Statistica Neerlandica (1996) Vol. 51, nr. 1, pp. 90–106

Optimal transformations for categorical autoregressive time series[1]

S. van Buuren*

Department of Statistics, TNO Prevention and Health, P.O. Box 2215, 2301 CE Leiden, The Netherlands

This paper describes a method for finding optimal transformations for analyzing time series by autoregressive models. 'Optimal' implies that the agreement between the autoregressive model and the transformed data is maximal. Such transformations help 1) to increase the model fit, and 2) to analyze categorical time series. The method uses an alternating least squares algorithm that consists of two main steps: estimation and transformation. Nominal, ordinal and numerical data can be analyzed. Some alternative applications of the general idea are highlighted: intervention analysis, smoothing categorical time series, predictable components, spatial modeling and cross-sectional multivariate analysis. Limitations, modeling issues and possible extensions are briefly indicated.

Key Words & Phrases: nonlinear transformations, quantification, qualitative data, canonical correlation, majorization, autoregressive model, predictable components, intervention analysis, smoothing.

1 Introduction

An important objective in the analysis of time series is to predict future from past data. To this end, a time series model is often fitted to the data. If this model holds well and if it is expected to hold in the future as well, then the model can be used to predict future observations. The ARIMA model developed by Box and Jenkins (1976) is a very popular linear model in this context. This paper describes a method to extend the autoregressive model with an optimal transformation of the data. 'Optimal' means that, within the allowable class, a transformation will be sought such that the agreement between the model and the data is maximal. This serves two purposes: 1) to increase the fit so that predictions can improve, and 2) to analyze and predict categorical time series. Purpose 1 is achieved by applying appropriate linearizing transformations to the data. Purpose 2 is achieved by assigning an optimal value to each category followed by estimating the parameters of interest from the quantified data. Both goals work together and are attained by the same method.

* S.vanBuuren@pg.tno.nl

[1] I am particularly indebted to Jan de Leeuw for suggesting and sketching the majorization method that I apply in this paper. I thank Jaap Brand, Catrien Bijleveld, Peter van der Heijden, Dirk Sikkel, the editor and several anonymous reviewers for their suggestions and encouragements.

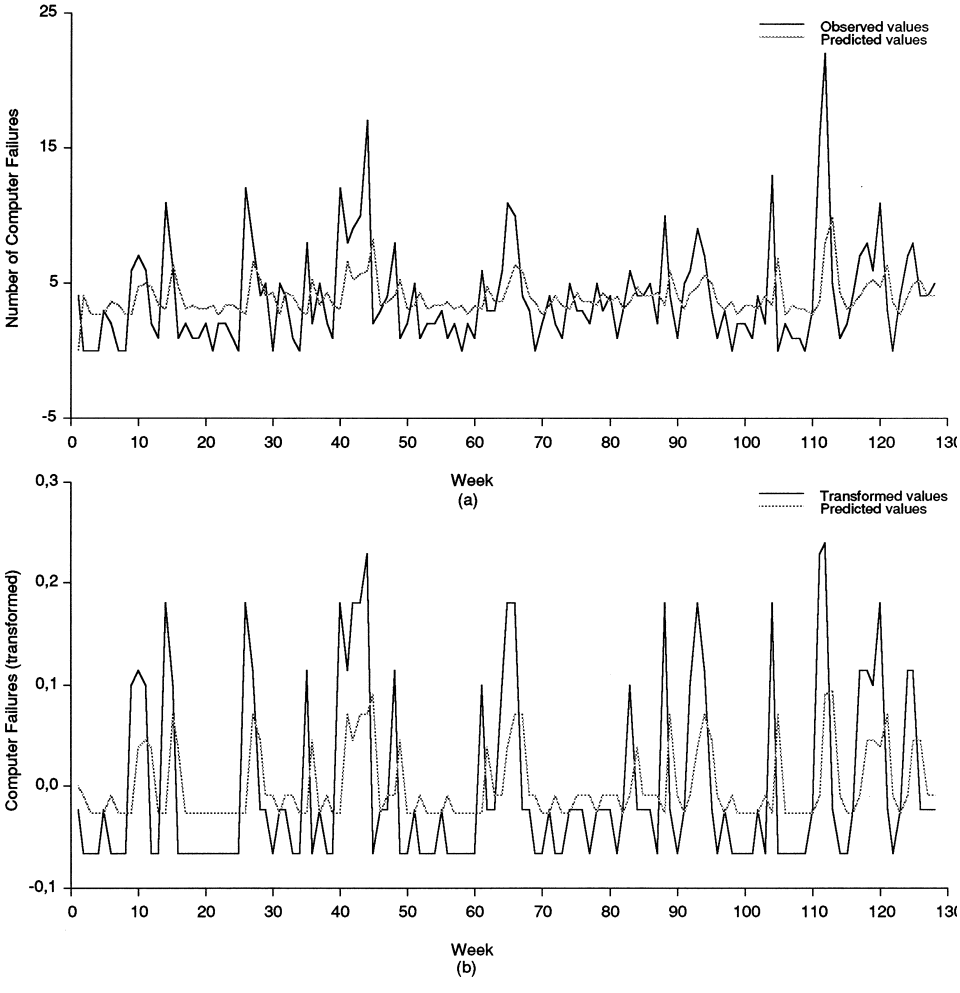

Fig. 1. Weekly number of computer failures ($T = 128$). Figure (a) plots the raw data (solid line) together with the fitted values according to the AR(1) model (dashed line). Figure (b) is the same plot for the optimally transformed data (Source: HAND et al., 1993, p. 109).

As an example of the type of problems for which the method may be useful, consider the empirical time series plotted in Figure 1a. These data are taken from HAND et al. (1993, dataset 141) and give the number of times that a DEC-20 computer of the Open University broke down in each of 128 consecutive weeks of operation, starting in late 1983. The pattern of autocorrelations suggest an autoregressive model of order 1, that is, $u_t = u_{t-1}\phi_1 + e_t$, where u_t denotes the mean deviation of the series of interest, and where e_t is an i.i.d. normal random variable ($t = 1, \ldots, T$). The least squares estimate is $\hat{\phi}_1 = 0.324$. The series of predicted values, $\hat{u}_t = u_{t-1}\hat{\phi}_1$ is also plotted. Large differences point to locations for which the model fits badly. The expected mean squared prediction error for standardized data is equal to

$E[(u-\hat{u})^2] = 0.911$. Notation $E[\cdot]$ indicates the expectation operator. Figure 1b plots the same series, but now after an optimal monotone transformation $x_t = f(u_t)$ has been applied. This transformation uses only the ranking order in the data. The estimate of ϕ is now $\hat{\phi}_1 = 0.394$. The comparable mean squared error is equal to $E[(x-\hat{x})^2] = 0.851$, which improves upon the model for the raw data. The plot demonstrates that the transformation effectively compresses a few large residual terms, thereby making the analysis more robust against outliers. In addition, the transformation reduces the mean prediction error and improves the fit to the data.

