

On Estimation and Prediction Procedures for AR(1) Models with Power Transformation

J. C. LEE* and S. L. TSAO

National Chiao Tung University, Hsinchu, Taiwan

and

AT&T Bell Laboratories, Holmdel, NJ, U.S.A.

ABSTRACT

The power transformation of Box and Cox (1964) has been shown to be quite useful in short-term forecasting for the linear regression model with AR(1) dependence structure (see, for example, Lee and Lu, 1987, 1989). It is crucial to have good estimates of the power transformation and serial correlation parameters, because they form the basis for estimating other parameters and predicting future observations. The prediction of future observations is the main focus of this paper. We propose to estimate these two parameters by minimizing the mean squared prediction errors. These estimates and the corresponding predictions compare favourably, via retests and simulated data, with those obtained by the maximum likelihood method. Similar results are also demonstrated in the repeated measurements setting.

KEY WORDS AR(1) dependence Box–Cox transformation Maximum likelihood Minimum prediction errors Simulations Technology penetration

INTRODUCTION

Through scientific discoveries and evolution, new products spread and grow to replace existing ones. This process is called technological substitution. Thus, as technology advances, steam/motor merchant vessels replace sail merchant vessels, colour televisions replace black-and-white televisions, digital telephone switching systems replace analog switching systems. The rate of the change, called technology penetration, is defined as the ratio of the number of new products to the combined total of new and old products. The penetration data tend to follow an S-shaped curve and exhibit very strong positive first-order autoregressive (or AR(1)) dependence. The S-shaped curve is typical in many types of growth curve models, such as Gompertz and logistic, among others. However, a typical growth curve such as Gompertz

* Research supported in part by NSC grant 82-208-M009-054

or logistic is not quite adequate in forecasting penetration at a future time. To illustrate the point, let $F(t)$ be the technology penetration at time t . Then if a logistic curve is assumed, we expect $\log(F(t)/(1 - F(t)))$ to be linear in time, i.e. a linear growth function. But when dealing with real data, we often found that the more flexible transformation of $Y_t = F(t)/(1 - F(t))$ gives much superior predictive accuracies. This is, perhaps, primarily due to the fact that the linearity assumption for the growth function can be enhanced considerably with the Box-Cox power transformation, along with the incorporation into the model of the proper dependence structure among the observations. For more details, see Lee and Lu (1987). Nevertheless, the maximum likelihood estimation method used by Lee and Lu (1987) could produce undesirable estimates to result in a poor forecast when the sample size is small. Therefore we propose in this paper an intuitively appealing method to deal with the prediction of future observations in small samples.

A family of data-based transformed models for forecasting technological substitutions has been empirically shown to be quite useful for short-term forecasts (see Lee and Lu, 1987, 1989). These models are more general and include as special cases the four well-known S-shaped growth curve models: Gompertz, logistic, normal, and Weibull. The transformation needed depends on the data at hand. Basically, the widely used Box-Cox (1964) transformation is applied to the AR(1)-dependent data. More specifically, let Y_1, \dots, Y_T be a set of T observations. The transformation is defined as

$$Y_t^{(\lambda)} = \begin{cases} \frac{(Y_t + \gamma)^\lambda - 1}{\lambda} & \text{when } \lambda \neq 0 \\ \log(Y_t + \gamma) & \text{when } \lambda = 0 \end{cases} \quad (1)$$

where γ is a known constant such that $Y_t + \gamma > 0$ for all t . In practice, $\gamma = 0$ if all Y_t 's are positive. Furthermore, $Y_t^{(\lambda)}$ ($t = 1, \dots, T$) are assumed to be normally distributed with the mean linear in the regression parameters and the covariance matrix $\Sigma = \sigma^2 V_{11}$, $V_{11} = (c_{ab})$, $c_{ab} = \rho^{|a-b|}$, for $a, b = 1, \dots, T$, and $-1 < \rho < 1$. In other words,

$$\mathbf{Y}^{(\lambda)} = (Y_1^{(\lambda)}, \dots, Y_T^{(\lambda)})' = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2)$$

where $\mathbf{X}\boldsymbol{\beta}$ is the growth function, $\boldsymbol{\beta} = (\alpha, \beta)'$, \mathbf{X} is the design matrix with 1's in the first column and t_1, \dots, t_T or $\log(t_1), \dots, \log(t_T)$ in the second column, depending on the growth curve model being considered (see Lee and Lu, 1987). The vector of disturbance terms $\boldsymbol{\varepsilon}$ is assumed to distribute as a multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2 V_{11}$. If a higher-degree polynomial is assumed for the growth function, the dimensions of \mathbf{X} and $\boldsymbol{\beta}$ are adjusted accordingly. For example, if a second-degree polynomial in time is assumed for the growth function, then the third column of \mathbf{X} is t_1^2 to t_T^2 . Model (2) has also been proved useful by Fiebig and Bewley (1987) in forecasting demand for international communications services as well as for other types of data.

The Box-Cox transformation has been considered for the general time series by several authors, including Granger and Newbold (1976), Hopwood *et al.* (1984), and Pankratz and Dudley (1987), among others. However, none of the above authors discussed the estimation of parameters from a prediction point of view. Model (2) has been extended to the situation in which several concurrent short time series are available (see Keramidias and Lee, 1990). We will discuss this extended model in more detail in the next section while the rest of this section is devoted to the single-series situation.

