Prediction and Modularity in Dynamical Systems

Artemy Kolchinsky[1,2] and Luis M. Rocha[1,2]

[1] School of Informatics and Computing, Indiana University, Bloomington IN 47401, USA
[2] FLAD Computational Biology Collaboratorium, Instituto Gulbenkian de Ciência, Portugal
{akolchin,rocha}@indiana.edu

Abstract

Identifying and understanding modular organizations is centrally important in the study of complex systems. Several approaches to this problem have been advanced, many framed in information-theoretic terms. Our treatment starts from the complementary point of view of statistical modeling and prediction of dynamical systems. It is known that for finite amounts of training data, simpler models can have greater predictive power than more complex ones. We use the trade-off between model simplicity and predictive accuracy to generate optimal multiscale decompositions of dynamical networks into weakly-coupled, simple modules. State-dependent and causal versions of our method are also proposed.

Introduction

The study of complex dynamical systems – such as gene regulatory networks (Han et al., 2004), structural and functional brain networks (Bullmore and Sporns, 2009), ecological food webs (Krause et al., 2003), and others (Hartwell et al., 1999, Schlosser and Wagner, 2004) – has frequently uncovered the presence of modularity. Broadly speaking, modular systems are composed of tightly-integrated subsystems, called *modules,* which are in turn weakly coupled to one another.

Numerous explanations have been proposed for the function of modularity in complex systems, only a few of which are mentioned here. Simon (1962) suggested that modularity can contain the effects of harmful perturbations and lead to greater developmental and operational robustness, especially when modules are hierarchically arranged. Kashtan and Alon (2005) argued that modular systems can take advantage of reusability when adapting to changing combinations of fixed environmental tasks. Tononi et al. (1998) proposed that modularity balances the conflicting needs for subsystems that are functionally specialized but also integrated into globally coherent states. Notably, it has also been shown to arise as a result of non-adaptive processes, such as neutral evolution of gene regulatory networks (Force et al., 2005, Solé and Valverde, 2008) and stochastic fluctuations in network connectivity patterns (Guimera et al., 2004).

Though the concept of modularity has acquired a central place in the study of complex systems, its meaning and operationalization varies widely between scientific paradigms, fields, and processes of interest. In the biological sciences alone, one can find references to *structural*, *developmental*, *physiological*, *variational*, and *functional modularity* (Winther, 2001, Wagner et al., 2007), among others. In this work, we propose a formal notion of modularity based on statistical modeling. Our approach applies to a broad class of discrete-time multivariate dynamics, whether represented by dynamic models, such as Boolean or dynamic Bayesian networks, or empirical distributions estimated from time series recordings. Unlike much recent work on community-structure in static graphs, we identify modularity in the organization of dynamically interacting components. We argue that in addition to being useful for analysis of real-life dynamical systems, our approach can shed light on connections between notions of modularity utilized in different domains, as well as the general role of modularity in modeling.

The next section provides a brief background on information theory. We then outline traditional information-theoretic approaches to modularity in dynamical systems, and develop our own treatment in terms of statistical modeling. After applying it to an example dynamical system, we consider state-dependent and causal versions of modular decompositions. We conclude by discussing issues of parameterization, directions for further work, and connections between our method and broader questions of modeling.

Information theory

Information theory provides principled measures of information transfer and statistical dependence in distributed systems. As such, it is well-suited for quantifying measures of coupling and modularity.

To review, Shannon *entropy* measures the uncertainty in the measurement outcomes of a random variable. If X is a discrete random variable with an associated probability distribution $P(X)$, then its entropy is:

$$H(X) = -\sum_{x \in X} P(x) \log P(x)$$

A random variable that takes a single value with probability 1 has an entropy of 0, while an equiprobable random variable assumes the maximum entropy of $\log |X|$, where $|X|$ is the number of possible outcomes. When the base of the logarithm is 2, as in this work, the units of entropy are *bits* (1 bit is the uncertainty in the choice between 2 equally possible outcomes). Because measuring a variable reduces uncertainty about its value, entropy can also be considered a measure of information.

When provided with a joint distribution over two random variables such as $P(X, Y)$, *conditional entropy* measures the expected uncertainty in the value of one variable given that the value of the other is known:

$$H(X|Y) = H(X, Y) - H(Y) = -\sum_{x,y} P(x, y) \log P(x|y)$$

Mutual information is a symmetric measure of nonlinear correlation between two random variables. Expressed as the difference between entropy and conditional entropy, it can be interpreted as the reduction in uncertainty about the value of one random variable provided by knowledge of the other:

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(Y, X) \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \end{aligned}$$

Mutual information captures the amount of constraint in the joint distribution of two variables not present in their marginal distributions. It is equal to 0 when two variables are statistically independent, and reaches its maximum possible value of $\min\{H(X), H(Y)\}$ when one variable is a function of the other.

