A SIMILARITY BASED MODEL FOR ORDERED CATEGORICAL DATA

By

Gaby Gayer, Offer Lieberman and Omer Yaffe

June 3, 2016

RESEARCH INSTITUTE FOR ECONOMETRICS

DISCUSSION PAPER NO. 3-14-R2

DEPARTMENT OF ECONOMICS
BAR-ILAN UNIVERSITY
RAMAT-GAN 5290002, ISRAEL

http://econ.biu.ac.il/en/node/2473

A Similarity Based Model for Ordered Categorical Data

Gabi Gayer*, Offer Lieberman†and Omer Yaffe‡

Revised, June 3, 2016

Abstract

In a large variety of applications the data for a variable we wish to explain is ordered and categorical. In this paper we present a new similarity-based model for the scenario and investigate its properties. We establish that the process is ψ-mixing and strictly stationary and derive the explicit form of the autocorrelation function (ACF) in some special cases. Consistency and asymptotic normality of the maximum likelihood estimator (MLE) of the model's parameters are proven. A simulation study supports our findings. The results are applied to the Netflix data set, comprised of a survey on users' grading of movies.

Key words and phrases: Consistency; Ergodicity; Mixing; Ordered Probit; Similarity; Stationarity.

JEL Classification: C22

*Department of Economics, Bar-Ilan University

†Department of Economics and Research Institute for Econometrics (RIE), Bar-Ilan University. Support from Israel Science Foundation grant No. 1082/14 and from the Sapir Center in Tel Aviv University are gratefully acknowledged. Correspondence to: Department of Economics, Bar-Ilan University, Ramat Gan 52900, Israel. E-mail: offer.lieberman@biu.ac.il

‡Department of Economics, Bar-Ilan University

1 Introduction

In a large number of applications the data for a variable we wish to explain is ordered and categorical. Examples include the level of education or income attained, the amount of insurance coverage purchased, voting for candidates that are positioned from left to right, corporate bond ratings and questionnaire based survey response coding. For all these examples and more, the statistical workhorse is undoubtedly the ordered probit model. In this paper we introduce a novel similarity based modeling alternative for such situations and investigate its properties.

Our model may be applied to circumstances where the decision of which product to purchase depends on recommendations of others. These situations often arise in connection with the consumption of a new product or of a product that is consumed only once and for which there is uncertainty about its quality. An example for such a situation is the Netflix data set, comprised of a survey on users' grading of movies. In order to predict the grade a user would assign to a particular movie at time t, one would analyze the grades assigned to this movie by other users with similar tastes. The prediction of a grade a user will assign to a movie at time t, is based on the grades assigned to this movie prior to this date, by other users who share similar tastes. Similarity of tastes of two users can be measured by their ranking of other movies prior to time t.

Similarity based models were introduced to economics by Gilboa *et. al.* (2006), applied in the context of real estate prices by Gayer *et. al.* (2007), and suggested as an approach for prediction by Gilboa *et. al.* (2011). Furthermore, Gilboa *et. al.* (2010) discussed the relevance of empirical similarity to the definition of objective probabilities. For the model

$$Y_t = \frac{\sum_{i<t} s\left(X_i, X_t; w\right) Y_i}{\sum_{i<t} s\left(X_i, X_t; w\right)} + \varepsilon_t, t = 2, ..., n, \tag{1}$$

where s is a similarity function, X_i is the ith observation on K explanatory variables, w is a $K \times 1$ parameter vector and ε_t is an iid random variable, the asymptotic theory of estimation was established by Lieberman (2010)

and this work has been extended by Lieberman (2012) and Lieberman and Phillips (2014) to the time-varying coefficient, non-stationary autoregression

$$Y_t = \mu + s_t\left(X_i, X_t; w\right) Y_{t-1} + \varepsilon_t, t = 2, ..., n,$$

where $s_t\left(\cdot\right)$ is possibly time varying. The latter model has been applied to Japanese dual stock data[1] and to international Exchange Traded Funds (ETFs). We remark that the issues that surface as well as the methods of proof used in Lieberman (2012) and Lieberman and Phillips (2014) are very different from the ones in the present paper, which deals with ordered categorical data. Finally, the concepts of similarity and contagion of views are central in Kapetanios *et. al.* (2013), who constructed a nonlinear panel data model of cross-sectional dependence.

For the ordered-probit model, applications are numerous and the topic is covered in many text books, see, for example, Maddala (1983), Cameron and Trivedi (2005), Greene (2008) and Greene and Hensher (2010). These applications include, among other things, the effects of family background on schooling choices (Cameron and Heckman, 1998), self-reports of levels of work disability in different countries (Kapteyn, Smith and Soest, 2007) and estimating the conditional distribution of trade-to-trade price changes that develop in discrete increments (Hausman, Lo, and MacKinlay, 1992). For an extensive list of applications the reader is referred to Greene and Hensher (2010) and the references therein.

Recently, De Jong and Woutersen (2011) investigated the binary time series

$$Y_t = 1\left\{\sum_{j=1}^{p} \rho_j Y_{t-j} + \gamma' x_n + \varepsilon_t > 0\right\} \tag{2}$$

where $1\left\{\cdot\right\}$ is the indicator function, taking the value of unity if the condition in the brackets is satisfied and zero otherwise, and $\varepsilon_t \overset{iid}{\sim} N\left(0, 1\right)$. They proved near epoch dependence and strong mixing of the process, as well as consistency and asymptotic normality of the MLE of $\beta = \left(\rho_1, ..., \rho_p, \gamma'\right)'$.

[1] This data set is on stocks which are traded in both Tokyo and New York.

The extension of De Jong and Woutersen's (2011) method of proof to our multi-category ordered setting, involving similarity weighted averages in place of the ρ_j's, does not appear to be trivial and our main proofs, especially on the ψ-mixing[2] and strict stationarity properties of the process, make extensive use of Markov chain theory.

The plan for this paper is as follows. In Section 2 we introduce the probabilistic model in detail and in Section 3 we investigate cases in which the ordered probit model may fail whereas our model is expected to deliver desirable predictions. The interpretation of partial derivatives in the model is discussed in Section 4. The model's assumptions are specified in Section 5. In Section 6 we use Markov chain theory to establish that the process is ψ-mixing and strictly stationary and derive an explicit form for the autocorrelation function (ACF) in the case where the response variable depends only on one lag. Consistency and asymptotic normality theorems of the MLE follow in Section 7. Simulations are presented in Section 8 and an application to the Netflix data set follows in Section 9. Section 10 concludes. The proofs related to Section 6 are provided in the Appendix, whereas those related to Section 7 as well as Tables and Figures relevant to Section 8 are given in an online supplement to the paper.

2 The Ordered Similarity Model

To fix ideas, we start by presenting a special case of our M-category model, when it is restricted to only two categories, 0 and 1, in which case,

$$Y_t = 1\left\{\bar{Y}_{t-1}^s + \varepsilon_t > \mu\right\}, (t = \lambda + 1, ..., n),$$

where

$$\bar{Y}_{t-1}^s = \rho \frac{\sum_{i=t-\lambda}^{t-1} s\left(X_i, X_t; w\right) Y_i}{\sum_{i=t-\lambda}^{i=t-1} s\left(X_i, X_t; w\right)}, \tag{3}$$

[2] We remark that, as is well known, ψ-mixing is stronger than uniform mixing, see, for instance, Fan and Yao (2005, pp. 68–69). For an extensive dicussion on properties of various mixing types and the relationships between them, the reader is refered to Bradley (2005).