Let \mathbf{y} , of dimension $k \times 1$, be the vector of future observations to be predicted on the basis of the past observations $\mathbf{Y} = (Y_1, \dots, Y_T)'$ such that $E(\mathbf{y}^{(\lambda)}) = \mathbf{x}\boldsymbol{\beta}$ where \mathbf{x} is the design matrix corresponding to the future observations. Furthermore, $\text{Cov}(\mathbf{Y}^{(\lambda)'}, \mathbf{y}^{(\lambda)'}) = \sigma^2 (c_{ab})$, where $c_{ab} = \rho^{|a-b|}$. When the unknown parameters are replaced by their maximum likelihood

estimates (MLEs), the forecasts of y , which is obtained from the conditional expectation of $y^{(\lambda)}$ given Y , can be written as

$$\hat{y} = \begin{cases} (1 + \hat{\lambda}[\mathbf{x}\hat{\beta} + \hat{\eta}(Y_T^{\hat{\lambda}}) - X_T\hat{\beta}])^{1/\hat{\lambda}} - \gamma & \text{when } \hat{\lambda} \neq 0 \\ \exp[\mathbf{x}\hat{\beta} + \hat{\eta}(Y_T^{\hat{\lambda}}) - X_T\hat{\beta}] - \gamma & \text{when } \hat{\lambda} = 0 \end{cases} \quad (3)$$

where $\hat{\beta} = (\mathbf{X}'\hat{\mathbf{V}}_{11}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}_{11}^{-1}\mathbf{Y}^{(\hat{\lambda})}$, $\hat{\eta} = (\hat{\rho}, \hat{\rho}^2, \dots, \hat{\rho}^k)'$, $\mathbf{1} = (1, \dots, 1)'$, $\hat{\mathbf{V}}_{11} = (\hat{c}_{ab})$, $\hat{c}_{ab} = \hat{\rho}^{|a-b|}$, X_T and $Y_T^{\hat{\lambda}}$ are the T th row of \mathbf{X} and $\mathbf{Y}^{(\hat{\lambda})}$, respectively, and $\hat{\lambda}$ and $\hat{\rho}$ are the MLEs of λ and ρ , respectively; $\mathbf{z} = (z_1, \dots, z_k)'$, $\mathbf{z}^d = (z_1^d, \dots, z_k^d)'$, for any constant d , and $\exp(\mathbf{z}) = (\exp(z_1), \dots, \exp(z_k))'$. For the remainder of the paper we will concentrate on the special case in which $k = 1$, i.e. we are interested only in a one-step-ahead forecast. This is due to our belief that a one-step-ahead forecast is less speculative and hence is more reliable for the purpose of comparing different methods.

The prediction of future observations y is the main focus of this paper. The forecasts \hat{y} , of y , depend very crucially on the quality of the estimates $\hat{\lambda}$ and $\hat{\rho}$. The MLEs of these two parameters have been used with some success by Lee and Lu (1987). Nevertheless, the method could produce undesirable estimates when the sample size is relatively small, as the MLEs are optimal in large samples only. This will in turn render poor forecasts in small samples.

The purpose of this paper is to propose an intuitively appealing method, minimum prediction error (MPE), for estimating these two parameters and to explore possible advantages of using this estimation method in the prediction of future observations. The MPE method, formulated to some extent in the spirit of Rissanen's (1986) accumulated prediction errors method and Lee's (1988) minimum prediction variance estimate, is more sensitive than the traditional MLEs to the prediction of future observations. Moreover, it is less dependent on the distributional assumption imposed on the observations, as it minimizes the mean squared prediction errors.

Since the MPE method will not give closed-form representations for the parameter estimates, only empirical comparisons of the proposed method with the maximum likelihood method will be made in terms of the predictive accuracy of future observations. These comparisons are made in both real and simulated data. In addition, empirical comparisons for the parameter estimates by both methods will also be made via simulation.

Details of the proposed method are given in the next section. Empirical comparisons in both real data and simulations are presented in the third section. Conclusions are given in the final section.

MINIMUM PREDICTION ERROR ESTIMATES

In this section we will consider the MPE method for a single series as well as for several concurrent short series, with the emphasis on the single-series situation.

A single series

In equation (3) the MLpEs of λ and ρ are obtained by maximizing the profile log-likelihood function (see Lee and Lu, 1987), which is defined as

$$l(\lambda, \rho) = -\frac{T}{2} \log \hat{\sigma}^2(\lambda, \rho) + (\lambda - 1) \sum_{t=1}^T \log(Y_t + \gamma) - \frac{(T-1)}{2} \log(1 - \rho^2) \quad (4)$$

where $\hat{\sigma}^2(\lambda, \rho) = (1/T)(\mathbf{Y}^{(\lambda)} - \mathbf{X}\hat{\beta}(\lambda, \rho))' \mathbf{V}_{11}^{-1}(\mathbf{Y}^{(\lambda)} - \mathbf{X}\hat{\beta}(\lambda, \rho))$ and $\hat{\beta}(\lambda, \rho)$ is the $\hat{\beta}$, given in equation (3), with $\hat{\lambda}$ and $\hat{\rho}$ replaced by λ and ρ , respectively. The reason for considering only the profile likelihood function of λ and ρ is that, given λ and ρ , $\hat{\beta}(\lambda, \rho)$ and $\hat{\sigma}^2(\lambda, \rho)$ are

functions of λ and ρ only. Once $\hat{\lambda}$ and $\hat{\rho}$ are obtained by maximizing $l(\lambda, \rho)$, then $\hat{\beta}(\hat{\lambda}, \hat{\rho})$, $\hat{\sigma}^2(\hat{\lambda}, \hat{\rho})$, $\hat{\lambda}$ and $\hat{\rho}$ are the MLEs of β , σ^2 , λ , and ρ .

Instead of using the MLEs for these two parameters, we can treat \hat{y} , defined in equation (3), as a function of λ and ρ , whose values are obtained by minimizing the mean squared prediction errors. More specifically, the predictions are made for Y_t , $t = 3, \dots, T$, since we need at least two observations to estimate parameters in a simple linear model. In predicting Y_t , the sample is $\mathbf{Y}_{t-1} = (Y_1, \dots, Y_{t-1})'$. Thus, the predictions are made sequentially and the sample on which each prediction is based increases in size one at a time. The approach is prequential in nature in that predictions of future observations are made sequentially (see Dawid, 1984; Keramidas and Lee, 1990). The objective function to be minimized is

$$\text{MSE}_1(\lambda, \rho) = \sum_{t=3}^T (Y_t - \hat{Y}_t)^2 / (T - 2) \tag{5}$$

where \hat{Y}_t , is defined in exactly the same form as \hat{y} , given in equation (3), with $k = 1$, T replaced by $t - 1$, \mathbf{x} by X_t , and \mathbf{X} by \mathbf{X}_{t-1} , the first $t - 1$ rows of \mathbf{X} .

The minimization of $\text{MSE}_1(\lambda, \rho)$ with respect to λ and ρ can be facilitated by an optimization routine in the Port 3 Mathematical Subroutine Library (1984). We will use the version requiring both the first and the second derivatives of the objective function. These derivatives are provided in the Appendix.

