CHAPTER 20

TREND AND SEASONALITY IN TIME SERIES

> "· · · both *London* and *Dublin* by reason of the great and casual Accession of *Strangers* who die therein, rendred them incapable of being Standards for this purpose; which requires, if it were possible, that the People we treat of should not at all be changed, but die where they were born, without any Adentitious Increase from Abroad, or Decay by Migration elsewhere."
>
> - - - Edmund Halley (1693)

The observed time series generated by the real world seldom appear to be "stationary" but exhibit more complicated behavior. In most series, particularly demographic or economic data, trend is the most common form of nonstationarity. Many economic time series are so dominated by trend (due, for example, to steadily rising population or inflation) that any attempt to detect other regularities like cyclical fluctuations or settling back after response to a shock, can be more misleading than helpful until we have a safe way of dealing with trend.

The problem has been with us from the very beginning, as our opening quotation shows. In that work Edmund Halley compiled the first tables of mortality; but he perceived that the data on births and deaths from London and Dublin were so dominated by trend (both cities were growing rapidly) that the information he needed could not be extracted from them. Instead he used data from the city of Breslau in Silesia (today called Wroclaw, in what is now Poland) because the people there were more meticulous in record keeping and less inclined to migrate.

Likewise, many time series are so dominated by cyclic fluctuations (seasonal effects in economic date, hum in electrical circuits) that it frustrates the attempt to extract an underlying "signal" such as a long-term trend from a short run of data. In the present Chapter we examine what probability theory has to say about the similar (logically, almost identical) problems of extracting the information one wants in spite of such contaminations.

Previous Methods

The traditional procedures do not apply probability theory to this problem; and indeed, do not even recognize the possibility that probability theory might be applied. Instead, one resorts to the same kind of intuitive *ad hockeries* that we have noted so often before. The usual ones are called "detrending" and "seasonal adjustment" in the economic literature, "filtering" in the electrical engineering literature. Like all such *ad hockeries* not derived from first principles, they capture enough of the truth to be usable in many problems, but they are less than optimal in most and dangerously misleading in some.

The almost universal detrending procedure in economics is to suppose the data (or the logarithm of the data) to be $y(t) = x(t) + Ct + e(t)$, composed additively of a linear "trend" Ct, a random "error" or "noise" $e(t)$, and the component of interest $x(t)$. We estimate the trend component, subtract it from the data, and proceed to analyze the resulting "detrended data" for other effects. However, many writers have noted that conventional detrending may introduce spurious artifacts that distort the evidence for other effects, and render suspect some of the conclusions that one tries to draw from the data. Detrending may even destroy the relevance of the data for our purposes.

Merely to recognize the unsatisfactory nature of this procedure does not in itself suggest an alternative that would be any better; and nothing better is to be found in the orthodox literature.

To find it we need a deeper theoretical analysis. Now very fundamental theorems indicate that Bayesian methods are the optimal way of dealing with any such problems of inference. Indeed, it may be that the Bayesian method of dealing with trend may prove to be the most important contribution of this work to practical econometrics.

Likewise, the traditional way of dealing with seasonal effects is to produce "seasonally adjusted" data, in which one subtracts an estimate of the seasonal component from the true data, then tries to analyze the adjusted data for other effects. Indeed, most of the economic time series data one can obtain have been rendered nearly useless because they have been seasonally adjusted in an irreversible way that has destroyed information which probability theory could have extracted from the raw data. We think it imperative that this be recognized, and that researchers be able to obtain the true, unmutilated data.

Electrical engineers would think instead in terms of fourier analysis and resort to "high–pass filters" and "band–rejection filters" to deal with trend and seasonality. Again, the philosophy is to produce a new time series (the output of the filter) which represents in some sense an estimate of what the real series would be if the contaminating influence were absent. Then choice of the "best" physically realizable filter is a difficult and basically indeterminate problem.

The Bayesian procedure (direct application of probability theory) leads us to an entirely different philosophy in that we do not seek to remove the trend or seasonal component from the data; that is fundamentally impossible because there is no way to know the "true" trend or seasonal term, and any assumption about them is almost certain to inject false information into the detrended, seasonally adjusted, or filtered series. Rather, we seek to remove the effect of trend or seasonality from our final conclusions, while leaving the actual data intact. We develop the Bayesian procedure for this and compare it in detail to the conventional one.