$s(\cdot)$ is a similarity function, $w = (w_1, ..., w_K)'$, X_i is the ith observation on a K-vector of explanatory variables, $\varepsilon_t \overset{iid}{\sim} N(0,1)$, μ is a cut-off parameter, which may or may not be known, $\lambda \geq 1$ is the lag-length and ρ is a free parameter which is allowed to be greater than-, equal to- or less than unity. For brevity, we have suppressed the dependence of $\bar{Y}^s_{t-1}$ on ρ, λ and w. The model is entirely analogous to the probit model apart from the fact that $X_t'\beta$ in the latter is replaced by a similarity weighted average, $\bar{Y}^s_{t-1}$, in the former. Conditions on $s(\cdot)$ will be given in Section 6. For instance, we may specify an exponential similarity, viz.,

$$s(X_i, X_t; w) = \exp\left(-\sum_{j=1}^{K} w_j (X_{ij} - X_{tj})^2\right), \tag{4}$$

so that, *ceteris paribus*, the closer X_i will be to X_t, the larger will be the weight that Y_i will receive, relative to other the Y_j's, in the construction of $\bar{Y}^s_{t-1}$.

Let

$$W_t = \left(Y_t, X_t'\right), t = 1, ..., n,$$

$$\zeta_t = \left(W_t, ..., W_{t-(\lambda-1)}\right), t \geq \lambda \tag{5}$$

and let $\mathcal{F}_{t-1}$ be the σ-field based on all the information in $(W_{t-1}, \ldots, W_1)$. We have,

$$\begin{aligned} \Pr\left(Y_t = 1 | X_t = x, \mathcal{F}_{t-1}; w\right) &= \Pr\left(Y_t = 1 | X_t = x, \zeta_{t-1} = \zeta; w\right) \\ &= \Phi\left(\bar{Y}^s_{t-1}(x, \zeta) - \mu\right), \end{aligned} \tag{6}$$

where $\Phi(\cdot)$ is the standard normal CDF and $\bar{Y}^s_{t-1}(x, \zeta)$ is $\bar{Y}^s_{t-1}$ evaluated at ($X_t = x$, $\zeta_{t-1} = \zeta$). It is emphasized that given the entire history, the conditional probability (6) depends only on the most recent λ-lags of W_t and on the present value X_t. It is evident from (6) that given a μ, the larger is the value of $\bar{Y}^s_{t-1}$, the higher is the probability that Y_t will be equal to unity. In the formula for $\bar{Y}^s_{t-1}$ (given in (3)), the weight assigned to each past Y_i, with $t - \lambda \leq i \leq t - 1$, is a function of the closeness of X_i to X_t (and of the

other X_j's $j \neq i$, through the normalizing factor in the denominator of (3)). The higher is the degree of the closeness through the similarity function, the larger is the weight assigned to Y_i. In contrast, in the dynamic model (2) the weights assigned to past Y_i's are fixed and in the ordered probit model the probability that Y_t is equal to unity is only a function of the current X_t and not a function of past Y_t's.

Extending the idea to M ordered categories, $j = 1, ..., M$, our model is

$$\begin{aligned} Y_t = j \times 1\left\{\bar{Y}_{t-1}^s + \varepsilon_t \in (\mu_{j-1}, \mu_j]\right\}, \ \ (j = 1, ..., M), \\ (\varepsilon_t, X_t) \ \text{ is iid, } \varepsilon_t \text{ is } N(0,1) \ \text{ conditional on all } X_t, \\ t = \lambda + 1, ..., n, \ \ \left(-\infty = \mu_0 < \mu_1 < \cdots < \mu_{M-1} < \mu_M = \infty\right). \end{aligned} \tag{7}$$

In addition, throughout the paper we shall assume that the X_t's are bounded, iid random variables, copies of X, with a finite state space $\mathcal{S}_X$, and a strictly positive probability of being at ν, $\forall \nu \in \mathcal{S}_X$. For simplicity of the exposition we shall assume that the distribution of X_t does not depend on unknown parameters, although this assumption can be easily relaxed. An assumption on the distribution of the initial value of the process, $(W_\lambda, W_{\lambda-1}, \ldots, W_1)$, which we shall denote by π, will be given in Section 5. Let $\theta = (\mu', w', \rho)'$, with $\mu = \left(\mu_1, ..., \mu_{M-1}\right)'$. We will treat in the analysis the whole vector θ as unknown, even though in some applications, the researcher may wish to assume that a subset of it is known, particularly μ and/or ρ. We have

$$\begin{aligned} \Pr\left(Y_t = j | X_t = x, \mathcal{F}_{t-1}; \theta\right) &= \Pr\left\{\bar{Y}_{t-1}^s + \varepsilon_t \in (\mu_{j-1}, \mu_j] | X_t = x, \zeta_{t-1} = \zeta\right\} \\ &= \Phi\left(\mu_j - \bar{Y}_{t-1}^s(x, \zeta)\right) - \Phi\left(\mu_{j-1} - \bar{Y}_{t-1}^s(x, \zeta)\right) \\ &= \Delta_j\left(X_t = x, \zeta_{t-1} = \zeta; \theta\right), \end{aligned} \tag{8}$$

say. For brevity, we will simply write $\Delta_{t,j}(\theta)$. The likelihood function is

given by

$$L_n(\theta) = \prod_{j_n \in \mathcal{S}_Y} \prod_{\nu_n \in \mathcal{S}_X} \cdots \prod_{j_1 \in \mathcal{S}_Y} \prod_{\nu_1 \in \mathcal{S}_X} \Pr(Y_n = j_n, X_n = \nu_n, \ldots, Y_1 = j_1, X_1 = \nu_1)^{\prod_{r=1}^{n} 1\{Y_r = j_r\} 1\{X_r = \nu_r\}}$$

$$= \prod_{j_n \in \mathcal{S}_Y} \prod_{\nu_n \in \mathcal{S}_X} \cdots \prod_{j_1 \in \mathcal{S}_Y} \prod_{\nu_1 \in \mathcal{S}_X} \{\pi(Y_\lambda = j_\lambda, X_\lambda = \nu_\lambda, ..., Y_1 = j_1, X_1 = \nu_1)$$

$$\prod_{t=\lambda+1}^{n} (\Delta_{j_t}(X_t = \nu_t, Y_{t-1} = j_{t-1}, X_{t-1} = \nu_{t-1}, ...,$$

$$Y_{t-\lambda} = j_{t-\lambda}, X_{t-\lambda} = \nu_{t-\lambda}) \Pr(X_t = \nu_t))\}^{\prod_{r=1}^{n} 1\{Y_r = j_r\} 1\{X_r = \nu_r\}}, \qquad (9)$$

where $\mathcal{S}_Y = \{1, ..., M\}$ and we have used the assumption that the X_t's are iid. Therefore, the log-likelihood can easily be shown to reduce to

$$l_n(\theta) = \sum_{t=\lambda+1}^{n} \sum_{j \in \mathcal{S}_Y} 1\{Y_t = j\} \log \Delta_{t,j}(\theta) + l_{n,X} + l\pi, \qquad (10)$$

where

$$l_{n,X} = \sum_{t=\lambda+1}^{n} \sum_{\nu \in \mathcal{S}_X} 1\{X_t = \nu\} \log \Pr(X_t = \nu)$$

and

$$l\pi = \sum_{j_\lambda \in \mathcal{S}_Y} \sum_{\nu_\lambda \in \mathcal{S}_X} \cdots \sum_{j_1 \in \mathcal{S}_Y} \sum_{\nu_1 \in \mathcal{S}_X} \left(\prod_{r=1}^{\lambda} 1\{Y_r = j_r\} 1\{X_r = \nu_r\} \right) \log \pi(Y_\lambda = j_\lambda, X_\lambda = \nu_\lambda, ..., Y_1 = j_1, X_1 = \nu_1).$$

3 Special Cases

The purpose of this section is to draw the connection between the similarity model and the probit model and indicate, where possible, for which scenarios our model is likely to be superior to the ordered probit model and vice versa. To do so, consider the case in which there are two categories:

the value of the lower category being 0 and of the upper category being 1. Furthermore, assume that there is only one X, which is fixed at unity, making the similarity function constant. When $\mu = 1/2$, the similarity model reduces to

$$\Pr\left(Y_n = 1 | X_n = 1, \mathcal{F}_{n-1}\right) = \Phi\left(\bar{Y}_{n-1} - 1/2\right).$$