The use of nonlinear data transformations in time series analysis has a long tradition. Log, square root and reciprocal transformations are routinely used. The Box–Cox transformation is another popular candidate. The shape parameter is sometimes optimized to increase the model fit, usually by grid search or by numerical optimization as in Ansley, Spivey and Wrobleski (1977). Differencing the data is another well known transformation. It forms an integral component of the ARIMA model. Methods for integrating nonlinear transformations and linear time series models by ACE can be found in Owen (1983) and Young (1990). The present work extends previous research by considering a different class of transformations. This class is less restrictive in the sense that the transformation function need not be smooth or monotone.

Methods for analyzing categorical time series come in several flavors. Bishop, Fienberg and Holland (1975, Ch. 7) and others have shown how Markov chains can be formulated as log-linear and logit models for transition matrices. These methods are often applied to panel data in which many individuals are observed during a small number of time points. By assuming stationarity it is also possible to estimate the transition matrix from a single categorical time series, which can be used for further analysis. In this way, one could not only study the transition from $t-1$ to t, but also second order transitions from $t-2$ and $t-1$ to t, and so on. Jacobs and Lewis (1978) proposed a quite different method, the DARMA model. Their model generalizes the ARMA model to sequences of discrete random variables by forming linear combinations of discrete random variables. Singh and Lemaitre (1987) adapted the method to panel data, but the DARMA model has not gained wide acceptance. Stoffer (1991) proposes Walsh-Fourier analysis for time series. This method decomposes the data into block waves, and can also handle categorical series. Harvey and Fernandes (1989) put forward structural models for categorical series. Fahrmeir (1992) describes a method for fitting categorical series to a discrete state space model by extended Kalman filtering. Raveh and Tapiero (1980) present interesting techniques for studying recurring patterns in categorical series. Gregson (1987) simply applies linear models to categorical daily self-report data. Deville and Saporta (1983) adapted correspondence analysis to nominal time series.

The present work adapts the classic linear autoregressive model to categorical data by means of optimal scaling. Optimal scaling as practiced here is equivalent to linear analysis with an added set of scaling parameters. Given the transformation, finding parameter estimates is a linear problem which can be solved by least squares.

However, since the transformation is unknown, iteration is typically needed to estimate both sets of parameters. Extensive experience with such methods for multivariate analysis has accumulated in GIFI (1990).

The method assumes that the data are classified into discrete and non-overlapping categories. Time series can be measured on nominal, ordinal or interval scales, or any mix of these. It is assumed that it makes sense to scale the categories on a line. If so, optimal scaling generally enhances the interpretation of the relations among the categories. If not, e.g. for truly nominal series like religion, scale values may have little meaning, but these cases are rather exceptional. Another assumption of the method is constancy of quantification, i.e. scaling does not change with time. This is a desirable property if the meaning of the category system is the same for all time points, which is typically the case. Observations are assumed to be equidistant in time. The number of needed time points grows, amongst others, with the number of free parameters. Unless the user is willing to accept very imprecise estimates, at least 50 observations should be available for fitting a simple autoregressive model. Of course, more complex models need more.

The structure of the text is as follows. Section 2 introduces the mathematical translation of the problem into the minimization of a loss function. Section 3 presents two examples of the approach. Section 4 sketches a number of other applications that are based on the same ideas. Finally, section 5 closes the paper and discusses some pitfalls and extensions.

2 Method

2.1 *Problem formulation*

The autoregressive model of order one as discussed in the introduction is a special case of the general stationary model of order P,

$$u_t = u_{t-1}\phi_1 + u_{t-2}\phi_2 + \cdots + u_{t2}\phi_P + e_t \tag{1}$$

where $E[(\Sigma_p u_{t-p}\phi_p)' e_t = 0$ and where e_t is a serially uncorrelated normal random variable with zero mean. Autoregressive models express the current score u_t as a linear combination of previous observations $u_{t-1}, u_{t-2}, \ldots, u_{t-P}$ plus an error component. This error term incorporates everything new in the series at time t that is not explained by the past data. If only one particular lag serves as a predictor, say u_{t-4} for quarterly series, a seasonal model is formed that can be used to portray periodic phenomena.

Suppose that the data u_t are transformed by a function $f(\,)$ as $x_t = f(u_t)$ then the autoregressive model of the transformed data turns into

$$\begin{aligned} f(u_t) &= f(u_{t-1})\phi_1 + f(u_{t-2})\phi_2 + \cdots + f(u_{t-P})\phi_P + \varepsilon_t \\ &= \sum_{p=1}^{P} f(u_{t-p})\phi_p + \varepsilon_t \end{aligned} \tag{2}$$

where ε_t consists of white noise. The class of transformations that will be considered in this paper can be written as $x_t = g_t y$. Here g_t is a time varying binary row vector of length K that indicates in which of K categories each u_t falls, and $y = [y_1, \ldots, y_K]'$ is a column vector containing unknown scaling weights, or category quantifications. If the measurement level of u_t is ordinal then it makes sense to restrict the sequence of y-values to be monotonically increasing so that the transformation preserves the ordering of the categories. Likewise, for interval variables requiring that the y-values increase in fixed increments maintains equal distances between the categories.