The Bayesian Procedure

First, we analyze the simplest possible nontrivial model, which can be solved completely and will enable us to understand the exact relation between the two procedures. Having this understanding, the generalization to the most complicated multivariate case will be straightforward, with no surprises.

Suppose the model consists of only a single sinusoid and a linear trend: $y(t) = A \sin \omega t + Bt + e(t)$ where A is the amplitude of interest to be estimated, and B is the unknown trend rate. If the data are monthly economic data and the sinusoid represents a seasonal effect, then ω will be $2\pi/12 = 0.524$. But, for example, if we are trying to detect a cycle with a period of twenty years, ω will be $.524/20 =, 00262$. Estimation of an unknown ω from such data is the very important problem of spectrum analysis, considered in Chapter 21; for the present we suppose ω known. Writing for brevity $s(t) \equiv sin(\omega t)$, our model equation is then:

$$y(t) = A\, s(t) + B\, t + e(t) \tag{20–1}$$

and the available data $D \equiv (y_1, \cdots, y_N)$ are values of this sampled at equal time intervals $t = 1, 2, \cdots, N$. Assigning the noise an *iid* gaussian prior probability density function $e_t \sim N(0, \sigma)$, the sampling *pdf* for the data is

$$p(y|A, B, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y_t - A\, s_t - B\, t)^2\right] \tag{20–2}$$

and as in any gaussian calculation, the first task is to rearrange the quadratic form

$$\begin{aligned} Q(A,B) &\equiv \sum_t (y_t - A\, s_t - B\, t)^2 \\ &= N\left[\overline{y^2} + A^2\,\overline{s^2} + B^2\overline{t^2} - 2A\overline{sy} - 2B\,\overline{ty} + 2AB\,\overline{ts}\right] \end{aligned} \tag{20–3}$$

where

$$\overline{y^2} \equiv \frac{1}{N}\sum_{t=1}^{N} y_t^2\,, \qquad\qquad \overline{sy} \equiv \frac{1}{N}\sum_{t=1}^{N} s_t y_t\,, \tag{20–4}$$

etc. denote averages over the data sample. Three of these averages, $(\overline{s^2},\ \overline{t^2},\ \overline{ts})$ are determined by the "design of the experiment" and can be known before one has the data. In fact, we have nearly

$$\overline{s^2} \simeq \frac{1}{2}\,, \qquad\qquad \overline{t^2} \simeq \frac{1}{3}N^2 \tag{20–5}$$

with errors of relative order $O(1/N)$, while $\overline{ts}$ is highly variable. It is certainly less than $N/2$, since that could be achieved only if $s(t) = 1$ at every sampling point. Generally, $\overline{ts}$ is much less than this, of the order $\overline{ts} \simeq 1/\omega$ due to near cancellation of positive and negative terms.

The other three averages $(\overline{y^2},\ \overline{sy},\ \overline{ty})$ depend on the data and are the "sufficient statistics" for our problem, to be calculated as soon as one has the data.

Suppose that it is the seasonal amplitude A that we wish to estimate, while the trend rate B is the nuisance parameter that makes the problem complicated. We want to make its effects disappear, as far as is possible. We shall do this by finding the joint posterior *pdf* for A and B

$$p(A,B|DI) \tag{20–6}$$

and integrating out B to get the marginal posterior *pdf* for A

$$p(A|DI) = \int p(A,B|DI)\,dB \tag{20–7}$$

This is the quantity that tells us everything the data D and prior information I have to say about A, whatever the value of B. Conversely, if we wanted to estimate B, then A would be the nuisance parameter, and we would integrate it out of (20—6) to get the marginal posterior $p(B|DI)$.

In the limit of diffuse priors for A and B (*i.e.*, their prior *pdf*'s do not vary appreciably over the region of high likelihood), the appropriate integration formula for (20—7) is

$$\int_{-\infty}^{\infty} \exp\left[-\frac{Q(A,B)}{2\sigma^2}\right] dB = (const.) \times \exp\left\{\frac{n}{2\sigma^2}\,(\overline{t^2}\,\overline{s^2} - \overline{ts}^2)\,(A-\hat{A})^2\right\} \tag{20–8}$$

where

$$\hat{A} \equiv \frac{\overline{t^2}\ \overline{sy}^2 - \overline{ts}^2\ \overline{ty}^2}{\overline{t^2}\ \overline{s^2} - \overline{ts}^2} \tag{20–9}$$

and the $(const.)$ is independent of A. Thus the marginal posterior *pdf* for A is proportional to (20—8), and the Bayesian estimate of A, regardless of the value of B, is

$$(A)_{est} = \hat{A} \pm \sigma\sqrt{\frac{\overline{t^2}}{N(\overline{t^2}\,\overline{s^2} - \overline{ts}^2)}} \tag{20–10}$$

