A Similarity Based Model for Ordered Categorical Data

Gabi Gayer*, Offer Lieberman† and Omer Yaffe‡

June 27, 2013

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 the rate of decay of the autocorrelation function (ACF) in the general case and derive its explicit form in some special cases. Stationarity and ergodicity of the process are proven, as well as consistency and asymptotic normality of the maximum likelihood estimator (MLE) of the model's parameters. 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. 396/10 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 econometric 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 (2013) 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 and to international Exchange Traded Funds (ETFs). 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 vast and the topic is covered in almost every microeconometrics text book, see, *inter alia*, Maddala (1983), Cameron and Trivedi (2005) and Greene (2008).

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

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

where $\varepsilon_t \overset{iid}{\sim} N\left(0,1\right)$. They proved near epoch dependence and strong mixing of the process, as well as consistency and asymptotic normality of the MLE of $\beta = \left(\rho_1, ..., \rho_p, \gamma'\right)'$. 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 rate of decay of the ACF, stationarity and ergodicity, adopt a different technique.

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. In Section 5 we establish the rate of decay of the ACF in the general case and derive its explicit form in some special cases. We show, in particular, that the ACF of the binary response model in which the response depends only on one lag, behaves in a completely analogous way

to the behavior of the ACF of the dynamic AR(1) model. This result does not appear to have been documented previously. Stationarity and ergodicity of the process are proven. The model's assumptions necessary for the proofs of consistency and asymptotic normality of the MLE are specified in Section 6 and the theorems 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 and proofs are provided in the Appendix.

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,

$$Y_1 = 1\{\varepsilon_1 > 0\}$$

and

$$Y_t = 1\left\{\bar{Y}_{t-1}^s + \varepsilon_t > \mu\right\}, (t = 2, ..., n),$$

where

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

$1\{\cdot\}$ is the indicator function, taking the value of unity if the condition in the brackets is satisfied and zero otherwise, $s(\cdot)$ 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 $\bar{Y}_{t-1}^s$ 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, $\bar{Y}_{t-1}^s$, in the former. Conditions on $s(\cdot)$ will be given in Section 5. For instance, we may specify an exponential similarity,

viz.,

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

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 $\bar{Y}_{t-1}^s$.

Let $\mathcal{F}_{t-1}$ be the σ-field based on all the information included up to time $t-1$. Then

$$\begin{aligned}\Pr\left(Y_t = 1 | \mathcal{F}_{t-1}, X_t; w\right) &= \Pr\left(\varepsilon_t > \mu - \bar{Y}_{t-1}^s\right) \\ &= \Phi\left(\bar{Y}_{t-1}^s - \mu\right).\end{aligned}$$

This means that, given a μ, the larger the value of $\bar{Y}_{t-1}^s$, the higher the probability that Y_t will be equal to unity. The model thus connects in a nonlinear way between the history of the Y_t's and the X_t's and the current value of Y_t, which is very different from the way in which the observations are generated in the probit specification.

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

$$Y_1 = j1\left\{\varepsilon_1 \in (\Phi^{-1}\left(\frac{j-1}{M}\right), \Phi^{-1}\left(\frac{j}{M}\right)]\right\}$$

and for $t = 2, ..., n$,

$$\begin{aligned}Y_t &= j1\left\{\bar{Y}_{t-1}^s + \varepsilon_t \in (\mu_{j-1}, \mu_j]\right\}, (j = 1, ..., M), \varepsilon_t \overset{iid}{\sim} N\left(0, 1\right), \\ &\quad \left(-\infty = \mu_0 < \mu_1 < \cdots < \mu_{M-1} < \mu_M = \infty\right). \qquad (2)\end{aligned}$$

Let $\theta = (\mu', w', \rho)'$, with $\mu = \left(\mu_1, ..., \mu_{M-1}\right)'$. We have

$$\begin{aligned}
\Pr\left(Y_t = j | \mathcal{F}_{t-1}, X_t; \theta\right) &= \Pr\left\{\bar{Y}^s_{t-1} + \varepsilon_t \in (\mu_{j-1}, \mu_j]\right\} \\
&= \Phi\left(\mu_j - \bar{Y}^s_{t-1}\right) - \Phi\left(\mu_{j-1} - \bar{Y}^s_{t-1}\right) \\
&= \Phi_{t,j}\left(X_1, ..., X_t, Y_1, ..., Y_{t-1}; \theta\right) \\
&\quad -\Phi_{t,j-1}\left(X_1, ..., X_t, Y_1, ..., Y_{t-1}; \theta\right) \\
&= \Delta_{t,j}\left(x_t, y_{t-1}; \theta\right), \qquad (3)
\end{aligned}$$

say, where $x_t = (X_1, ..., X_t)$ and $y_{t-1} = (Y_1, ..., Y_{t-1})$. For brevity, we will simply write $\Delta_{t,j}(\theta)$. The likelihood function is given by

$$L_n(\theta) = \prod_{t=1}^{n} \prod_{j=1}^{M} \left(\Pr\left(Y_t = j | \mathcal{F}_{t-1}; \theta\right)\right)^{1\{Y_t = j\}}$$

and therefore, the log-likelihood is

$$l_n(\theta) = \sum_{t=1}^{n} \sum_{j=1}^{M} 1\{Y_t = j\} \ln \Delta_{t,j}(\theta).$$

3 Special Cases

In this section we shall draw a connection between the similarity model and the probit model through a simple example and then proceed to demonstrate certain circumstances in which the former performs better than the latter.

To do so, consider the case in which there two ordered categories: the value of the lower category being 0 and of the upper category being 1. Furthermore, assume that there is only one X (that is equal to 1 for example) making the similarity function constant. Finally, when $\mu = 1/2$, the similarity model reduces to

$$\begin{aligned}
\Pr\left(Y_n = 1 | \mathcal{F}_{n-1}, X_n\right) &= \Pr\left(\bar{Y}_{n-1} + \varepsilon_n > 1/2\right) \\
&= \Phi\left(\bar{Y}_{n-1} - 1/2\right).
\end{aligned}$$

Thus, if $\bar{Y}_{n-1} > 1/2$, we set the predicted value for Y_n, $\hat{Y}^s_n$, to be $\hat{Y}^s_n = 1$

and zero otherwise.

On the other hand, the ordered probit model predicts

$$\begin{aligned}\Pr\left(Y_n = 1|X_n\right) &= \Pr\left(\beta + \varepsilon_n > 1/2\right) \\ &= \Phi\left(\beta - 1/2\right).\end{aligned}$$

The solution to the score equation in the probit model for this setting is easily seen to satisfy

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

or

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

The ordered probit prediction is thus given by

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

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

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

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.

Suppose that $X_t = 0$, -1, or 1 and $Y_t = X_t^2$, $t = 1, ..., n$. Consequently,

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

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 $Y_t = 1\{X_t = \pm 1\}$, $\forall t$.

Now, for some $\mu \in \mathbb{R}$ the probit model postulates

$$\Pr(Y_i = 1|X_i) = \Phi(X_i\beta - \mu).$$

The solution to the 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

$$\hat{Pr}(Y_{2n+1} = 1|X_i) = \Phi(-\mu), \forall X_i$$

and with $\mu = 1/2$, the rule here would be to set

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

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

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

and setting the rule

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

gives a prediction which is always correct.

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.

4 Partial Derivatives

We have

$$\begin{aligned}\frac{\partial \Pr\left(Y_t = j|\mathcal{F}_{t-1};\theta\right)}{\partial s\left(X_k, X_t; w\right)} &= \left(\phi\left(\mu_{j-1} - \bar{Y}^s_{t-1}\right) - \phi\left(\mu_j - \bar{Y}^s_{t-1}\right)\right) \\ &\quad \times \frac{Y_k - \bar{Y}^s_{t-1}}{\sum_{i<t} s\left(X_i, X_t; w\right)},\end{aligned} \tag{4}$$

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

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

so that,

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

Consider, for instance, the two-category case. In the notation of eq'n (2), $\mu_0 = 0$ and $\mu_2 = \infty$, so that (4) becomes

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

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, (4) becomes

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

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 The ACF and Ergodic Stationarity

In this section we establish the rate of decay of the ACF in the general case and derive its explicit form in the $\lambda = 1$ case. We show, in particular, that the ACF of the model 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. This result does not appear to have been documented previously. We further establish stationarity and ergodicity.

5.1 Population Moments

Evidently, the conditional distribution (3) depends on t and therefore the process is not stationary for finite n. For the binary case, De Jong and Woutersen (2011) proved that the process is near epoch dependent and strong mixing. For the more general M-category ordered process, we have

$$\begin{aligned}\lambda_{t,j}^{w_0} &\equiv \Pr_{w_0}(Y_t = j) = \sum_{k \neq j} \Pr_{w_0}(Y_t = j | Y_{t-1} = k)\, \lambda_{t-1,k}^{w_0} \\ &\quad + \Pr_{w_0}(Y_t = j | Y_{t-1} = j)\, \lambda_{t-1,j}^{w_0} \\ &= a_t^{w_0} + b_t^{w_0} \lambda_{t-1,j}^{w_0},\end{aligned}$$

say, which converges to $\lambda_{\infty,j}^{w_0}$ because $0 < \delta_1 < b_t^{w_0} < 1 - \delta_2 < 1$ for some δ_1, $\delta_2 > 0$ and $\forall t$. It follows that for any $1 \leq h < \infty$,

$$E_{w_0}\left(Y_t^h\right) = \sum_{j=1}^{M} j^h \Pr_{w_0}(Y_t = j) \to_{n\to\infty} \sum_{j=1}^{M} j^h \lambda_{\infty,j}^{w_0}. \tag{5}$$

In words, all moments of the distribution of Y_t are asymptotically independent of n, a result which is in line with the findings of De Jong and Woutersen (2011).