Mutual information can be extended to the case of more than two variables. Let random vector $\mathbf{X}$=$(X_1, X_2, \ldots, X_L)$ with distribution $P(\mathbf{X})$ represent the state of a system composed of L distinct variables. The total constraint in this system not present in any single variable is measured by a multivariate version of mutual information, often called *multi-information* (Studeny and Vejnarova, 1998) or *integration* (Tononi et al., 1994):

$$\begin{aligned} \mathcal{I}(\mathbf{X}) &= \sum_{i=1}^{L} H(X_i) - H(\mathbf{X}) \qquad (1) \\ &= \sum_{\mathbf{x}} P(\mathbf{x}) \log \frac{P(\mathbf{x})}{\prod_{i=1}^{L} P(x_i)} \end{aligned}$$

Kullback-Leibler (KL) divergence is a measure of the difference between two distributions:

$$\mathrm{KL}(P\|Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \qquad (2)$$

It is always positive and 0 iff $P = Q$, though it is not a distance because it is not symmetric. Importantly, many information-theoretic measures can be restated in terms of KL divergence. For example, the multi-information of eq. 1 is equal to the KL divergence between the distribution of $\mathbf{X}$ and a product of the marginal distributions over the individual variables of $\mathbf{X}$.

Modularity in multivariate dynamics

As previously mentioned, multi-information measures the total amount of higher-order constraint present among the variables of a multivariate system. It is 0 when these variables are independent, and increases when more statistical interaction between variables is present (Studeny and Vejnarova, 1998). For this reason, many formal approaches to modularity search for system transformations that minimize this measure.

Several kinds of transformations can be investigated. *Independent component analysis* attempts to minimize multi-information over the space of linear mappings (coordinate changes) of a multivariate system (Hyvärinen and Oja, 2000). A different approach, closer to the one pursued here, looks for *partitions* of system variables with low multi-information.

A partition π of set S is a set of mutually exclusive, nonempty subsets $B \subseteq S$, called *blocks*, such that $\bigcup_{B\in\pi} B = S$. For example, $\{\{1\}, \{2, 3\}\}$ and $\{\{1, 2, 3\}\}$ are two possible partitions of the set $\{1, 2, 3\}$. We also use a more concise notation: the two partitions above, for example, can be referred to as 1/23 and 123 respectively. Additionally, π_0 is used to indicate the *total partition*, which includes the entire set in a single block, i.e. $\pi_0 \equiv \{S\}$.

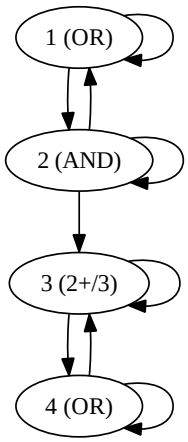
We look at partitions of $V = \{1, \ldots, L\}$, the set of indexes of the variables of random vector $\mathbf{X}$. For partition π and block $B \in \pi$, $P(\mathbf{X}_B)$ indicates the marginalization of $P(\mathbf{X})$ onto the variables whose indexes are in B. For example, $P\big(\mathbf{X}_{\{1,2\}}\big)$ is the marginal distribution of the first two variables of $\mathbf{X}$.

We define the multi-information of partition π as:

$$\mathcal{I}_\pi(\mathbf{X}) = \sum_{B\in\pi} H(\mathbf{X}_B) - H(\mathbf{X})$$

This measure quantifies the amount of constraint holding among the blocks of π. Finding partitions with low multi-information corresponds to identifying weakly-coupled subsystems. Variations on this theme appear in information-theoretic treatments of modularity starting from early cybernetics (Conant, 1972) to more recent approaches in computational neuroscience (Tononi and Sporns, 2003).

Multi-information is defined over a time-invariant distribution of system states. Though it does not account for the dynamic flow of information within a system, it can be generalized to this case. Assume a multivariate system with Markovian dynamics represented by $P(\mathbf{X}' = \mathbf{x}'|\mathbf{X} = \mathbf{x})$, the conditional probability distribution of transitioning to



Starting state	Future state
0000	0000
0001	0001
0010	0001
0011	0011
0100	1000
0101	1011
0110	1011
0111	1011
1000	1000
1001	1001
1010	1001
1011	1011
1100	1100
1101	1111
1110	1111
1111	1111

π	$\mathcal{I}_\pi(\mathbf{X}'\|\mathbf{X})$
1234	0.00
12 / 34	0.50
1 / 234	1.00
123 / 4	1.00
134 / 2	1.25
124 / 3	1.31
12 / 3 / 4	1.31
1 / 2 / 34	1.50
14 / 23	2.00
1 / 23 / 4	2.00
13 / 24	2.16
13 / 2 / 4	2.16
14 / 2 / 3	2.31
1 / 24 / 3	2.31
1 / 2 / 3 / 4	2.31

Figure 1: A simple four node Boolean network (nodes 1, 2, 3, and 4 perform OR, AND, majority, and OR update functions respectively). Its full state transition table is shown in center. On the right, the stochastic interaction of every possible partition of the network.

each future state $\mathbf{x}'$ given starting state $\mathbf{x}$, as well as $P(\mathbf{X} = \mathbf{x})$, the distribution over starting states.[1] The amount of information flowing dynamically among the blocks of π is called *stochastic interaction* (Ay and Wennekers, 2003). It is a conditional version of KL divergence between the transition distribution of the whole system and the product of marginal transition distributions of the variable blocks specified by partition π:

$$\begin{aligned} \mathcal{I}_\pi(\mathbf{X}'|\mathbf{X}) &= \sum_{B\in\pi} H(\mathbf{X}'_B|\mathbf{X}_B) - H(\mathbf{X}'|\mathbf{X}) \qquad (3) \\ &= \mathrm{KL}\left[P(\mathbf{X}'|\mathbf{X}) \middle\| \prod_{B\in\pi} P(\mathbf{X}'_B|\mathbf{X}_B)\right] \end{aligned}$$

These kinds of dynamic generalizations of multi-information have recently been proposed as measures of system-wide coupling in brain dynamics (Balduzzi and Tononi, 2008, Barrett et al., 2011).

