STATISTICALLY OPTIMAL LINEAR ESTIMATORS

Statistical models of phenomena can be developed directly from observations without appeal to detailed dynamical arguments. In fact, in most cases even a prediction based on perfect dynamical knowledge can be improved by blending with a statistical model. The primary uses of statistical models are to interpolate or extrapolate observations (prediction is time extrapolation) and to test general hypotheses that certain processes are the cause of variations of some variable. These models are most easily developed when there is an approximately linear relation between the predictor variables and the thing to be predicted. For example, we believe that coastal sea-level elevation, y, is influenced in an approximately linear way by atmospheric pressure, $x(1)$, and alongshore wind stress, $x(2)$. Thus we might try to build a linear model relating y to $x(1)$ and $x(2)$. If reasonable predictability is achieved then practical predictions can be made without appeal to numerical models and the hypothesis that pressure and wind stress are important causes of sea level variation will be supported. Notice that we have been clever in setting up the linear model by using wind stress rather than wind velocity; any known nonlinear relation can be expressed as a linear relation of nonlinear functionals.

1. Single variable estimators

To begin, let us review the procedure of predicting the fluctuations y' of a single variable from observations x' of another (by definition $\langle y' \rangle = \langle x' \rangle = 0$) using the model

$$\hat{y} \;=\; \alpha\, x' \tag{1}$$

Clearly we want the error $\epsilon \;=\; \hat{y} - y'$ to usually be as small as possible, but unless a linear response to x' is the only cause of y' variations, this error can not be made to vanish. We will a convenient and sensible measure of typical error, the Mean Square Error (MSE). Here "mean" is defined as the average over an infinite number of predictions in each of which the same model (here the same value of α) is used but x' and y' vary. This mean is equivalent to averaging over many realizations just as assumed in defining the true mean $\langle\rangle$. For the model (1)

$$\text{MSE} = \langle (\hat{y} - y')^2 \rangle \;=\; \langle (\alpha x' - y')^2 \rangle \;=\; \alpha^2 \langle x'^2 \rangle - 2\alpha \langle x'y' \rangle + \langle y'^2 \rangle \tag{2}$$

The model with minimum MSE (as a function of α) is

$$\alpha \;=\; \frac{\langle x'y' \rangle}{\langle x'^2 \rangle}, \qquad \text{MSE} = \langle y'^2 \rangle \, (1 - \rho^2), \qquad \rho = \frac{\langle x'y' \rangle}{(\langle x'^2 \rangle \langle y'^2 \rangle)^{1/2}} \tag{3}$$

Here $\langle x'y' \rangle$ is the covariance of x and y and ρ is their correlation. The error predicting y without knowledge of x is $\langle y'^2 \rangle$; the reduction of mean-square error from using the model is $\rho^2 \langle y'^2 \rangle$. The fractional error reduction is called the prediction **skill** and is equal to ρ^2. This is the most important meaning of correlation - ρ^2 is the fraction of one variable's variance which is determined by the linear relation with another.

From our model we can say that $y = \alpha x + \epsilon$ where (you can easily show) ϵ is uncorrelated with x, the **explained variance** is $\alpha^2 \langle x'^2 \rangle \;=\; \rho^2 \langle y'^2 \rangle$, and the **unexplained variance** is $\langle \epsilon^2 \rangle \;=\; [1 - \rho^2] \langle y'^2 \rangle$. Note that $\rho^2 \langle y^2 \rangle$ is both the portion of $\langle y'^2 \rangle$ explained and the variance of the estimate $\hat{y}$, itself.

In general, a linear statistical estimator depends on the statistical relation between the data variables and the relation between the data and the estimand (variable to be estimated). Above, the data are specified by $\langle x'^2 \rangle$ and and the data-estimand relation is determined by $\langle x'y' \rangle$. A general

feature of such estimators is that as the data-estimand relation becomes weaker, the estimator pays less attention to the data. This can be seen in the simplest estimation problem - estimating the variable y from noisy observations $x = y + \epsilon$. By "noise" we here mean that $\langle\epsilon\rangle = 0$ and that ϵ is unrelated to y so that $\langle y\epsilon\rangle = 0$. In this case the estimator (1-3) leads to $\alpha = \langle xy\rangle/\langle x^2\rangle = \langle y^2\rangle/(\langle y^2\rangle + \langle\epsilon^2\rangle)$. As the **signal-to-noise ratio** $\langle y^2\rangle/\langle\epsilon^2\rangle$ becomes smaller so does α; then the "optimal" estimate $\hat{y}$ for the same observation "fades" toward the value with the smallest mean square, *i.e.* $\hat{y} = 0$. The associated MSE is

$$\mathrm{MSE} = \langle\epsilon^2\rangle \frac{y^2}{\langle y^2\rangle + \langle\epsilon^2\rangle} \ . \tag{4}$$

