A SIMILARITY BASED MODEL FOR ORDERED CATEGORICAL DATA

By

Gaby Gayer, Offer Lieberman and Omer Yaffe

October 29, 2014

RESEARCH INSTITUTE FOR ECONOMETRICS

DISCUSSION PAPER NO. 3-14-R

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, October 29, 2014

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 (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 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 almost every microeconometrics text book, see, *inter alia*, 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 refered to Greene and Hensher (2010) and the references therein.

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

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

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, $\varepsilon_t \overset{iid}{\sim} N\left(0,1\right)$. 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.

In the following we introduce the model under consideration. Let

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

$s\left(\cdot\right)$ is a similarity function, $w = \left(w_1, ..., w_K\right)'$, X_i is the ith observation on a K-vector of explanatory variables, $\varepsilon_t \overset{iid}{\sim} N\left(0,1\right)$, $\lambda \geq 1$ is the lag-length, taken to be known and ρ is a free parameter which is allowed to be greater than-, equal to- or less than unity. For brevity, we have suppressed in the notation the dependence of $\bar{Y}_{t-1}^{s}$ on ρ, λ and w. Conditions on $s\left(\cdot\right)$ will be given in Section 3. For instance, we may specify an exponential similarity, viz.,

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

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 $\bar{Y}_{t-1}^{s}$. For categories $j = 1, ..., M$, the model is

$$Y_1 = j1\left\{\varepsilon_1 \in (\Phi^{-1}(\left(\frac{j-1}{M}\right), \Phi^{-1}\left(\frac{j}{M}\right))\right\} \tag{3}$$

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

$$\begin{aligned} Y_t &= j1\left\{\bar{Y}_{t-1}^{s} + \varepsilon_t \in \left(\mu_{j-1}, \mu_j\right]\right\}, \left(j = 1, ..., M\right), \varepsilon_t \overset{iid}{\sim} N\left(0,1\right), \\ &\quad \left(-\infty = \mu_0 < \mu_1 < \cdots < \mu_{M-1} < \mu_M = \infty\right). \end{aligned} \tag{4}$$

We remark that $\left\{X_t, Y_t\right\}_{t=1}^{n}$ is a vector of time series in which the observations are ordered at a given frequency, as is common in most conventional times series models. The assumption on the starting value (3) implies that each of the M categories is equally likely at the outset. Any other reasonable assumption is likely to affect the numerical values of the estimates in small

samples, but not the asymptotic results of the paper.

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. Let $\mathcal{F}_{t-1}$ be the σ-field based on all the information included up to time $t-1$ and $\Phi(\cdot)$ be the standard normal cdf. We have

$$\begin{aligned}
\Pr\left(Y_t = j | \mathcal{F}_{t-1}, X_t; \theta\right) &= \Pr\left\{\bar{Y}^s_{t-1} + \varepsilon_t \in (\mu_{j-1}, \mu_j]\right\} \\
&= \Phi\left(\mu_j - \bar{Y}^s_{t-1}\right) - \Phi\left(\mu_{j-1} - \bar{Y}^s_{t-1}\right) \\
&= \Phi_{t,j}\left(X_1, ..., X_t, Y_1, ..., Y_{t-1}; \theta\right) \\
&\quad -\Phi_{t,j-1}\left(X_1, ..., X_t, Y_1, ..., Y_{t-1}; \theta\right) \\
&= \Delta_{t,j}\left(x_t, y_{t-1}; \theta\right), \qquad (5)
\end{aligned}$$

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

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

The plan for this paper is as follows. In Section 2 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 3 and the theorems follow in Section 4. Simulations are presented in Section 5 and an application to the Netflix

data set follows in Section 6. Section 7 concludes and proofs are provided in the Appendix.

2 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.

2.1 Population Moments

Evidently, the conditional distribution (5) 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

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

say, now, by virtue of the assumption that the μ_j's are all different (see eq'n (4)), $\Pr\{Y_t = j | Y_{t-1} = j\}$ cannot be zero or one. In other words:

$$0 < \delta_1 < b_t^{w_0} < 1 - \delta_2 < 1 \tag{6}$$

for some δ_1, $\delta_2 > 0$ and $\forall t$,. By backward substitution,

$$\begin{aligned}
\xi_{t,j}^{w_0} &= a_t^{w_0} + b_t^{w_0}\xi_{t-1,j}^{w_0} \\
&= a_t^{w_0} + b_t^{w_0}\left(a_{t-1}^{w_0} + b_{t-1}^{w_0}\xi_{t-2,j}^{w_0}\right) \\
&= a_t^{w_0} + b_t^{w_0}\left(a_{t-1}^{w_0} + b_{t-1}^{w_0}\left(a_{t-2}^{w_0} + b_{t-2}^{w_0}\xi_{t-3,j}^{w_0}\right)\right) \\
&\cdots \\
&= (a_t^{w_0} + b_t^{w_0}a_{t-1}^{w_0} + b_t^{w_0}b_{t-1}^{w_0}a_{t-2}^{w_0} + ... + a_2^{w_0}\Pi_{j=0}^{t-2}b_{t-j}^{w_0}) \\
&\quad + b_t^{w_0}b_{t-1}^{w_0}b_{t-2}^{w_0}\cdots b_2^{w_0}\xi_{1,j}^{w_0}. \qquad (7)
\end{aligned}$$

In view of (6), the second term in (7) converges to zero as $t \to \infty$. Letting $\eta = \sup_t b_t^{w_0}$, as $\sup_t a_t^{w_0} \leq M - 1$, we see that the first term on the rhs of (7) is bounded by

$$(M-1)\left(1 + \eta + \eta^2 + \cdots\right) \to_{t\to\infty} \frac{M-1}{1-\eta}.$$

Finally, convergence to $\xi_{\infty,j}^{w_0}$ is assured because the summation is over positive elements. It follows that for any $1 \leq h < \infty$,

$$E_{w_0}\left(Y_t^h\right) = \sum_{j=1}^{M} j^h \Pr_{w_0}\left(Y_t = j\right) \to_{n\to\infty} \sum_{j=1}^{M} j^h \xi_{\infty,j}^{w_0}. \qquad (8)$$

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).

