Simulation of Time Series by Distorted Gaussian Processes

C. A. Greenhall
TDA Engineering Office

Distorted stationary gaussian processes can be used to provide computer-generated imitations of experimental time series. A method of analyzing a source time series and synthesizing an imitation is shown, and an example using X-band radiometer data is given.

I. Motivation

The simulation technique described here was motivated by the problem of weather-induced degradation of X- and K-band communication. A sequence of X-band noise temperature measurements is being gathered at Deep Space Station 13 (Goldstone); it is desired to use these data to study the effects of weather fluctuations on space communication. One can, for example, predict the percentage of time (out of a given year-quarter, for example) that the noise temperature exceeds a given level, but then no information about the variation of noise temperature with time is used. To study the effect of these fluctuations, one would like to have typical samples of noise temperature data to serve as inputs to communication system models. Computer-generated pseudorandom synthetic data have advantages over the real data, provided that the synthetic data preserve essential features of the real data. Such synthetic data can be controlled simply by changing parameters. As much (and only as much) data as needed can be generated. It is free of the inevitable bugs that infect the data-gathering process (but see Sections V and VI). Moreover, the synthetic data are more *random* than the original data, in the sense that the original data, once gathered and plotted, are *known*, whereas the exact course of the synthetic data is not. To change the sample function entirely, one need only start the program's pseudorandom-number generator at a different place. As M. Easterling put it, the real data are never *typical.*[1]

The noise temperature data are nowhere near gaussian. (Fig. 1.) This makes them more difficult to simulate. They motivate the search for a *general-purpose* simulation method, one that can be used to imitate a wide variety of time series. The method used here is: generate a stationary gaussian process having certain correlations, and distort it by a zero-memory nonlinearity so that the result has a desired marginal distribution. The problem is how to choose the correlations.

[1] Private communication.

II. Ideas Leading to Present Method

We wish to produce a random process $Y_0, Y_1, \ldots$ that in some way imitates a source time series $y_1, y_2, \ldots, y_N$. There are two broad questions:

(1) What statistical properties of the y_i are to be duplicated?

(2) What class of processes shall the Y_i belong to?

The answers to these two questions depend on each other, of course. Let us start with some tentative answers: The process (Y_i) is stationary. Its marginal distribution function

$$P\{Y_i \leqslant y\}$$

and a certain number of correlations

$$\rho(Y_i, Y_{i+t}), \qquad t = 1 \text{ to } n$$

agree with the sample distribution function and correlations of the source series. This leads to another question: Is there a stationary process having *arbitrary* prescribed marginal distribution function F (with finite second moment) and (nonnegative definite) correlation function λ_1 to λ_n? (Here, λ_t is the correlation for lag t.) The answer is quickly no, for if the distribution described by F is not symmetric about its mean[2], then there is a number $\lambda_{\min} > -1$ such that

$$\rho(Y_1, Y_2) \geqslant \lambda_{\min}$$

for any random variables Y_1, Y_2 whose marginal distributions are both F. For such F, the λ_t cannot be allowed to get too close to -1. Given F, then, there is a set C_F of permissible correlation sequences $(\lambda_1, \ldots, \lambda_n)$. However, it is unlikely, perhaps even impossible, that the sample correlations of (y_i) fall outside C_F for the sample distribution function F. So far, the proposed answer to questions (1) and (2) seems feasible.

It is time to restrict the answer to (2) to a class of processes that are easy to generate to order. We say that (Y_i) is a *stationary distorted gaussian process* if, for some function g,

$$Y_i = g(X_i), \tag{1}$$

where $X_1, X_2, \ldots$ is a stationary *standard* (mean 0, variance 1) gaussian process. The function g can be chosen to give the Y_i any desired marginal distribution function F. In fact, we can take

$$g(x) = F^{-1}(\Phi(x)), \tag{2}$$

where F^{-1} is the (generalized) inverse of F, and Φ is the standard gaussian distribution function. Then g is nondecreasing. There is an invertible function Λ_g such that the correlations λ_t of the Y_i are related to the correlations ρ_t of the X_i by

$$\lambda_t = \Lambda_g(\rho_t) \tag{3}$$

(Ref. 1). The program now appears to be: Given a distribution function F and a correlation sequence $(\lambda_1, \ldots, \lambda_n)$ in C_F, let

$$\rho_t = \Lambda_g^{-1}(\lambda_t), \qquad t = 1 \text{ to } n \tag{4}$$

Construct a stationary standard gaussian process (X_i) with correlations ρ_1 to ρ_n. Then the Y_i have the desired marginal F and correlations λ_t.