A simple demonstration is provided by the Boolean network in fig. 1. It has four nodes, whose update functions are OR, AND, majority rule, and OR respectively. The stochastic interaction of each possible partition is provided, assuming a uniform distribution over starting states. For example, the partition 12/34 is the bi-partition having the lowest stochastic interaction: the block $\{1, 2\}$ has conditional entropy $H\big(\mathbf{X}'_{\{1,2\}}|\mathbf{X}_{\{1,2\}}\big) = 0$ (nodes 1 and 2 do not depend on the rest of the system, so their marginalized dynamics are deterministic), while block $\{3, 4\}$ has conditional entropy $H\big(\mathbf{X}'_{\{3,4\}}|\mathbf{X}_{\{3,4\}}\big) = 0.5$. Because the system as a whole is deterministic, $H(\mathbf{X}'|\mathbf{X}) = 0$ and the total stochastic interaction of partition 12/34 is $H\big(\mathbf{X}'_{\{1,2\}}|\mathbf{X}_{\{1,2\}}\big) + H\big(\mathbf{X}'_{\{3,4\}}|\mathbf{X}_{\{3,4\}}\big) - H(\mathbf{X}'|\mathbf{X}) = 0.5$.

Unfortunately, stochastic interaction is not a suitable cost function for identifying modular partitions of a multivariate dynamical system (similarly for multi-information and multivariate non-dynamical systems). In any such system, a minimal stochastic interaction of 0 will be assigned to the total partition π_0, and generally a partition will never have a greater stochastic interaction than any of its refinements (where one partition is a *refinement* of another if every block of the former is a subset of some block of the latter). Selecting partitions using stochastic interaction will thus favor partitions with large blocks, the total partition being a (possibly non-unique) global minimum.

Due to this, several authors have proposed normalizing factors that penalize large partitions (Conant, 1972, Balduzzi and Tononi, 2008). However, the derivation and justification of these normalizing terms is ad hoc. In this work, we approach the problem of identifying modules from the point of view of statistical prediction. This yields principled penalization terms for large partitions and leads us to uncover modular decompositions with clear interpretations in terms of statistical modeling.

Statistical modeling and modular decompositions

Information theory is intimately connected with statistical modeling (Rissanen, 2007). For example, assume a model that assigns a probability value to data $\mathbf{x}$:

$$Q(\mathbf{x}) = \int_\Theta Q(\mathbf{x}|\theta)\omega(\theta)\mathrm{d}\theta \qquad (4)$$

This term, called the *marginal likelihood* in the Bayesian literature, is the expectation of the likehood function $Q(\mathbf{x}|\theta)$ with respect to distribution $\omega(\theta)$ over parameter values.

$Q(\mathbf{x})$ is a measure of predictive fit to data, and its logarithm is often maximized over parameter distributions or model choices. Equivalently, one can minimize the negative of its logarithm, a measure of predictive error called *log loss*. If data samples are drawn from some true probability distribution $P(\mathbf{X} = \mathbf{x})$, then the expectation of the log loss of the marginal likelihood is:

$$-\sum_{\mathbf{x}\in\mathbf{X}} P(\mathbf{x}) \log Q(\mathbf{x}) = \mathrm{KL}(P\|Q) + H(P(\mathbf{X}))$$

The KL term (from eq. 2) is non-negative, and reaches its minimum of 0 when the model is perfectly fit, i.e. $Q = P$. It is a measure of excess prediction error of the model above the minimum possible. This minimum is specified by the entropy term, and depends only on the true distribution $P(\mathbf{X})$ and not on model or parameter choices.

A similar situation holds in the dynamic setting. We call

[1] We assume that the dynamics are stationary, in that the transition probability distribution does not change through time. Our analysis can also be applied to higher-order Markovian systems, though for simplicity they are not considered here.

dynamic models those that generate conditional distributions of multivariate future states $\mathbf{x}'$ given starting states $\mathbf{x}$:

$$Q(\mathbf{x}'|\mathbf{x}) = \int_{\Theta} Q(\mathbf{x}'|\mathbf{x},\theta)\omega(\theta)\mathrm{d}\theta$$

We look at statistical prediction of dynamical systems from the perspective of an agent who does not possess a perfectly fit model, but must learn a dynamic model given previous observations. The agent is provided with a set of factorized models: for each partition of system variables π, there is a dynamic model Q_π whose parameters and marginal likelihood obey the independence conditions imposed by the block structure of π:

$$Q_\pi(\mathbf{x}'|\mathbf{x}) \quad = \quad \prod_{B\in\pi} Q_\pi(\mathbf{x}'_B|\mathbf{x}_B) \tag{5}$$