Substituting for x_t in (2) results in the generalized autoregressive model

$$\begin{aligned} g_t y &= g_{t-1} y\phi_1 + g_{t-2} y\phi_2 + \cdots + g_{t-P} y\phi_P + \varepsilon_t \\ &= \sum_{p=1}^{P} g_{t-p} y\phi_p + \varepsilon_t \end{aligned} \tag{3}$$

in which $\phi_1, \ldots, \phi_P$ and $y_1, \ldots, y_K$ are two sets of unknown parameters. Estimating these parameters from the data by least squares can be done by minimizing the loss function

$$\sigma^*(y_1, \ldots, y_K, \phi_1, \ldots, \phi_P) = \sum_{t=1+P}^{T} \left(g_t y - \sum_{p=1}^{P} g_{t-p} y\phi_p \right)^2 \tag{4}$$

The first P observations are deleted from the function because g_t for $t < 1$ is generally not known. GIFI (1990, p. 241) remarks that minimizing $\sigma^*(\,)$ maximizes the multiple correlation between $g_t y$ and $\Sigma_p g_{t-p} y\phi_p$. Although it is possible to minimize $\sigma^*(\,)$ directly, it is convenient to reformulate (4) as

$$\sigma(y_1, \ldots, y_K, a_0, \ldots, a_P, z_t) = \sum_{t=1+P}^{T} \left((z_t - g_t y a_0)^2 + \left(z_t - \sum_{p=1}^{P} g_{t-p} y a_p \right)^2 \right) \tag{5}$$

where z_t is an auxiliary observation from a latent variable, with $E[z_t] = 0$ and $E[z_t' z_t] = 1$, and where $a_0, \ldots, a_P$ are unknown weights. It is known that $\sigma(\,)$ and $\sigma^*(\,)$ lead to the same solution for y (GIFI, 1990, p. 220). The main reason for using $\sigma(\,)$ instead of $\sigma^*(\,)$ is to facilitate generalizations to the multivariate case. In practice, one finds the y-estimates by minimizing (5), followed by least squares regression of $g_t y$ on $[g_{t-1} y, \ldots, g_{t-P} y]$ to estimate $\phi_1, \ldots, \phi_P$. To identify the solution the transformation $x_t = g_t y$ is normalized such that $E[x_t] = 0$ and $E[x_t' x_t] = 1$. The next section addresses the problem of minimizing (5).

2.2 *Parameter estimation*

Let $G = [g_1', \ldots, g_T']'$ denote the $T \times K$ indicator matrix of the data. Let $x = Gy$ represent the quantified time series. The matrix B is defined as the $T \times T$ matrix with ones on the subdiagonal, zeroes elsewhere (the backshift operator). The product Bx is called a lagged variable of order one. The product $B^p = BB \ldots B$ is the pth order

backshift matrix. Define $\text{ssq}(z) = \text{tr}(z'z)$. Loss function (5) can now be written in matrix notation as

$$\begin{aligned}\sigma(y, a_0, \ldots, a_P, z) &= \text{ssq}(z - Gya_0) + \text{ssq}\left(z - \sum_p B^p Gya_p\right) \\ &= \text{ssq}(z - xa_0) + \text{ssq}\left(z - \sum_p B^p xa_p\right)\end{aligned} \tag{6}$$

where the first P rows are deleted from each matrix. This function is minimized by an alternating least squares algorithm that consists of three main steps. Each of these steps decreases $\sigma(\,)$, or at least does not increase it, over a specific set of parameters. Iterating the steps while conditioning on the most recent estimates produces, under mild regularity conditions, a convergent algorithm. See GIFI (1990, p. 58) for a discussion of the general idea. The paper by DE LEEUW, YOUNG and TAKANE (1976) contains additional theoretical properties of alternating least squares procedures. The steps are

a) minimization over z for fixed y and $a_0, \ldots, a_P$ by least squares;
b) minimization over $a_0, \ldots, a_P$ for fixed z and y by least squares;
c) minimization over y for fixed z and $a_0, \ldots, a_P$ by majorization.

The minimum of $\sigma(\,)$ over z is found by setting $z = xa_0 + \Sigma_p B^p xa_p$, followed by scaling it to zero mean and unit variance. The minimum over a_0 is obtained by projection, that is, by setting $a_0 = x'z/x'x$. Likewise, the minimum over $[a_1, \ldots, a_P]$ is obtained as $[Bx, \ldots, B^P x]^+ z$, where the superscript '+' indicates the Moore-Penrose inverse of a matrix. The procedure for finding y is more complicated because it is difficult to derive a closed form expression for y conditional on the other parameters. A solution is to replace the minimization of (6) by the iterative minimization of a sequence of simpler loss functions whose values are always in excess of those of the complicated function. This idea is called majorization, and it was applied first within the context of multidimensional scaling (DE LEEUW, 1977; DE LEEUW and HEISER, 1980). The appendix outlines how majorization is applied to the present problem.

Since steps a, b, and c all decrease, or at least do not increase the loss, iterating them also decreases the loss until it cannot be lowered any further. The algorithm stops if the loss difference between two subsequent main iterations becomes less than 0.0005. About 15 to 50 iterations are typically needed for fitting a univariate autoregressive model. The algorithm is quite stable in the sense that tight stopping criteria like 0.0000001 do not affect convergence.

Not much is known about the occurrence of local minima. Using different starting configurations, the algorithm may produce dissimilar solutions. Sometimes it helps to use tight convergence criteria that prevent the algorithm from stopping at an almost flat region of the loss function. In other cases, the solutions may still be qualitatively different, for example if two solutions tie together two different combinations of

categories. These cases seem harder to preclude. A moderately successful strategy is to start from a reasonable initial solution in the hope that it will be close to the final solution. Experience suggests that executing a preliminary homogeneity analysis tends to compress the solution towards the optimum. As a last resort, one may compute many solutions, each using different starting configurations, and pick out the solution with the best fit. Of course, this is computationally intensive. Moreover, even this does not guarantee that the global optimum will be found within reasonable time.

3 Examples

3.1 *Box–Jenkins Series D*

The Box and Jenkins (1976) series D consists of 310 observations of hourly viscosity readings of an uncontrolled chemical process. The series is plotted in Figure 2a.