A program similar to this was carried out by Posner and Zeigler (Ref. 2), using $g(x) = |x|$ (not a monotonic function), and specifying two nonnegative correlations λ_1, λ_2. They noted that they had not proven that *any* (nonnegative) (λ_1, λ_2) in C_F (where F is the distribution function of a "half-gaussian") can be reached by Eq. (3), where ρ_1, ρ_2 are the correlations of a gaussian process. This same question must be asked of the general program above. Another way to put it is: Given $(\lambda_1, \ldots, \lambda_n)$ in C_F, define $(\rho_1, \ldots, \rho_n)$ by Eq. (4), and let $\rho_{-t} = \rho_t$. Is the sequence

$$\rho_{-n}, \ldots, \rho_{-1}, 1, \rho_1, \ldots, \rho_n$$

nonnegative definite? If it is not, the process (X_i) does not exist.

E. Rodemich found a counterexample that shows that the program fails in general. Consider the three-point sample space $\{1, 2, 3\}$, with $P\{i\} = 1/3$. For i, = 1 to 3, let $Y_i(\omega) = \sqrt{2}$ if $\omega = i$, and $Y_i(\omega) = -1/\sqrt{2}$ otherwise. Then Y_i is standard, and $EY_iY_j = -\frac{1}{2}$ if $i \neq j$. Suppose that X_1, X_2, and X_3 are *jointly* gaussian with standard marginals, and that g is a function such that $g(X_1)$, $g(X_2)$, $g(X_3)$ have the same marginals and correlations as the Y_i. Then with probability 1, $g(X_i)$ only takes values $\sqrt{2}$ and $-1/\sqrt{2}$, and

$$g(X_1) + g(X_2) + g(X_3) = 0 \tag{5}$$

[2] We owe this observation to E. Rodemich.

Let $A = \{x{:}g\ (x) = \sqrt{2}\}$. Then $P\{X_i \epsilon A\} = 1/3$. The rank of the distribution of (X_1, X_2, X_3) must be 1, for suppose, say, that (X_1, X_2) had a density. Then $P\{X_i \epsilon\ A,\ X_2 \epsilon\ A\} > 0$, which is impossible, for if X_1 takes a value in A, then Eq. (5) implies that X_2 does not. On the other hand, if the rank is 1, then $X_2 = \pm X_1$, $X_3 = \pm X_1$: therefore at least two of the X_i are equal, which is again impossible. The process (Y_i) cannot be simulated by a stationary distorted Gaussian process. Notice, however, that we used a *process* instead of a *time series* for this example. The *finite* time series

$$(1/\sqrt{2})\,(2, -1, -1, 2, -1, -1, \ldots, -1) \qquad (6)$$

won't work for the example unless correlations are computed cyclically, which we don't want to do. We conjecture that the sample distribution and correlations of any finite time series *can* be obtained by a stationary distorted Gaussian process.

For the time being, we are stepping around the problem, instead of surmounting it. One can always specify the correlations of a gaussian process, as long as they are nonnegative definite. Therefore, let us take the source time series (y_i), measure its sample distribution function F, and use F to *compress* the y_i into a time series (x_i) whose sample distribution function is approximately Φ. In fact, let

$$x_i = g^{-1}\,(y_i) = \Phi^{-1}\,(F\,(y_i)). \qquad (7)$$

Measure the sample correlations $\rho_1, \ldots, \rho_n$ of the x_i. These will be nonnegative definite (if defined correctly), so a gaussian process (X_i) can be generated having these correlations. Finally, use Eq. (1) to *expand* the X_i into a process (Y_i) whose marginal distribution approximately equals the sample distribution of the y_i. Moreover, when (Y_i) and (y_i) are compressed by g^{-1} to (X_i) and (x_i), respectively, the correlations of the X_i equal the sample correlations of the x_i. Note that (X_i) is a true gaussian process, whereas the *most* that can be said about the x_i is that their sample distribution is approximately gaussian.

III. Analysis of Source Time Series

The sample distribution function of the source time series $y_1, \ldots, y_N$ is given by

$$F\,(y) = (\text{number of } y_i \leqslant y)/N.$$

A binsort of the data will yield the values of F at the bin boundaries. If one agrees to interpolate F linearly between the boundaries, then Eq. (7) yields a sequence (x_i) whose sample distribution function is a step-function approximation to Φ. To make this approximation good, the jumps of F should be small, and the bin boundaries close enough together to allow linear interpolation. In implementing Eq. (7), it is also a good idea to truncate Φ^{-1} at 4 and −4, say. This avoids the possibility of the linear interpolation producing grossly oversize x_i for those y_i that are very close to the maximum bin boundary.

