Reduction of State Variable Dimension in Stochastic Dynamic Optimization Models which Use Time-Series Data

Oscar R. Burt and C. Robert Taylor

Statistical procedures are developed for reducing the number of autonomous state variables in stochastic dynamic optimization models when these variables follow a stationary process over time. These methods essentially delete part of the information upon which decisions are based while maintaining a logically consistent model. The relatively simple linear autoregressive process as well as the general case is analyzed and the necessary formulae for practical application are derived. Several applications in agricultural economics are discussed and results presented which quantify the relative amount of information sacrificed with the reduction in number of state variables.

Key words: dynamic programming, dynamic optimization, time series, curse of dimensionality.

Empirical stochastic dynamic optimization problems frequently become unwieldy with too many state variables required to fully summarize the entire history of the decision process. Bellman has called this problem the "curse of dimensionality." Although continuing advances in computing power and available memory cause this curse to fade appreciably, empirical practitioners must still use considerable ingenuity as well as good judgment in arriving at computationally operational yet acceptably accurate models. Burt discussed various ways of reducing the dimensionality problem, one of which was to deliberately discard part of the information contained in the full set of state variables. This paper provides a practical methodology for implementing a reduction in the number of state variables when a subset of these variables emanates from time-series data.

In many actual and potential applications of dynamic optimization to problems in agricultural economics, there are some sets of state variables that are unaffected by the decisions. For example, prices are obvious state variables in most firm level decision models, yet in competitive markets we do not expect an individual firm's decisions to affect price. Thus, prices can be viewed as "autonomous" state variables in the decision model. It is not uncommon to find that autonomous state variables such as prices appear to be generated by second- or higher-order autoregressive (AR) processes. In such cases, the number of state variables pertaining to this component of a fully specified decision model is equal to the order of the AR process. Often, computational considerations dictate use of fewer state variables in the decision model and thus possible use of a lower-order AR process in the model.

The following pest management problem taken from Danielson illustrates the ideas involved. A soil-born disease (Cephalosporium stripe) in winter wheat is controlled by not planting winter wheat for one or more years. Alternative land uses are summer fallow or barley, so it is clear that expected wheat and barley prices are important information in the decision process. Necessary state variables for the dynamic decision process are the number of years since winter wheat was grown, the level of infection in that last crop of winter

Oscar R. Burt and C. Robert Taylor are professors of agricultural economics at the University of California, Davis, and Auburn University, respectively.

Giannini Foundation Research Paper No. 903.

wheat, and however many lagged wheat and barley prices are needed to forecast these two prices. A statistical analysis of the bivariate time series of these two prices suggested that the best model was a separate equation for each price, namely, second- and first-order AR processes for wheat and barley prices, respectively. Two questions naturally come to mind in this case: (*a*) Can we use a first-order process to approximate the time-series structure of wheat prices without losing much information? (*b*) How do we estimate the unknown parameters for the first-order process when the actual process is second order?

This article presents the statistical procedure for reducing the order of a Markov process[1] for use in a decision model as well as providing a useful quantitative answer to the first question posed above. Our approach works only in those cases where the decision variables do not affect the subset of state variables to be reduced in dimension either directly or indirectly; that is, the approach works for a subset of autonomous state variables. The focus of the paper is on Markov processes estimated with time-series data and decision models that will be solved using stochastic dynamic programming (DP); nevertheless, many of the derivations are appropriate for dimensional compromises necessitated by other solution techniques.[2]

Four cases are treated in the article. The first case is for a linear stochastic difference equation (AR process) with an intertemporally uncorrelated and homoskedastic disturbance term. This case applies to the example of wheat prices given above. The second case is for two interdependent linear difference equations with intertemporally uncorrelated, but contemporaneously cross correlated, homoskedastic disturbance terms. This case would have applied to the example above if the disturbance terms in the barley and wheat price regression equations had been correlated significantly. The third case is for a single continuous random variable with a general second-order Markov probability distribution, and the fourth case is for a single discrete random variable following a general finite Markov process. These last two cases encompass general nonlinear relationships among time dated values of the random variable. We begin with a detailed statement of the problem.

The Problem

Reducing state variable dimensions in stochastic dynamic optimization models is largely an art as opposed to a science in formulating and solving stochastic DP models (Dreyfuss and Law, p. xi). The problem is largely one of structuring the discrete stochastic dynamic decision process so that the most useful information is used for a given level of computational difficulty. The quantity and quality of information used determines the maximized value of the expected value criterion function used in stochastic DP. When the naturally occurring state variables are continuous, information is lost in the transformation to a discrete approximation for computational purposes as well as from reduction in the number of state variables by approximating a given order Markov process by one of a lower order.

For given computational time and storage, the issue is often whether it is best to have n discrete classes on each of say y_{t-1} and y_{t-2} as state variables, thus giving a total of n^2 discrete states, or have n^2 discrete classes on y_{t-1} as a single state variable for this component of the model. Since we usually have additional state variables to represent other dynamic processes, we are often restricted to small values of n. It is often the case that using a finer grid on y_{t-1} only rather than using a coarse grid on y_{t-1} and y_{t-2} would lead to a more nearly optimal solution of the stochastic dynamic optimization model, especially when the second-order effect is in some sense small relative to the first-order effect. Thus, reducing the order of Markov processes as a means for state variable reduction in stochastic dynamic optimization models is pertinent to many empirical applications in agricultural economics.