Grubb (1992) noted that it may seem perverse to force this series into a categorical scale, but since the observations happen to fall into a small number (26) of discrete

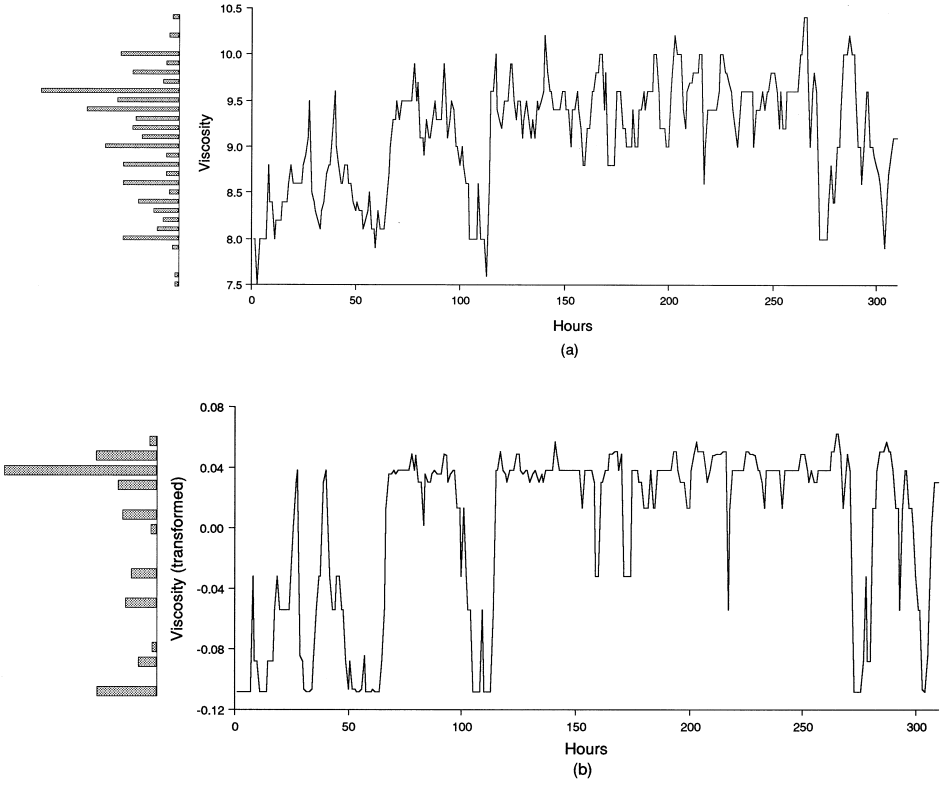

Fig. 2. Box–Jenkins Series D: Hourly viscosity readings from an uncontrolled chemical process ($T = 310$). Figure (a) plots the original data. Figure (b) plots the series after optimal transformation by the AR(1) model. The diagram on the left estimates the density along the scale (Source: Box and Jenkins, 1976, p. 529).

values no information is lost. Box and Jenkins identified two candidate models for this series: a nonstationary differencing model with an MA-component, and a stationary AR(1) model. To replicate the latter analysis, equation (6) was fitted to series D under a linear transformation function. The minimum loss was $\sigma_{\text{lin}} = 0.1385$, with $a_0 = a_1 = 0.96$ and $\hat{\phi}_1 = 0.87$, which is the same as in Box and Jenkins. An important model evaluation criterion in time series analysis is whether the residuals from the model effectively conform to a white noise process. This was stressed by GHADDAR and TONG (1981). The value of Box–Pierce's Q, a statistic that measures residual autocorrelation, is equal to $Q = 10.2$ with 24 degrees of freedom. This value is not significantly larger, so it is concluded that the residuals could have been generated by a white noise process.

Next the series was analyzed under a monotone transformation of the data which preserves the ordering of the categories. The results for this ordinal analysis are $\sigma_{\text{ord}} = 0.0975$ with $a_0 = a_1 = 0.98$ and $\hat{\phi}_1 = 0.91$. The autocorrelation of the transformed series is equal to $r_{\text{ord}} = 0.91$. The residuals do not correlate (the Box–Pierce statistic is 19.5 with df = 24). Figure 2b plots the transformed series. In general, the series is flatter. The method compresses most values into the extremes of the scale. This can best be seen from the transformation plot. This plot is given in Figure 3 and graphs the observed values against their optimally scaled counterparts.

Figure 3 clearly demonstrates the increasing score pattern that preserves the ordering. In standard numerical time series analysis the scores would all have been located on a straight line. As noted, the monotone transformation tends to cluster the extremes of the scale. This effect is especially visible on the lower side: scores 7.2 to 8.2 obtain identical quantifications. This implies that, given an AR(1) model, the extremes of the scale do not discriminate very much among the measurements, i.e. it matters little whether a score of 7.2 or a score of 8.2 is observed. One possible interpretation of the phenomenon is that the physical process moves back and forward between two points of attraction, located at about 8.2 and 9.6. It is not known whether there is a physical basis for such an interpretation. What happens is that maximizing predictability minimizes the variation within each level, thereby making the points of attraction more visible. An alternative explanation is that optimal scaling blows up the agreement at the ends of the scale just to increase the autocorrelation, almost regardless of the data. This cannot be done without limits however. For example, coding the data into two categories never produces a first-order autocorrelation that exceeds 0.83. This is even lower than the raw value. It is not easy to beat the predictability of the transformed series by other methods. For example, regression analysis on the original data including 4 or 5 polynomial degrees comes close to producing a multiple correlation of 0.91, but does not exceed it.

It is not easy to say whether the ordinal analysis is 'better' than the linear one. The difference in terms of explained variance are not very large. For one thing, if the relationship between x_t and x_{t-1} is truly linear (and most time series models describe only linear relationships), the transformation plot shows a straight line. This is clearly