Several concurrent short series

When several short time series are available, Keramidas and Lee (1990) extended model (2) to the general growth curve model with serial covariance structure for forecasting technological substitutions. The model is

$$\mathbf{Y}^{(\lambda)} = \mathbf{X} \tau \mathbf{A} + \boldsymbol{\varepsilon} \tag{6}$$

$\begin{matrix} T \times N & T \times m & m \times r & r \times N & T \times N \end{matrix}$

where $\mathbf{Y}^{(\lambda)} = (\mathbf{Y}_1^{(\lambda)}, \dots, \mathbf{Y}_N^{(\lambda)})$, $\mathbf{Y}_i^{(\lambda)} = (Y_{1i}^{(\lambda)}, \dots, Y_{Ti}^{(\lambda)})'$ with $Y_{ji}^{(\lambda)}$ defined as in model (1); τ is unknown, \mathbf{X} is the known design matrix characterizing the type of growth function, and \mathbf{A} is the known design matrix characterizing the grouping among the N independent series into r different curves. Furthermore, the columns of $\boldsymbol{\varepsilon}$ are independent T -variate normal with mean vector $\mathbf{0}$ and common covariance matrix $\sigma^2 V_{11}$. For the rest of the paper we will consider a particularly practical situation in which there are only r different λ 's. This means that we will apply the same power transformation to each of the observations in the same group. The MLpEs of λ and ρ have been considered by Keramidas and Lee (1990).

For the MPE estimates of these parameters, we consider the following extended prediction problem. Let \mathbf{y} , of dimension $k \times n$, be a set of $n (\leq N)$ future k -dimensional observations whose previous T -dimensional observations are a subset of \mathbf{Y} . Let \mathbf{x} , of dimension $k \times m$, be the design matrix corresponding to \mathbf{y} , $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_N)$, $\mathbf{A} = (\mathbf{A}_1, \dots, \mathbf{A}_N)$, $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ and assume that for $i \leq n$,

$$E \begin{bmatrix} \mathbf{Y}^{(\lambda_i)} \\ \mathbf{y}_i^{(\lambda_i)} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{x} \end{bmatrix} \tau \mathbf{A}_i \tag{7}$$

$$\text{cov} \begin{bmatrix} \mathbf{Y}^{(\lambda_i)} \\ \mathbf{y}_i^{(\lambda_i)} \end{bmatrix} = \sigma^2 \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \tag{8}$$

where

$$\begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} = (\rho^{|a-b|})$$

for $a, b = 1, \dots, T + k$. When the MLEs of λ and ρ are used in place of these unknown parameters, the forecasts of y_i is the same as \hat{y} , given in equation (3), with $\hat{\lambda}$ replaced by $\hat{\lambda}_i$, $\hat{\beta}$ by $\hat{\tau}A_i$ where $\hat{\tau} = (\mathbf{X}'\hat{\mathbf{V}}_{11}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}_{11}^{-1}\mathbf{Y}(\hat{\mathbf{A}}\mathbf{A}\mathbf{A}')^{-1}$.

As in the single-series situation, we can treat the predictor \hat{y}_i as a function of λ_i and ρ , whose estimates are derived by minimizing the mean squared prediction errors. Thus, the objective function to be minimized is

$$\text{MSE}_2(\lambda, \rho) = \sum_{i=1}^N \sum_{j=m+1}^T (Y_{ij} - \hat{Y}_{ij})^2 / (T - m)N \tag{9}$$

The minimization of $\text{MSE}_2(\lambda, \rho)$ is similar to that of $\text{MSE}_1(\lambda, \rho)$. The derivatives of $\text{MSE}_2(\lambda, \rho)$ can be obtained in exactly the same manner as those of $\text{MSE}_1(\lambda, \rho)$ and hence are omitted.

EMPIRICAL COMPARISONS

In this section we will compare empirically the predictive accuracy of the linear models in which the power transformation and serial correlation parameters are estimated by the maximum likelihood (ML) and the MPE methods. For real data examples, we will consider four sets of observations on technology penetration,¹ three in telephone-switching systems and one in television receivers. Since in the area of technological substitutions the logistic growth curve (or Fisher–Pry model, 1971) is the most popular, we will confine our attention to this particular model. For the logistic curve, if $F(t)$ is the new technology penetration at time t , then setting $Y_t = F(t)/(1 - F(t))$ we have $Y_t^{(\lambda)} = \log(F(t)/(1 - F(t)))$ when $\lambda = 0$, which is exactly the logistic or the Fisher–Pry model if the transformed variables are assumed independent. For the simulation study, however, we will be dealing with Y_t directly and hence a particular growth curve is irrelevant to our discussion.

In order to compare the predictive accuracy when the unknown power transformation and serial correlation parameters are estimated by the ML and the MPE methods, we will employ three different measures: the mean absolute deviation (MAD), the mean absolute relative deviation (MARD), and the fraction of accurate prediction (FAP). The MARD is perhaps a better measure than MAD for real data examples because the time series observations are monotonically increasing and the comparison of forecasts is conducted in a sequential manner. On the other hand, for the simulation study, we will fix the sample size and hence only MAD and FAP are employed for the comparison purpose. The definition of FAP is simply the fraction with more accurate forecast for a given method. We next define the other two measures.

Let $\hat{F}(t)$ be the forecast of the new technology penetration $F(t)$ and $e(t) = F(t) - \hat{F}(t)$, for $t = 1, \dots, a$. Then the MAD and MARD are defined as

$$\text{MAD} = \sum_{t=1}^a |e(t)| / a \tag{10}$$

$$\text{MARD} = \sum_{t=1}^a (|e(t)| / F(t)) / a \tag{11}$$

¹ The technology penetration is defined as the ratio of the number of new products to the combined total of new and old products.

The above definitions are relevant for real data examples. For the simulation the corresponding definitions are easily seen by replacing $F(t)$ by Y_t and $\hat{F}(t)$ by \hat{Y}_t .