[1] A generalized Markov process refers to any stochastic process in which a random variable is related to a finite number of previous levels of that random variable. Markovian relationships can, but need not, be conditional on exogenous variables. Thus, Markovian processes embrace random walk, rational expectations, autoregressive, and many other conditional models that are often used to model economic and technical variables of interest in agricultural economics.

[2] Although the phrase "curse of dimensionality" was coined in the context of using DP to numerically solve problems, the curse also manifests itself with dynamic programming or control theory used to analytically solve problems. In the numerical case, the curse refers to the computational and storage problems, while in the analytical case, the curse refers to the problem of analytically solving a set of difference or differential equations. In the analytical case, reducing the number of state variables thus reduces the number of dynamic equations to be simultaneously solved.

A Linear Stochastic Difference Equation

The typical problem in practice will be a process described by a second-order linear difference equation which implies two state variables in a dynamic optimization model. This stochastic process is represented by

$$(1) \qquad y_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + e_t,$$

where y_t and e_t are random variables measured at time t, and b_0, b_1, and b_2 are fixed parameters. Of course, the units of y can just as well be after a logarithmic transformation if such a model better fits the empirical situation. This is the classic second-order autoregressive process in univariate time-series models for which the following assumptions are made: (*a*) e_t is independently and identically distributed with expectation zero, and (*b*) the roots of the characteristic equation, $x^2 - b_1 x - b_2 = 0$, lie inside the unit circle. The latter condition can be summarized by the three inequalities

$$\begin{aligned} b_1 + b_2 &< 1, \\ b_2 - b_1 &< 1, \\ |b_2| &< 1, \end{aligned}$$

which are useful for verification of the stability of a fitted equation.

The above two assumptions guarantee that the process is stationary in the sense that the joint distribution of y_t and $y_{t\pm i}$ depends only on i and not on t. Also, e_t is independent of y_{t-1} and y_{t-2}, which can be demonstrated by expressing y_t in its equivalent so-called moving-average representation where it can be seen that y_t is correlated with current and past but not future values of e (Nelson).

Before proceeding, it is important to note that where the underlying economic or technical process is second order, it is statistically inappropriate to simply reestimate a linear difference equation with the second-order term deleted from the model. This approach is undesirable because the associated parameter estimators of b_0 and b_1 do not have desirable properties such as unbiasedness and consistency. Thus, basing transition probabilities on a model obtained in this manner will result in biased and inconsistent probabilities for use in the stochastic DP model. Also, the problem is not overcome by using a first-order process in y_{t-1} with a first- or higher-order moving-average error structure as an approximation, because the presence of the error structure would imply that the error term, say u_{t-1}, as well as y_{t-1} should be treated as state variables in the decision model.

Conditional Distribution for the Second-Order Process

Before turning to reducing the order of the Markovian process, it is instructive to first consider derivation of the conditional distribution of the second-order process for the linear model, (1), estimated by regression. For the second-order process, we are viewing y_{t-1} and y_{t-2} in (1) as conditionally fixed, and y_t is a random variable linearly related to the random error, e_t. From (1) we can obtain the conditional expectation of y_t, which is

$$(2) \qquad E(y_t \mid y_{t-1}, y_{t-2}) = b_0 + b_1 y_{t-1} + b_2 y_{t-2},$$

where $E(\cdot)$ is the expectation operator. Similarly, we can obtain the conditional variance of y_t from (1) as

$$(3) \qquad \mathrm{Var}(y_t \mid y_{t-1}\, y_{t-2}) = \mathrm{Var}\,(e_t) = \sigma_e^2.$$

By assuming a particular form for the probability distribution of e_t, such as the normal, one can calculate the discrete conditional probabilities of y_t over a rectangular grid for the two state variables, y_{t-1} and y_{t-2}. This is the type of data required in the stochastic DP model. More specifically, let $F(y_t \mid y_{t-1}, y_{t-2})$ be the cumulative distribution function, and the cells of the grid on y_{t-1} and y_{t-2} are denoted by coordinates of the center of a cell (a_i, a_j). For one of the discrete outcomes on y_t, we use a_k as the midpoint and a_{k-} and a_{k+} for the lower and upper boundaries of the class interval, respectively. Then the transition probabilities for the second-order process are given by the approximation,

$$\begin{aligned} (4) \qquad &\mathrm{pr}(a_{k-} \leq y_t < a_{k+} \mid y_{t-1} = a_j,\ y_{t-2} = a_i) \\ &\quad = F(a_{k+} \mid a_j,\ a_i) - F(a_{k-} \mid a_j,\ a_i). \end{aligned}$$

Conditional Distribution for the Reduced Process

From a conceptual standpoint, we are throwing away information by transforming the state vector into a lower dimension, and to some extent, we are creating an artificial Markov process for the decision model. It is artificial

in that we must visualize a random drawing at each stage of the decision process from some underlying event space (time series) which only exists in principle. The created process is not, strictly speaking, capable of tracking the system from stage to stage as a Markov process to meet the definition that the current state summarizes all the history of that process with respect to making optimal decisions from this stage forward. Instead, the current state summarizes all the information which will be used to make the current and future decisions in a decision criterion restricted to that information set.