However, orthodox writers have railed against this process of integrating out nuisance parameters – in spite of the fact that it is uniquely determined by the rules of probability theory as the correct procedure – on the ideological grounds that the probability of a parameter is meaningless because the parameters are not 'random' variables; and even worse, in the integration we introduced a prior that they consider arbitrary. But, independently of all such philosophical hangups, we can examine how the Bayesian and orthodox procedures are related mathematically.

How are they related?

The integration of a nuisance parameter may be related to the detrending procedure as follows. The joint posterior *pdf* may be factored into marginal and conditional *pdf*'s in two different ways:

$$p(A, B, |DI) = p(A|DI)\, p(B|A, DI) \tag{20–11}$$

or equally well,

$$p(A, B|DI) = p(A|BDI)\, p(B|DI) \tag{20–12}$$

From (20—11) we see that (20—7) follows at once. From (20—12) we see that (20—89) can be written as

$$p(A|DI) = \int p(A|BDI)\, p(B|DI)\, dB \tag{20–13}$$

Thus the marginal *pdf* for A is a weighted average of the conditional *pdf*'s with B known:

$$p(A|BDI) \tag{20–14}$$

But if B is known, then (20—14), in its dependence on A, is just (20—2) with B held fixed. This is, from (20—3),

$$P(A|BDI) \propto \exp\left[\frac{N\overline{s^2}}{2\sigma^2}\,(A - A^*)^2\right] \tag{20–15}$$

where

$$A^* \equiv \frac{\overline{sy} - B\overline{ts}}{\overline{s^2}} \tag{20–16}$$

But this just the estimate that one would make by ordinary least squares (OLS) fitting of $As(t)$ to the detrended data $y(t)_{det} \equiv y(t) - Bt$

$$A^* = \frac{\overline{sy}_{det}}{\overline{s^2}} \tag{20–17}$$

That is, A^* is the estimate the orthodoxian would make if he estimated the trend rate to be B. If his estimate was exactly correct, then he would indeed find the best estimate possible; but any error in his estimate of the trend rate will bias his estimate of A.

The Bayesian estimate of A obtained from (20—13/ does not assume any particular trend rate B; it is a weighted average over all possible values that the trend rate might have, weighted according to their respective probabilities. Thus if the trend rate is very well determined by the data, so that the probability $p(B|DI)$ in (20—26) has a very sharp peak, then the Bayesian and orthodoxian will be in essential agreement on the estimate of A. If the trend rate is not well determined by the data, then the Bayes estimate is a more cautious, conservative one that "hedges its bets" by taking into account all possible values of trend rate.

But while an orthodoxian would presumably accept what we have done as mathematically correct, this argument would not convince him of the superiority of the Bayesian estimate, because he judges estimates by a different criterion. It is the sampling distribution for the estimate that is, for him, all-important. So let us investigate this.

Comparison of Bayesian and Conventional Estimates

Having found a Bayesian estimator, which theorems demonstrate to be optimal by the Bayesian criterion of performance, nothing prevents us from examining its performance from the "orthodox" sampling theory viewpoint and comparing it with orthodox estimates. Then let $\tilde{A}$ and $\tilde{B}$ be the unknown true values of the parameters, and let us describe the situation as it would appear to one who already knew $\tilde{A}$ and $\tilde{B}$, but not what data we have found. As he would know, but unknown to us, our data vector will in fact be

$$y_t = \tilde{A}\, s_t + \tilde{B}\, t + e_t \tag{20–18}$$

and we shall calculate the statistic

$$\overline{sy} = \tilde{A}\, \overline{s^2} + \tilde{B}\, \overline{ts} + \overline{es} \tag{20–19}$$

in which the first two terms are fixed (i.e. independent of the noise) and only the last varies with different noise samples.