Real data with a single series

Two sets of technology penetration data, analysed by Lee and Lu (1987, 1989), are listed in Table I. We will consider the empirical comparison of the predictive accuracy in one-step-ahead forecasts between the two estimation methods for the power transformation and serial correlation parameters. Also, in the comparison the initial sample size is five, instead of eight as in the previous studies. The feasibility of estimating these two parameters in small samples is one of the features of the proposed method.

Table II shows the comparison of the predictive accuracy in terms of the three measures mentioned earlier. In terms of MARD, the improvements (IMP) of the forecasts based on the MPE method over the ML method for these two parameters are 27% for the TV data and 51% for the telephone data. The corresponding improvements in terms of MAD are 11% and 33%,

Table I. Technology penetration data

Colour TV receivers		Telephone-switching systems	
Year	Penetrations	Year	Penetrations
1955	0.00016	1965	0.00009
1956	0.00052	1966	0.00015
1957	0.00219	1967	0.00157
1958	0.00394	1968	0.00707
1959	0.00569	1969	0.01205
1960	0.00743	1970	0.02234
1961	0.00943	1971	0.03578
1962	0.01208	1972	0.05746
1963	0.01889	1973	0.08495
1964	0.03120	1974	0.11700
1965	0.05332	1975	0.14995
1966	0.09694	1976	0.18482
1967	0.16325	1977	0.23308
1968	0.24175	1978	0.28858
1969	0.32034	1979	0.35238
1970	0.35744	1980	0.42244
1971	0.41015	1981	0.47951
1972	0.48631		
1973	0.55355		
1974	0.62251		
1975	0.68394		
1976	0.73606		
1977	0.77065		
1978	0.77984		
1979	0.80926		
1980	0.83028		
1981	0.82916		
1982	0.87595		
1983	0.88703		
1984	0.90465		
1985	0.91519		

Table II. Comparison of one-step-ahead forecast accuracy for two single series

	Colour TV receivers			Telephone-switching systems		
	MPE	ML	IMP(%)	MPE	ML	IMP(%)
MAD	0.0112	0.0126	11	0.0061	0.0090	33
MARD	0.0586	0.0800	27	0.0383	0.0788	51
FAP(%)	73	27		67	33	

respectively. Meanwhile, for the MPE method, the FAP achieves 73% for the TV and 67% for the telephone data. These improvements are indeed impressive.

Real data with several concurrent short series

The two data sets (Regions A and B), given in Tables 1 and 4 of Keramidas and Lee (1988), are used here for the comparison of the predictive accuracy of the two estimation methods for the power transformation and serial correlation parameters. Region B was further analysed in Keramidas and Lee (1990).

In the comparison the initial sample sizes for Regions A and B are 3 and 4, respectively. The higher initial sample size for Region B reflects the inclusion of the quadratic growth function. In Region A only a linear growth function is used while both the linear and the quadratic growth functions are explored for Region B. It is found that only the linear growth function shows some improvement for the MPE estimates over the MLEs for Region B. Table III exhibits the comparison of the predictive accuracy in terms of MAD, MARD, and FAP. There are some improvement of the forecasts based on the MPE Method over the ML method. The improvement, however, is not as pronounced as in the single series studied earlier.

Simulation

The main focus of this section is on the simulation results for a single series. We will consider the comparison of the predictive accuracy of the forecasts based on both estimation methods. In addition, we will also consider the comparison of the estimates of the power transformation and serial correlation parameters.

In the simulation for a single series we have

$$Y_t^{(\lambda)} = \alpha + \beta t + \varepsilon_t \tag{12}$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + \delta_t \tag{13}$$

where the disturbance terms $\delta_t (t = 1, \dots, T)$ are independent normal, Student's t , or gamma,

Table III. Comparison of one-step-ahead forecast accuracy for two sets of concurrent short series

	Region A			Region B		
	MPE	ML	IMP(%)	MPE	ML	IMP(%)
MAD	0.0151	0.0198	24	0.0199	0.0216	8
MARD	0.0491	0.0336	32	0.0589	0.0620	5
FAP(%)	53	47		65	35	

Table IV. Comparisons of one-step-ahead forecasts and parameter estimates for simulated data

	$N(0, 1)$	t_{10}	$G(1, 1)$
MAD-ML	0.01597	0.01573	0.01556
MPE	0.01482	0.01482	0.01448
IMP(%)	7	6	7
FAP(%)	65	60	65
Mean (ρ)-ML	0.346	0.346	0.360
MPE	0.541	0.539	0.550
Var (ρ)-ML	0.053	0.051	0.047
MPE	0.072	0.069	0.068
Mean (λ)-ML	0.9985	1.0024	0.9951
MPE	1.0001	1.0038	0.9958
Var (λ)-ML	0.00095	0.00185	0.00189
MPE	0.00088	0.00174	0.00168

each with mean σ and variance σ^2 . Also, $G(\theta_1, \theta_2)$ denotes the gamma distribution with the shape parameter θ_1 and the scale parameter $1/\theta_2$, and t_ν denotes the Student's t distribution with ν degrees of freedom. In the simulation, we set $\alpha = 0$, $\beta = 0.2$, $\gamma = 0$, $\sigma = 0.03$, $\lambda = 1$, $\rho = 0.85$, and $T = 20$. The number of realizations of model (12) is 1000 for each distribution. Forecasts up to five steps ahead are performed and compared with the actuals. A summary of the comparisons of one-step-ahead forecast accuracy and parameter estimates are given in Table IV.

From this table it is evident that some improvement in one-step-ahead forecast accuracy has been achieved in terms of MAD and FAP. Improvements for higher step forecasts are similar. Also, there is clearly a significant improvement of the MPE estimate of the serial correlation over the MLE. This means that the MPE method could provide a viable alternative to the MLE of the serial correlation alone even when no power transformation is applied to a linear model with AR(1) dependence structure. This special situation is very important because the model is useful in time-series data. A more detailed comparison of the estimation of the serial correlation parameter by both methods is depicted in Figure 1 in which the distributions of the 1000 estimates of ρ by both estimation methods for normal, t_{10} , and $G(1, 1)$ are shown. First, the distributions of $\hat{\rho}$, for a given estimation method, is essentially independent of the distributional assumption on the disturbance terms δ_t . However, there is a significant difference in these distributions between the two estimation methods. Even though the true value of ρ is 0.85, its ML estimate never goes beyond 0.8, while the MPE estimate of ρ is between -1 and 1. This explains why the variance of the MLEs is smaller than that of the MPE estimates.