The Markov requirement is met by thinking of a sampling experiment over a population of all possible historical outcomes which could have given rise to the current state as described by the limited information being used (reduced state dimension). For time series such as prices, the population would be over all possible sequences of years in some ensemble of possible time-series samples with the same structure being assumed as that given by the empirically estimated equation. Using (1) as an example with y_{t-1} taken as given information, but not y_{t-2}, we are interested in the conditional joint distribution of y_t and y_{t-2} as a means to ultimately get the marginal distribution of y_t conditional on y_{t-1} only.

The linearity of (1) and associated assumptions of the second-order autoregressive process let us directly derive the conditional mean and variance of y_t, given y_{t-1}. Taking the conditional expectation of (1) gives

$$(5)\quad E(y_t \mid y_{t-1}) = b_0 + b_1 y_{t-1} + b_2 E(y_{t-2} \mid y_{t-1}).$$

Stationarity in the time series makes $E(y_{t-2} \mid y_{t-1}) = E(y_t \mid y_{t-1})$ because in each case, the subscript of the random variable y is one period removed from the same given value, y_{t-1}, albeit reversed in the time sequence. Therefore, (5) reduces to

$$(6)\quad E(y_t \mid y_{t-1}) = (b_0 + b_1 y_{t-1})/(1 - b_2).$$

Proceeding directly with the formula for the variance of linear combinations of random variables,

$$\mathrm{Var}(y_t \mid y_{t-1}) = b_2^2 \,\mathrm{Var}(y_{t-2} \mid y_{t-1}) + \mathrm{Var}(e_t).$$

There is no covariance term because e_t is independent of y_{t-2}, and by the stationarity of the process, $\mathrm{Var}(y_{t-2} \mid y_{t-1}) = \mathrm{Var}(y_t \mid y_{t-1})$. Therefore, we can write

$$(7)\quad \mathrm{Var}(y_t \mid y_{t-1}) = \sigma_e^2/(1 - b_2^2).$$

Using (6) and (7) for the first two moments of y_t conditional on y_{t-1}, a unique probability distribution is determined if we assume e_t is normal. With the implied cumulative distribution, the same procedure as (4) can be used to calculate discrete probabilities for y_t but with one small change. Now only a value of $y_{t-1} = a_i$ defines the given condition instead of a joint condition involving both y_{t-1} and y_{t-2}.

In the above discussion, the unknown parameters, b_0, b_1, b_2, and σ^2, have been treated as known values, but in applications they will be estimated values from sample data. It is assumed that these estimates will be treated as if they were the parameters in the stochastic optimization problem under consideration. To do otherwise would require a Bayesian approach and much more complicated analysis.

The above analysis assumes a stable structure on the difference equation in (1), but the results can be extended to the case where the characteristic equation has a unit root. It is well known in the time-series literature that differencing such an equation yields a stable first-order process if there is a single unit root (Nelson). In this case (1) is replaced by

$$(8)\quad (y_t - y_{t-1}) = c(y_{t-1} - y_{t-2}) + e_t$$

which can be written as

$$(9)\quad y_t = (1 + c)y_{t-1} - cy_{t-2} + e_t,$$

where we have dropped the constant term b_0 because there is no convergent steady state for this process.

Although the unconditional variance of y_t is undefined in (9) because of the unit root for the characteristic equation, the conditional variance given y_{t-1} is defined. In fact, the derivation to get (6) and (7) is unchanged except for b_0 not appearing in the conditional expectation. Therefore, (6) and (7) are replaced by

$$(10)\quad \begin{aligned} E(y_t \mid y_{t-1}) &= y_{t-1}, \\ \mathrm{Var}(y_t \mid y_{t-1}) &= \sigma_e^2/(1 - c^2), \end{aligned}$$

as the dynamic relationships for a stochastic DP model.

This single-unit root model is quite frequently applicable, e.g., in a time-series representation of farmland prices. Although Phipps used a second-differenced model with a first-order moving-average disturbance to represent aggregate U.S. farmland prices, analysis by one of the authors suggested that a first-differenced model in a first-order autoregres-

sive process is about equally plausible for aggregate data and more plausible for individual state data. State variables to represent farmland prices are important in stochastic DP models of farm firm growth and finance (Schnitkey).

Quantification of Discarded Information

It would be useful to have at least a rough idea of the amount of information lost by the reduction of a second-order to a first-order process. This can be done by a comparison of the residual variances in an analogous way to the construction of R^2 in a multiple regression context. The "total" variance as a basis for comparison is the unconditional variance of the series, that is,

$$\text{(11)}\quad \text{Var}(y_t) = (1 - b_2)\sigma_e^2/(1 + b_2)[(1 - b_2)^2 - b_1^2],$$

which can be derived directly from the Yule-Walker equations (Box and Jenkins, p. 62). The two conditional variances for comparison are given in (3) and (7), and each of these can be viewed as a "residual" variance. A measure of the variance removed by means of the conditional distribution is total minus residual variance, and the analogue of R^2 would be that difference divided by the total variance.