If $\bar{Y}_{n-1} > 1/2$, we set the predicted value for Y_n, $\hat{Y}_n^s$, to be $\hat{Y}_n^s = 1$, and zero otherwise. On the other hand, the ordered probit model predicts

$$\Pr\left(Y_n = 1 | X_n = 1\right) = \Pr\left(\beta + \varepsilon_n > 1/2\right) = \Phi\left(\beta - 1/2\right).$$

The solution to the score equation in the probit model for this setting is easily seen to be $\beta = \Phi^{-1}\left(\bar{Y}_n\right) + 1/2$. The ordered probit prediction is thus given by

$$\hat{Y}_n = 1 \Leftrightarrow \Phi\left(\hat{\beta} - 1/2\right) > 1/2 \Leftrightarrow \bar{Y}_n > 1/2.$$

Both models predict $\hat{Y}_n = 1$ if the sample mean is greater than $1/2$, but the predicted probabilities are different. In a sample of $n-1$ data points the similarity model's predicted probability is

$$\widehat{\Pr}\left(Y_n = 1 | X_n = 1, \mathcal{F}_{n-1}\right) = \Phi\left(\bar{Y}_{n-1} - 1/2\right),$$

whereas the ordered probit model's predicted probability amounts to

$$\widehat{\Pr}\left(Y_n = 1 | X_n = 1\right) = \bar{Y}_{n-1}.$$

The next example demonstrates certain circumstances in which the probit model fails but the similarity model succeeds. Typically, this would be the case when the ordered probit linear specification, $X_n'\beta$, in the conditional probability of Y_t, is not consistent with the underlying data generating process. To this end, suppose that $X_t = 0, -1$, or 1 and $Y_t = X_t^2$, $t = 1, ..., n$. Consequently,

$$Y_t = \frac{\sum_{i<t} 1\left\{X_i = X_t\right\} Y_i}{\sum_{i<t} 1\left\{X_i = X_t\right\}} + \varepsilon_t, t = n+1, ..., 2n,$$

where we have implicitly let $\lambda = t - 1$. Here, the similarity process starts after an initial 'learning set' which is based on n observations. Assume for simplicity that $\varepsilon_t = 0$, $\forall t$ and that the first n sample points gave exactly $n/4$ times $X_i = -1$, $n/2$ times $X_i = 0$ and $n/4$ times $X_i = 1$. Then it is clear that $Y_t = 1\{X_t = \pm 1\}$, $\forall t$. For some $\mu \in \mathbb{R}$, the solution to probit's score function in this case is

$$\frac{\left(1 - \Phi\left(\hat{\beta} - \mu\right)\right)\phi\left(\hat{\beta} - \mu\right)}{\Phi\left(\hat{\beta} - \mu\right)\left(1 - \Phi\left(\hat{\beta} - \mu\right)\right)} - \frac{\left(1 - \Phi\left(-\hat{\beta} - \mu\right)\right)\phi\left(-\hat{\beta} - \mu\right)}{\Phi\left(-\hat{\beta} - \mu\right)\left(1 - \Phi\left(-\hat{\beta} - \mu\right)\right)} = 0,$$

yielding $\hat{\beta} = 0$. Thus, the probit predicted probabilities are

$$\hat{Pr}\left(Y_{2n+1} = 1 | X_{2n+1}\right) = \Phi\left(-\mu\right).$$

With $\mu = 1/2$, the prediction rule here would be to set

$$\hat{Y}_{2n+1} = 1\left\{\Phi\left(-1/2\right) > 1/2\right\},$$

implying that the prediction is always $\hat{Y}_{2n+1} = 0$, which is correct for 50% of the observations. On the other hand, the similarity model predicts

$$\begin{aligned}
\widehat{\Pr}\left(Y_{2n+1} = 1 | X_{2n+1}, \mathcal{F}_{2n}\right) &= \Phi\left(\frac{\sum_{i<2n+1} 1\left\{X_i = X_{2n+1}\right\} Y_i}{\sum_{i<2n+1} 1\left\{X_i = X_{2n+1}\right\}} - \frac{1}{2}\right) \\
&= \Phi\left(\frac{1}{2}\right) 1\left\{X_{2n+1} = \pm 1\right\} \\
&\quad + \Phi\left(-\frac{1}{2}\right) 1\left\{X_{2n+1} = 0\right\}.
\end{aligned}$$

Now, set the rule:

$$\hat{Y}_{2n+1} = 1\left\{\widehat{\Pr}\left(Y_{2n+1} = 1 | X_{2n+1}, \mathcal{F}_{2n}\right) > \frac{1}{2}\right\}.$$

It follows that if $X_{2n+1} = \pm 1$, then $\widehat{\Pr}\left(Y_{2n+1} = 1 | X_{2n+1} = \pm 1, \mathcal{F}_{2n}\right) > 1/2$, giving $\hat{Y}_{2n+1} = 1$. But $Y_t = 1\{X_t = \pm 1\}$, $\forall t$, so the prediction is always correct when $X_{2n+1} = \pm 1$. If $X_{2n+1} = 0$, then $\widehat{\Pr}\left(Y_{2n+1} = 1 | X_{2n+1} = 0, \mathcal{F}_{2n}\right) <$

1/2, giving $\hat{Y}_{2n+1} = 0$. But, again, $Y_t = 1\{X_t = \pm 1\}$, $\forall t$, so the prediction is always correct when $X_{2n+1} = 0$ as well. In summary, the similarity based prediction is always correct in this example.

This example is indicative of the possible failure of the ordered probit model when the latent process is nonlinear in X and the similarity model is expected to outperform it in these scenarios. We emphasize that it is not expected that this will always be the case. Obviously, when the dgp is consistent with the probit specification, this model is expected to be superior to the similarity model.

4 Partial Derivatives

We have

$$\begin{aligned} \frac{\partial \Pr\left(Y_t = j | X_t, \mathcal{F}_{t-1}; \theta\right)}{\partial s\left(X_k, X_t; w\right)} &= \left(\phi\left(\mu_{j-1} - \bar{Y}_{t-1}^s\right) - \phi\left(\mu_j - \bar{Y}_{t-1}^s\right)\right) \quad (11) \\ &\quad \times \frac{Y_k - \bar{Y}_{t-1}^s}{\sum_{i<t} s\left(X_i, X_t; w\right)}, \end{aligned}$$

where $\phi(\cdot)$ is the standard normal pdf. Note that

$$\phi\left(\mu_{j-1} - \bar{Y}_{t-1}^s\right) \geq \phi\left(\mu_j - \bar{Y}_{t-1}^s\right) \text{ iff } \left|\mu_{j-1} - \bar{Y}_{t-1}^s\right| \leq \left|\mu_j - \bar{Y}_{t-1}^s\right|,$$

so that,

$$\frac{\partial \Pr\left(Y_t = j | X_t, \mathcal{F}_{t-1}; \theta\right)}{\partial s\left(X_k, X_t; w\right)} > 0 \text{ iff } \left|\mu_{j-1} - \bar{Y}_{t-1}^s\right| \leq \left|\mu_j - \bar{Y}_{t-1}^s\right| \text{ and } Y_k > \bar{Y}_{t-1}^s.$$

Consider, for instance, the two-category case. In the notation of equation (7), $\mu_0 = -\infty$ and $\mu_2 = \infty$, so that (11) becomes

$$\frac{\partial \Pr\left(Y_t = 1 | \mathcal{F}_{t-1}; \theta\right)}{\partial s\left(X_k, X_t; w\right)} = -\phi\left(\mu_1 - \bar{Y}_{t-1}^s\right) \frac{Y_k - \bar{Y}_{t-1}^s}{\sum_{i<t} s\left(X_i, X_t; w\right)} > 0, \text{ iff } Y_k < \bar{Y}_{t-1}^s.$$

In words, if Y_k is smaller than the (historical) similarity weighted average, increasing the similarity between Y_k and Y_t will increase the probability that

Y_t is equal to the lower category, 1, as expected. Similarly, for the higher category, 2, equation (11) becomes

$$\frac{\partial \Pr\left(Y_t = 2 | X_t, \mathcal{F}_{t-1}; \theta\right)}{\partial s\left(X_k, X_t; w\right)} = \phi\left(\mu_1 - \bar{Y}^s_{t-1}\right) \frac{Y_k - \bar{Y}^s_{t-1}}{\sum_{i<t} s\left(X_i, X_t; w\right)} > 0, \text{iff } Y_k > \bar{Y}^s_{t-1},$$

as expected.