5.2 The Autocovariance Function

In Theorem 1 a bound is placed on the rate of decay of the autocovariance function (ACV) and is proven in the Appendix.

Theorem 1 *For the model (2),* $\forall m \in \mathbb{N}$, $\exists x \in (0,1)$ *such that*

$$|Cov\left(Y_{s+m}, Y_s\right)| \leq x^m.$$

The implication of the result is that the ACF is absolutely summable and the process is covariance stationary. In the case $\lambda = 1$, we are able to provide the precise form of the ACF.

Theorem 2 *For the model (2) with* $\lambda = 1$,

$$\begin{aligned} Cov\left(Y_{s+m}, Y_s\right) &= \sum_{l=1}^{M} l \Pr\left(Y_s = l\right)\left(1 - \Pr\left(Y_s = l\right)\right) \\ &\quad \times \sum_{j=1}^{M} j \Lambda_{s+k,s+k-1,j,l} \prod_{k=1}^{m-1} \Lambda_{s+k,s+k-1,l,l}, \end{aligned} \tag{6}$$

where

$$\Lambda_{t,s,j,l} = \left\{\Pr\left(Y_t = j | Y_s = l\right) - \Pr\left(Y_t = j | Y_s \neq l\right)\right\}. \tag{7}$$

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

$$Cor\left(Y_{s+m}, Y_s\right) = Cor^m\left(Y_{s+1}, Y_s\right) = \left\{\Phi\left(\rho - \mu\right) - \Phi\left(-\mu\right)\right\}^m, m \in \mathbb{N}. \tag{8}$$

The result (8) is therefore completely analogous to the result for the ACF of a linear AR(1) process.

5.3 Ergodic Stationarity

The ACF of the Y_t's is absolutely summable, as has been established and therefore the process is ergodic for the mean. See, for instance, Hamilton (1994, pp. 46–47). For Gaussian processes, the absolute summability of the ACF is sufficient for complete ergodicity (of all moments). As Y_t is not

Gaussian, for ergodicity for all the moments we need to prove that for any bounded functions $f : \mathbb{R}^{k+1} \to \mathbb{R}$ and $g : \mathbb{R}^{l+1} \to \mathbb{R}$,

$$\begin{aligned} & \lim_{n\to\infty} (|E\left[f\left(Y_s, ..., Y_{s+k}\right) g\left(Y_{s+n}, ..., Y_{s+n+l}\right)\right]| \\ & - |E\left[f\left(Y_s, ..., Y_{s+k}\right)\right]| \, |E\left[g\left(Y_{s+n}, ..., Y_{s+n+l}\right)\right]|) \\ = \; & 0. \end{aligned} \tag{9}$$

Theorem 3 *: The process (2) is ergodic stationary.*

It is clear from the proof of Theorem 3, specifically, the bound placed on (13), that the same method of proof can be used to establish that the process is mixing. In turn, stationarity and mixing imply ergodicity, see, for instance, White (2001, Theorem 3.44). Ergodic stationarity will be used in the proofs of consistency and asymptotic normality of the MLE in Section 7.

6 Assumptions

In this section we set the assumptions which will be used in the proofs of consistency and asymptotic normality of the MLE. The parameter space is given by $\Theta = \Theta_1 \times \Theta_2 \times \Theta_3$, where Θ_1, Θ_2 are the spaces in which μ, w and ρ are assumed to lie, respectively. The true value of θ is denoted by θ_0. By 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: $\{\varepsilon_t\}_{t=1}^n$ is a sequence of $NID\left(0, 1\right)$. For each $t = 1, ..., n$, the $K \times 1$ vector X_t is nonstochastic, real and finite and $Y_t \in \{1, ..., M\}$. If $w \neq w'$, $\Pr_w\left(\bar{Y}_{t-1}^s\left(w\right) \neq \bar{Y}_{t-1}^s\left(w'\right)\right) = 1$, $\forall t$.

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,

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

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

Assumption (A4): The derivatives $\partial \bar{Y}^s_{t-1}/\partial \theta_k$, $k = M, ..., M+K$, are linearly independent.

The last part of Assumption A0 is an identification condition. Assumption A1 is a compactness assumption and Assumption A2 is satisfied for the exponential and inverse similarity functions. For the exponential similarity, for instance,

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

and the inequality holds because X_t is bounded $\forall t$ under Assumption A0. Similarly, for the inverse similarity function

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

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

and the inequality holds in this case as well. Finally, 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. Both assumptions hold for the exponential and inverse similarity functions.

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 4 *Under Assumptions A0-A1,* $\hat{\theta}_n \rightarrow_p \theta_0$.

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 5 *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

The correlograms of the process are depicted in Figures 1-6. In each case 10000 Y_t's were generated from i.i.d. standard normal ε_t's. We set $\lambda = 1, 2, 5$, $M = 2, 3$ and for simplicity, $\bar{Y}^s_{t-1} = \lambda^{-1} \sum_{i=t-\lambda}^{t-1} Y_i$. It is obvious that the correlograms decay rapidly, supporting Theorem 1. Moreover, as stated in Theorem 2, the correlogram in the $\lambda = 1$ case (Figures 1-2) fade in a

very similar fashion to the decay of the theoretical ACF of the linear AR(1) model.

In Tables 1-4 and in Figures 7-14 we summarize the simulation results for the performance of the MLE's of w and μ. Each setting consists of 1500 replications of the Y data series, generated from $N(0,1)$ ε_t's and with $X \sim [-1,1]$, which was generated once and consequently was held fixed in each iteration, with $n = 250,\ 500,\ 1000,\ 2000$, $w_0 = 1,\ 3,\ 5$, $\mu_0 = 0.3,\ 0.5$, and $\lambda = 2,\ 5$. 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 1500 replications converge to the true parameter values and their standard deviations decline, as expected. This holds also 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 7–14 were constructed in MATLAB using a Gaussian kernel and Silverman's optimal bandwidth. Figures 7–10 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 11-14, corresponding to the case $\mu_0 = 0.5$, $w_0 = 3$ and $\lambda = 5$. Overall, the simulations very much support the analytical results concerning the properties of the MLE.

9 An Empirical Application

The data on Netflix compiled by the authors 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.

9.1 Data

In 2006 the online DVD rental service Netflix ran a competition for the best algorithm to predict customer ratings of films. 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.

The goal of this exercise is to compare the performance of the ordered probit model to that of the similarity-based model. We started out with a subset of the Netflix data set, containing ratings made by 13,000 viewers of 99 movies,[1] of which only 14 were rated by all users. For the purpose of this exercise we estimated the model with only five explanatory variables as 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 condition were excluded from the database. Sweet Home

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

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 models were estimated on the first 1,000 observations.

9.2 Model Estimation

A standard method for comparison of two non-nested models is to estimate the models on a train data set, and then evaluate their predictions on a separate test data set. Indeed this is the technique used in the Netflix contest. In our case both the similarity-based model and the probit model use the entire train data to estimate the parameters. However, the similarity-based model, being a weighted average of past observations, continues to use observations in the test data set for prediction, giving it an advantage over the probit model. To eliminate this advantage we estimated the model on a data set of size t in order to predict the $t+1$ observation. This way both models used the same t observations for prediction. This was repeated for $t = 900, ..., 999$, so that each 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 average estimates of the parameters of the models appear in Table 6. Interestingly, the estimated coefficients of Pretty Woman, $\hat{w}_2$ for the similarity-based model and $\hat{\beta}_2$ for the ordered probit model, are the largest of all coefficient estimates, implying that both models have identified this movie to be the most suitable for predicting Sweet Home Alabama. Indeed, out of the six movies, these two are the closest in terms of category classification.

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

1) $\hat{Y}_{t+1} = j1\left\{\bar{Y}_t^s(\hat{w},\hat{\rho}) \in (\hat{\mu}_{j-1},\hat{\mu}_j]\right\}$, $(j = 1, ..., M)$, for the similarity-based model and $\hat{Y}_{t+1} = j1\left\{X'_{t+1}\hat{\beta} \in (\hat{\mu}_{j-1},\hat{\mu}_j]\right\}$, $(j = 1, ..., M)$ for the ordered probit model;

2) $\tilde{Y}_{t+1} = \arg\max_{j=1,...,M} \Pr\left(\bar{Y}_t^s(\hat{w},\hat{\rho}) + \varepsilon_{t+1} \in (\hat{\mu}_{j-1},\hat{\mu}_j]\right)$ for the similarity-based model and $\tilde{Y}_{t+1} = \arg\max_{j=1,...,M} \Pr\left(X'_{t+1}\hat{\beta} + \varepsilon_{t+1} \in (\hat{\mu}_{j-1},\hat{\mu}_j]\right)$ for the ordered probit model.

Two measures of model performance were considered:

1) Hit– the ratio of correct predictions to the total number of observations;

2) RMSE– root mean squared error of prediction;

These two measures were computed both for the train data and the test data. The results are provided in Table 5. As can be seen from the table, the similarity-based model has a slight disadvantage on the train data having a lower hit percent and a higher RMSE for both methods of prediction. However this disadvantage is offset by the advantage of the similarity-based model has when using the test data which yields a lower RMSE for both methods of prediction. The hit percent measure in the test data has mixed results, with the similarity-based model having a higher rate than the ordered probit model for one method of prediction, but a lower one for the other. These results hold for all lag-lengths, with the similarity model gaining a slightly bigger advantage as λ grows. It should be noted that all these advantages and disadvantages are of low magnitude. Overall, these results indicate that the performances of the similarity-based model and the ordered probit model are comparable, at least in this application.

10 Conclusions

In the context of decision making the data are 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 pro-

bit 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.

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

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

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, No.1, 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*, 7nd Edition. Prentice Hall.

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

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

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 (2013) Norming rates and limit theory for some time-varying coefficient autoregressions. Submitted for publication.