These R^2 analogues are denoted $\tilde{R}_1^2$ and $\tilde{R}_2^2$ for y_{t-1} given and both y_{t-1} and y_{t-2} given, respectively. The formulae are derived directly from (3), (7), and (11) to get

$$\text{(12)}\quad \tilde{R}_1^2 = b_1^2/(1 - b_2)^2,$$
$$\tilde{R}_2^2 = b_2^2 + b_1^2(1 + b_2)/(1 - b_2).$$

But we are primarily interested in the relative amount of variance removed by the two conditional distributions, second- versus first-order difference equations. A relative measure bounded between zero and one is the ratio $\tilde{R}_1^2/\tilde{R}_2^2$ which simplifies to

$$\text{(13)}\quad \theta = b_1^2/[(1 - b_2)^2 b_2^2 + b_1^2(1 - b_2^2)].$$

A high value of θ implies little information would be lost by the first-order approximation.

In many of the cases where a second-order process fits an empirical situation, $b_1 > 0$ and $b_2 < 0$. An explanation of this phenomenon is suggested by the following reparameterization of (1),

$$\text{(14)}\quad y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2(y_{t-1} - y_{t-2}) + e_t,$$

Table 1. Comparison of Conditional Variances in Second-Order Difference Equations

β_1	β_2	b_1	b_2	$\tilde{R}_1^2$	$\tilde{R}_2^2$	θ
.5	.3	.8	−.3	.38	.44	.87
.5	.5	1.0	−.5	.44	.58	.76
.6	.4	1.0	−.4	.51	.59	.87
.6	.6	1.2	−.6	.56	.72	.78
.7	.3	1.0	−.3	.59	.63	.94
.7	.5	1.2	−.5	.64	.73	.88
.7	.7	1.4	−.7	.68	.84	.81
.8	.2	1.0	−.2	.69	.71	.98
.8	.4	1.2	−.4	.74	.78	.95
.8	.6	1.4	−.6	.77	.85	.90
.8	.7	1.5	−.7	.78	.89	.88
.8	.8	1.6	−.8	.79	.92	.86
.9	.3	1.2	−.3	.85	.87	.99
.9	.4	1.3	−.4	.86	.88	.98
.9	.5	1.4	−.5	.87	.90	.96
.9	.6	1.5	−.6	.88	.92	.95
.9	.7	1.6	−.7	.89	.94	.94
.9	.8	1.7	−.8	.89	.96	.93
.9	.9	1.8	−.9	.90	.98	.92

where we would expect $\beta_1 > 0$ and $\beta_2 > 0$. The intuitive interpretation of (14) is that β_1 extrapolates the level of y, and β_2 extrapolates the changes in y. A typical application with this structure is annual prices of cattle. In terms of the parameters in (1), $b_1 = \beta_1 + \beta_2$ and $b_2 = -\beta_2$.

Results are reported in table 1 for an array of combinations for β_1 and β_2 with the restriction $\beta_1 < \beta_2$, which was thought to be the more common situation. The last three columns are $\tilde{R}_1^2$, $\tilde{R}_2^2$, and θ, respectively. Notice that θ is seldom below .9 when $\beta_1 > .7$. Also, θ is monotonically decreasing with an increase in β_2 while holding β_1 constant, and the $\tilde{R}^2$s are monotonically increasing with β_1 and β_2.

Three Applications

The first case is the one discussed in the introduction dealing with disease control in winter wheat. The statistically estimated second-order difference equation with wheat prices (1984 dollars) in natural logarithms was (numbers in parentheses are t-ratios)

$$\hat{y}_t = .47 + \underset{(7.2)}{1.151}y_{t-1} - \underset{(2.9)}{.457}y_{t-2},$$

with $\hat{\sigma}_e = .152$. Application of (6) gives the first-order equation

$$\hat{y}_t = .33 + .79y_{t-1},$$

and (7) yields $\sqrt{\text{var}(y_t \mid y_{t-1})} = .171$ which is

only 13% larger than $\hat{\sigma}_e$. From (13), θ is calculated to be .89.

The second case is an empirical dynamic hedging model for Montana winter wheat producers (research still in progress).[3] Important state variables in this stochastic DP model are lagged basis variables for May and December futures contracts (futures price at Kansas City minus cash price in Montana). Statistical criteria for the December contract suggested a second-order stochastic difference equation, containing time to contract maturity as an exogenous variable. The estimated equation for monthly data (1977–87) was

$$\hat{y}_t = \underset{}{.0090} + \underset{(7.1)}{.646}y_{t-1} + \underset{(2.4)}{.226}y_{t-2} + \underset{(2.4)}{.015}TM_t,$$

where TM denotes time to maturity. The May contract had essentially the same regression coefficients on lagged basis. The first-order equation given by (6) is

$$\hat{y}_t = .0116 + .835y_{t-1},$$

and θ is .978. This very favorable result for the first-order approximation tends to hold whenever both b_1 and b_2 are positive.