For M categories the idea is similar. Increasing the similarity between the k-th and t-th observations will increase the probability of categories with values above the sample's similarity weighted average if and only if the value of Y_k itself is above this average and decrease the probability of categories with values below the sample's similarity weighted average.

5 Assumptions

In this section we set the assumptions which will be used in the various proofs of the paper. The parameter space is given by $\Theta = \Theta_1 \times \Theta_2 \times \Theta_3$, where Θ_1, Θ_2 and Θ_3 are the spaces in which μ, w and ρ are assumed to lie, respectively, μ is $(M-1)\times 1$, w is $K \times 1$ and ρ is a scalar. The true value of θ is denoted by θ_0. By $\bar{K}$ we denote a generic bounding constant, independent of n, which may vary from step to step. For the proof of consistency of the MLE, we shall require the following Assumptions.

Assumption A0: (ε_t, X_t) is iid, ε_t is iid $N(0,1)$ conditional on all X_t, for each $t = 1, ..., n$, the $K \times 1$ vector X_t are bounded, iid random variables copies of X, with a finite state space $\mathcal{S}_X$, and a strictly positive probability for being at ν, $\forall \nu \in \mathcal{S}_X$, and $Y_t \in \mathcal{S}_Y$, where $\mathcal{S}_Y$ is defined immediately following (9). If $w \neq w'$, $\Pr_w\left(\bar{Y}^s_{t-1}(w) \neq \bar{Y}^s_{t-1}(w')\right) > 0$, $\forall t$.

Assumption A1: The μ-vector satisfies

$$\left(-\infty = \mu_0 < \mu_1 < \cdots < \mu_{M-1} < \mu_M = \infty\right)$$

and there exist w_L, w_H, ρ_L and ρ_H such that for each $i = 1, ..., K$, $w_{i,0} \in [w_L, w_H]$, with $-\infty < w_L < w_H < \infty$ *and* $\rho \in [\rho_L, \rho_H]$, *with* $-\infty < \rho_L < \rho_H < \infty$.

For the derivation of the asymptotic distribution of the score and Hessian, we require the following additional assumptions:

Assumption (A2): *For all* i, t, k,

$$\sup_{i,t,k,\Theta} |\dot{s}_{w_k}(X_i, X_t; w)| < \bar{K} s(X_i, X_t; w) < \infty,$$

where $\dot{s}_{w_k}$ is the partial derivative of s wrt w_k.

Assumption (A3): The function $s_w(\cdot)$ is twice continuously differentiable in w for all X and Y.

We denote by $l_t(\theta)$ the log-likelihood per observation (excluding the terms corresponding to $l_{n,X}$ and $l\pi$), viz.,

$$l_t(\theta) = \sum_{j \in \mathcal{S}_Y} 1\{Y_t = j\} \log \Delta_{t,j}(\theta).$$

Assumption (A4): The vectors $\{\partial l_t(\theta)/\partial \theta_r\}_{t=\lambda+1}^{n}$, $r = 1, ..., M+K$, are linearly independent.

The last part of Assumption A0 is an identification condition. It is superfluous for $\lambda = 1$, in which case $\bar{Y}_{t-1}^{s} = Y_{t-1}$, which does not depend on w. In the proof of Theorem 1 it is shown that for any $\lambda > 1$, $\Pr_w(Y_t = j) > 0$ $\forall j \in \mathcal{S}_Y$ and $\forall t \geq \lambda$ and therefore, $\Pr_w\left(\bar{Y}_{t-1}^{s}(w) \neq \bar{Y}_{t-1}^{s}(w')\right) > 0$. A1 is a compactness assumption and Assumptions A2-A4 are satisfied for the exponential and inverse similarity functions as established in the following.

For the exponential similarity function (4),

$$\dot{s}_{w_k}(X_i, X_t; w) = -(X_{ik} - X_{tk})^2 s(X_i, X_t; w), \tag{12}$$

The inequality in A2 holds for (12) because X_t is bounded $\forall t$ under Assumption A0. Similarly, for the inverse similarity function

$$s(X_i, X_t; w) = \frac{1}{1 + \sum_{j=1}^{K} w_j (X_{ij} - X_{tj})^2}, \tag{13}$$

we have

$$\dot{s}_{w_k}(X_i, X_t; w) = -(X_{ik} - X_{tk})^2 s^2(X_i, X_t; w),$$

and A2 holds in this case as well. Assumptions A3 and A4 are analogous to Assumptions (2) and (5) of Proposition 7.9 of Hayashi (2000), respectively, the latter to ensure that the expected value of the normalized Hessian is nonsingular. Assumption A3 obviously holds for both (4) and (13). As for A4, it is clear from the proof of Lemma 1 of the Appendix, particularly equations (4), (6) and (7), that the assumption holds for (4) and (13), except for pathological cases, such as X_{it} is a constant, $\forall i$ and $\forall t$. In this case $s\left(X_i, X_t; w\right) = 1\ \forall i$ and $\forall t$, so that $\bar{Y}_{t-1}^{s} = \lambda^{-1} \sum_{i=t-\lambda}^{t-1} Y_i$, which does not depend on w.

6 ψ-Mixing, Strict Stationarity and the ACF

In this section we establish that the process is ψ-mixing and strictly stationary. Furthermore, we discuss the behavior of the ACF in the general case and derive its explicit form in the $\lambda = 1$ case. In particular, we show that the ACF in the $\lambda = 1$ and $M = 2$ case behaves in a completely analogous way to the behavior of the ACF of the dynamic AR(1) model.

In Theorem 1 we give the main result of this section by appealing to Markov chain theory. Starting from the conditional distribution of Y_t given X_t and the entire history of W_t, which depends only on X_t and on the most recent λ-lags of W_t, we first show that ζ_t, defined in (5), is a homogeneous Markov chain. The trick to transform what is essentially a λ-step Markov chain into a first order chain is standard, see for instance, Meyn and Tweedie (2009, pp. 24-25), or Bougerol and Picard (1992), the latter in the GARCH(p, q) context. Furthermore, the representation of a bivariate Markov chain leading to (5) is demonstrated in Lloyd (1974, Sec. 9).

Theorem 1 *For the model (7), under Assumptions A0-A1: (i) ζ_t is a homogenous Markov chain which admits a unique stationary distribution π. (ii) Under the assumption that the initial distribution of the Markov chain is π, the process is ψ-mixing and strictly stationary.*

A few comments are in place. First, in order to prove part (i) of Theorem 1, we show in the Appendix that $\{\zeta_t\}_{t \geq \lambda}$ is aperiodic and positive recurrent.