One computes the sample mean, covariances, and correlations of the x_i:

$$\mu = (1/N) \sum x_i,$$

$$r_t = (1/N) \sum_{i=1}^{N-t} (x_i - \mu)\,(x_{i+t} - \mu),$$

$$\rho_t = r_t / r_0, \qquad t = 0 \text{ to } n.$$

If the previous work was done well, μ should be close to 0, and r_0 to 1. The ρ_t are guaranteed to be nonnegative definite.

The functions F and ρ serve as inputs to the synthesis algorithm given in the next section.

IV. Synthesis of Artificial Time Series

We are given tabulated values of a distribution function F, and a sequence of correlations $\rho_0, \rho_1, \ldots, \rho_n$, where $\rho_0 = 1$. The main job is to generate a stationary standard gaussian process $X_0, X_1, \ldots$ such that

$$EX_i X_{i+t} = \rho_t, \qquad t = 0 \text{ to } n.$$

Then the process $Y_0, Y_1, \ldots$ that we seek is given by

$$Y_i = F^{-1}\,(\Phi\,(X_i)), \qquad (8)$$

where F^{-1} is executed by linear interpolation in the F-table.

Here is an algorithm that generates the X_i as an autoregressive scheme. Let $Z_0, Z_1, \ldots$ be a sequence of independent standard gaussians. Execute the following steps in order:

Step 0. Set

$$g_0 = 1, c_{00} = 1, X_0 = Z_0$$

Step i, $i = 1$ to n. Set

$$a_j = \frac{1}{g_j} \sum_{k=0}^{j} \rho_{i-k} c_{jk}, \qquad j = 0 \text{ to } i-1$$

$$g_i = 1 - \sum_{j=0}^{i-1} a_j^2 g_j$$

$$c_{ij} = - \sum_{k=j}^{i-1} a_k c_{kj}, \qquad j = 0 \text{ to } i-1$$

$$c_{ii} = 1$$

$$X_i = \sqrt{g_i} Z_i - \sum_{j=0}^{i-1} c_{ij} X_j \tag{9}$$

Step i, $i > n$. Set

$$X_i = \sqrt{g_n} Z_i - \sum_{j=0}^{n-1} c_{nj} X_{i-n+j} \tag{10}$$

This algorithm constructs a lower triangular matrix $C = (c_{ij})$ and a nonnegative diagonal matrix $G = \text{diag}\,(g_0, \ldots, g_n)$ such that $CRC^T = G$, where $R = (\rho_{i-j})$, with $\rho_{-i} = \rho_i$. The vectors (a_j) are the rows of C^{-1}.

The algorithm goes through if and only if R is nonnegative definite and has rank n or $n+1$. Otherwise, the algorithm will run into a negative g_i, in which case R is indefinite, or a $g_m = 0$ for some $m < n$, in which case $R_m = (\rho_{i-j}: i, j = 0 \text{ to } m)$ is nonnegative definite and singular, and the full matrix R may or may not be nonnegative definite.

The synthesis procedure has been realized in a documented MBASIC program TSS (Time Series Synthesis).

V. Example

Figure 1 shows a plot of X-band noise temperature measurements made by a radiometer at the Goldstone DSCC from Day 207 to Day 214 of 1976. The data have preprocessed so that they represent noise temperature above quiescent as seen at zenith. The data come once every two minutes, but the plot samples them only once every 20 minutes. The gaps indicate missing data; the program that computes correlations maintains the correct time relationships among the rest of the data. The sharp negative peaks are caused by equipment malfunction; nevertheless, for the purpose of this exercise, they were not excised.

Figure 2 shows five weeks worth of output of the Time Series Synthesis program, whose inputs were the distribution function and correlations obtained by the procedure of Section III from the data of Fig. 1. Only the correlations $\rho_{20k \text{ min}}$, $k = 1$ to 13, were used; thus the order n of the autoregression is 13, and the output of the program represents 20-minute samples.

VI. Remarks

Comparing Figs. 1 and 2, we see that the synthesis program does produce sharp irregular peaks resembling those of the source data. The peaks of the synthetic data seem to be more clumped together than those of the source data, and the quiet periods of the synthetic data are noisier than those of the source. The spurious negative peaks of the source cause strange-looking clumps of negative excursions in the synthetic data. Obviously, bad points should be removed when putting a time series through the analysis procedure; alternatively, the distribution function can be fixed before giving it to the synthesis program.