The third case is a firm growth model for an Illinois feeder pig operation (research still in progress).[4] One of the primary economic measures is monthly gross margins (revenue minus variable cost) per hog barn. The time-series statistical structure of this variable determines a subset of state variables in the DP model. A monthly data series for 1974–87 was analyzed and the structure selected on statistical criteria was second-order AR,

$$\hat{y}_t = \underset{}{2.43} + \underset{(16.5)}{1.177}y_{t-1} - \underset{(6.3)}{.439}y_{t-2}.$$

The coefficient on y_{t-1} for the first-order approximation using (6) is .82, and θ equals .91. These results are surprisingly close to those for the first case involving wheat prices.

Higher-Order Processes

The same procedure as used to analyze the linear second-order process can be applied to third- or higher-order processes, but the algebra gets cumbersome. An empirically operational method to handle such extensions is to assume normality and exploit the matrix methods of multivariate normal theory (Anderson). Reduction of the order of a process by one, such as a third to second order, is straightforward from the results presented above.

Interdependent Linear Stochastic Difference Equations

Now consider two linear, second-order interdependent processes, say price of corn, y_{1t} and the price of soybeans, y_{2t}. The general second-order bivariate autoregressive process is

$$\begin{aligned} y_{1t} &= \gamma_1 + \alpha_{11}y_{1,t-1} + \alpha_{12}y_{2,t-1} \\ &\quad + \beta_{11}y_{1,t-2} + \beta_{12}y_{2,t-2} + e_{1t}, \\ y_{2t} &= \gamma_2 + \alpha_{21}y_{1,t-1} + \alpha_{22}y_{2,t-1} \\ &\quad + \beta_{21}y_{1,t-2} + \beta_{22}y_{2,t-2} + e_{2t}, \end{aligned} \tag{15}$$

or in matrix notation,

$$Y_t = \gamma + AY_{t-1} + BY_{t-2} + e_t, \tag{16}$$

where

$$Y_t = \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix}, \quad \gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}, \quad e_t = \begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix},$$

$$A = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \quad B = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix}.$$

The vector e_t is assumed to have zero mean and covariance matrix Ω. Each term e_{it} is assumed to be independent over t and the pairs (e_{it}, e_{jk}) are independent for $t \neq k$. Given these assumptions,

$$E(Y_t \mid Y_{t-1}, Y_{t-2}) = \gamma + AY_{t-1} + BY_{t-2}, \tag{17}$$

and the conditional covariance matrix of Y_t, given Y_{t-1} and Y_{t-2}, is

$$V(Y_t \mid Y_{t-1}, Y_{t-2}) = \Omega. \tag{18}$$

As in the single variate case, to reduce the order of the process from second to first order, we must view Y_t and Y_{t-2} as jointly distributed vector random variables with Y_{t-1} fixed (nonstochastic). Recognizing the stationarity assumption and the equality of moments which have all subscripts equidistant in time,

$$\begin{aligned} E(Y_t \mid Y_{t-1}) &= \gamma + AY_{t-1} + BE(Y_{t-2} \mid Y_{t-1}) \\ &= \gamma + AY_{t-1} + BE(Y_t \mid Y_{t-1}) \\ &= (I - B)^{-1}(\gamma + AY_{t-1}). \end{aligned} \tag{19}$$

In analyzing the conditional variance, we

[3] Russell Tronstad, personal communication.

[4] Frank Novak, personal communication.

use the well-known result that if z is a vector random variable with covariance matrix Σ, and A is a matrix of fixed coefficients, then the covariance matrix of Az is $A\Sigma A'$. To obtain the conditional covariance matrix of Y_t, given Y_{t-1}, observe that $\gamma + AY_{t-1}$ will be treated as a vector of constants in (16) and thus not affect the outcome. Also, e_t is independent of Y_{t-2} for any Y_{t-1} and thus is independent of Y_{t-2} in the conditional distribution. Therefore, using an obvious notation,

$$V(Y_t \mid Y_{t-1}) = V(BY_{t-2} \mid Y_{t-1}) + V(e_t), \tag{20}$$

where $V(\cdot)$ denotes the covariance matrix. Stationarity of the vector process implies

$$V(BY_{t-2} \mid Y_{t-1}) = V(BY_t \mid Y_{t-1}),$$

which lets us write (20) as

$$V(Y_t \mid Y_{t-1}) = BV(Y_t \mid Y_{t-1})B' + \Omega. \tag{21}$$

Let W denote the unknown conditional covariance matrix $V(Y_t \mid Y_{t-1})$, then W is the unknown in the system of linear equations,

$$W - BWB' = \Omega. \tag{22}$$

Symmetry of a covariance matrix makes the number of unknowns in the $\{w_{ij}\}$ equal to $N(N+1)/2$ when the vector Y_t has N components, e.g., the two-equation system of (15) implies three unknowns, w_{11}, w_{22}, and $w_{12} = w_{21}$. With the assumption of normality, the transition probabilities for the bivariate first-order process can be obtained in a conceptually straightforward way,[5] but note that the first-order system will exhibit dependence across equations even if $\Omega = I$. The number of simultaneous equations in the $\{w_{ij}\}$ to be solved in (22) grows rapidly with the dimension of the vector Y_t in (16), i.e., the number of joint second-order autoregressive equations comparable to the pair in (15).