Similarly, he knows what is unknown to us; that we shall find the statistic

$$\overline{ty} = \tilde{A}\, \overline{ts} + \tilde{B}\, \overline{t^2} + \overline{et} \tag{20–20}$$

Substituting (2) and (3) into (3) we find that $\tilde{B}$ cancels out and the Bayes estimate reduces to

$$(A)_{Bayes} = \tilde{A} + \frac{\overline{t^2}\ \overline{es} - \overline{ts}\ \overline{et}}{\overline{t^2}\ \overline{s^2} - \overline{ts}^2} \tag{20–21}$$

which is exactly independent of the true trend rate $\tilde{B}$. Therefore the Bayesian estimate does indeed eliminate the effect of trend; one could hardly hope to do so more completely than that.

On the other hand, if one uses the conventional OLS estimator ( ) with detrended data $[y_t - \hat{B}t]$ based on any estimate $\hat{B}$, he will find instead

$$(A)_{orthodox} = \tilde{A} + \frac{\overline{es}}{\overline{s^2}} + \left[\tilde{B} - (\hat{A})\right] \frac{\overline{ts}}{;\,\overline{s^2}} \tag{20–22}$$

and any error in estimation of the trend contributes an error in the estimate of the seasonal. But if one uses the OLS estimate of the trend,

$$\hat{B} = \frac{\overline{ty}}{\overline{t^2}}$$

we find

$$\begin{aligned} (A)_{orth} &= \tilde{A} + \frac{\overline{es}\ \overline{s^2}}{-} \frac{\overline{ts}^2\ \tilde{A} + \overline{ts}\,\overline{et}}{\overline{t^2}\ \overline{s^2}} \\ &= (1 - r^2)\tilde{A} + \frac{\overline{t^2}\ \overline{s} - \overline{ts}\ \overline{et}}{\overline{t^2}\ \overline{s^2}} \end{aligned} \tag{20–23}$$

where

$$r \equiv \frac{\overline{ts}}{\sqrt{\overline{t^2}\,\overline{s^2}}} \tag{20–24}$$

is the sample correlation coefficient of t and $s(t)$. Thus (6) is also exactly independent of the true trend rate $\tilde{B}$; but orthodox teaching would hold that the estimator (6) has a negative bias.

But in further comparison of (4) and (6) we see that in fact

$$(A)_{orth} = (1 - r^2)(A)_{Bayes} \tag{20–25}$$

and so if the othodoxian corrected the bias simply by multiplying the detrended estimator (6) by $(1 - r^2)$ he would be led to exactly the Bayes estimate.

However, having recognized what he would consider a shortcoming of (6) and perceiving that the Bayesian result (4) has at least the merit (from his viewpoint) of being unbiased, it does not follow that the Bayesian solution is the best possible one. It is far from clear that the optimal estimator can be found merely by multiplying the OLS estimate by a constant. Indeed, one who has absorbed a strong anti-Bayesian indoctrination would, we suspect, reject any such suggestion and would say that we should be able to correct the defects of (6) by a little more careful thinking about the problem from the orthodox viewpoint. Let us try.

An Improved Orthodox Estimate

Orthodox reasoning runs about as follows. If one had in mind only the seasonal term and was not aware of trend, one would be led to estimate the cyclic amplitude as

$$\hat{A}^{(0)} = \frac{\overline{sy}}{\overline{s^2}}, \tag{20–26}$$

the conventional regression solution. Many different lines of reasoning, including Ordinary Least Squares (OLS) fitting of the data, lead us to this result.

But then one realizes that (20—26) is not a very good estimate because it ignores the disturbing effect of trend. A better seasonal estimate could be made from the detrended data

$$(y_t)_{det} \equiv y_t - \hat{B}\,t \tag{20–27}$$

where $\hat{B}$ is an estimate of the trend rate, and it seems natural to estimate it by the conventional regression rule

$$\hat{B}^{(0)} = \frac{\overline{ty}}{\overline{t^2}} \tag{20–28}$$

from OLS fitting of a straight line to the data. Using the detrended data (20—27) in (20—26) yields the corrected cyclic amplitude estimate

$$\hat{A}^{(1)} = \frac{\overline{sy} - \overline{ts}\,\hat{B}^{(0)}}{\overline{s^2}} \tag{20–29}$$

or

$$\hat{A}^{(1)} = \frac{\overline{t^2}\,\overline{sy} - \overline{ts}\,\overline{ty}}{\overline{t^2}\,\overline{s^2}} \tag{20–30}$$

which is the conventional orthodox result for the problem.