Kapetanios, G., J. Mitchell & Y. Shin (2013) A nonlinear panel data model of cross-sectional dependence. Mimeo.

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, No. 4, 620–628.

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

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: For $t > s$, we have

$$\begin{aligned} Cov\left(Y_t, Y_s\right) &= \sum_{j,l=1}^{M} jl\left(\Pr\left(Y_t = j, Y_s = l\right) - \Pr\left(Y_t = j\right)\Pr\left(Y_s = l\right)\right) \\ &= \sum_{j,l=1}^{M} jl\Pr\left(Y_s = l\right)\left(\Pr\left(Y_t = j|Y_s = l\right) - \Pr\left(Y_t = j\right)\right) \\ &= \sum_{j,l=1}^{M} jl\Pr\left(Y_s = l\right)\{\Pr\left(Y_t = j|Y_s = l\right) \\ &\quad - \Pr\left(Y_t = j|Y_s = l\right)\Pr\left(Y_s = l\right) \\ &\quad - \Pr\left(Y_t = j|Y_s \neq l\right)\Pr\left(Y_s \neq l\right)\} \\ &= \sum_{j,l=1}^{M} jl\Pr\left(Y_s = l\right)\left(1 - \Pr\left(Y_s = l\right)\right)\Lambda_{t,s,j,l}, \end{aligned} \tag{10}$$

where $\Lambda_{t,s,j,l}$ is defined in (7). For $t > s + \lambda$, let

$$A_t = \left\{Y_{t-1}^l = Y_{t-1}^{l^c}, Y_{t-2}^l = Y_{t-2}^{l^c}, ..., Y_{t-\lambda}^l = Y_{t-\lambda}^{l^c}\right\},$$

where Y_t^l, $Y_t^{l^c}$ are the Y_t's which were generated given $Y_s = l$ and $Y_s \neq l$, respectively. Notice that

$$A_t \Longrightarrow \left\{\left(\bar{Y}_{t-1}^s|Y_s = l\right) = \left(\bar{Y}_{t-1}^s|Y_s \neq l\right)\right\} \Longrightarrow \left\{Y_t^l = Y_t^{l^c}\right\}$$

and therefore,

$$A_t \Longrightarrow A_{t+1}, t > s + \lambda. \tag{11}$$

In other words, if $\exists T > s+\lambda$ such that the two series, $\left(Y_{T-1}^l, Y_{T-2}^l, ..., Y_{T-\lambda}^l\right)$ and $\left(Y_{T-1}^{l^c}, Y_{T-2}^{l^c}, ..., Y_{T-\lambda}^{l^c}\right)$, coincide, it will follow that $Y_t^l = Y_t^{l^c}$ $\forall t \geq T$. Hence,

$$A_T \Longrightarrow \Lambda_{t,s,j,l} = 0, \forall t \geq T. \tag{12}$$

Furthermore, as $\bar{Y}_{t-1}^{s} \in [\rho, \rho M]$ and $\varepsilon_t \in \mathbb{R}$ and in view of the restriction on the μ_j's implied by (2), for fixed $\lambda \in (0, \infty)$, $\exists x_U$ such that

$$\Pr\left(A_t^c\right) < x_U < 1, \forall t > s + \lambda.$$

Using (11),

$$\begin{aligned}
\Pr\left(A_{t+1}^c\right) &= \Pr\left(A_{t+1}^c | A_t^c\right) \Pr\left(A_t^c\right) + \Pr\left(A_{t+1}^c, A_t\right) \\
&= \Pr\left(A_{t+1}^c | A_t^c\right) \Pr\left(A_t^c\right) + \Pr\left(A_{t+1}^c, A_t, A_{t+1}\right) \\
&= \Pr\left(A_{t+1}^c | A_t^c\right) \Pr\left(A_t^c\right).
\end{aligned}$$

Consider the case $\lambda = 1$, $M = 2$. We have $Y_{s+1}^{l} = 1 + 1\{\rho Y_s^l + \varepsilon_{s+1} > \mu_1\}$ and $Y_{s+1}^{l^c} = 1 + 1\{\rho Y_s^{l^c} + \varepsilon_{s+1} > \mu_1\}$, so both $A_{s+2} = \left\{Y_{s+1}^{l} = Y_{s+1}^{l^c}\right\}$ and $A_{s+2}^c = \left\{Y_{s+1}^{l} \neq Y_{s+1}^{l^c}\right\}$ have positive probability. For the latter case, we can have $A_{s+3} = \left\{Y_{s+2}^{l} = Y_{s+2}^{l^c}\right\}$ or $A_{s+3}^c = \left\{Y_{s+2}^{l} \neq Y_{s+2}^{l^c}\right\}$, both with positive probability. More generally, $\exists z_U \in (0, 1)$ such that for each $t > s + \lambda$, $\Pr\left(A_{t+1}^c | A_t^c\right) < z_U < 1$ and therefore $\Pr\left(A_{t+1}^c\right) \leq z_U x_U$. This implies, in particular, that

$$\Pr\left(A_{s+\lambda+2}\right) = 1 - \Pr\left(A_{s+\lambda+2}^c\right) \geq 1 - x^2, x = \max\left\{z_U, x_U\right\}$$

and more generally,

$$\Pr\left(A_{s+\lambda+m}\right) \geq 1 - x^m, m \in \mathbb{N}, x \in (0, 1).$$

In view of (12),

$$\Pr\left(\cap_{t=s+\lambda+m}^{\infty}\left\{\Lambda_{t,s,j,l} = 0\right\}\right) \geq \Pr\left(A_{s+\lambda+m}\right) \geq 1 - x^m, m \in \mathbb{N}, x \in (0, 1),$$

implying that

$$\Pr\left(\cup_{t=s+\lambda+m}^{\infty}\left\{\Lambda_{t,s,j,l} \neq 0\right\}\right) = 1 - \Pr\left(\cap_{t=s+\lambda+m}^{\infty}\left\{\Lambda_{t,s,j,l} = 0\right\}\right) \leq x^m, (m = 1, 2, ...).$$

■

Proof of Theorem 2: For the $\lambda = 1$ case,

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

and so (6) follows on using (10).

In the special case where $\lambda = 1$ and $M = 2$, recoding the categories to be 0 (lower) and 1 (higher) and setting $\mu_1 = \mu$, we obtain

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

The autocovariance in this case reduces to

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

Together with eq'n (5), for large enough s, this implies (8). ■

Proof of Theorem 3: In order to verify (9), we write:

$$
\begin{aligned}
& |E\left[f\left(Y_s,...,Y_{s+k}\right)g\left(Y_{s+n},...,Y_{s+n+l}\right)\right]| \\
& -|E\left[f\left(Y_s,...,Y_{s+k}\right)\right]|\,|E\left[g\left(Y_{s+n},...,Y_{s+n+l}\right)\right]| \\
\leq\ & |E\left[f\left(Y_s,...,Y_{s+k}\right)g\left(Y_{s+n},...,Y_{s+n+l}\right)\right] \\
& -E\left[f\left(Y_s,...,Y_{s+k}\right)\right]E\left[g\left(Y_{s+n},...,Y_{s+n+l}\right)\right]| \\
=\ & |\sum_{\substack{j_1,...,j_{k+1}\\ m_1,...,m_{l+1}}} f\left(B^s_{j_1,...,j_{k+1}}\right)g\left(C^{s+n}_{m_1,...,m_{l+1}}\right) \\
& \times[\Pr\left(B^s_{j_1,...,j_{k+1}},C^{s+n}_{m_1,...,m_{l+1}}\right)-\Pr\left(B^s_{j_1,...,j_{k+1}}\right)\Pr\left(C^{s+n}_{m_1,...,m_{l+1}}\right)]| \\
=\ & |\sum_{\substack{j_1,...,j_{k+1}\\ m_1,...,m_{l+1}}} f\left(B^s_{j_1,...,j_{k+1}}\right)g\left(C^{s+n}_{m_1,...,m_{l+1}}\right) \\
& \times\Pr\left(B^s_{j_1,...,j_{k+1}}\right)\left(1-\Pr\left(B^s_{j_1,...,j_{k+1}}\right)\right)[\Pr\left(C^{s+n}_{m_1,...,m_{l+1}}|B^s_{j_1,...,j_{k+1}}\right) \\
& -\Pr\left(C^{s+n}_{m_1,...,m_{l+1}}|\left(B^s_{j_1,...,j_{k+1}}\right)^c\right)]|, \qquad (13)
\end{aligned}
$$

where

$$B^s_{j_1,...,j_{k+1}}=\{Y_s=j_1,...,Y_{s+k}=j_{k+1}\}$$

and

$$C^{s+n}_{m_1,...,m_{l+1}}=\{Y_{s+n}=m_1,...,Y_{s+n+l}=m_{l+1}\}.$$

For $t>s+k+\lambda$ we construct the event

$$A_t=\left\{Y^B_{t-1}=Y^{B^c}_{t-1},...,Y^B_{t-\lambda}=Y^{B^c}_{t-\lambda}\right\}.$$

where, for brevity, the superscript B stands for $B^s_{j_1,...,j_{k+1}}$ and B^c is its complement. It follows that

$$A_t\Longrightarrow\left\{\left(\bar{Y}^s_{t-1}|B\right)=\left(\bar{Y}^s_{t-1}|B^c\right)\right\}\Longrightarrow\left\{Y^B_t=Y^{B^c}_t\right\}.$$

Hence,

$$A_t\Longrightarrow A_{t+1},t>s+k+\lambda.$$

The rest of the proof is very similar to the proof of Theorem 1 and is omitted.■

Appendix B: Consistency and Asymptotic Normality

Proof of Theorem 4: The proof can be made by either checking the conditions of Proposition 7.5 of Hayashi (2000), Theorem 2.7 of Newey and McFadden (1994), or by directly verifying Wu's (1981) criterion. For any $\delta_1 > 0$, denote by $B_{\delta_1}(\theta_0)$ the ball $\{\theta \in \Theta : ||\theta - \theta_0|| \leq \delta_1\}$ and by $B^c_{\delta_1}(\theta_0)$ the complement of $B_{\delta_1}(\theta_0)$ in Θ. For any $\theta \in \Theta$, let

$$D_n(\theta_0, \theta_1) = \frac{1}{n}\left(l_n(\theta_0) - l_n(\theta_1)\right).$$

To establish consistency, we must prove that $\forall \delta_1 > 0$,

$$\lim \inf_{n\to\infty} \inf_{B^c_{\delta_1}(\theta_0)} D_n(\theta_0, \theta_1) \tag{14}$$

is strictly positive in probability. See, for instance, Wu (1981).