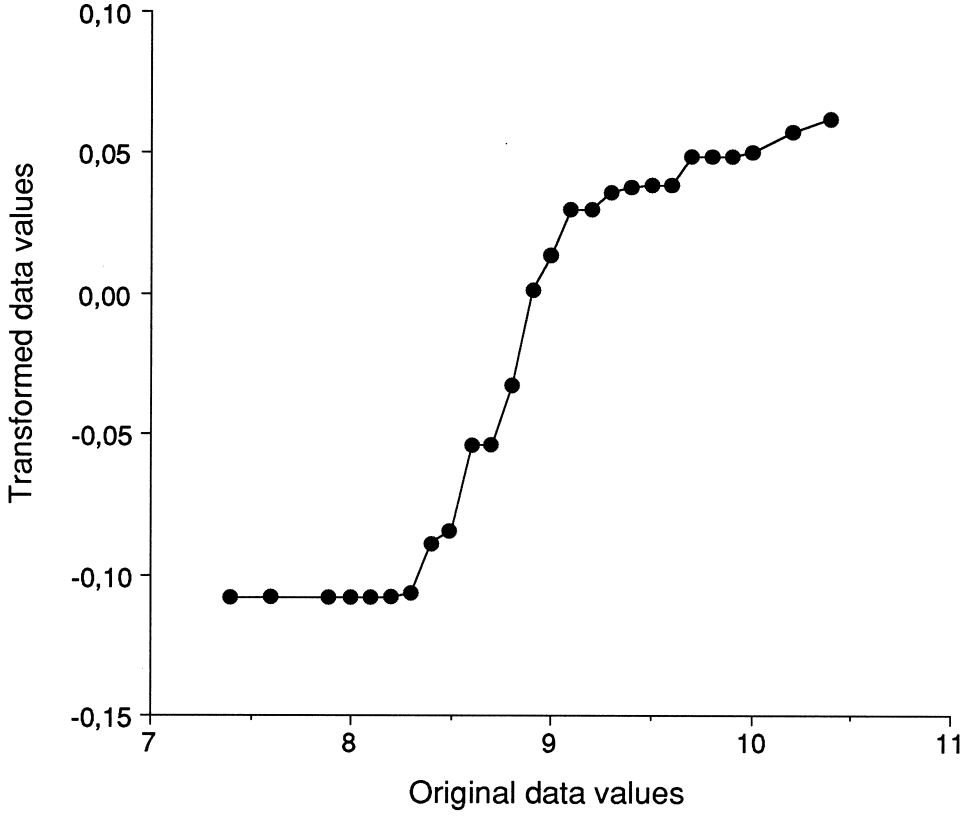

Fig. 3. Transformation function for Box–Jenkins Series D. The horizontal axis corresponds to the series in Figure 2a. The vertical axis corresponds to the transformed series of Figure 2b.

not the case here, but only a specialist in this field may be able to determine whether the grouping effect indicates a real world physical process or not.

3.2 *Rainfall data*

Different optimization criteria can lead to different data transformations. This subsection shows that the precise form of the transformation depends on the goal of the analysis. The method was applied to thirty successive values of March precipitation for Minneapolis/St. Paul listed in HAND et al. (1993, dataset 412). HINKLEY (1977) attempted to find a transformation that symmetrizes the distribution by minimizing the deviation between the mean and the median.

Figure 4 compares the optimal transformation functions for Hinkley's criterion, for the AR(1) model with ordinal data and for the AR(1) model with nominal data. In the latter two analyses, the data were coded into four categories. Hinkley's transformation is plotted in Figure 4a. It is equal to $(u_t^{0.25} - 1)/0.25$ and its form resembles that of the natural logarithm. Figure 4b gives the optimal monotone

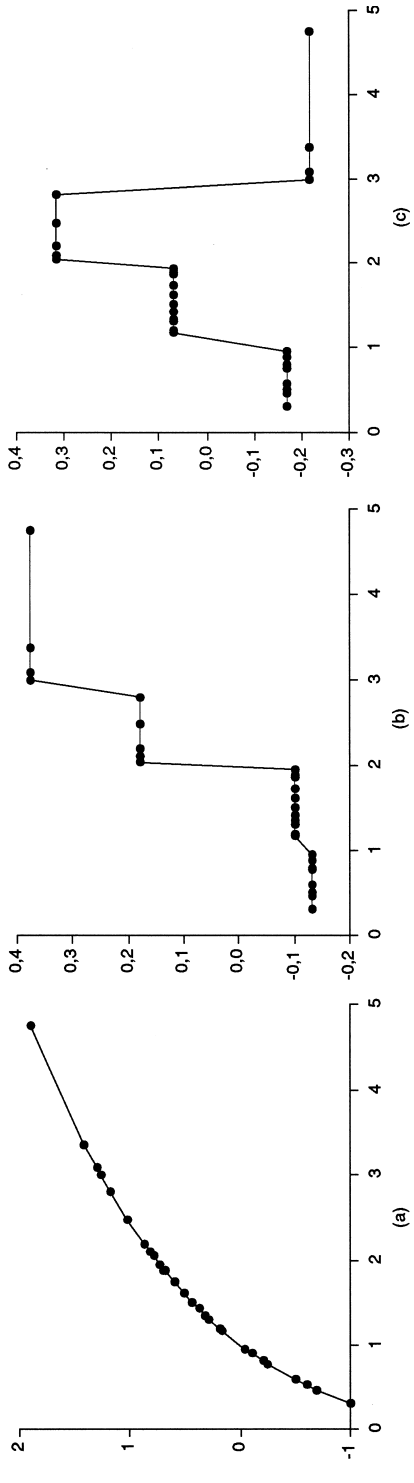

Fig. 4. Minneapolis/St. Paul rainfall data. Figure (a) plots Hinkley's symmetrizing power transformation. Figure (b) and (c) are the transformation functions for ordinal and nominal measurement levels respectively. (Source: HAND et al., 1993, p. 336).

transformation. This function is quite different from Hinkley's transform, and expands high values instead of compressing them. Figure 4c graphs the optimal transformation for a nominal measurement level. This is again quite different from the others. It almost equalizes the extremes. All three optimal transformations are strikingly different.

To gain some insight into the factors that govern these differences consider the summary statistics in Table 1.

Table 1. Summary statistics of the raw rainfall data plus three transformations (Hinkley's optimal power transform, AR(1) optimal ordinal, AR(1) optimal nominal). The symmetry criterion is equal to (Mean − Median)/SD. 'Autocor' means the first-order autocorrelation.

	Raw	Power	Ordinal	Nominal
Mean	1.675	0.404	0	0
Median	1.470	0.404	0.100	−0.067
SD	0.984	0.669	0.183	0.183
Criterion	0.208	0.000	−0.549	0.372
Autocor	0.197	0.213	0.303	−0.397

The power transformation brings the median close to the mean, as intended. Both other transforms actually induce asymmetry and are thus sub-optimal with respect to Hinkley's criterion. On the other hand, the power transformation hardly improves upon the predictability of the series if modelled by an AR(1) model. The first-order autocorrelation rises from 0.197 to just 0.213. The ordinal transform increases it to 0.303, while the nominal transform doubles the magnitude of the autocorrelation to (minus!) −0.397.

