

# An Alternative to Unit Root Tests: Bridge Estimators Differentiate between Nonstationary versus Stationary Models and Select Optimal Lag

Mehmet Caner  
North Carolina State University \*

Keith Knight  
University of Toronto

March 4, 2011

## Abstract

This paper introduces a novel way of differentiating a unit root from stationary alternatives using so-called “Bridge” estimators; this estimation procedure can potentially generate exact zero estimates of parameters. We exploit this property and treat this as a model selection problem. We show that Bridge estimators can select the correct model with probability tending to 1. They estimate “zero” parameter on the lagged dependent variable as zero (nonstationarity), if this is nonzero (stationary), estimate the coefficient with standard normal limit. In this sense, we extend the statistics literature as well, since that literature only deals with model selection among only stationary variables.

The reason that our methodology can outperform the existing unit root tests with lag selection methods stems from the two-step nature of existing unit root tests. In our method, we select the optimal lag length and unit root simultaneously. We show that in simulations, this makes big difference in terms of size and power.

---

\*Mehmet Caner: Department of Economics, 4168 Nelson Hall, Raleigh, NC 27518. email: mcaner@ncsu.edu. Keith Knight: Department of Statistics, University of Toronto, 100 St. George St., Toronto, Ont. M5S 3G3 e-mail: keith@utstat.toronto.edu. We thank the participants at Nonlinear Time Series Conference in Xiamen, China, May 2008, and Duke-UNC-NCSU econometrics seminar in December 2008. The research of Keith Knight was supported by a grant from NSERC of Canada.

# 1 Introduction

Unit root testing has been a cornerstone of applied econometrics research. There are a variety of unit root tests, most notably those proposed by Dickey and Fuller (1979), Phillips (1987), Elliot, Rothenberg, and Stock (1996), and for panel data, the test of Bai and Ng (2004). These tests depend on choosing the correct lag length given a maximum amount of possible lags. AIC, BIC and t-tests, as well as sophisticated Modified AIC of Ng and Perron (2001) are proposed for lag selection. Related to this point, in model selection, Breiman (1996) shows that subset selection methods such as AIC, BIC can often select the wrong model. This problem is more acute with large number of parameters. Problems with lag selection in unit root testing is also observed by Hall (1994). Breiman (1996) further shows that ridge-regression based model selection gives good results, and is not affected by the number of the parameters.

Related to ridge-regression, a method that penalizes the parameters less, and has no asymptotic bias gained prominence in the statistics. These are Bridge/Lasso type estimators. In the last couple of years, there has been a huge interest in Bridge/Lasso estimation in the statistical literature. Knight and Fu (2000) establish the limit law for Lasso in a LS framework. Fan and Li (2001, 2002) develop penalized LS with a smooth penalty function. They show that we can derive the limit of the LS estimators as if they were known in advance. In other words, these methods simultaneously select the model and estimate the coefficients. In econometrics Knight (2008) and Caner (2009) consider LS and GMM based Lasso estimators respectively. Caner (2009) also finds that in small samples Lasso estimator have smaller mean square error and bias than the ones chosen by AIC, BIC and sequential testing procedures. To this point, Bridge/Lasso methods have only been used as a model selection tool in stationary models.

In this paper we propose a novel way of using Bridge estimators to distinguish between stationary and unit root models. The proposed methods can be used either as a complement or as a substitute for unit root tests for which lag selection is necessary. The idea is to use the Bridge to do model selection while simultaneously determining whether the parameter of the lagged dependent variable is zero or non-zero. If the true parameter of the lagged dependent variable is nonzero, then with high probability, the Bridge estimates it as nonzero, and hence finds stationarity. This is a somewhat nonstandard approach to determining whether the variable is stationary or nonstationary. The classical approach to this problem has been to first select the lag length, and then to apply a unit root test. Our procedure eliminates the two-step nature of detection. By doing that, it is possible to have power gains and thus we are able to differentiate better between stationary and unit root behavior.

We should note that Bridge estimation is not a testing procedure *per se*. It is, in fact,

a model selection procedure that can also estimate the parameters. It also involves penalty parameters that can be tuned to help in selecting between different models. The penalties are not estimated but rather are determined theoretically. We use the probability of correct model selection as a basis to compare Bridge with unit root test results.

In this paper, we derive the limit laws for the Bridge estimators in a number of possible models, including the case of mixed nonstationary and stationary regressors. We also show that we need the exponent of the penalty function for the lagged dependent variable to be less than  $1/2$  (as opposed to 1, which is typically the case in the statistical literature) to avoid asymptotic bias, in addition to differentiating between stationary and non-stationary models. We show that we can estimate "zero" parameters as "zeros" and the estimators of nonzero parameters have the standard limiting distribution. Essentially, we are using the main idea of model selection in Bridge estimation to select between unit root versus stationary behavior as well as simultaneously estimate the coefficients of the lags.

We also analyze a model with near-integrated properties, and a model with time trend. Simulations show that our method generally displays superior behavior compared with Dickey-Fuller GLS, (DFGLS), and the Augmented Dickey-Fuller unit root tests (ADF). Much of the theoretical hoopla given Bridge-Lasso (and other shrinkage-based) methods is criticized by Leeb and Pötscher (2008). In particular, they show that at certain parameter values (moderately away from zero), these methods produced estimators that have high mean squared errors in small samples. However, our idea is to compare and contrast our method with unit root tests with the MAIC lag selection method, which face the same criticisms as shrinkage-based methods.

Section 2 introduces the simple model without lags, and introduces the main idea. Section 3 adds lags to the main model. Section 4 considers the most general model with lags and time trend. We present some Monte Carlo evidence in Section 5 and provide some concluding remarks and future directions in Section 6.

## 2 The Simple Model

The following is the benchmark model. This is without any lags and illustrates the main points of the paper. We do not know *a priori* whether the following  $y_t$  is stationary or nonstationary. Our aim is to show that when the true model is nonstationary ( $\rho_0 = 0$ ), the estimator of  $\rho_0$  will converge to zero in probability and therefore  $y_{t-1}$  can be dropped from the regression, and if it is stationary, the estimator will converge in a probability to the appropriate for the stationary AR(1) model. In other words, we are looking for an oracle property. The model is:

$$\Delta y_t = \rho_0 y_{t-1} + e_t, \tag{1}$$

where  $e_t$  are iid with mean 0 and variance  $\sigma^2$ , and has finite fourth moments. We estimate the parameter of interest  $\rho_0$  by minimizing the following objective function given  $\lambda_T, \gamma$ :

$$Z_T(\rho) = \sum_{t=1}^T (\Delta y_t - \rho y_{t-1})^2 + \lambda_T |\rho|^\gamma. \quad (2)$$

In the remaining parts of the paper,  $\lambda_T$ , and  $\gamma$  will be specified. We first consider the consistency of our estimator. In that respect, when  $\rho_0 = 0$ , (nonstationary model) we benefit from the following function

$$Z_{T1}(\rho) = \frac{1}{T^2} \sum_{t=1}^T (\Delta y_t - \rho y_{t-1})^2 + \frac{\lambda_T}{T^2} |\rho|^\gamma. \quad (3)$$

When  $\rho_0 < 0$ , (stationary), we use the following

$$Z_{T2}(\rho) = \frac{1}{T} \sum_{t=1}^T (\Delta y_t - \rho y_{t-1})^2 + \frac{\lambda_T}{T} |\rho|^\gamma. \quad (4)$$

Now we show that regardless of whether  $\rho_0 = 0$  (nonstationary model) or  $\rho_0 < 0$  (stationary model), we can consistently estimate the parameter by  $\hat{\rho}$  which minimizes (2).

**Theorem 1.**

(i). For  $\rho_0 = 0$  and  $\lambda_T/T^2 \rightarrow \lambda_0 \geq 0$ ,

$$\hat{\rho} \xrightarrow{p} \operatorname{argmin}_{\rho \in A} Z_1(\rho),$$

where

$$Z_1(\rho) = \rho^2 \sigma^2 \int_0^1 W^2(r) dr + \lambda_0 |\rho|^\gamma,$$

where  $W(r)$  is the standard Brownian Motion and  $\int_0^1 W(r)^2 dr$  is a nonstandard distribution mentioned in Chapter 17 of Hamilton (1994). "A" is a compact set specified in the proofs.

(ii). For  $\rho_0 < 0$  (stationary case) and  $\lambda_T/T \rightarrow \lambda_0 \geq 0$ ,

$$\hat{\rho} \xrightarrow{p} \operatorname{argmin}_{\rho \in A} Z_2(\rho),$$

where

$$Z_2(\rho) = \sigma^2 + (\rho - \rho_0)^2 \Gamma_0 + \lambda_0 |\rho|^\gamma,$$

and  $\Gamma_0 = E y_{t-1}^2$ ,  $\Gamma_0 > 0$ .

(iii). If  $\lambda_T = o(T)$ , then estimators in both cases (i)-(ii) are consistent.

Remark. This theorem states that regardless of stationarity or nonstationarity we can estimate  $\rho_0$  consistently as long as  $\lambda_T = o(T)$ . This also shows that the criterion for consistency  $\lambda_T = o(T)$  in the stationary case (ii), is the same one for the nonstationary case as well. This is the same criterion that is found in LS stationary case of Knight and Fu (2000).

This fact also can be seen by comparing (2) with (3) and (2) with (4). Note that  $\hat{\rho}$  is the minimizer of (2) and converges in probability to  $\rho_0$  regardless of  $\rho_0 = 0$  or  $\rho_0 < 0$ . This is clear by (2)-(4)

$$\hat{\rho} = \operatorname{argmin} Z_T(\rho) = \operatorname{argmin} Z_{T_1}(\rho) = \operatorname{argmin} Z_{T_2}(\rho).$$

The following theorem displays the limits when  $\rho_0 = 0$  (nonstationary case), and when  $\rho < 0$ , (stationary case). This is a new result both in the statistics and econometrics literature. By using the Bridge estimators as model selection devices, we can differentiate between the stationarity versus nonstationarity.

**Theorem 2.** *Suppose  $0 < \gamma < 1/2$ , and  $\lambda_T/T^\gamma \rightarrow \lambda_0 \geq 0$ , then*

(i). *If  $\rho_0 = 0$  then*

$$\hat{u} = T\hat{\rho} \xrightarrow{d} u_0 = \operatorname{argmin} V_1(u),$$

where

$$V_1(u) = -2u\sigma^2 \int_0^1 W(r)dW(r) + u^2\sigma^2 \int_0^1 W(r)^2 dr + \lambda_0|u|^\gamma,$$

where  $\int_0^1 W(r)dW(r)$  is the stochastic integral and the distribution is described in equation 17.3.26 of Hamilton (1994).

(ii). *If  $\rho_0 < 0$ , and  $\Gamma_0 > 0$ , then*

$$\hat{u} = T^{1/2}(\hat{\rho} - \rho_0) \xrightarrow{d} u_0 = \operatorname{argmin} V_2(u),$$

where  $V_2(u) = -2uL + u^2\Gamma_0$ , and  $L \equiv N(0, \sigma^2\Gamma_0)$ , so  $u_0 \equiv N(0, \sigma^2\Gamma_0^{-1})$ .

Remarks. 1. We clearly see that by (i), we can shrink the estimates of zero parameter (nonstationary case) to zero with positive probability and nonzero regression parameter, (ii), is estimated via standard limit (stationary case). This shows that Bridge estimator basically can select between stationary and nonstationary models. The estimator acts like a unit root test in that sense. Note that we setup  $T\hat{\rho}$  and if the true model is  $\rho_0 < 0$  (stationary  $y_t$ ), then  $\hat{u} = T\hat{\rho} \xrightarrow{P} -\infty$ , as can be seen from Theorem 2ii, instead of shrinking to zero with positive probability. This shows that two cases can be differentiated through Bridge estimators.

2. Note that if  $\rho_0 = 0$ , then as mentioned we obtain the result with positive probability. For  $\rho_0 < 0$ , since the parameter is nonzero we do not have the penalty term in the limit and  $\lambda_T$  is chosen accordingly to achieve this ( $\lambda_T/T^\gamma \rightarrow \lambda_0$  where  $0 < \gamma < 1/2$ ).

3. The  $\gamma$  coefficient should be smaller than 1/2 unlike LS case of Knight and Fu (2000). This is due to the differences between stationary and nonstationary cases. For the stationary case,  $\lambda_T = o(T^{1/2})$  is needed, and for the nonstationary case  $\lambda_T = O(T^\gamma)$  is needed to differentiate between the two cases by Bridge estimator we require  $0 < \gamma < 1/2$ . In our setup, by the switch of the coefficient from zero to negative, the variable  $y_t$  switches from

nonstationary to stationary. This has not been analyzed before in the context Bridge estimation; shrinkage based estimation has, to our knowledge, been employed only for model selection in the case of stationary variables.

4. One of the main ingredients of the proof is if  $\lambda_T = o(T^{1/2})$ , the penalty term converges to zero when there is stationarity ( $\rho_0 < 0$ ), but for  $\rho_0 = 0$  we need to define the penalty term so that  $\lambda_T/T^\gamma \rightarrow \lambda_0$ , where  $0 < \gamma < 1/2$ .

5. One of the most important issues is a local analysis. To do this, we set  $\rho_0 = -C/T$ , where  $C$  is a positive constant. This is also called the near-integrated framework in the unit root literature. Under Assumptions of Theorem 2, following the proof of Theorem 2i, we can show

$$\hat{u} = T[\hat{\rho} - (-C/T)] \xrightarrow{d} \operatorname{argmin} V_c(u),$$

where

$$V_c(u) = -2u\sigma^2 \int_0^1 W_c(r) dW_c(r) + u^2\sigma^2 \int_0^1 W_c(r)^2 dr + \lambda_0|u - C|^\gamma, \quad (5)$$

where  $W_c(r)$  represents an Ornstein Uhlenbeck process. We see two differences with the limit in Theorem 2i. First, Brownian motions are replaced with Ornstein Uhlenbeck processes. Second, and most important, the penalty term is centered differently. This shows that the limit distribution of  $T(\hat{\rho} - (-C/T))$  puts positive probability on  $C$ . This last point clearly demonstrates that we can differentiate between nonstationary and near integrated behavior as well as the stationary one in Theorem 2ii.

The limit distribution of  $T\hat{\rho}$  (not shown here, but easy transformation from (5)) puts positive probability on 0. This probability decreases as  $C$  increases. The near-integrated results apply to case with lags and time trends below. We do not show them below.

In fact, a slightly better result than Theorem 2 can be obtained by choosing  $\lambda_T = O(T^\alpha)$  where  $\gamma < \alpha < 1/2$ . This is given below in Corollary 1. We show that the estimator converges to zero with probability one when we penalize slightly more than Theorem 2. The result in Theorem 2i changes and Theorem 2ii stays the same.

**Corollary 1.** *Suppose  $0 < \gamma < \alpha < 1/2$ ,  $\lambda_T/T^\alpha \rightarrow \lambda_0 \geq 0$ , then if  $\rho_0 = 0$  then*

$$\hat{u} = T\hat{\rho} \xrightarrow{wp1} 0.$$

### 3 Model with Lags

The model in this section extends (1) with the addition of lags to account for serial correlation. This setup is used widely in time series econometrics in unit root testing, see Chapter 17 of Hamilton (1994) for a wide number of references. The following is equation (17.7.6) of Hamilton (1994).

$$\Delta y_t = \rho_0 y_{t-1} + \zeta_{1,0} \Delta y_{t-1} + \cdots + \zeta_{p-1,0} \Delta y_{t-(p-1)} + e_t, \quad (6)$$

where  $e_t$  is iid, and distributed with mean 0 and variance of  $\sigma^2$ , and with finite fourth moments. We can simplify the notation somewhat by defining  $M_{t-1} = (\Delta y_{t-1}, \dots, \Delta y_{t-(p-1)})'$  which is a  $(p-1) \times 1$  vector and  $C_M = \frac{1}{T} \sum_{t=2}^T M_{t-1} M_{t-1}'$ , and  $\Gamma_M = \lim_{T \rightarrow \infty} C_M$ . We also define  $w_{t-1} = (y_{t-1}, M_{t-1}')'$ . We now estimate  $\rho_0, \zeta_{1,0}, \dots, \zeta_{p-1,0}$  by minimizing

$$\sum_{t=1}^T (\Delta y_t - \rho y_{t-1} - \zeta_1 \Delta y_{t-1} - \cdots - \zeta_{p-1} \Delta y_{t-(p-1)})^2 + \lambda_T |\rho|^{\gamma_1} + b_T \sum_{j=1}^{p-1} |\zeta_j|^{\gamma_2}, \quad (7)$$

where  $\lambda_T, b_T, \gamma_1, \gamma_2$  will be specified in the assumptions below and the theorem statements. Before stating the consistency theorem we need the following assumptions.

### Assumptions

1.  $e_t$  is iid, has zero mean and  $E(e_t^2 | w_{t-1}) = \sigma^2$ , and has finite fourth moments with  $\sigma^2 > 0$ .
2. For the nonstationary case,  $\Gamma_M$  is nonsingular, and maximum eigenvalue of  $\Gamma_M$  is bounded away from infinity. For the stationary case  $\Gamma_W = E w_{t-1} w_{t-1}'$ , is nonsingular and finite.
3. True values of the parameters  $\zeta_{1,0}, \dots, \zeta_{p-1,0}$  are bounded away from infinity, and are in a compact set in  $R^{p-1}$ .
4.  $\rho_0 \in A$ , where  $A = [-2 + a, 0]$  where  $a$  is a small positive constant.
5.  $\max(\lambda_T, b_T) = o(T)$ .

