Information Theory and Predictability
Lecture 10: Predictability of realistic models and application to the atmosphere.

1. Limitations of ensembles

A prediction ensemble is a sample drawn from the prediction probability distribution function (*p.d.f.*). It therefore represents in some sense an estimate of this underlying function. As such it has an associated sampling error which will influence its usefulness as an estimate. As we shall see shortly such errors can become a serious issue the more complex the dynamical system becomes.

The simplest way of viewing this sample estimate of the *p.d.f.* is to partition the state space of the system under consideration and then use the sample count in each partition element as an estimate of the probability that the system at that particular time lies within the particular subregion. We can make this process quite precise (see [3]) as follows:

Let us suppose that we have a complete partitioning of state space by subsets X_i with $i = 1, ..., m$. Given such a partitioning then an ensemble implies a frequency count f_i associated with every element X_i. Hopefully many of the f_i may be of significant size, a point we come to in detail below. The ensemble size N is obviously given by $\sum_{i=1}^{m} f_i$.

Consider now the prediction *p.d.f.* p on this reduced state-space. If we integrate over each partition element X_i we obtain the coarse-grained discrete probability vector element p_i. Evidently we could estimate such a vector using the f_i. Consider now the conditional probability $P(\mathbf{f}|\mathbf{p})$ that we observe f_i given that p_i holds. It follows from elementary probability theory that

$$P(\mathbf{f}|\mathbf{p}) = \frac{1}{Z(\mathbf{f})} \prod_{i=1}^{m} p_i^{f_i}$$

$$Z(\mathbf{f}) = \frac{\prod_{i=1}^{m} (f_i)!}{(N+m-1)!}$$

Now Bayes theorem [1] gives

$$P(\mathbf{p}|\mathbf{f}) \propto P(\mathbf{f}|\mathbf{p}) P_{pr}(\mathbf{p})$$

where $P_{pr}(\mathbf{p})$ is the prior[1] probability that a particular set of p_i occurs. Without any evidence of what values p_i take it is reasonable to take this prior probability as uniform i.e. in the absence of evidence there is no reason to expect that any one set of p_i is any more likely than any other. With this assumption we obtain

$$P(\mathbf{p}|\mathbf{f}) = \frac{C}{Z(\mathbf{f})} \prod_{i=1}^{m} p_i^{f_i} \equiv \Phi_{\mathbf{f}+1}(\mathbf{p})$$

[1] In other words prior to the ensemble observation of the frequencies f_i

where Φ is the so-called Dirichlet distribution. The first moment of this distribution then gives the expected p_i given the observed f_i and may be shown easily to be given by

$$\langle p_i \rangle = \frac{f_i + 1}{N + m}$$

Now that we have calculated the distribution of the p_i we can calculate the expected information loss in assuming that $p_i = \langle p_i \rangle$. This is clearly

$$EL(\mathbf{f}) = \int \Phi_{\mathbf{f}+1}(\mathbf{p}) D(\mathbf{p}, \langle \mathbf{p} \rangle) d\mathbf{p}$$

where $D(\mathbf{p}, \langle \mathbf{p} \rangle)$ is the relative entropy of the coarse-grained $p.d.f.'s$ $\mathbf{p}$ and $\langle \mathbf{p} \rangle$. Using known analytical expression for moments of the Dirichlet distribution and the expected values of their logarithms it is possible to evaluate EL analytically with the result:

$$EL(\mathbf{f}) = \sum_{i=1}^{m} \langle p_i \rangle \left(\psi(f_i + 2) - \psi(n + m + 1) - \ln \langle p_i \rangle \right) \tag{1}$$

where ψ is the digamma function.

It may also be shown relatively straightforwardly that the expected information loss is minimized by choosing $\langle p_i \rangle$ as our estimator of the coarse-grained $p.d.f.$ p_i. The relative entropy of the coarse grained optimal prediction and climatological $p.d.f.s$ $\langle p \rangle$ and $\langle q \rangle$ may be easily evaluated:

$$D(\langle p \rangle, \langle q \rangle) = \sum_{i=1}^{m} \langle p_i \rangle \ln \left(\frac{\langle p_i \rangle}{\langle q_i \rangle} \right) \tag{2}$$

In general, in many practical contexts the loss of information due to sampling of the climatological distribution is considerably smaller than that due to the prediction ensemble sampling since it is much larger. An intuitively appealing definition then of the utility of an ensemble prediction is therefore:

$$EU(\langle p \rangle, \langle q \rangle) \equiv max \left\{ D(\langle p \rangle, \langle q \rangle) - EL(\langle p \rangle), 0.0 \right\} \tag{3}$$

i.e. one reduces the coarse grained information content by the information loss due to the prediction distribution sampling.