Note that nominal transformation distinguishes the extremes from the middle values. Apparently, it is easier to predict rapid alterations of mean deviations, than it is to model the series by flattening the original. Whether the transformation can be defended on scientific grounds is of course another matter. In some applications, it might be more important to minimize prediction errors than it is to understand the phenomenon. In such cases, a transform like that in Figure 4c might just be all that is needed.

4 Other applications

Optimal scaling is not limited to autoregressive models. This section describes a number of other applications of the same idea.

4.1 *Intervention analysis*

The goal of intervention analysis is to infer whether a specific event has an effect on the level of the series. A standard reference is Glass, Willson and Gottman (1975). Suppose that x_1 denotes a series of outcome values and that x_2 is a binary series that codes the presence and absence of an event, for example indicating whether a specific treatment was given at time point t. It would be convenient to use the t-test,

conditional on the level of x_2. Such a procedure is questionable however if x_1 is autocorrelated. The t-test is therefore only used after the serial correlation has been removed by a time series model. For autoregressive models the function

$$\sigma(z; x_1, x_2; a_0, \ldots, a_P, \beta) = \text{ssq}(z - x_1 a_0) + \text{ssq}\left(z - x_2\beta - \sum_{p=1}^{P} B^p x_1 a_p\right)$$

can be minimized over the relevant parameters. After optimal transforms are obtained and after checking the residuals for white noise, regression weights are found by projecting x_1 onto the space spanned by $[x_2, Bx_1, \ldots, B^P x_P]$. Coefficient β indicates the corrected change in mean level and can then be tested for significance. Van Buuren (1990, p. 98–103; 1996) contain examples of this application.

4.2 *Smoothing categorical time series*

Suppose that x is smoothed by a running mean smoother

$$v_t = \sum_{p=-P}^{P} x_{t-p} a_p$$

where $a_{-P}, \ldots, a_P$ are known filter weights that determine the precise properties of the smoother. Some well known choices correspond to the running average filter, the Hanning filter, the Spencer 15-point filter and the Gaussian kernel. A nice overview of such techniques can be found in Goodall (1990). Minimizing

$$\sigma(z; x) = \sum_{p=-P}^{P} \text{ssq}(z - B^p x a_p)$$

over z and $x = Gy$ then defines $v = \sum_{p=-P}^{P} B^p x a_p$ as the filtered series. This technique seems especially useful to quantify univariate series for which no additional information is available, other than being smooth. No applications of this technique have yet seen the light.

4.3 *Predictable components*

Box and Tiao (1977) proposed a canonical analysis that extracts predictable components from multivariate time series. The first predictable component is a linear combination of the original series that forecasts itself as well as possible. Like principal components analysis, the second component optimizes the same criterion, but under the condition that it is orthogonal to the first. The technique can be used as a dimension reduction device to bring out the major time dependent characteristics of a multivariate data set.

Let X of order $T \times M$ contain M quantified series sampled at T points of time. Suppose that x_t can be modelled by the multivariate autoregressive process $x_t = \Phi_1 x_{t-1} + \cdots + \Phi_P x_{t-P} + e_t$, where $\Phi_1, \ldots, \Phi_P$ are $M \times M$ matrices, and where e_t is an M-component white noise process. Box and Tiao (1977) show that this

multivariate autoregressive process can be reparametrized as a collection of M uncoupled univariate autoregressive processes on some new series $v_{1t}, \ldots, v_{Mt}$. The transforms works by finding linear combinations $v_j = Xa_j$ for $j = 1, \ldots, M$ that are contemporaneously independent, that is, $E[v_j' v_{j'}] = 0$ for $j \neq j'$, and that are ordered according to their predictability. The predictability measure γ_j reflects how much the jth component can predict itself by a univariate Pth order autoregressive model $v_{t,j} = f_1 v_{t-1,j} + \cdots + f_P v_{t-P,j} + \tilde{e}_{t,j} = \hat{v}_{t,j} + \tilde{e}_{t,j}$, where f_p are scalar autoregressive weights. Let $\hat{\sigma}_j^2 = E[\hat{v}_{t,j}^2]$, and let $\sigma_j^2 = E[v_{t,j}^2]$, then the predictability for v_j is equal to $\gamma_j = \hat{\sigma}_j^2 / \sigma_j^2$, which is the proportion of variance of v_j explained by the systematic part $\hat{v}_j$. For the first predictable component, the goal is to find a weight vector a_1 such that the linear combination $v_1 = Xa_1$ has maximum predictability γ_1. Next, a second predictable component, orthogonal to the first, can be identified, and so on. To see how this problem can be solved within the present framework write all components simultaneously as

$$V = \sum_{p=1}^{P} B^p V F_p + E$$

Since $V = XA$, this can be rephrased in terms of the observed data as

$$XA = \sum_{p=1}^{P} B_p XAF_p + E = \sum_{p=1}^{P} B^p XA_p + E$$

where $A_p = AF_p$. It is now easy to see that the problem of determining maximum predictability is equivalent to finding the largest canonical correlations between X and $[BX, \ldots, B^P X]$. The relationship with canonical correlation analysis has been studied by PARZEN and NEWTON (1980) and VELU, REINSEL, and WICHERN (1986). The latter authors found that γ_j is equal to the squared canonical correlation. For categorical data, the problem is now to minimize the loss

$$\sigma(Z; X; A_0, \ldots, A_P) = \text{ssq}(Z - XA_0) + \text{ssq}\left(Z - \sum_{p=1}^{P} B_p XA_p\right)$$

over Z, X and $A_0, \ldots, A_P$, with orthogonal Z. The predictable components are then equal to $V = XA_0$. Since V approaches the orthogonal matrix Z the components themselves are nearly orthogonal. An application of this technique to multivariate categorical time series can be found in VAN BUUREN (1992).

4.4 *Spatial models*

In time series analysis, observations are linked in the direction of time by means of the backshift matrix. Spatial dependency on a two-dimensional surface is more complex since observations may influence each other in several directions simultaneously. Examples of spatial dependency occur in agriculture, where experiments plots have

common borders, in the analysis of social networks, and in the study of environmental pollution. It is often possible to code dependencies among analysis units by means of an adjacency matrix, a more general form of the backshift matrix. Using the loss function as before is straightforward since the minimization procedure does not use the fact that B is a special matrix. The algorithm holds for any real $T \times T$ matrix B. There is no experience with this particular application.