Because of fading, this is smaller than the error $\langle\epsilon^2\rangle$ for the simplest estimate $y = x$.

One use of single variable estimators is extrapolating stationary time series $z(t)$, a simple form of prediction. Here we want to estimate $y = z(t + t_0)$ from knowledge of $x = z(t_0)$. Assuming the mean of z vanishes, the minimum MSE linear prediction is

$$\hat{z}(t + t_0) = \rho(t)z(t_0), \qquad \mathrm{MSE} = \langle z^2\rangle[1 - \rho^2(t)], \qquad \rho(t) = \frac{\langle z(0)z(t)\rangle}{\langle z^2\rangle} \tag{5}$$

Thus the time lagged correlation is a measure of the predictability of a variable from knowledge of its present value. It is instructive to contrast the minimum MSE predictor with the "persistence" forecast $z(t) = z(0)$, which corresponds to $\alpha = 1$. The MSE of persistence is (from (1-3) $2\langle z^2\rangle[1 - \rho(t)]$; at all t this error exceeds (2). At large t, where ρ has approached zero, the minimum MSE is $\langle z^2\rangle$ and is achieved by taking $\hat{z} = 0$. Persistence guesses something different than the mean $\langle z\rangle = 0$ and this results in a larger error of $2\langle z^2\rangle$. The lesson is that if you don't know anything, guess the mean value.

The example above shows why the most accurate simulation of a process is not the best predictive model! A faithful dynamical model of a stationary process would predict a time-independent variance. Unless the model is perfect, this will not be the best prediction and as predictability decreases with forecast range the dynamically faithful model becomes less and less optimal. A very simple method of improving on the predictability of a dynamical model is to use its output, x, as data for a linear minimum MSE statistical model of the form (1-3). The example of estimating y from a noisy datum shows that a statistical estimator will not degrade a perfect dynamical prediction but will improve an imperfect one. As input for such statistical improvement, the best dynamical model is the one having the greatest correlation with the predictand (not necessarily the one with the smallest error).

The seemingly paradoxical relation between the best dynamical model and the best predictive model is related to the development of irreversibility in statistical models of reversible processes such as the diffusion equation model of random walks. In that case the detailed dynamics are reversible and statistically stationary but diffusion is irreversible and has statistics with strong time dependence. For example, the area-average square of a tracer's concentration only decreases under diffusion. In essence, the particles' positions are can not be known in detail we seek the concentration which is essentially a mean description of the particles. As this mean describes less about the particles the dynamics become irreversible. Optimal linear estimators have the same property. With loss of predictability leading to irreversibility as the estimate "fades" towards zero just as the tracer field diffuses away.

On occasion the model (1-3) is called a maximum likelihood model. This follows when x' and y' are normally distributed. Then any model producing an estimate $\hat{y}$ will have an error $\epsilon = \hat{y} - y'$ which, for a given x', will be normally distributed. The estimate which minimizes the mean square

error for a normal variable is the mean y' for that x'; this is also the most probable value of y' for that x'.

2. Multi-variate estimators

Now suppose that that y depends approximately linearly on M variables $x(1),\ x(2), ... x(M)$. We might then employ the model

$$\hat{y} = \sum_{n=1}^{M} \alpha(n)\ x(n) \tag{6}$$

without making any restriction on the mean values. Even when y and the $x(n)$ are complex valued, calculation of the MSE is straightforward

$$\langle|\hat{y} - y|^2\rangle \;=\; \sum_{n=1}^{M}\sum_{m=1}^{M} \alpha(n)\langle x(n)\ x^*(m)\rangle\alpha^*(m) \tag{7}$$

$$-\sum_{n=1}[\alpha(n)\langle x(n)\ y^*\rangle + \langle y\ x^*(n)\rangle\alpha^*(n)] + \langle y^2\rangle \tag{8}$$