The predictive performance of our agent depends on the chosen model and the amount of previously observed data. It can be quantified with a *risk function*, which here is the KL divergence between the true distribution $P(\mathbf{X}'|\mathbf{X})$ and the distribution predicted by a dynamic model (Haussler and Opper, 1997). The risk of model Q_π on the next sample, after observing N previous samples, is:

$$r_{N,Q_\pi} = \mathrm{KL}\big[P(\mathbf{X}'|\mathbf{X})\big\|Q_\pi\big(\mathbf{X}'\big|\mathbf{X},\mathbf{X}'^{1..N},\mathbf{X}^{1..N}\big)\big] \tag{6}$$

The expectation in the KL term is taken over the next sample of $\mathbf{X}', \mathbf{X}$, as well as N previous i.i.d. samples $\mathbf{X}'^{1..N}, \mathbf{X}^{1..N}$. The Bayesian *posterior predictive distribution*:

$$Q_\pi\big(\mathbf{x}'\big|\mathbf{x},\mathbf{x}'^{1..N},\mathbf{x}^{1..N}\big) = \int Q_\pi(\mathbf{x}'|\mathbf{x},\theta)Q_\pi\big(\theta\big|\mathbf{x}'^{1..N},\mathbf{x}^{1..N}\big)\mathrm{d}\theta$$

is the marginal likelihood of eq. 4, with the distribution over parameter values conditioned on N previous data samples. From the point of view of machine learning, such Bayesian updating of parameters in light of observed data corresponds to model *training*, while evaluating the expected model risk on new samples corresponds to model *testing*. More concretely, our dynamic models can be considered *supervised learners*: given data, they infer probabilistic mappings from inputs (starting states $\mathbf{X}$) to outputs (future states $\mathbf{X}'$).

Given the independence assumption of eq. 5, risk r_{N,Q_π} becomes:

$$\mathcal{I}_\pi(\mathbf{X}'|\mathbf{X}) + \sum_{B\in\pi} \mathrm{KL}\big[P(\mathbf{X}'_B|\mathbf{X}_B)\big\|Q_\pi\big(\mathbf{X}'_B\big|\mathbf{X}_B,\mathbf{X}'^{1..N}_B,\mathbf{X}^{1..N}_B\big)\big]$$

This form draws attention to the two components that contribute to risk (that is, predictive error). The *stochastic interaction term* (see also eq. 3) arises as a consequence of ignoring dynamic coupling between variables in different blocks. It is the minimal excess error of a factorized model (in which the dynamics of the variable blocks induced by partition π are independent) above an optimally fit whole-system model (where interactions between all variables can be captured).

The second term, called the *complexity term*, reflects the excess predictive error of a trained model above the minimum possible. It arises because a model trained on a finite amount of data maintains some uncertainty about optimal parameter values. For a given amount of training data, complex models (with larger parameter spaces) will have greater parameter uncertainty than simpler models, resulting in more excess predictive error. As $N \to \infty$, the complexity term can be asymptotically approximated by $\frac{d_\pi}{2N}$, where d_π refers to the number of parameters of model Q_π (Komaki, 1996, Barron and Hengartner, 1998). This yields:[2]

$$r_{N,Q_\pi} \approx \mathcal{I}_\pi(\mathbf{X}'|\mathbf{X}) + \frac{d_\pi}{2N} \tag{7}$$

For a given amount of training data N, the model with the lowest risk,

$$Q^\star(N) = \arg\min_{Q_\pi} r_{N,Q_\pi}$$

corresponds to the partition providing an optimal predictive decomposition of the system. Models that minimize risk offer a balance between two conflicting constraints: on one hand, low stochastic interaction (better predictions under optimal fit), on the other, low model complexity (easier parameter estimation with limited training data). Because partitions with smaller blocks (which have smaller-state-space dynamics representable by fewer parameters) generally induce simpler models, risk presents a principled cost function for identifying small, weakly-coupled modules. The amount of data N parameterizes this trade-off: as N increases, emphasis is shifted from the complexity term to the stochastic interaction term, and groups of variables whose dynamic interactions carry the most information while being easiest to learn are first to coalesce into multivariate blocks of the optimal model.[3] Thus, selecting optimal decompositions while increasing the amount of training data generates a modular multiscale decomposition of system variables. In the infinite data limit, the risk of each model Q_π reaches its minimum of $\mathcal{I}_\pi(\mathbf{X}'|\mathbf{X})$, and the partition corresponding to $Q^\star$ becomes the one with lowest stochastic integration (the total partition being a possibly non-unique minimum).

Decomposing a dynamical system

The complexity term in eq. 7 depends on the parametric form of the dynamic model. Though a variety of possibili-

[2] This approximation assumes continuously-parameterized models and standard regularity conditions. It also assumes that, for all π, some parameterization of Q_π offers a perfect fit to the factorized $\Pi_{B\in\pi}P(\mathbf{X}'_B|\mathbf{X}_B)$. It is possible to generalize beyond this case, where the factorizations of the true distribution are 'out-of-class' of the models Q_π.

[3] Minimizing risk can be seen as a form of information bottleneck (Tishby et al., 1999): it searches for factorized models whose parameters minimize information about training data while maximizing information about system dynamics; the size of the training data serves as a trade-off parameter.