These assumptions are standard in the Bridge/Lasso estimation literature as in Knight and Fu (2000), Huang, Horowitz and Ma (2008). Assumption 4, we have a specific compact just to avoid nonstationarity due to seasonal behavior, where the left bound is slightly larger than  $-2$ . The important assumption is the penalty rates. We have two penalties and two different coefficients. These will be needed to obtain the appropriate limits. The penalty rate for the consistency is true regardless of stationary or nonstationary behavior (i.e.  $\rho_0 < 0$ , or  $\rho_0 = 0$ ). Compared to stationary case of Knight and Fu (2000), we penalize at the same rate for the consistency proof. This changes in the case of limit law proofs. The following is a new result in the literature and shows that even there is nonstationarity bridge estimators are consistent. Furthermore, we do not know whether the data are stationary or nonstationary. The estimators converge to the true values in both cases.

**Theorem 3.** *Under Assumptions 1-5,*

$$\hat{\rho} \xrightarrow{P} \rho_0,$$

$$\hat{\zeta}_j - \zeta_{j,0} \xrightarrow{P} 0, \quad j = 1, \dots, p-1,$$

where these are true regardless of the nonstationarity where  $\rho_0 = 0$ , or stationary behavior of  $y_t$ , where  $\rho_0 < 0$ , ( $\rho_0 \in A - \{0\}$ ).

**Remark.** Since the behavior of the stationary and nonstationary random variables differ, the proof technique in Knight and Fu (2000) does not work for both cases. We extend and modify Huang, Horowitz and Ma (2008) consistency proof from stationary variable to a mixed set of nonstationary/stationary variables for our case. This is especially important when there is nonstationarity ( $\rho_0 = 0$ ).

The following result provides the limits in the cases of nonstationarity ( $\rho_0 = 0$ ) and stationarity ( $\rho_0 < 0$ ). In each case, true first differenced lag coefficients ( $\zeta_{10}, \dots, \zeta_{p-1,0}$ ) can be either zero or nonzero. Our Bridge estimators converge to zero with positive probability if the true coefficients are zero and have the normal limits when they are nonzero. This is true also for  $\rho$  coefficient as well. This is the one of the main results of the paper.

**Theorem 4.** Suppose  $0 < \gamma_1 < 1/2$ ,  $0 < \gamma_2 < 1$ , and  $\lambda_T/T^{\gamma_1} \rightarrow \lambda_0$ ,  $\lambda_0 \geq 0$ ,  $b_T/T^{\gamma_2/2} \rightarrow b_0$ ,  $b_0 \geq 0$ ,  $\theta = (\rho, \zeta)'$ ,  $\zeta = (\zeta_1, \dots, \zeta_{p-1})'$ ,  $\theta_0 = (\rho_0, \zeta_0)'$ , with Assumptions 1-4,

(i). If  $\rho_0 = 0$ , and then we can write  $\Delta y_t = \sum_{j=0}^{\infty} \psi_j e_{t-j}$ , (Chapter 17, Hamilton, 1994) assume  $\sum_{j=0}^{\infty} j|\psi_j| < \infty$ , then

$$\hat{u} = T\hat{\rho} \xrightarrow{d} \operatorname{argmin}_{u \in S} V_1(u),$$

where

$$V_1(u) = u^2 \iota^2 \int_0^1 W(r)^2 dr - u \sigma \iota (W(1)^2 - 1) + \lambda_0 |u|^{\gamma_1},$$

$$\iota = \sigma \sum_{j=0}^{\infty} \psi_j.$$

Also we have (by definition  $\zeta_{10} = \theta_{20}, \dots, \zeta_{p-1,0} = \theta_{p,0}$ )

$$\hat{l} = T^{1/2}(\hat{\zeta} - \zeta_0) \xrightarrow{d} \operatorname{argmin}_{l \in K} V_2(l),$$

$$V_2(l) = l' \Gamma_M l - 2l' N(0, \Gamma_M) + b_0 \sum_{j=1}^{p-1} |l_j|^{\gamma_2} 1_{\{\zeta_{j0}=0\}},$$

where  $\Gamma_M = EM_{t-1}M'_{t-1}$  nonsingular  $p-1 \times p-1$  matrix and  $M_{t-1} = (\Delta y_{t-1}, \dots, \Delta y_{t-(p-1)})'$ .

$S$  and  $K$  are compact subsets in  $R^1, R^{p-1}$  respectively in compliance with Assumptions 1-4.

Also  $\hat{u}, \hat{l}$  are asymptotically independent. We could have written the results in the following way

$$(\hat{u}, \hat{l}) \xrightarrow{d} \operatorname{argmin}_{u \in S, l \in K} V_1(u) + V_2(l).$$

(ii). If  $\rho_0 < 0$ , with  $\Gamma_W = Ew_{t-1}w'_{t-1}$ ,  $w_{t-1} = (y_{t-1}, M'_{t-1})'$ , assume  $\Gamma_W$  is nonsingular and bounded away from infinity, all the variables  $y_{t-1}, M_{t-1}$  are stationary, then for all of the coefficients we denote  $\hat{u}$  as the minimizer of the objective function

$$\hat{u} = T^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} \operatorname{argmin}_{u \in S} V(u),$$

$$V(u) = u' \Gamma_W u - 2u' N(0, \sigma^2 \Gamma_W) + b_0 \sum_{j=2}^p |u_j|^{\gamma_2} 1_{\{\theta_{j0}=0\}}.$$

Remarks.

1. Note that this theorem clearly shows we can simultaneously select the lags as well as distinguish between stationary and nonstationary behavior. In other words, compared to unit root testing, where the optimal lag length is determined first and then unit root test is conducted, Bridge estimation permits simultaneous selection of both. This approach is entirely new in the unit root literature.

2. Similar to Theorem 2, Remark 1, if we put  $\hat{u} = T\hat{\rho}$  when  $\rho_0 < 0$ , by using Theorem 4ii, we see that in that case  $\hat{u} \rightarrow -\infty$ , and thus nonstationary and stationary behavior can be differentiated.

3. Note that in (ii), since  $\rho_0 < 0$  the penalty term involves only  $\theta_{j0}$ ,  $j = 2, \dots, p$ , where  $\theta_{20} = \zeta_{10}, \dots, \theta_{p0} = \zeta_{p-1,0}$ . So it correctly selects  $\rho_0 < 0$ , and has a limiting normal distribution for its estimator.

4. In (i),  $\hat{\rho}$  and  $(\hat{\zeta} - \zeta_0)$  are asymptotically independent. We see that in that case  $\hat{\rho}$  converges to zero with positive probability and  $\hat{\zeta}$  estimates the zero coefficients asymptotically as zero, and nonzero ones with an asymptotically normal limiting distribution.

5. Specifically, the reason that we have  $0 < \gamma_1 < 1/2$  whereas  $0 < \gamma_2 < 1$  is the following. First, in the nonstationary case ( $\rho_0 = 0$ ), we need to obtain the estimators tending to 0 in probability. Thus we need a penalty term in the limit for  $\rho$  parameter. To achieve this,  $\lambda_T$  should grow at rate  $T_1^\gamma$ . On the other hand if  $\rho_0 < 0$ , we need to get nonzero estimates, so then in that case we do not need the penalty attached to  $\rho$ . Then to show that the penalty converges to zero, we need  $\lambda_T$  to grow at a rate smaller than  $T^{1/2}$ . Combining those two, we need  $0 < \gamma_1 < 1/2$ . For coefficients on dynamic differenced regressors we need  $0 < \gamma_2 < 1$ . This is the case since regardless of  $y_{t-1}$  they are stationary, so as in the Lasso literature for iid variables we have  $0 < \gamma_2 < 1$ .

6. Similarly, in Theorem 4i the penalty involves only  $l_1, \dots, l_{p-1}$ . These correspond to  $\zeta_{1,0}, \dots, \zeta_{p-1,0}$  (i.e.  $\hat{\zeta}_1 = \zeta_{1,0} + \frac{l_1}{T^{1/2}}$ ). These are the coefficients on dynamic differenced regressors  $(\Delta y_{t-1}, \dots, \Delta y_{t-(p-1)})$ . In Theorem 4ii, the penalty only involves again all the coefficients on dynamic differenced regressors. There is no penalty corresponding to the term with  $y_{t-1}$  since  $\rho_0 < 0$ . This shows that in Theorem 4i we have the penalty term attached in  $V_1(u)$  for  $\rho_0 = 0$ , but for the  $\rho_0 < 0$  case in Theorem 4ii, there is no penalty term.

A sharper result can be derived instead of Theorem 4i, when  $\rho_0 = 0$ . This is obtained exactly in the same way as in Corollary 1. Corollary 1 extended Theorem 2i. So the estimator of  $\rho$  when  $\rho_0 = 0$  with converges to zero with probability one. This is due

to the use of slightly “larger” penalization parameter. We choose  $\lambda_T = O(T^{\alpha_1})$  where  $\gamma_1 < \alpha_1 < 1/2$ .

**Corollary 2.** *Suppose  $0 < \gamma_1 < \alpha_1 < 1/2, \lambda_T/T^\alpha \rightarrow \lambda_0, \lambda_0 \geq 0$ , then with Assumptions 1-4,*

*(If  $\rho_0 = 0$ , and then  $\Delta y_t = \sum_{j=0}^{\infty} \psi_j e_{t-j}$ , (Chapter 17, Hamilton, 1994) assume  $\sum_{j=0}^{\infty} j|\psi_j| < \infty$ )*

$$\hat{u} = T\hat{\rho} \xrightarrow{wp1} 0.$$

Remark. The zero coefficients on lags (stationary variables) can also be obtained as zero with probability one like Corollary 2. They are asymptotically independent of the nonstationary variable coefficient estimators. This is seen in p.1361 of Knight and Fu (2000) for iid variables. We can extend this to stationary time series case (given Theorem 2) with ease with the new penalty rate  $b_T/T^{\alpha_2/2} \rightarrow b_0, b_0 \geq 0$ , where  $0 < \gamma_2 < \alpha_2 < 1$ .

## 4 Model With Time Trend

We now extend the previous model to one with an intercept and a time trend. (This is designated Case 4 in Chapter 17 of Hamilton (1994).) There are other ways of modeling the series with time trend; however, this approach is slightly richer and Bridge estimators show that there is no time trend in the first differenced model (nonstationary  $y_t$  case) when the model is misspecified with a time trend. We consider the estimation of the model

$$\Delta y_t = \rho y_{t-1} + \alpha + \delta t + \zeta_1 \Delta y_{t-1} + \dots + \zeta_{p-1} \Delta y_{t-(p-1)} + e_t, \quad (8)$$

where  $e_t$  is iid with  $E(e_t^2 | w_{t-1}) = \sigma^2$ , for all  $t = 1, \dots, T$ ,  $w_{t-1} = (y_{t-1}, 1, t, M'_{t-1})'$ . We can estimate the true values  $\rho_0, \delta_0, \alpha_0, \zeta_{1,0}, \dots, \zeta_{p-1,0}$  by minimizing

$$\sum_{t=p}^T (\Delta y_t - \rho y_{t-1} - \alpha - \delta t - \zeta_1 \Delta y_{t-1} - \dots - \zeta_{p-1} \Delta y_{t-(p-1)})^2 + \lambda_T |\rho|^{\gamma_1} + \nu_T |\alpha|^{\gamma_2} + \tau_T |\delta|^{\gamma_3} + b_T \sum_{j=1}^{p-1} |\zeta_j|^{\gamma_4}. \quad (9)$$

We can comment now on the form of the objective function under the nonstationary  $y_t, \rho_0 = 0$ . First of all as in Hamilton (1994) in the case of  $\rho_0 = 0$  we see that there is a quadratic time trend in  $y_t$ . To prevent that  $\delta_0 = 0$ . So a misspecified model is fit, and we show that Bridge estimator of  $\delta_0$  can converge to zero with positive probability. Second, in both in nonstationary and stationary cases we use transformed variables. This can be seen at the beginning of the proofs. In order to do this, we need to modify Assumptions 2, 3, and 5 to deal with the time trend.

**Assumption 2\*.**

(a). For the nonstationary case,  $\Gamma_M^* = EM_{t-1}^* M_{t-1}^{*'}$  is nonsingular and maximal eigenvalue of  $\Gamma_M^*$  is bounded away from infinity, where  $M_{t-1}^* = (u_{t-1}, \dots, u_{t-(p-1)})'$ ,  $u_t = \Delta y_t - \mu$ ,  $\mu = \alpha / (1 - \zeta_1 - \dots - \zeta_{p-1})$  (note that the denominator in  $\mu$  definition cannot be zero as explained in Chapter 17 of Hamilton (1994)).

(b). For the stationary case,  $\lim_{T \rightarrow \infty} \Sigma_T = \Sigma_M$  is nonsingular and the maximal eigenvalue is bounded away from infinity where

$$\Sigma_T = \begin{bmatrix} \frac{\sum_{t=2}^T y_{t-1}^{*2}}{T} & \frac{\sum_{t=2}^T y_{t-1}^*}{T} & \frac{\sum_{t=2}^T y_{t-1}^* t}{T^2} & \frac{\sum_{t=2}^T y_{t-1}^* M_{t-1}^{*'}}{T} \\ \frac{\sum_{t=2}^T y_{t-1}^*}{T} & 1 & \frac{\sum_{t=2}^T t}{T^2} & \frac{\sum_{t=2}^T M_{t-1}'}{T} \\ \frac{\sum_{t=2}^T t y_{t-1}^*}{T^2} & \frac{\sum_{t=2}^T t}{T^2} & \frac{\sum_{t=2}^T t^2}{T^3} & \frac{\sum_{t=2}^T t M_{t-1}^*}{T^2} \\ \frac{\sum_{t=2}^T M_{t-1}^* y_{t-1}^{*'}}{T} & \frac{\sum_{t=2}^T M_{t-1}^*}{T} & \frac{\sum_{t=2}^T M_{t-1}^* t}{T^2} & \frac{\sum_{t=2}^T M_{t-1}^* M_{t-1}^{*'}}{T} \end{bmatrix},$$

where  $y_{t-1}^* = y_{t-1} - \alpha - \delta(t-1)$ ,  $M_{t-1}^* = (\Delta y_{t-1}^*, \dots, \Delta y_{t-(p-1)}^*)'$ ,  $\Sigma_M$  is described in the proof of Theorem 5.

**Assumption 3\***. True value of  $\delta_0 = 0$  in the nonstationary case. Otherwise this is in a compact set in  $\mathbb{R}$ . Note that  $\alpha_0, \zeta_{1,0}, \dots, \zeta_{p-1,0}$  are in compact subsets of  $\mathbb{R}$ , and  $\mathbb{R}^{p-1}$  respectively.

**Assumption 5\***.  $\max(\lambda_T, \iota_T, \tau_T, b_T)/T \rightarrow 0$ .

The next Theorem provides the consistency of the estimators and extends Theorem 3 to the intercept and time trend case.

**Theorem 5**. Under Assumptions 1, 2\*, 3\*, 4 and 5\*,

$$\begin{aligned} \hat{\rho} &\xrightarrow{P} \rho_0, \\ \hat{\zeta}_j - \zeta_{j0} &\xrightarrow{P} 0, \quad j = 1, \dots, p-1, \\ \hat{\alpha} - \alpha_0 &\xrightarrow{P} 0, \\ \hat{\delta} &\xrightarrow{P} \delta_0. \end{aligned}$$

These are true regardless of nonstationarity ( $\rho_0 = 0$ ) or stationarity ( $\rho_0 < 0$ ) of  $y_t$ . Under nonstationarity, the proof shows that the estimators  $\hat{\rho}$ ,  $\hat{\delta}$  converge in probability to zero.

To derive the limits we rotate our variables as in Hamilton (1994) both in the nonstationary and stationary cases. Recall that the model that has to be estimated is

$$\Delta y_t = \rho y_{t-1} + \alpha + \delta t + \zeta_1 \Delta y_{t-1} + \dots + \zeta_{p-1} \Delta y_{t-1} + e_t. \quad (10)$$

For the nonstationary case (as in Hamilton (1994, p.498)), we use the following rotation, which can be obtained by simple addition and subtraction

$$\Delta y_t = \rho \xi_{t-1} + \mu^* + \delta^* t + \zeta_1 u_{t-1} + \dots + \zeta_{p-1} u_{t-(p-1)} + e_t, \quad (11)$$

where  $u_t = \Delta y_t - \mu$ ,  $\mu = \alpha/(1 - \zeta_1 - \dots - \zeta_{p-1})$ ,

$$\mu^* = (1 - \rho)\mu. \quad (12)$$

$$\xi_{t-1} = y_{t-1} - \mu(t-1),$$

and

$$\delta^* = \delta + \rho\mu. \quad (13)$$

Note that under  $\rho_0 = 0$ , and  $\delta_0 = 0$ , we obtain  $u_t = e_t/(1 - \zeta_1 L - \dots - \zeta_{p-1} L^{p-1})$ , and  $\xi_{t-1} = u_1 + u_2 + \dots + u_{t-1}$ . The important issue is whether Bridge estimators can show that  $\hat{\rho}, \hat{\delta}$  (or  $\hat{\delta}^*$ ) converge to zero in positive probability, and obtain the limits for nonzero coefficients as derived in Hamilton (1994). This will also indicate whether  $y_t$  is nonstationary.