This is a positive quadratic form in the $\alpha(n)$ so its extremum with respect to variations of α is a minimum. Thus the optimal weights are found by differentiation to satisfy

$$\sum_{m=1}^{M}\langle x^*(n)x(m)\rangle\alpha(m) = \langle y\ x(n)^*\rangle \tag{9}$$

In terms of the data-data mean product $D(n,m) = \langle x^*(n)\ x(m)\rangle$ and the solution is

$$\alpha(n) = \sum_{m=1}^{M} D^{-1}(n,m)\langle y\ x(m)^*\rangle \quad \text{where} \quad \sum_k D^{-1}(n,k)\ D(k,m) \;=\; \delta(n,m)\ , \tag{10}$$

that is $\mathbf{D}^{-1}\mathbf{D}$ and $\mathbf{D}^{-1}$ is the inverse of $\mathbf{D}$.

The above is a **multi-variate linear estimator** and can be employed for a wide variety of purposes. Simplest is allowing for nonzero mean values in the estimator (1-3). To do this we define $x(1) = x$ and $x(2) = 1$, that is take the constant unity as a "datum". Substituting into (6-10) and expressing $x(1) = x' + x$ shows that

$$\alpha(1) = \langle x'y'\rangle/\langle x'^2\rangle, \qquad \alpha(2) = \langle y\rangle - \alpha(1)\langle x\rangle, \tag{11}$$

This estimate of y from x is identical to that made from (1-3) by adding the mean values back onto y' and x'. The "trick" of adding a constant as a predictor datum simply allows the effects of mean values to be accounted for by the same methodology used for other data.

The form of the multi-variate estimator of (6-10) is useful but perhaps more important is its measure of performance since this tells us what kind of variables, $x(n)$, lead to good estimates.

$$\text{MSE} = \langle y^2\rangle\left(1 - \sum_n\sum_m \frac{\langle y\ x(n)^*\rangle D^{-1}(n,m)\langle x(m)\ y^*\rangle}{\langle y^2\rangle}\right) \tag{12}$$

or

$$\text{MSE} = \langle y^2\rangle\left(1 - \frac{\mathbf{C}_{xy}^T\mathbf{D}^{-1}\mathbf{C}_{xy}}{\langle y^2\rangle}\right) \tag{13}$$

where $C_{xy}(n) = \langle x(n)\ y^*\rangle$. Here $C_{xy}^T\mathbf{D}^{-1}\mathbf{C}_{xy}/\langle Y^2\rangle$ is the fractional reduction of error and, in comparison with ρ^2 of (3), is called the square of the **multiple correlation** between y and the

vector $\mathbf{x}$. The multi-variate model is similar to the single variable model (1-3) in that the mean square of the estimate is also the reduction of estimation error below that for guessing $y = 0$, that is

$$\langle \hat{y}^2 \rangle = \langle y^2 \rangle - \langle (\hat{y} - y)^2 \rangle \tag{14}$$

The multiple correlation is most easily interpreted when the data variables are uncorrelated so that $\langle x^*(n)\, x(m) \rangle = D(n,m) = \delta(n,m) \langle |x(n)|^2 \rangle$ and

$$\text{MSE} = \langle y^2 \rangle \left[1 - \sum_m^M \frac{|\langle y^* x(m) \rangle|^2}{\langle x(m)^2 \rangle \langle y^2 \rangle} \right]. \tag{15}$$

In this case the squared multiple correlation is simply the sum of the squares of each variable's correlation with y. There is generally no disadvantage in working with somewhat correlated data but uncorrelated data are easier to understand. You can orthogonalize a set of data using a strictly mathematical procedure or by finding the Empirical Orthogonal Functions of the data field $\mathbf{x}$ (more on that later).

The best estimates $\hat{y}$ are obtained when the squared multiple correlation $\mathbf{C}_{xy}^T \mathbf{D}^{-1} \mathbf{C}_{xy}$, or in the case of uncorrelated data $\sum |\langle x(m) y^* \rangle|^2 / (\langle x^2(m) \rangle \langle y^2 \rangle)$, is large. Note that, in theory, adding a new parameter, $x(m)$ can never hurt the estimate. We will see in a subsequent section that when the estimator is based on measured (and hence inaccurate) statistics, adding new data variables can decrease accuracy. Note also that data that are poorly correlated with other variables are weighted differently than independent data and, for a given correlation with y are less useful than independent data. Adding a completely redundant datum does no good at all.