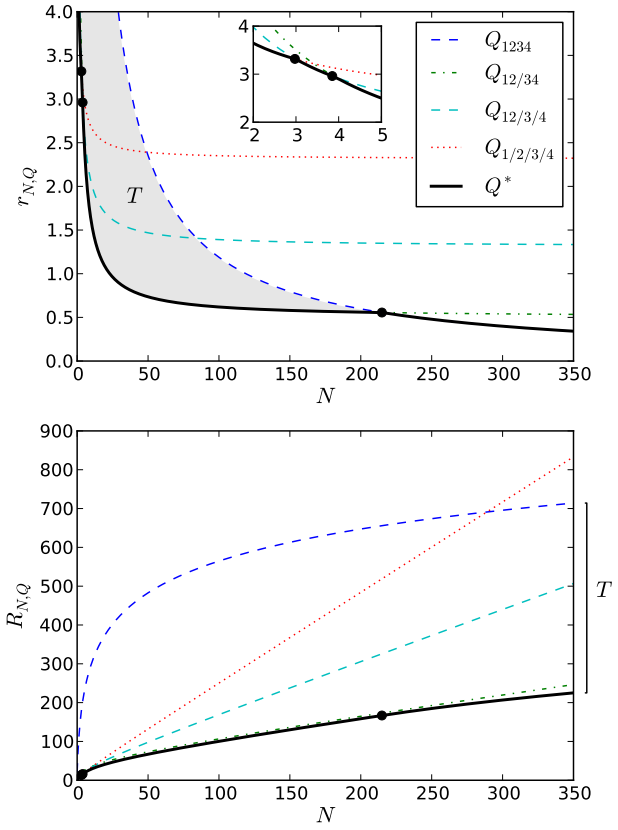


Figure 2: Top: approximate risk for optimally-predictive models of the Boolean network from fig. 1. Dots mark switches of the optimal model $Q^\star$; inset shows first two switches. Bottom: cumulative risk, or total accumulated prediction error for models plotted in the top graph. Total modularity (T) is asymptotic difference between cumulative risks of Q_{1234} and $Q^\star$ or, alternatively, area between lines corresponding to (non-cumulative) risks of Q_{1234} and $Q^\star$.

ties exist, here our dynamic models are assumed to be products of first-order Markov chains with Dirichlet priors. The number of parameters of model Q_π from this class is:

$$d_\pi = \sum_{B \in \pi} |\mathbf{X}_B|(|\mathbf{X}'_B| - 1) \tag{8}$$

where $|\mathbf{X}_B|$ is the number of supported starting state outcomes and $|\mathbf{X}'_B|$ is the number of possible future state outcomes of the variables with indexes in block B. For example, for a single block of Boolean variables with a fully supported starting state distribution, these are both equal to $2^{|B|}$. For this model class, the complexity term scales exponentially with the number of variables in each block.

As an example, we look at optimal decompositions of the network in fig. 1. Its risk, calculated using the approximation of eq. 7 and parameter counts of eq. 8, is shown at the top of fig. 2.[4] The risk is plotted for those models which reach minimum risk at some point of the training process, as well as that of the overall minimal risk model $Q^\star$ at each N. Predictive power is initially optimized by the model corresponding to partition 1/2/3/4 (the simplest model which treats all nodes independently). At $N \approx 3$ (inset), it is replaced by the model corresponding to partition 12/3/4 (variables 1 and 2 now merged into a single block); at $N \approx 4$ (inset), by the model corresponding to partition 12/34; and finally at $N \approx 215$, the most predictive model becomes the one corresponding to the total partition 1234.

Total modularity

So far, our measure of modularity has been parameterized by N, the amount of training data. Here, we derive a parameter-free measure of the *total modularity* in a dynamical system.

In our definition of risk (eq. 6), we used the *posterior predictive distribution* $Q_\pi(\mathbf{X}'|\mathbf{X}, \mathbf{X}'^{1..N}, \mathbf{X}^{1..N})$, the probability assigned to the next data sample by a model trained on N previous data samples. Given our assumptions, the following relationship holds between the *prior predictive distribution*, the probability an untrained model assigns to N data samples, and the posterior predictive distribution:

$$Q_\pi(\mathbf{X}'^{1..N}|\mathbf{X}^{1..N}) = \prod_{n=0}^{N-1} Q_\pi(\mathbf{X}'^{n+1}|\mathbf{X}^{n+1}, \mathbf{X}'^{1..n}, \mathbf{X}^{1..n})$$

This suggests the *prequential* interpretation of Bayesian prediction (Dawid, 1992): the expected predictive error of a model on N samples is the sum of the expected predictive errors on each successive sample after training on the previous samples. This accumulated prediction error is termed *cumulative risk* (Haussler and Opper, 1997):

$$R_{N,Q_\pi} = \sum_{n=0}^{N-1} r_{n,Q_\pi}$$

The risk of eq. 6 can be seen as the rate of change of the cumulative risk as the amount of training data grows.

Total modularity is the total gain in predictive accuracy (i.e., decrease in cumulative risk) provided by the optimally predictive models $Q^\star(N)$ versus the unfactorized, total-partition model Q_{π_0}. Let $R_{N,Q^\star} = \sum_{n=0}^{N-1} r_{n,Q^\star_\pi(n)}$ be the cumulative risk of an agent who selects the risk-minimal model at each N. The total modularity is then:

$$T = \lim_{N\to\infty} \left(R_{N,Q_{\pi_0}} - R_{N,Q^\star}\right) \tag{9}$$

Total modularity measures the overall predictive advantage gained by using factorized models, and is not a function of a particular N. High values of total modularity indicate

[4] In general, the approximation of eq. 7 is only accurate for large N. However, it suffices for our explanatory purposes.