In the stationary case, we need the following rotation.

$$\delta^* = \delta(1 + \rho),$$

$$\mu^* = \alpha(1 + \rho) - \delta(\rho - \zeta_1 - \dots - \zeta_{p-1}),$$

$$y_{t-1}^* = y_{t-1} - \alpha - \delta(t-1).$$

Equivalent to the model we are estimating, when  $\rho_0 < 0$ , we use the following

$$\Delta y_t = \rho y_{t-1}^* + \mu^* + \delta^* t + \zeta_1 \Delta y_{t-1}^* + \dots + \Delta y_{t-(p-1)}^* + e_t. \quad (14)$$

We now provide the limit theorems for both the nonstationary as well as the stationary case. This is a new result in the model selection literature and is one of the main results of the paper. We show that we can estimate time trend coefficient.

**Theorem 6.** *Suppose  $0 < \gamma_1 < 1/2, 0 < \gamma_2 < 1, 0 < \gamma_3 < 2/3, 0 < \gamma_4 < 1$ , and  $\lambda_T/T^{\gamma_1} \rightarrow \lambda_0, \lambda_0 \geq 0, \iota_T/T^{\gamma_2/2} \rightarrow \iota_0, \iota_0 \geq 0, \tau_T/T^{3\gamma_3/2} \rightarrow \tau_0, \tau_0 \geq 0, b_T/T^{\gamma_4/2} \rightarrow b_0, b_0 \geq 0$ .*

*Let  $\theta = (\rho, \mu^*, \delta^*, \zeta')'$ , and  $\theta_0 = (\rho_0, \mu_0^*, \delta_0^*, \zeta_0^*)'$ . Under Assumptions 1, 2\*, 3\*, 4,*

*(i). If  $\rho_0 = 0$ , with  $\delta_0 = 0$  (i.e.  $\delta_0^* = 0$ ), then*

$$\hat{v} = (T\hat{\rho}, T^{1/2}(\hat{\mu}^* - \mu_0^*), T^{3/2}\hat{\delta}^*)' \xrightarrow{d} \operatorname{argmin}_{v \in S} V_1(v),$$

where

$$V_1(v) = v' \Sigma_W v - 2v' h_{2\omega} + \lambda_0 |v_1|^{\gamma_1} + \iota_0 \mathbf{1}_{\{\mu_0^* = 0\}} |v_2|^{\gamma_2} + \tau_0 |v_3|^{\gamma_3},$$

and  $v = (v_1, v_2, v_3)$  correspond to respective cells in  $\hat{v}$ . The notation can be described as in (62)

$$\Sigma_W = \begin{bmatrix} \kappa^2 \int_0^1 W(r)^2 dr & \kappa \int_0^1 W(r) dr & \kappa \int_0^1 r W(r) dr \\ \kappa \int_0^1 W(r) dr & 1 & 1/2 \\ \kappa \int_0^1 r W(r) dr & 1/2 & 1/3 \end{bmatrix}.$$

$$h_{2\omega} = \begin{bmatrix} 1/2\sigma\kappa[W(1)^2 - 1] \\ \sigma W(1) \\ \sigma(W(1) - \int_0^1 W(r)dr) \end{bmatrix}.$$

For the stationary dynamic regressors we have

$$\hat{l} = \sqrt{T}(\hat{\zeta} - \zeta_0)' \xrightarrow{d} \operatorname{argmin}_{l \in K} V_2(l),$$

$$V_2(l) = l' \Gamma_M l - 2l' N(0, \Gamma_M) + b_0 \sum_{j=1}^{p-1} |l_j|^{\gamma_4} 1_{\{\zeta_{j0}=0\}},$$

where  $\Gamma_M$  is described in Theorem 4i.  $\hat{v}, \hat{l}$  are asymptotically independent.

(ii). If  $\rho_0 < 0$ , Let  $\hat{v}$  be

$$\hat{v} = (T^{1/2}\hat{\rho}, T^{1/2}(\hat{\mu}^* - \mu_0^*), T^{3/2}(\hat{\delta}^* - \delta_0^*), T^{1/2}(\hat{\zeta} - \zeta_0)')' \xrightarrow{d} \operatorname{argmin}_{v \in S_1} V(v),$$

where  $S_1$  is a compact subset of  $R^{p+2}$ , and

$$\begin{aligned} V(v) &= v' \Sigma_M v - 2\sigma^2 v' N(0, \Sigma_M) + \iota_0 |v_2|^{\gamma_2} 1_{\{\mu_0^*=0\}} \\ &+ \tau_0 |v_3|^{\gamma_3} 1_{\{\delta_0^*=0\}} + b_0 \sum_{j=4}^{p+2} |v_j|^{\gamma_4} 1_{\{\zeta_{j-3,0}=0\}}, \end{aligned}$$

where  $v_1, v_2, v_3, v_4$  represent the limits for  $\hat{v}_1 = T^{1/2}\hat{\rho}$ ,  $\hat{v}_2 = T^{1/2}(\hat{\mu}^* - \mu_0^*)$ ,  $\hat{v}_3 = T^{3/2}(\hat{\delta}^* - \delta_0^*)$ ,  $\hat{v}_4 = T^{1/2}(\hat{\zeta} - \zeta_0)'$  respectively.

Remarks.

1. The results here largely reflect Theorem 4. The main difference is that we can estimate the time trend coefficient as zero with positive probability in the nonstationary system. The penalty exponent is between 0 and 2/3, in contrast to the stationary regressors' exponents, which lie between 0 and 1.

2. The method can simultaneously select the optimal lag and differentiate the unit root case from the stationary case in the time trend case as well. Related to this point, if we substitute  $T\hat{\rho}$  when  $\rho_0 < 0$  by Theorem 6ii,  $\hat{u} \rightarrow -\infty$ , and so nonstationary and stationary behavior can be differentiated. Also  $\hat{\rho}$  converges to zero with positive probability when the nonstationarity holds (with  $\rho_0 = 0$ ), otherwise if  $\rho_0 < 0$ , by Theorem 6ii, it converges to the standard limit.

3. We can also obtain a better result for  $\rho_0$  when it is zero as in Corollary 2.

4. Note also that an alternative approach to modeling the time trend case is as follows.

Set

$$z_t = \alpha + \delta t + y_t,$$

$$y_t = \rho_1 y_{t-1} + \zeta_1 \Delta y_{t-1} + \cdots + \zeta_{p-1} \Delta y_{t-(p-1)} + e_t,$$

where the true  $\rho_{1,0} = 1$ . Then through transformations we are able to convert the system into an equation in  $z_t$  only. But since  $z_t$  carries the time trend, the proof demands an auxiliary regression with a detrended  $z_t$ , but the auxiliary regression unfortunately does not involve time trend because of the transformations, and hence does not correspond well with the main equation in  $z_t$  variable which contains a time trend. The details of the argument can be obtained from authors on demand, to save space this is not included here.

## 5 Monte Carlo

In this section we try to answer a basic question. How do Bridge Estimators compare with Dickey-Fuller GLS test (DFGLS from now on)? This is one of the most widely used tests and in the case of models with time trends has better power than Augmented Dickey-Fuller (ADF) test. The size and power results of DFGLS test can be seen in Ng and Perron (2001) with a new way of lag selection: Modified Akaike Information Criterion (MAIC). With MAIC lag selection, DFGLS test has very good size and power properties than ADF test with AIC, or BIC. In the case of no time trend and no intercept, the DFGLS is very similar to ADF test, so we consider ADF with AIC, BIC in that simple case.

Specifically, for Bridge Estimators we are interested in estimation of  $\rho_0$  coefficient. We want to see whether  $\rho_0 = 0$  or  $\rho_0 < 0$ . We report the percentage of correct model selection. Eventually we will relate that to size and power of the unit root tests below in this section.

There are two setups that we generate the data. The first one is used by Chapter 17, Case 4 of Hamilton (1994). This is favorable to Bridge. The second one is used by Elliot, Rothenberg, and Stock (1996), and this is favorable to DFGLS with MAIC.

The first one is given by

$$\Delta y_t = \rho_0 y_{t-1} + \alpha_0 + \delta_0 t + \zeta_{1,0} \Delta y_{t-1} + \dots + \zeta_{4,0} \Delta y_{t-4} + e_t, \quad (15)$$

where  $e_t$  is iid and  $N(0,1)$ . Under the null hypotheses of  $\rho_0 = 0$  we set  $\delta_0 = 0$ , otherwise  $\delta_0 \neq 0$ .

The second setup is as follows:

$$y_t = \alpha_0 + \delta_0 t + u_t, \quad (16)$$

and

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \phi_3 u_{t-3} + \phi_4 u_{t-4} + e_t, \quad (17)$$

In setup 1, described by equation (15), we have the following design.

Design 1: (when  $\rho_0 = 0$ , we set  $\delta_0 = 0$ )

$$\Delta y_t = \rho_0 y_{t-1} + 0.5 + 0.1t - 0.4 \Delta y_{t-1} + 0.7 \Delta y_{t-3} + e_t, \quad (18)$$

so  $\zeta_2 = \zeta_4 = 0$ , this is the model with holes and 3 lags,  $\rho_0$  can take values of 0, -0.05, -0.1, -0.2.

Design 2: (when  $\rho_0 = 0$ , we set  $\delta_0 = 0$ )

$$\Delta y_t = \rho_0 y_{t-1} + 0.5 + 0.1t - 0.4\Delta y_{t-1} - 0.2\Delta y_{t-2} + e_t, \quad (19)$$

so  $\zeta_3 = \zeta_4 = 0$ , this is the model with 2 lags, no holes,  $\rho_0$  can take values of 0, -0.05, -0.1, -0.2.

Design 3: (when  $\rho_0 = 0$ , we set  $\delta_0 = 0$ )

$$\Delta y_t = \rho_0 y_{t-1} + 0.5 + 0.1t - 0.65\Delta y_{t-1} + e_t, \quad (20)$$

so  $\zeta_2 = \zeta_3 = \zeta_4 = 0$ , this is the model with one lag, no holes,  $\rho_0$  can take the values of 0, -0.05, -0.1, -0.2.

The second setup is used by Elliot, Rothenberg, and Stock (1996). We have the following design.

Design 1: For the size part

$$y_t = 0.5 + 0.1t + u_t, \quad (21)$$

$$u_t = 0.6u_{t-1} + 0.4u_{t-2} + 0.7u_{t-3} - 0.7u_{t-4} + e_t, \quad (22)$$

For the power, the coefficient on  $u_{t-1}$  is changed to 0.55, 0.50, 0.40.

Design 2: For the size part, same as (21) but instead of (22) we have

$$u_t = 0.6u_{t-1} + 0.2u_{t-2} + 0.2u_{t-3} + e_t, \quad (23)$$

For the power exercise, the coefficient on  $u_{t-1}$  in (23) is 0.55, 0.50, 0.40.

Design 3: For the size part, same as (21) but instead of (22) we have

$$u_t = 0.35u_{t-1} + 0.65u_{t-2} + e_t, \quad (24)$$

where for the power the coefficient on  $u_{t-1}$  is 0.30, 0.25, 0.15.

In setup 1, the researcher starts with 4 lag structure with no holes, and then uses Bridge estimation. But the true data generating processes are Designs 1-3. When  $\rho_0 = 0$  this is nonstationary  $y_{t-1}$  and the other values of  $\rho_0 = -0.05, -0.1, -0.2$  correspond to stationary  $y_{t-1}$ . For Bridge estimators, we set  $\gamma_1 = 1/4, \gamma_2 = 1/2, \gamma_3 = 1/3, \gamma_4 = 1/2$ . The choice of  $\lambda_T, \nu_T, \tau_T, b_T$  are done in two different ways. We try cross-validation as in Fan and Li (2001) and the universal  $\lambda_T, \nu_T, \tau_T, b_T$  choices as suggested by Donoho and Johnstone (1994). Donoho and Johnstone (1994) approach works better in simulations so we set  $\lambda_T = \sqrt{2\log 2} T^{1/3}, \nu_T = \tau_T = b_T = \sqrt{2\log 4} T^{1/3}$ . These rates are in line with Corollary

2, and used in Caner (2008) in GMM-Lasso context for stationary but endogenous variables. Since the objective function is nondifferentiable we use local quadratic approximation as suggested by Fan and Li (2001, 2002) and used in the literature. The choice for thresholding rule to set zero coefficients to zero is when  $\hat{\rho} < \frac{-6}{10^2}$ , and  $|\hat{\zeta}_j| < \frac{6}{10^2}$ ,  $j = 1, \dots, p-1$ . These are similar to thresholding rules in Huang, Horowitz and Ma (2008), but since we are selecting between stationary versus nonstationary models the magnitude of the threshold is larger. Minor changes in the thresholding rule do not affect our results in simulations.

In setup 2, we start with 5 maximum lags. The rest is the same as in setup 1 for Bridge.

Now we describe how we set up DFGLS test with Modified Akaike Information Criterion (MAIC) of Ng and Perron (2001). Before lag selection we need the following

1.  $y^\alpha = (y_1, y_2 - \bar{\alpha}y_1, \dots, y_T - \bar{\alpha}y_{T-1})$ , where  $\bar{\alpha} = -13.5$  for the time trend,  $\bar{\alpha} = -7$  for the intercept case. The same is true for  $z_t = (1, t)$ .  $z^\alpha = (z_1, z_2 - \bar{\alpha}z_1, \dots, z_T - \bar{\alpha}z_{T-1})$ .
2. Then run least squares regression  $y^\alpha$  on  $z^\alpha$  to have the coefficients  $\hat{\phi}$ .
3. Use  $y_t^\alpha - \hat{\phi}z_t^\alpha = \tilde{y}_t$ .
4. Then run least squares for the following

$$\Delta\tilde{y}_t = \rho_0\tilde{y}_{t-1} + \sum_{j=1}^p \zeta_j\Delta\tilde{y}_{t-j} + e_{tp}, \quad (25)$$

The first step is the optimal lag selection. We use equation (12) of Ng and Perron (2001) for each  $p = 1, 2, \dots, p_{max}$

$$MAIC(p) = \ln(\hat{\sigma}_p^2) + 2\frac{g(p) + p}{T - p_{max}},$$

where

$$g(p) = (\hat{\sigma}_p^2)^{-1}\hat{\rho}_0^2 \sum_{t=p_{max}+1}^T \tilde{y}_{t-1}^2,$$

$$\hat{\sigma}_p^2 = (T - p_{max})^{-1} \sum_{t=p_{max}+1}^T \hat{e}_{tp}^2,$$

where  $\hat{\rho}$ , and  $\hat{e}_{tp}$  are obtained for LS in equation (25). The lag  $p$  that minimizes  $MAIC(p)$  will be used in the next step of DFGLS. The DFGLS test is basically a t-test for  $\rho_0 = 0$  in equation (25) given optimal  $p$ . The limit critical values from the distribution of the test statistic is given by Ng and Perron (2001).

The exercise we run first reports percentage of correct model selection via Bridge on  $\rho_0$  coefficient, whether that is  $\rho_0 = 0$  or  $\rho_0 < 0$ . We take  $T = 100$ , and use 10000 iterations. Tables 1a-b report the size. Since Bridge is only estimating, correct percentage of zero coefficients on  $\rho_0$ , we take the probability of incorrectly estimating zero coefficients as the "size" of Bridge. We should remind ourselves that Bridge is a model selection procedure and does also estimation simultaneously. So there is no testing involved in Bridge. This is

the reason that we put "size" in quotation marks. Indeed with "size", we are just reporting the frequency of wrong model selection in percentage terms. For "power", this will be percentage of correct models selected. Also an important reminder is that the penalty parameters  $\lambda_T = \sqrt{2\log 2} T^{1/3}$ ,  $\iota_T = \tau_T = b_T = \sqrt{2\log 4} T^{1/3}$  are not estimated in limits of Bridge estimators. These are sequences with special convergence rates. This is related to model selection.

Both in setups 1 and 2, Bridge has better size than DFGLS. Especially in the more complex, Designs such as 1 and 2, we see that DFGLS has size of 21-38% at 5% level in the case of time trend. We also see some power declines in DFGLS when we are away from nonstationarity, this is due to being in a Design (Design 1) that is far away from the model of DFGLS tests. In the same cases Bridge has 5-8% size.

The next exercise deals with the power issue. These are in Tables 2-3. This is done in the following way. For Bridge, we first find the threshold values that will give the actual size that is found in Table 1a-b for DFGLS case. Then by using these threshold values we look at the percentage of nonzero coefficients for  $\rho_0$ . This is compared with the power of DFGLS test.

We have also done the following exercise, but did not report. We calculate the size adjusted power of DFGLS test (at 95% critical values from the empirical distribution) and compare it with the correct percentage of nonzero estimates of  $\rho_0$  when the Bridge has 5% "size" as described above. This is very similar to the results in Tables 2-3 and are not reported. We see that in our setup (setup 1), the power of DFGLS is less than Bridge (Tables 2a-2b). On the other hand, when we follow setup 2, except from Design 1, DFGLS does better.

We also see that when the coefficient is near zero ( $\rho_0 = -0.05$ ), Bridge is especially good in terms of power in time trend case, we see that in Designs 1-3 in Table 2a, Bridge has 27-74% power, compared with 0-43% power of DFGLS.