These properties imply, among other things, that starting from any given state, the chain will return to that state at some stage with probability one and that the mean return time is finite. This also implies that the chain does not have absorbing states, that is, there are no states from which the chain cannot escape. Secondly, ψ-mixing is stronger than uniform mixing and the latter together with strict stationarity imply ergodicity, see, for instance, White (2001, Theorem 3.44). We shall use ergodic stationarity in the proofs of consistency and asymptotic normality of the MLE in Section 7. Third, it is obvious from the proof that the results of the theorem are still valid if instead of normality it was assumed that ε_t has a continuous and strictly increasing CDF. Finally, a consequence of Theorem 1 is that ACF of the process decays geometrically. See, for instance, Lloyd (1974, Section 10). In the $\lambda = 1$ case we are able to provide the precise form of the ACF.[3]

Theorem 2 *For the model (7) with $\lambda = 1$, under Assumptions A0-A1,*

$$\begin{aligned} Cov\left(Y_{s+m}, Y_s\right) &= \sum_{l=1}^{M} l \Pr\left(Y_s = l\right)\left(1 - \Pr\left(Y_s = l\right)\right) \\ &\quad \times \sum_{j=1}^{M} j \Lambda_{s+m,s+m-1,j,l} \prod_{k=1}^{m-1} \Lambda_{s+k,s+k-1,l,l}, \end{aligned} \tag{14}$$

where

$$\Lambda_{t,s,j,l} = \left\{\Pr\left(Y_t = j | Y_s = l\right) - \Pr\left(Y_t = j | Y_s \neq l\right)\right\}. \tag{15}$$

If, in addition, $M = 2$ and s is large,

$$Cor\left(Y_{s+m}, Y_s\right) = Cor^m\left(Y_{s+1}, Y_s\right) = \left\{\Phi\left(\rho - \mu\right) - \Phi\left(-\mu\right)\right\}^m, m \in \mathbb{N}. \tag{16}$$

The result (16) is therefore completely analogous to the ACF of a linear AR(1) process and is of interest by its own right. Furthemore, it is obvious from the proof that the results (14) and $Cor\left(Y_{s+m}, Y_s\right) = Cor^m\left(Y_{s+1}, Y_s\right)$ still hold if the normality assumption is replaced by the assumption that ε_t has a continuous and strictly increasing CDF.

[3] We remark that in the $\lambda = 1$ case $\bar{Y}_{t-1} = Y_{t-1}$, which does not depend on X.

7 Consistency and Asymptotic Normality of the MLE

In this section we establish consistency and asymptotic normality of the MLE Our first result is consistency.

Theorem 3 *Under Assumptions A0-A1,* $\hat{\theta}_n \to_p \theta_0$.

We remark that the normality assumption of ε_t is not necessary for the proof of Theorem 3 and it can be relaxed so that it has a continuous and strictly increasing CDF. Ignoring $l_{n,X} + l\pi$ in (10), we denote the normalized score and Hessian components by

$$\begin{aligned} z_{n,\mu_k}(\theta) &= \frac{1}{\sqrt{n}} \frac{\partial l_n(\theta)}{\partial \mu_k}, (k = 1, ..., M-1), \\ z_{n,w_k}(\theta) &= \frac{1}{\sqrt{n}} \frac{\partial l_n(\theta)}{\partial w_k}, (k = 1, ..., K), \\ z_{n,\rho}(\theta) &= \frac{1}{\sqrt{n}} \frac{\partial l_n(\theta)}{\partial \rho}, \end{aligned}$$

$z_n(\theta) = \left(z_{n,w_1}(\theta), ..., z_{n,w_K}(\theta), z_{n,\mu_1}(\theta), z_{n,\mu_{M-1}}(\theta), z_{n,\rho}(\theta)\right)'$ and

$$H_{n,\theta_j,\theta_k}(\theta) = \frac{1}{n} \frac{\partial^2 l_n(\theta)}{\partial \theta_j \partial \theta_k},$$

respectively. Let $V(\theta_0)$ be the asymptotic Fisher's information matrix, with an n^{-1} normalization.

Asymptotic normality of the MLE is stated in the following theorem.

Theorem 4 *Under Assumptions A0-A4,* $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, V^{-1}(\theta_0)\right)$.

8 Simulations

All tables and figures relevant to this section are placed in an online supplement. The correlograms of the process for $\lambda = 1$ and $M = 2, 3$ are depicted

in Figures 1-2, respectively. In each case 10000 Y_t's were generated from i.i.d. standard normal ε_t's. It is obvious that the correlograms fade in a very similar fashion to the decay of the theoretical ACF of the linear AR(1) model, thus, supporting the result of Theorem 2.

In Tables 1-6 and in Figures 3-14 we summarize the simulation results for the performance of the MLE's of w and μ. Each setting consists of 2500 replications of the Y data series, generated from $N(0,1)$ ε_t's and with $X \sim [-1,1]$, with $n = 250,\ 500,\ 1000,\ 2000$, $w_0 = 1,\ 3,\ 5$, $\mu_0 = 0.3,\ 0.5$, and $\lambda = 2,\ 5,\ 10$. The choice of the lag-length covers at least part of the range that would be mostly used in applied work and its sole purpose in this context is to illustrate and support our analytical findings. In each case we report in the Tables the sample means, their standard deviations, the trimmed means with symmetric 5% trimming together with their standard deviations, the medians, first- and third quartiles.

Uniformly in all cases, as n increases the sample means over the 2500 replications converge to the true parameter values and their standard deviations decline, as expected. This also holds for the trimmed means and for both the estimates of w and of μ. The medians appear to be very close to the parameter values and the interquartile range becomes tighter in all settings as n increases.

The density estimates displayed in Figures 3-14 were constructed in MATLAB using a Gaussian kernel and Silverman's optimal bandwidth. Figures 3-6 correspond to the kernel density estimates for $\hat{w}$ in the case $\mu_0 = 0.3$, $w_0 = 1$ and $\lambda = 2$. Clearly, as n increases from 250 to 2000, the density becomes more symmetric around 1 and with much fewer outliers. The same conclusions hold qualitatively in Figures 7-10, corresponding to the case $\mu_0 = 0.5$, $w_0 = 3$ and $\lambda = 5$, and in Figures 11-14, corresponding to the $\mu_0 = 0.5$, $w_0 = 5$ and $\lambda = 10$ case. Overall, the simulations very much support the analytical results concerning the properties of the MLE.

9 An Empirical Application

The data was released by Netflix in order to improve their algorithm to predict customer rating. It consists of a survey of viewers' ranking of movies from 1998 to 2005. Movies belong to a class of items whose various components do not necessarily translate into success, therefore it is hard to find a general formula for tastes or rating of movies. However, it is reasonable to assume that people who shared similar tastes in the past will continue to do so, making the rating of movies a suitable application for a similarity-based model.

This evaluation process may be applied to the rating of other cultural items, such as works of art, music, literature, etc. Indeed this appears to be the rationale for Amazon's provision of information to potential customers on purchases made by other customers. For example, a customer considering the purchase of a particular book is given a list of other books that were also purchased by the purchasers of this book. Thus a customer is able to see whether his tastes are similar to those of the other purchasers of this book. Bobadilla *et. al.* (2013) reviews recent developments in recommendation systems.

9.1 Data

In 2006 the online DVD rental service Netflix ran a competition for the best algorithm to predict customer ratings of films (see Koren *et. al.*, 2009, for a description of the winning algorithm, which was further improved upon in Takacs *et. al.* 2008) The data set consists of four variables: user ID, movie title, the date on which the movie was rated, and the movie's rating - an integer between 1 and 5, with 1 corresponding to the lowest rating and 5 corresponding to the highest.

We started out with a subset of the Netflix data set, containing ratings made by 13,000 viewers of 99 movies,[4] of which only 14 were rated by all users. For the purpose of this exercise we estimated the model with only five

[4]The original database contains approximately 100 million ratings of 18,000 movies made by 500,000 viewers.

explanatory variables because it considerably simplifies the computations. Six movies out of the 14 were chosen arbitrarily, where one movie (Sweet Home Alabama) acted as the Y variable and the remaining 5 movies acted as the X variables (Independence Day, Pretty Woman, Forrest Gump, The Green Mile, and Con Air). The observations were ordered by the date Y was ranked. Moreover, at time t, the viewer must have watched all movies corresponding to the X variables in order to be able to make similarity comparisons. We further restricted the viewer of time t to have watched the movies corresponding to the X variables before the viewer of time $t' > t$. Those observations that did not satisfy these conditions were excluded from the database. Sweet Home Alabama was chosen to be the dependent variable as it was released much later than the other movies, making it more likely to be viewed last. The model was estimated on the first 1,000 observations. We remark that the observations are not all equidistant from each other. However, for this application it is inconsequential because once a rating is made, it is unlikely to change over time. Therefore, we did not require the data to satisfy this condition. We further note that only 1086 days have passed between the first and 1000th observation, where 85% of the observations were made within one day from each other, 97% within a week and less than 1% of the observations' time difference exceeded two weeks.