As a rule of thumb in practical situations one finds that ensuring that "most" of the f_i are larger than around 5 is sufficient to keep EL small relative to common values of the relative entropy. Notice that implicit in this statement is the fact that one has a choice in the degree of coarseness chosen for the partitioning of state space. Thus we see that there is a trade-off involved between the relative entropy which decreases with coarser partitions (recall Lecture 2 and 5) and the sample loss EL which increases as finer partitions are chosen. This effect is illustrated in Figure 1.

2. High dimensional systems

As we have seen in the previous section a direct estimate of the coarse grained $p.d.f.$ requires that there be sufficient sample members in most of the partition elements. This becomes a problem as the dimensionality of the desired state space increases. Suppose for the sake of argument that we require that each dimension have d partitions then the total number of partitions increases as d^n where n is the dimension of the state space. Given the derivation in the previous section regarding

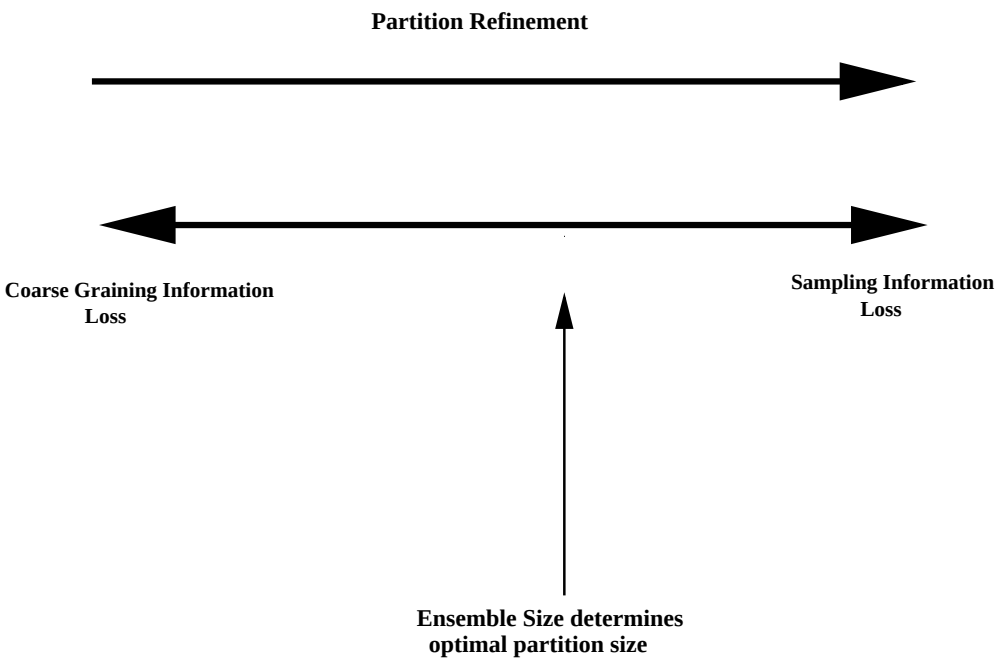

FIGURE 1. The effect on information of different partition refinements for a particular ensemble

EL it is clear that the ensemble size must be at least of this order to avoid significant sampling loss of information. Clearly as the dimension increases we rapidly require unrealistic ensemble sizes to properly sample the entire multivariate $p.d.f.$. This issue is known commonly as the "curse of dimensionality". We take two different approaches to this.

2.1. **Marginal entropy.** Although the full multivariate $p.d.f.$ may often not practically available from an ensemble it is clear that marginal distributions of sufficiently low dimension may still be accessible. This motivates the concept of *marginal* relative entropy (see again [3]) which is the average of relative entropy with respect to all possible marginal distributions of a given dimension.

Definition. Suppose we have n random variables X_i with corresponding multivariate distribution $p(X_1, X_2, \ldots, X_n)$ and all possible marginal distributions of order $m < n$ $p(X_{j_1}, X_{j_2}, \ldots, X_{j_m})$ with $j_k \leq n$ and distinct then the **marginal relative entropy** of order m is given by

$$D^m(p \parallel q) \equiv \frac{1}{C_m^n} \sum_{j_1,,j_2,\ldots,j_m} D\left(p(X_{j_1}, X_{j_2}, \ldots, X_{j_m}) \parallel q(X_{j_1}, X_{j_2}, \ldots, X_{j_m})\right)$$

The marginal entropy is thus the average relative entropy of all possible marginal distributions of order m. Marginal entropies have been used in statistical mechanics most particularly in connection with liquids where correlations between molecules are of significance.