In the case of no time trend and no intercept, we compare Bridge with ADF test. For the ADF unit root tests we use AIC or SIC to pick up the optimal lags. Since there are no time trend and no intercept, the setup1 and setup 2 structure is the same. True data generating processes are Designs 1-3 without the time trend and the intercept. When  $\rho_0 = 0$ , then there is nonstationarity in the data, and with  $\rho_0 = -0.05, -0.10, -0.20$ , there is stationarity. For Bridge we choose  $\gamma_1 = 1/4, \gamma_2 = 1/2$ , the Bridge threshold value is the same as in the time trend/intercept case, as well as the  $\lambda_T, b_T$  penalty terms.

We minimize the following over  $p= 2, 3, 4, 5$

$$T\log\hat{s}_p^2 + C_T,$$

Table 1a: Size, Time Trend Case

	Setup 1			Setup 2		
	Design 1	Design 2	Design 3	Design 1	Design 2	Design 3
Bridge	5.71	5.22	1.66	4.63	7.55	2.53
DFGLS	38.41	21.15	6.99	37.45	26.03	6.96

Note: The size of DFGLS test is at 5% nominal level. For Bridge, we use threshold value of  $-6/10^2$ . The size figures for Bridge are 1- probability of correct model selection on  $\rho_0$ . Setup 1, Designs 1-3, represent the equations (18)-(20). Setup 2, Designs 1-3 represent equations (21)-(24).

where

$$\hat{s}_p^2 = T^{-1} \sum_{t=1}^T (\Delta y_t - \rho y_{t-1} - \zeta_1 \Delta y_{t-1} - \dots - \zeta_{p-1} \Delta y_{t-(p-1)})^2,$$

$C_T = 2(p+2)$  for AIC  $C_T = (p+2)\log T$  for SIC.

Given the optimal p, the we run ADF test for unit roots, where  $H_0 : \rho_0 = 0$  and this is a simple t-ratio

$$t = \frac{\hat{\rho}_p}{\hat{s}_p (M_p^{11})^{1/2}},$$

where  $\hat{\rho}_p$  is the least squares estimator for  $\rho_0$  given the lag p that is chosen by AIC or SIC, and  $M_p = \sum_{t=1}^T (y_{t-1}, \Delta y_{t-1}, \Delta y_{t-(p-1)})' (y_{t-1}, \Delta y_{t-1}, \Delta y_{t-(p-1)})$ , where  $M_p^{11}$  is the (1,1) element of  $M_p^{-1}$ . The limit is given in Chapter 17 of Hamilton (1994) and tabulated in Table B.6. This is known as Dickey-Fuller distribution. The results are in Tables 4-5.

AIC is usually known to overfit the models. In our simulations both AIC and SIC pick 4 lags at each design, clearly overfitting the models. Bridge find the places of zero parameters in Design 1 almost near 100% and the same for other designs. This again illustrates superior model selection properties of Bridge Estimators.

The "size" of the Bridge is the one minus the probability of correctly selecting the zero of  $\rho_0$ . With true  $\rho_0 = 0$  in Design 1, the "size" of the Bridge estimator is 5.16%. In Design 2, and 3 the rates are 5.72% and 3.30% respectively (a threshold of  $-6/10^2$  is used as in the time trend case). The size of the ADF test is very good at 5.00%, 5.21%, 5.13% levels at nominal 5% level. In terms of the power, the bridge looks at probably of correct model selection (in this case selecting  $\rho_0 < 0$  when this is the case). To make things comparable we use a threshold of  $-5/10^2$  for Design 3 to get the same size as Design 3 in ADF. The power results clearly show that Bridge dominates ADF in terms of power except in Design 1 when the coefficient is -0.05.

Table 1b: Size, Intercept Case

	Setup 2		
	Design 1	Design 2	Design 3
Bridge	2.56	4.74	1.98
DFGLS	30.41	15.64	7.00

Note: The size of DFGLS test is at 5% nominal level. For Bridge, we use threshold value of  $-6/10^2$ . The size figures for Bridge are 1- probability of correct model selection on  $\rho_0$ . (15) is used for Bridge/testing unit roots in Setup 1. Setup 2, Designs 1-3 represent equations (21)-(24). (25) is used for Bridge/testing unit roots in Setup 2.

Table 2a: Power, Time Trend Case, Setup 1

	Design 1			Design 2			Design 3		
$\rho_0 =$	-0.05	-0.1	-0.2	-0.05	-0.1	-0.2	-0.05	-0.10	-0.20
Bridge	74.40	65.15	99.80	36.24	32.92	44.83	26.50	18.74	13.20
DFGLS	43.63	28.64	40.20	0.00	17.11	75.83	0.00	3.71	37.91

Note: The power of DFGLS test is at 5% nominal level. For Bridge, we use threshold values that will generate the size of DFGLS tests in Table 1a, Setup 1 ( $-4/10^2$ ,  $-5/10^2$ ,  $-5.2/10^2$  respectively for Designs 1-3). The power figures for Bridge are probability of correct model selection on  $\rho_0$ . Setup 1, Designs 1-3, represent the equations (18)-(20).

## 6 Empirical Results

In this section we consider a data set with our technique, and compare with ADF and DFGLS unit root tests. In the past years, unemployment rate in US is tested with different unit root tests before year 2000, and they were unable to reject the null of unit roots, for this issue see Caner and Hansen (2001). The unemployment rate that we analyze is from Bureau of Labor Statistics, dating from January 1969 to July 2009, monthly data (487 observations). The same data is also available at Economagic. The unemployment rates are for 16 years and older persons. In our model we set the maximum number of lags possible as either 24 or 36. Since there is no time trend in the unemployment rate, we use the model with intercept. This is equation (8) with  $\alpha \neq 0$ , and  $\delta = 0$  there. We analyze specifically ADF test with either AIC, or SIC, and DFGLS with MAIC lag selection criterion. These are described in detail in simulation section. In Table 6, we report the results from the unit root tests. For ADF we report AIC results, SIC result is the same. It is clear that at 5% level, none of them were able to reject the null. Since unemployment rate is a bounded variable, the expectation is that it should be stationary.

Next, we employ Bridge estimator. This is the model with intercept in equation (8)

Table 2b: Power, Intercept Case, Setup 1

	Design 1			Design 2			Design 3		
$\rho_0 =$	-0.05	-0.1	-0.2	-0.05	-0.1	-0.2	-0.05	-0.10	-0.20
Bridge	54.15	80.00	96.27	33.96	57.30	83.64	15.89	31.95	54.87
DFGLS	35.27	28.93	5.05	13.06	22.95	60.82	3.74	6.42	25.28

Note: The power of DFGLS test is at 5% nominal level. For Bridge, we use threshold values that will generate the size of DFGLS tests in Table 1a, Setup 1 ( $-4/10^2$ ,  $-5/10^2$ ,  $-5.2/10^2$  respectively for Designs 1-3). The power figures for Bridge are probability of correct model selection on  $\rho_0$ . Setup 1, Designs 1-3, represent the equations (18)-(20).

Table 3a: Power, Time Trend Case, Setup 2

	Design 1			Design 2			Design 3		
$\rho_0 =$	-0.05	-0.1	-0.2	-0.05	-0.1	-0.2	-0.05	-0.10	-0.20
Bridge	56.15	79.84	96.95	33.47	48.22	77.27	10.96	16.95	36.92
DFGLS	44.99	27.88	7.21	29.66	46.02	79.33	9.55	16.99	40.70

Note: The power of DFGLS test is at 5% nominal level. For Bridge, we use threshold values that will generate the size of DFGLS tests in Table 1a, Setup 2 ( $-3.7/10^2$ ,  $-5/10^2$ ,  $-5.4/10^2$  respectively for Designs 1-3). The power figures for Bridge are probability of correct model selection on  $\rho_0$ . Setup 2, Designs 1-3, represent the equations (22)-(24).

Table 3b: Power, Intercept Case, Setup 2

	Design 1			Design 2			Design 3		
$\rho_0 =$	-0.05	-0.1	-0.2	-0.05	-0.1	-0.2	-0.05	-0.10	-0.20
Bridge	60.87	86.46	98.44	22.93	37.96	73.54	9.29	14.80	39.80
DFGLS	61.79	34.15	23.60	43.95	70.31	93.95	22.46	40.87	65.13

Note: The power of DFGLS test is at 5% nominal level. For Bridge, we use threshold values that will generate the size of DFGLS tests in Table 1b, Setup 2 ( $-3/10^2$ ,  $-4.8/10^2$ ,  $-5/10^2$  respectively for Designs 1-3). The power figures for Bridge are probability of correct model selection on  $\rho_0$ . Setup 2, Designs 1-3, represent the equations (22)-(24).

Table 4: Size, No Intercept, No Time Trend Case

	Design 1	Design 2	Design 3
Bridge	5.16	5.72	3.30
ADF	5.00	5.21	5.13

Note: The size of ADF test is at 5% nominal level. For Bridge, we use threshold value of  $-6/10^2$ . The size figures for Bridge are 1- probability of correct model selection on  $\rho_0$ .

Table 5: Power, No Time Trend/Intercept Case

	Design 1			Design 2			Design 3		
$\rho_0 =$	-0.05	-0.1	-0.2	-0.05	-0.1	-0.2	-0.05	-0.10	-0.20
Bridge	40.16	83.66	99.52	21.54	50.66	91.89	19.66	44.71	86.32
ADF	41.66	76.60	96.79	17.15	37.00	74.82	16.62	35.84	77.76

Note: The power of ADF test is at 5% nominal level. For Bridge, we use threshold values that will generate the size of ADF tests ( $-6/10^2$ ,  $-6/10^2$ ,  $-5/10^2$  respectively for Designs 1-3). The power figures for Bridge are probability of correct model selection on  $\rho_0$ .

Table 6: Unit Root Tests on Unemployment Rate,  $H_0 : \rho = 0$ 

Tests	$p_{max} = 24$	$p_{max} = 36$
DFGLS	-2.76	-2.43
ADF	-1.52	-1.43

Note: The 5% critical values are -2.86 for ADF test, and -1.98 from DFGLS with MAIC from Table 1 in Ng, Perron (2001).  $p_{max}$  represents the maximum number of lags fitted to the model.

with  $\delta = 0$ . The tuning parameter is chosen as in the simulation section. This is also true for truncation setup. So if the Bridge estimator  $\hat{\rho}$  is less than -0.06, we set that coefficient to zero, and the unemployment rate is a nonstationary variable. If  $\hat{\rho}$  is less than -0.06, then the unemployment rate itself is stationary. When we run Bridge estimator we find that  $\hat{\rho} = -0.13$  when  $p_{max} = 24$ . With  $p_{max} = 36$ , then  $\hat{\rho} = -0.18$ . So in both cases  $\hat{\rho} < -0.06$ , and the unemployment rate is stationary. This shows that Bridge can give sensible results when unit root tests cannot. And an important issue is this is pretty robust to truncation value of -0.06. With slightly smaller or larger values the stationarity of the unemployment rate does not change.

## 7 Conclusion

In this paper we show that Bridge estimators can differentiate between unit roots and stationary variables and select the optimal lag simultaneously. This is unlike the existing methods where first lag length is selected according to AIC, BIC, and then unit root is tested.

In future research, one possible extension is to allow lags to converge to infinity. This is shown in Huang, Horowitz, and Ma (2008) in a stationary regressor environment. We believe that this extension is possible in our case as well. The next project is to apply these methods to forecasting by using the averaging estimator introduced by Hansen (2007, 2008).

We can select the weights according to Mallows criterion, and then instead of pretesting via unit root tests, we use Bridge estimator in weighted forecasts.

## APPENDIX

**Proof of Theorem 1i.** First set  $A = [-2 - \epsilon, 0]$  where  $\epsilon$  is a small positive number, this is done to prevent nonstationarities due to other frequencies (e.g. seasonal unit roots). This is a compact set in the real line. Uniformly over  $A$  we have to prove the following

$$Z_{T1}(\rho) \xrightarrow{d} Z_1(\rho), \quad (26)$$

where  $Z_1(\rho) = \rho^2 \sigma^2 \int_0^1 W(r)^2 dr + \lambda_0 |\rho|^\gamma$ .

To derive (26), see that by (1) at  $\rho_0 = 0$ ,

$$\begin{aligned} \frac{1}{T^2} \sum_{t=1}^T (\Delta y_t - \rho y_{t-1}^2) &= \frac{1}{T^2} \sum_{t=1}^T (e_t - \rho y_{t-1})^2 \\ &= \frac{1}{T^2} \sum_{t=1}^T e_t^2 - 2 \frac{1}{T^2} \rho \sum_{t=1}^T e_t y_{t-1} \\ &\quad + \frac{1}{T^2} \rho^2 \sum_{t=1}^T y_{t-1}^2. \end{aligned}$$

The first term on the right hand side

$$\frac{1}{T^2} \sum_{t=1}^T e_t^2 \xrightarrow{p} 0,$$

by law of large numbers. Then by Proposition 17.1 of Hamilton (1994), we have

$$\frac{1}{T^2} \rho \sum_{t=1}^T e_t y_{t-1} \xrightarrow{p} 0,$$

$$\frac{1}{T^2} \rho^2 \sum_{t=1}^T y_{t-1}^2 \xrightarrow{d} \rho^2 \sigma^2 \int_0^1 W(r)^2 dr.$$

Combine the above results to have

$$\frac{1}{T^2} \sum_{t=1}^T (e_t - \rho y_{t-1})^2 \xrightarrow{d} \rho^2 \sigma^2 \int_0^1 W(r)^2 dr. \quad (27)$$

Then

$$(\lambda_T / T^2) |\rho|^\gamma \rightarrow \lambda_0 |\rho|^\gamma.$$

So combine the last result with (27) to have (26). Next, we need to show that

$$\hat{\rho} = O_p(1). \quad (28)$$

To prove that, see

$$Z_{T1}(\rho) \geq \frac{1}{T^2} \sum_{t=1}^T (\Delta y_t - \rho y_{t-1})^2 = Z_{T0}(\rho).$$

Since by p. 486-488 of Hamilton (1994),  $\operatorname{argmin} Z_{T_0}(\rho) = O_p(1)$ , it follows that  $\hat{\rho} = \operatorname{argmin} Z_{T_1}(\rho) = O_p(1)$ . Then by (26)(28), uniformly over  $\rho \in A$ ,

$$\operatorname{argmin} Z_{T_1}(\rho) \xrightarrow{p} \operatorname{argmin} Z_1(\rho).$$

**Q.E.D**

**Proof of Theorem 1ii.** Now we analyze the stationary case. See that if  $\rho_0 < 0$ ,

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T (\Delta y_t - \rho y_{t-1})^2 &= \frac{1}{T} \sum_{t=1}^T [e_t - (\rho - \rho_0) y_{t-1}]^2 \\ &= \frac{1}{T} \sum_{t=1}^T e_t^2 + (\rho - \rho_0)^2 \frac{1}{T} \sum_{t=1}^T y_{t-1}^2 - 2(\rho - \rho_0) \frac{1}{T} \sum_{t=1}^T e_t y_{t-1}. \end{aligned}$$

Then by Law of large numbers,

$$\begin{aligned} T^{-1} \sum_{t=1}^T e_t^2 &\xrightarrow{p} \sigma^2, \\ T^{-1} \sum_{t=1}^T y_{t-1}^2 &\xrightarrow{p} \Gamma_0, \end{aligned}$$

Also by Proposition 17.3, b and c of Hamilton (1994)

$$T^{-1} \sum_{t=1}^T e_t y_{t-1} \xrightarrow{p} 0.$$

So

$$T^{-1} \sum_{t=1}^T (\Delta y_t - \rho y_{t-1})^2 \xrightarrow{p} \sigma^2 + (\rho - \rho_0)^2 \Gamma_0.$$

Then uniformly over  $\rho \in A$ , if  $\lambda_T = O(T)$ ,

$$Z_{T_2}(\rho) \xrightarrow{p} \sigma^2 + (\rho - \rho_0)^2 + \lambda_0 |\rho|^\gamma \equiv Z_2(\rho). \quad (29)$$

Clearly,

$$\hat{\rho} = O_p(1), \quad (30)$$

following the same logic as in the proof of Theorem 1i, which is due to the simple least squares estimator being stochastically bounded in the stationary case. These prove the desired result. **Q.E.D.**

**Proof of Theorem 1iii.** This is analyzed for both cases. First start with the case of nonstationary  $y_t$ . There by (3), and  $\lambda_T = o(T)$  we have the consistency. This can be seen easily since in that scenario,  $\lambda_T/T^2 \rightarrow 0$ , and the penalty limit term converges to zero as well on the right hand side of (3). So  $Z_1(\rho) = \rho^2 \sigma^2 \int W(r)^2 dr$ . Note that  $\rho = 0$  is the unique minimum of that function. Via argmax continuous mapping theorem (Theorem 3.2.2) in van der Vaart and Wellner (1996) we have the desired result.