9.2 Model Estimation

The similarity-based model, being a weighted average of past observations, uses the last λ observations for prediction. In the estimation we used the exponential similarity function (4) that was shown to satisfy the assumptions in Section 3. We refer to the first n observations as the train set and the $(n+1)$th observation as the test set. This was repeated for $n = 900, ..., 999$, so that the model was estimated 100 times making a one-step ahead prediction each time. The similarity model was estimated with λ set to 5, 10, and 20. The parameter estimates together with their associated t-ratios based on $n = 1,000$ are provided in Table 8. Interestingly, the estimated coefficient of Pretty Woman, $\hat{w}_2$, was larger than those of the other movies,

so this movie was found to be the most suitable for predicting Sweet Home Alabama. Moreover, it was found to be significant for λ equal to 10, and 20, whereas the other movies turned out to be insignificant for all λ-choices. Indeed, out of the six movies, these two are the closest in terms of category classification. Finally, The estimate $\hat{\rho}$ was significant for all λ-choices.

The study uses two methods to generate one-step ahead predictions:

1) $\hat{Y}_{t+1} = j \times 1\left\{\bar{Y}_t^s\left(\hat{w},\hat{\rho}\right) \in (\hat{\mu}_{j-1},\hat{\mu}_j]\right\}$, $(j = 1, ..., M)$

2) $\tilde{Y}_{t+1} = \arg\max_{j=1,...,M} \Pr\left(\bar{Y}_t^s\left(\hat{w},\hat{\rho}\right) + \varepsilon_{t+1} \in (\hat{\mu}_{j-1},\hat{\mu}_j]\right)$

These were compared to predicting the outcome of Y_{t+1} according the sample's mode in the first n observations. The hit percent, defined as the ratio of correct predictions to the total number of observations, was computed for the predictions based on $\hat{Y}$, $\tilde{Y}$ and the mode. Table 7 contains their values for $\lambda = 5$, 10, 20. These estimates were computed for both the train data set that contains the first 900 observations and the test set that contains the remaining 100 observations. As can be seen from table 7, the hit percent of the similarity based-model was considerably larger than that of the mode prediction both in the train- and in the test set, representing an improvement of 3% to 27% across the different settings, with the similarity model gaining more advantage as λ increases, although the increase in this advantage appears to be diminishing with the increase in λ.

10 Conclusions

In the context of decision making the data is frequently ordered and categorical, as in the choice of education level and consumer satisfaction surveys. In this paper we presented a similarity-based model that can be applied to this type of ordered data. Its key aspect is that the dependent variable Y is assumed to be determined by outcomes of similar past observations, as opposed to the ordered probit model which typically assumes that Y only depends on the independent variables. It seems reasonable that if the evaluating agent has a well-defined method for rating, the ordered probit model would better explain the data. However, if the objects that the evaluating agent is rating are abstract (making the ranking process more complicated),

then the agent may very well rely on other people's evaluations. Gilboa *et. al.* (2006), Gayer *et. al.* (2007), and Gilboa *et. al.* (2013) refer to a similarity-based model as case-based reasoning and to the ordered probit model as rule-based reasoning and discuss the circumstances of when one mode of reasoning will dominate the other. The results of this paper suggest that the similarity-based model provides a potentially very useful framework for analyzing and forming accurate predictions for data formed by case-based reasoning.

References

Amemiya, T. (1985) *Advanced Econometrics*. Cambridge: Harvard University Press.

Athreya K.B. & S.G. Pantula (1986) Mixing properties of Harris chains and autoregressive processes. *Journal of Applied Probability* 23, 880-892.

Bobadilla J. , F. Ortega, A. Hernando & A. Gutierrez (2013) Recommender systems survey. *Knowledge-Based Systems* 46, 109-132.

Bougerol, P. & N. Picard (1992) Stationarity of GARCH processes and some nonnegative time series. *Journal of Econometrics* 52, 115–127.

Bradley, R.C. (2005) Basic properties of strong mixing conditions: A survey and some open questions. *Probability Surveys* 2, 107–144.

Cameron, A.C. & P.K. Trivedi (2005) *Microeconometrics: Methods and Applications*. Cambridge University Press: New York.

Cameron, S. & J. Heckman (1998) Life cycle schooling and dynamic selection bias: models and evidence for five cohorts of American males. *Journal of Political Economy* 106, 262–333.

De Jong, R.M. & T. Woutersen (2011) Dynamic time series binary choice. *Econometric Theory* 27, 673–702.

Fan, J. & Q. Yao (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer: New York.

Gayer, G., I. Gilboa & O. Lieberman (2007) Rule-based and case-based reasoning in real estate prices. *The B.E. Journals in Theoretical Economics* 7, No. 1 (Advances), Article 10.

Gilboa, I., O. Lieberman & D. Schmeidler (2006) Empirical similarity. *The Review of Economics and Statistics* 88, 433–444.

Gilboa, I., O. Lieberman & D. Schmeidler (2010) On the definition of objective probabilities by empirical similarity. *Synthese* 172, 79–95.

Gilboa, I., O. Lieberman & D. Schmeidler (2011) A similarity-based approach to prediction. *Journal of Econometrics* 162, 124–131.

Gilboa, I., L. Samuelson, & D. Schmeidler (2013) Dynamics of inductive inference in a unified model. *Journal of Economic Theory* 148, 1399–1432.

Greene, W.H. (2008) *Econometric Analysis*, 7th Edition. Prentice Hall.

Greene, W.H. & D.A. Hensher (2010) *Modeling Ordered Choices: A Primer*. Cambridge University Press.

Hamilton, J. (1994) *Time Series Analysis*. Princeton University Press.

Hausman, J.A, A.W. Lo & A.C MacKinlay (1992) An ordered probit analysis of transaction stock prices. *Journal of Financial Economics* 31, 319–379.

Hayashi, F. (2000) *Econometrics*. Princeton University Press.

Kapteyn, A., J. Smith & A. Van Soest (2007) Vignettes and self-reports of work disability in the United States and the Netherlands. *American Economic Review* 97, 461–473.

Koren, Y., B. Robert & C. Volinsky (2009) Matrix factorization techniques for recommender systems. *IEEE, Computer Journal*, 42(8), 30-37.

Lieberman, O. (2010) Asymptotic theory for empirical similarity models. *Econometric Theory* 26, 1032–1059.

Lieberman, O. (2012) A similarity-based approach to time-varying coefficient nonstationary autoregression. *Journal of Time Series Analysis* 33, 484–502.

Lieberman, O. & P.C.B. Phillips (2014) Norming rates and limit theory for some time-varying coefficient autoregressions. *Journal of Time Series Analysis* 35, 592–623.

Lloyd, E.H. (1974) What is, and what is not, a Markov chain? *Journal of Hydrology*, 22, 1–28.

Kapetanios, G., J. Mitchell & Y. Shin (2013) A nonlinear panel data model of cross-sectional dependence. *Journal of Econometrics* 179, 134–157.

Maddala, G.S. (1983) *Limited-Dependent and Qualitative Variables in Economics*. Cambridge University Press: New York.

McLeish, D.L. (1974) Dependent central limit theorems and invariance principles. *The Annals of Probability* 2, 620–628.

Meyn, S. & R.L. Tweedie (2009) *Markov Chains and Stochastic Stability*, second edition. Cambridge University Press: New York.

Newey, W.K. & D. McFadden (1994) Large sample estimation and hypothesis testing. In R.F. Engle & D. McFadden (eds.), *Handbook of Econometrics* 4, North-Holand.

Resnick, S.I. (2002) *Adventures in Stochastic Processes*. Springer: New York.

Takacs G., I. Pilaszy, B. Nemeth & D. Tikk (2008) Matrix factorization and neighbor based algorithms for the netflix prize problem. *In Proceedings of the 2008 ACM conference on Recommender systems*, 267-274.

