p \), the rank of \( R \), so the number of independent equations, equals the number of outputs, and \( m := q - p \) the number of inputs. In this representation, we denote with \( n \) the sum of the degrees of the \( p \) polynomial equations. Those degrees are the longest lags contained in each equation, so \( n \), being the sum of these separate lags, can be viewed as the ‘total lag’. For that reason, a further interpretation of \( n \) is the minimal number of initial conditions required to express future outputs in terms of future inputs.
Simplicity in a behavioural, non-parametric context

Alternatively, \( \hat{w} \) can be represented in state space form:

\[
x_{t+1} = A \cdot x_t + B \cdot v_t, \quad \hat{w} = C \cdot x_t + D \cdot v_t
\]

Here \( v \) is an \( m \)-dimensional auxiliary vector of unrestricted and unobserved driving variables (inputs), and \( x \) is an \( n \)-dimensional vector of unobserved state-variables. Both representations are highly non-unique. We reduce the non-uniqueness by focusing on minimal representations: representations that have fewer driving variables \( (m) \), or the same number of driving variables but fewer states \( (n) \), than any other representation.

A third representation is in terms of transfer functions:

\[
\hat{w} = G(L)u
\]

In this representation, \( m \) is the rank of the transfer function \( G \), and \( n \) the so-called McMillan degree, the sum of the Kronecker indices of the \( q \) polynomials (see Heij et al., 1997).

The pair \((m, n)\) of a minimal representation is used as a measure of the complexity of the system \( B \). The intuition behind this choice of the two indices determining complexity will be clear from the preceding discussion. The complexity measure \((m, n)\) is directly related to the dimension of the linear space \( B_T \), the restriction of the behaviour \( B \) to the time interval \([1..T]\). That dimension equals \( mT + n \). So the complexity is defined by the number of inputs and the number of states of the system when referring to a state space representation. Or by the number of inputs and the total lag (number of initial conditions) in case of a polynomial representation. The simpler a system, the fewer degrees of freedom it has. There are two different kinds of degrees of freedom: the initial ones, and the degrees of freedom at each time instant. The first equals the dimension of the state, the second equals the dimension of the input vector.

**Deterministic modelling**

In the static case, the modelling problem is to describe a finite number of points in \( \mathbb{R}^d \) by means of a linear subspace. So data set \( D \) consists of the finite subsets of \( \mathbb{R}^d \) and the model class \( M \) consists of the linear subspaces of \( \mathbb{R}^d \). In this case, the state dimension \( n = 0 \) and the measure of complexity of a model \( M \in M \) reduces to its dimension: \( c^d(M) = \dim(M) \). So the simpler the model, the lower its dimension, the more it excludes, the easier it is to falsify. In this static case, our complexity definition results in a total ordering, thereby fulfilling the requirements of the two procedures of minimizing misfit under a complexity constraint \( P_{\text{tol}} \) and minimizing complexity under a misfit constraint \( P_{\text{tol}} \). Those procedures are based on singular value decomposition of the empirical covariance matrix.
(see Willems, 1986a, or Heij, 1989). There is a close relationship between these procedures and total least squares.

In the dynamic case, the modelling problem is to describe a finite number of q-valued vectors in $(\mathbb{R}^q)^T$ by means of a linear, time-invariant and finite-dimensional system. As a measure of complexity, we take the pair $(m, n)$. High complexity implies few and high-order laws: complex systems impose few restrictions on the behaviour, and are thus characterized by many degrees of freedom. In dynamic systems, this can be the result of a small number of laws, and laws that are 'loose' in nature, in the sense that they involve large lags and therefore only constrain the behaviour if long time series are considered. However, this definition of complexity only induces a partial ordering, not total ordering. So the procedure of minimizing misfit under a complexity constraint $P_{e/tot}$ and that minimizing complexity under a misfit constraint $P_{e/tot}$ has to contain a grid-search step: for a range of complexities $(m, n)$, the accuracy of the model is calculated and traded-off against the simplicity. $P_{e/tot}$ first finds the maximal number of zero-order relations under the misfit constraint (as is the static problem). Among the solutions, which are in general non-unique, those with minimal misfit are preferred. Subsequently the number of first-order relations is maximized, again under the misfit constraint. And so on. $P_{e/tot}$ first minimizes the misfit of the zero-order laws that are compatible with the tolerated complexity, then continues with the truly first-order laws, and so on.