General Case for Continuous Random Variables

Using probability theory, the general case of reducing the second-order Markov process to the first order under any distributional form can be developed for both the case of continuous and discrete random variables. The continuous case appears to offer pedagogical advantages, while the discrete case offers computational advantages. Consequently, the continuous case is developed before presentation of the discrete case.

Consider the identity that a conditional probability distribution function (pdf) is given by the ratio of a joint pdf and a marginal pdf,

$$h(y_t \mid y_{t-1}) \equiv w(y_t, y_{t-1})/\phi(y_{t-1}), \tag{23}$$

where $w(\cdot)$ is the joint pdf of y_t and y_{t-1}, $h(\cdot)$ is the pdf of y_t conditioned on y_{t-1}, and $\phi(\cdot)$ is the marginal pdf of y_{t-1}. From the definition of marginal probability, we know that

$$\phi(y_{t-1}) \equiv \int w(y_t, y_{t-1})\, dy_t. \tag{24}$$

Substituting (24) into the right-hand side of (23) for $\phi(\cdot)$ gives

$$h(y_t \mid y_{t-1}) \equiv \frac{w(y_t, y_{t-1})}{\int w(y_t, y_{t-1})\, dy_t}. \tag{25}$$

Note that if we are given $w(\cdot)$, the desired pdf $h(\cdot)$ can be obtained using identity (25). Taking this approach, the question is how to obtain $w(\cdot)$ from $f(y_t \mid y_{t-1}, y_{t-2})$ which is the pdf associated with a second-order Markov process.

As a means of obtaining $w(\cdot)$, consider the following identity

$$g(y_t, y_{t-1}, y_{t-2}) \equiv f(y_t \mid y_{t-1}, y_{t-2})w(y_{t-1}, y_{t-2}), \tag{26}$$

where $g(\cdot)$ is the joint pdf of y_t, y_{t-1}, and y_{t-2}, and $f(\cdot)$ and $w(\cdot)$ are as defined previously. We assume time invariance of all density functions; therefore, the functional forms of $w(y_t,y_{t-1})$ and $w(y_{t-1},y_{t-2})$ are the same (although they have different arguments), and the functions $f(y_t \mid y_{t-1})$ and $f(y_{t-1} \mid y_{t-2})$ are the same but with different arguments.

Where three random variables are involved, the joint pdf for two of the three random variables can be obtained by integrating the joint pdf over the range of the third random variable; hence,

$$w(y_t, y_{t-1}) \equiv \int g(y_t, y_{t-1}, y_{t-2})\, dy_{t-2}. \tag{27}$$

Substituting (26) into the right-hand side of (27) for $g(\cdot)$ gives

[5] As in the univariate case, there is no closed-form expression in the bivariate normal distribution function. Hence, transition probabilities for an empirical problem cannot be obtained by differences in the distribution function as in (4), but instead, empirical transition probabilities must be obtained by numerical integration of the bivariate conditional probability density function with the mean given by (19) and covariance matrix W.

$$(28)\quad w(y_t, y_{t-1}) \equiv \int f(y_t \mid y_{t-1}, y_{t-2}) w(y_{t-1}, y_{t-2})\, dy_{t-2},$$

which is a functional equation that can, in principle, be solved for the unknown function $w(\cdot)$ when $f(y_t \mid y_{t-1}, y_{t-2})$ is known. Once the function $w(\cdot)$ is obtained, the pdf of interest, $h(\cdot)$, can be obtained using identity (25).

With most mathematical forms for $f(\cdot)$, analytical solution of functional equation (28) for $w(\cdot)$ is quite difficult if not impossible. Thus, we see that working with the continuous random variable version of the model is often impractical.[6]

General Case with Discrete Random Variables

In empirical applications of nearly all stochastic dynamic optimization techniques, computational considerations dictate discrete valued approximations of the state space. Hence, the discrete equivalent of (28) is pertinent for empirical application. The first step is to show how to obtain the analogue of the functional equation in (28) as a system of equations in discrete probabilities. The following discrete probabilities are defined as approximations to the continuous random variables in the last section,

$$(29)\quad \text{pr}(y_t \mid y_{t-1}) = \text{pr}(a_{k-} \le y_t < a_{k+} \mid y_{t-1} = a_j),$$

$$(30)\quad \text{pr}(y_t, y_{t-1}) = \text{pr}(a_{j-} \le y_t < a_{j+} \text{ and } a_{i-} \le y_{t-1} < a_{i+}),$$

$$(31)\quad \text{pr}(y_t \mid y_{t-1}, y_{t-2}) = \text{pr}(a_{k-} \le y_t < a_{k+} \mid y_{t-1} = a_j, y_{t-2} = a_i),$$

where pr$(\cdot)$ denotes probability of the outcome; k, j, and i are indices for the discrete states assumed by y; a_i is the midpoint while a_{i-} and a_{i+} are the lower and upper boundaries of state i, respectively. We use midpoints instead of the intervals as an approximation for given states of the process at a point in time because that is commonly done in computing transition probabilities in applied work. If the intervals are not "too large," this approximation is satisfactory, but a serious bias can be introduced when the intervals become so large that there is a high probability of staying in the same interval (see Burt, p. 390, for a method to overcome this problem).