Let

$$l_{n,j}(\theta) \equiv \frac{1}{n}\sum_{t=1}^{n} 1\{y_t = j\} \ln \Delta_{t,j}(\theta).$$

The series $\{l_{n,j}(\theta)\}$ is nonpositive and uniformly bounded from below and by ergodicity of the process, it is convergent $w.p.1$. We shall denote this limit by $l_j(\theta)$. This implies that $\forall \theta \in \Theta$, $l_n(\theta) \to_{a.s.} \sum_{j=1}^{M} l_j(\theta) \equiv l(\theta)$. Using Jensen's inequality and the fact that $\sum_{j=1}^{M} \Delta_{t,j}(\theta_0) = 1$,

$$\begin{aligned}
E_{\theta_0}\left(D_n(\theta_1, \theta_0)\right) &= \frac{1}{n} E_{\theta_0} \sum_{t=1}^{n}\sum_{j=1}^{M} E_{\theta_0}\left(1\{y_t = j\} \ln \frac{\Delta_{t,j}(\theta_1)}{\Delta_{t,j}(\theta_0)} | \mathcal{F}_{t-1}\right) \\
&= \frac{1}{n} E_{\theta_0} \sum_{t=1}^{n}\sum_{j=1}^{M} \Delta_{t,j}(\theta_0) \ln \frac{\Delta_{t,j}(\theta_1)}{\Delta_{t,j}(\theta_0)} \\
&\leq \frac{1}{n} E_{\theta_0} \sum_{t=1}^{n} \ln(1) \qquad (15) \\
&= 0.
\end{aligned}$$

If $\mu_0 \neq \mu_1$, $\Delta_t(\mu_0, w, \rho) \neq \Delta_t(\mu_1, w, \rho)$, $\forall t$ and if $w_0 \neq w_1$, $\bar{Y}_{t-1}^s \neq \bar{Y}_{t-1}^s$ $w.p.$ 1 $\forall t$ under Assumption A0, which also implies $\Delta_t(\mu, w_0) \neq \Delta_t(\mu, w_1)$ $w.p.$ 1 $\forall t$. Furthermore, if $\rho_0 \neq \rho_1$, $\Delta_t(\mu, w, \rho_0) \neq \Delta_t(\mu, w, \rho_1)$, $\forall t$. Hence, as $n \to \infty$, equality in (15) holds iff $\theta_0 = \theta$ and the proof of the Theorem is completed. ■

In order to prove Theorem 5, we shall require the following lemmas.

Lemma 6 *Under Assumptions A0-A2,* $z_n(\theta_0) \xrightarrow{d} N(0, V(\theta_0))$.

Proof of Lemma 6: Let

$$f_{t,k}(\theta) = \phi\left(\mu_k - \bar{Y}_{t-1}^s\right),$$

where ϕ is the standard normal PDF. As

$$\dot{\Delta}_{t,j}^{\mu_k}(\theta) \equiv \frac{\partial \Delta_{t,j}(\theta)}{\partial \mu_k} = f_{t,k}(\theta)\left(1\{j = k\} - 1\{j = k+1\}\right),$$

we have

$$z_{n,\mu_k}(\theta) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} W_t^{\mu_k}(\theta), \tag{16}$$

where

$$W_t^{\mu_k}(\theta) = f_{t,k}(\theta)\left(\frac{1\{Y_t = k\}}{\Delta_{t,k}} - \frac{1\{Y_t = k+1\}}{\Delta_{t,k+1}}\right).$$

We notice that

$$E_{\theta_0}\left(W_t^{\mu_k}(\theta) \,|\mathcal{F}_{t-1}\right) = 0$$

so that $W_t^{\mu_k}$ is an m.d.s.. Furthermore,

$$\dot{\Delta}_{t,j}^{w_k}(\theta) \equiv \frac{\partial \Delta_{t,j}(\theta)}{\partial w_k} = -\delta_{t,j}(\theta)\,\dot{h}_t^{w_k}(\theta)$$

where

$$\delta_{t,j}(\theta) = f_{t,j}(\theta) - f_{t,j-1}(\theta)$$

and

$$\dot{h}_t^{w_k}(\theta) = \frac{\partial}{\partial w_k} \bar{Y}_{t-1}^s.$$

Thus,

$$z_{n,w_k}(\theta) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} W_t^{w_k}(\theta), \tag{17}$$

where

$$W_t^{w_k}(\theta) = -\dot{h}_t^{w_k}(\theta) \sum_{j=1}^{M} 1\{Y_t = j\} \frac{\delta_{t,j}(\theta)}{\Delta_{t,j}(\theta)}. \tag{18}$$

We have,

$$\begin{aligned} E_{\theta_0}\left(W_t^{w_k}(\theta) \,|\mathcal{F}_{t-1}\right) &= -\dot{h}_{t,k}^{w}(\theta) \sum_{j=1}^{M} \delta_{t,j}(\theta) \\ &= -\dot{h}_{t,k}^{w}(\theta)\left(f_{t,M}(\theta) - f_{t,0}(\theta)\right) \\ &= 0, \end{aligned}$$

so that $W_t^{w_k}(\theta)$ is also an m.d.s.. Finally,

$$z_{n,\rho}(\theta) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} W_t^{\rho}(\theta),$$

where

$$W_t^{\rho}(\theta) = -\rho^{-1} \bar{Y}_{t-1}^{s} \sum_{j=1}^{M} 1\{Y_t = j\} \frac{\delta_{t,j}(\theta)}{\Delta_{t,j}(\theta)},$$

which is also an m.d.s.. For asymptotic normality of the score function, it will thus be sufficient to verify conditions (2.3) of McLeish (1974). Let $\sigma_n^{i_k}(\theta)^2 = \sum_{t=1}^{n} \left(W_t^{i_k}(\theta)\right)^2$, $i = \mu$ with $k = 1, ..., M-1$, $i = w$ with $k = 1, ...K$, or $i = \rho$ with the k-index suppressed. We need to show that for each $\theta \in \Theta$,

$$\frac{\sigma_n^{i_k}(\theta)^2}{n} \xrightarrow{p} V^{i_k}(\theta) < \infty \tag{19}$$

and that $\forall \varepsilon > 0$, i and k,

$$\frac{1}{\sigma_n^{i_k}(\theta)^2} \sum_{t=1}^{n} \left(W_t^{i_k}(\theta)\right)^2 1\left\{\left|W_t^{i_k}(\theta)\right| > \varepsilon \sigma_n^{i_k}(\theta)\right\} \xrightarrow{p} 0. \tag{20}$$

As $f_{t,k}(\theta) < \infty$, uniformly in n, k and Θ,

$$\frac{1}{n}\sum_{t=1}^{n}\left(W_t^{\mu_k}(\theta)\right)^2 = \frac{1}{n}\sum_{t=1}^{n} f_{t,k}(\theta)^2 \left(\frac{1\{Y_t = k\}}{\Delta_{t,k}} - \frac{1\{Y_t = k+1\}}{\Delta_{t,k+1}}\right)^2 < K \tag{21}$$

and convergence is assured by ergodicity, the limit of which is denoted by $V^{\mu_k}(\theta)$. Also, because $W_t^{\mu_k}(\theta)$ is uniformly bounded and $\sigma_n^{\mu_k}(\theta)$ behaves as $\sqrt{n}$ in probability, condition (20) trivially holds and we are done for $z_{n,\mu_k}(\theta)$.

For $W_t^{w_k}(\theta)$, observe that

$$\dot{h}_t^{w_k}(\theta) = \rho\left(\frac{\sum_{i<t}\dot{s}_{w_k}(X_i, X_t; w)\, Y_i}{\sum_{i<t} s(X_i, X_t; w)} - \frac{\sum_{i<t} s(X_i, X_t; w)\, Y_i \sum_{i<t}\dot{s}_{w_k}(X_i, X_t; w)}{\left(\sum_{i<t} s(X_i, X_t; w)\right)^2}\right), \tag{22}$$

where $\dot{s}_{w_k}(X_i, X_t; w) = \partial s(X_i, X_t; w)/\partial w_k$. It follows from (22) that under Assumptions A1-A2,

$$\sup_{t,k,\Theta}\left|\dot{h}_t^{w_k}(\theta)\right| < 2KM.$$

In view of (18) and the last inequality

$$\sup_{t,k,n,\Theta}\left|W_t^{w_k}(\theta)\right| < K,$$

so that, together with ergodicity,

$$\frac{1}{n}\sum_{t=1}^{n}\left(W_t^{w_k}(\theta)\right)^2 \xrightarrow{p} V^{w_k}(\theta) < \infty.$$

Condition (20) also holds because $W_t^{w_k}(\theta)$ is uniformly bounded and $\sigma_n^{w_k}(\theta)$ behaves as $\sqrt{n}$ in probability. Similar reasoning follows for $W_t^{\rho}(\theta)$ and the proof of the Lemma 6 is therefore completed.■

Lemma 7 *Under Assumptions A0-A4, $\forall\theta \in \Theta$,*

$$\lim_{n\to\infty} E_\theta\left(\left(H_{n,\theta_j,\theta_k}(\theta)\right)_{1\le j,k\le K+M}\right)$$

is finite and nonsingular.