But now we see that this is not the end of the story; for A and B enter into the model on just the same footing. If it is true that we should estimate the cyclic amplitude A from detrended data (16-30), surely it is equally true that we should estimate the trend rate B from the decyclized, or "seasonally adjusted" data:

$$y_t - \hat{A}^{(0)}\, s_t \tag{20–31}$$

Thus a better estimate of trend than (20—28) would be

$$\hat{B}^{(1)} = \frac{\overline{ty} - \overline{ts}\ \hat{A}^{(0)}}{\overline{t^2}} \tag{20–32}$$

or with the OLS estimate (16–29),

$$\hat{B}^{(1)} = \frac{\overline{s^2}\ \overline{ty} - \overline{ts}\ \overline{sy}}{\overline{t^2}\ \overline{s^2}} \tag{20–33}$$

But now, with this better estimate of trend, we can get a better estimate of the seasonal than (20—29) by using (20—33):

$$\hat{A}^{(2)} = \frac{\overline{sy} - \overline{ts}\, \hat{B}^{(1)}}{\overline{s^2}} \tag{20–34}$$

But this improved estimate of the seasonal amplitude will in turn enable us to get a still better estimate of trend $\hat{B}^{(2)} \cdots$; and so on forever!

Therefore, the reasoning underlying the conventional detrending procedure, if applied consistently, does not stop at the conventional result (20—30). It leads us into an infinite sequence of back-and-forth revisions of our estimates, each set $[\hat{A}^{(n)}, \hat{B}^{(n)}]$ better than the last $[\hat{A}^{(n-1)}, \hat{B}^{(n-1)}]$.

Then does this infinite sequence converge to a final "best of all" set of estimates $[\hat{A}^{(\infty)}\hat{B}^{(\infty)}]$? If so, this is surely the optimal way of dealing with a nuisance parameter from the orthodox viewpoint. But can we calculate these final optimal estimates directly without going through the infinite sequence of updatings?

To answer this define the (2×1) vector of n'th order estimates:

$$V_n \equiv \begin{pmatrix} \hat{A}^{(n)} \\ \hat{B}^{(n)} \end{pmatrix} \tag{20–35}$$

Then the general recursion relation is, as we see from (20—29), (20—89), (20—34),

$$V_{n+1} = V_0 + M\, V_n \tag{20–36}$$

where the matrix M is

$$M = \begin{pmatrix} 0 & -\overline{ts}/\overline{s^2} \\ \overline{ts}/\overline{t^2} & 0 \end{pmatrix} \tag{20–37}$$

The solution of (19) is

$$V_n = (1 + M + M^2 + \cdots + M^n)\, V_0 \tag{20–38}$$

and since, by Schwartz inequality, the eigenvalues of M are less than unity, this infinite series sums to

$$V_\infty = (I - M)^{-1} V_0 \tag{20–39}$$

Now we find readily that

$$(I - M)^{-1} = \frac{1}{\overline{t^2}\,\overline{s^2} - \overline{ts}^2} \begin{pmatrix} \overline{t^2}\,\overline{s^2} & -\overline{t^2}\,\overline{ts} \\ -\overline{s^2}\,\overline{ts} & \overline{t^2}\,\overline{s^2} \end{pmatrix} \tag{20–40}$$

and so our final best of all estimate is

$$\begin{aligned} \hat{A}^{(\infty)} &= \overline{t^2}\,\overline{s^2}\,\hat{A}^{(0)} - \overline{t^2}\,\overline{ts}\,\hat{B}^{(0)} \overline{t^2}\,\overline{s^2} - \overline{ts}^2 \\ &= \frac{\overline{t^2}\,\overline{sy} - \overline{ts}\,\overline{ty}}{\overline{t^2}\,\overline{s^2} - \overline{ts}^2} \end{aligned} \tag{20–41}$$

But this is precisely the Bayesian estimate that we calculated far more easily in (20—10). Likewise, the final best estimate of trend rate is

$$\hat{B}^{(\infty)} = \frac{\overline{s^2}\,\overline{ty} - \overline{ty}\,\overline{sy}}{\overline{t^2}\,\overline{s^2} - \overline{ts}^2} \tag{20–42}$$

which is just the Bayesian estimate that we get by integrating out A as a nuisance parameter from (20—6).

This is another example of what we found before (Chapter 13); if the orthodoxian will think his estimation problems through to the end, he will find himself obliged to use the mathematical form of the Bayesian solution, even if his ideology still leads him to reject the Bayesian rationale for it; this mathematical form is required by elementary requirements of rationality and consistency, quite independently of all philosophical stances.