Note that (31) can be obtained from the continuous representation of the conditional pdf for the second-order process from

$$(32)\quad \text{pr}(y_t \mid y_{t-1}, y_{t-2}) = \int_{a_{k-}}^{a_{k+}} f(y_t \mid y_{t-1}, y_{t-2})\, dy_t .$$

The discrete representation of (28) is

$$(33)\quad \text{pr}(y_t, y_{t-1}) = \sum_{y_{t-2}} \text{pr}(y_t \mid y_{t-1}, y_{t-2}) \text{pr}(y_{t-1}, y_{t-2}),$$

where

$$\text{pr}(y_t, y_{t-1}) = \text{pr}(y_{t-1}, y_{t-2})$$

by virtue of stationarity in the stochastic process.

A special characteristic of Markov chains allows us to obtain pr(y_t, y_{t-1}) by solving a set of simultaneous linear equations for the discrete case as contrasted to solving the functional equation (28) in the continuous case. Before considering the set of linear equations, it is instructive to consider the classical "transition matrix" representation of a second-order Markov process.

The transition matrix of probabilities for a first-order process gives the conditional probabilities of going from the ith to the jth state, with the passage of one unit of time. If the continuous variable y is divided into N discrete intervals, the transition matrix will be of dimension N. For a second-order process, the given outcomes are described by a pair of integers, one each for y_{t-1} and y_{t-2} where t is the year of interest for a random outcome. We now need a triple subscript on the elements of the transition matrix, say p_{ijk}, where $y_{t-2} = a_i$, $y_{t-1} = a_j$, and $y_t = a_k$, and the transition matrix is of dimension N^2. With the progress of one more time unit, let $y_{t+1} = a_m$. The transition probabilities, with actual outcomes substituted for the sequence of events, for periods t and $t+1$ are p_{ijk} and p_{jkm}, respectively. Note how the first subscript is dropped, the second and third become the first and second, and the third subscript takes on a new integer in going from period t to $t+1$. However, the basic structure of the transition matrix is not dependent on time per se, which reflects the stationarity of the stochastic process.

[6] Subtle problems of existence also have to be considered with respect to $w(\cdot)$. This is handled in the next section for discrete processes by assuming ergodicity.

Elements of the transition matrix P are defined by

$$(34)\quad p_{ijk} = \text{pr}(a_{k-} \leq y_t < a_{k+} \mid y_{t-1} = a_j, y_{t-2} = a_i).$$

Note that the transition matrix of the second-order Markov process, P, can be constructed from $f(y_t \mid y_{t-1}, y_{t-2})$ using (32). We are now ready to derive the set of unconditional probabilities, $\text{pr}(y_t, y_{t-1})$, by solving (33). For a completely ergodic process,[7] we can obtain this set of joint probabilities as follows. Following the standard Markov chain theory (Howard), we define a state probability, $\pi_{ij}(t)$, for a second-order process as the probability that the system will occupy the state given by the combination $y_{t-2} = a_i$, $y_{t-1} = a_j$, after t transitions, given that its state is known at $t = 0$. It follows from this definition that

$$(35)\quad \sum_i \sum_j \pi_{ij}(t) = 1,$$

and that

$$(36)\quad \pi_{jk}(t+1) = \sum_i \pi_{ij}(t) p_{ijk}.$$

If we define a row vector of state probabilities, $\Pi(t)$ with element $\pi_{ij}(t)$, then in matrix form we have the recursive relationship

$$(37)\quad \Pi(t+1) = \Pi(t)P.$$

For a completely ergodic process, the vector $\Pi(t)$ can be shown to approach a limit as t approaches infinity (Howard). In Markov chain literature, this limiting vector, say Π, is referred to as the limiting or steady-state probability vector. From (37), it follows that this limiting vector is given by the equation

$$(38)\quad \Pi = \Pi P,$$

which is equivalent to (33) in a different notation.

Equation (38), along with the constraint that the elements of the vector Π sum to one, can be used to solve for the limiting state probabilities. It can be shown that the matrix $(I - P)$, where I is the identity matrix, has rank one less than the dimension of P. Therefore, one arbitrary row of (38) can be replaced by (35) and the unique solution for Π calculated numerically from the system of linear equations. The steady-state equation, (38), is the discrete variable analogue of (28).

[7] An ergodic process is defined to be any Markov process whose limiting state probability distribution is independent of starting conditions (Howard, p. 6).

It is crucial here to note that the vector of limiting or steady-state probabilities does not depend on the starting state of the system. Thus, it can be seen that Π is a vector of unconditional state probabilities, while the original transition matrix, P, gives conditional state probabilities. Since Π is unconditional, it can be seen that the elements of Π are the joint probabilities, (30).