It is intuitive that the actual minimum MSE estimate does not depend on whether or not the data are rotated to be uncorrelated so long as the transformation is one-to-one. Consequently, we know that the multi-variate estimator separates the effect of correlated variables. In essence the procedure finds that part of $x(n)$ which is not correlated with any of the other data and finds the $\alpha(n)$ by correlating y with that part of $x(n)$.

A very important aspect of minimum MSE estimation models is that they exactly recover the true relation between y and $\mathbf{x}$ if that relation is linear, that is of the form

$$y = \mathbf{a}^T \mathbf{x} + \epsilon \qquad \text{or} \qquad y = \sum_m a(m)\, x(m) + \epsilon \tag{16}$$

where ϵ is uncorrelated with $\mathbf{x}$ (i.e. $\langle x(m)\, \epsilon \rangle = 0$). It is easy to verify that the statistical model (610) leads to a α equal to the dynamical transfer function $\mathbf{a}$ so long as the data-data matrix $\mathbf{D}$ is not singular.

If data are redundant in the sense that one could be exactly predicted from others, then the rank of $\mathbf{D}$ is less than its order and $\mathbf{D}^{-1}$ does not exist. In this case the statistical analysis can not separate which of the redundant data are really related to y, the weights α are not uniquely specified, and there are an infinite number of α's which will give the same estimate. Thus when $\mathbf{D}$ is singular the solution of (9) is not unique but all solutions give the same estimate, $\hat{y}$, and the same MSE. In the real world it is pathologic to obtain data which are exactly redundant so $\mathbf{D}$ usually has an inverse.

Recovery of the true linear relation hinges critically on the error or noise, ϵ, being uncorrelated with the data $x(n)$. This restriction is often not met in practice because the error ϵ is caused by some factor not included in the set of x's but correlated with some of them. It is violation of this restriction which leads to nonsensical statistical "models" such as baseball games cause the ocean to warm. A model with $x =$"number of baseball games" does not include the true cause of ocean

warming but would show some predictive ability because baseball is mostly played when solar heating is strong. The associated weight α would, however, have nothing to do with the dynamics of the ocean.

The substance of minimum MSE estimation is contained in (6-10) and (12-15). Subsequently we will address the practical reality of applying this formalism; the problems are in choosing appropriate data and estimating the statistics involved in (6-10) and (12-refeq4c). But it is equally important to understand how an ideal linear estimator behaves and this is most easily done by example.

3. Examples

Interpolation. Numerical analysis provides formulas for interpolating a continuous function $X(t)$ between points where its value is known; so does linear estimation. Suppose we know $x(1) = X(-\delta)$ and $x(2) = X(\delta)$ and wish to find $y = X(t)$ when it is known that X has stationary statistics with $R(t) = \langle X(0)X(t)\rangle = \langle X^2\rangle C(t)$ (note that we have not specified $\langle X\rangle = 0$ so R is not necessarily a covariance and C is not a correlation). The estimator (6-10) is obtained from

$$\alpha(1) + \alpha(2)\, C(2\delta) = C(t+\delta), \qquad \alpha(1)\, C(2\delta) + \alpha(2) = C(t-\delta) \tag{17}$$

with solution

$$\hat{X}(t) = X(\delta)\frac{C(t-\delta) - C(2\delta)C(t+\delta)}{1 - C^2(2\delta)} + X(-\delta)\frac{C(t+\delta) - C(2\delta)C(t-\delta)}{1 - C^2(2\delta)} \tag{18}$$

Note that at the data points $\hat{X}(\pm\delta) = X(\pm\delta)$, that is, the exact answer is recovered.