2.2 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 (4) with $\lambda \geq 1$ and $M \geq 2$, $\forall m \in \mathbb{N}$, $\exists x \in (0,1)$, 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 (4) with* $\lambda = 1$*,*

$$\begin{aligned} Cov\left(Y_{s+m}, Y_s\right) &= \sum_{l=1}^{M} l \Pr\left(Y_s = l\right)\left(1 - \Pr\left(Y_s = l\right)\right) \\ &\quad \times \sum_{j=1}^{M} j \Lambda_{s+k,s+k-1,j,l} \prod_{k=1}^{m-1} \Lambda_{s+k,s+k-1,l,l}, \end{aligned} \tag{9}$$

where

$$\Lambda_{t,s,j,l} = \left\{\Pr\left(Y_t = j | Y_s = l\right) - \Pr\left(Y_t = j | Y_s \neq l\right)\right\}. \tag{10}$$

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

$$Cor\left(Y_{s+m}, Y_s\right) = Cor^m\left(Y_{s+1}, Y_s\right) = \left\{\Phi\left(\rho - \mu\right) - \Phi\left(-\mu\right)\right\}^m, m \in \mathbb{N}. \tag{11}$$

The result (11) is therefore completely analogous to the result for the ACF of a linear AR(1) process.

2.3 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 $g : \mathbb{R}^{l+1} \to \mathbb{R}$,

$$\begin{aligned} &\lim_{n\to\infty} \left(\left|E\left[f\left(Y_s, ..., Y_{s+k}\right) g\left(Y_{s+n}, ..., Y_{s+n+l}\right)\right]\right| \right. \\ &\left. - \left|E\left[f\left(Y_s, ..., Y_{s+k}\right)\right]\right| \left|E\left[g\left(Y_{s+n}, ..., Y_{s+n+l}\right)\right]\right|\right) \\ = \;& 0. \end{aligned} \tag{12}$$

Theorem 3 *: The process (4) is ergodic stationary.*

It is emphasized that Theorem 3 holds for all $\lambda \geq 1$ and $M \geq 2$. It is clear from the proof of Theorem 3, specifically, the bound placed on (21), 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.

3 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, 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: $\{\varepsilon_t\}_{t=1}^n$ is a sequence of $NID(0,1)$. For each $t = 1, ..., n$, the $K \times 1$ vector X_t is nonstochastic, real and finite and $Y_t \in \{1, ..., M\}$. If $w \neq w'$, $\Pr_w\left(\bar{Y}_{t-1}^s(w) \neq \bar{Y}_{t-1}^s(w')\right) > 0$, $\forall t$.

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.

Assumption (A4): (i) For $k = 1, ..., K$, $(X_{ik} - X_{tk})^2$ are linearly independent; (ii) The derivatives $\partial \bar{Y}^s_{t-1}/\partial w_k$, $k = 1, ..., K$, are linearly independent.

The last part of Assumption A0 is an identification condition. It is superflous for $\lambda = 1$, in which case $\bar{Y}^s_{t-1} = Y_{t-1}$, which does not depend on w. For any $\lambda > 1$, $\Pr_w(Y_t = j) > 0 \ \forall j \in \{1, ..., M\}$ and $\forall t$ and therefore $\Pr_w\left(\bar{Y}^s_{t-1}(w) \neq \bar{Y}^s_{t-1}(w')\right) > 0$. 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 (2),

$$\dot{s}_{w_k}(X_i, X_t; w) = -(X_{ik} - X_{tk})^2 s(X_i, X_t; w), \tag{13}$$

The inequality in A2 holds for (13) because X_t is bounded $\forall t$ under Assumption A0. Similarly, for the inverse similarity function

$$s(X_i, X_t; w) = \frac{1}{1 + \sum_{j=1}^K w_j (X_{ij} - X_{tj})^2}, \tag{14}$$

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 (2) and (14). As for A4, under (2)

$$\frac{\partial \bar{Y}^s_{t-1}}{\partial \theta_k} = \frac{\sum_{i=t-\lambda}^{i=t-1} s(X_i, X_t; w)(X_{ik} - X_{tk})^2 \left(\bar{Y}^s_{t-1} - Y_i\right)}{\sum_{i=t-\lambda}^{i=t-1} s(X_i, X_t; w)} \tag{15}$$

whereas under (14),

$$\frac{\partial \bar{Y}_{t-1}^{s}}{\partial \theta_k} = \frac{\sum_{i=t-\lambda}^{i=t-1} s^2 \left(X_i, X_t; w\right) \left(X_{ik} - X_{tk}\right)^2 \left(\bar{Y}_{t-1}^{s} - Y_i\right)}{\sum_{i=t-\lambda}^{i=t-1} s \left(X_i, X_t; w\right)}. \tag{16}$$

It follows from (15) and (16) that given A4(i), Assumption A4(ii) is satisfied for (2) and (14).

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

5 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 correlograms 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 1-8 and in Figures 7-18 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 $X \sim [-1, 1]$, which was generated once and then consequently 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, 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 7–18 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 $\lambda = 5$, and in Figures 15-18, 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.

6 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.

6.1 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.

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 because it considerably simplifies the computations. Six movies out of the 14 were chosen arbitrarily, where one movie (Sweet

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

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.

6.2 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 (2) 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) $\hat{Y}_{t+1} = j1\left\{\bar{Y}_t^s\left(\hat{w},\hat{\rho}\right) \in (\hat{\mu}_{j-1},\hat{\mu}_j]\right\}$, $(j = 1, ..., M)$

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.

7 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.

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.

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.

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

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

Hausman, J.A, A.W. Lo, A.C MacKinlay (1992), An ordered probit analysis of transaction stock prices, J*ournal of Financial Economics*, Vol. 31, 3, 319-379.

Hayashi, F. (2000) *Econometrics*, Princeton: 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, 1, 461-473.

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. Forthcoming in *Journal of Time Series Analysis*.

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: The proof of this Theorem holds for all $\lambda \geq 1$ and $M \geq 2$. For $t > s$, we have

$$\begin{aligned} Cov\left(Y_t, Y_s\right) &= \sum_{j,l=1}^{M} jl \left(\Pr\left(Y_t = j, Y_s = l\right) - \Pr\left(Y_t = j\right)\Pr\left(Y_s = l\right)\right) \\ &= \sum_{j,l=1}^{M} jl \Pr\left(Y_s = l\right)\left(\Pr\left(Y_t = j | Y_s = l\right) - \Pr\left(Y_t = j\right)\right) \\ &= \sum_{j,l=1}^{M} jl \Pr\left(Y_s = l\right)\{\Pr\left(Y_t = j | Y_s = l\right) \\ &\quad - \Pr\left(Y_t = j | Y_s = l\right)\Pr\left(Y_s = l\right) \\ &\quad - \Pr\left(Y_t = j | Y_s \neq l\right)\Pr\left(Y_s \neq l\right)\} \\ &= \sum_{j,l=1}^{M} jl \Pr\left(Y_s = l\right)\left(1 - \Pr\left(Y_s = l\right)\right)\Lambda_{t,s,j,l}, \end{aligned} \tag{17}$$

where $\Lambda_{t,s,j,l}$ is defined in (10). For $t > s + \lambda$, let

$$A_t = \left\{Y_{t-1}^{l} = Y_{t-1}^{l^c}, Y_{t-2}^{l} = Y_{t-2}^{l^c}, ..., Y_{t-\lambda}^{l} = Y_{t-\lambda}^{l^c}\right\}, \tag{18}$$