Now we see the relation between the orthodox and Bayesian procedures in an entirely different light. The procedure of integrating out a nuisance parameter sums an infinite series of mutual updatings for us, and does it in such a simple, unobtrusive way that to the best of our knowledge, no orthodox writer has yet noticed that this is what is happening. What we have just found will generalize effortlessly to far more complex problems.

As we noted before (Jaynes, 1976) in many other cases, it is a common phenomenon that orthodox results, when improved to the maximum possible extent, become mathematically equivalent to the results that Bayesian methods give us far more easily. Indeed, it is one of the problems we have that Bayesian and Maximum Entropy methods are so slick and efficient that orthodoxians, unaccustomed to getting results so easily, accuse us of claiming to get something for nothing.

Thus in the long run, attempts to evade the use of Bayes' theorem do not lead to different final results; they only make us work harder to get them. So much harder that many important Bayesian results – even some that were given already by Jeffreys (1939) – are still unknown in the orthodox literature.

The Orthodox Criterion of Performance

In our endeavor to understand this situation fully, let us examine it from a different viewpoint. According to orthodox theory, the accuracy of an estimation procedure is to be judged by the sampling distribution of the estimator, while in Bayesian theory it should be judged from the posterior *pdf* for the parameter. Let us compare these. For the orthodox analysis, note that in both (4) and (6) the terms containing the noise vector e combine to make a linear combination of the form

$$\overline{ge} \equiv \frac{1}{N} \sum_{t=1}^{N} g_t e_t \tag{20–43}$$

Then over the sampling *pdf* for the noise we have

$$E(\overline{ge}) = \frac{1}{N} \sum_t g_t E(e_t) = 0 \tag{20–44}$$

$$E[(\overline{ge})^2] = \frac{1}{N} \sum g_t g_{t'} E(e_t e_{t'}) = \overline{g^2}\, \sigma^2 \tag{20–45}$$

since $E(e_t e_{t'}) = \sigma^2 \delta(t, t')$. Thus, the sampling *pdf* would estimate this error term by (mean $\pm$ standard deviation):

$$(\overline{ge})_{est} = 0 \pm \sigma \sqrt{\overline{g^2}} \tag{20–46}$$

For the Bayes estimator (4)

$$g_t = \frac{\overline{t^2}\, s_t - \overline{ts}\, t}{\overline{t^2}\; \overline{s^2} - \overline{ts}^2} \tag{20–47}$$

and after some algebra we find

$$\overline{g^2} = \frac{\overline{t^2}(\overline{t^2 s^2} - \overline{ts}^2)}{\overline{s^2}\,(1 - r^2)} \tag{20–48}$$

where r is the correlation coefficient defined before. Thus the sampling distribution for the Bayes estimator (4) has mean $\pm$ standard deviation of

$$\tilde{A} \pm \sigma \sqrt{\hat{N} s^2 (1 - r^2)} \tag{20–49}$$

while for the orthodox estimator this is

$$(1 - r^2)\tilde{A} \pm \sigma \sqrt{\frac{1 - r^2}{N \overline{s^2}}} \tag{20–50}$$

The General Case

Having shown the nature of the Bayesian results from several different viewpoints, we now generalize them to a fairly wide class of useful problems. We assume that the data are not necessarily uniformly spaced in time, that the noise probability distribution, although Gaussian, is not necessarily stationary or white (uncorrelated) and that the prior probabilities for the parameters are not necessarily independent. It turns out that the computer programs to take all this into account are not appreciably more difficult to write, if the most general analytical formulas are in view when we write them.

So now we have the model

$$y_t = T(t) + F(t) + e_t \tag{20–51}$$

where $T(t)$ is the trend function, $f(t)$ is the seasonal function and e_t is the irregular component. We define

$$T(t) = \sum \gamma_k \Phi_k(t) \tag{20–52}$$

$$f(t) = \sum [A_k C(kt) + B_k S(kt)] \tag{20–53}$$

The joint likelihood of all the parameters is

$$L(\gamma, A, B, \sigma) = p(y_1, \cdots, y_N | \gamma A B \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left\{\frac{1}{2\sigma^2} \sum_{t=1}^{N} [y(t) - T(t) - f(t)]^2\right\} \tag{20–54}$$