Proof of Lemma 7: We have

$$\begin{aligned}\frac{\partial^2 l_n(\theta)}{\partial\mu_j\partial\mu_k} &= [\sum_{t=1}^{n}\dot{f}_{t,j}(\theta)\left(\frac{1\{Y_t=k\}}{\Delta_{t,k}}-\frac{1\{Y_t=k+1\}}{\Delta_{t,k+1}}\right)\\ &\quad -\sum_{t=1}^{n}f_{t,k}^2(\theta)\left(\frac{1\{Y_t=k+1\}}{\Delta_{t,k+1}^2(\theta)}+\frac{1\{Y_t=k\}}{\Delta_{t,k}^2(\theta)}\right)]1\{j=k\}\\ &\quad +\sum_{t=1}^{n}f_{t,j}(\theta)f_{t,j+1}(\theta)\frac{1\{Y_t=j+1\}}{\Delta_{t,j+1}^2}1\{j=k-1\},\end{aligned}$$

with $\dot{f}_{t,j}(\theta)=\partial f_{t,j}(x;\theta)/\partial x$. Hence,

$$\begin{aligned}E_\theta\left(\frac{\partial^2 l_n(\theta)}{\partial\mu_j\partial\mu_k}|\mathcal{F}_{t-1}\right) &= \left(-\sum_{t=1}^{n}f_{t,k}^2(\theta)\left(\frac{1}{\Delta_{t,k+1}(\theta)}+\frac{1}{\Delta_{t,k}(\theta)}\right)\right)1\{j=k\}\\ &\quad +\sum_{t=1}^{n}\frac{f_{t,j}(\theta)f_{t,j+1}(\theta)}{\Delta_{t,j+1}}1\{j=k-1\}.\end{aligned}$$

In view of (18) and under Assumption A3,

$$\begin{aligned}\frac{\partial^2 l_n(\theta)}{\partial w_l\partial w_k} &= -\sum_{t=1}^{n}\ddot{h}_t^{w_k,w_l}(\theta)\sum_{j=1}^{M}1\{Y_t=j\}\frac{\delta_{t,j}(\theta)}{\Delta_{t,j}(\theta)}\\ &\quad -\sum_{t=1}^{n}\dot{h}_t^{w_k}(\theta)\sum_{j=1}^{M}1\{Y_t=j\}\left(\frac{\dot{\delta}_{t,j,l}(\theta)}{\Delta_{t,j}(\theta)}+\frac{\delta_{t,j}^2(\theta)\dot{h}_t^{w_l}(\theta)}{\Delta_{t,j}^2(\theta)}\right),\end{aligned}$$

where

$$\dot{\delta}_{t,j,l}(\theta)=\frac{\partial\delta_{t,j}(\theta)}{\partial w_l}=-\left(\dot{f}_{t,j}(\theta)-\dot{f}_{t,j-1}(\theta)\right)\dot{h}_t^{w_l}(\theta)=-\rho_{t,j}(\theta)\dot{h}_t^{w_l}(\theta),$$

say. We have,

$$E_{\theta_0}\left(\frac{\partial^2 l_n(\theta)}{\partial w_k\partial w_l}|\mathcal{F}_{t-1}\right)=-\sum_{t=1}^{n}\dot{h}_t^{w_k}(\theta)\dot{h}_t^{w_l}(\theta)\sum_{j=1}^{M}\left(-\rho_{t,j}(\theta)+\frac{\delta_{t,j}^2(\theta)}{\Delta_{t,j}(\theta)}\right).$$

For the normal distribution, $\dot{\phi}(x) = -x\phi(x)$ so that $\sum_{j=1}^{M} \rho_{t,j}(\theta) = 0$ and we are left with

$$E_\theta\left(\frac{\partial^2 l_n(\theta)}{\partial w_k \partial w_l}|\mathcal{F}_{t-1}\right) = -\sum_{t=1}^{n} \dot{h}_t^{w_k}(\theta)\,\dot{h}_t^{w_l}(\theta) \sum_{j=1}^{M} \frac{\delta_{t,j}^2(\theta)}{\Delta_{t,j}(\theta)}.$$

Similarly, with $\dot{\delta}_{t,j,\rho}(x;\theta) = \partial\delta_{t,j}(\theta)/\partial\rho = -\rho_{t,j}(\theta)\,\rho^{-1}\bar{Y}_{t-1}^s$,

$$\begin{aligned}\frac{\partial^2 l_n(\theta)}{\partial \rho^2} &= -\sum_{t=1}^{n} \rho^{-1}\bar{Y}_{t-1}^s \sum_{j=1}^{M} 1\{Y_t = j\}\left(\frac{\dot{\delta}_{t,j}(\theta)}{\Delta_{t,j}(\theta)} + \frac{\delta_{t,j}^2(\theta)\,\rho^{-1}\bar{Y}_{t-1}^s}{\Delta_{t,j}^2(\theta)}\right) \\ &= -\sum_{t=1}^{n} \left(\rho^{-1}\bar{Y}_{t-1}^s\right)^2 \sum_{j=1}^{M} 1\{Y_t = j\}\left(-\frac{\rho_{t,j}(\theta)}{\Delta_{t,j}(\theta)} + \frac{\delta_{t,j}^2(\theta)}{\Delta_{t,j}^2(\theta)}\right),\end{aligned}$$

giving

$$E_\theta\left(\frac{\partial^2 l_n(\theta)}{\partial \rho^2}|\mathcal{F}_{t-1}\right) = -\sum_{t=1}^{n} \left(\rho^{-1}\bar{Y}_{t-1}^s\right)^2 \sum_{j=1}^{M} \frac{\delta_{t,j}^2(\theta)}{\Delta_{t,j}(\theta)}.$$

Because

$$\frac{\partial l_n(\theta)}{\partial w_k} = -\sum_{t=1}^{n} \dot{h}_t^{w_k}(\theta) \sum_{j=1}^{M} 1\{Y_t = j\}\frac{\delta_{t,j}(\theta)}{\Delta_{t,j}(\theta)},$$

$$\begin{aligned}\frac{\partial^2 l_n(\theta)}{\partial w_k \partial \mu_l} &= -\sum_{t=1}^{n} \dot{h}_t^{w_k}(\theta) \sum_{j=1}^{M} 1\{Y_t = j\}\frac{1\{j=l\} - 1\{j=l+1\}}{\Delta_{t,j}(\theta)} \\ &\quad \times\left(\dot{f}_{t,l}(\theta) - \frac{\delta_{t,j}(\theta)\,f_{t,l}(\theta)}{\Delta_{t,j}(\theta)}\right).\end{aligned}$$

Thus,

$$E_\theta\left(\frac{\partial^2 l_n(\theta)}{\partial w_k \partial \mu_l}|\mathcal{F}_{t-1}\right) = \sum_{t=1}^{n} \dot{h}_t^{w_k} f_{t,l}(\theta)\left(\frac{\delta_{t,l}(\theta)}{\Delta_{t,l}(\theta)} - \frac{\delta_{t,l+1}(\theta)}{\Delta_{t,l+1}(\theta)}\right).$$

Also,

$$\begin{aligned}\frac{\partial^2 l_n(\theta)}{\partial w_l \partial \rho} &= -\sum_{t=1}^{n} \ddot{h}_t^{w_k,\rho}(\theta) \sum_{j=1}^{M} 1\{Y_t = j\} \frac{\delta_{t,j}(\theta)}{\Delta_{t,j}(\theta)} \\ &\quad -\sum_{t=1}^{n} \dot{h}_t^{w_k}(\theta) \sum_{j=1}^{M} 1\{Y_t = j\} \left(\frac{\dot{\delta}_{t,j,\rho}(\theta)}{\Delta_{t,j}(\theta)} + \frac{\delta_{t,j}^2(\theta) \rho^{-1} \bar{Y}_{t-1}^s}{\Delta_{t,j}^2(\theta)} \right),\end{aligned}$$

and

$$E_\theta \left(\frac{\partial^2 l_n(\theta)}{\partial w_k \partial \rho} | \mathcal{F}_{t-1} \right) = -\rho^{-1} \sum_{t=1}^{n} \dot{h}_t^{w_k}(\theta) \bar{Y}_{t-1}^s \sum_{j=1}^{M} \frac{\delta_{t,j}^2(\theta)}{\Delta_{t,j}(\theta)}.$$

Finally,

$$\begin{aligned}\frac{\partial^2 l_n(\theta)}{\partial \rho \partial \mu_l} &= -\rho^{-1} \sum_{t=1}^{n} \bar{Y}_{t-1}^s \sum_{j=1}^{M} 1\{Y_t = j\} \frac{1\{j = l\} - 1\{j = l+1\}}{\Delta_{t,j}(\theta)} \\ &\quad \times \left(\dot{f}_{t,l}(\theta) - \frac{\delta_{t,j}(\theta) f_{t,l}(\theta)}{\Delta_{t,j}(\theta)} \right)\end{aligned}$$

and

$$E_\theta \left(\frac{\partial^2 l_n(\theta)}{\partial \rho \partial \mu_l} | \mathcal{F}_{t-1} \right) = \rho^{-1} \sum_{t=1}^{n} \bar{Y}_{t-1}^s f_{t,l}(\theta) \left(\frac{\delta_{t,l}(\theta)}{\Delta_{t,l}(\theta)} - \frac{\delta_{t,l+1}(\theta)}{\Delta_{t,l+1}(\theta)} \right).$$

