Useful references


The problem of information flow

Suppose we have a countable set of random variables \( X^k_m \) where the lower index refers to space and the upper index refers to time and the spatial grid spacing is \( \Delta s \) while the temporal grid spacing is \( \Delta t \).

In general the uncertainty or entropy of a variable for a given spatial location will be time dependent but may also depend on the time evolution of the random variables at other spatial locations.

Conceptually we shall consider the information flow between two spatial locations to be the time rate of change in entropy or uncertainty at a particular location due to the influence of another spatial location. If all spatial locations evolve independently then the information flow within the system is zero. In most interesting dynamical systems however this will not be the case.

The motivation for studying this is that uncertainty created at one location may influence uncertainty at another location at a later time. This has clear implications for the problem of prediction.
Problem studied first by physicists in the 1980s. Kaneko (1986) studied the propagation of perturbations in simple non-linear dynamical systems using a "moving frame" or co-moving Lyupanov exponent:

\[ \lambda(v; x_1, x_2) = \lim_{t \to \infty} \frac{1}{t} \ln \left[ \frac{\Delta(v, x_1, x_2, t)}{\Delta(v, x_1, x_2, 0)} \right] \]

\[ \Delta(v, x_1, x_2, t) \equiv \left[ \int_{x_1 + vt}^{x_2 + vt} |\delta \psi(x, t)|^2 \, dx \right]^{\frac{1}{2}} \]

Maximization of this showed the preferred velocity of growing perturbations. Since regular Lyupanov exponents are often related to entropy production it was natural to try to find an information theoretic counterpart for co-moving exponents. This turned out empirically and in the systems studied, to be the time lagged mutual information of random variables:

\[ I(X^j_n; X^k_m) \equiv \sum_{x \in X^j_n} \sum_{y \in X^k_m} p(x, y) \ln \frac{p(x, y)}{p(x)p(y)} \]

Which has the natural velocity scale

\[ v = \Delta s(n - m) / \Delta t(j - k) \]

The lagged mutual entropy turned out to be maximized when this velocity scale matched that which maximized the co-moving Lyupanov exponent.
In addition to this match of physical propagation scales, mutual information has an appealing interpretation as the reduction in uncertainty of $X_m^k$ due to perfect knowledge of $X_n^j$ i.e. roughly speaking, the contribution of uncertainty in the former due to uncertainty in the latter. This follows from the identity

$$I(X_m^k; X_n^j) = H(X_m^k) - H(X_m^k|X_n^j)$$

This measure of information flow was further verified as physically plausible in more complex and realistic dynamical systems by Vastano and Swinney (1988). It was however shown to give misleading results in certain pathological situations by Schreiber (1990). In particular when both $X_m^k$ and $X_n^j$ are subjected to a synchronized source of uncertainty then unphysical transfers are possibly indicated by the lagged mutual information. This is somewhat analogous to over interpreting a correlation as causitive. Note that the mutual information reduces to a simple function of correlation in the case that the distributions are Gaussian.
Intuitively defined measures of transfer

Schreiber (2000) suggested a new information theoretic measure of flow which overcame the problem of lack of causality identified in his 1990 study. He considered the situation where each spatial location was a Markov process of order $q$. Thus the probability function at any particular point depends only on the values for the previous $q$ random variables. In such a situation there can, by construction, be no information flow between spatial points. He then tested the deviation from this null hypothesis using a (conditional) relative entropy functional. For an order 1 Markov process this transfer entropy is defined as

$$T(j \rightarrow i, n) \equiv \sum p(i_{n+1}, i_n, j_n) \ln \frac{p(i_{n+1}|i_n, j_n)}{p(i_{n+1}|i_n)}$$

Here the $i$ and $j$ indices refer to different spatial random variable values at various times given by the subscripts. This formula can be extended in an obvious manner to a Markov process of order $q$. For the order one case one can also show that

$$T = H(X_{i}^{n+1}|X_{i}^{n}) - H(X_{i}^{n+1}|X_{i}^{n}, X_{j}^{n})$$

Notice that from the practical computational viewpoint (to be visited shortly) the transfer entropy $(TE)$ is defined with respect to trivariate distributions whereas the lagged mutual information functional is defined with respect to bivariate distributions. The order $q$ TE is defined on distributions of dimension $q+2$. 
Consider a situation where one wishes to reduce the uncertainty of a prediction at a very particular spatial location. An example might be the prediction (or otherwise) of a very intense storm in a significant location.

How might the uncertainty of this prediction be reduced? One strategy is evidently to improve the observing system which defines the initial conditions for the prediction. There are however usually time constraints involved in such an improvement. For example one might choose to send out an aircraft to improve observations in a "critical" region. But where might such a region be located? For hurricane prediction this is usually fairly obvious but for mid-latitude storms much less so.

This has led to the concept of targeted observations and much sensitivity analysis of the complex dynamical models underlying the atmosphere has been undertaken. This has often had the limitation however that it is linear in nature and beyond a certain prediction time frame, this is a highly suspect assumption (3-4 days typically).
Information flow offers a natural way of analyzing this problem: If one is able to calculate the uncertainty flow from initial time spatial points to prediction time spatial points then it becomes clear how the uncertainty at the latter times may be improved by reducing the uncertainty at the initial time.

Moreover if we have information on how this flow connects all spatial points then we can identify how the uncertainty may be optimally reduced at the initial time in order to reduce errors in predictions at particular spatial locations of interest.

Information flow calculations make no assumptions about linearity but do require a statistically significant number of "ensemble" predictions.

The latter restriction is important practically and requires the use of functionals of a low degree of multivariateness.
Kleeman (2007) carried out this program in a (somewhat) realistic model of the atmosphere using both intuitive measures (lagged mutual information and transfer entropy) of information flow discussed above. There are plans to extend the computation to the more formal measures later once some technical issues are resolved.

A T42 primitive equation dry dynamical core model with realistic depiction of mid-latitude jets and storms was utilized. A simple multivariate Gaussian distribution was assumed for initial conditions and sample trajectories of length 10 days generated using this distribution with means drawn from a very long model run (ergodicity is assumed). The ensembles were of size 10,000 which enabled highly statistically significant (non-noisy) estimates of the information flow functionals.
Application to observing networks for prediction

Temperature to Temperature TLMI at 1 Day

Zonal Velocity TLMI (One Day)

Temperature to Temperature TE at 1 Day

Zonal Velocity TE (One Day)
Application to observing networks for prediction
Application to observing networks for prediction. Practical issues

The ensemble sizes used above are not yet practical for a real world application. The ensembles are however quasi-Gaussian meaning that analytical expressions for the relevant functionals can be easily derived and used as good approximations. Since these involve only the first two moments of the samples/ensembles much smaller ensembles are required (order 200 rather than 10,000).

In all practical cases there was little qualitative difference between the two intuitive measures except for the persistence effect. This suggests that the theoretical debate between different flow functionals may not be important for this particular application. Of course in other applications that may not be the case. This is under active current investigation.
Conclusions

- Significant theoretical progress has been made in defining information flow. More work is still required.

- It has been shown that these concepts can be applied to the practical prediction problem of targetted observing networks. Unlike all other proposed techniques the present methods do not make any assumptions of linearity. Their accuracy however does depend on two important things: A reasonable estimate of the initial condition distribution from a data assimilation methodology; Confidence that the dynamical model captures most relevant aspects of the systems evolution.

- Many other applications of this theory are potentially possible in other areas.