In the stationary case, the consistency holds by  $\lambda_T = o(T)$ , then by (29)(30). This is easy to see since the limit is  $Z_2(\rho) = \sigma^2 + (\rho - \rho_0)^2 \Gamma_0$ , due to penalty limit term converging to zero by  $\lambda_T = o(T)$ . But this expression is uniquely minimized at  $\rho = \rho_0$ , so we have the desired result due to Theorem 3.2.2 of van der Vaart and Wellner (1996). **Q.E.D.**

To derive the limit theory we need to explain the reparametrized objective function. We write the optimized function,  $0 < \gamma < 1/2$ ,

$$\sum_{t=1}^T (\Delta y_t - \hat{\rho} y_{t-1})^2 + \lambda_T |\hat{\rho}|^\gamma.$$

Note that regardless of the stationary or nonstationary cases,  $\hat{\rho}$  minimizes the above function. We define the following function that will be used in the derivation of the limit for the estimator. The derivation of the function from the objective function is explained immediately below.

$$V_T(u) = \left[ \sum_{t=1}^T \left( e_t - \frac{u y_{t-1} 1_{\{\rho_0 < 0\}}}{T^{1/2}} - \frac{u y_{t-1} 1_{\{\rho_0 = 0\}}}{T} \right)^2 - e_t^2 \right] + \lambda_T |\rho_0 + \frac{u 1_{\{\rho_0 < 0\}}}{T^{1/2}} + \frac{u 1_{\{\rho_0 = 0\}}}{T}|^\gamma - \lambda_T |\rho_0|^\gamma.$$

Note that  $V_T(u)$  is minimized at  $u = T^{1/2}(\hat{\rho} - \rho_0)$  when  $\rho_0 < 0$ , and  $V_T(u)$  is minimized at  $u = T\hat{\rho}$  when  $\rho_0 = 0$ .

Specifically for  $\rho_0 = 0$  case

$$V_T(u) = \left[ \sum_{t=2}^T \left( e_t - \frac{u y_{t-1}}{T} \right)^2 - e_t^2 \right] + \lambda_T |\rho_0 + \frac{u}{T}|^\gamma.$$

For  $\rho_0 < 0$ ,

$$V_T(u) = \left[ \sum_{t=1}^T \left( e_t - \frac{u y_{t-1}}{\sqrt{T}} \right)^2 - e_t^2 \right] + \lambda_T |\rho_0 + \frac{u}{T^{1/2}}|^\gamma - \lambda_T |\rho_0|^\gamma.$$

We now provide how  $V_T(u)$  expression is derived in the case for  $\rho_0 < 0$ . First

$$\hat{\rho} = \operatorname{argmin}_\rho \sum_{t=2}^T (\Delta y_t - \rho y_{t-1})^2 + \lambda_T |\rho|^\gamma.$$

But it is true that

$$\hat{\rho} = \operatorname{argmin}_\rho \left[ \sum_{t=2}^T (\Delta y_t - \rho y_{t-1})^2 + \lambda_T |\rho|^\gamma - \left( \sum_{t=2}^T (\Delta y_t - \rho_0 y_{t-1})^2 + \lambda_T |\rho_0|^\gamma \right) \right]. \quad (31)$$

Then rewrite (31) via

$$\begin{aligned} \Delta y_t - \rho y_{t-1} &= (\Delta y_t - \rho_0 y_{t-1}) - (\rho - \rho_0) y_{t-1} \\ &= e_t - (\rho - \rho_0) y_{t-1} \\ &= e_t - T^{1/2} (\rho - \rho_0) y_{t-1} / T^{1/2} \\ &= (e_t - u y_{t-1} / T^{1/2}), \end{aligned}$$

where  $u = T^{1/2}(\rho - \rho_0)$ . Then use the above equality in (31) to have

$$\hat{u} = \arg \min_u \left[ \sum_{t=2}^T (e_t - uy_{t-1}/T^{1/2})^2 - e_t^2 \right] + \lambda_T |\rho_0 + \frac{u}{T^{1/2}}|^2 - \lambda_T |\rho_0|^2. \quad (32)$$

Similarly we can derive the  $V_T(u)$  expression above for  $\rho_0 = 0$ .

**Proof of Theorem 2.** For parts i, and ii, the following few derivations simplify the proof. After following the three equations below we provide the proofs for i and ii. First note that since  $1_{\{\rho_0=0\}}1_{\{\rho_0<0\}} = 0$ , and  $\Gamma_0 = Ey_{t-1}^2$  when  $\rho_0 < 0$ .

$$\begin{aligned} \sum_{t=1}^T \left( e_t - \frac{uy_{t-1}1_{\{\rho_0<0\}}}{T^{1/2}} - \frac{uy_{t-1}1_{\{\rho_0=0\}}}{T} \right)^2 & - e_t^2 = [-2u \left( \frac{\sum_{t=1}^T e_t y_{t-1}}{T^{1/2}} \right) 1_{\{\rho_0<0\}}] \\ & - [2u \left( \frac{\sum_{t=1}^T e_t y_{t-1}}{T} \right) 1_{\{\rho_0=0\}}] \\ & + [u^2 \left( \frac{\sum_{t=1}^T y_{t-1}^2}{T} \right) 1_{\{\rho_0<0\}}] + [u^2 \left( \frac{\sum_{t=1}^T y_{t-1}^2}{T^2} \right) 1_{\{\rho_0=0\}}] \\ & \xrightarrow{d} [-2uL + u^2\Gamma_0] 1_{\{\rho_0<0\}} \\ & + [-2u\sigma^2 \int_0^1 W(r)dW(r) + u^2\sigma^2 \int_0^1 W(r)^2 dr] 1_{\{\rho_0=0\}} \end{aligned} \quad (33)$$

The limits are obtained uniformly over  $u$  in a compact set  $K$ . In the proof of (33), the case for  $\rho_0 = 0$  is derived by Proposition 17.1 in Hamilton (1994), and the stationary case ( $\rho_0 < 0$ ) is obtained by Proposition 17.3 b, and c in Hamilton (1994). Then

Then if  $\rho_0 = 0$ , uniformly over  $u$  in a compact set, since  $\lambda_T/T^\gamma \rightarrow \lambda_0$

$$\lambda_T \left| \frac{u}{T} \right|^\gamma \rightarrow \lambda_0 |u|^\gamma. \quad (34)$$

For  $\rho_0 < 0$  case, uniformly over  $u$  in a compact set, since  $\lambda_T = O(T^\gamma)$ ,  $\lambda_T/T^{1/2} \rightarrow 0$ ,  $0 < \gamma < 1/2$

$$\lambda_T \left( \left| \rho_0 + \frac{u}{T^{1/2}} \right|^\gamma - |\rho_0|^\gamma \right) \rightarrow 0. \quad (35)$$

Now we analyze the specific cases.

**Proof of Theorem 2i.** Combine (33)(34), for  $\rho_0 = 0$ , uniformly over  $u$

$$V_T(u) \xrightarrow{d} -2u \int_0^1 W(r)dW(r) + u^2 \int_0^1 W(r)^2 dr + \lambda_0 |u|^\gamma = V_1(u). \quad (36)$$

Then we need to show  $\text{argmin} V_T(u) = O_p(1)$  when  $\rho_0 = 0$ . In that respect, with  $\lambda_T/T^\gamma \rightarrow \lambda_0 \geq 0$ , (in the case of  $\rho_0 = 0$ )

$$\begin{aligned} V_T(u) & \geq \left[ \sum_{t=1}^T \left( e_t - \frac{uy_{t-1}}{T} \right)^2 - e_t^2 \right] - \frac{\lambda_T}{T^\gamma} |u|^\gamma \\ & \geq \left[ \sum_{t=1}^T \left( e_t - \frac{uy_{t-1}}{T} \right)^2 - e_t^2 \right] - (\lambda_0 + \delta) |u|^\gamma \\ & = V_T^{l1}(u), \end{aligned} \quad (37)$$

for all  $u$ , and sufficiently large  $n$ , with  $\delta > 0$ . Note that the quadratic terms in  $V_T^{l1}(u)$  grow faster than  $|u|^\gamma$  term in (37), so  $\operatorname{argmin} V_T^{l1}(u) = O_p(1)$ . From that and  $V_T(u) \geq V_T^{l1}(u)$ , we have  $\operatorname{argmin} V_T(u) = O_p(1)$  in the nonstationary case ( $\rho_0 = 0$ ). Since  $u_0 = \operatorname{argmin} V_1(u)$  is unique in (36), we have the desired result,  $\operatorname{argmin} V_T(u) \xrightarrow{d} \operatorname{argmin} V_1(u)$ , when  $\rho_0 = 0$ .

**Proof of Theorem 2ii.** In this case note that by (33) and (35)

$$V_T(u) \xrightarrow{d} -2uL + u^2\Gamma_0 = V_2(u), \quad (38)$$

uniformly over  $u$ . Then we need to show that  $\operatorname{argmin} V_T(u) = O_p(1)$  when  $\rho_0 < 0$ . To see that

$$\begin{aligned} V_T(u) &\geq \left[ \sum_{t=1}^T \left( e_t - \frac{uy_{t-1}}{T^{1/2}} \right)^2 - e_t^2 \right] \\ &\quad - \frac{\lambda_T}{T^{1/2}} |u| |\rho_0|^{\gamma-1} \\ &\geq \left[ \sum_{t=1}^T \left( e_t - \frac{uy_{t-1}}{T^{1/2}} \right)^2 - e_t^2 \right] - \delta |u| |\rho_0|^{\gamma-1} \\ &= V_T^{l2}(u), \end{aligned} \quad (39)$$

for all  $u$ , and sufficiently large  $n$ , with  $\delta > 0$ . Note that the quadratic terms in  $V_T^{l2}(u)$  grow faster than  $|u|$  term in (39), so  $\operatorname{argmin} V_T^{l2}(u) = O_p(1)$ . From that and  $V_T(u) \geq V_T^{l2}(u)$ , we have  $\operatorname{argmin} V_T(u) = O_p(1)$  in the stationary case ( $\rho_0 < 0$ ). Since  $u_0 = \operatorname{argmin} V_2(u)$  is unique in (38), we have the desired result,  $\operatorname{argmin} V_T(u) \xrightarrow{d} \operatorname{argmin} V_2(u)$ , when  $\rho_0 < 0$ .

**Q.E.D.**

**Proof of Corollary 1.** If  $\rho_0 = 0$ , then the definition of convergence of  $V_T(u)$  to the limit should use the notion of epiconvergence as in Knight and Fu (2000), since this will be on extended real line rather than the compact set. By (36)

$$\sum_{t=2}^T \left( e_t - \frac{uy_{t-1}}{T} \right)^2 - e_t^2 \xrightarrow{d} -2u \int_0^1 W(r) dW(r) + u^2 \int_0^1 W(r)^2 dr,$$

Then (34) changes due to new penalty factor if  $u \neq 0$ ,

$$\lambda_T \left| \frac{u}{T} \right|^\gamma \rightarrow +\infty,$$

and it is 0 if  $u = 0$

$$\lambda_T \left| \frac{u}{T} \right|^\gamma = 0.$$

So if  $u \neq 0$  (epiconvergence)

$$V_T(u) \xrightarrow{d} +\infty,$$

So clearly in that case  $u = 0$  is the minimizer of the limit. **Q.E.D.**

**Proof of Theorem 3.** We start with the nonstationary case. First write the objective function (7) as

$$\sum_{t=2}^T (\Delta y_t - \theta' w_{t-1})^2 + \lambda_T |\theta_1|^{\gamma_1} + b_T \sum_{j=2}^p |\theta_j|^{\gamma_2},$$

where  $\theta = (\rho, \zeta)'$ ,  $\zeta = (\zeta_1, \dots, \zeta_{p-1})'$ , hence  $\theta_1 = \rho, \theta_2 = \zeta_1, \dots, \theta_p = \zeta_{p-1}$ , as well as  $w_{t-1} = (y_{t-1}, M'_{t-1})'$ . By  $\hat{\theta} = (\hat{\rho}, \hat{\zeta})'$  definition and using  $\theta_{10} = \rho_0 = 0$  under nonstationarity

$$\begin{aligned} \sum_{t=1}^T (\Delta y_t - \hat{\theta}' w_{t-1})^2 + \lambda_T |\hat{\theta}_1|^{\gamma_1} + b_T \sum_{j=2}^p |\hat{\theta}_j|^{\gamma_2} \\ \leq \sum_{t=2}^T (\Delta y_t - \theta'_0 w_{t-1})^2 + b_T \sum_{j=2}^p |\theta_{j0}|^{\gamma_2}. \end{aligned}$$

Set  $\eta_T = b_T \sum_{j=2}^p |\theta_{j0}|^{\gamma_2}$ , then we can simplify the above expression as

$$\begin{aligned} \eta_T &\geq \sum_{t=2}^T (\Delta y_t - \hat{\theta}' w_{t-1})^2 - \sum_{t=2}^T (\Delta y_t - \theta'_0 w_{t-1})^2 \\ &= \sum_{t=2}^T (w'_{t-1} (\hat{\theta} - \theta_0))^2 + 2 \sum_{t=2}^T e_t w'_{t-1} (\theta_0 - \hat{\theta}). \end{aligned} \quad (40)$$

Next set  $K_T = \Sigma_T^{-1/2} D_T^{-1} X'$ ,  $\delta_T = \Sigma_T^{1/2} D_T (\hat{\theta} - \theta_0)$ . Specifically

$$X = \begin{bmatrix} w'_1 \\ \vdots \\ w'_{t-1} \\ \vdots \\ w'_{T-1} \end{bmatrix},$$

$$\Sigma_T = \begin{bmatrix} \frac{\sum y_{t-1}^2}{T^2} & \frac{\sum y_{t-1} M'_{t-1}}{T^{3/2}} \\ \frac{\sum M_{t-1} y_{t-1}}{T^{3/2}} & \frac{\sum M_{t-1} M'_{t-1}}{T} \end{bmatrix},$$

$$D_T = \begin{bmatrix} T & 0'_{p-1} \\ 0_{p-1} & T^{1/2} I_{p-1} \end{bmatrix},$$

where  $X$  is  $(T-1) \times p$ , and the other two matrices are of  $p \times p$  dimension. Given the notation we can rewrite (40) as

$$\delta'_T \delta_T - 2(K_T e)' \delta_T - \eta_T \leq 0,$$

where  $e = (e_2, \dots, e_T)'$ . Next as in p.21 of Huang, Horowitz, Ma (2008) we have

$$\|\delta_T\|^2 \leq 6\|K_T e\|^2 + 3\eta_T. \quad (41)$$

See that

$$\begin{aligned}
E\|K_T e\|^2 &= \sigma^2 \text{tr}[K_T K_T'] \\
&= \sigma^2 \text{tr}[\Sigma_T^{-1/2} D_T^{-1} X' X D_T^{-1} \Sigma_T^{-1/2}] \\
&= \sigma^2 \text{tr} I_p = \sigma^2 p,
\end{aligned}$$

by  $D_T^{-1} X' X D_T^{-1} = \Sigma_T$  by  $X, D_T, \Sigma_T$  definitions above. So by (41) and  $\delta_T$  definition

$$E\|(\hat{\theta} - \theta_0)' D_T \Sigma_T D_T (\hat{\theta} - \theta_0)\| \leq 6\sigma^2 p + 3\eta_T. \quad (42)$$

First see that by Proposition 17.3c,h of Hamilton (1994)

$$\Sigma_T \xrightarrow{d} \begin{bmatrix} \int_0^1 W(r)^2 dr & 0'_{p-1} \\ 0_{p-1} & \Gamma_M \end{bmatrix}. \quad (43)$$

Note that  $\Gamma_M : p-1 \times p-1$  matrix and

$$\Gamma_M = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-2} \\ \cdots & \gamma_0 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \cdots & \cdots & \cdots & \gamma_0 \end{bmatrix},$$

where  $\gamma_0 = E\Delta y_{t-1}^2, \gamma_j = E\Delta y_{t-1} \Delta y_{t-1-j}, j = 1, \dots, p-2$ . Then by (43)(42), and  $\Sigma_T = O_p(1)$  and  $\eta_T = O(b_T)$ , with  $D_T$  definition

$$\hat{\rho} = O_p\left(\frac{\sqrt{b_T}}{T}\right), \|\hat{\zeta} - \zeta_0\| = O_p\left(\frac{\sqrt{b_T}}{\sqrt{T}}\right).$$

So with Assumption 5, both of them are consistent in the nonstationary case. **Q.E.D.**

Now we prove the consistency of the estimators in stationary case. This is basically the proof in Huang, Horowitz and Ma (2008). As before  $\theta_1 = \rho, \theta_2 = \zeta_1, \dots, \theta_p = \zeta_{p-1}$ . Notation in both proofs are the same. From the definition of  $\hat{\theta}$  and using the objective function we have

$$\begin{aligned}
\sum_{t=2}^T (\Delta y_t - \hat{\theta}' w_{t-1})^2 &+ \lambda_T |\hat{\theta}_1|^{\gamma_1} + b_T \sum_{j=2}^p |\hat{\theta}_j|^{\gamma_2} \\
&\leq \sum_{t=2}^T (\Delta y_t - \theta_0' w_{t-1})^2 \\
&+ \lambda_T |\theta_{10}|^{\gamma_1} + b_T \sum_{j=2}^p |\theta_{j0}|^{\gamma_2}.
\end{aligned}$$

Clearly,

$$\begin{aligned}
\sum_{t=2}^T (\Delta y_t - \hat{\theta}' w_{t-1})^2 &\leq \sum_{t=2}^T (\Delta y_t - \theta_0' w_{t-1})^2 \\
&+ \lambda_T |\theta_{10}|^{\gamma_1} + b_T \sum_{j=2}^p |\theta_{j0}|^{\gamma_2}.
\end{aligned}$$

Denote  $\eta_T = \lambda_T |\theta_{10}|^{\gamma_1} + b_T \sum_{j=2}^p |\theta_{j0}|^{\gamma_2}$ , and note the one difference between the nonstationary case and the stationary one: here we have an extra penalty term due to nonzero nature of  $\rho_0 = \theta_{10} < 0$  in the stationary case. So

$$\begin{aligned}\eta_T &\geq \sum_{t=2}^T (\Delta y_t - \hat{\theta}' w_{t-1})^2 - \sum_{t=2}^T (\Delta y_t - \theta_0' w_{t-1})^2 \\ &= \sum_{t=2}^T [w_{t-1}'(\hat{\theta} - \theta_0)]^2 + 2 \sum_{t=2}^T e_t w_{t-1}'(\theta_0 - \hat{\theta}).\end{aligned}$$