Similar comparison of the estimation of the power transformation parameter by ML and MPE methods is depicted in Figure 2. There seems to be no significant difference between the two methods in estimating this parameter.

We note that when the power transformation parameter λ is fixed at either 0 or 1 the results for the normal distribution are not much different from those presented in Table IV. Similar results are obtained when $\lambda = 0$ but estimates of both λ and ρ are made for the purpose of forecasting.

If ρ is restricted to $[0, 1)$, a possible situation in which the serial correlation is non-negative, the results for the one-step-ahead forecasts and parameter estimates are similar to those presented in Table IV. In general, the mean values of $\hat{\rho}$ are larger while the variances are

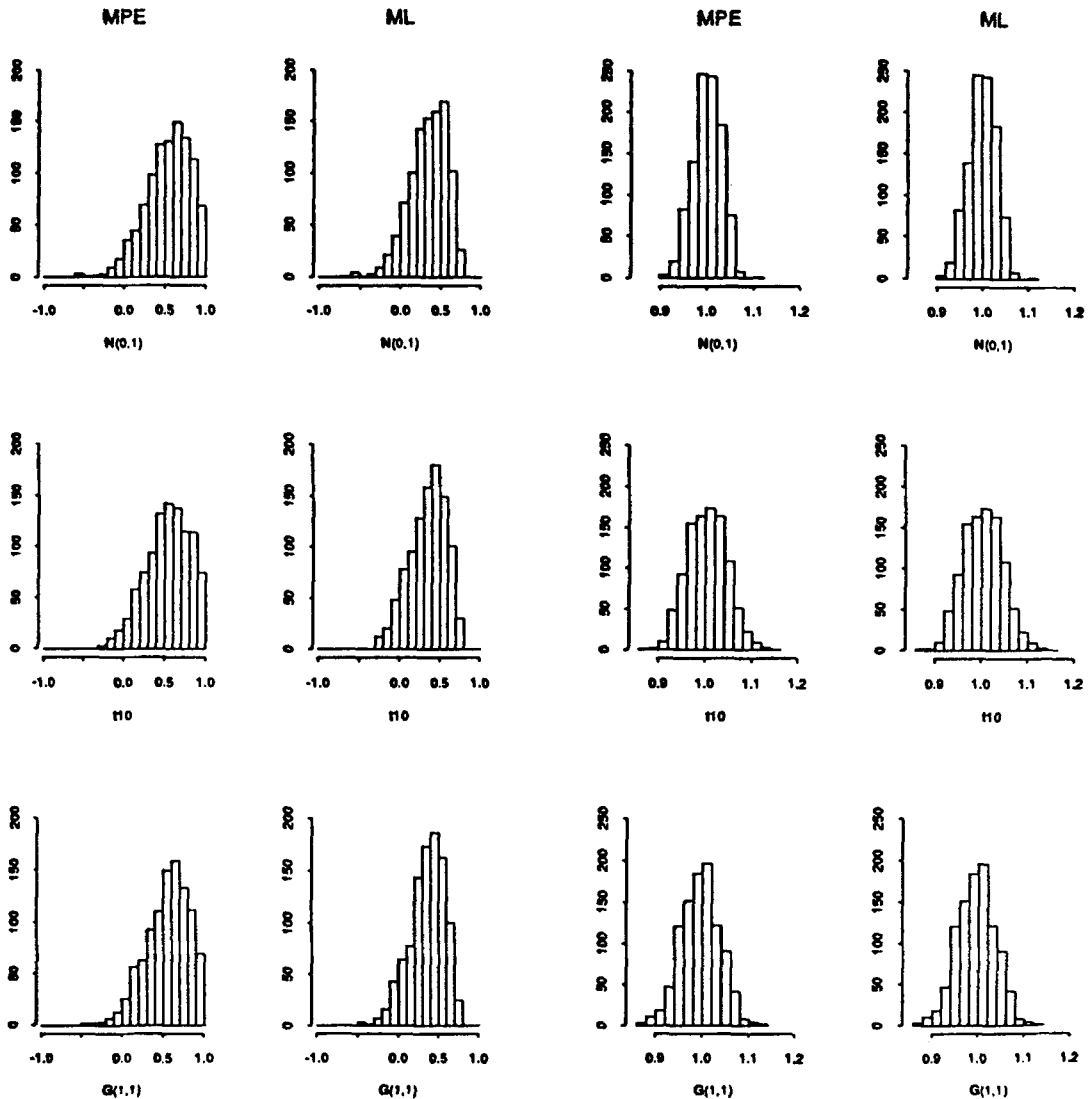


Figure 1. Distributions of $\hat{\rho}$ by MPE and ML methods

Figure 2. Distributions of $\hat{\lambda}$ by MPE and ML methods

slightly smaller, because the negative values of $\hat{\rho}$ are not allowed. The decrease in the size of the variance of $\hat{\rho}$ for each distribution is more appreciable in the ML method than in the MPE method, which may be an argument for considering only the nonnegative value for ρ , when the ML method is used.

CONCLUSIONS

From the real and simulated data we found that the minimum prediction error estimates of the power transformation and serial correlation parameters are viable alternatives to the MLEs

of these two parameters. This is especially true with a single time series when the sample size is relatively small. For the situation in which several concurrent short time series are available, the results are still encouraging, even though they are not as convincing. Finally, it is emphasized that although all the real data are from the area of technological substitutions, the proposed method is useful for any type of data in which a power transformation of the dependent variable can be applied to a linear model with AR(1) dependence structure. The method is, of course, equally suitable when no power transformation is needed.