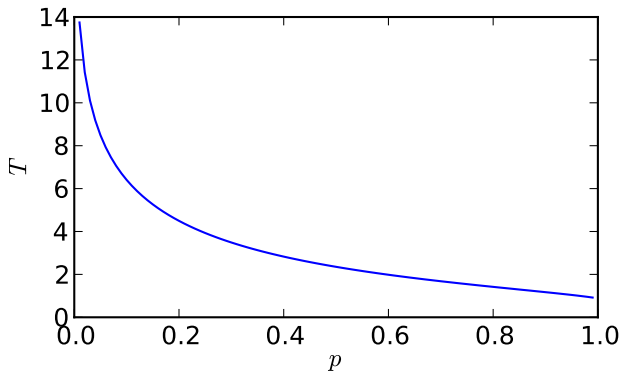


Figure 3: Total modularity of two binary variables which copy each others' state with probability p and maintain their own state with probability $1-p$. Total modularity increases as coupling decreases, and diverges as $p \to 0$.

that simpler models have significantly improved predictive performance during earlier stages of the learning process.[5] To use the previous example, the cumulative risk of the models plotted at the top of fig. 2 is shown at the bottom of that figure. The total modularity of the dynamic network shown in fig. 1 is equal to the asymptotic difference between the cumulative risks of Q_{1234} $(= Q_{\pi_0})$ and $Q^\star$. Equivalently, it is also the total area between the lines corresponding to the (non-cumulative) risks of Q_{1234} and $Q^\star$.

For another illustration of total modularity, we consider a simple dynamical system composed of two binary variables. Each variable is parameterized in the following manner: at each time step, with probability p it assumes the value of the other variable in the previous time step, and with probability $1-p$ it maintains its own value from the previous time step. The amount of dynamic coupling between the two nodes increases with p: at $p=0$ the variables have no interaction, while at $p=1$ their values are completely correlated (with a one timestep lag). This dynamic coupling is illustrated in fig. 3, which plots the total modularity of this system against the coupling parameter p. The total modularity monotonically decreases as p increases, showing that greater coupling leads to lower total modularity. As $p \to 0$, the two variables become completely independent and total modularity diverges (in this case, it grows without bound at a rate proportional to $\log N$).

State-dependent and causal modularity

The way information flows within a dynamical system can depend on the system's state. For example, a partition's stochastic interaction can be different in different attractors. We can quantify this by different choices of the starting state distribution, $P(\mathbf{X})$. Though we have generally taken $P(\mathbf{X})$ to be a fully-supported uniform distribution, it can be weighted preferentially over some subset of starting states.

[5] Minimization of accumulated error by online switching from simpler to more complex models is related to a learning framework recently proposed by van Erven et al. (2007)

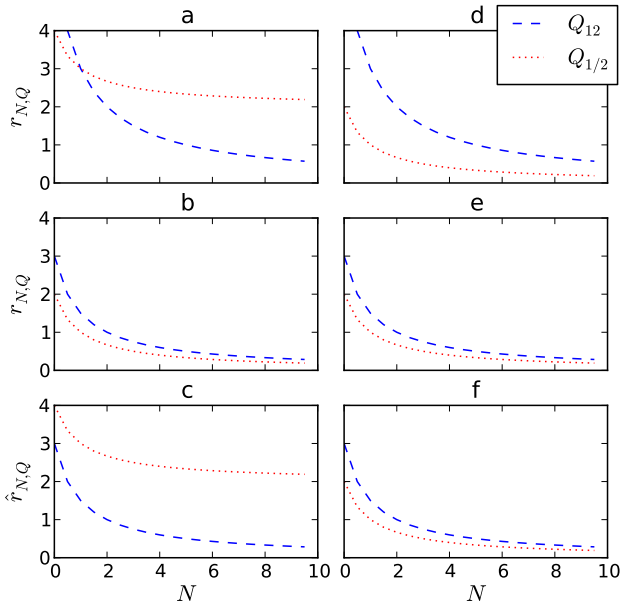


Figure 4: Risk for two systems, each having two binary variables: in system A (left column) each variable copies previous value of the other, in system B (right column) each variable takes opposite of its own previous state. a) and d): Risk under uniform starting state distribution. Lowest risk model of A becomes the total one, while factorized model remains optimal for B. b) and e): Risk and optimal decompositions depend on the starting state distribution. Computed over $P(\mathbf{X}=(0,1))=0.5, P(\mathbf{X}=(1,0))=0.5$, risk and optimal decompositions become the same for A and B, though their causal organization is different. c) and f): *Causal risk* leads to different decompositions of A and B, even when computed over same starting state distribution as in b) and e).

For example, consider two systems, each composed of two binary variables. In system A, each variable copies the previous value of the other, while in system B, each variable takes the opposite of its own previous state. Fig. 4 shows the risk plots for both A (left column) and B (right column), where 4a and 4d are calculated for a uniform starting state distribution. The risk, as well as the optimal decompositions, is different between the two systems: A (which performs the copy operation) eventually chooses the total partition $\{\{1,2\}\}$ as the most predictive, while B (whose variables perform independent state flips) never does.