White, H. (2001) *Asymptotic Theory for Econometricians*, revised edition. Academic Press.

Wu, C.F. (1981) Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics* 9, 501–513.

Appendix A: Stationarity and Ergodicity

Proof of Theorem 1: The proof of the theorem proceeds along the following lines. First, we prove that ζ_t is a homogeneous Markov chain. Next, we show that the chain is irreducible, and use this to establish that the chain is positive recurrent and aperiodic. In turn, positive recurrence imply that the chain admits a unique stationary distribution π and it is well known that as a process, it is strictly stationary as long as the initial distribution of the process is π. The result on ψ-mixing of the process follows from the fact that the chain is irreducible and aperiodic.

Let

$$\gamma = (\gamma_1, ..., \gamma_\lambda)\,,\ \tau = (\tau_1, ..., \tau_\lambda)\,,$$

where $\gamma_1 = (\gamma_{11}, \gamma_{12})$, the γ_j's are $1 \times (K+1)$, $j = 1, ..., \lambda$, γ_{11} is a scalar and γ_{12} is $1 \times K$. By (5), (7) and (8), for any $\gamma, \tau_j \in (\mathcal{S}_Y \times \mathcal{S}_X)^\lambda$, $j = 1, ..., t-1$, and for any $t \geq \lambda + 1$,

$$\Pr\left(W_t = \gamma_1, ..., W_{t-(\lambda-1)} = \gamma_\lambda | W_{t-1} = \tau_1, ..., W_1 = \tau_{t-1}\right), \qquad (17)$$

is zero, unless $(\gamma_2, ..., \gamma_\lambda) = (\tau_1, ..., \tau_{\lambda-1})$. Equation (8) implies that (17) is equal to

$$\begin{aligned}
&\Pr\left(W_t = \gamma_1, W_{t-1} = \tau_1, ..., W_{t-(\lambda-1)} = \tau_{\lambda-1} | W_{t-1} = \tau_1, ..., W_{t-\lambda} = \tau_\lambda\right) \\
&= \Pr\left(\zeta_t = (\gamma_1, \tau_1, ..., \tau_{\lambda-1}) \,|\zeta_{t-1} = (\tau_1, ..., \tau_\lambda)\right) \\
&= \Pr\left(W_t = \gamma_1 | W_{t-1} = \tau_1, ..., W_{t-\lambda} = \tau_\lambda\right) \\
&= \Pr\left(Y_t = \gamma_{11} | X_t = \gamma_{12}, W_{t-1} = \tau_1, ..., W_{t-\lambda} = \tau_\lambda\right) \Pr\left(X_t = \gamma_{12}\right) \\
&= \Delta_{\gamma_{11}}\left(X_t = \gamma_{12}, \zeta_{t-1} = \tau\right) \Pr\left(X = \gamma_{12}\right),
\end{aligned}$$

where we have used the assumption that the X_t's are iid, copies of X. In other words, the conditional probability of ζ_t given the past depends only on ζ_{t-1} and is not a function of t. Thus, ζ_t is a homogeneous Markov chain. Next, we shall prove that $\{\zeta_t\}_{t \geq \lambda}$ is aperiodic and positive recurrent. For

any $\gamma, \tau \in (\mathcal{S}_Y \times \mathcal{S}_X)^\lambda$, we have

$$\begin{aligned} p_\lambda^{\lambda+s} &\equiv \Pr\left(\zeta_{\lambda+s} = \gamma | \zeta_\lambda = \tau\right) \\ &= \Pr\left(W_{\lambda+s} = \gamma_1, ..., W_{s+1} = \gamma_\lambda | W_\lambda = \tau_1, ..., W_1 = \tau_\lambda\right). \end{aligned} \tag{18}$$

We consider the following cases.

(a) The case $\lambda = 1$ and $s \geq 1$. Here,

$$\begin{aligned} p_1^{1+s} = &\sum_{\nu_1,...,\nu_{s-1} \in \mathcal{S}_Y \times \mathcal{S}_X} \Pr\left(W_{1+s} = \gamma | W_s = \nu_1\right) \Pr\left(W_s = \nu_1 | W_{s-1} = \nu_2\right) \\ &\cdots \Pr\left(W_2 = \nu_{s-1} | W_1 = \tau\right). \end{aligned} \tag{19}$$

Now,

$$\begin{aligned} \Pr\left(W_{1+s} = \gamma | W_s = \nu_1\right) &= \Pr\left(Y_{1+s} = \gamma_{11} | X_{1+s} = \gamma_{12}, W_s = \nu_1\right) \\ &\quad \times \Pr\left(X_{1+s} = \gamma_{12}\right) \\ &= \Delta_{\gamma_{11}}\left(X_{1+s} = \gamma_{12}, \zeta_s = \nu_1\right) \Pr\left(X = \gamma_{12}\right), \end{aligned}$$

which, by (7) and (8), is strictly in $(0, 1)$. Similarly, each of the probabilities in (19) is strictly in $(0, 1)$. Hence, $0 < p_1^{1+s} < 1$.

(b) The case $\lambda \geq 2$ and $1 \leq s \leq \lambda - 1$. It is obvious from (18) that $p_\lambda^{\lambda+s}$ is zero unless $\gamma_{s+1} = \tau_1$, $\gamma_{s+2} = \tau_2$,...,$\gamma_\lambda = \tau_{\lambda-s}$. For this subcase,

$$\begin{aligned} &\Pr\left(W_{\lambda+s} = \gamma_1, ..., W_{s+1} = \gamma_\lambda | W_\lambda = \tau_1, ..., W_1 = \tau_\lambda\right) \\ = \;&\Pr\left(W_{\lambda+s} = \gamma_1 | W_{\lambda+s-1} = \gamma_2, ..., W_{\lambda+s-(s-1)} = \gamma_s,\right. \\ &\left. W_{\lambda+s-(s)} = \tau_1, \ldots, W_{\lambda+s-\lambda} = \tau_s\right) \cdots \\ &\Pr\left(W_{\lambda+s-(s-1)} = \gamma_s | W_{\lambda+s-s} = \tau_\lambda, ..., W_{\lambda+s-(s+\lambda-1)} = \tau_1\right). \end{aligned}$$

Each of the probabilities above is strictly in $(0, 1)$ and so, for this subcase, $0 < p_1^{1+s} < 1$ as well.

(c) The case $\lambda \geq 2$ and $s = \lambda$. Here

$$\begin{aligned} p_{\lambda}^{\lambda+s} &= \Pr\left(W_{\lambda+s} = \gamma_1 | W_{\lambda+s-1} = \gamma_2, ..., W_{\lambda+1} = \gamma_s, W_{\lambda} = \tau_1\right) \\ &\quad \cdots \Pr\left(W_{\lambda+1} = \gamma_s | W_{\lambda} = \tau_1, ..., W_1 = \tau_{\lambda}\right), \end{aligned}$$

which is strictly in $(0,1)$ $\forall \gamma, \tau \in (\mathcal{S}_Y \times \mathcal{S}_X)^{\lambda}$.

(d) The case $\lambda \geq 2$, $s > \lambda$. We have

$$\begin{aligned} p_{\lambda}^{\lambda+s} &= \sum_{\nu_1,...,\nu_{s-\lambda} \in \mathcal{S}_Y \times \mathcal{S}_X} \Pr\left(W_{\lambda+s} = \gamma_1 | W_{\lambda+s-1} = \gamma_2, ..., \right. \\ &\quad \left. W_{\lambda+s-(\lambda-1)} = \gamma_{\lambda}, W_{\lambda+s-\lambda} = \nu_1\right) \\ &\quad \Pr\left(W_{\lambda+s-1} = \gamma_2 | W_{\lambda+s-2} = \gamma_3, ..., W_{\lambda+s-(\lambda-1)} = \gamma_{\lambda}, \right. \\ &\quad \left. W_{\lambda+s-\lambda} = \nu_1, W_{\lambda+s-(\lambda+1)} = \nu_2\right) \\ &\quad \cdots \Pr\left(W_{\lambda+1} = \nu_{s-\lambda} | W_{\lambda} = \tau_1, ...W_1 = \tau_{\lambda}\right), \end{aligned}$$

which is strictly in $(0,1)$ $\forall \gamma, \tau \in (\mathcal{S}_Y \times \mathcal{S}_X)^{\lambda}$.