APPENDIX: THE DERIVATIVES OF $MSE_1(\lambda, \rho)$

In this appendix we will give the first and the second derivatives of the objective function $MSE_1(\lambda, \rho)$ with respect to λ and ρ .

$$\frac{\partial}{\partial \rho} MSE_1(\lambda, \rho) = - \left[\frac{2}{T-2} \right] \sum_{t=3}^T (Y_t - \hat{Y}_t) \frac{\partial}{\partial \rho} \hat{Y}_t \tag{A1}$$

$$\frac{\partial}{\partial \rho} \hat{Y}_t = \begin{cases} (1 + \lambda \hat{Y}_t^{(\lambda)})^{(1/\lambda)-1} \frac{\partial}{\partial \rho} \hat{Y}_t^{(\lambda)} & \text{when } \lambda \neq 0 \\ \exp(\hat{Y}_t^{(\lambda)}) \frac{\partial}{\partial \rho} \hat{Y}_t^{(\lambda)} & \text{when } \lambda = 0 \end{cases} \tag{A2}$$

$$\frac{\partial}{\partial \rho} \hat{Y}_t^{(\lambda)} = (X_t - \rho X_{t-1}) \frac{\partial}{\partial \rho} \hat{\beta}(\lambda, \rho) + (Y_{t-1}^{(\lambda)} - X_{t-1} \hat{\beta}(\lambda, \rho)) \tag{A3}$$

$$\frac{\partial}{\partial \rho} \hat{\beta}(\lambda, \rho) = - W^{-1} \left[\frac{\partial}{\partial \rho} W \right] W^{-1} \mathbf{X}'_{t-1} V_{11}^{-1} \mathbf{Y}_{t-1}^{(\lambda)} + W^{-1} \mathbf{X}'_{t-1} \left[\frac{\partial}{\partial \rho} V_{11}^{-1} \right] \mathbf{Y}_{t-1}^{(\lambda)} \tag{A4}$$

where $W = \mathbf{X}'_{t-1} V_{11}^{-1} \mathbf{X}_{t-1}$

$$\frac{\partial}{\partial \rho} W = \mathbf{X}'_{t-1} \left[\frac{\partial}{\partial \rho} V_{11}^{-1} \right] \mathbf{X}_{t-1} \tag{A5}$$

$$\frac{\partial}{\partial \rho} V_{11}^{-1} = (1 - \rho^2)^{-2} \begin{bmatrix} 2\rho & -1 - \rho^2 & 0 & \dots & 0 \\ -1 - \rho^2 & 4\rho & -1 - \rho^2 & \dots & \cdot \\ 0 & -1 - \rho^2 & 4\rho & \dots & \cdot \\ \cdot & 0 & -1 - \rho^2 & \dots & \cdot \\ \cdot & \cdot & 0 & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ 0 & \cdot & \cdot & \dots & 2\rho \end{bmatrix} \tag{A6}$$

$$\frac{\partial}{\partial \lambda} MSE_1(\lambda, \rho) = - \left[\frac{2}{T-2} \right] \sum_{t=3}^T (Y_t - \hat{Y}_t) \frac{\partial}{\partial \lambda} \hat{Y}_t \tag{A7}$$

$$\frac{\partial}{\partial \lambda} \hat{Y}_t = \begin{cases} \hat{Y}_t \left[\lambda^{-1} (1 + \lambda \hat{Y}_t^{(\lambda)})^{-1} \left[\hat{Y}_t^{(\lambda)} + \lambda \frac{\partial}{\partial \lambda} \hat{Y}_t^{(\lambda)} \right] \right] - \hat{Y}_t [\lambda^{-2} \log(1 + \lambda \hat{Y}_t^{(\lambda)})] & \text{when } \lambda \neq 0 \\ \exp(\hat{Y}_t^{(\lambda)}) \frac{\partial}{\partial \lambda} \hat{Y}_t^{(\lambda)} & \text{when } \lambda = 0 \end{cases} \tag{A8}$$

$$\frac{\partial}{\partial \lambda} \hat{Y}_t^{(\lambda)} = (X_t - \rho X_{t-1}) \frac{\partial}{\partial \lambda} \hat{\beta}(\lambda, \rho) + \rho \frac{\partial}{\partial \lambda} Y_{t-1}^{(\lambda)} \tag{A9}$$

$$\frac{\partial}{\partial \lambda} \hat{\beta}(\lambda, \rho) = W^{-1} \mathbf{X}'_{t-1} V_{11}^{-1} \frac{\partial}{\partial \lambda} \mathbf{Y}_{t-1}^{(\lambda)} \tag{A10}$$

$$\frac{\partial}{\partial \lambda} Y_{t-1}^{(\lambda)} = \begin{cases} \lambda^{-1}(Y_{t-1}^\lambda \log Y_{t-1} - Y_{t-1}^{(\lambda)}) & \text{when } \lambda \neq 0 \\ 1/2(\log Y_{t-1})^2 & \text{when } \lambda = 0 \end{cases} \quad (\text{A11})$$

$$\frac{\partial}{\partial \rho^2} \text{MSE}_1(\lambda, \rho) = \left[\frac{2}{T-2} \right] \sum_{t=3}^T \left[\frac{\partial}{\partial \rho} \hat{Y}_t \right]^2 - \left[\frac{2}{T-2} \right] \sum_{t=3}^T (Y_t - \hat{Y}_t) \frac{\partial^2}{\partial \rho^2} \hat{Y}_t \quad (\text{A12})$$

$$\frac{\partial^2}{\partial \rho^2} \hat{Y}_t = \begin{cases} (1-\lambda)/(1+\lambda \hat{Y}_t)^{(1/\lambda)-2} \left[\frac{\partial}{\partial \rho} \hat{Y}_t^{(\lambda)} \right]^2 + (1+\lambda \hat{Y}_t^{(\lambda)})^{(1/\lambda)-1} \frac{\partial^2}{\partial \rho^2} \hat{Y}_t^{(\lambda)} & \text{when } \lambda \neq 0 \\ \exp(\hat{Y}_t^{(\lambda)}) \left[\frac{\partial}{\partial \rho} \hat{Y}_t^{(\lambda)} \right]^2 + \exp(\hat{Y}_t^{(\lambda)}) \frac{\partial^2}{\partial \rho^2} \hat{Y}_t^{(\lambda)} & \text{when } \lambda = 0 \end{cases} \quad (\text{A13})$$

$$\frac{\partial^2}{\partial \rho^2} \hat{Y}_t^{(\lambda)} = -2X_{t-1} \frac{\partial}{\partial \rho} \hat{\beta}(\lambda, \rho) + (X_t - \rho X_{t-1}) \frac{\partial^2}{\partial \rho^2} \hat{\beta}(\lambda, \rho) \quad (\text{A14})$$