Marginal relative entropies can be shown to satisfy an inequality hierarchy:

Theorem. *Marginal relative entropies with respect to n random variables X_i and the same partitioning of state space satisfy the following chain of inequalities*

$$D^1(p \parallel q) \leq D^2(p \parallel q) \leq \ldots \leq D^n(p \parallel q) = D(p \parallel q)$$

Proof. Use the notation $D(Y_1, Y_2, \ldots, Y_k)$ to denote the relative entropy of $p(Y_1, Y_2, \ldots, Y_k)$ and $q(Y_1, Y_2, \ldots, Y_k)$ (Note order of random variables here is immaterial). The chain rule of relative entropy shows that

$$D(Y_1, Y_2, \ldots, Y_k) \leq D(Y_1, Y_2, \ldots, Y_k, Y_{k+1})$$

which also shows that

$$D(Y_1, Y_2, \ldots, Y_k) \geq \frac{1}{k}\left\{D(Y_2, \ldots, Y_k) + D(Y_1, Y_3, \ldots, Y_k) + \ldots + D(Y_1, Y_2, \ldots, Y_{k-1})\right\}$$

If this inequality is applied term by term to the sum $C_k^n D^k(p \parallel q)$ and repetitions of the smaller order relative entropies collected we obtain

$$C_k^n D^k(p \parallel q) \geq \frac{n-k+1}{k} C_{k-1}^n D^{k-1}(p \parallel q)$$

or using the properties of C_k^n

$$D^k(p \parallel q) \geq D^{k-1}(p \parallel q)$$

□

This hierarchy has the natural interpretation that more information is apparent as the higher order multivariate behaviour of the distribution is taken into account.

2.2. **Maximum Entropy approach.** Another view of this problem is that the ensemble only provides us with limited moment information and that we should make the least assumption regarding the unknown moments. We can do this by finding the maximum entropy distribution consistent with the moments we believe are well defined by the ensemble. Since many practical ensembles have marginal distributions which are close to Gaussian and all second moments are accurately defined, it seems reasonable to assume based on this philosophy that the full distribution is also Gaussian with the appropriate first and second moments.

Of course the problem with this approach is that in some systems higher order moments may indeed be important and also well defined (at least for finite subsets of variables) by the ensemble. The problem then becomes in deciding which particular set of higher order moments may be assumed accurately estimated by the ensemble. Since in our application below such moments are apparently very seldom important we shall assume only that we have no relevant information on them from the ensemble and that so therefore a Gaussian distribution is an appropriate maximum entropy estimate.

3. Application to the atmosphere

3.1. **Equations and basic approach.** The equations governing the atmosphere are the three dimensional Navier Stokes fluid equations with gravitation and rotation. These are simplified often by assuming that in the vertical direction there is a steady balance between gravitation and vertical pressure gradient. This is called the hydrostatic approximation and the resulting fluid equations are called the primitive equations. Due to advection of fluid parcels the equations are fundamentally non-linear and turbulence is a defining characteristic of the dynamical system. In

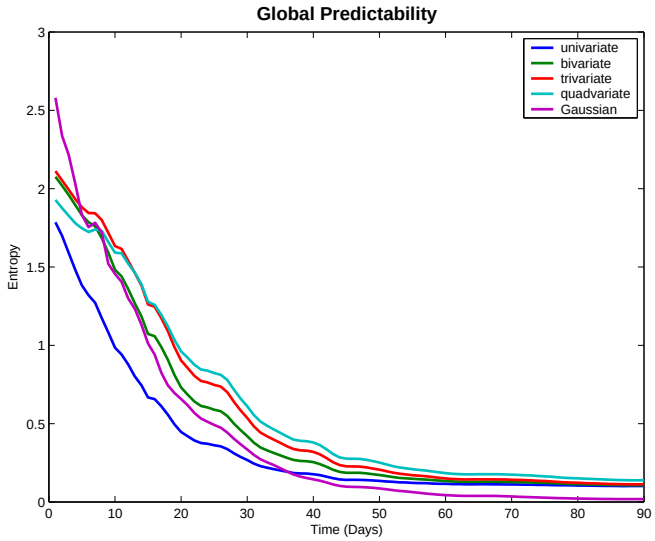

FIGURE 2. Marginal relative entropies of various orders for the atmosphere. Also displayed is the Gaussian maxent estimate divided by the reduced state space dimension.

the Earth's atmosphere the temperature gradient between equator and pole generates a large mean vertical shear called the jet stream in the region between. Such a shear generates strong turbulence whose basic element is the common weather disturbance which has a typical horizontal scale of some hundreds of kilometers. A good theoretically oriented introduction to this subject may be found in the book by Salmon [4].