The quadratic form is

$$Q(\alpha_k, \gamma_j) \equiv \sum_{t=1}^{N} \left[y_t - \sum_{j=1}^{r} \gamma_j T_j - \sum_{k=1}^{m} \alpha_k G_k(t) \right]^2 \tag{20–55}$$

where, in the seasonal adjustment problem, $m = 12$ and

$$\{\alpha_1, \cdots, \alpha_m\} = \{A_0, A_1, \cdots, A_6, B_1, B_2, \cdots, B_5\} \tag{20–56}$$

Likewise,

$$G_K(t) = \begin{cases} cos\, k\omega t, & \text{for } 0 \le k \le 6; \\ sin\,(k-6)\omega t, & \text{for } 7 \le k \le 12 \end{cases} \tag{20–57}$$

But if we combine α, γ into a single vector of dimension $n = m + r$:

$$q \equiv (\, \alpha \, \gamma \,) \tag{20–58}$$

$$F_k(t) = \begin{cases} G_k(t), & \text{for } 1 \le k \le m; \\ T_k(t), & \text{for } m+1 \le k \le n \end{cases} \tag{20–59}$$

The model is then in the form

$$y(t) = \sum_{j=1}^{m} q_j F_j(t) + e(t) \tag{20–60}$$

The data vector is

$$y_i = \sum_{j=1}^{m} q_j F_j(t_i) + e(t_i) \tag{20–61}$$

or

$$y = Fq + e \tag{20–62}$$

where

$$F_{ij} \equiv F_j(t_i) \qquad 1 \le j \le n \qquad 1 \le i \le N \tag{20–63}$$

The "noise" values $e_t = e(t_i)$ have prior probability density

$$p(e_1, \cdots, e_N) = \frac{\sqrt{det K}}{(2\pi)^{N/2}} \exp\left\{-\frac{1}{2} e^T K e\right\} , \tag{20–64}$$

where K^{-1} is the $(N \times N)$ noise prior covariance matrix. For "stationary white noise", it reduces to

$$K^{-1} = \sigma^2 \, \delta_{ij} \, , \qquad 1 \le i, j \le N \, . \tag{20–65}$$

Given K and the parameters $\{q_j\}$, the sampling *pdf* for the data takes the form

$$p(y_1, \cdots, y_N | qKI) = \frac{\sqrt{det(K)}}{(2\pi)^{N/2}} \exp\left\{-\frac{1}{2}(y - Fy)^T K (y - Fy)\right\} \tag{20–66}$$

Likewise, a very general form of joint prior *pdf* for the parameters is

$$p(A, \cdots, q_m | I) = \frac{\sqrt{det(L)}}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}(q - q_0)^T L (q - q_0)\right\} \tag{20–67}$$

where L^{-1} is the $(n \times n)$ prior covariance matrix and q_0 the vector of prior estimates. Almost always we shall take L to be diagonal:

$$L_{ij} = \sigma_j^2 \, \delta_{ij} \, , \qquad 1 \le i, j \le n \tag{20–68}$$

and q_0 to be zero. But the general formulas without these simplifying assumptions are readily found and programmed.

The joint posterior *pdf* for the parameters $\{q_j\}$ is then

$$p(q|yI) = \frac{exp(-\frac{Q}{2})}{\int exp(-\frac{Q}{2}) \, dA \cdots dq_n} \tag{20–69}$$

where Q is the quadratic form

$$Q \equiv (y - Fq)^T K (y - Fq) + (q - q_0)^T L (q - q_0) \tag{20–70}$$

which we may expand into eight terms:

$$Q = y^T K y - y^T K F q - q^T F^T K y + q^T F^T K F q + q^T L q - q^T L q_0 - q_0^T L q + q_0^T L q_0 \tag{20–71}$$

We want to bring out the dependence on q by writing this in the form

$$Q = (q - \hat{q})^T M (q - \hat{q}) + Q_0 \tag{20–72}$$

where Q_0 is independent of q. Writing this out and comparing with (20—71), we have

$$\begin{aligned} M &= F^T K F + L \,, \\ M\hat{q} &= F^T K y + L q_0 \,, \quad . \\ \hat{q}^T M \hat{q} + Q_0 &= y^T K y + q_0^T L q_0 \end{aligned} \tag{20–73}$$