It is obvious that for any θ_k, θ_l, all the second-order derivatives may be written as

$$H_{n,\theta_j,\theta_k}(\theta) = \frac{1}{n} \sum_{t=1}^{n} z_t(\theta),$$

where, under Assumptions A1-A2, $z_t(\theta)$ are uniformly bounded. By the ergodicity, $H_{n,\theta_j,\theta_k}(\theta)$ converges $w.p.1$ to a nonstochastic function, say $H_{\theta_j,\theta_k}(\theta)$. Moreover, the Cauchy Schwartz inequality implies that the determinant of $E_\theta \left(H_{n,\theta_j,\theta_k}(\theta) \right)$ is non-negative for all n, θ_j, θ_k, with equality holding iff the terms in $\left(\dot{h}_t^{w_k} \right)_{1 \le k \le K}$ are linearly dependent. This possibility is precluded by Assumption A4 and thus, the proof of Lemma 7 is complete.■

Proof of Theorem 5: It is straightforward to verify that the second-order Bartlett identity holds for all the second-order partial derivatives. As $H_{n,\theta_j,\theta_k}(\theta)$ converges $w.p.1$ to $H_{\theta_j,\theta_k}(\theta)$, it also converges in probability. Because $\hat{\theta}_n \rightarrow_p \theta_0$ and because $H_{\theta_j,\theta_k}(\theta)$ is continuous, it follows from Theorem 4.1.5 of Amemiya (1985) that $\text{plim}\Big(H_{n,\theta_j,\theta_k}\left(\hat{\theta}_n\right)\Big) = H_{\theta_j,\theta_k}(\theta_0)$. This, together with Lemma 6 and the mean value Theorem, as in eq'n (7.3.7) of Hayashi (2000), completes the proof. ■

Table 1. Simulated MLE point estimates for $\mu_0 = 0.5$ and $\lambda = 2$.

	n	250		500		1000		2000	
		$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$
	Mean	4.341	0.499	1.468	0.499	1.115	0.500	1.050	0.500
	Std	24.613	0.083	3.896	0.058	0.598	0.041	0.359	0.029
	Trim	1.584	0.499	1.193	0.499	1.085	0.500	1.038	0.500
$w_0 = 1$	Std Trim	1.942	0.072	0.763	0.051	0.476	0.035	0.301	0.025
	Median	1.053	0.499	1.004	0.499	1.009	0.499	1.010	0.501
	Q_1	0.498	0.443	0.629	0.459	0.723	0.473	0.796	0.480
	Q_3	1.976	0.559	1.626	0.540	1.394	0.525	1.243	0.519
	Mean	12.815	0.500	6.542	0.500	3.806	0.500	3.300	0.500
	Std	43.107	0.084	22.169	0.059	4.147	0.041	1.581	0.029
	Trim	6.495	0.501	3.908	0.500	3.420	0.500	3.198	0.500
$w_0 = 3$	Std Trim	13.165	0.073	3.261	0.051	1.642	0.036	0.973	0.025
	Median	2.850	0.502	2.999	0.502	3.073	0.500	3.054	0.500
	Q_1	1.440	0.444	1.929	0.460	2.226	0.473	2.440	0.481
	Q_3	5.716	0.560	4.660	0.540	4.177	0.526	3.770	0.519
	Mean	18.778	0.499	12.923	0.501	8.262	0.500	5.962	0.500
	Std	49.275	0.083	35.147	0.058	19.249	0.040	5.747	0.029
	Trim	12.716	0.499	8.031	0.501	6.170	0.500	5.449	0.500
$w_0 = 5$	Std Trim	29.233	0.072	10.037	0.050	3.924	0.034	2.075	0.025
	Median	4.457	0.498	5.022	0.502	5.084	0.501	4.929	0.500
	Q_1	2.281	0.442	3.083	0.464	3.611	0.474	3.918	0.480
	Q_3	10.487	0.555	8.818	0.538	7.366	0.527	6.518	0.519

Note: λ is the lag-length; Std is the standard deviation of the mean; Trim is the trimmed mean with 5% symmetric trimming; Std Trim is the standard deviation of the trimmed mean; Q_1 and Q_3 are the first and third quartiles, respectively.

Table 2. Simulated MLE point estimates for $\mu_0 = 0.5$ and $\lambda = 5$.

	n	250		500		1000		2000	
		$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$
	Mean	1.705	0.501	1.164	0.500	1.070	0.501	1.030	0.500
	Std	6.347	0.081	0.973	0.057	0.591	0.040	0.384	0.029
	Trim	1.261	0.501	1.104	0.500	1.045	0.501	1.022	0.500
$w_0 = 1$	Std Trim	1.261	0.070	0.717	0.050	0.476	0.035	0.328	0.025
	Median	0.969	0.501	0.994	0.498	0.987	0.501	0.995	0.500
	Q_1	0.413	0.448	0.578	0.462	0.682	0.474	0.777	0.481
	Q_3	1.763	0.556	1.514	0.537	1.356	0.528	1.248	0.518
	Mean	4.579	0.500	3.673	0.500	3.288	0.499	3.129	0.500
	Std	8.305	0.080	3.118	0.057	1.439	0.040	0.827	0.028
	Trim	3.726	0.800	3.398	0.500	3.211	0.499	3.108	0.500
$w_0 = 3$	Std Trim	2.885	0.070	1.738	0.050	1.065	0.035	0.703	0.024
	Median	2.965	0.500	3.012	0.501	3.053	0.501	3.036	0.500
	Q_1	1.770	0.444	2.120	0.461	2.389	0.472	2.553	0.482
	Q_3	4.874	0.555	4.262	0.539	3.934	0.526	3.634	0.518
	Mean	9.791	0.500	6.836	0.500	5.465	0.500	5.142	0.500
	Std	33.429	0.082	18.622	0.059	2.709	0.040	1.452	0.029
	Trim	6.505	0.500	5.620	0.500	5.289	0.500	5.088	0.500
$w_0 = 5$	Std Trim	5.838	0.072	2.905	0.051	1.851	0.034	1.208	0.025
	Median	4.858	0.503	4.973	0.499	4.923	0.500	4.912	0.500
	Q_1	3.002	0.443	3.496	0.461	3.845	0.473	4.147	0.481
	Q_3	7.996	0.555	7.042	0.540	6.385	0.527	5.975	0.520

Note: λ is the lag-length; Std is the standard deviation of the mean; Trim is the trimmed mean with 5% symmetric trimming; Std Trim is the standard deviation of the trimmed mean; Q_1 and Q_3 are the first and third quartiles, respectively.

Table 3. Simulated MLE point estimates for $\mu_0 = 0.3$ and $\lambda = 2$.

	n	250		500		1000		2000	
		$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$
	Mean	3.787	0.299	2.073	0.299	1.152	0.301	1.063	0.300
	Std	21.847	0.086	12.133	0.060	0.769	0.042	0.409	0.030
	Trim	1.606	0.300	1.238	0.298	1.097	0.301	1.046	0.300
$w_0 = 1$	Std Trim	2.229	0.074	0.899	0.052	0.491	0.037	0.317	0.026
	Median	0.988	0.301	1.022	0.299	1.008	0.301	1.014	0.300
	Q_1	0.483	0.240	0.622	0.258	0.724	0.272	0.803	0.280
	Q_3	1.874	0.357	1.571	0.338	1.394	0.329	1.267	0.320
	Mean	11.79	0.297	6.198	0.298	4.204	0.301	3.325	0.301
	Std	38.258	0.083	18.497	0.059	9.583	0.043	1.780	0.030
	Trim	6.393	0.297	4.197	0.298	3.466	0.301	3.192	0.301
$w_0 = 3$	Std Trim	11.471	0.072	3.877	0.051	1.880	0.37	1.065	0.026
	Median	3.011	0.298	2.959	0.299	2.964	0.301	2.993	0.301
	Q_1	1.605	0.242	1.938	0.258	2.188	0.273	2.362	0.280
	Q_3	6.192	0.351	4.863	0.339	4.263	0.330	3.818	0.322
	Mean	19.307	0.301	12.174	0.302	7.902	0.301	5.904	0.301
	Std	51.551	0.082	30.598	0.058	17.282	0.042	5.552	0.023
	Trim	12.944	0.301	8.084	0.302	6.006	0.301	5.466	0.301
$w_0 = 5$	Std Trim	30.462	0.072	10.360	0.050	3.691	0.036	2.024	0.025
	Median	4.646	0.303	4.817	0.303	4.945	0.300	5.075	0.301
	Q_1	2.3207	0.247	3.066	0.263	3.598	0.273	3.973	0.281
	Q_3	10.323	0.354	8.557	0.340	7.182	0.328	6.502	0.320

Note: λ is the lag-length; Std is the standard deviation of the mean; Trim is the trimmed mean with 5% symmetric trimming; Std Trim is the standard deviation of the trimmed mean; Q_1 and Q_3 are the first and third quartiles, respectively.