$$\begin{aligned} \frac{\partial^2}{\partial \rho^2} \hat{\beta}(\lambda, \rho) = & \left[\frac{\partial^2}{\partial \rho^2} W^{-1} \right] \mathbf{X}'_{t-1} V_{11}^{-1} \mathbf{Y}_{t-1}^{(\lambda)} + 2 \left[\frac{\partial}{\partial \rho} W^{-1} \right] \mathbf{X}'_{t-1} \left[\frac{\partial}{\partial \rho} V_{11}^{-1} \right] \mathbf{Y}_{t-1}^{(\lambda)} \\ & + W^{-1} \mathbf{X}'_{t-1} \left[\frac{\partial^2}{\partial \rho^2} V_{11}^{-1} \right] \mathbf{Y}_{t-1}^{(\lambda)} \end{aligned} \quad (\text{A15})$$

$$\frac{\partial^2}{\partial \rho^2} W^{-1} = - \left[\frac{\partial}{\partial \rho} W^{-1} \right] \left[\frac{\partial}{\partial \rho} W \right] W^{-1} - W^{-1} \left[\frac{\partial^2}{\partial \rho^2} W \right] W^{-1} - W^{-1} \left[\frac{\partial}{\partial \rho} W \right] \frac{\partial}{\partial \rho} W^{-1} \quad (\text{A16})$$

$$\frac{\partial^2}{\partial \rho^2} W = \mathbf{X}'_{t-1} \left[\frac{\partial^2}{\partial \rho^2} V_{11}^{-1} \right] \mathbf{X}_{t-1} \quad (\text{A17})$$

$$\frac{\partial^2}{\partial \rho^2} V_{11}^{-1} = (1-\rho^2)^{-3} \begin{bmatrix} 2+6\rho^2 & -6\rho-2\rho^3 & \dots & 0 \\ -6\rho-2\rho^3 & 4+12\rho^2 & \dots & \cdot \\ 0 & -6\rho-2\rho^3 & \dots & \cdot \\ \cdot & 0 & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ 0 & \cdot & \dots & 2+6\rho^2 \end{bmatrix} \quad (\text{A18})$$

$$\frac{\partial^2}{\partial \lambda^2} \text{MSE}_1(\lambda, \rho) = \left[\frac{T}{T-2} \right] \sum_{t=3}^T \left[\frac{\partial}{\partial \lambda} \hat{Y}_t \right]^2 - \left[\frac{2}{T-2} \right] \sum_{t=3}^T (Y_t - \hat{Y}_t) \frac{\partial^2}{\partial \lambda^2} \hat{Y}_t \quad (\text{A19})$$

$$\frac{\partial^2}{\partial \lambda^2} \hat{Y}_t = \begin{cases} \lambda^{-1} \left[\frac{\partial}{\partial \lambda} \hat{Y}_t \right] [1 + \lambda \hat{Y}_t^{(\lambda)}]^{-1} \left[\hat{Y}_t^{(\lambda)} + \lambda \frac{\partial}{\partial \lambda} \hat{Y}_t^{(\lambda)} \right] \\ - \lambda^{-2} \left[\frac{\partial}{\partial \lambda} \hat{Y}_t \right] \log [1 + \lambda \hat{Y}_t^{(\lambda)}] \\ + \lambda^{-1} \hat{Y}_t [1 + \lambda \hat{Y}_t^{(\lambda)}]^{-1} \left[2 \frac{\partial}{\partial \lambda} \hat{Y}_t^{(\lambda)} + \lambda \frac{\partial^2}{\partial \lambda^2} \hat{Y}_t^{(\lambda)} \right] \\ - \lambda^{-1} \hat{Y}_t [1 + \lambda \hat{Y}_t^{(\lambda)}]^{-2} \left[\hat{Y}_t^{(\lambda)} + \lambda \frac{\partial}{\partial \lambda} \hat{Y}_t^{(\lambda)} \right]^2 \\ - 2\lambda^{-2} \hat{Y}_t [1 + \lambda \hat{Y}_t^{(\lambda)}]^{-1} \left[\hat{Y}_t^{(\lambda)} + \lambda \frac{\partial}{\partial \lambda} \hat{Y}_t^{(\lambda)} \right] \\ + 2\lambda^{-3} \hat{Y}_t \log [1 + \lambda \hat{Y}_t^{(\lambda)}] & \text{when } \lambda \neq 0 \\ \exp[\hat{Y}_t^{(\lambda)}] \left[\frac{\partial}{\partial \lambda} \hat{Y}_t^{(\lambda)} \right]^2 + \exp[\hat{Y}_t^{(\lambda)}] \frac{\partial^2}{\partial \lambda^2} \hat{Y}_t^{(\lambda)} & \text{when } \lambda = 0 \end{cases} \quad (\text{A20})$$

$$\frac{\partial^2}{\partial \lambda^2} \hat{Y}_t^{(\lambda)} = (X_t - \rho X_{t-1}) \frac{\partial^2}{\partial \lambda^2} \hat{\beta}(\lambda, \rho) + \rho \frac{\partial^2}{\partial \lambda^2} Y_{t-1}^{(\lambda)} \quad (\text{A21})$$

$$\frac{\partial^2}{\partial \lambda^2} \hat{\beta}(\lambda, \rho) = W^{-1} X'_{t-1} V_{11}^{-1} \frac{\partial^2}{\partial \lambda^2} Y_{t-1}^{(\lambda)} \quad (\text{A22})$$

$$\frac{\partial^2}{\partial \lambda^2} Y_{t-1}^{(\lambda)} = \begin{cases} \lambda^{-1} \left[Y_{t-1}^{\lambda} (\log Y_{t-1})^2 - 2 \frac{\partial}{\partial \lambda} Y_{t-1}^{(\lambda)} \right] & \text{when } \lambda \neq 0 \\ \frac{1}{3} (\log Y_{t-1})^3 & \text{when } \lambda = 0 \end{cases} \quad (\text{A23})$$

$$\frac{\partial^2}{\partial \lambda \partial \rho} \text{MSE}_1(\lambda, \rho) = \left[\frac{2}{T-2} \right] \sum_{t=3}^T \left[\frac{\partial}{\partial \lambda} \hat{Y}_t \right] \left[\frac{\partial}{\partial \rho} \hat{Y}_t \right] - \left[\frac{2}{T-2} \right] \sum_{t=3}^T (Y_t - \hat{Y}_t) \frac{\partial^2}{\partial \lambda \partial \rho} \hat{Y}_t \quad (\text{A24})$$