Conditional probabilities for the process reduced to first order are

$$(39)\quad \text{pr}(y_t \mid y_{t-1}) = \frac{\text{pr}(y_t, y_{t-1})}{\sum_{y_t} \text{pr}(y_t, y_{t-1})},$$

which is the discrete counterpart of (25) in the continuous variable case. Using the notation p_{jk} for the conditional probability of outcome k in period t, given the outcome j in period $t - 1$, (39) can be written

$$(40)\quad p_{jk} = \pi_{jk} / \sum_k \pi_{jk},$$

where the $\{\pi_{jk}\}$ are elements of the steady-state probability vector Π.

Computational steps for reducing a discrete second-order process to first order are: (*a*) calculate the steady-state probability vector Π for the second-order process, and (*b*) apply (40) for $j = 1, 2, \ldots, N$, where N is the number of discrete intervals on y used in the discrete approximation to the continuous variable process. These results readily extend to higher-order processes as well.

For example, consider a third-order process where the transition probabilities have four subscripts, p_{ijkm}, and the steady-state probabilities have three subscripts, π_{ijk}. Reduction from a third-order process to a second order is accomplished by means of the formula,

$$(41)\quad p_{ijk} = \pi_{ijk} / \sum_k \pi_{ijk},$$

where p_{ijk} is the conditional probability of outcome k in period t, given outcomes j and i in periods $t - 1$ and $t - 2$, respectively. Reduction to a first-order process uses the formula

$$(42)\quad p_{jk} = \left(\sum_i \pi_{ijk}\right) / \sum_k \left(\sum_i \pi_{ijk}\right),$$

because the numerator is the joint probability of outcomes j and k in periods $t - 1$ and t, while the denominator is the marginal probability of outcome j in period $t - 1$.

Concluding Remarks

The methods of discrete distribution approximations for reduction of higher-order processes to lower order are particularly useful if one does not wish to assume normality of the stochastic process. In such cases, the hyperbolic tangent method of Taylor can be applied to estimate a closed functional form for the conditional cumulative distribution function of the process. This flexible approach permits the order of the process to be arbitrary, and computation of the matrix of transition probabilities is practical, thus permitting application of the results above to get a lower-order process (first order in particular).

As stated earlier, the order of the process must be specified correctly in performing the statistical estimation of unknown parameters, whether using the linear stochastic difference equation model or an alternative like the polynomial imbedded hyperbolic tangent method. However, the linear difference equation framework might be used advantageously jointly with the hyperbolic tangent method because of the ease with which the former can be used to identify the maximum order of the stochastic process. The latter would serve as the final model after the order of dependence had been determined, at least tentatively so.

Estimates of the amount of information sacrificed in reducing the process to a lower order in the discrete variable case can be estimated by ratios of variances as was suggested for the linear difference equation model. Variances can be calculated from the alternative reduced-order processes for comparison with the variance of the highest-order process. Using the third-order Markov chain in the last section, the least residual variance is associated with the set of probabilities $\{p_{ijk}\}$ which utilize the full set of information on the history of the process, while the least information (most residual variance) is associated with the first-order set of probabilities in (42). Each conditional probability distribution permits calculation of the conditional mean and variance of the stochastic process, and the "total" variance is given by the steady-state probabilities. These different measures of variance can be used to construct variance ratios paralleling those developed for the linear difference equation model.

It is noted that the assumption of an infinite ergodic stochastic process does not prohibit application of these procedures to a finite horizon DP problem. The infinite process only has to exist in principle to justify the methods used to discard information and derive "less conditional" probability distributions. An intraseasonal component imbedded in the process should not cause a problem, but it effectively introduces another state variable because the seasonal period must be identifiable at each stage.

[*Received February 1989; final revision received May 1989.*]

References

Anderson, T. W. *An Introduction to Multivariate Statistical Analysis,* 2nd ed. New York: John Wiley and Sons, 1984.

Bellman, R. *Dynamic Programming.* Princeton: Princeton University Press, 1957.

Box, G. E. P., and G. M. Jenkins. *Time Series Analysis: Forecasting and Control,* rev. ed. San Francisco: Holden and Day, 1976.

Burt, O. R. "Dynamic Programming: Has Its Day Arrived?" *West. J. Agr. Econ.* 7(1982):381–93.

Danielson, J. G. *Optimal Crop Sequences to Control Cephalosporium Stripe in Winter Wheat.* MS thesis, Montana State University, 1987.

Dreyfuss, S. E., and A. M. Law. *The Art and Theory of Dynamic Programming.* New York: Academic Press, 1977.

Howard, R. A. *Dynamic Programming and Markov Processes.* Cambridge MA: MIT Press, 1960.

Nelson, C. R. *Applied Time Series Analysis for Managerial Forecasting.* San Francisco: Holden and Day, 1973.

Novak, F. Assistant professor of agricultural economics, University of Alberta. Personal communication.

Phipps, T. T. "Land Prices and Farm-Based Returns." *Amer. J. Agr. Econ.* 66(1984):422–29.

Schnitkey, G. *A Dynamic Programming Analysis of Optimal Farmland Investment Decisions: An Application to Central Illinois High-Quality Farmland.* Ph.D. thesis, University of Illinois, 1987.

Taylor, C. R. "A Flexible Method for Empirically Estimating Probability Functions." *West. J. Agr. Econ.* 9(1984):66–76.

Tronstad, R. Assistant professor of agricultural economics, University of Arizona. Personal communication.