If, however, a non-uniform starting state distribution is chosen, risk and optimal decompositions can change. The risk for starting state distribution $P(\mathbf{X}=(0,1)) = 0.5, P(\mathbf{X}=(1,0)) = 0.5$ are shown in fig. 4b and 4e (for systems A and B respectively). Different parts of the start-

ing state space induce different risk values and optimal decompositions: for this distribution, fig. 4b shows that the total partition $\{\{1,2\}\}$ is never chosen as the optimally predictive one for system A.

Additionally, for these starting states the transition distributions of A and B are identical: if either system is started in state $(0,1)$, it deterministically transitions to state $(1,0)$, and similarly for the transition from $(1,0)$ to $(0,1)$. Because the observed dynamics of the two systems are identical, the risk functions and optimal decompositions are also equal. Though systems A and B are defined using different causal architectures, here their modular organizations are indistinguishable. Specifically, A is postulated to have a causal connection among its variables but – for this starting state distribution – they display no stochastic interaction.

This example highlights the difference between statistical correlation and causal interaction. To properly handle the latter, we utilize a notion of causality based on semantics of intervention (Pearl, 2000), recently developed in an information-theoretic direction by Ay and Polani (2008). In Pearl's treatment, conditional probability distributions represent not only correlations, but also responses of variables to externally-imposed interventions. This is especially natural when dynamics of interest are generated by causal models, such as dynamic causal Bayesian or Boolean network models frequently used in artificial life and systems biology.

In our example, the functional organization of systems A and B can be differentiated – even within the non-uniform starting state distribution mentioned above – if the starting states of the systems can be intervened upon. This is because in system A – but not system B – changing the starting state of one variable can change the other variable's future state.

We consider interventions formally by noting that the risk r_{N,Q_π} of eq. 6 need not take the same starting state distribution for training data as for the testing data. Instead, we take the starting state distribution for training data to be drawn i.i.d. from a fully-supported and uniform distribution $\hat{P}(\mathbf{X})$ (the distribution of interventions), while the testing starting states can be drawn from any $P(\mathbf{X})$ of interest. We refer to risk evaluated under this learning scenario as *causal risk*:

$$\hat{r}_{N,Q_\pi} = \sum_{\mathbf{x},\mathbf{x}'} P(\mathbf{x})P(\mathbf{x}'|\mathbf{x})\Big[\log P(\mathbf{x}'|\mathbf{x}) - \sum_{\mathbf{x}^{1..N},\mathbf{x}'^{1..N}} \hat{P}\big(\mathbf{x}^{1..N}\big) P\big(\mathbf{x}'^{1..N}\big|\mathbf{x}^{1..N}\big) \log Q_\pi\big(\mathbf{x}'\big|\mathbf{x},\mathbf{x}'^{1..N},\mathbf{x}^{1..N}\big)\Big]$$

As $N \to \infty$, the posterior predictive distribution of model Q_π approaches $\prod_{B\in\pi} \hat{P}(\mathbf{X}'_B|\mathbf{X}_B)$, where $\hat{P}(\mathbf{X}'_B|\mathbf{X}_B)$ is the whole-system transition distribution $P(\mathbf{X}'|\mathbf{X})$ marginalized onto variables in block B using $\hat{P}(\mathbf{X})$. Then, $\hat{r}_{N,Q_\pi}$ can be approximated by:

$$\mathcal{I}_\pi(\mathbf{X}'|\mathbf{X}) + \sum_{B\in\pi} \mathrm{KL}\Big[P(\mathbf{X}'_B|\mathbf{X}_B)\Big\|\hat{P}(\mathbf{X}'_B|\mathbf{X}_B)\Big] + \frac{d_\pi}{2N}$$

where $\mathcal{I}_\pi$, d_π, and the expectations in the KL terms use the testing starting state distribution. The KL divergence between $P(\mathbf{X}'_B|\mathbf{X}_B)$ (the whole-system transition distribution marginalized onto variables in block B using $P(\mathbf{X})$) and $\hat{P}(\mathbf{X}'_B|\mathbf{X}_B)$ reflects the amount of extra perturbation that active interventions inject into block dynamics. The two distributions need not be equal, unless $P(\mathbf{X}) = \hat{P}(\mathbf{X})$ or the partition under consideration is the total one. Because KL divergence is non-negative, causal risk $\hat{r}_{N,Q_\pi}$ is not less than the statistical risk r_{N,Q_π}(compare above to eq. 7).

Fig. 4c and 4f show the causal risk for systems A and B (respectively) with $P(\mathbf{X}=(0,1))=0.5, P(\mathbf{X}=(1,0))=0.5$. In 4c – but not 4f – the total partition model assumes a lower risk than the factorized model, indicating that for the starting states in question, system A – but not system B – has causal interactions between its variables.

Conclusion

Modularity is normally treated as an objective property of a system's organization. Our approach instead considers from the perspective of modeling and prediction. In the context of inferring dynamic models from limited data, modularity allows for models that are predictive but simple, with the amount of training data controlling the trade-off. Our statistical treatment connects to previous information-theoretic approaches, but goes further by providing principled terms for identifying small modules.