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. In words, A_t is the event that given a Y_s (with $t > s + \lambda$), each of the last λ lags Y_{t-j} $(j = 1, ..., \lambda)$ is unaffected by the outcome of Y_s. This event implies that the similarity weighted average $\bar{Y}_{t-1}^{s}$ is unaffected by the outcome of Y_s. Notice that

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

and therefore,

$$A_t \Rightarrow A_{t+1}, t > s + \lambda, \tag{19}$$

where '$\Rightarrow$' denotes implication. 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,

$$A_T \Longrightarrow \Lambda_{t,s,j,l} = 0, \forall t \geq T. \tag{20}$$

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 (4), 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 (19),

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

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 (20),

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

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 (9) follows on using (17).

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 (8), for large enough s, this implies (11). ■

Proof of Theorem 3: The proof holds for all $\lambda \geq 1$ and $M \geq 2$. In order to verify (12), we write:

$$\begin{aligned}
& \left|E\left[f\left(Y_s,...,Y_{s+k}\right)g\left(Y_{s+n},...,Y_{s+n+l}\right)\right]\right| \\
& -\left|E\left[f\left(Y_s,...,Y_{s+k}\right)\right]\right|\left|E\left[g\left(Y_{s+n},...,Y_{s+n+l}\right)\right]\right| \\
\leq\ & \left|E\left[f\left(Y_s,...,Y_{s+k}\right)g\left(Y_{s+n},...,Y_{s+n+l}\right)\right]\right. \\
& \left.-E\left[f\left(Y_s,...,Y_{s+k}\right)\right]E\left[g\left(Y_{s+n},...,Y_{s+n+l}\right)\right]\right| \\
=\ & \Big|\sum_{\substack{j_1,...,j_{k+1}\\ m_1,...,m_{l+1}}} f\left(B^s_{j_1,...,j_{k+1}}\right)g\left(C^{s+n}_{m_1,...,m_{l+1}}\right) \\
& \times\left[\Pr\left(B^s_{j_1,...,j_{k+1}},C^{s+n}_{m_1,...,m_{l+1}}\right)-\Pr\left(B^s_{j_1,...,j_{k+1}}\right)\Pr\left(C^{s+n}_{m_1,...,m_{l+1}}\right)\right]\Big| \\
=\ & \Big|\sum_{\substack{j_1,...,j_{k+1}\\ m_1,...,m_{l+1}}} f\left(B^s_{j_1,...,j_{k+1}}\right)g\left(C^{s+n}_{m_1,...,m_{l+1}}\right) \\
& \times\Pr\left(B^s_{j_1,...,j_{k+1}}\right)\left(1-\Pr\left(B^s_{j_1,...,j_{k+1}}\right)\right)\left[\Pr\left(C^{s+n}_{m_1,...,m_{l+1}}|B^s_{j_1,...,j_{k+1}}\right)\right. \\
& \left.-\Pr\left(C^{s+n}_{m_1,...,m_{l+1}}|\left(B^s_{j_1,...,j_{k+1}}\right)^c\right)\right]\Big|, \qquad (21)
\end{aligned}$$

where

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

and

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

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 $\forall \delta_1 > 0$,

$$\lim \inf_{n\to\infty} \inf_{B^c_{\delta_1}(\theta_0)} D_n(\theta_0, \theta_1) \tag{22}$$

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

Let

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

By Assumption A1, all the μ_j's are different from each other and therefore,

$$0 < \Delta_{t,j}(\theta) < 1$$

for all t, j and θ. Hence,

$$-\infty < \ln \Delta_{t,j}(\theta) < 0.$$

The series $\{l_{n,j}(\theta)\}$ is evidently nonpositive and uniformly bounded from below and by Theorem 3, 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 $\sum_{j=1}^{M} \Delta_{t,j}(\theta_0) = 1$,

$$\begin{aligned} E_{\theta_0}(D_n(\theta_1,\theta_0)) &= \frac{1}{n} E_{\theta_0} \sum_{t=1}^{n} \sum_{j=1}^{M} E_{\theta_0}\left(1\{y_t = j\} \ln \frac{\Delta_{t,j}(\theta_1)}{\Delta_{t,j}(\theta_0)} | \mathcal{F}_{t-1}\right) \\ &= \frac{1}{n} E_{\theta_0} \sum_{t=1}^{n} \sum_{j=1}^{M} \Delta_{t,j}(\theta_0) \ln \frac{\Delta_{t,j}(\theta_1)}{\Delta_{t,j}(\theta_0)} \\ &\leq \frac{1}{n} E_{\theta_0} \sum_{t=1}^{n} \ln(1) \qquad (23) \\ &= 0. \end{aligned}$$

If $\mu_0 \neq \mu_1$, $\Delta_t(\mu_0, w, \rho) \neq \Delta_t(\mu_1, w, \rho)$, $\forall t$ and if $w_0 \neq w_1$, $\bar{Y}_{t-1}^s \neq \bar{Y}_{t-1}^s$ with positive probability $\forall t$ under Assumption A0, which also implies $\Delta_t(\mu, w_0) \neq \Delta_t(\mu, w_1)$ with positive probability $\forall t$. 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 (23) 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

$$z_{n,\mu_k}(\theta) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} W_t^{\mu_k}(\theta), \qquad (24)$$

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,

$$z_{n,w_k}(\theta)=\frac{1}{\sqrt{n}}\sum_{t=1}^{n}W_t^{w_k}(\theta)\,, \tag{25}$$

where

$$W_t^{w_k}(\theta)=-\dot{h}_t^{w_k}(\theta)\sum_{j=1}^{M}1\{Y_t=j\}\,\frac{\delta_{t,j}(\theta)}{\Delta_{t,j}(\theta)}. \tag{26}$$

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 $\theta \in \Theta$,

$$\frac{\sigma_n^{i_k}(\theta)^2}{n} \xrightarrow{p} V^{i_k}(\theta) < \infty \tag{27}$$

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

$$\frac{1}{\sigma_n^{i_k}(\theta)^2} \sum_{t=1}^n \left(W_t^{i_k}(\theta)\right)^2 1\left\{\left|W_t^{i_k}(\theta)\right| > \varepsilon \sigma_n^{i_k}(\theta)\right\} \xrightarrow{p} 0. \tag{28}$$

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

$$\frac{1}{n}\sum_{t=1}^n \left(W_t^{\mu_k}(\theta)\right)^2 = \frac{1}{n}\sum_{t=1}^n f_{t,k}(\theta)^2 \left(\frac{1\{Y_t = k\}}{\Delta_{t,k}} - \frac{1\{Y_t = k+1\}}{\Delta_{t,k+1}}\right)^2 < \bar{K} \tag{29}$$

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 (28) trivially holds and we are done for $z_{n,\mu_k}(\theta)$.