Table 4. Simulated MLE point estimates for $\mu_0 = 0.3$ and $\lambda = 5$.

	n	250		500		1000		2000	
		$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$
	Mean	1.417	0.296	1.196	0.300	1.084	0.300	1.037	0.300
	Std	2.636	0.085	1.632	0.059	0.622	0.041	0.377	0.029
	Trim	1.212	0.297	1.094	0.300	1.053	0.300	1.030	0.300
$w_0 = 1$	Std Trim	1.194	0.074	0.731	0.052	0.466	0.036	0.325	0.026
	Median	0.940	0.298	0.961	0.299	1.001	0.300	1.011	0.300
	Q_1	0.934	0.240	0.555	0.260	0.693	0.273	0.765	0.281
	Q_3	1.740	0.355	1.488	0.342	1.380	0.328	1.270	0.320
	Mean	6.369	0.298	4.091	0.300	3.281	0.299	3.126	0.299
	Std	26.611	0.083	11.170	0.058	1.461	0.041	0.919	0.029
	Trim	3.904	0.298	3.452	0.300	3.208	0.299	3.092	0.299
$w_0 = 3$	Std Trim	3.353	0.072	1.894	0.050	1.138	0.036	0.751	0.025
	Median	3.023	0.300	3.025	0.300	3.015	0.298	2.979	0.300
	Q_1	1.682	0.246	2.050	0.259	2.322	0.272	2.500	0.279
	Q_3	4.986	0.354	4.510	0.338	3.994	0.325	3.644	0.319
	Mean	9.592	0.297	6.992	0.299	5.459	0.299	5.243	0.300
	Std	30.564	0.083	17.844	0.059	2.489	0.041	1.482	0.029
	Trim	6.423	0.297	5.701	0.299	5.307	0.299	5.190	0.300
$w_0 = 5$	Std Trim	5.329	0.073	3.039	0.052	1.805	0.036	1.243	0.025
	Median	5.019	0.298	4.985	0.300	5.000	0.299	5.049	0.300
	Q_1	2.942	0.238	3.414	0.259	3.957	0.272	4.214	0.281
	Q_3	8.076	0.351	7.366	0.337	6.372	0.326	6.089	0.320

Note: λ is the lag-length; Std is the standard deviation of the mean; Trim is the trimmed mean with 5% symmetric trimming; Std Trim is the standard deviation of the trimmed mean; Q_1 and Q_3 are the first and third quartiles, respectively.

Table 5. Hit and RMSE for the similarity based model and the ordered probit model.

λ	**Database**	**Prediction Method**	**Criteria**	**Similarity**	**Ordered Probit**
5	Train	$\hat{Y}$	Hit	0.35	0.41
	Train	$\hat{Y}$	RMSE	1.0323	0.9499
	Train	$\tilde{Y}$	Hit	0.36	0.43
	Train	$\tilde{Y}$	RMSE	1.1226	1.0022
	Test	$\hat{Y}$	Hit	0.37	0.41
	Test	$\hat{Y}$	RMSE	1.0296	1.1662
	Test	$\tilde{Y}$	Hit	0.39	0.36
	Test	$\tilde{Y}$	RMSE	1.1136	1.3267
10	Train	$\hat{Y}$	Hit	0.35	0.41
	Train	$\hat{Y}$	RMSE	1.0203	0.9499
	Train	$\tilde{Y}$	Hit	0.38	0.43
	Train	$\tilde{Y}$	RMSE	1.1107	1.0022
	Test	$\hat{Y}$	Hit	0.37	0.41
	Test	$\hat{Y}$	RMSE	1.0734	1.1662
	Test	$\tilde{Y}$	Hit	0.4	0.36
	Test	$\tilde{Y}$	RMSE	1.1747	1.3267
20	Train	$\hat{Y}$	Hit	0.36	0.41
	Train	$\hat{Y}$	RMSE	1.0171	0.9499
	Train	$\tilde{Y}$	Hit	0.38	0.43
	Train	$\tilde{Y}$	RMSE	1.0889	1.0022
	Test	$\hat{Y}$	Hit	0.38	0.41
	Test	$\hat{Y}$	RMSE	1.0149	1.1662
	Test	$\tilde{Y}$	Hit	0.42	0.36
	Test	$\tilde{Y}$	RMSE	1.118	1.3267

Note: λ is the lag-length; $\hat{Y}$, $\tilde{Y}$, Hit and RMSE are given in Section 9.2.

Table 6. Estimated coefficients for the similarity-based model and the ordered probit model on the entire Netflix database

Similarity				Ordered Probit	
	$\lambda = 5$	$\lambda = 10$	$\lambda = 20$		
$\hat{w}_1$	0.0000	0.0000	0.0000	$\hat{\beta}_1$	0.1380
$\hat{w}_2$	18.4440	2.7903	1.8517	$\hat{\beta}_2$	0.5051
$\hat{w}_3$	0.5268	0.0246	0.0000	$\hat{\beta}_3$	-0.1203
$\hat{w}_4$	4.8465	0.3260	0.3547	$\hat{\beta}_4$	0.1854
$\hat{w}_5$	3.6719	0.0536	0.1812	$\hat{\beta}_5$	0.1278
$\hat{\mu}_1$	-1.0382	-0.3238	0.5063	$\hat{\mu}_1$	1.1059
$\hat{\mu}_2$	-0.1945	0.5405	1.3805	$\hat{\mu}_2$	2.0648
$\hat{\mu}_3$	0.7573	1.5042	2.3573	$\hat{\mu}_3$	3.1319
$\hat{\mu}_4$	1.6733	2.4264	3.2973	$\hat{\mu}_4$	4.1516
$\hat{\rho}$	0.2765	0.4732	0.7022		

Note: λ is the lag-length, the $\hat{w}$, $\hat{\mu}$ and $\hat{\rho}$ are the estimated coefficients of the similarity-based model and $\hat{\beta}$ and $\hat{\mu}$ are the estimated coefficients of the ordered probit model

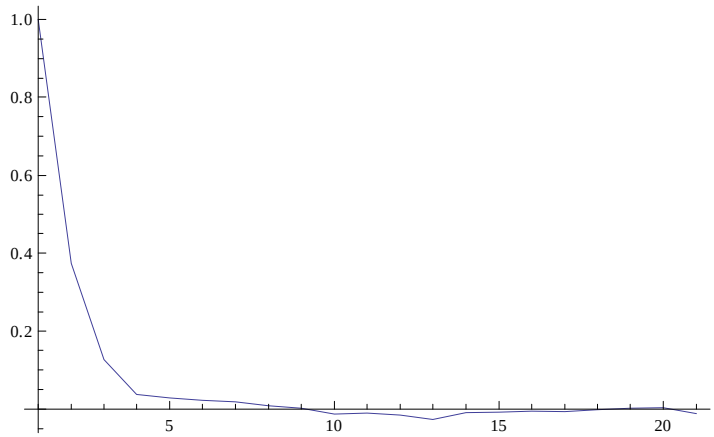

Figure 1. Correlogram of the process in the case $\lambda = 1$, $M = 2$, $n = 10000$.

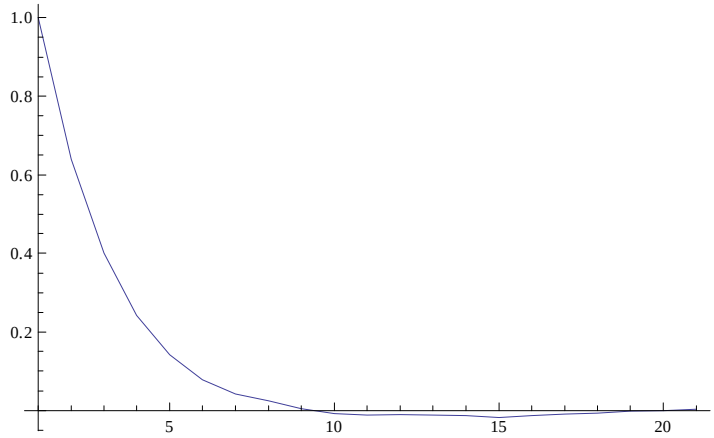

Figure 2. Correlogram of the process in the case $\lambda = 1$, $M = 3$, $n = 10000$.

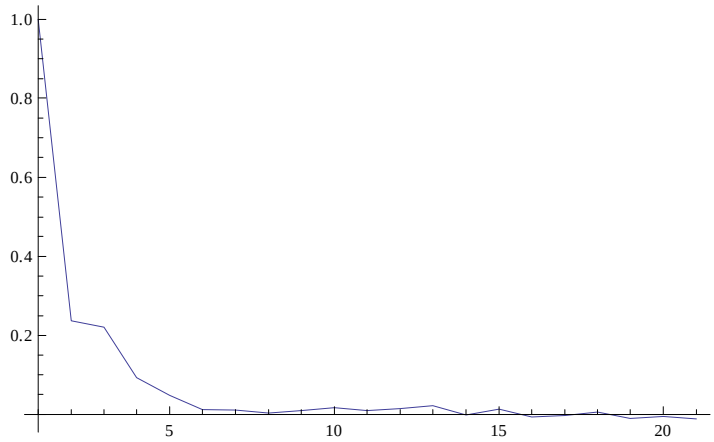

Figure 3. Correlogram of the process in the case $\lambda = 2$, $M = 2$, $n = 10000$.

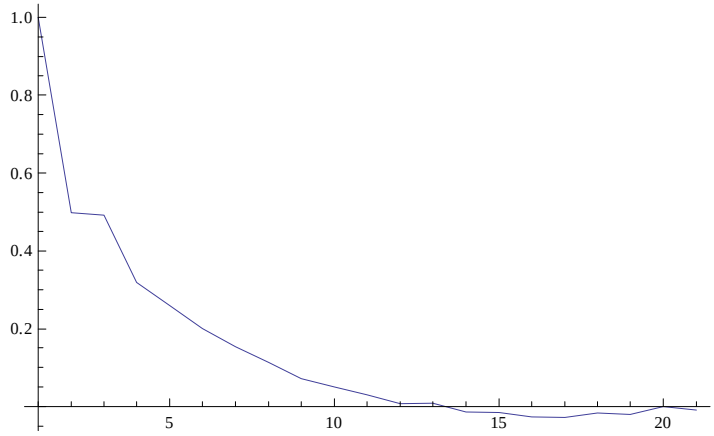

Figure 4. Correlogram of the process in the case $\lambda = 2$, $M = 3$, $n = 10000$.