Our approach can also be used to find state-dependent modular organizations, both in statistical and causal (interventional) senses: models trained on interventional dynamics but tested on arbitrary distributions give rise to a measure that identifies causal modules. This is related to existing information-theoretic measures of causal interactions between subsystems (Tononi and Sporns, 2003), but here emerges naturally from the framework of statistical modeling. This framework also produces a measure of total modularity present in the system, which quantifies the overall predictive advantage that modularity provides through the entire model inference process.

As a side note, if the learning of real-world cognitive systems (such as scientists or organisms) proceeds in a manner somewhat similar to the statistical framework presented here, our approach suggests why such systems may infer modular organizations in the external world: under conditions of limited data, this assumption can simplify learning and lead to gains in predictive power.

One important issue with our treatment is its model-dependence. The complexity penalization term of eq. 6 depends on the model class, and different model classes may have different parameterizations and functional forms. Our examples employed products of Markov chain models, a rather general dynamic model class but one heavily parameterized; others could be used. The choice of model class can be thought of as a null model of system dynamics.

Several generalizations suggest themselves. For example, it is possible to infer module timescales by searching not only over decompositions, but also model orders (numbers of previous states on which transition probabilities depend; for inferring Markov chain order, see Strelioff et al., 2007). Fuzzy modular organizations, in which a variable can belong to more than one module, can be accommodated by allowing partially-overlapping blocks. More generally, the model search space could include other structures besides partitions (e.g. trees or networks) to impose independence constraints on information flow between blocks.

Identifying modularity in dynamical systems is important in complex systems research in general, and biological systems modeling in particular. Our method differs from recent community-detection methods that find modularity in static graphs, in that it focuses on the organization of interactions between dynamic system components. In future work, we hope to apply it to the analysis of regulatory and signaling control in biochemical networks, as well as inference of functional neural organization from brain recordings.

Acknowledgments

Thanks to Randy Beer, Paul Williams, Olaf Sporns, and the participants of the *Guided Self-Organization 3* workshop for useful feedback and encouragement.

References

Ay, N. and Polani, D. (2008). Information flows in causal networks. *Advances in Complex Systems*, 11(1).

Ay, N. and Wennekers, T. (2003). Dynamical properties of strongly interacting Markov chains. *Neural Networks*, 16(10).

Balduzzi, D. and Tononi, G. (2008). Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput Biol*, 4(6).

Barrett, A., Seth, A., and Sporns, O. (2011). Practical Measures of Integrated Information for Time-Series Data. *PLoS Comput Biol*, 7(1).

Barron, A. and Hengartner, N. (1998). Information theory and superefficiency. *Ann Stat*, 26(5).

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3).

Conant, R. (1972). Detecting subsystems of a complex system. *IEEE Trans on Systems, Man, and Cybernetics*.

Dawid, A. (1992). Prequential analysis, stochastic complexity and Bayesian inference. *Bayesian statistics*, 4.

Force, A., Cresko, W., Pickett, F., Proulx, S., Amemiya, C., and Lynch, M. (2005). The origin of subfunctions and modular gene regulation. *Genetics*, 170(1).

Guimera, R., Sales-Pardo, M., and Amaral, L. (2004). Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2).

Han, J. et al. (2004). Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*.

Hartwell, L. et al. (1999). From molecular to modular cell biology. *Nature*, 402(6761).

Haussler, D. and Opper, M. (1997). Mutual information, metric entropy and cumulative relative entropy risk. *Ann Stat*, 25(6).

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5).

Kashtan, N. and Alon, U. (2005). Spontaneous evolution of modularity and network motifs. *PNAS*, 102(39).

Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika*, 83(2).

Krause, A. et al. (2003). Compartments revealed in food-web structure. *Nature*, 426(6964).

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge Univ Pr.

Rissanen, J. (2007). *Information and complexity in statistical modeling*. Springer Verlag.

Schlosser, G. and Wagner, G. (2004). *Modularity in development and evolution*. University of Chicago Press.

Simon, H. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6).

Solé, R. and Valverde, S. (2008). Spontaneous emergence of modularity in cellular networks. *J R Soc Interface*, 5(18).

Strelioff, C., Crutchfield, J., and Hübler, A. (2007). Inferring Markov chains. *Physical Review E*, 76(1).

Studeny, M. and Vejnarova, J. (1998). The multiinformation function as a tool for measuring stochastic dependence. *Learning in graphical models*, 261.

Tishby, N., Pereira, F., and Bialek, W. (1999). The information bottleneck method. In *37th Allerton Conf on Communication*.

Tononi, G., Edelman, G., and Sporns, O. (1998). Complexity and coherency: integrating information in the brain. *Trends in cognitive sciences*, 2(12).

Tononi, G. and Sporns, O. (2003). Measuring information integration. *BMC Neuroscience*, 4(1).

Tononi, G., Sporns, O., and Edelman, G. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *PNAS*, 91(11).

van Erven, T., Grünwald, P., and de Rooij, S. (2007). Catching up faster in Bayesian model selection and model averaging. *NIPS*, 20.

Wagner, G., Pavlicev, M., and Cheverud, J. (2007). The road to modularity. *Nature Reviews Genetics*, 8(12).

Winther, R. (2001). Varieties of modules: kinds, levels, origins, and behaviors. *J of Experimental Zoology*, 291(2).