Some objections to the technique come to mind.

First, there is no objective criterion for acceptance of the output of the synthesis program. It does have certain statistical properties in common with the source time series, the ones it was designed to have, but other than that, one can perhaps only ask whether it "looks right."

The second objection applies to radiometer data. The important features of Fig. 1 are the large positive peaks, for during these periods, X-band communication is considerably degraded. This time series is severely compressed by Eq. (7) into a gaussian mold; the peaks become insignificant and can affect the correlations of the compressed series only very little. Most of the information in the correlations comes from the uninteresting quiet periods. Yet, these correlations strongly affect the peaks of the synthetic data. Perhaps this is why the peaks of the synthetic data tend to come in clumps. Actually, the peaks of the source data are probably caused by phenomena (clouds or rain) that are independent of the phenomena that cause the small fluctuations of the quiet periods. We may be fooling ourselves if we treat these data as a *single* time series.

Third, it has been objected that the compression-plus-correlation technique requires that the source time series be saved in case one wants to improve the distribution function and correlations by using additional source series. As an illustration, consider two source series $y_1, \ldots, y_N$ and $y'_1, \ldots, y'_N$ of the same length, which give rise to compressed series (x_i) and (x'_i), distribution functions F and F', and correlations ρ and ρ'. The distribution function of the combined source series is $F'' = (F + F')/2$. One should use F'' to compress $y_1, \ldots, y_N, y'_1, \ldots, y'_N$ to a series

$$x''_1, \ldots, x''_N - \text{gap} - x''_{N+1}, \ldots, x''_{2N}.$$

Then the correlation function ρ'' of the x''_i would be computed. However, this may not be necessary. The x_i and x'_i series both have an approximately gaussian sample distribution; it seems reasonable to use the series

$$(x'''_i) = (x_1, \ldots, x_N - \text{gap} - x'_1, \ldots, x'_N)$$

in place of (x''_i) for computing the new correlations. If the sample means and variances of (x_i) and (x'_i) are close to 0 and 1, respectively, then the sample correlation of (x'''_i) is close to $(\rho + \rho')/2$. In this case, it is sufficient to save only F and ρ, instead of the source series.

Finally, it must be admitted that the idea of computing correlations after compressing is an expediency created to dodge a difficult mathematical problem, the relationship between marginals and correlations for nongaussian processes and finite time series. In fact, the dodge itself may be illusory; our "gaussian correlation" technique will not work at all well on the time series of Eq. (6). But this time series comes from the very counterexample which we used as an excuse to go to the gaussian correlation technique. Perhaps what we are *really* dodging is the numerical evaluation of Λ_g^{-1} in Eq. (4). More mathematical effort is needed to clarify the situation. In the meantime, the present method works in practice.

The conjecture that appears after (6) is false. Provided that (6) is long enough, no stationary distorted gaussian process with nondecreasing distortion function g can have the same distribution and correlations as (6). This fact makes the compression-plus-correlation technique more attractive, for we can guarantee that the analysis and synthesis procedures can be carried out, whatever the source time series (y_i). However, for pathological examples such as (6), the sample distribution function of the compressed series (x_i) will not be close to Φ.

References

1. Rodemich, E. R., "Spectral Estimates Using Nonlinear Functions," *Annals of Mathematical Statistics*, Vol. 37, No. 5, October 1965.
2. Posner, E. C., and Zeigler, F. J., "A Technique for Generating Correlated X-band Weather Degradation Statistics," in *The Deep Space Network Progress Report* 42-35, Jet Propulsion Laboratory, Pasadena, Calif., Oct. 15, 1976.