Set  $\delta_T = T^{1/2} \Sigma_T^{-1/2}(\hat{\theta} - \theta_0)$ , where

$$\Sigma_T = \frac{\sum_{t=2}^T w_{t-1} w_{t-1}'}{T},$$

a  $p \times p$  matrix. Then define  $K_T = T^{-1/2} \Sigma_T^{-1/2} W'$ , where  $W = (w_1', \dots, w_{T-1}')'$ . Rewrite the above equation as

$$\delta_T' \delta_T - 2[K_T e]' \delta_T - \eta_T \leq 0,$$

where  $e = (e_2, \dots, e_T)'$ . Then follow p.21 of Huang, Horowitz, and Ma (2008), and this can be simplified as

$$\|\delta_T\|^2 \leq 6\|K_T e\|^2 + 3\eta_T.$$

Clearly, by  $E e_i e_j = 0$  conditioned on  $y_{t-1}, \Delta y_{t-1}, \dots$  on all regressors. So

$$E\|K_T e\|^2 = \sigma^2 \text{tr}[K_T K_T'] = \sigma^2 p.$$

Since  $W'W/T = \Sigma_T$ , we have  $\text{tr}[K_T K_T'] = \text{tr}[\Sigma_T^{-1/2} \Sigma_T \Sigma_T^{-1/2}] = p$ . So

$$E\|\delta_T\|^2 = TE\|(\hat{\theta} - \theta_0)' \Sigma_T (\hat{\theta} - \theta_0)\|.$$

By stationarity of all regressors  $(y_{t-1}, M_{t-1}')$ , via law of large numbers through Proposition 17.3b,c of Hamilton (1994)  $\Sigma_T \rightarrow \Sigma$ ,  $\Sigma < \infty$ . . Then via

$$\begin{aligned}E\|\delta_T\|^2 &\leq 6\sigma^2 p + 3\eta_T, \\ \|\hat{\theta} - \theta_0\| &= O_p\left(\frac{\max(\lambda_T, b_T)^{1/2}}{T^{1/2}}\right) = o_p(1),\end{aligned}$$

when  $\max(\lambda_T, b_T) = o(T)$ . **Q.E.D.**

**Proof of Theorem 4i.** We start with the nonstationary case. We benefit from the following expression by using (7) as in (31)(32). Specifically

$$\begin{aligned}(\hat{\rho}, \hat{\zeta}') &= \underset{t=2}{\text{argmin}} \sum^T (\Delta y_t - \rho y_{t-1} - \zeta' M_{t-1}')^2 + \lambda_T |\rho|^{\gamma_1} + b_T \sum_{j=1}^{p-1} |\zeta_j|^{\gamma_2} \\ &= \underset{\rho, \zeta}{\text{argmin}} Z_T(\rho, \zeta).\end{aligned}\tag{44}$$

Then it is also true that

$$(\hat{\rho}, \hat{\zeta}') = \operatorname{argmin}[Z_T(\rho, \zeta) - Z_T(\rho_0, \zeta_0)]. \quad (45)$$

Reparametrizing  $u = T\rho, l = T^{1/2}(\zeta - \zeta_0)$  as

$$(\hat{u}, \hat{l}') = \operatorname{argmin}_{u,l} V_T(u, l). \quad (46)$$

$$\begin{aligned} V_T(u, l) &= \sum_{t=2}^T \left[ e_t - \frac{uy_{t-1}}{T} - \frac{l' M_{t-1}}{T^{1/2}} \right]^2 - \sum_{t=2}^T e_t^2 \\ &+ \lambda_T \left| \frac{u}{T} \right|^{\gamma_1} + b_T \sum_{j=1}^{p-1} \left| \zeta_{j0} + \frac{l_j}{T^{1/2}} \right|^{\gamma_2} - b_T \sum_{j=1}^{p-1} |\zeta_{j0}|^{\gamma_2}, \end{aligned} \quad (47)$$

where  $\hat{u} = T\hat{\rho}, \hat{l}' = (\hat{l}_1, \dots, \hat{l}_{p-1})' = T^{1/2}(\hat{\zeta}_1, \dots, \hat{\zeta}_{p-1} - \zeta_{10}, \dots, \zeta_{p-1,0})'$  minimize the function  $V_T(u, l)$  when  $\rho_0 = 0$ . So

$$(\hat{u}, \hat{l}') = \operatorname{argmin}_{u \in S, l \in K} V_T(u, l),$$

where  $S, K$  are suitable compact sets in  $R^1, R^{p-1}$ . These are compatible with Assumptions 3-4. We need to prove uniformly over  $u \times l \in S \times K$

$$V_T(u, l) \xrightarrow{d} V(u, l) = V_1(u) + V_2(l), \quad (48)$$

$$\hat{u} = O_p(1), \quad (49)$$

$$\hat{l}' = O_p(1). \quad (50)$$

We show first (48). Note that

$$\begin{aligned} \sum_{t=2}^T \left( e_t - \frac{uy_{t-1}}{T} - \frac{l' M_{t-1}}{T^{1/2}} \right)^2 &- \sum_{t=2}^T e_t^2 \\ &= u^2 \left[ \frac{\sum_{t=2}^T y_{t-1}^2}{T^2} \right] + l' \left[ \frac{\sum_{t=2}^T M_{t-1} M'_{t-1}}{T} \right] l \\ &- 2u \frac{\sum_{t=2}^T y_{t-1} e_t}{T} - 2l' \frac{\sum_{t=2}^T M_{t-1} e_t}{T^{1/2}} \\ &+ 2l' \left[ \frac{\sum_{t=2}^T M_{t-1} y_{t-1}}{T^{3/2}} \right] u. \end{aligned}$$

First see that uniformly over  $l \times u$

$$l' \left[ \frac{\sum_{t=2}^T M_{t-1} y_{t-1}}{T^{3/2}} \right] u \xrightarrow{p} 0, \quad (51)$$

by Proposition 17.3e in Hamilton (1994). Clearly, after that result, we can obtain the limits of  $\hat{u}$ , and  $\hat{l}$  asymptotically independent from each other. Note that by Proposition 17.3 d, h of Hamilton (1994) (uniformly over  $u$ )

$$u^2 \frac{\sum_{t=2}^T y_{t-1}^2}{T^2} \xrightarrow{d} u^2 \iota^2 \int_0^1 W(r)^2 dr,$$

$$u \frac{\sum_{t=2}^T y_{t-1} e_t}{T} \xrightarrow{d} \left(\frac{u}{2}\right) \sigma \iota [W(1)^2 - 1],$$

where  $\iota = \sigma \sum_{j=0}^{\infty} \psi_j$ , under  $\rho_0 = 0$ ,  $\Delta y_t = \Psi(L)e_t = \sum_{j=0}^{\infty} \psi_j e_{t-j}$  by (17.3.10) of Hamilton (1994). Note that  $W(1) \equiv N(0,1)$  which is the standard normal distribution. Next uniformly over  $l$

$$l' \left[ \frac{\sum_{t=2}^T M_{t-1} M'_{t-1}}{T} \right] l \xrightarrow{p} l' \Gamma_M l,$$

where  $\Gamma_M$  and its components are described in equation after (43).

Then by Proposition 17.3b of Hamilton (1994), uniformly over  $l$

$$l' \left[ \frac{\sum_{t=2}^T M_{t-1} e_t}{\sqrt{T}} \right] \xrightarrow{d} l' N(0, \Gamma_M).$$

Now we consider the penalty terms in (47). Given  $\lambda_T/T^{\gamma_1} \rightarrow \lambda_0$ , since  $\rho_0 = 0$ ,

$$\lambda_T \left| \frac{u}{T} \right|^{\gamma_1} = \frac{\lambda_T}{T^{\gamma_1}} |u|^{\gamma_1} \rightarrow \lambda_0 |u|^{\gamma_1}.$$

For the others, when  $\zeta_{j0} \neq 0$ ,

$$b_T \left( \sum_{j=1}^{p-1} \left| \zeta_{j0} + \frac{l_j}{T^{1/2}} \right|^{\gamma_2} - |\zeta_{j0}|^{\gamma_2} \right) \rightarrow 0,$$

by  $b_T = O(T^{\gamma_2/2})$  ( $0 < \gamma_2 < 1$ ). When  $\zeta_{j0} = 0$ ,  $b_T/T^{\gamma_2/2} \rightarrow b_0 \geq 0$ ,

$$b_T \sum_{j=1}^{p-1} \left| \frac{l_j}{T^{1/2}} \right|^{\gamma_2} \rightarrow b_0 \sum_{j=1}^{p-1} |l_j|^{\gamma_2}.$$

Combining the above results in (47)

$$V_T(u, l) \xrightarrow{d} V_1(u) + V_2(l),$$

where

$$\begin{aligned} & V_1(u) + V_2(l) \\ &= [u^2 \iota^2 \int_0^1 W(r)^2 dr - u \iota [W(1)^2 - 1] + \lambda_0 |u|^{\gamma_1}] \\ &+ [l' \Gamma_M l + l' N(0, \Gamma_M) + b_0 \sum_{j=1}^{p-1} |l_j|^{\gamma_2} 1_{\{\zeta_{j0}=0\}}]. \end{aligned} \tag{52}$$

$V_1(u)$  and  $V_2(l)$  show first and second square bracketed terms in (52) respectively. Next we want to show (49)(50). Define the following estimators:

$$\tilde{u} = \operatorname{argmin} V_{T1}(u),$$

where

$$V_{T1}(u) = \sum_{t=2}^T (e_t - \frac{uy_{t-1}}{T})^2 - e_t^2 + \lambda_T |\frac{u}{T}|^{\gamma_1}.$$

Also define

$$\tilde{l} = \operatorname{argmin} V_{T2}(l),$$

where

$$V_{T2}(l) = \sum_{t=1}^T (e_t - \frac{l' M_{t-1}}{T^{1/2}})^2 - \sum_{t=2}^T e_t^2 + b_T (\sum_{j=1}^{p-1} |\zeta_{j0} + \frac{l_j}{T^{1/2}}|^{\gamma_2} - \sum_{j=1}^{p-1} |\zeta_{j0}|^{\gamma_2}).$$

See that by (51)

$$\tilde{u} - \hat{u} \xrightarrow{p} 0,$$

$$\tilde{l} - \hat{l} \xrightarrow{p} 0.$$

But  $\tilde{u} = O_p(1)$  by Theorem 2i. Then  $\tilde{l} = O_p(1)$  can be shown easily by substituting  $M_{t-1}$  instead of  $y_{t-1}$  in (39). So  $\hat{u} = O_p(1), \hat{l} = O_p(1)$ . Then note that

$$u_o = \operatorname{argmin} V_1(u),$$

$$l_o = \operatorname{argmin} V_2(l),$$

are unique minimums. So we have the desired results. **Q.E.D**

**Proof of Theorem 4ii.** We use (7). Note that  $\theta = (\rho, \zeta)'$  and we can rewrite (7) as

$$\sum_{t=2}^T (\Delta y_t - w'_{t-1} \theta)^2 + \lambda_T |\theta_1|^{\gamma_1} + b_T \sum_{j=2}^p |\theta_j|^{\gamma_2},$$

where  $\theta_1 = \rho, \theta_2 = \zeta_1, \dots, \theta_p = \zeta_{p-1}, w_{t-1} = (y_{t-1}, M'_{t-1})'$ . For the stationary case here,  $\rho_0 < 0$  ( $\theta_{10} < 0$ ), and  $\hat{u} = T^{1/2}(\hat{\theta} - \theta_0)$  minimizes the objective function above. All the other coefficients can be zero or nonzero. This is the reason we have different rates on  $\theta_1$  than all the rest. Proceed in (31)(32) (or in (44)-(46))

$$\begin{aligned} V_T(u) &= \sum_{t=2}^T (e_t - \frac{u' w_{t-1}}{T^{1/2}})^2 - \sum_{t=2}^T e_t^2 \\ &+ \lambda_T (|\theta_{10} + \frac{u_1}{T^{1/2}}|^{\gamma_1} - |\theta_{10}|^{\gamma_1}) \\ &+ b_T (\sum_{j=2}^p |\theta_{j0} + \frac{u_j}{T^{1/2}}|^{\gamma_2} - |\theta_{j0}|^{\gamma_2}). \end{aligned} \tag{53}$$

We consider each one of the terms above,

$$\begin{aligned} \sum_{t=2}^T \left( e_t - \frac{u' w_{t-1}}{T^{1/2}} \right)^2 &= \sum_{t=2}^T e_t^2 \\ &= u' \left( \frac{\sum_{t=2}^T w_{t-1} w'_{t-1}}{T} \right) u - 2u' \left( \frac{\sum_{t=2}^T w_{t-1} e_t}{T^{1/2}} \right) \end{aligned}$$

Then by Law of Large numbers for stationary variables by Proposition 17.3 of Hamilton (1994)

$$\frac{\sum_{t=2}^T w_{t-1} w'_{t-1}}{T} \xrightarrow{p} \Gamma_W = E w_{t-1} w'_{t-1}.$$

Central Limit Theorem in Proposition 17.3b in Hamilton (1994) proves that

$$\frac{\sum_{t=2}^T w_{t-1} e_t}{T^{1/2}} \xrightarrow{d} N(0, \sigma^2 \Gamma_w).$$

We consider the penalty terms now. Given  $\lambda_T/T^{\gamma_1} \rightarrow \lambda_0 \geq 0$ ,  $0 < \gamma_1 < 1/2$ , since  $\theta_{10} = \rho_0 < 0$ ,

$$\lambda_T \left[ \left| \theta_{10} + \frac{u_1}{T^{1/2}} \right|^{\gamma_1} - |\theta_{10}|^{\gamma_1} \right] \rightarrow 0.$$

Next, with  $b_T = O(T^{\gamma_2/2})$ , for the remaining coefficients which are zero, for  $j = 2, \dots, p$ ,  $0 < \gamma_2 < 1$ ,

$$\begin{aligned} b_T \left( \left| \theta_{j0} + \frac{u_j}{T^{1/2}} \right|^{\gamma_2} - |\theta_{j0}|^{\gamma_2} \right) \\ = b_T \left| \frac{u_j}{T^{1/2}} \right|^{\gamma_2} \rightarrow b_0 |u_j|^{\gamma_2}. \end{aligned}$$

For the nonzero elements of  $\theta_2 = \dots = \theta_p$ , (for  $j = 2, \dots, p$ ), since  $b_T = O(T^{\gamma_2/2})$

$$b_T \left( \left| \theta_{j0} + \frac{u_j}{T^{1/2}} \right|^{\gamma_2} - |\theta_{j0}|^{\gamma_2} \right) \rightarrow 0.$$

Combining the above results in (53) we have, uniformly over  $u$

$$V_T(u) \xrightarrow{d} u' \Gamma_W u - 2u' N(0, \sigma^2 \Gamma_W) + b_0 \sum_{j=2}^p |u_j|^{\gamma_2} 1_{\{\theta_{j0}=0\}} = V(u).$$

Note that in the penalty term we do not have terms related to  $\theta_1 = \rho$  since  $\theta_{10} = \rho_0 < 0$ . Then clearly since  $y_{t-1}, M_{t-1}$  are stationary, by following the proof of Theorem 2ii

$$\hat{u} = \operatorname{argmin} V_T(u) = O_p(1).$$

Note that  $u_o = \operatorname{argmin} V(u)$  is the unique minimizer so

$$\hat{u} = \sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} \operatorname{argmin} V(u),$$

$$V(u) = u' \Gamma_W u - 2u' N(0, \sigma^2 \Gamma_W) + b_0 \sum_{j=2}^p |u_j|^{\gamma_2} 1_{\{\theta_{j0}=0\}}.$$

So the first element of  $\theta_0(\rho_0)$  is estimated as nonzero with standard limit and the all the other elements of  $\theta_0$  can be estimated as zero or nonzero with normal limit. **Q.E.D**

**Proof of Theorem 5.** We start the proof of consistency for the nonstationary  $y_t$  case. We start with the estimation of

$$\Delta y_t = \rho y_{t-1} + \alpha + \delta t + \zeta_1 \Delta y_{t-1} + \cdots + \zeta_{p-1} \Delta y_{t-(p-1)} + e_t.$$