The conclusion is that the chain can reach any state with positive probability from any starting state in $s = \lambda$-steps. In other words, the chain is irreducible (see, for instance, Meyn and Tweedie (2009, p. 14)). It is well known that an irreducible finite-state Markov chain is always positive recurrent. Moreover, cases (a)-(d) above show that $p_{\lambda}^{\lambda+s}$ is strictly positive for all $s \geq \lambda \geq 1$ and therefore, by Definition 8.2 of Fan and Yao (2005), the chain is aperiodic.

Next, by Theorem 10.0.1, of Meyn and Tweedie (2009), if the chain is positive recurrent, then it admits a unique stationary distribution (invariant measure) π. In turn, if the Markov chain has a stationary distribution, then as process, it is strictly stationary, as long as the initial distribution of the chain is the stationary distribution. See, for instance, Meyn and Tweedie (2009, p. 231), or Proposition 2.12.1 of Resnick (2002). Furthermore, by Theorem 15.0.1, of Meyn and Tweedie (2009), a positive recurrent chain is geometrically ergodic. Athreya and Pantula (1986) showed that an ergodic chain is always strong mixing (see also p. 419 of Meyn and Tweedie (2009)), and by Theorem 3.1 of Bradley (2005) the Markov chain is in fact ψ-mixing

iff the chain is irreducible and aperiodic. This completes the proof of the Theorem. ■

Proof of Theorem 2: For $t > s$, we have

$$\begin{aligned}
Cov\left(Y_t, Y_s\right) &= \sum_{j,l=1}^{M} jl \left(\Pr\left(Y_t = j, Y_s = l\right) - \Pr\left(Y_t = j\right)\Pr\left(Y_s = l\right)\right) \\
&= \sum_{j,l=1}^{M} jl \Pr\left(Y_s = l\right)\left(\Pr\left(Y_t = j | Y_s = l\right) - \Pr\left(Y_t = j\right)\right) \\
&= \sum_{j,l=1}^{M} jl \Pr\left(Y_s = l\right)\{\Pr\left(Y_t = j | Y_s = l\right) \\
&\quad - \Pr\left(Y_t = j | Y_s = l\right)\Pr\left(Y_s = l\right) \\
&\quad - \Pr\left(Y_t = j | Y_s \neq l\right)\Pr\left(Y_s \neq l\right)\} \\
&= \sum_{j,l=1}^{M} jl \Pr\left(Y_s = l\right)\left(1 - \Pr\left(Y_s = l\right)\right)\Lambda_{t,s,j,l}, \qquad (20)
\end{aligned}$$

where $\Lambda_{t,s,j,l}$ is defined in (15). For the $\lambda = 1$ case,

$$\begin{aligned}
\Lambda_{s+2,s,j,l} &= \Pr\left(Y_{s+2} = j | Y_s = l\right) - \Pr\left(Y_{s+2} = j | Y_s \neq l\right) \\
&= \Pr\left(Y_{s+2} = j | Y_{s+1} = l, Y_s = l\right)\Pr\left(Y_{s+1} = l | Y_s = l\right) \\
&\quad + \Pr\left(Y_{s+2} = j | Y_{s+1} \neq l, Y_s = l\right)\Pr\left(Y_{s+1} \neq l | Y_s = l\right) \\
&\quad - \Pr\left(Y_{s+2} = j | Y_{s+1} = l, Y_s \neq l\right)\Pr\left(Y_{s+1} = l | Y_s \neq l\right) \\
&\quad - \Pr\left(Y_{s+2} = j | Y_{s+1} \neq l, Y_s \neq l\right)\Pr\left(Y_{s+1} \neq l | Y_s \neq l\right) \\
&= \Pr\left(Y_{s+2} = j | Y_{s+1} = l\right)\Pr\left(Y_{s+1} = l | Y_s = l\right) \\
&\quad + \Pr\left(Y_{s+2} = j | Y_{s+1} \neq l\right)\Pr\left(Y_{s+1} \neq l | Y_s = l\right) \\
&\quad - \Pr\left(Y_{s+2} = j | Y_{s+1} = l\right)\Pr\left(Y_{s+1} = l | Y_s \neq l\right) \\
&\quad - \Pr\left(Y_{s+2} = j | Y_{s+1} \neq l\right)\Pr\left(Y_{s+1} \neq l | Y_s \neq l\right) \\
&= \{\Pr\left(Y_{s+2} = j | Y_{s+1} = l\right) - \Pr\left(Y_{s+2} = j | Y_{s+1} \neq l\right)\} \\
&\quad \times \{\Pr\left(Y_{s+1} = l | Y_s = l\right) - \Pr\left(Y_{s+1} = l | Y_s \neq l\right)\} \\
&= \Lambda_{s+2,s+1,j,l}\Lambda_{s+1,s,l,l}
\end{aligned}$$

and so (14) follows on using (20).

In the special case where $\lambda = 1$ and $M = 2$, there are no X_t's in the model. Recoding the categories to be 0 (lower) and 1 (higher) and setting $\mu_1 = \mu$, we obtain

$$\Pr(Y_{s+1} = 1|Y_s = 1) - \Pr(Y_{s+1} = 1|Y_s = 0) = \Phi(\rho - \mu) - \Phi(-\mu).$$

The autocovariance in this case reduces to

$$Cov(Y_{s+m}, Y_s) = \Pr(Y_s = 1)(1 - \Pr(Y_s = 1))\{\Phi(\rho - \mu) - \Phi(-\mu)\}^m, m \in \mathbb{N}.$$

By Theorem 1(ii), for large enough s, $\Pr(Y_s = 1)$ is independent of s, which implies (16). ■

Table 7. Hit for the similarity based model and mode predictions.

Database	**Prediction Method**	$\lambda = 5$	$\lambda = 10$	$\lambda = 20$	Mode
Train	$\hat{Y}$	0.35	0.35	0.36	0.34
Train	$\tilde{Y}$	0.36	0.38	0.38	
Test	$\hat{Y}$	0.37	0.37	0.38	0.33
Test	$\tilde{Y}$	0.39	0.4	0.42	

Note: λ is the lag-length; $\hat{Y}$, $\tilde{Y}$, and Mode are given in Section 6.2.

Table 8. Estimates and t-ratios for the similarity-based model.

	$\lambda = 5$	$\lambda = 10$	$\lambda = 20$
$\hat{w}_1$	0.0000	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0001)
$\hat{w}_2$	18.4492	2.7920	1.8512
	(1.4423)	(2.9672)	(3.7347)
$\hat{w}_3$	0.5268	0.0244	0.0000
	(0.9121)	(0.1131)	(0.0000)
$\hat{w}_4$	4.8478	0.3254	0.3548
	(1.4674)	(0.9862)	(1.9125)
$\hat{w}_5$	3.6731	0.0533	0.1808
	(1.3778)	(0.3939)	(0.9316)
$\hat{\mu}_1$	-1.0384	-0.3233	0.5061
	(-6.3958)	(-1.2432)	(1.3206)
$\hat{\mu}_2$	-0.1949	0.5408	1.3802
	(-1.3119)	(2.1211)	(3.5986)
$\hat{\mu}_3$	0.7564	1.5041	2.3567
	(5.0802)	(5.8409)	(6.0686)
$\hat{\mu}_4$	1.6742	2.4283	3.2986
	(10.9186)	(9.2974)	(8.3998)
$\hat{\rho}$	0.2766	0.4735	0.7022
	(7.2111)	(7.0345)	(6.9148)

Note: λ is the lag-length; t-ratios are given in brackets.