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

$$\dot{h}_t^{w_k}(\theta) = \rho\left(\frac{\sum_{i<t} \dot{s}_{w_k}(X_i, X_t; w) Y_i}{\sum_{i<t} s(X_i, X_t; w)} - \frac{\sum_{i<t} s(X_i, X_t; w) Y_i \sum_{i<t} \dot{s}_{w_k}(X_i, X_t; w)}{\left(\sum_{i<t} s(X_i, X_t; w)\right)^2}\right), \tag{30}$$

where $\dot{s}_{w_k}(X_i, X_t; w) = \partial s(X_i, X_t; w)/\partial w_k$. It follows from (30) that under Assumptions A1-A2,

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

In view of (26) and the last inequality

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

so that, together with ergodicity,

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

Condition (28) also holds because $W_t^{w_k}\left(\theta\right)$ is uniformly bounded and $\sigma_n^{w_k}\left(\theta\right)$ behaves as $\sqrt{n}$ in probability. Similar reasoning follows for $W_t^{\rho}\left(\theta\right)$ and the proof of the Lemma 6 is therefore completed.■

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

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

is finite and nonsingular.

Proof of Lemma 7: We have

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

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

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

In view of (26) 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 $\left(\dot{h}_t^{w_k}\right)_{1\leq k\leq K}$ 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 $\operatorname{plim}\left(H_{n,\theta_j,\theta_k}\left(\hat{\theta}_n\right)\right)=H_{\theta_j,\theta_k}(\theta_0)$. 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	4.299	0.498	1.531	0.499	1.129	0.499	1.043	0.500
	Std	25.187	0.083	5.967	0.059	0.905	0.041	0.364	0.029
	Trim	1.605	0.498	1.177	0.500	1.077	0.499	1.031	0.500
$w_0 = 1$	Std Trim	2.014	0.073	0.774	0.051	0.484	0.035	0.305	0.025
	Median	1.033	0.498	0.998	0.499	0.995	0.499	1.000	0.500
	Q_1	0.488	0.441	0.609	0.459	0.706	0.473	0.792	0.480
	Q_3	1.983	0.557	1.581	0.540	1.384	0.525	1.235	0.519
	Mean	12.624	0.500	6.916	0.500	3.980	0.500	3.363	0.500
	Std	41.375	0.083	23.756	0.058	7.520	0.041	2.676	0.029
	Trim	6.640	0.500	4.019	0.500	3.438	0.500	3.199	0.500
$w_0 = 3$	Std Trim	13.148	0.073	3.588	0.051	1.629	0.036	0.980	0.025
	Median	2.914	0.501	3.041	0.501	3.097	0.501	3.063	0.499
	Q_1	1.546	0.444	1.932	0.459	2.230	0.472	2.440	0.481
	Q_3	5.971	0.557	4.727	0.540	4.256	0.528	3.775	0.518
	Mean	16.587	0.499	13.074	0.501	7.995	0.500	5.916	0.500
	Std	44.573	0.074	35.555	0.058	17.764	0.041	5.138	0.029
	Trim	10.481	0.499	8.097	0.501	6.146	0.500	5.463	0.500
$w_0 = 5$	Std Trim	20.133	0.063	10.217	0.050	3.696	0.035	1.999	0.025
	Median	4.815	0.500	5.118	0.502	5.177	0.500	5.037	0.500
	Q_1	2.593	0.451	3.117	0.463	3.636	0.473	3.967	0.480
	Q_3	9.791	0.546	8.946	0.538	7.429	0.527	6.550	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 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	1.601	0.500	1.171	0.500	1.072	0.501	1.033	0.500
	Std	5.120	0.082	0.968	0.057	0.579	0.040	0.381	0.029
	Trim	1.253	0.500	1.111	0.500	1.048	0.501	1.024	0.500
$w_0 = 1$	Std Trim	1.218	0.071	0.713	0.050	0.462	0.035	0.324	0.025
	Median	0.977	0.500	0.992	0.500	0.995	0.500	0.997	0.500
	Q_1	0.396	0.445	0.584	0.463	0.693	0.474	0.777	0.481
	Q_3	1.790	0.555	1.533	0.537	1.351	0.527	1.248	0.519
	Mean	4.986	0.499	3.699	0.500	3.246	0.500	3.098	0.500
	Std	14.116	0.081	5.228	0.057	1.390	0.040	0.814	0.028
	Trim	3.733	0.499	3.339	0.500	3.173	0.500	3.078	0.500
$w_0 = 3$	Std Trim	2.990	0.071	1.661	0.050	1.045	0.035	0.698	0.024
	Median	2.915	0.499	2.986	0.500	3.028	0.501	3.004	0.500
	Q_1	1.745	0.444	2.098	0.461	2.373	0.473	2.520	0.482
	Q_3	4.843	0.555	4.219	0.539	3.881	0.526	3.586	0.518
	Mean	9.723	0.500	6.539	0.501	5.457	0.500	5.147	0.500
	Std	32.849	0.083	14.780	0.059	2.576	0.040	1.438	0.029
	Trim	6.529	0.500	5.621	0.501	5.296	0.500	5.096	0.500
$w_0 = 5$	Std Trim	5.770	0.072	2.934	0.051	1.845	0.034	1.201	0.025
	Median	4.858	0.502	4.963	0.501	4.913	0.500	4.928	0.500
	Q_1	2.990	0.444	3.442	0.462	3.838	0.473	4.151	0.481
	Q_3	8.037	0.555	7.048	0.540	6.437	0.526	5.943	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 $\mu_0 = 0.5$ and $\lambda = 10$.