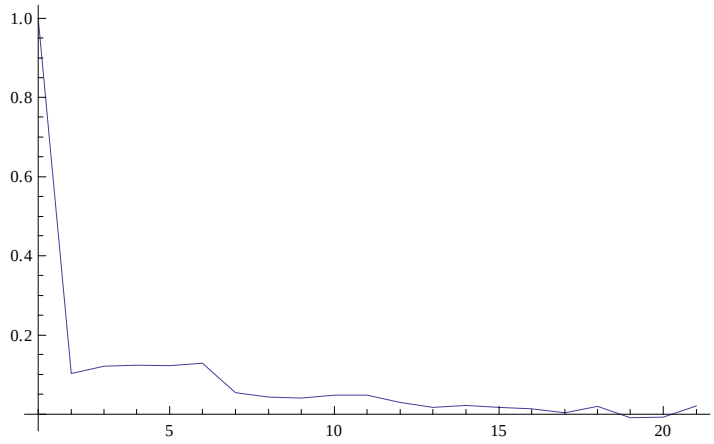

Figure 5. Correlogram of the process in the case $\lambda = 5$, $M = 2$, $n = 10000$.

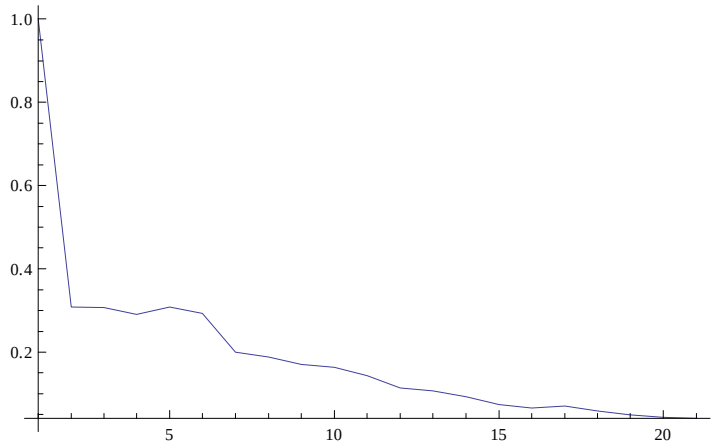

Figure 6. Correlogram of the process in the case $\lambda = 5$, $M = 3$, $n = 10000$.

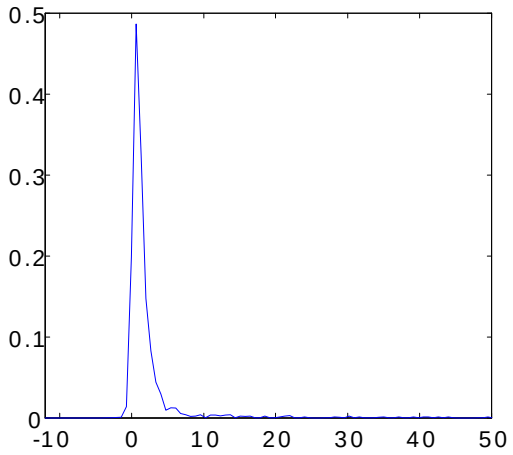

Figure 7. Kernel density estimate for $\hat{w}$, $w_0 = 1$, $\mu_0 = 0.3$, $\lambda = 2$, $n = 250$.

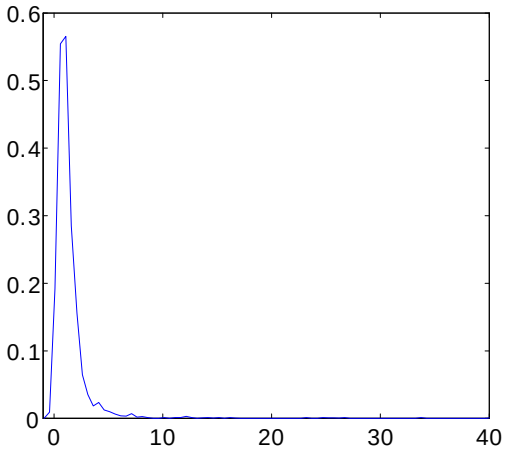

Figure 8. Kernel density estimate for $\hat{w}$, $w_0 = 1$, $\mu_0 = 0.3$, $\lambda = 2$, $n = 500$.

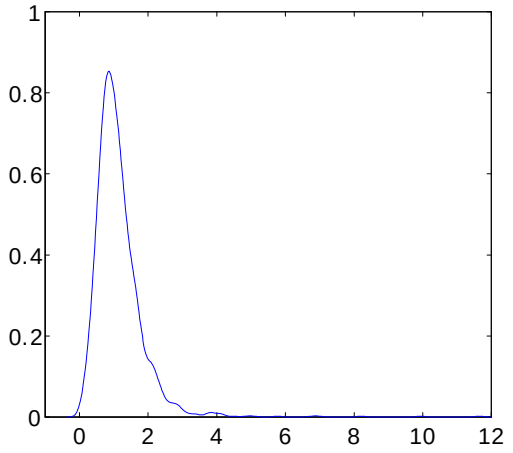

Figure 9. Kernel density estimate for $\hat{w}$, $w_0 = 1$, $\mu_0 = 0.3$, $\lambda = 2$, $n = 1000$.

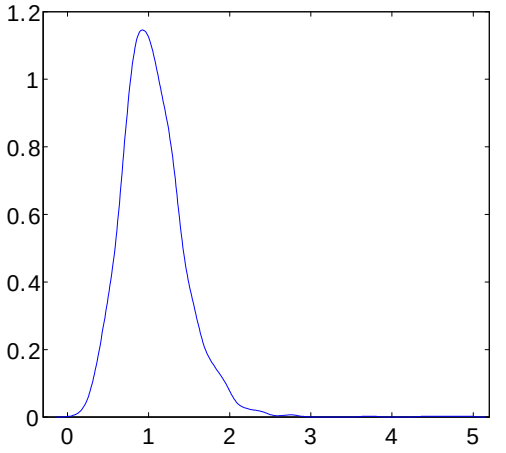

Figure 10. Kernel density estimate for $\hat{w}$, $w_0 = 1$, $\mu_0 = 0.3$, $\lambda = 2$, $n = 2000$.

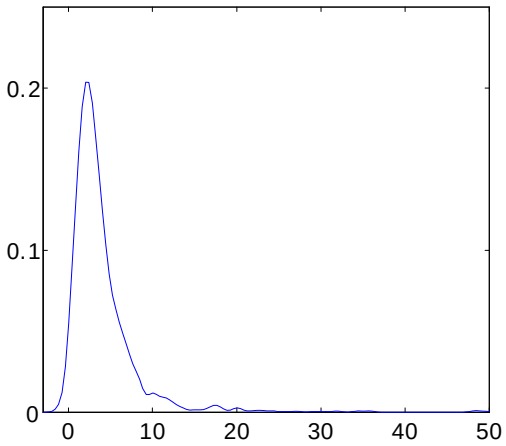

Figure 11. Kernel density estimate for $\hat{w}$, $w_0 = 3$, $\mu_0 = 0.5$, $\lambda = 5$, $n = 250$.

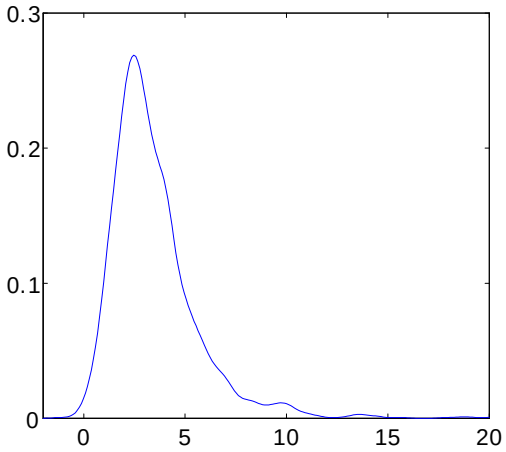

Figure 12. Kernel density estimate for $\hat{w}$, $w_0 = 3$, $\mu_0 = 0.5$, $\lambda = 5$, $n = 500$.

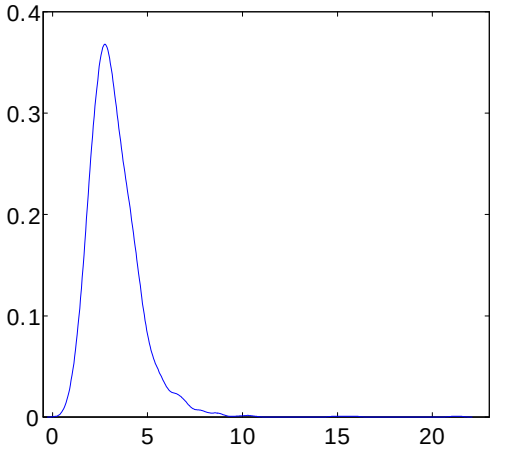

Figure 13. Kernel density estimate for $\hat{w}$, $w_0 = 3$, $\mu_0 = 0.5$, $\lambda = 5$, $n = 1000$.

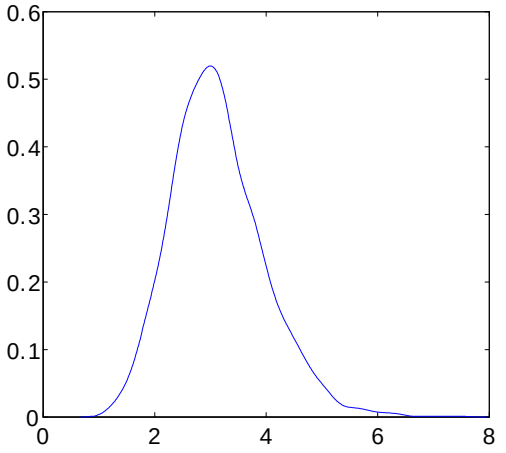

Figure 14. Kernel density estimate for $\hat{w}$, $w_0 = 3$, $\mu_0 = 0.5$, $\lambda = 5$, $n = 2000$.