To see how the estimator (18) behaves, first consider the case where both t and δ are small compared with the scale of $R(t)$ so that this mean product can be expanded in a Taylor series about $t = 0$. By definition $R(0) = \langle X^2\rangle$. Differentiating, $\dot{R}(t) = \frac{d}{dt}\langle X(0)X(t)\rangle = \langle X(0)\dot{X}(t)\rangle$; evaluated at $t = 0$ this gives $\dot{R}(0) = \langle X\dot{X}\rangle = \frac{1}{2}\frac{d}{dt}\langle X^2\rangle = 0$. Noting that by stationarity $\langle X(0)\dot{X}(t)\rangle = \langle X(t)\dot{X}(0)\rangle$ gives $\ddot{R}(0) = -\langle \dot{X}^2\rangle$ so that

$$C(t) = 1 - \frac{\langle \dot{X}^2\rangle}{\langle X^2\rangle}\, t^2/2 + O(t^4). \tag{19}$$

In the case that $O(t^4)$ can be neglected (18) becomes

$$\hat{X}(t) = X(\delta)\frac{t+\delta}{2\delta} + X(-\delta)\frac{\delta - t}{2\delta} = \frac{X(\delta) + X(-\delta)}{2} + \frac{X(\delta) - X(-\delta)}{2\delta}t \tag{20}$$

which will be recognized as a straight-line fit. This would have been obtained by the usual numerical analysis procedure of Taylor series expanding $X(t)$ around $t = 0$ and fitting the first two terms to the data. When t or δ is not small, the linear estimator (18) tells how to improve on straight-line interpolation. If the mean $\langle X\rangle$ were zero and the scale of X were so short that $C(t+\delta) = C(t-\delta) = 0$ then $\hat{X} = 0$, showing again the general property that as predictability is lost the estimate of a fluctuation "fades" toward zero.

It is more interesting to note that even when the scale of $X(t)$ is so short that a Taylor series is invalid, linear estimation works fine. Suppose $C(t) = \exp(-|t|)$; this corresponds to a process with infinite $\langle \dot{X}^2\rangle = -\frac{d^2}{dt^2}\, C(t)\Big|_{t=0}$. Neither $X(t)$ nor its time-lagged correlation can be Taylor series expanded so the usual numerical analysis methods of interpolation can not be used. The minimum MSE estimate (18) is

$$\hat{X}(t) = \frac{X(\delta) + X(-\delta)}{2}\, \frac{\exp(-|t-\delta|) + \exp(-|t+\delta|)}{1 + \exp(-2\delta)} + \frac{X(\delta) - X(-\delta)}{2}\, \frac{\exp(-|t-\delta|) - \exp(-|t+\delta|)}{1 - \exp(-2\delta)} \tag{21}$$

Taylor series expansion of this for small t and δ shows that even though the series is undifferentiable, one recovers straight-line interpolation as $\delta \to 0$. The clever student may notice that when $|t| > \delta$ the prediction is based solely on the nearest data point. For example, for $t > \delta$ the estimate is $\hat{X}(t) = X(\delta)\ \exp(-[t-\delta])$ regardless of $X(-\delta)$. This peculiarity results from the particular covariance $C(t) = \exp(-|t|)$ which is the covariance of a **first order Markov process**, *i.e.* one of the form

$$\frac{d\,X(t)}{dt} = \beta X(t) + \epsilon(t) \tag{22}$$

where ϵ is a white noise process having no serial correlation and a finite frequency spectrum (hence infinite variance). Because this is a first order equation with unpredictable forcing, it is not surprising that the latest value contains all the useful information about the history of $X(t)$.

Linear Functionals. Suppose we want to estimate the integral I of a function X known at discrete points. There are two ways to approach the linear estimation: (a) the function could be estimated from (6-10) with $y = Y(t)$ and then integrated or (b) the integral could be estimated directly using (6-10) with y being the integral I. Both approaches give the same result! In either case the data $\mathbf{x}$ are the values of Y at the points it is known. Thus from (6-10) the two estimates would be

$$\hat{I}_1 = \int dt\ \hat{Y}(t) = \int dt\ \sum_n\ x(n) \left[\sum_m D^{-1}(n,m) \langle x(m) Y(t) \rangle \right] , \tag{23}$$

$$\hat{I}_2 = \sum_n x(n) \left[\sum_m D^{-1}(n,m) \langle x(m)\ I \rangle \right] . \tag{24}$$

Note that the integral over t passes through most of (23) so that

$$\langle x(m)\ I \rangle = \langle x(m)\ \int dt\ Y(t) \rangle = \int dt \langle x(m) Y(t) \rangle \tag{25}$$

so that $\hat{I}_1 = \hat{I}_2$.