	n	250		500		1000		2000	
		$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$
	Mean	1.697	0.500	1.269	0.500	1.091	0.499	1.035	0.499
	Std	3.620	0.081	1.505	0.057	0.794	0.040	0.504	0.028
	Trim	1.362	0.500	1.155	0.500	1.052	0.499	1.019	0.499
$w_0 = 1$	Std Trim	1.731	0.071	0.965	0.050	0.602	0.035	0.411	0.024
	Median	0.953	0.498	0.986	0.499	0.986	0.498	0.975	0.500
	Q_1	0.219	0.445	0.433	0.461	0.587	0.471	0.701	0.480
	Q_3	1.991	0.554	1.729	0.539	1.454	0.526	1.297	0.518
	Mean	4.766	0.500	3.579	0.500	3.265	0.499	3.086	0.499
	Std	10.682	0.081	3.009	0.056	1.678	0.040	1.016	0.028
1	Trim	3.878	0.500	3.345	0.499	3.178	0.499	3.058	0.499
$w_0 = 3$	Std Trim	3.852	0.071	2.035	0.049	1.297	0.035	0.868	0.024
	Median	2.796	0.500	2.905	0.500	3.005	0.498	2.940	0.499
	Q_1	1.328	0.446	1.827	0.462	2.203	0.471	2.381	0.480
	Q_3	5.216	0.555	4.513	0.538	4.019	0.526	3.663	0.519
	Mean	7.608	0.500	5.824	0.499	5.399	0.498	5.116	0.499
	Std	15.915	0.081	4.416	0.0579	2.524	0.041	1.542	0.028
	Trim	6.182	0.499	5.506	0.499	5.280	0.498	5.068	0.499
$w_0 = 5$	Std Trim	5.574	0.070	3.117	0.050	2.014	0.035	1.289	0.025
	Median	4.654	0.498	4.852	0.499	4.984	0.498	4.947	0.499
	Q_1	2.234	0.445	3.215	0.462	3.728	0.471	4.056	0.480
	Q_3	8.383	0.554	7.151	0.538	6.5777	0.525	5.967	0.518

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 $\lambda = 2$.

	n	250		500		1000		2000	
		$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$
	Mean	3.572	0.297	2.110	0.298	1.143	0.300	1.059	0.300
	Std	21.488	0.085	12.308	0.060	0.755	0.042	0.409	0.030
	Trim	1.603	0.297	1.215	0.298	1.090	0.300	1.043	0.300
$w_0 = 1$	Std Trim	2.115	0.074	0.868	0.052	0.494	0.037	0.326	0.026
	Median	1.010	0.298	1.022	0.299	0.998	0.301	1.003	0.299
	Q_1	0.483	0.238	0.608	0.256	0.707	0.271	0.791	0.280
	Q_3	1.893	0.354	1.561	0.337	1.394	0.329	1.267	0.320
	Mean	11.653	0.295	6.481	0.298	4.281	0.300	3.338	0.300
	Std	38.079	0.085	20.200	0.059	10.253	0.042	1.929	0.030
	Trim	6.241	0.295	4.203	0.298	3.483	0.300	3.204	0.300
$w_0 = 3$	Std Trim	10.761	0.074	3.914	0.052	1.890	0.037	1.078	0.026
	Median	3.008	0.297	2.982	0.299	2.990	0.300	2.997	0.301
	Q_1	1.620	0.238	1.942	0.258	2.177	0.272	2.352	0.280
	Q_3	6.039	0.353	4.882	0.339	4.316	0.329	3.849	0.321
	Mean	20.497	0.298	12.716	0.301	8.237	0.300	5.975	0.300
	Std	53.766	0.0842	32.829	0.059	18.844	0.041	7.266	0.030
	Trim	14.160	0.298	8.172	0.301	6.144	0.300	5.451	0.300
$w_0 = 5$	Std Trim	34.347	0.073	10.250	0.051	3.982	0.036	2.023	0.026
	Median	4.808	0.301	4.961	0.302	5.007	0.299	5.013	0.300
	Q_1	2.327	0.242	3.076	0.261	3.593	0.273	3.978	0.281
	Q_3	10.661	0.356	8.712	0.339	7.306	0.328	6.481	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. 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	2.125	0.299	1.196	0.300	1.086	0.300	1.036	0.300
	Std	17.826	0.084	1.407	0.059	0.614	0.041	0.374	0.029
	Trim	1.235	0.299	1.109	0.300	1.057	0.300	1.028	0.300
$w_0 = 1$	Std Trim	1.211	0.073	0.713	0.052	0.462	0.036	0.320	0.026
	Median	0.952	0.299	0.988	0.300	0.999	0.300	1.011	0.300
	Q_1	0.413	0.244	0.586	0.260	0.704	0.272	0.776	0.281
	Q_3	1.762	0.356	1.516	0.341	1.369	0.329	1.263	0.320
	Mean	6.215	0.297	3.922	0.299	3.244	0.299	3.111	0.299
	Std	24.331	0.084	9.208	0.059	1.410	0.041	0.883	0.029
	Trim	3.928	0.297	3.407	0.300	3.179	0.299	3.083	0.299
$w_0 = 3$	Std Trim	3.457	0.072	1.836	0.051	1.129	0.035	0.734	0.025
	Median	2.990	0.299	3.001	0.299	2.994	0.299	2.982	0.299
	Q_1	1.682	0.243	2.047	0.260	2.307	0.272	2.504	0.279
	Q_3	4.977	0.353	4.413	0.338	3.917	0.326	3.621	0.319
	Mean	9.562	0.298	6.755	0.300	5.385	0.301	5.201	0.301
	Std	29.605	0.084	16.343	0.059	2.531	0.041	1.504	0.029
	Trim	6.369	0.298	5.658	0.300	5.222	0.301	5.139	0.301
$w_0 = 5$	Std Trim	5.269	0.074	3.076	0.051	1.776	0.035	1.192	0.025
	Median	4.990	0.300	4.960	0.302	4.928	0.301	5.001	0.301
	Q_1	2.931	0.239	3.392	0.260	3.892	0.274	4.207	0.282
	Q_3	8.091	0.355	7.241	0.340	6.340	0.327	5.991	0.321

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 6. Simulated MLE point estimates for $\mu_0 = 0.3$ and $\lambda = 10$.