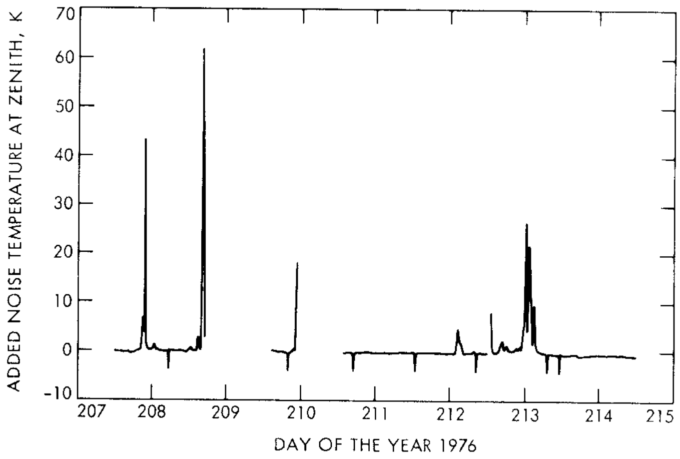

Fig. 1. One week of X-band radiometer data

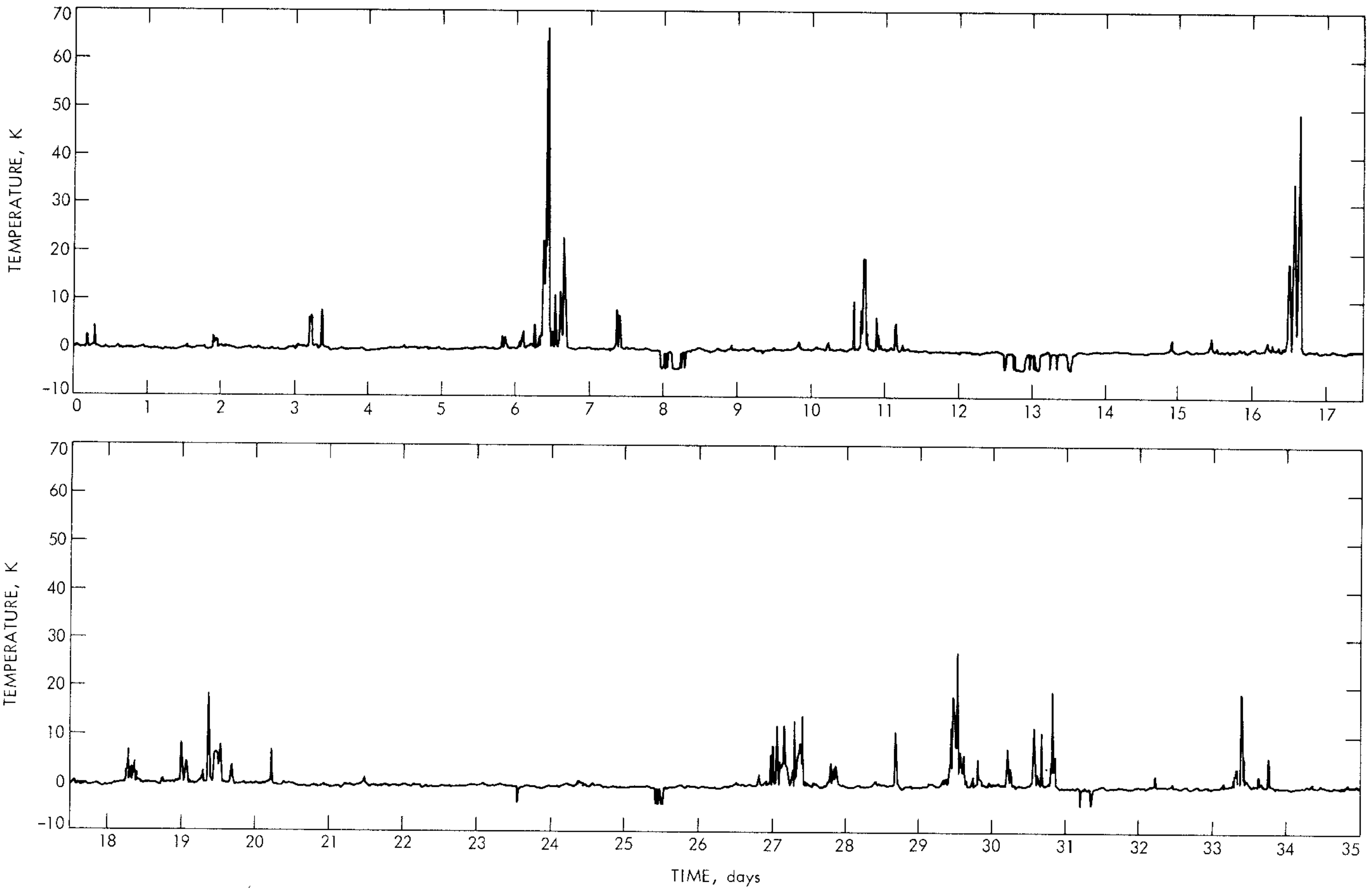

Fig. 2. Five weeks of synthetic data