As discussed in Lecture 8 the dimensionality of the atmospheric dynamical system is rather large and in fact too large to make even the approaches of marginal relative entropy above practical. In order ensure that our theoretical program is tractable we consider a subspace (often referred to as a reduced state space) which represents well most of the important variability of the system. We do this by identifying the leading principal components (or EOFs) of the equilibrium/climatological distribution. For a reasonable model resolution approximately 100 such components are sufficient to explain over 90% of all variability within the system. We therefore confine our attention to such a subspace. In addition like the simple examples of the previous Lecture we use a highly idealised Gaussian distribution to define the initial condition distribution. We insure that assumed errors there are an order of magnitude less than the equilibrium spread.

3.2. **Predictability results.** We summarize here results from [2]. The atmospheric model chosen allowed us to construct ensembles of size 10^4 which meant that marginal relative entropies up to order 4 were calculable. Such an ensemble also allowed very accurate estimates for the mean and covariance of the principal components so we were also able to calculate a maximum entropy Gaussian estimate of relative entropy as well.

Typical results are shown in Figure 2 and show a somewhat quasi-linear decline to a convergence point at around 45 days. Notice that the value never approaches

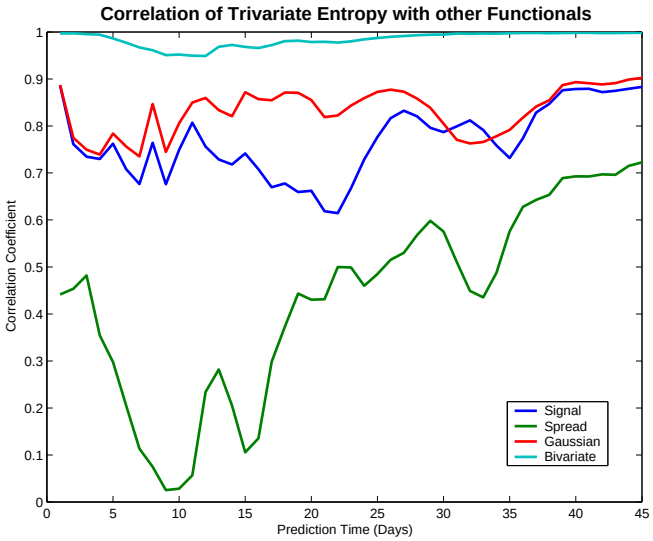

FIGURE 3. The correlation (over initial conditions) between various entropic functionals for different prediction times (see text).

zero exactly because the EL value is of order 0.15 nats. The convergence time represents the point at which initial condition information become useless to prediction. This identification of a well defined temporal cutoff point for prediction is a new result. The basic behavior shown is robust to details of the calculation i.e. it is not significantly affected by model resolution or reduced space dimension. Other features to note are the apparent convergence with order of the marginal relative entropies as well as the close resemblance of these to the Gaussian maxent curve. The latter is considerably larger (by a factor equal to the reduced space dimension) than the marginal relative entropies since it is derived from a maxent estimate of the converged probability density rather than being derived directly from a coarse grained estimate.

Just as in the simple model results considered previously we can choose a representative selection of initial conditions from the equilibrium distribution and ask the question whether the Gaussian signal and dispersion components are strongly related to variations in predictability. In the present case such an analysis appears particularly appropriate since nearly all marginal distributions are close to Gaussian. To ascertain this we took a sample of 50 initial conditions with distribution means taken from the equilibrium and correlated the two Gaussian components with the marginal relative entropy of order 3. Results are shown in Figure 3 for time points on the route to convergence of the prediction and equilibrium ensembles. Notice firstly the very high correlation between marginal entropies and the quite high correlation with the Gaussian maxent estimate. The latter is evidence of the close to Gaussian nature of most ensembles. Also notice like the stochastic oscillator example of the previous Lecture the signal component is the dominant "control" over predictability variation. Unlike that case however variations in the dispersion term are also related (albeit at a reduced level relative to signal) to the predictability.

References

[1] J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley and Sons, 1994.

[2] R. Kleeman. Limits, variability and general behaviour of statistical predictability of the mid-latitude atmosphere. *J. Atmos Sci*, 2006. Accepted with minor revisions.

[3] R. Kleeman. Statistical predictability in the atmosphere and other dynamical systems. *Physica D*, 2006. Special issue on data assimilation. In press.

[4] R. Salmon. *Lectures on Geophysical Fluid Dynamics*. Oxford Univ. Press, New York, 1998.