	n	250		500		1000		2000	
		$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$	$\hat{w}$	$\hat{\mu}$
	Mean	2.119	0.296	1.308	0.297	1.129	0.298	1.048	0.299
	Std	9.604	0.082	1.612	0.056	0.846	0.041	0.526	0.029
	Trim	1.460	0.296	1.188	0.298	1.090	0.298	1.033	0.299
$w_0 = 1$	Std Trim	1.949	0.071	1.055	0.049	0.661	0.035	0.441	0.025
	Median	0.980	0.296	1.007	0.298	1.011	0.298	0.987	0.299
	Q_1	0.164	0.242	0.416	0.259	0.582	0.272	0.695	0.280
	Q_3	2.086	0.350	1.751	0.336	1.525	0.325	1.340	0.318
	Mean	5.723	0.297	3.597	0.298	3.233	0.299	3.093	0.299
	Std	18.085	0.081	3.128	0.056	1.659	0.040	1.029	0.029
	Trim	4.085	0.297	3.351	0.299	3.162	0.299	3.069	0.299
$w_0 = 3$	Std Trim	4.199	0.071	2.086	0.049	1.331	0.035	0.885	0.025
	Median	2.835	0.296	2.916	0.299	2.966	0.299	2.976	0.299
	Q_1	1.354	0.245	1.825	0.260	2.120	0.272	2.364	0.280
	Q_3	5.545	0.351	4.463	0.337	4.038	0.327	3.731	0.317
	Mean	8.488	0.297	5.931	0.299	5.400	0.299	5.141	0.299
	Std	21.108	0.081	4.701	0.056	2.602	0.041	1.588	0.029
	Trim	6.438	0.297	5.574	0.299	5.272	0.299	5.096	0.299
$w_0 = 5$	Std Trim	6.131	0.071	3.254	0.049	2.050	0.035	1.330	0.025
	Median	4.682	0.297	4.905	0.300	4.975	0.299	5.031	0.299
	Q_1	2.201	0.244	3.139	0.261	3.676	0.272	4.033	0.279
	Q_3	8.578	0.351	7.360	0.337	6.605	0.3267	6.019	0.318

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 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.

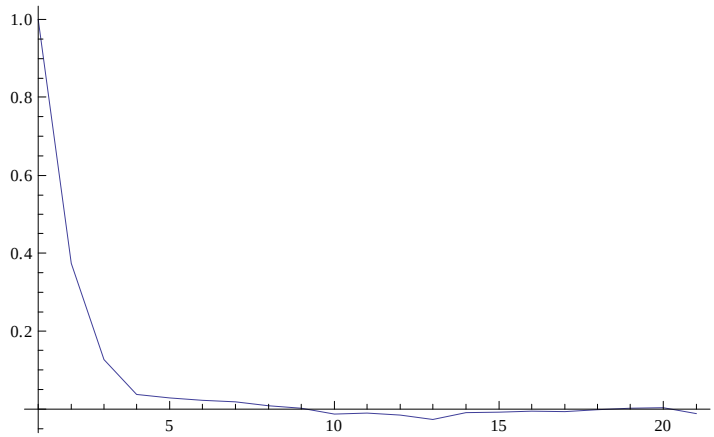

Figure 1. Correlogram of the process in the case $\lambda = 1$, $M = 2$, $n = 10000$.

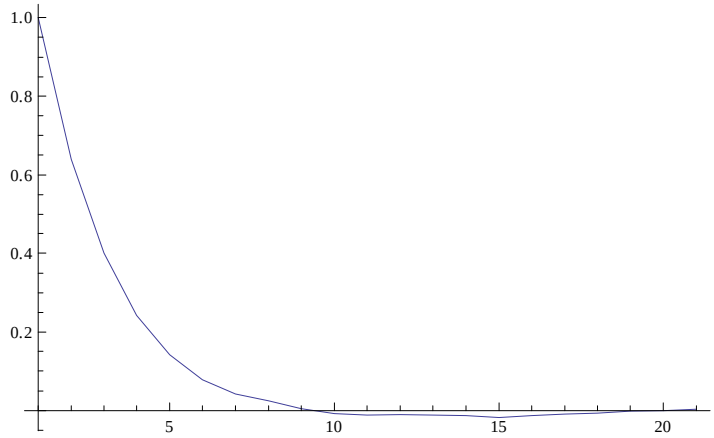

Figure 2. Correlogram of the process in the case $\lambda = 1$, $M = 3$, $n = 10000$.

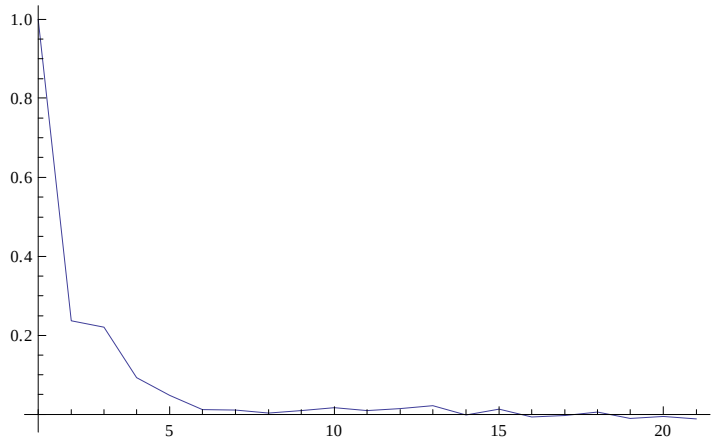

Figure 3. Correlogram of the process in the case $\lambda = 2$, $M = 2$, $n = 10000$.

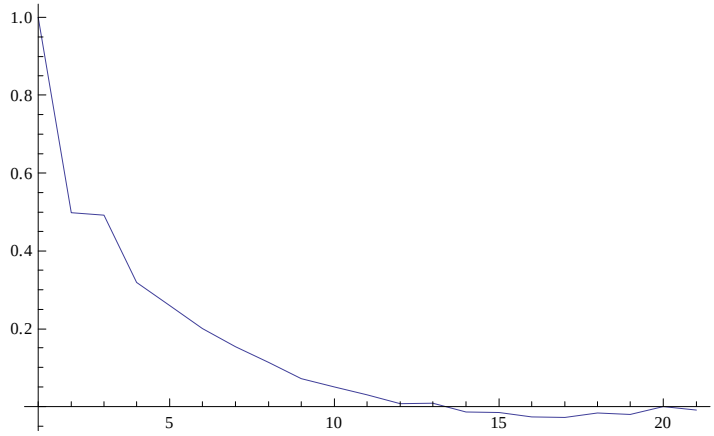

Figure 4. Correlogram of the process in the case $\lambda = 2$, $M = 3$, $n = 10000$.

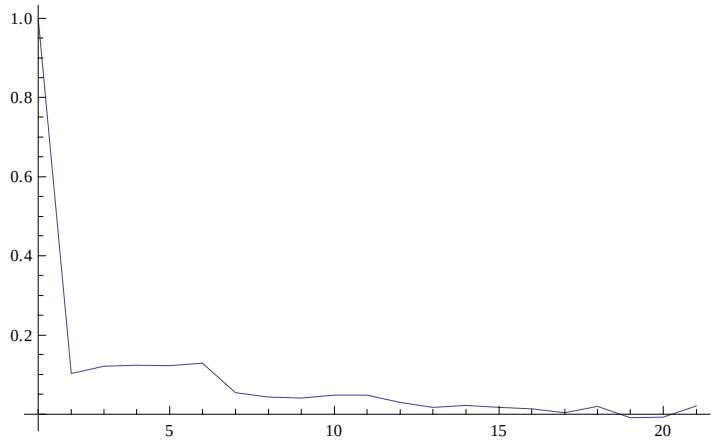

Figure 5. Correlogram of the process in the case $\lambda = 5$, $M = 2$, $n = 10000$.