Simple transformations in p.540 of Hamilton (1994) show that an identical equation is

$$\Delta y_t = \rho \xi_{t-1} + \mu^* + \delta^* t + \zeta_1 u_{t-1} + \cdots + \zeta_{p-1} u_{t-(p-1)} + e_t, \quad (54)$$

where

$$\begin{aligned} u_t &= \Delta y_t - \mu, \\ \mu &= \alpha / (1 - \zeta_1 - \cdots - \zeta_{p-1}), \\ \mu^* &= (1 - \rho)\mu. \end{aligned} \quad (55)$$

$$\begin{aligned} \xi_{t-1} &= y_{t-1} - \mu(t-1), \\ \delta^* &= \delta + \rho\mu, \end{aligned} \quad (56)$$

furthermore under  $\rho_0 = 0, \delta_0 = 0(\delta_0^* = 0)$ , without losing any generality by assigning  $y_0 = 0$ ,

$$u_t = e_t / (1 - \zeta_1 L - \cdots - \zeta_{p-1} L^{p-1}),$$

$$\xi_{t-1} = u_1 + \cdots + u_{t-1}.$$

To simplify the proofs we further need to rewrite our objective function in the following way

$$\sum_{t=2}^T (\Delta y_t - \theta' w_{t-1})^2 + \lambda_T |\theta_1|^{\gamma_1} + \nu_T |\theta_2|^{\gamma_2} + \tau_T |\theta_3|^{\gamma_3} + b_T \sum_{j=4}^{p+2} |\theta_j|^{\gamma_4}, \quad (57)$$

where  $\theta_1 = \rho, \theta_2 = \mu^*, \theta_3 = \delta^*, (\theta_4, \cdots, \theta_{p+2}) = (\zeta_1, \cdots, \zeta_{p-1})$ . Also  $w_{t-1} = (\xi_{t-1}, 1, t, M_{t-1}^*)'$ .  $M_{t-1}^* = (u_{t-1}, \cdots, u_{t-(p-1)})'$ . Use the definition of  $\hat{\theta}$  which is the minimizer of the objective function,

$$\begin{aligned} \sum_{t=2}^T (\Delta y_t - \hat{\theta}' w_{t-1})^2 &+ \lambda_T |\hat{\theta}_1|^{\gamma_1} + \nu_T |\hat{\theta}_2|^{\gamma_2} + \tau_T |\hat{\theta}_3|^{\gamma_3} + b_T \sum_{j=4}^{p+2} |\hat{\theta}_j|^{\gamma_4} \\ &\leq \sum_{t=2}^T (\Delta y_t - \theta_0' w_{t-1})^2 + \nu_T |\theta_{2,0}|^{\gamma_2} + b_T \sum_{j=4}^{p+2} |\theta_{j,0}|^{\gamma_4}. \end{aligned}$$

Set

$$\eta_T = \nu_T |\theta_{2,0}|^{\gamma_2} + b_T \sum_{j=4}^{p+2} |\theta_{j,0}|^{\gamma_4}. \quad (58)$$

Then

$$\begin{aligned}
\eta_T &\geq \sum_{t=2}^T (\Delta y_t - \hat{\theta}' w_{t-1})^2 - (\Delta y_t - \theta_0' w_{t-1})^2 \\
&= \sum_{t=2}^T (w'_{t-1} (\hat{\theta} - \theta_0))^2 + 2 \sum_{t=2}^T e_t w'_{t-1} (\theta_0 - \hat{\theta}).
\end{aligned} \tag{59}$$

Next set, for the nonstationary case

$$D_T = \begin{bmatrix} T & 0 & 0 & 0'_{p-1} \\ 0 & T^{1/2} & 0 & 0'_{p-1} \\ 0 & 0 & T^{3/2} & 0'_{p-1} \\ 0 & 0 & 0 & T^{1/2} I_{p-1} \end{bmatrix}.$$

$$\Sigma_T = \begin{bmatrix} \frac{\sum_{t=2}^T \xi_{t-1}^2}{T^2} & \frac{\sum_{t=2}^T \xi_{t-1}}{T^{3/2}} & \frac{\sum_{t=2}^T \xi_{t-1} t}{T^{5/2}} & \frac{\sum_{t=2}^T \xi_{t-1} M_{t-1}^*}{T^{3/2}} \\ \frac{\sum_{t=2}^T \xi_{t-1}}{T^{3/2}} & 1 & \frac{\sum_{t=2}^T t}{T^2} & \frac{\sum_{t=2}^T M_{t-1}^*}{T} \\ \frac{\sum_{t=2}^T t \xi_{t-1}}{T^{5/2}} & \frac{\sum_{t=2}^T t}{T^2} & \frac{\sum_{t=2}^T t^2}{T^3} & \frac{\sum_{t=2}^T t M_{t-1}^*}{T^2} \\ \frac{\sum_{t=2}^T M_{t-1}^* \xi_{t-1}}{T^{3/2}} & \frac{\sum_{t=2}^T M_{t-1}^*}{T} & \frac{\sum_{t=2}^T M_{t-1}^* t}{T^2} & \frac{\sum_{t=2}^T M_{t-1}^* M_{t-1}^*}{T} \end{bmatrix},$$

$X = [w_1, \dots, w_{T-1}]$ , where  $D_T, \Sigma_T$  are square matrices of  $p + 2$  dimension, and  $X$  is a  $(T - 1) \times (p + 2)$  dimensional matrix.

Rewrite (59) as

$$\delta_T' \delta_T - 2(K_T e)' \delta_T - \eta_T \leq 0,$$

where  $e = (e_2, \dots, e_T)$ ,  $K_T = \Sigma_T^{-1/2} D_T^{-1} X'$ ,  $\delta_T = \Sigma_T^{1/2} D_T (\hat{\theta} - \theta_0)$ . Then by p.21 of Huang, Horowitz, and Ma (2008)

$$\|\delta_T\|^2 \leq 6\|K_T e\|^2 + 3\eta_T, \tag{60}$$

See that by  $D_T^{-1} X' X D_T^{-1} = \Sigma_T$ , we have

$$\begin{aligned}
E\|K_T e\|^2 &= \text{tr}[K_T K_T'] \\
&= \sigma^2 \text{tr}[\Sigma_T^{-1/2} D_T^{-1} X' X' D_T^{-1} \Sigma_T^{-1/2}] \\
&= \sigma^2 \text{tr}[I_{p+2}] = \sigma^2(p + 2).
\end{aligned}$$

Then by

$$E\|(\hat{\theta} - \theta_0)' D_T \Sigma_T D_T (\hat{\theta} - \theta_0)\| \leq 6\sigma^2(p + 2) + 3\eta_T. \tag{61}$$

Then by proposition 17.3 of Hamilton (1994),

$$\Sigma_T \xrightarrow{d} \begin{bmatrix} \kappa^2 \int_0^1 W(r)^2 dr & \kappa \int_0^1 W(r) dr & \kappa \int_0^1 r W(r) dr & 0'_{p-1} \\ \kappa \int_0^1 W(r) dr & 1 & 1/2 & 0'_{p-1} \\ \kappa \int_0^1 r W(r) dr & 1/2 & 1/3 & 0'_{p-1} \\ 0_{p-1} & 0_{p-1} & 0_{p-1} & \Gamma_M \end{bmatrix} \equiv \Sigma, \tag{62}$$

$\kappa = \alpha/(1 - \zeta_1 - \dots - \zeta_{p-1})$ ,  $\gamma_j = Eu_t u_{t-j}$ ,  $j = 0, \dots, p-2$ .

$$\Gamma_M = \begin{bmatrix} \gamma_0 & \gamma_1 & \dots & \gamma_{p-2} \\ \dots & \gamma_0 & \dots & \dots \end{bmatrix}. \quad (63)$$

Then seeing that  $\eta_T = O(\max\{\nu_T, b_T\})$ , using  $D_T$  definition  $\hat{\rho} = O_p(\frac{\sqrt{\eta_T}}{T})$ ,  $\|\hat{\zeta} - \zeta_0\| = O_p(\frac{\sqrt{\eta_T}}{\sqrt{T}})$ ,  $\hat{\delta}^* = O_p(\frac{\sqrt{\eta_T}}{T^{3/2}})$ ,  $\hat{\mu}^* - \mu_0 = O_p(\frac{\sqrt{\eta_T}}{\sqrt{T}})$ . So by Assumption 5\*, and the transformations (55)(56) we obtain the consistency for the case of nonstationarity.

To show how the transformations provide the consistency for some of the estimators, we proceed in the following way. First, from the results above

$$\begin{aligned} \hat{\mu}^* - \mu_0^* &\xrightarrow{p} 0, \\ \hat{\delta}^* &\xrightarrow{p} 0. \end{aligned}$$

We know through (55)(56) that

$$\hat{\mu} = \hat{\mu}^*/(1 - \hat{\rho}) \xrightarrow{p} \mu_0^*,$$

since  $\rho_0 = 0$ , from (55)  $\mu_0 = \mu_0^*$ . So

$$\begin{aligned} \hat{\mu} &\xrightarrow{p} \mu_0. \\ \hat{\mu} &= \frac{\hat{\alpha}}{1 - \hat{\zeta}_1 - \dots - \hat{\zeta}_{p-1}}, \end{aligned}$$

Note that we also have for the original estimator  $\hat{\alpha}$

$$\hat{\alpha} = \hat{\mu}(1 - \hat{\zeta}_1 - \dots - \hat{\zeta}_{p-1}),$$

where we know that  $\hat{\mu} \xrightarrow{p} \mu_0$ , and  $1 - \hat{\zeta}_1 - \dots - \hat{\zeta}_{p-1} \xrightarrow{p} 1 - \zeta_{1,0} - \dots - \zeta_{p-1,0}$ . So  $\hat{\alpha} \xrightarrow{p} \alpha_0 = \mu_0(1 - \zeta_{1,0} - \dots - \zeta_{p-1,0})$ .

Next clearly, with  $\hat{\rho} \xrightarrow{p} 0$ , we have by the transformation in (56)

$$\hat{\delta}^* = (\hat{\delta} + \hat{\rho}\hat{\mu}) \xrightarrow{p} \delta_0 = 0. \quad (64)$$

Note that the original parameters and estimates, under the nonstationary of  $y_t$ , have the following property

$$\hat{\delta} - \hat{\delta}^* \xrightarrow{p} 0.$$

This is due to first equality in (64) and  $\hat{\rho} \xrightarrow{p} 0$ .

We now start the proof of consistency for the stationary case. This is very similar to nonstationary case. From (8) transform the variables and parameters similar to p.464, Chapter 16 of Hamilton (1994). Let

$$\delta^* = \delta(1 + \rho), \quad (65)$$

$$\mu^* = \alpha(1 + \rho) - \delta(\rho - \zeta_1 - \cdots - \zeta_{p-1}), \quad (66)$$

$y_{t-1}^* = y_{t-1} - \alpha - \delta(t-1)$ . So

$$\Delta y_t = \rho y_{t-1}^* + \zeta_1 \Delta y_{t-1}^* + \cdots + \zeta_{p-1} \Delta y_{t-(p-1)}^* + \mu^* + \delta^* t + e_t.$$

The objective function in transformed variables are

$$\sum_{t=2}^T (\Delta y_t - \theta' w_{t-1})^2 + \lambda_T |\theta_1|^{\gamma_1} + \iota_T |\theta_2|^{\gamma_2} + \tau_T |\theta_3|^{\gamma_3} + b_T \sum_{j=4}^{p+2} |\theta_j|^{\gamma_4},$$

where  $\theta_1 = \rho$ ,  $\theta_2 = \mu^*$ ,  $\theta_3 = \delta^*$ ,  $(\theta_4, \dots, \theta_{p+2}) = (\zeta_1, \dots, \zeta_{p-1})$ , and  $w_{t-1} = (y_{t-1}^*, 1, t, M_{t-1}^{*'})'$  and  $M_{t-1}^{*'} = (\Delta y_{t-1}^*, \dots, \Delta y_{t-(p-1)}^*)'$ . Then by definition of  $\hat{\theta}$  (with  $\rho_0 < 0$ ) (stationarity of  $y_t$ ) we have

$$\begin{aligned} \sum_{t=2}^T (\Delta y_t - \hat{\theta}' w_{t-1})^2 + \lambda_T |\hat{\theta}_1|^{\gamma_1} + \iota_T |\hat{\theta}_2|^{\gamma_2} + \tau_T |\hat{\theta}_3|^{\gamma_3} + b_T \sum_{j=4}^{p+2} |\hat{\theta}_j|^{\gamma_4} \\ \leq \sum_{t=2}^T (\Delta y_t - \theta'_0 w_{t-1})^2 + \lambda_T |\theta_{1,0}|^{\gamma_1} + \iota_T |\theta_{2,0}|^{\gamma_2} + \tau_T |\theta_{3,0}|^{\gamma_3} + b_T \sum_{j=4}^{p+2} |\theta_{j,0}|^{\gamma_4}. \end{aligned}$$

Then set , for the stationary case

$$D_T = \begin{bmatrix} T^{1/2} & 0 & 0 & 0'_{p-1} \\ 0 & T^{1/2} & 0 & 0'_{p-1} \\ 0 & 0 & T^{3/2} & 0'_{p-1} \\ 0 & 0 & 0 & T^{1/2} I_{p-1} \end{bmatrix}. \quad (67)$$

$$\Sigma_T = \begin{bmatrix} \frac{\sum_{t=2}^T y_{t-1}^{*2}}{T} & \frac{\sum_{t=2}^T y_{t-1}^*}{T} & \frac{\sum_{t=2}^T y_{t-1}^* t}{T^2} & \frac{\sum_{t=2}^T y_{t-1}^* M_{t-1}^{*'}}{T} \\ \frac{\sum_{t=2}^T y_{t-1}^*}{T} & 1 & \frac{\sum_{t=2}^T t}{T^2} & \frac{\sum_{t=2}^T M_{t-1}^{*'}}{T} \\ \frac{\sum_{t=2}^T t y_{t-1}^*}{T^2} & \frac{\sum_{t=2}^T t}{T^2} & \frac{\sum_{t=2}^T t^2}{T^3} & \frac{\sum_{t=2}^T t M_{t-1}^{*'}}{T^2} \\ \frac{\sum_{t=2}^T M_{t-1}^{*'} y_{t-1}^*}{T} & \frac{\sum_{t=2}^T M_{t-1}^{*'}}{T} & \frac{\sum_{t=2}^T M_{t-1}^{*'} t}{T^2} & \frac{\sum_{t=2}^T M_{t-1}^{*'} M_{t-1}^{*'}}{T} \end{bmatrix},$$

$X = [w_1, \dots, w_{T-1}]$ , where  $D_T, \Sigma_T$  are square matrices of  $p+2$  dimension, and  $X$  is a  $(T-1) \times (p+2)$  dimensional matrix.

$$\eta_T = \lambda_T |\theta_{1,0}|^{\gamma_1} + \iota_T |\theta_{2,0}|^{\gamma_2} + \tau_T |\theta_{3,0}|^{\gamma_3} + b_T \sum_{j=4}^{p+2} |\theta_{j,0}|^{\gamma_4}. \quad (68)$$

Then as in the nonstationary case

$$\begin{aligned} \eta_T &\geq \sum_{t=2}^T (\Delta y_t - \hat{\theta}' w_{t-1})^2 - (\Delta y_t - \theta'_0 w_{t-1})^2 \\ &= \sum_{t=2}^T (w'_{t-1} (\hat{\theta} - \theta_0))^2 + 2 \sum_{t=2}^T e_t w'_{t-1} (\theta_0 - \hat{\theta}). \end{aligned} \quad (69)$$

In order to simplify the last equation introduce  $\delta_T = \Sigma_T^{1/2} D_T (\hat{\theta} - \theta_0)$ ,  $K_T = \Sigma_T^{-1/2} D_T^{-1} X'$ ,

$$\delta_T' \delta_T - 2(K_T' e) \delta_T - \eta_T \leq 0,$$

As in p.21 of Huang, Horowitz and Ma (2008), also in the nonstationary case above

$$\|\delta_T\|^2 \leq 6\|K_T e\|^2 + 3\eta_T, \quad (70)$$

where exactly as in the nonstationary case, via  $D_T^{-1} X' X D_T^{-1} = \Sigma_T$  we find

$$E\|K_T e\|^2 = \sigma^2(p+2).$$

See that by Appendix 16.A of Hamilton (1994), or via law of large numbers for stationary variables

$$\Sigma_T \xrightarrow{p} \Sigma_M, \quad (71)$$

where

$$\Sigma_M = \begin{bmatrix} \gamma_0^* & 0 & 0 & \gamma^{*'} \\ 0 & 1 & 1/2 & 0'_{p-1} \\ 0 & 1/2 & 1/3 & 0'_{p-1} \\ \gamma^* & 0_{p-1} & 0_{p-1} & \Gamma_{p-1,p-1} \end{bmatrix},$$

where  $\gamma_0^* = E y_{t-1}^{*2}$ ,  $\gamma_j^* = E y_{t-1}^* y_{t-j-1}^*$  for  $j = 1, \dots, p$ ,  $\gamma^* = (\gamma_0^* - \gamma_1^*, \dots, \gamma_{p-2}^* - \gamma_{p-1}^*)'$ .  $\Gamma_{p-1,p-1} = E M_{t-1}^* M_{t-1}^{*'}$ . Clearly using the above equations we can write

$$E\|(\hat{\theta} - \theta_0)' D_T \Sigma_T D_T (\hat{\theta} - \theta_0)\| \leq 6\sigma^2(p+2) + 3\eta_T,$$

Then seeing that  $\eta_T = O(\max\{\lambda_T, \nu_T, \tau_T, b_T\})$ , using  $D_T$  definition in (67)  $\hat{\rho} - \rho_0 = O_p(\frac{\sqrt{\eta_T}}{\sqrt{T}})$ ,  $\|\hat{\zeta} - \zeta_0\| = O_p(\frac{\sqrt{\eta_T}}{\sqrt{T}})$ ,  $\hat{\delta}^* - \delta_0^* = O_p(\frac{\sqrt{\eta_T}}{T^{3/2}})$ ,  $\hat{\mu}^* - \mu_0^* = O_p(\frac{\sqrt{\eta_T}}{\sqrt{T}})$ . So by Assumption 5\*, and the transformations (65)(66) we obtain the consistency for the case of stationary data.