M, $\hat{q}$, and Q_0 are thus uniquely determined, because the equality of (20—71) and (20—72) must be an identity in q:

$$\hat{q} = M^{-1}\left[F^T K y + L q_0\right] \tag{20–74}$$

$$Q_0 = y^T K y + q_0^T L q_0 - \hat{q}^T M \hat{q} \tag{20–75}$$

The denominator of (20—69) is then found using ( ), with the final result

$$p(A, \cdots, q_n | yKL) = \frac{\sqrt{det(M)}}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}(q-\hat{q})^T M (q-\hat{q})\right\} \tag{20–76}$$

The components $A, \cdots, q_m$ are the seasonal amplitudes we wish to estimate, while $(q_{m+1}, \cdots, q_n)$ are the trend nuisance parameters to be eliminated. From ( ) the marginal *pdf* we want is

$$\begin{aligned} p(A, \cdots, q_m \mid yKL) &= \int \cdots \int dq_{m+1} \cdots dq_n p(A, \cdots q_n | yKL) \\ &= \frac{\sqrt{det(M)}}{(2\pi)^{n/2}} \frac{(2\pi)^{(n-m)/2}}{\sqrt{det(\overline{W})}} \exp\left\{-\frac{1}{2}(u-\hat{u})^T U (u-\hat{u})\right\} \\ &= \frac{\sqrt{det(U)}}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}(u-\hat{u})^T U (u-\hat{u})\right\} \end{aligned} \tag{20–77}$$

where U, V, W, u are defined by ( ), ( ), ( ). From the fact that they are normalized, we see that

$$det(M) = det(W) det(U) \tag{20–78}$$

a remarkable theorem not at all obvious from ( ) and ( ) except in the case $V = 0$. This is another good example of the power of probabilistic reasoning to prove purely mathematical theorems.

Thus, the most general solution consists, computationally, of a string of elementary matrix operations and is readily programmed. To summarize the final computation rules:

K^{-1} is the $N \times N$ prior covariance matrix for the "noise".

L^{-1} is the $n \times n$ prior covariance matrix for the parameters.

F is the $N \times n$ matrix of model functions.

First, calculate the $(n \times n)$ matrix

$$M \equiv F^T K F + L \tag{20–79}$$

and decompose it into block form representing the interesting and uninteresting subspaces:

$$M = \begin{pmatrix} U_0 & V \\ V^T & W_0 \end{pmatrix} \tag{20–80}$$

Then calculate the $(m \times m)$ and $(r \times r)$ renormalized matrices

$$U \equiv U_0 - V W_0^{-1} V^T \tag{20–81}$$

$$W \equiv W_0 - V^T U_0^{-1} V \tag{20–82}$$

This much is determined by the definition of the model, and the computer can work all this out in advance, before the data are known.

Now given y, the $(N \times 1)$ data vector and q_0, the $(n \times 1)$ vector of prior estimates, the computer should calculate the $(n \times 1)$ vector

$$\hat{q} = M^{-1} \left[F^T K y + L q_0 \right] \tag{20–83}$$

of "best" estimates of the parameters. Actually, the first m of them are the interesting ones wanted, and the remaining $r = n - m$ components are not needed unless one also wants an estimate of the trend function. Then we can use the following result.

The inverse M^{-1} can be written in the same block form as M:

$$M^{-1} = \begin{pmatrix} U^{-1} & -U_0 V W^{-1} \\ -W_0 V^T U^{-1} & W^{-1} \end{pmatrix} \tag{20–84}$$

where, analogous to U,

$$W \equiv W_0 - V^T U_0^{-1} V \tag{20–85}$$

Then F^T has the same block form with respect to its rows:

$$(F^T)_{ji} = \left(G_j(t_i) \quad T_i(t_i) \right), \qquad \begin{array}{c} 1 \le j \le m, \\ 1 \le i \le N, \\ (m+1) \le K \le n \end{array} \tag{20–86}$$

where $G_j(t)$ are the seasonal sinusoids and $T_k(t)$ the trend functions.

Almost always, $q_0 = 0$ and so the "interesting" seasonal amplitudes are given by

$$\hat{q} = R K y \tag{20–87}$$

where R is the reduced $(m \times N)$ matrix

$$R \equiv U^{-1} G - U_0^{-1} V W^{-1} T \tag{20–88}$$

and U^{-1} is the joint posterior covariance matrix for the interesting parameters $\{A, \cdots, q_m\}$. Note that R and U^{-1} are determined by the model, so the computer can calculate them once and for all before the data are available, and then use them for any number of data sets.

**************************** MUCH MORE TO COME! ********************