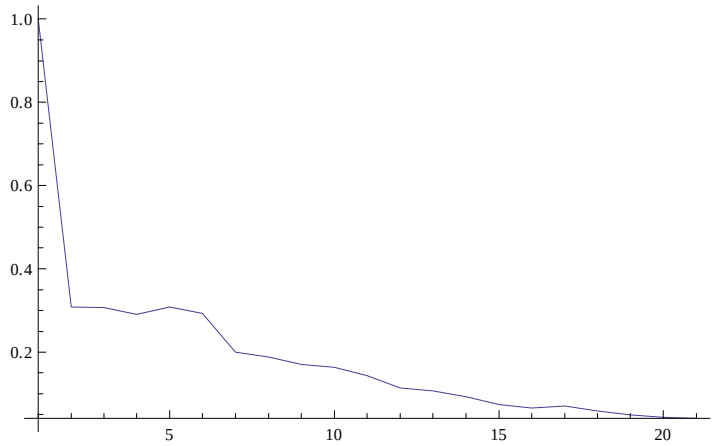

Figure 6. Correlogram of the process in the case $\lambda = 5$, $M = 3$, $n = 10000$.

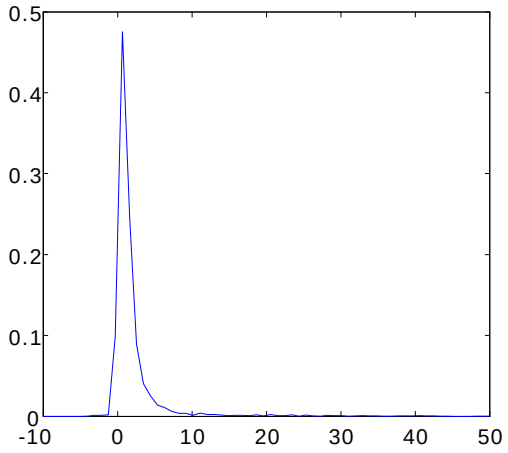

Figure 7. Kernel density estimate for $\hat{w}$, $w_0 = 1$, $\mu_0 = 0.3$, $\lambda = 2$, $n = 250$.

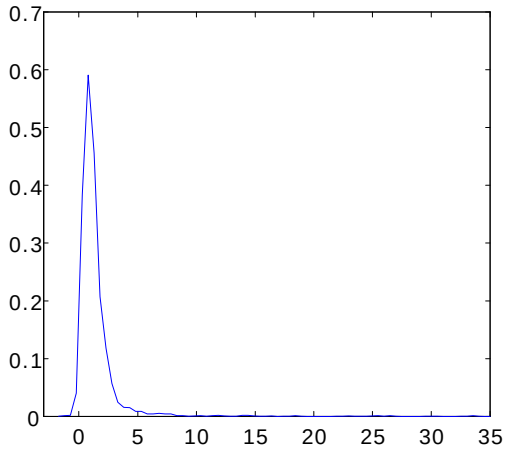

Figure 8. Kernel density estimate for $\hat{w}$, $w_0 = 1$, $\mu_0 = 0.3$, $\lambda = 2$, $n = 500$.

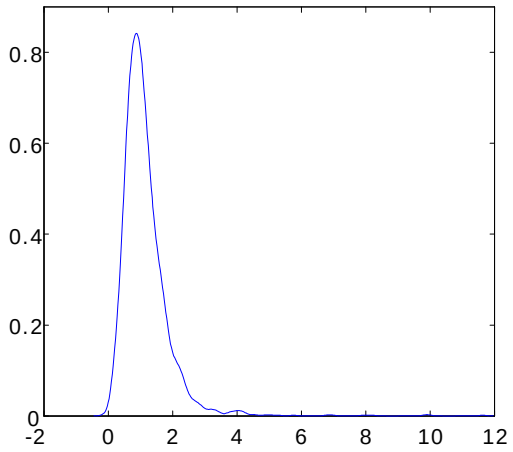

Figure 9. Kernel density estimate for $\hat{w}$, $w_0 = 1$, $\mu_0 = 0.3$, $\lambda = 2$, $n = 1000$.

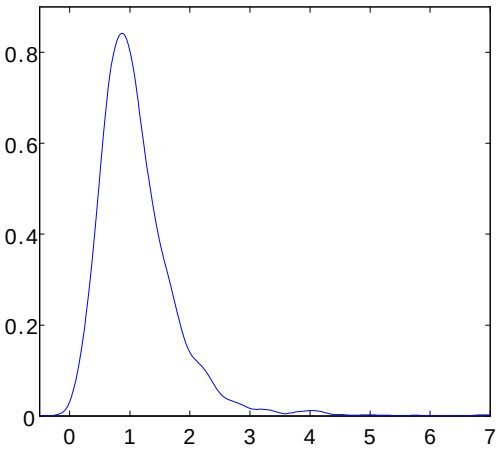

Figure 10. Kernel density estimate for $\hat{w}$, $w_0 = 1$, $\mu_0 = 0.3$, $\lambda = 2$, $n = 2000$.

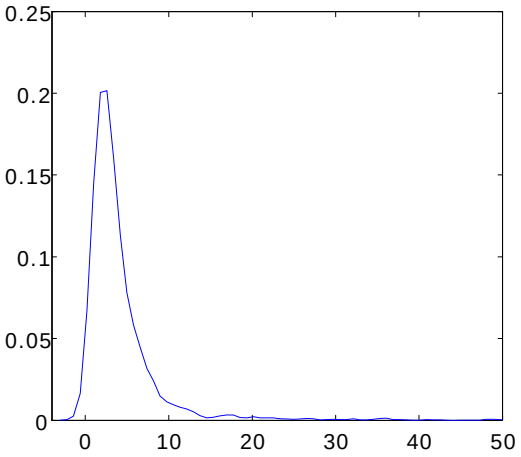

Figure 11. Kernel density estimate for $\hat{w}$, $w_0 = 3$, $\mu_0 = 0.5$, $\lambda = 5$, $n = 250$.