To show that how the transformations work in the stationary framework we proceed as follows. The transformations are

$$\mu^* = \alpha(1 + \rho) - \delta(\rho - \zeta_1 - \dots - \zeta_{p-1}),$$

$$\delta^* = \delta(1 + \rho).$$

We know from the results for the stationary case above that

$$\hat{\mu}^* \xrightarrow{p} \mu_0^*,$$

$$\hat{\delta}^* \xrightarrow{p} \delta_0^*.$$

So clearly,

$$\hat{\delta} = \frac{\hat{\delta}^*}{1 + \hat{\rho}} \xrightarrow{p} \frac{\delta_0^*}{1 + \rho_0} = \delta_0,$$

and also

$$\begin{aligned} \hat{\alpha} &= \frac{\hat{\delta}(\hat{\rho} - \hat{\zeta}_1 - \cdots - \hat{\zeta}_{p-1})}{1 + \hat{\rho}} + \frac{\hat{\mu}^*}{1 + \hat{\rho}} \\ &\xrightarrow{p} \frac{\delta_0(\rho_0 - \zeta_{1,0} - \cdots - \zeta_{p-1,0})}{1 + \rho_0} + \frac{\mu_0^*}{1 + \rho_0} = \alpha_0. \end{aligned}$$

So consistency of all parameters in all various setups have been shown. **Q.E.D.**

**Proof of Theorem 6i.** This is similar to proof of Theorem 4i. With  $\rho_0 = 0, \delta_0^* = 0$ , define the following as in (31)(32)

$$V_T(v, l) = \sum_{t=2}^T (e_t - v'z_{t-1,T} - l'M_{t-1,T})^2 - \sum_{t=2}^T e_t^2 \quad (72)$$

$$\begin{aligned} &+ \lambda_T \left| \frac{v_1}{T} \right|^{\gamma_1} + \left( \nu_T |\mu_0 + \frac{v_2}{\sqrt{T}}|^{\gamma_2} - \nu_T |\mu_0|^{\gamma_2} \right) \\ &+ \tau_T \left| \frac{v_3}{T^{3/2}} \right|^{\gamma_3} + b_T \sum_{j=1}^{p-1} \left| \zeta_{j0} + \frac{l_j}{\sqrt{T}} \right|^{\gamma_4} - |\zeta_{j0}|^{\gamma_4}, \end{aligned} \quad (73)$$

where  $v = (T\rho, T^{1/2}(\mu^* - \mu_0^*), T^{3/2}\delta^*)' = (v_1, v_2, v_3)'$ ,  $l = (\sqrt{T}(\zeta_1 - \zeta_{1,0}, \cdots, \zeta_{p-1} - \zeta_{p-1,0}))' = (l_1, \cdots, l_{p-1})'$ , and  $z_{t-1,T} = (\frac{\xi_{t-1}}{T}, \frac{1}{\sqrt{T}}, \frac{t}{T^{3/2}})'$ ,  $M_{t-1,T} = T^{-1/2}(u_1, \cdots, u_{p-1})'$ . Then  $\hat{v}, \hat{l}$  minimize (73).

$$(\hat{v}, \hat{l}) = \operatorname{argmin}_{v \in S, l \in K} V_T(v, l),$$

where  $S, K$  are compact subsets in  $R^3, R^{p-1}$ .

First we prove

$$V_T(v, l) \xrightarrow{d} V(v, l) = V_1(v) + V_2(l). \quad (74)$$

We define  $V(v, l), V_1(v), V_2(l)$  below in the proof.

To prove (74) we evaluate

$$\begin{aligned} \sum_{t=2}^T (e_t - v'z_{t-1,T} - l'M_{t-1,T})^2 &- \sum_{t=2}^T e_t^2 = \sum_{t=2}^T v'z_{t-1,T}z'_{t-1,T}v + \sum_{t=2}^T l'M_{t-1,T}M'_{t-1,T}l \\ &+ 2 \sum_{t=2}^T v'z_{t-1,T}M'_{t-1,T}l - 2 \sum_{t=2}^T v'z_{t-1,T}e_t - 2 \sum_{t=2}^T l'M_{t-1,T}e_t \end{aligned} \quad (75)$$

To analyze (75), by p.541 of Hamilton (1994) or via Proposition 17.3 of Hamilton (1994), uniformly over  $v$  and  $l$

$$v' \left[ \sum_{t=2}^T z_{t-1,T}z'_{t-1,T} \right] v \xrightarrow{d} v' \Sigma_W v, \quad (76)$$

where

$$\Sigma_W = \begin{bmatrix} \kappa^2 \int_0^1 W(r)^2 dr & \kappa \int_0^1 W(r) dr & \kappa \int_0^1 r W(r) dr \\ \kappa \int_0^1 W(r) dr & 1 & 1/2 \\ \kappa \int_0^1 r W(r) dr & 1/2 & 1/3 \end{bmatrix}.$$

Next again by Proposition 17.3, and  $\Gamma_M$  is defined in (63)

$$l' \sum_{t=2}^T M_{t-1,T} M'_{t-1,T} l \xrightarrow{p} l' \Gamma_M l, \quad (77)$$

As can be seen from the consistency proof here, so

$$v' \sum_{t=2}^T z_{t-1,T} M'_{t-1,T} l \xrightarrow{p} 0,$$

where we see that  $\hat{v}, \hat{l}$  are asymptotically independent. Next we see that via Proposition 17.3 of Hamilton (1994),

$$\begin{aligned} v' \sum_{t=2}^T z_{t-1,T} e_t &= v_1 \sum_{t=2}^T \frac{\xi_{t-1} e_t}{T} + v_2 \sum_{t=2}^T \frac{e_t}{\sqrt{T}} + v_3 \sum_{t=2}^T \frac{e_t t}{T^{3/2}} \\ &\xrightarrow{d} v_1 [1/2 \sigma \kappa [W(1)^2 - 1]] + v_2 [\sigma W(1)] + v_3 [\sigma [W(1) - \int_0^1 W(r) dr]] \\ &\equiv v' h_{2\omega}. \end{aligned} \quad (78)$$

Then also by p.541 of Hamilton (1994),

$$l' \sum_{t=2}^T M_{t-1,T} e_t \xrightarrow{d} l' N(0, \sigma^2 \Gamma_M). \quad (79)$$

From (73), we consider the penalty terms, first since  $\rho_0 = 0$ ,

$$\lambda_T \left| \frac{v_1}{T} \right|^{\gamma_1} \rightarrow \lambda_0 |v_1|^{\gamma_1},$$

with  $\lambda_T / T^{\gamma_1} \rightarrow \lambda_0$ . Then if  $\mu_0^* = 0$ , then

$$\iota_T \left| \frac{v_2}{\sqrt{T}} \right|^{\gamma_2} \rightarrow \iota |v_2|^{\gamma_2},$$

with  $\iota_T / T^{\gamma_2/2} \rightarrow \iota_0$ . If  $\mu_0^* \neq 0$ , by  $\iota_T / T^{\gamma_2/2} \rightarrow \iota_0$ , where  $0 < \gamma_2 < 1$ ,

$$\iota \left[ \left| \mu_0^* + \frac{v_2}{\sqrt{T}} \right|^{\gamma_2} - |\mu_0^*|^{\gamma_2} \right] \rightarrow 0.$$

Since  $\delta_0^* = \delta_0 = 0$ , by  $\tau_T / T^{3\gamma_3/2} \rightarrow \tau_0$ , we have

$$\tau_T \left| \frac{v_3}{T^{3/2}} \right|^{\gamma_3} \rightarrow \tau_0 |v_3|^{\gamma_3}.$$

The other penalty on lagged coefficients are proved in Theorem 4i, with  $b_T/T^{\gamma_2/2} \rightarrow b_0$ , we have

$$b_T \left( \sum_{j=1}^{p-1} \left| \zeta_{j0} + \frac{l_j}{\sqrt{T}} \right|^{\gamma_4} - |\zeta_{j0}|^{\gamma_4} \right) \rightarrow b_0 \sum_{j=1}^{p-1} |l_j|^{\gamma_4} 1_{\{\zeta_{j0}=0\}}.$$

Then we need  $\hat{v} = O_p(1), \hat{l} = O_p(1)$ . These follow from the proof of Theorem 4i with  $y_{t-1}$  is replaced by  $z_{t-1,T}$  for  $\hat{v}$  proof. The other proof is the same as in Theorem 4i. So we derive

$$V_T(v, l) \xrightarrow{d} V_1(v) + V_2(l),$$

and  $\hat{v} = \operatorname{argmin} V_1(v), \hat{l} = \operatorname{argmin} V_2(l)$ . The limit is

$$\hat{v} = \begin{pmatrix} T\hat{\rho} \\ T^{1/2}(\hat{\mu}^* - \mu_0^*) \\ T^{3/2}\hat{\delta}^* \end{pmatrix} \xrightarrow{d} \operatorname{argmin}_{v \in S} V_1(v),$$

where  $v = (v_1, v_2, v_3)'$ , and

$$V_1(v) = v' \Sigma_W v - 2v' h_{2\omega} + \lambda_0 |v_1|^{\gamma_1} + \iota_0 1_{\{\mu_0^*=0\}} |v_2|^{\gamma_2} + \tau_0 |v_3|^{\gamma_3},$$

and for

$$\hat{l} = \sqrt{T}(\hat{\zeta} - \zeta_0) \xrightarrow{d} l' \Gamma_M l - 2N(0, \sigma^2 \Gamma_M) + b_0 \sum_{j=1}^{p-1} |l_j|^{\gamma_4} 1_{\{\zeta_{j0}=0\}}.$$

**Q.E.D.**

**Proof of Theorem 6ii.** Now we go through the proof when there is stationarity ( $\rho_0 < 0$ ) as in (31)(32)

We use the following function to derive the limit law

$$\begin{aligned} V_T(v) &= \sum_{t=2}^T (e_t - v' z_{t-1,T})^2 - \sum_{t=2}^T e_t^2 \\ &+ \lambda_T (|\rho_0 + \frac{v_1}{\sqrt{T}}|^{\gamma_1} - |\rho_0|^{\gamma_1}) + \iota_T (\mu_0^* + \frac{v_2}{\sqrt{T}})^{\gamma_2} - |\mu_0^*|^{\gamma_2} \\ &+ \tau_T (|\delta_0^* + \frac{v_3}{T^{3/2}}|^{\gamma_3} - |\delta_0^*|^{\gamma_3}) + b_T \sum_{j=4}^{p+2} \left| \zeta_{j-3,0} + \frac{v_j}{\sqrt{T}} \right|^{\gamma_4} - |\zeta_{j-3,0}|^{\gamma_4}. \end{aligned}$$

This time we use

$$v = (\sqrt{T}(\rho - \rho_0), \sqrt{T}(\mu^* - \mu_0^*), T^{3/2}(\delta^* - \delta_0^*), \sqrt{T}(\zeta_1 - \zeta_{1,0}), \dots, \sqrt{T}(\zeta_{p-1} - \zeta_{p-1,0}))'.$$

$$z_{t-1,T} = \left( \frac{y_{t-1}^*}{\sqrt{T}}, \frac{1}{\sqrt{T}}, \frac{t}{T^{3/2}}, \frac{\Delta y_{t-1}^*}{\sqrt{T}}, \dots, \frac{\Delta y_{t-(p-1)}^*}{\sqrt{T}} \right)'.$$

We know that

$$\hat{v} = \operatorname{argmin}_{v \in S} V_T(v),$$

and we show that

$$V_T(v) \xrightarrow{d} V(v),$$

$$\hat{v} = O_p(1).$$

We define the limit  $V(v)$  later. To derive the above results we benefit from

$$\sum_{t=2}^T (e_t - v'z_{t-1,T})^2 - e_t^2 = v' \left( \sum_{t=2}^T z_{t-1,T} z'_{t-1,T} \right) v - 2v' \sum_{t=2}^T z_{t-1,T} e_t.$$

First see that as in the consistency proof (equation (71))

$$v' \left( \sum_{t=2}^T z_{t-1,T} z'_{t-1,T} \right) v \xrightarrow{p} \Sigma_M.$$

Then follow Appendix 16.A of Hamilton (1994) or simply using a central limit theorem

$$v' \sum_{t=2}^T z'_{t-1,T} e_t \xrightarrow{d} \sigma^2 N(0, \Sigma_M).$$

Then consider the penalty terms, with  $\lambda_T/T^{\gamma_1} \lambda_0$ ,  $0 < \gamma_1 < 1/2$ , when  $\lambda_T/\sqrt{T} \rightarrow 0$

$$\lambda_T (|\rho_0 + \frac{v_1}{\sqrt{T}}|^{\gamma_1} - |\rho_0|^{\gamma_1}) \rightarrow 0,$$

since  $\rho_0 < 0$ . Next, if  $\mu_0^* = 0$ , then with  $\iota_T/T^{\gamma_2/2} \rightarrow \iota_0$ ,  $0 < \gamma_2 < 1$ ,

$$\iota_T (|\frac{v_2}{\sqrt{T}}|^{\gamma_2}) \rightarrow \iota_0 |v_2|^{\gamma_2},$$

if  $\mu_0^* \neq 0$ , then due to  $\iota_T/T^{1/2} \rightarrow 0$ ,

$$\iota_T (|\mu_0^* + \frac{v_2}{\sqrt{T}}|^{\gamma_2} - |\mu_0^*|^{\gamma_2}) \rightarrow 0.$$

For the time trend variable, if  $\delta_0^* = 0$ , with  $0 < \gamma_3 < 2/3$ ,

$$\tau_T (|\frac{v_3}{T^{3/2}}|^{\gamma_3}) \rightarrow \tau_0 |v_3|^{\gamma_3}.$$

If  $\delta_0^* \neq 0$

$$\tau_T (|\delta_0^* + \frac{v_3}{T^{3/2}}|^{\gamma_3} - |\delta_0^*|^{\gamma_3}) \rightarrow 0,$$

with  $\tau_T/T^{3/2} \rightarrow 0$ . Next dynamic regressors behave exactly as in the same way in the nonstationary case so we have

$$V_T(v) \xrightarrow{d} V(v)$$

$$\equiv v' \Sigma_M v - 2\sigma^2 v' N(0, \Sigma_M) + \iota_0 |v_2|^{\gamma_2} 1_{\{\mu_0^*=0\}} + \tau_0 |v_3|^{\gamma_3} 1_{\{\delta_0^*=0\}} + b_0 \sum_{j=4}^{p+2} |v_j| 1_{\{\zeta_{j-3,0}=0\}}.$$

Next as in Theorem 2ii proof,  $\hat{v} = O_p(1)$ . This can be proved with (least squares estimators)  $\hat{v}_{LS} = O_p(1)$ , which is already in Appendix 16.A of Hamilton (1994). **Q.E.D.**

## REFERENCES

- BAI, J. AND S. NG (2004): "A Panic Attack on Unit Roots and Cointegration" *Econometrica*, 72, 1127-1177.
- BREIMAN, L. (1996): "Heuristics of instability and stabilization in model selection," *Annals of Statistics*, 24, 2350-2383.
- DICKEY, D. A. AND W. A. FULLER (1979): "Distribution of the estimators for autoregressive time series with a unit root," *Journal of The American Statistical Association*, 74, 427-431.
- CANER, M. (2009): "Lasso-type GMM estimator," *Econometric Theory*, 25, 1-21.
- CANER, M. AND B.E. HANSEN (2001): "Threshold Autoregression with a Unit Root," *Econometrica*, 69, 1555-1597.
- DONOHO, D.L. AND I.M. JOHNSTONE (1994): "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, 81, 425-455.
- ELLIOT, G., T.J. ROTHENBERG, AND J.H. STOCK (1996): "Efficient tests for an autoregressive unit root," *Econometrica*, 64, 813-836.
- FAN, J. AND R. LI (2001): "Variable selection via concave penalized likelihood and its oracle properties," *Journal of The American Statistical Association*, 96, 1348-1360.
- FAN, J. AND R. LI (2002): "Variable selection for Cox's proportional hazards model and frailty model," *Annals of Statistics*, 30, 74-99.
- HALL, A. (1994): "Testing for a unit root in time series with pretest data-based model selection," *Journal of Business and Economic Statistics*, 12, 461-470.
- HAMILTON, J. (1994): *Time Series Analysis*, Princeton University Press.
- HANSEN B.E. (2007): "Least squares model averaging," *Econometrica*, 75, 1175-1189.
- HANSEN, B.E. (2008): "Averaging estimators for autoregressions with a near unit root," Working Paper, Department of Economics. University of Wisconsin-Madison.
- HUANG, J., J.L. HOROWITZ, AND S. MA (2008): "Asymptotic properties of bridge estimators in sparse high-dimensional regression models," *Annals of Statistics*, 36, 587-613.
- KNIGHT, K. (2008): "Shrinkage estimation for nearly-singular designs," *Econometric Theory*, 24, 323-338.
- KNIGHT, K. AND W. FU (2000): "Asymptotics for lasso-type estimators," *Annals of Statistics*, 28, 1356-1378.
- LEEB, H. AND B. PÖTSCHER (2008): "Sparse estimators and the oracle property, or the return of Hodges' estimator," *Journal of Econometrics*, 142, 201-211.
- NG, S. AND P. PERRON (2001): "Lag length selection and the construction of unit root tests with good size and power," *Econometrica*, 69, 1519-1554.
- PHILLIPS, P.C.B. (1987): "Time Series Regression with a Unit Root" *Econometrica*, 55, 227-301.

VAN DER VAART, A., AND J. WELLNER (1996): *Weak Convergence and Empirical Processes*, Springer-Verlag.