4.5 *Cross-sectional multivariate analysis*

The loss function approach as outlined in section 2 can also be extended to multivariate analysis for cross-sectional data. For example, this can be done by replacing the lagged variables by conventional multi-attribute data. Actually, all methods described in this paper are special cases of the canonical class as defined in VAN BUUREN (1990, p. 128). This class also generalizes OVERALS, the most flexible of Gifi's techniques, and thus automatically covers the special cases like nonlinear discriminant analysis, regression analysis, homogeneity analysis, principal components analysis, canonical correlation analysis and MANOVA (cf. GIFI, 1990, p. 329). The estimation procedure for the canonical class is known. The major technical contribution of the method is that it allows to equate the transformations of an arbitrary subset of variables. This property was used in autoregression to equalize the transformation functions of the $P + 1$ different lags of the same series. It is also useful for the analysis of ranking data, missing data and event history data (VAN BUUREN and DE LEEUW, 1992).

5 Discussion

This paper describes a method for transforming categorical time series that are measured on nominal, ordinal and numerical scales. The underlying model consists of two main pieces: a scaling component and a linear time series component. An alternating least squares algorithm was derived that transforms the data such that it is optimal with respect to the linear model. Some extensions to other than autoregressive models were also indicated.

Questions related to the minimum order of the model can be handled by the iterative Box–Jenkins strategy based on autocorrelations and partial autocorrelations. A complication is that transforming a series changes its autocorrelation. Therefore autocorrelations are not comparable across models that depend on different transformations. Thus, using autocorrelations for identification is questionable. Experience shows that transformations primarily depend on one or two best fitting time lags. Therefore, cautioned use of Box–Jenkins identification techniques is possible as long as the most influential predictors are preserved.

Stochastic variation due to finite sampling has played no role so far. The method does not provide standard errors of the estimates, so it is not possible to assess the accuracy of the results or to test for statistical significance. One approach would be to obtain asymptotic standard errors from conventional linear time series analysis

applied to the optimally scaled data. This method effectively uses the quantified data as if they had been real. Despite some work in this area (e.g. DE LEEUW, 1988), such a two-step procedure is not recommended in general since the effects that scaling may have on the estimates and their standard errors are not well understood. A safer way is to use bootstrap methods for the autoregressive model as in EFRON and TIBSHIRANI (1993, p. 92–102). It is not correct to resample from the individual observations as in conventional bootstrap methods since this destroys the serial correlation. Instead, one may use a moving block bootstrap, a model-free resampling method in which not individual observations, but entire blocks of observations are sampled. The method assumes that the length of the block is sufficient to preserve the time dependent relations. This technique is promising, but not much practical experience is currently available. HJORTH (1994) contains further methods for resampling time series.

The method is currently limited to categorical data. For continuous data, it is certainly worthwhile to replace G in $x = Gy$ by a matrix of B-splines. An advantage of this is that the transformation function becomes smoother by borrowing strength from adjacent values. Furthermore, no arbitrary coding of continuous variables into categories is needed anymore. The y-values in such a setup correspond to the knot locations of the spline. These locations can be optimized to fit a linear model. For homogeneity analysis, this was done by VAN RIJCKEVORSEL (1987). The generalization is straightforward and easy to compute. It is more difficult to incorporate moving average terms into the model. The major problem is that this leads to a nonlinear optimization problem that cannot be solved by least squares. Thus, it is not yet possible to integrate optimal scaling and full ARMA modeling. Another generalization would be to include not only to analyze lags of x, but also lags of z. For example, optimizing over lagged z open up a whole box of interesting techniques like exponential smoothing, (replicated) dynamic factor analysis and state space models. A more elaborate account of this potential is nonetheless beyond the scope of this paper.

References

ANSLEY, C. F., W. A. SPIVEY and W. J. WROBLESKI (1977), A class of transformations for Box-Jenkins seasonal models, *Applied Statistics* **26**, 173–178.

BISHOP, Y. M. M., S. E. FIENBERG and P. W. HOLLAND (1975), *Discrete multivariate analysis: theory and practice*, MIT Press, Cambridge, MA.

BOX, G. E. P. and G. W. JENKINS (1976), *Time series analysis: forecasting and control*, Revised edition, Holden-Day, San Francisco.

BOX, G. E. P. and G. C. TIAO (1977), A canonical analysis of multiple time series, *Biometrika* **64**, 355–365.

VAN BUUREN, S. (1990), Optimal scaling of time series, Dissertation, University of Utrecht. DSWO Press, Leiden.

VAN BUUREN, S. (1992), Predictable linear combinations from categorical time series, *Statistica Applicata: Italian Journal of Applied Statistics* **4**, 743–751.

VAN BUUREN, S. (1996), Fitting ARMA time series by structural equation models, *Psychometrika*, in press.

VAN BUUREN, S. and J. DE LEEUW (1992), Equality constraints in multiple correspondence analysis, *Multivariate Behavioral Research* **27**, 567–583.