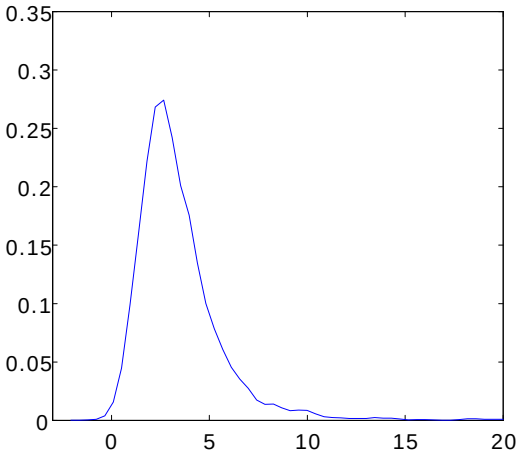

Figure 12. Kernel density estimate for $\hat{w}$, $w_0 = 3$, $\mu_0 = 0.5$, $\lambda = 5$, $n = 500$.

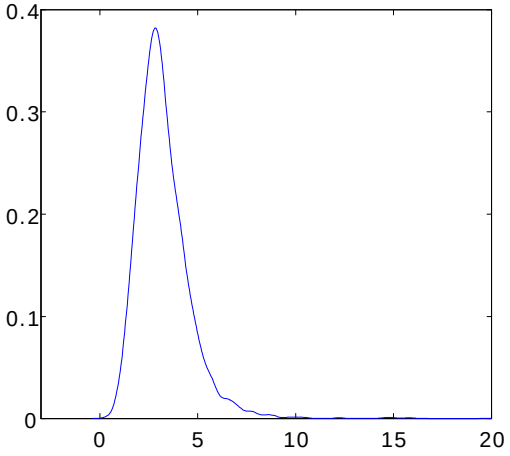

Figure 13. Kernel density estimate for $\hat{w}$, $w_0 = 3$, $\mu_0 = 0.5$, $\lambda = 5$, $n = 1000$.

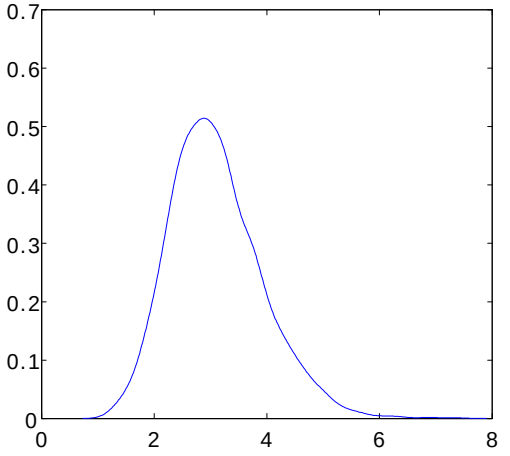

Figure 14. Kernel density estimate for $\hat{w}$, $w_0 = 3$, $\mu_0 = 0.5$, $\lambda = 5$, $n = 2000$.

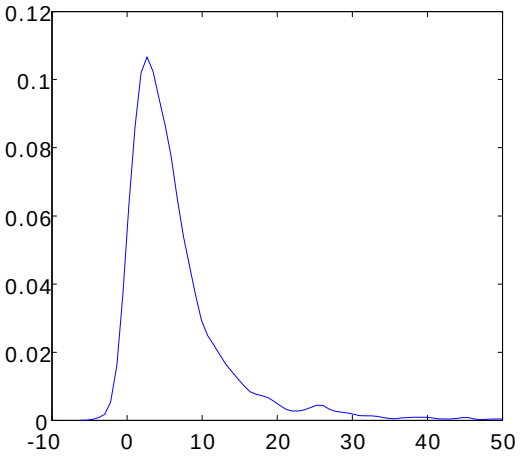

Figure 15. Kernel density estimate for $\hat{w}$, $w_0 = 5$, $\mu_0 = 0.5$, $\lambda = 10$, $n = 250$.

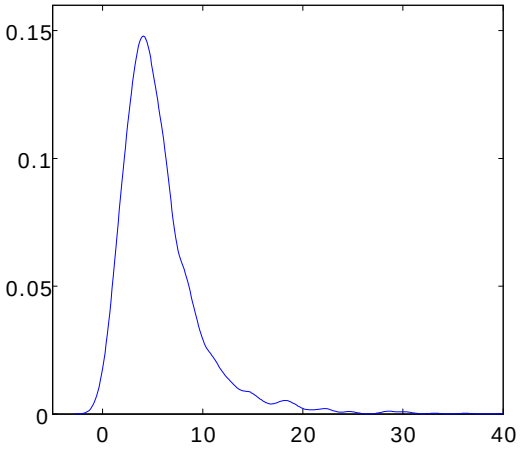

Figure 16. Kernel density estimate for $\hat{w}$, $w_0 = 5$, $\mu_0 = 0.5$, $\lambda = 10$, $n = 500$.

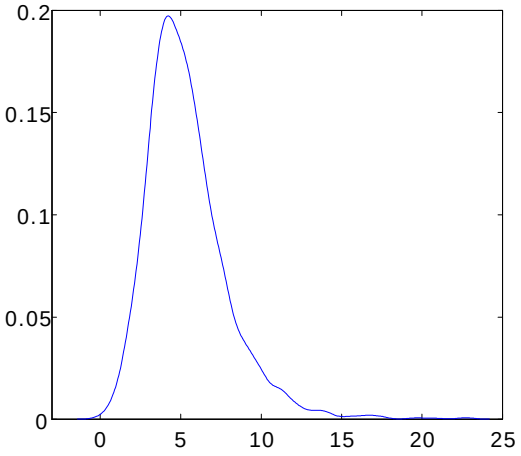

Figure 17. Kernel density estimate for $\hat{w}$, $w_0 = 5$, $\mu_0 = 0.5$, $\lambda = 10$, $n = 1000$.

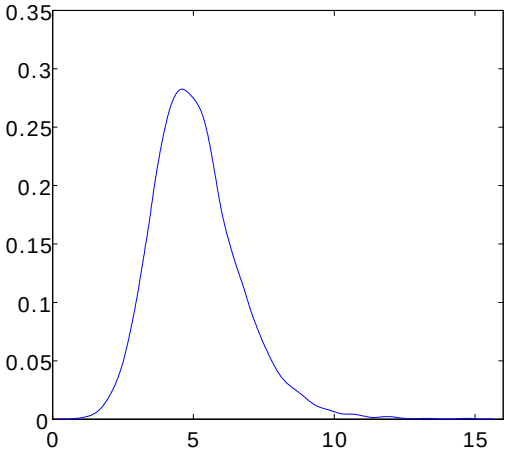

Figure 18. Kernel density estimate for $\hat{w}$, $w_0 = 5$, $\mu_0 = 0.5$, $\lambda = 10$, $n = 2000$.