**Dynamic factor models**

A direct extension of the above modelling procedure is to consider stochastic systems. The extension is bought at some cost, e.g. in the form of making assumptions on the stochastic processes $w$, $\hat{w}$ and $\tilde{w}$. But the cost brings some additional returns: the model of behaviour $\hat{w}$ can be interpreted as a factor model. For an elaboration of this stochastic case, we refer to Heij et al. (1997) and Scherrer and Heij (1997).

**Conclusion and comparisons**

In this contribution an (informal) introduction to the behavioural systems framework and approximate modelling is given. This approach allows for a non-parametric definition of the concepts of simplicity, complexity and accuracy. Among other things, this approach seems to be a way out of the apparent paradox built in parametric applications of the concept of simplicity: they require parameters to be important aspects of
the modelling process, but will only contribute to that process in cases where there are no obvious candidates for such structural parameters.

The second advantage of the behavioural approach is its capacity to reconcile Popper's view on simplicity and this framework of modelling (dynamic) phenomena, thereby rejecting Jeffreys' simplicity postulate. The hypothesis that all variation is random, in the sense of being unexplained by the model, is not the simplest hypothesis one can think of, it is just the reverse: it is the most complex hypothesis one can possibly suggest. The more structure is added, the more manifestations are excluded, the simpler the hypothesis gets, and the more easily it is falsified.

If we regard the opening paragraph of Keuzenkamp and McAleer (1997): 'it seems natural to consider a trade-off between simplicity and empirical validity. In order to determine such a trade-off, simplicity has to be defined precisely, and the same holds for empirical validity', as summarizing the major objective of any theory on simplicity, the behavioural approach performs its duty. And how does it relate to other contributions on the concept of simplicity? In the framework proposed by Sober (2001), behavioural simplicity is a semantic construal. Our definition even allows us to answer one of the (rhetorical) questions posed in Sober's contribution: the hypothesis of egoism is indeed simpler than the hypothesis of motivational pluralism, not because of the smaller number of postulated types of ultimate desire, but because it is more restrictive with regard to the behaviour that is compatible with the hypothesis: it restricts more, and thus is easier to falsify (the other examples in Sober's contribution are more difficult to integrate into our behavioural framework). Spanos (2001) offers a further binary classification of simplicity concepts (albeit with less appealing names). Behavioural simplicity falls into his first category.

Although the above quote from Keuzenkamp and McAleer isn't precise enough to rigorously deduce the following principle, it seems at least to suggest it: in trading off accuracy and simplicity, we trade off a property of the model alone (simplicity) and a property of both model and data (accuracy). There seem to be some attractive intuitive arguments for using a data-independent definition of simplicity, for example, why should it be obvious that adding one observation to the data sample still using the same model increases the complexity? The large watershed between the several contributions on simplicity is to be found in this characteristic: is simplicity a property of the model (so data-independent), or of the data and the model together (so data-dependent)? All information-theory-based definitions of simplicity fall into this second category: stochastic complexity (Rissanen, 2001), Kolmogorov complexity (Vitányi and Li, 2001) and Simon's parsimony principle (Simon, 2001). Given the
strong ties of information theory and the application areas of coding, speech processing and signal recognition, this does not surprise. The main objective of modelling in these areas is to achieve efficiency gains in transferring data. But after the transfer has taken place, the original data is to be reconstructed. This gives a very special meaning to the word noise (see e.g. Rissanen, 2001): it is that part of the data for which no efficiency gains can be achieved. But in this area there isn't such a thing as an approximate model, so the code length of the noise is an integral part of Rissanen’s stochastic complexity, and even has the same weight as the code length of the model. But then, there is in fact no real trade-off between accuracy and simplicity: the definition of simplicity by its very nature presumes full accuracy, and no accuracy-loss is tolerated. In contrast, data-independent definitions of simplicity seem to be more natural in application areas where noise is seen as error: inevitable but uninteresting, since it disturbs our view on the ‘true underlying system’. In that case, the only objective is to eliminate this noise, and in no way to reconstruct it.

The view that simplicity is a model property, accuracy is a property of the relation between model and data, and the essence of modelling is the trade-off between those two concepts, shares the behavioural approach with interpretations of simplicity based on ‘counting parameters’ (see Keuzenkamp and McAleer, 1997, p. 553). But besides this common background, the two approaches choose sharply diverging routes in their definition of simplicity, as is illustrated in this chapter.

REFERENCES


Simplicity in a behavioural, non-parametric context