This integration example is a special case of the general rule that "the best estimate of a linear functional is the linear functional of the best estimate". Specifically, if $y = L(z)$ where L is some deterministic linear operator and $\hat{y}$ is the minimum MSE estimate of y and $\hat{z}$ is the estimate of z, both based on the same data, then $\hat{y} = L(\hat{z})$. This follows directly from (6-refeq3d) since

$$L(\hat{z}) = L[\mathbf{x}^T D^{-1} \langle \mathbf{x} z \rangle] = \mathbf{x}^T \mathbf{D}^{-1} L[\langle \mathbf{x} z \rangle] = \ \mathbf{x}^T \mathbf{D}^{-1} \langle \mathbf{x} L(z) \rangle = \ \mathbf{x}^T \mathbf{D}^{-1} \langle \mathbf{x} y \rangle = \hat{y} \tag{26}$$

For this to hold we need only be able to pass the operator L through the average $\langle \cdot \rangle$; this is possible whenever L is linear and deterministic (the same in every realization over which $\langle\rangle$ is defined).

Now consider estimating $y = \int_{-T}^{T} dt\ X(t)$ when $\langle X(t_1) X(t_2) \rangle = \exp(-|t_1 - t_2|)$. First suppose the only datum is $x(1) = X(t_0)$. Note that $\langle x(1)\ y \rangle = \int_{-T}^{T} dt\ \exp(-|t - t_0|) = 2 - \exp(-T - t_0) - \exp(-T + t_0)$ is maximized when $t_0 = 0$. Thus the best single datum is the mid-interval value $X(0)$. As $T \to 0$ the estimate based on $X(0)$ approaches $2T \cdot X(0)$ but if $T > 1$ this "finite difference" formula is an overestimate and the optimal α is less than $2T$.

Next suppose $x(1) = X(-\delta),\ x(2) = X(0),\ x(3) = X(\delta)$ for the same problem. By symmetry $\alpha(1) = \alpha(3)$ which simplifies the requisite arithmetic and gives

$$\alpha(1) = 2\frac{e^{\delta} - 1}{e^{\delta} - e^{-\delta}} - e^{\delta - T} \qquad \alpha(2) = 2\frac{e^{\delta} - 2 + e^{-\delta}}{e^{\delta} - e^{-\delta}} \tag{27}$$

The surprising feature here is that in the limit $T \to \infty$ the weight given to $x(2)$ becomes smaller than $\alpha(1) = \alpha(3)$ even though of the data taken one at a time $x(2)$ is best related to y. This

underscores that there are two things that make a datum useful: (a) it must be well correlated with the quantity to be estimated and (b) it should be poorly correlated with the other data, that is nonredundant. Here $x(2)$ is quite predictable from $x(1)$ and $x(3)$, much more predictable, for example, than $x(1)$ is from $x(2)$ and $x(3)$. Consequently, $x(2)$ tells less new about y than does $x(1)$ or $x(3)$.

4. Problems

(1) Use the multi-variate formalism to find the linear estimator of y from x accounting for mean values. Show this is what would be obtained directly from (1-3).

(2) Work out the $M = 2$ case of the estimator (6-10) and find its skill.

(3) For problem 2 take the correlations of y with $x(1)$ and of y and $x(2)$ to be fixed, and find how the estimator skill changes as the correlation of $x(1)$ and $x(2)$ changes.

(4) Verify that the estimators based on the data $\mathbf{x}$ and $\mathbf{z} = \mathbf{R} \cdot \mathbf{x}$ are in fact the same. Assume $\mathbf{R}$ has an inverse.

(5) A numerical model of tides will involve the large scale flow field U but not the small scale components u. In that model the bottom drag $\tau = C_D(U + u)\,|U + u|$ must be parameterized in terms of U. Consider the approximation $\hat{\tau} = \gamma(U) \cdot U$ in one dimensional flow with $u = 0$. (a) Find equations determining the optimal functions $\gamma(U)$ for the "beauty principle" of minimizing the mean square error in tidal dissipation τU. (b) Find γ if u is normally distributed with known variance. (c) Compare the mean square dissipation error using the answer (b) with that using the dumb modeler's choice $\hat{\tau} = C_D U\,|U|$.