DEVILLE, J-C. and G. SAPORTA (1983), Correspondence analysis with an extension towards nominal time series, *Journal of Econometrics* **22**, 169–189.
EFRON, B. and R. J. TIBSHIRANI (1993), *An introduction to the bootstrap*, Chapman and Hall, London.
FAHRMEIR, L. (1992), State space modeling and conditional mode estimation for categorical time series, in: D. R. Brillinger et al. (eds.), *New directions in time series analysis, part I* (pp. 87–109), Springer, Berlin.
GHADDAR, D. K. and H. TONG (1981), Data transformation and self-exciting threshold autoregression, *Applied Statistics* **30**, 238–248.
GIFI, A. (1990), *Nonlinear multivariate analysis*, Wiley, Chichester.
GLASS, G. V., V. L. WILLSON and J. M. GOTTMAN (1975), *Design and analysis of time series experiments*, Colorado Associated University Press, Boulder, Colorado.
GOODALL, C. (1990), A survey of smoothing techniques, in: J. Fox and J. Scott Long (eds.), *Modern methods of data analysis*, Sage, London, 126–176.
GREGSON, R. A. M. (1987), The time-series analysis of self-reported headache sequences, *Behaviour Change* **4**, 6–13.
GRUBB, H. (1992), Review of "Optimal scaling of time series" by S. van Buuren, *Journal of the Royal Statistical Society, Series A* **155**, 179–180.
HAND, D. J., F. DALY, A. D. LUNN, K. J. MCCONWAY and E. OSTROWSKI (1993), *A handbook of small data sets*, Chapman and Hall, London.
HARVEY, A. C. and C. FERNANDES (1989), Time series models for count of qualitative observations (with discussion), *Journal of Business and Economics Statistics* **7**, 407–422.
HINKLEY, D. (1977), On quick choice of power transformation, *Applied Statistics* **26**, 67–69.
HJORTH, J. S. U. (1994), *Computer intensive statistical methods*, Chapman and Hall, London.
JACOBS, P. A. and P. A. W. LEWIS (1978), Discrete time series generated by mixtures, I: correlational and runs properties, *Journal of the Royal Statistical Society, Series B* **40**, 94–105.
DE LEEUW, J. (1977), Applications of convex analysis to multidimensional scaling, in: J. R. Barra, F. Brodeau, G. Romier and B. van Cutsum (eds.), *Recent developments in statistics*, North-Holland, Amsterdam, 137–145.
DE LEEUW, J. (1988), Multivariate analysis with linearizable regressions, *Psychometrika* **53**, 437–454.
DE LEEUW, J., F. W. YOUNG and Y. TAKANE (1976), Additive structure in qualitative data: an alternating least squares method with optimal scaling features, *Psychometrika* **41**, 471–503.
DE LEEUW, J. and W. J. HEISER (1980), Multidimensional scaling with restrictions on the configuration, in: P. R. Krishnaiah (ed.), *Multivariate analysis V*, North-Holland, Amsterdam, 501–522.
OWEN, A. (1983), Optimal transformations for autoregressive time series models, Report LCSD013, Stanford University, Dept. of Statistics.
PARZEN, E. and H. J. NEWTON (1980), Multiple time series modelling II, In: P. R. Krishnaiah (ed.), *Multivariate analysis V*, North Holland, Amsterdam, 107–197.
RAVEH, A. and C. S. TAPIERO (1980), Periodicity, constancy, heterogeneity and the categories of qualitative time series, *Ecology* **61**, 715–719.
VAN RIJCKEVORSEL, J. L. A. (1987), The application of fuzzy coding and horseshoes in multiple correspondence analysis, Dissertation, University of Leiden, DSWO Press, Leiden.
SINGH, A. C. and G. LEMAITRE (1987), Categorical auto-regressive models for analysing panel survey data, *ASA Proceedings of the Section of Survey Research Methods*, 390–394.
STOFFER, D. S. (1991), Walsh-Fourier analysis and its statistical applications, *Journal of the American Statistical Association* **86**, 461–479.
VELU, R. P., G. C. REINSEL and D. W. WICHERN (1986), Reduced rank regression models for multiple time series, *Biometrika* **73**, 105–118.
YOUNG, M. R. (1990), Estimating optimal transformations for correlation and coherence, in: C. Page and R. LePage (eds.), *Computing science and statistics: Proceedings of the Symposium on the Interface*, Springer, Berlin, 571–575.

Received: June 1994. Revised: April 1996.

Appendix: Procedure to minimize (6) over y

Let $C^s = B^s G$ and let $y = y_o + (y - y_o) = y + \delta$, where y_o is some old solution satisfying all appropriate constraints. Then (6) can be written as a function of y only as

$$\begin{aligned}\sigma(y) &= \mathrm{ssq}(z - C^0(y_o + \delta)a_0) + \mathrm{ssq}\left(z - \sum_p C^p(y_o + \delta)a_p\right) \\ &= \mathrm{ssq}((z - C^0 y_o a_0) - C^0\delta a_0) + \mathrm{ssq}\left(\left(z - \sum_p C^p y_o a_0\right) - \sum_p C^p \delta a_0\right)\end{aligned}$$

Let $p_1 = z - C^0 y_o a_0$ and $p_2 = z - \sum_p C^p y_o a_0$ then

$$\begin{aligned}\sigma(y) &= \sigma(y_o) - p_1' C^0 \delta a_0 - p_2'\left(\sum_p C^p \delta a_p\right) + \mathrm{ssq}(C^0\delta a_0) + \mathrm{ssq}\left(\sum_p C^p \delta a_p\right) \\ &= \sigma(y_o) - 2\delta' u + \delta' W \delta,\end{aligned}$$

where $u = C^{0\prime} p_1 a_0 + \Sigma_p C^{p\prime} p_2 a_p$ and $W = (a_0 \otimes C^0)'(a_0 \otimes C^0) + (\Sigma_p a_p \otimes C^p)'(\Sigma_p a_p \otimes C^p)$, and where $\otimes$ stands for the Kronecker product. Let $\lambda^2(W)$ denote the largest eigenvalue of the symmetric matrix W, and choose a constant $\alpha \geq \lambda^2(W)$. Because $\delta' W \delta \leq \alpha \delta' \delta$ for symmetric W it follows that $\sigma(y) \leq \sigma(y_o) - 2\delta' u + \alpha\delta'\delta$. The problem is now to minimize the quantity $-2\delta' u + \alpha\delta'\delta$ over δ, as this lowers the upper bound on $\sigma(y)$. Remember that $\delta = y - y_o$, and define $y_u = u/\alpha$. Then

$$\begin{aligned}\alpha\delta'\delta - 2\delta' u &= \alpha\big((y - y_o)'(y - y_o) - 2(y - y_o)' y_u + y_{u'} y_u - y_{u'} y_u\big) \\ &= \alpha\big(y - (y_o + y_u)\big)'\big(y - (y_o + y_u)\big) - \alpha y_{u'} y_u.\end{aligned}$$

Since $y_u y_u$ is fixed, the problem reduces to finding the minimum of

$$\big(y - (y_o + y_u)\big)'\big(y - (y_o + y_u)\big)$$

over y. If y is not subject to constraints, the solution of this problem is simply to set $y = y_o + y_u$. For ordinal and numerical scales, monotone and linear constraints on y are applied by regression as in Gifi (1990, p. 169–170). Constant α is often taken as $\alpha = \lambda^2(W)$ since this gives maximal update increments.