$$\frac{\partial^2}{\partial \lambda \partial \rho} \hat{Y}_t = \begin{cases} \left[\frac{\partial}{\partial \lambda} \hat{Y}_t \right] [1 + \lambda \hat{Y}_t^{(\lambda)}]^{-1} \frac{\partial}{\partial \rho} \hat{Y}_t^{(\lambda)} \\ - \hat{Y}_t [1 + \lambda \hat{Y}_t^{(\lambda)}]^{-2} \left[\hat{Y}_t^{(\lambda)} + \lambda \frac{\partial}{\partial \lambda} \hat{Y}_t^{(\lambda)} \right] \frac{\partial}{\partial \rho} \hat{Y}_t^{(\lambda)} \\ + \hat{Y}_t [1 + \lambda \hat{Y}_t^{(\lambda)}]^{-1} \frac{\partial^2}{\partial \lambda \partial \rho} \hat{Y}_t^{(\lambda)} & \text{when } \lambda \neq 0 \\ \exp[\hat{Y}_t^{(\lambda)}] \left[\frac{\partial}{\partial \rho} \hat{Y}_t^{(\lambda)} \right] \left[\frac{\partial}{\partial \rho} \hat{Y}_t^{(\lambda)} \right] \\ + \exp[\hat{Y}_t^{(\lambda)}] \frac{\partial^2}{\partial \lambda \partial \rho} \hat{Y}_t^{(\lambda)} & \text{when } \lambda = 0 \end{cases} \quad (\text{A25})$$

$$\frac{\partial^2}{\partial \lambda \partial \rho} \hat{Y}_t^{(\lambda)} = (X_t - \rho X_{t-1}) \frac{\partial^2}{\partial \lambda \partial \rho} \hat{\beta}(\lambda, \rho) + \left[\frac{\partial}{\partial \lambda} \hat{Y}_{t-1}^{(\lambda)} - X_{t-1} \frac{\partial}{\partial \lambda} \hat{\beta}(\lambda, \rho) \right] \quad (\text{A26})$$

$$\frac{\partial^2}{\partial \lambda \partial \rho} \hat{\beta}(\lambda, \rho) = \left[\frac{\partial}{\partial \rho} W^{-1} \right] X'_{t-1} V_{11}^{-1} \frac{\partial}{\partial \lambda} Y_{t-1}^{(\lambda)} + W^{-1} X'_{t-1} \left[\frac{\partial}{\partial \rho} V_{11}^{-1} \right] \frac{\partial}{\partial \lambda} Y_{t-1}^{(\lambda)} \quad (\text{A27})$$

REFERENCES

- Box, G. E. P. and Cox, D. R., 'An analysis of transformations', *J. Roy Statist. Soc., Ser. B*, **26** (1964), 211-252.
- Dawid, A. P., 'Present position and potential developments: some personal views', *J. R. Statist. Soc. A*, (1984), 278-292.
- Fisher, J. C. and Pry, R. H., 'A simple substitution model of technological change', *Technological Forecasting and Social Change*, **3** (1971), 75-88.
- Fiebig, D. G. and Bewley, R., 'International telecommunications forecasting: an investigation of alternative functional forms', *Applied Economics*, **19** (1987), 949-960.
- Granger, C. W. J. and Newbold, P., 'Forecasting transformed series', *Journal of Royal Statist. Soc., Ser. B*, **38** (1976), 189-203.
- Hopwood, W. S., McKeown, J. C. and Newbold, P., 'Time series forecasting involving power transformation', *Journal of Forecasting*, **3** (1984), 57-61.
- Keramidas, E. M. and Lee, J. C., 'Forecasting technological substitutions with short time series', *Proceedings of the Business and Economic Statistic Section*, American Statistical Association, 1988, pp. 1-10.

- Keramidas, E. M. and Lee, J. C., 'Forecasting technological substitutions with concurrent short time series', *J. American Statistical Association*, **85** (1990), 625–632.
- Lee, J. C. and Lu, K. W., 'On a family of data-based transformed models useful in forecasting technological substitutions', *Technological Forecasting and Social Change*, **31** (1987), 61–78.
- Lee, J. C., 'Prediction and estimation of growth curves with special covariance structures', *Journal of American Statistical Association*, **83** (1988), 432–440.
- Lee, J. C. and Lu, K. W., 'Algorithm and practice of forecasting technological substitutions with data-based transformed model', *Technological Forecasting and Social Change*, **36** (1989), 401–414.
- Pankratz, A. and Dudley, U., 'Forecasts of power-transformed series', *Journal of Forecasting*, **6** (1987), 239–248.
- Port 3 Mathematical Subroutine Library, AT&T Bell Telephone Laboratories, Inc. 1984.
- Rissanen, J., 'Order estimation by accumulated prediction errors', in Gani, J. and Priestley, M. B. (eds), *Essays in Time Series and Allied Processes—Papers in Honor of E. J. Hannan*, Sheffield, 1986.

Authors' biographies:

Jack C. Lee is Professor and Director, Institute of Statistics, National Chiao Tung University, Taiwan, Republic of China. He is currently on leave of absence from the Statistics and Econometrics Research Group in the Applied Research Area of Bellcore. He received a BA in Business from the National Taiwan University, an MA in economics from the University of Rochester and a PhD in statistics from SUNY at Buffalo. His areas of research include growth curves, forecasting, and applications of statistics.

S. Lee Tsao is currently a Member of Technical Staff of AT&T Bell Laboratories at Holmdel, New Jersey and was previously a consultant at Bellcore, Morristown, New Jersey. She holds a BS in psychology from the National Taiwan University and a PhD in mathematics from Rensselaer Polytechnic Institute, Troy, New York. Her research interests include differential equations, forecasting, cluster analysis, and telecommunications technology.

Authors' addresses:

Jack C. Lee, Institute of Statistics, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan, Republic of China.

S. Lee Tsao, AT&T Bell Labs, Crawford Corner Road, Holmdel, NJ 07733. USA.